

Krassimir Markov, Vladimir Ryazanov,  
Vitalii Velychko, Levon Aslanyan  
(editors)

**New Trends**  
**in**  
**Classification and Data Mining**

**I T H E A**  
**SOFIA**  
**2010**

**Krassimir Markov, Vladimir Ryazanov, Vitalii Velychko, Levon Aslanyan (ed.)**  
**New Trends in Classification and Data Mining**

ITHEA®

Sofia, Bulgaria, 2010

First edition

Recommended for publication by The Scientific Council of the Institute of Information Theories and Applications FOI ITHEA

This book maintains articles on actual problems of classification, data mining and forecasting as well as natural language processing:

- new approaches, models, algorithms and methods for classification, forecasting and clusterisation. Classification of non complete and noise data;
- discrete optimization in logic recognition algorithms construction, complexity, asymptotically optimal algorithms, mixed-integer problem of minimization of empirical risk, multi-objective linear integer programming problems;
- questions of complexity of some discrete optimization tasks and corresponding tasks of data analysis and pattern recognition;
- the algebraic approach for pattern recognition - problems of correct classification algorithms construction, logical correctors and resolvability of challenges of classification, construction of optimum algebraic correctors over sets of algorithms of computation of estimations, conditions of correct algorithms existence;
- regressions, restoring of dependences according to training sampling, parametrical approach for piecewise linear dependences restoration, and nonparametric regressions based on collective solution on set of tasks of recognition;
- multi-agent systems in knowledge discovery, collective evolutionary systems, advantages and disadvantages of synthetic data mining methods, intelligent search agent model realizing information extraction on ontological model of data mining methods;
- methods of search of logic regularities sets of classes and extraction of optimal subsets, construction of convex combination of associated predictors that minimizes mean error;
- algorithmic constructions in a model of recognizing the nearest neighbors in binary data sets, discrete isoperimetry problem solutions, logic-combinatorial scheme in high-throughput gene expression data;
- researches in area of neural network classifiers, and applications in finance field;
- text mining, automatic classification of scientific papers, information extraction from natural language texts, semantic text analysis, natural language processing.

It is represented that book articles will be interesting as experts in the field of classifying, data mining and forecasting, and to practical users from medicine, sociology, economy, chemistry, biology, and other areas.

General Sponsor: Consortium FOI Bulgaria ([www.foibg.com](http://www.foibg.com)).

Printed in Bulgaria

**Copyright © 2010 All rights reserved**

© 2010 ITHEA® – Publisher; Sofia, 1000, P.O.B. 775, Bulgaria. [www.ithea.org](http://www.ithea.org) ; e-mail: [info@foibg.com](mailto:info@foibg.com)

© 2010 Krassimir Markov, Vladimir Ryazanov, Vitalii Velychko, Levon Aslanyan – Editors

© 2010 Ina Markova – Technical editor

© 2010 For all authors in the book.

® ITHEA is a registered trade mark of FOI-COMMERCE Co.

**ISBN 978-954-16-0042-9**

© Jusaautor, Sofia, 2010

## PREFACE

ITHEA International Scientific Society (**ITHEA ISS**) is aimed to support growing collaboration between scientists from all over the world.

The scope of the books of the ITHEA ISS covers the area of Informatics and Computer Science.

ITHEA ISS welcomes scientific papers and books connected with any information theory or its application.

ITHEA ISS rules for preparing the manuscripts are compulsory.

ITHEA Publishing House is the official publisher of the works of the members of the ITHEA ISS.

Responsibility for papers and books published by ITHEA belongs to authors.

This book maintains articles on actual problems of classification, data mining and forecasting as well as natural language processing:

- new approaches, models, algorithms and methods for classification, forecasting and clusterisation. Classification of non complete and noise data;
- discrete optimization in logic recognition algorithms construction, complexity, asymptotically optimal algorithms, mixed-integer problem of minimization of empirical risk, multi-objective linear integer programming problems;
- questions of complexity of some discrete optimization tasks and corresponding tasks of data analysis and pattern recognition;
- the algebraic approach for pattern recognition - problems of correct classification algorithms construction, logical correctors and resolvability of challenges of classification, construction of optimum algebraic correctors over sets of algorithms of computation of estimations, conditions of correct algorithms existence;
- regressions, restoring of dependences according to training sampling, parametrical approach for piecewise linear dependences restoration, and nonparametric regressions based on collective solution on set of tasks of recognition;
- multi-agent systems in knowledge discovery, collective evolutionary systems, advantages and disadvantages of synthetic data mining methods, intelligent search agent model realizing information extraction on ontological model of data mining methods;
- methods of search of logic regularities sets of classes and extraction of optimal subsets, construction of convex combination of associated predictors that minimizes mean error;
- algorithmic constructions in a model of recognizing the nearest neighbors in binary data sets, discrete isoperimetry problem solutions, logic-combinatorial scheme in high-throughput gene expression data;
- researches in area of neural network classifiers, and applications in finance field;
- text mining, automatic classification of scientific papers, information extraction from natural language texts, semantic text analysis, natural language processing.

It is represented that book articles will be interesting as experts in the field of classifying, data mining and forecasting, and to practical users from medicine, sociology, economy, chemistry, biology, and other areas.

The book is recommended for publication by the Scientific Council of the Institute of Information Theories and Applications FOI ITHEA.

Papers in this book are selected from the ITHEA ISS Joint International Events of Informatics "ITA 2010":

<b>CFDM</b>	Second International Conference "Classification, Forecasting, Data Mining"
<b>i-i-i</b>	International Conference "Information - Interaction – Intellect"
<b>i.Tech</b>	Eight International Conference "Information Research and Applications"
<b>ISK</b>	V-th International Conference "Informatics in the Scientific Knowledge"
<b>MeL</b>	V-th International Conference "Modern (e-) Learning"
<b>KDS</b>	XVI-th International Conference "Knowledge - Dialogue – Solution"
<b>CML</b>	XII-th International Conference "Cognitive Modeling in Linguistics"
<b>INFOS</b>	Thirth International Conference "Intelligent Information and Engineering Systems"
<b>NIT</b>	International Conference "Natural Information Technologies"
<b>GIT</b>	Eight International Workshop on General Information Theory
<b>ISSI</b>	Forth International Summer School on Informatics

ITA 2010 took place in Bulgaria, Croatia, Poland, Spain and Ukraine. It has been organized by

ITHEA International Scientific Society

in collaboration with:

- ITHEA International Journal "Information Theories and Applications"
- ITHEA International Journal "Information Technologies and Knowledge"
- Institute of Information Theories and Applications FOI ITHEA
- Dorodnicyn Computing Centre of the Russian Academy of Sciences (Russia)
- Universidad Politecnica de Madrid (Spain)
- Institute of Linguistics, Russian Academy of Sciences (Russia)
- Association of Developers and Users of Intelligent Systems (Ukraine)
- V.M.Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine
- Institute of Mathematics and Informatics, BAS (Bulgaria)
- Institute of Mathematics of SD RAN (Russia)
- Taras Shevchenko National University of Kiev (Ukraine)
- New Bulgarian University (Bulgaria)
- The University of Zadar (Croatia)
- Rzeszow University of Technology (Poland)
- BenGurion University (Israel)
- University of Calgary (Canada)
- University of Hasselt (Belgium)
- Kharkiv National University of Radio Electronics (Ukraine)
- Kazan State University (Russia)
- Alexandru Ioan Cuza University (Romania)
- Moscow State Linguistic University (Russia)
- Astrakhan State University (Russia)

as well as many other scientific organizations. For more information: [www.ithea.org](http://www.ithea.org) .

We express our thanks to all authors, editors and collaborators as well as to the General Sponsor.

The great success of ITHEA International Journals, International Books and International Conferences belongs to the whole of the ITHEA International Scientific Society.

*Sofia – Moscow – Kiev - Erevan,*

*June 2010*

*K. Markov, V. Ryazanov, V. Velychko, L. Aslanyan*

## TABLE OF CONTENT

Preface .....	3
Table of Content.....	5
Index of Authors.....	7
<b>Minimization of Empirical Risk in Linear Classifier Problem</b>	
Yurii I. Zhuravlev, Yury Laptin, Alexander Vinogradov .....	9
<b>Restoring of Dependences on Samplings of Precedents with Usage of Models of Recognition</b>	
V.V.Ryazanov, Ju.I.Tkachev .....	17
<b>Composite Block Optimized Classification Data Structures</b>	
Levon Aslanyan, Hasmik Sahakyan .....	25
<b>Synthesis of Corrector Family with High Recognition Ability</b>	
Elena Djukova, Yurii Zhuravlev, Roman Sotnezov.....	32
<b>Methods for Evaluating of Regularities Systems Structure</b>	
Kostomarova Irina, Kuznetsova Anna, Malygina Natalia, Senko Oleg .....	40
<b>Growing Support Set Systems in Analysis of High-Throughput Gene Expression Data</b>	
Arsen Arakelyan, Anna Boyajian, Hasmik Sahakyan, Levon Aslanyan, Krassimira Ivanova, Iliya Mitov .....	47
<b>On the Complexity of Search for Conjunctive Rules in Recognition Problems</b>	
Elena Djukova, Vladimir Nefedov.....	54
<b>Optimal Forecasting Based on Convexcorrecting Procedures</b>	
Oleg Senko, Alexander Dokukin .....	62
<b>Reference-Neighbourhood Scalarization for Multiobjective Integer Linear Programming Problems</b>	
Krassimira Genova, Mariana Vassileva .....	73
<b>Multiagent Applications in Security Systems: New Proposals and Perspectives</b>	
Vladimir Jotsov .....	82

<b>Numeric-Lingual Distinguishing Features of Scientific Documents</b>	
Vladimir Lovitskii, Ina Markova, Krassimir Markov, Iliia Mitov.....	94
<b>Data and Metadata Exchange Repository Using Agents Implementation</b>	
Tetyana Shatovska, Iryna Kamenieva .....	102
<b>LSPL-Patterns as a Tool for Information Extraction From Natural Language Texts</b>	
Elena Bolshakova, Natalia Efremova, Alexey Noskov .....	110
<b>Computer Support of Semantic Text Analysis of a Technical Specification on Designing Software .....</b>	
Alla V. Zabolieva-Zotova, Yulia A. Orlova .....	119
<b>Linguistics Research and Analysis of the Bulgarian Folklore. Experimental Implementation of Linguistic Components in Bulgarian Folklore Digital Library</b>	
Konstantin Rangochev, Maxim Goynov, Desislava Paneva-Marinova, Detelin Luchev .....	131
<b>Natural Interface to Election Data</b>	
Elena Long, Vladimir Lovitskii, Michael Thrasher.....	138
<b>Analysis of Natural Language Objects</b>	
Oleksii Vasylenko .....	147
<b>Базовые структуры евклидовых пространств: конструктивные методы описания и использования</b>	
Владимир Донченко, Юрий Кривонос, Виктория Омардибирова.....	155
<b>Нейросетевая архитектура на частичных обученях</b>	
Николай Мурга .....	170
<b>Разработка автоматизированной процедуры совмещения изображений произведений живописи в видимом И рентгеновском спектральных диапазонах</b>	
Дмитрий Мурашов .....	185
<b>Распознавание объектов с неполной информацией и искаженных преобразованиями из заданной группы в рамках логико-предметной распознающей системы</b>	
Татьяна Косовская .....	196

---

**INDEX OF AUTHORS**

Alexander Dokukin	62	Mariana Vassileva	73
Alexander Vinogradov	9	Maxim Goynov	131
Alexey Noskov	110	Michael Thrasher	138
Alla Zaboлева-Zotova	119	Natalia Efremova	110
Anna Boyajian	47	Natalia Malygina	40
Anna Kuznetsova	40	Oleg Senko	40, 62
Arsen Arakelyan	47	Roman Sotnezov	32
Desislava Paneva-Marinova	131	Tetyana Shatovska	102
Detelin Luchev	131	V.V.Ryazanov	17
Elena Bolshakova	110	Vasylenko Oleksii	147
Elena Djukova	32, 54	Vladimir Jotsov	82
Elena Long	138	Vladimir Lovitskii	94, 138
Hasmik Sahakyan	25, 47	Vladimir Nefedov	54
Iliya Mitov	47, 94	Yulia Orlova	119
Ina Markova	94	Yurii Zhuravlev	9, 32
Irina Kostomarova	40	Yury Laptin	9
Iryna Kamenieva	102	Виктория Омардибирова	155
Ju.I.Tkachev	17	Владимир Донченко	155
Konstantin Rangochev	131	Дмитрий Мурашов	185
Krassimir Markov	94	Мурга Николай	170
Krassimira Genova	73	Татьяна Косовская	196
Krassimira Ivanova	47	Юрий Кривонос	155
Levon Aslanyan	25, 47		





## MINIMIZATION OF EMPIRICAL RISK IN LINEAR CLASSIFIER PROBLEM

Yurii I. Zhuravlev, Yury Laptin, Alexander Vinogradov

**Abstract:** Mixed-integer formulation of the problem of minimization of empirical risk is considered. Some possibilities of decision of the continuous relaxation of this problem are analyzed. Comparison of the proposed continuous relaxation with a similar SVM problem is performed too.

**Keywords:** cluster, decision rule, discriminant function, linear and non-linear programming, non-smooth optimization

**ACM Classification Keywords:** G.1.6 Optimization - Gradient methods, I.5 Pattern Recognition; I.5.2 Design Methodology - Classifier design and evaluation

**Acknowledgement:** This work was done in the framework of Joint project of the National Academy of Sciences of Ukraine and the Russian Foundation for Fundamental Research No 08-01-90427 'Methods of automatic intellectual data analysis in tasks of recognition objects with complex relations'.

---

### Introduction

Recently considerable number of researches are devoted to problems of construction of linear algorithms of classification (classifiers). In many cases such problems are considered for classification of two sets. Usually linear classifier problems are formulated for the case of linearly separable sets. In separable case the mentioned problems can be efficiently solved [1–4]. The concept of optimality for two linearly separable sets has a simple geometrical sense – the optimum classifier defines the strip of maximal width separating these sets.

For linear separability of two finite sets it is necessary and sufficient for convex envelopes of these sets don't intersect each other. But this condition is not sufficient in the case of more than two sets. In [5–7] some sufficient conditions of linear separability of any number of finite sets are formulated.

Minimization of the empirical risk is the natural criterion of choice of the classifier in case of linearly inseparable sets. In this paper, a mixed-integer formulation of the problem of minimization of empirical risk is considered, and some possibilities of decision of the continuous relaxation of this problem are analyzed. Comparison of the proposed continuous relaxation with a similar SVM problem is performed too.

---

### 1. Problem formulation

Let a set of linear functions is defined  $f_i(x, W^i) = (w^i, x) + w_0^i$ ,  $i = 1, \dots, m$ , where  $x \in R^n$  is attribute vector, and  $W^i = (w_0^i, w^i) \in R^{n+1}$ ,  $i = 1, \dots, m$ , are vectors of parameters. We denote  $W = (W^1, \dots, W^m)$ ,  $W \in R^L$ ,  $L = m(n+1)$ . Let's consider linear algorithms of classification (linear classifiers) of the following kind

$$a(x, W) = \arg \max_i \left\{ f_i(x, W^i) : i = 1, \dots, m \right\}; x \in R^n; W \in R^L \quad (1)$$

In [6] also classifiers, in which  $f_i$  are convex piece-wise linear functions, were investigated.

Here it is considered a family of finite not intersected sets  $\Omega_i, i=1, \dots, m$ . We will say that the classifier  $a(x, W)$  separates correctly points from  $\Omega_i, i=1, \dots, m$ , if  $a(x, W) = i$  for all  $x \in \Omega_i, i=1, \dots, m$ .

Sets  $\Omega_i, i=1, \dots, m$  are called *linearly separable* if there is a linear classifier correctly separating points from these sets.

Each set  $\Omega_i, i=1, \dots, m$  is a training sample of points from some class  $\bar{\Omega}_i$  known only on these sample units. The training process for classifier  $a(x, W)$  consists in selection of parameters  $W$  at which classes  $\bar{\Omega}_i, i=1, \dots, m$  are separated in the best way (in some sense). For definition of the quality of separation various approaches are used.

Let  $\Omega = \bigcup_{i=1}^m \Omega_i$ , points of the set  $\Omega$  are enumerated,  $T$  is the set of indices,  $\Omega = \{x^t : t \in T\}$ ,  $T_i$  is a

subset of indices corresponding to points from  $\Omega_i$ ,  $\Omega_i = \{x^t : t \in T_i\}$ ,  $T = \bigcup_{i=1}^m T_i$ . Let function  $i(t)$  returns

the index of the set  $\Omega_i$ , to which the point  $x^t$  belongs,  $t \in T$ . The value

$$\begin{aligned} g^t(W) &= \min \left\{ f_i(x^t, W^i) - f_j(x^t, W^j) : j \in \{1, \dots, m\} \setminus i, i = i(t) \right\} = \\ &= \min \left\{ (w^i - w^j, x^t) + w_0^i - w_0^j : j \in \{1, \dots, m\} \setminus i, i = i(t) \right\} \end{aligned} \quad (2)$$

is called as a *margin* or a *gap* of the classifier  $a(x, W)$  on the point  $x^t, t \in T$ .

The classifier  $a(x, W)$  makes a mistake in a point  $x^t$  iff the gap  $g^t(W)$  is negative.

The value  $g(W) = \min \{g^t(W) : t \in T\}$  is called as a gap of the classifier  $a(x, W)$  on the family of sets  $\Omega_i, i=1, \dots, m$ .

The classifier  $a(x, W)$  correctly separates points from  $\Omega_i, i=1, \dots, m$ , if  $g(W) > 0$ .

**Remark 1.** The classifier  $a(x, W)$  is invariant with respect to multiplication of all functions  $f_i$  (vectors  $W^i$ ) by positive number, and the gap  $g(W)$  is linear with respect to such multiplication. The classifier  $a(x, W)$  and the gap  $g(W)$  are invariant concerning to addition of any real number to all  $f_i$ .

The value  $g(W)$  can be used as a criterion of quality of the classifier  $a(x, W)$  (the more value  $g(W)$ , the more reliably points from  $\Omega_i, i=1, \dots, m$  are separated). However, it is necessary to take into account a norm for the family of vectors  $W$  which we denote  $\eta(W)$  and name *norm* of the classifier  $a(x, W)$ .

Let's use the following function:

$$\eta(W) = \sqrt{\sum_{i=1}^m \sum_{k=1}^n (w_k^i)^2} \quad (3)$$

Other functions also can be used as a norm  $\eta(W)$  [6].

Let the family of sets  $\Omega_i, i = 1, \dots, m$  is given. Taking into account the introduced notations the optimal classifier problem we write as following: find

$$g^* = \max_W \{g(W) : \eta(W) \leq 1, W \in R^L\} \quad (4)$$

Since the vector  $W = 0$  is feasible, the problem (4) always has a solution, and  $g^* \geq g(0) = 0$ . Let's notice that  $g^* > 0$  if sets  $\Omega_i, i = 1, \dots, m$  are linearly separable, i.e. there is the linear classifier correctly separating these sets. We will consider also a problem: find

$$\eta^* = \min_V \{\eta(V) : g(V) \geq 1, V \in R^L\} \quad (5)$$

Similar problems were considered by different authors (see, e.g., [4, 8]).

**Lemma 1.** Let  $W^*$  be an optimal solution to the problem (4). Then

1) if  $g^* > 0$ , the problem (5) also has the optimum solution  $V^*$ , and  $V^* = \frac{W^*}{g^*}, \eta^* = \frac{1}{g^*}$ ;

2) if  $g^* = 0$ , the problem (5) has no feasible solutions.

The proof is simple (see [6]).

Let's consider in more details problems of construction of linear classifiers for the family of sets  $\Omega_i = \{x^t, t \in T_i\}, i = 1, \dots, m$ . It is easy to see that the problem (4) can be represented as a LP- problem with additional quadratic constraint: find

$$g^* = \max_{w, \delta} \delta \quad (6)$$

subject to

$$(w^i - w^j, x^t) + w_0^i - w_0^j \geq \delta, \quad j \in \{1, \dots, m\} \setminus i, t \in T_i, i = 1, \dots, m \quad (7)$$

$$\sum_{i=1}^m \sum_{k=1}^n (w_k^i)^2 \leq 1 \quad (8)$$

The problem (5) is a quadratic programming problem: find

$$\eta^* = \min_v \sum_{i=1}^m \sum_{k=1}^n (v_k^i)^2 \quad (9)$$

subject to

$$(v^i - v^j, x^t) + v_0^i - v_0^j \geq 1, \quad j \in \{1, \dots, m\} \setminus i, t \in T_i, i = 1, \dots, m \quad (10)$$

It is possible to show that in case  $m = 2$  the problem (9) – (10) is equivalent to the problem which is used for construction of the strip of the maximum width separating some linearly separable sets  $\Omega_1, \Omega_2$ .

Existing efficient software packages for optimization problems of general purpose can be used for considered problems, if the number of points in training sample is not too large [6]. For a large number of points in training sample, it is appropriate to use non-smooth optimization methods [8, 9].

Problems (6) – (8) and (9) – (10) allow to find the optimum linear classifier only for linearly separable sets. For linearly inseparable sets the problem should be formulated in other way.

---

## 2. Empirical risk minimization

---

In the case of linearly inseparable training sample a natural criterion for choosing classifier is that of minimizing empirical risk, i.e. the number of training sample points which are separated by the classifier incorrectly.

Suppose that a reliability parameter  $\bar{\delta} > 0$  is fixed for separation of points of the training sample  $\Omega_i, i = 1, \dots, m$ . We say that the points  $x^t, t \in T$  are separated by the classifier unreliably, if  $g^t(W) < \bar{\delta}$ .

Below the value of empirical risk will be determined by reliability, characterized by parameter  $\bar{\delta}$ , i.e. the empirical risk is equal to the number of points of the training sample, which are separated by the classifier incorrectly or unreliably.

**Lemma 3 [6].** Let  $x^\alpha \in \Omega_i, x^\beta \in \Omega_j$ , classifier  $a(x, W)$  separates these points correctly, and for the norm of the classifier the constraint (8) is valid. Then

$$-R \leq w_0^i - w_0^j \leq R \quad (11)$$

where  $R = \max \{ \|x\| : x \in \Omega_i, i = 1, \dots, m \}$ .

Let  $\Omega_i = \{x^t, t \in T_i\}, i = 1, \dots, m, T = \bigcup_{i=1}^m T_i$ . To each point  $x^t, t \in T$  we associate a variable  $y_t = 0 \vee 1$

so that  $y_t = 0$ , if the point  $x^t$  is considered in the problem (6)–(8), and  $y_t = 1$  – otherwise.

Let a large positive number  $B$  be given. Empirical risk minimization problem based on reliability parameter  $\bar{\delta}$  has the following form: find

$$Q^* = \min_{w, y} \left\{ \sum_{t \in T} y_t \right\} \quad (12)$$

subject to

$$(w^i - w^j, x^t) + w_0^i - w_0^j \geq \bar{\delta} - B \cdot y_t, \quad j \in \{1, \dots, m\} \setminus i, \quad t \in T_i, \quad i = 1, \dots, m \quad (13)$$

$$\eta(W) \leq 1 \quad (14)$$

$$\sum_{t \in T_i} y_t \leq |T_i| - 1, \quad i = 1, \dots, m \quad (15)$$

$$0 \leq y_t \leq 1, \quad t \in T \quad (16)$$

$$y_t = 0 \vee 1, \quad t \in T \quad (17)$$

From (14), (11) follows that if  $y_t = 1$ , then for sufficiently large values  $B$  the corresponding inequalities of the form (13) are always valid, i.e. point  $x^t$  is excluded from the problem. Constraints (15) mean that at least one point from each set  $\Omega_i$  must be included in the problem.

The optimal value  $Q^*$  is equal to the minimum empirical risk based on reliability  $\bar{\delta}$ . Problem (12)-(17) is *NP*-hard; the branch and bound method can be used to solve it. To calculate the lower bounds for  $Q^*$  (minimum empirical risk), let's consider the continuous relaxation of the mentioned above problem – the problem (12)–(16). The optimum value of the relaxed problem is denoted  $q^*$ . To solve this problem we use decomposition on the variables  $W$ . Let variables  $W$  are fixed. Given (2), the problem of minimizing on the variables  $y$  takes the following form: find

$$q(W) = \min_y \left\{ \sum_{t \in T} y_t \right\} \quad (18)$$

subject to

$$y_t \geq \frac{1}{B} \left( \bar{\delta} - g^t(W) \right), t \in T \quad (19)$$

$$\eta(W) \leq 1 \quad (20)$$

$$\sum_{t \in T_i} y_t \leq |T_i| - 1, i = 1, \dots, m \quad (21)$$

$$0 \leq y_t \leq 1, t \in T \quad (22)$$

Denote  $d^t(W) = \max \left( 0, \frac{1}{B} \left( \bar{\delta} - g^t(W) \right) \right)$ . Obviously, if the problem (18)-(22) has a solution, then

$y^t = d^t(W)$ . So, we get the minimization problem on variables  $W$ : find

$$q^* = \min \sum_{t \in T} d^t(W) \quad (23)$$

subject to

$$\eta(W) \leq 1 \quad (24)$$

$$\sum_{t \in T_i} d^t(W) \leq |T_i| - 1, i = 1, \dots, m \quad (25)$$

$$d^t(W) \leq 1, t \in T \quad (26)$$

Functions  $d^t(W)$  are convex piecewise-linear,  $\eta(W)$  is quadratic and positively defined. To solve the problem (23)-(26) it is appropriate to apply efficient methods of nonsmooth optimization [9].

### 3. Comparison with support vector machine

Let us consider the case of two classes. Suppose, as previously,  $\Omega_i = \{x^t, t \in T_i\}$ ,  $i = 1, 2$ ,  $T = T_1 \cup T_2$ . In the method of support vectors (see eg [1]) to build a classifier which separates the two linearly inseparable sets, one has to solve the following problem: find

$$\min_{u, u_0, \xi} \left\{ \frac{1}{2} (u, u) + C \cdot \sum_{t \in T} \xi^t \right\} \quad (27)$$

subject to

$$(u, x^t) + u_0 \geq 1 - \xi^t, \quad t \in T_1 \quad (28)$$

$$(-u, x^t) - u_0 \geq 1 - \xi^t, \quad t \in T_2 \quad (29)$$

$$\xi^t \geq 0, \quad t \in T \quad (30)$$

where  $u \in R^n$ ,  $u_0 \in R$ ,  $\xi^t \in R$ ,  $t \in T$ .

To compare these approaches we consider an analogue of (12)–(16) for the case of two sets (in the case of two sets  $\Omega_i = \{x^t, t \in T_i\}$ ,  $i = 1, 2$  to build a linear classifier we need only two functions  $f_i(x, W^i) = (w^i, x) + w_0^i$ ,  $i = 1, 2$ , where  $f_1(x) = -f_2(x)$ ): find

$$q^* = \min_{w, w_0, y} \left\{ \sum_{t \in T} y_t \right\} \quad (31)$$

subject to

$$(w, x^t) + w_0 \geq \bar{\delta} - B \cdot y_t, \quad t \in T_1 \quad (32)$$

$$(-w, x^t) - w_0 \geq \bar{\delta} - B \cdot y_t, \quad t \in T_2 \quad (33)$$

$$(w, w) \leq 1 \quad (34)$$

$$\sum_{t \in T_i} y_t \leq |T_i| - 1, \quad i = 1, 2 \quad (35)$$

$$0 \leq y_t \leq 1, \quad t \in T \quad (36)$$

Change of variables in the problem (31)–(36):  $w = \bar{\delta}u$ ,  $w_0 = \bar{\delta}u_0$ ,  $\xi^t = \frac{By_t}{\bar{\delta}}$ ,  $t \in T_1 \cup T_2$ , gives

$$q^* = \frac{\bar{\delta}}{B} \cdot \min_{u, u_0, \xi} \left\{ \sum_{t \in T} \xi^t \right\} \quad (37)$$

subject to

$$(u, x^t) + u_0 \geq 1 - \xi^t, \quad t \in T_1 \quad (38)$$

$$(-u, x^t) - u_0 \geq 1 - \xi^t, \quad t \in T_2 \quad (39)$$

$$(u, u) \leq \frac{1}{\bar{\delta}^2} \quad (40)$$

$$\xi^t \geq 0, t \in T \quad (41)$$

$$\xi^t \leq \frac{B}{\bar{\delta}}, t \in T \quad (42)$$

$$\sum_{t \in T_i} \xi^t \leq \frac{B}{\bar{\delta}} (|T_i| - 1), i = 1, 2 \quad (43)$$

Denote  $\chi, \gamma_i, i = 1, 2$  the dual variables for constraints (40), (43) and consider the Lagrangian

$$L(\chi, \gamma, \xi, u) = \frac{\bar{\delta}}{B} \sum_{t \in T} \xi^t + \chi \cdot ((u, u) - \frac{1}{\bar{\delta}^2}) + \sum_{i=1}^2 \gamma_i \left( \sum_{t \in T_i} \xi^t - \frac{B}{\bar{\delta}} (|T_i| - 1) \right)$$

Let:

$$\varphi(\chi, \gamma) = \min_{u, u_0, \xi} L(\chi, \gamma, \xi, u) \quad (44)$$

subject to (38), (39), (41), (42).

Suppose a penalty factor  $C$  in the problem (27)-(30) is given. It is easy to see that, if we take  $\gamma = 0$  and choose

$\chi$  from the condition  $\frac{\bar{\delta}}{2\chi B} = C$ , we obtain

$$L(\chi, \gamma, \xi, u) = 2\chi \left\{ \frac{1}{2} (u, u) + C \cdot \sum_{t \in T} \xi^t \right\} - \frac{\chi}{\bar{\delta}^2}.$$

So, the problem (44), (38), (39), (41) is equivalent to (27)-(30) for the dual variables chosen above. Constraints (42) can be neglected at small  $\bar{\delta}$  and large  $B$ .

Thus, the SVM problem is a special case of (44), (38), (39), (41).

## References

1. Воронцов К.В. Машинное обучение. – [http://www.machinelearning.ru/wiki/index.php?title=Машинное\\_обучение\\_\(курс\\_лекций%2C\\_К.В.Воронцов\)](http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_(курс_лекций%2C_К.В.Воронцов)) – Последнее изменение: 30 мая 2009
2. Местецкий Л.М. Математические методы распознавания образов. – <http://www.intuit.ru/department/graphics/imageproc/> – Опубликовано 30.04.2008
3. Гупал А.М., Сергиенко И.В. Оптимальные процедуры распознавания. - Киев: Наук.думка, 2008. - 232 с.
4. Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию. – К.: Наукова думка, 2004. – 545 с.
5. Laptin Yu., Vinogradov A. Exact discriminant function design using some optimization techniques // "Classification, Forecasting, Data Mining" International Book Series "INFORMATION SCIENCE & COMPUTING", Number 8, Sofia, Bulgaria, 2009. – Pages 14-19.
6. Лаптин Ю.П., Виноградов А.П., Лиховид А.П. О некоторых подходах к проблеме построения линейных классификаторов в случае многих классов // Pattern Recognition and Image Analysis, 2010 (Сдана в печать)
7. Петунин Ю.И., Шульдешов Г.А. Проблемы распознавания образов с помощью линейных дискриминантных функций Фишера – Кибернетика, 1979, № 6, с. 134-137.
8. Методи негладкої оптимізації у спеціальних задачах класифікації, Стецюк П.І., Березовський О.А., Журбенко М.Г., Кропотов Д.О. – Київ, 2009. – 28 с. – (Препр./НАН України. Ін-т кібернетики ім. В.М.Глушкова; 2009–1)

9. Shor N.Z. Nondifferentiable Optimization and Polynomial Problems. – Dordrecht, Kluwer, 1998. – 394 p.
10. Koel Das, Zoran Nenadic. An efficient discriminant-based solution for small sample size problem // Pattern Recognition – Volume 42, Issue 5, 2009, Pages 857-866.
11. Juliang Zhang, Yong Shi, Peng Zhang. Several multi-criteria programming methods for classification // Computers & Operations Research – Volume 36, Issue 3, 2009, Pages 823-836.
12. E. Dogantekin, A. Dogantekin, D. Avci Automatic Hepatitis Diagnosis System based on Linear Discriminant Analysis and Adaptive Network Based Fuzzy Inference System // Expert Systems with Applications, In Press, 2009.

---

**Information about authors**

---

**Yurii I. Zhuravlev** – Academician, Deputy Director, Dorodnicyn Computing Centre of the RAS, Vavilova 40, 119333 Moscow, Russian Federation

**Yury Laptin** – Senior Researcher, V.M.Glushkov Institute of Cybernetics of the NASU, Prospekt Akademika Glushkova, 40, 03650 Kyiv, Ukraine; e-mail: [laptin\\_yu\\_p@mail.ru](mailto:laptin_yu_p@mail.ru)

**Alexander Vinogradov** – Senior Researcher, Dorodnicyn Computing Centre of the RAS, Vavilova 40, 119333 Moscow, Russian Federation; e-mail: [vngccas@mail.ru](mailto:vngccas@mail.ru)



---

## RESTORING OF DEPENDENCES ON SAMPLINGS OF PRECEDENTS WITH USAGE OF MODELS OF RECOGNITION

V.V.Ryazanov, Ju.I.Tkachev

**Abstract.** Two approaches to solution of the task of restoring of dependence between a vector of independent variables and a dependent scalar according to training sampling are considered. The first (parametrical) approach is grounded on a hypothesis of existence of piecewise linear dependence, and private linear dependences correspond to some intervals of change of the dependent parameter. The second (nonparametric) approach consists in solution of main task as search of collective solution on set of tasks of recognition

**Keywords:** dependence restoring, regression, algorithm of recognition, piecewise linear function, feature, dynamic programming

**ACM Classification Keywords:** A.0 General Literature - Conference proceedings, G.1.2 Approximation: Nonlinear approximation, H.4.2 Types of Systems: Decision support, I.2 Artificial intelligence, I.5 Pattern recognition

---

### Introduction

The task of restoring of dependence between a vector of variable (features)  $\mathbf{x} = (x_1, x_2, \dots, x_n)$ ,  $x_i \in M_i, i = 1, 2, \dots, n$ , where  $M_i$  are sets of any nature, and a scalar  $y$  on sampling  $\{(y_i, \mathbf{x}_i)\}_{i=1}^m$  is considered. Assuming existence between them functional link  $y = f(\mathbf{x})$ , on sampling function from some parametrical class of functions or the algorithm is selected, allowing to calculate for a vector of variables  $\mathbf{x}$  appropriate value of the dependent value  $y$ . The given task in statistical setting is known as the task of restoring of regression - functions of a conditional expectation. Now there are various parametrical and nonparametric approaches to restoring of regressions [1,2]. Parametrical approaches demand a priori knowledge of analytical sort of functions. Nonparametric approaches use as a rule methods of a frequency estimation and functions of distances. The given approaches have the essential limitations linked to such properties of the real data as heterogeneity of features, various informativity of features, co-ordination of metrics of various features, etc. At the same time, for a case of the discrete value  $y \in \{1, 2, \dots, l\}$  (the standard task of recognition [3,4]) the given limitations are not critical. Enumerated above difficulty are successfully overcome, for example, in the logical models of recognition [3-8] which are not demanding solution of additional task of preprocessing of partially contradictory polytypic no presentable data.

The nonparametric method of restoring of dependence assumes "carrying over" of all marked above problems with features on recognition level. According to training sampling,  $N$  tasks of recognition are formed and, respectively,  $N$  algorithms of recognition are constructed.  $N$  recognition tasks are solved independently for any vector  $\mathbf{x}$  of features, and value of the dependent value  $y = f(\mathbf{x})$  is calculated as collective solution over recognition tasks.

The parametrical approach assumes that to some segments of a range of the dependent value there correspond the linear dependences from features. The task of restoring of piecewise linear dependence is reduced to solution of the task of dynamic programming variables in which correspond to points of splitting of a range of the

dependent value on intervals, and addends of function to be optimized define quality of approximating of dependence in an appropriate interval by linear function.

Results of practical approbation are performed.

## 1. Restoring of piecewise linear dependences on samplings of precedents

We consider that training sampling  $\{(y_i, \mathbf{x}_i)\}_{i=1}^m$ , in the form of the training table  $T_{nm}$  where each string is a vector of values of features and to it appropriate value of the dependent value  $y$  is set,

$$T_{nm} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} & | & y_1 \\ x_{21} & x_{22} & \cdots & x_{2n} & | & y_2 \\ \cdots & \cdots & \cdots & \cdots & | & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} & | & y_m \end{pmatrix}. \text{ In the present section we consider that } x_i \in R, i = 1, 2, \dots, n. \text{ Without loss}$$

of generality we can consider that all  $y_i, i = 1, 2, \dots, m$  are various ones and arranged on increase:  $y_i < y_{i+1}, i = 1, 2, \dots, m-1$ . Let us divide a change interval  $[y_1, y_m]$  of  $y(\mathbf{x})$  on  $l \geq 2$  intervals  $\Delta_1 = [y_1, y_{i_1}), \Delta_2 = [y_{i_1}, y_{i_2}), \dots, \Delta_l = [y_{i_{l-1}}, y_m]$ . Any splitting is set by a vector of parameters  $z = (z_0, z_1, z_2, \dots, z_l)$ ,  $z_i \in \{y_1, y_2, \dots, y_m\}, z_i < z_{i+1}, i = 0, 2, \dots, l-1, z_0 \equiv y_1, z_l \equiv y_m$ .

For any segment  $\Delta_i$  by data  $\{(y_j, \mathbf{x}_j), j = 1, 2, \dots, m\}: y_j \in \Delta_i$  there is by means of the least squares

method a function  $g_i(\mathbf{x}) = \sum_{t=1}^n a_t^i x_t + b^i$  in the best way approximating given subsample. Quality of the present

function (and the segment  $\Delta_i$ ) is estimated as  $f_i(z_{i-1}, z_i) = \frac{1}{|\Delta_i|} \sum_{j: y_j \in \Delta_i} (y_j - g_i(\mathbf{x}_j))^2$ . Then the task of

search of the best piecewise linear approximating with  $l$  components is formulated in the form of the following task of dynamic programming:

$$\Phi(z_1, z_2, \dots, z_{l-1}) = f_1(y_1, z_1) + \sum_{i=2}^{l-1} f_i(z_{i-1}, z_i) + f_l(z_{l-1}, y_m) \quad (1)$$

$$z_i \in \{y_2, y_3, \dots, y_{m-1}\}, z_i < z_{i+1}, i = 1, 2, \dots, l-2, \quad (2)$$

Solving the dynamic programming task an optimal piecewise linear dependence is found. Thus the subsamples of  $\{(y_j, \mathbf{x}_j), j = 1, 2, \dots, m\}: y_j \in \Delta_i$ , are set which sets of vectors  $\mathbf{x}_j$  we consider as the description of some classes. Under descriptions of classes  $K_i, i = 1, 2, \dots, l$ , some algorithm of recognition  $A$  is calculated which is applied to classification of any new objects  $\mathbf{x}$ .

Finally, the task of  $y = f(\mathbf{x})$  calculation for any  $\mathbf{x}$  is solved in a two stages: object  $\mathbf{x}$  classification

$A: \mathbf{x} \rightarrow K_i$  is carried out, further  $y = f(\mathbf{x}) = g_i(\mathbf{x}) = \sum_{t=1}^n a_t^i x_t + b^i$  is calculated.

## 2. The nonparametric method of regression based on construction of collective solutions on set of tasks of recognition

In the present section we consider that  $x_i \in M_i, i = 1, 2, \dots, n$ ,  $M_i$  are sets of any nature. We will take number  $v \leq m$  and  $v+1$  points from a segment  $R = [y_1, y_m]$ :  $d_0 = y_1$ ,  $d_0 < d_1 < \dots < d_{v-1} < d_v = y_m$ . We receive set  $R$  splitting into  $v$  intervals  $\Delta_1 = [d_0, d_1), \Delta_2 = [d_1, d_2), \dots, \Delta_v = [d_{v-1}, d_v]$ ,  $\Delta = \{\Delta_1, \dots, \Delta_v\}$ .

We will put  $c_k = \frac{d_{k-1} + d_k}{2}$  - interval  $\Delta_k, k = 1, 2, \dots, v$  centre.

Let's take number  $2 \leq l \leq v$  and we will define  $N$  segment  $R$  splittings into  $l$  intervals having put to each splitting in correspondence a vector  $\mathbf{k}_i = (k_i^{(1)}, k_i^{(2)}, \dots, k_i^{(l-1)}), i = 1, 2, \dots, N, k_i^{(j)} < k_i^{(j+1)} < n$  with integer positive components. The vector data sets labels intervals in appropriate splitting. Intervals  $\Delta_1, \dots, \Delta_{k_i^{(1)}}$  flagged by a label «1», intervals  $\Delta_{k_i^{(1)+1}}, \dots, \Delta_{k_i^{(2)}}$  - a label «2», intervals  $\Delta_{k_i^{(l-2)+1}}, \dots, \Delta_{k_i^{(l-1)}}$  - a label « $l-1$ », intervals  $\Delta_{k_i^{(l-1)+1}}, \dots, \Delta_n$  - a label « $l$ ». Each splitting of a segment  $R$  defines set  $\mathbf{M} = M_1 \times \dots \times M_n$  splitting into  $l$  subsets (classes)  $K_1^i, \dots, K_l^i$ :  $\mathbf{M} = \bigcup_{i=1}^l K_i^i, v \neq \mu \Rightarrow K_v^i \cap K_\mu^i = \emptyset$ . Splitting  $\mathbf{K}^i = \{K_1^i, \dots, K_l^i\}$  is set by a rule: the object  $\mathbf{x}$  belongs to a class  $K_j^i$  if and only if  $y = f(\mathbf{x}) \in \Delta_k$  and the interval  $\Delta_k$  is flagged by a label « $j$ ». Each splitting  $\mathbf{K}^i$  defines the standard task of recognition  $Z_i$  with  $l$  classes.

Let  $A_i$  - some algorithm of solution of the task  $Z_i$ , carrying any object  $\mathbf{x}$  to one of classes  $K_{a_i}^i, a_i \in \{1, 2, \dots, l\}$ .

Let's consider the direct product  $\mathbf{K}^1 \times \dots \times \mathbf{K}^l \times \Delta$  as set of simple events. Event  $(K_{a_1}^1, \dots, K_{a_N}^N, \Delta_j)$  means reference of the object  $\mathbf{x}$  by algorithm  $A_1$  in a class  $K_{a_1}^1$ , ..., by algorithm  $A_N$  in a class  $K_{a_N}^N$ , thus  $y = f(\mathbf{x}) \in \Delta_j$ . Probability of such event we will designate as  $P(a_1, \dots, a_N, \Delta_j)$ .

According to the formula of Bayes, we have

$$P(\Delta_j | a_1, \dots, a_N) = \frac{P(a_1, \dots, a_N, \Delta_j)}{P(a_1, \dots, a_N)} = \frac{P(\Delta_j)}{P(a_1, \dots, a_N)} P(a_1, \dots, a_N | \Delta_j) \quad (3)$$

If algorithms are statistically independent, we have  $P(a_1, \dots, a_N | \Delta_j) = \prod_{i=1}^N P(K_{a_i}^i | \Delta_j)$ ,

$P(a_1, \dots, a_N) = \prod_{i=1}^N P(K_{a_i}^i)$  and the formula (3) becomes (4).

$$P(\Delta_j | a_1, \dots, a_N) = \frac{P(\Delta_j)}{\prod_{i=1}^N P(K_{a_i}^i)} \prod_{i=1}^N P(K_{a_i}^i | \Delta_j) \quad (4)$$

Let's enter notations  $p_k = P(\Delta_k | a_1, \dots, a_N), k = 1, \dots, v$ .

Function  $F : (p_1, \dots, p_v) \rightarrow R$  where  $p_1, \dots, p_v$  are received according to (3) is named as the Bayesian corrector. Function  $F : (p_1, \dots, p_v) \rightarrow R$  where  $p_1, \dots, p_v$  are received according to (4) is named as the naive Bayesian corrector. We will consider later the naive Bayesian corrector.

**Note.** We will notice that in cases when  $y = f(\mathbf{x}) \in \{1, 2, \dots, l\}$ , the value  $y$  means a class label at classification of the object  $\mathbf{x}$  and the primary goal is the recognition task. Here, all splittings  $\mathbf{K}^i$  are coincide ones. For the set of collection of recognition algorithms  $A_i, i = 1, 2, \dots, N$ , the model of the naive Bayesian corrector  $F : (p_1, \dots, p_v) \rightarrow \{1, 2, \dots, l\}$  in the recognition task is known (see, for example, [10]).

**Definition 1.** «As the answer on an average» of the Bayesian corrector for the object  $\mathbf{x}$ , we will name the value

$$\tilde{y} = \sum_{k=1}^v p_k c_k.$$

**Definition 2.** «As the answer on a maximum» of the Bayesian corrector for the object  $\mathbf{x}$ , we will name the value  $\tilde{y} = c_k$ , where  $k = \arg \max_j p_j$ .

Let's describe the common algorithm of restoring of dependence  $y = f(\mathbf{x})$  according to the training sampling, based on usage of Bayesian model, and algorithm of calculation of value of the dependent value  $y$  for any  $\mathbf{x}$ .

Algorithm of restoring of dependence  $y = f(\mathbf{x})$ .

1. Splitting of a segment  $R = [y_1, y_m]$  into intervals  $\Delta_k, k = 1, 2, \dots, v$ .
2. The calculation of splittings  $\mathbf{K}^i, i = 1, \dots, N$ , setting of tasks of recognition  $Z_i, i = 1, \dots, N$ , a choice of algorithms of recognition  $A_i, i = 1, \dots, N$ , and their training.
3. Calculation of estimations of probabilities  $P(K_j^i | \Delta_k), P(\Delta_k), P(K_j^i), i = 1, \dots, N, j = 1, \dots, l, k = 1, \dots, v$ .

Algorithm of calculation of value of the dependent value.

1. Classification of the object  $\mathbf{x}$  by algorithms  $A_i, i = 1, \dots, N$ .
2. Calculation of values of probabilities  $p_1, \dots, p_v$  according to (4).
3. Calculation of  $\tilde{y} = \sum_{k=1}^v p_k c_k$  or  $\tilde{y} = c_k, k = \arg \max_j p_j$ .

Practical implementation of the model of restoring of dependence set stated above and its application demands a concrete definition of all resulted parameters of model and algorithms of recognition. Here can be used any algorithms of recognition on precedents, and for calculation of estimations of probabilities – approaches and methods of mathematical statistics. In the present paper, the use of special collection of logical algorithms of recognition (test algorithms, algorithms of voting by representative sets, algorithms of voting by systems of logical regularities) and heuristic estimation of probabilities is considered and proved. The common feature of the given algorithms is faultless recognition of objects of no contradictory training sampling.

### 3. Restoring of dependences on the basis of application of collections of logical algorithms of recognition

Let's put  $l = 2$ . For simplicity we consider that all values  $y_i$  in training sampling are various and  $y_i < y_{i+1}, i = 1, \dots, m - 1$ . We will consider next two ways of construction of intervals  $\Delta_k$ .

1. We will take  $v = m, N = v - 1$ . We will put  $d_0 = y_1, d_1 = \frac{y_1 + y_2}{2}, \dots, d_{m-1} = \frac{y_{m-1} + y_m}{2}, d_m = y_m$ . For algorithm  $A_i$  intervals  $\Delta_1, \dots, \Delta_i$  are marked by a label «1», the others - «2».

2. The minimum value  $\varepsilon = y_{i+1} - y_i, i = 1, \dots, m - 1$ , is founded. We will put  $v = 2m - 2, N = v - 1 = 2m - 3$ , and

$$d_0 = y_1 - \frac{\varepsilon}{2}, d_1 = y_1 + \frac{\varepsilon}{2}, d_2 = y_2 - \frac{\varepsilon}{2}, d_3 = y_2 + \frac{\varepsilon}{2}, \dots,$$

$d_{2i} = y_{i+1} - \frac{\varepsilon}{2}, d_{2i+1} = y_{i+1} + \frac{\varepsilon}{2}, \dots, d_{v-1} = y_m - \frac{\varepsilon}{2}, d_v = y_m + \frac{\varepsilon}{2}$ . For algorithm of recognition  $A_i, i = 1, \dots, N$ , we will mark intervals  $\Delta_1, \dots, \Delta_i$  by label «1», the others – by «2».

As frequency estimation  $P(K_j^i), i = 1, \dots, N, j = 1, \dots, l$ , we will name a share of the objects belonging to a class  $K_j^i$  in the task  $Z_i$ . As frequency estimations  $P(K_j^i | \Delta_k), i = 1, \dots, N, j = 1, \dots, l, k = 1, \dots, v$ , we will name the ratio  $\frac{m_{ij}^{(k)}}{m_{ij}}$ , where  $m_{ij} = |\{\mathbf{x}_t : \mathbf{x}_t \in K_j^i\}|$ ,  $m_{ij}^{(k)} = |\{\mathbf{x}_t : \mathbf{x}_t \in K_j^i, y_t \in \Delta_k\}|$ . As a frequency

estimation  $P(\Delta_k), k = 1, \dots, v$ , we will name the value  $\frac{|\{\mathbf{x}_t, t = 1, \dots, m : y_t \in \Delta_k\}|}{m}$ .

**Definition 3.** The model of restoring of dependence is named correct if for training sampling  $\{(y_i, \mathbf{x}_i)\}_{i=1}^m$  takes place  $\tilde{y}_i = y_i, i = 1, \dots, m$ .

**Definition 4.** The Bayesian corrector over logical algorithms of recognition with frequency estimations of probabilities is named as model  $A_1$  at usage of the second way of construction of intervals and «the answer on a maximum».

**Theorem 1.** The model  $A_1$  at the no contradictory training information is correct.

At practical calculation of values  $P(K_j^i | \Delta_k)$  there are situations, when  $p_k = P(\Delta_k | a_1, \dots, a_N) = 0, k = 1, \dots, v$ . Let  $d$  is a natural number. We will consider

$$P'(K_j^i | \Delta_k) = \sum_{t=\max(k-d, 1)}^{\min(k+d, n)} w_{t-k} P(K_j^i | \Delta_t), \text{ and } \tilde{P}(K_j^i | \Delta_k) = \frac{P'(K_j^i | \Delta_k)}{\sum_{t=1}^v \frac{P(\Delta_t)}{P(K_j^i)} P'(K_j^i | \Delta_t)}$$

of a window of smoothing, and nonnegative values  $w_{-d}, \dots, w_d$  are smoothing scales. It is visible that,

$$\tilde{P}(K_j^i | \Delta_k) \geq 0, \sum_{k=1}^v P(\Delta_k) \tilde{P}(K_j^i | \Delta_k) = P(K_j^i), \text{ i.e. } \tilde{P}(K_j^i | \Delta_k) \text{ are conditional probabilities formally.}$$

Replacement process  $P(K_j^i|\Delta_k) \rightarrow \tilde{P}(K_j^i|\Delta_k)$  we will name as smoothing. In addition, we will set superimpose limitation of symmetry  $w_{-t} = w_t, t = 1, \dots, d$ , and character of decrease  $w_0 > 2w_1 \geq 4w_2 \geq \dots \geq 2^d w_d$ . We will name модель  $A_1$  as model  $A_2$  at usage of procedure of smoothing with exponential function of scales. The Theorem 2 is correct.

**Theorem 2.** The model  $A_2$  at the no contradictory training information is correct one

#### 4. Experimental matching of models of restoring on the model data

It is spent three sorts of experiments for matching of the nonparametric method of restoring of dependences offered in the present article with linear regression and the nonlinear regression grounded on nuclear smoothing.

**1. Matching on elementary functions.** For the first experiment the elementary nonlinear one-dimensional dependences (a parabola and a sinusoid) are used. On fig. 1, 2 results of matching of the restored dependences with true on some segment on an example of one-dimensional dependences  $y = x^2$  and  $y = \sin(x)$  are resulted. In the task "parabola" training sampling was under construction by a rule  $x_i = 2i, y_i = (x_i)^2, i = 1, \dots, 50$ , and control sampling – by a rule  $x_i = i, y_i = (x_i)^2, i = 1, \dots, 100$ . In the task

"sinusoid" training sampling was under construction by a rule  $x_i = 2i, y_i = \sin(\frac{2\pi}{100} x_i), i = 1, \dots, 50$ , and

control – by a rule  $x_i = i, y_i = \sin(\frac{2\pi}{100} x_i), i = 1, \dots, 100$ . In figures 1,2 plots of dependences and results of their restoring are shown, in table 1 norms of vectors of errors of each of algorithms on the given tasks are presented.

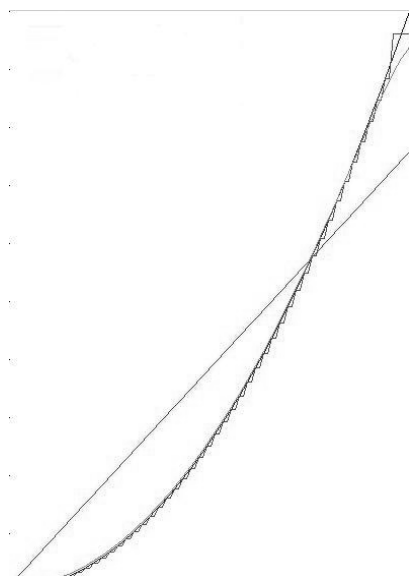


Figure 1. The model problem "parabola"

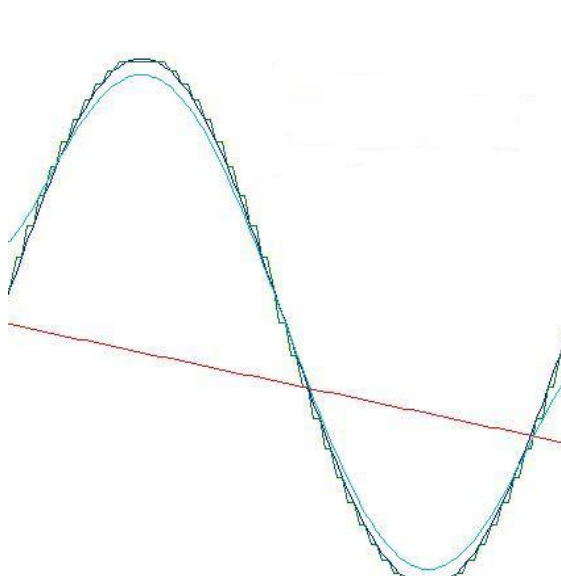


Figure 2. The model problem "sinusoid"

Table 1.

Task	Bayesian corrector	Linear regression	Nuclear smoothing
Parabola	0.31	6.52	0.64
Sinusoid	810	11321	1138

**2. Matching on the data with noise.** There were considered 10 various two-dimensional normal distributions. Means and dispersions of distributions were selected according to the uniform law of distribution. According to each of normal laws of distribution samplings of 12 points which have been united in one  $\{(x_i^{(1)}, x_i^{(2)})\}_{i=1}^{120}, x_i^{(1)}, x_i^{(2)} \in R$  are received. The target variable  $y$  coincides to within the sign with density of appropriate normal distribution. For the first 5 normal samples value  $y$  was the density, for the others – density with the sign a minus. Variables  $x^{(1)}, x^{(2)}$  have been added noise  $x^{(3)}, \dots, x^{(49)}$ , subordinates to the uniform law of distribution. The sampling  $\{(x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(49)}, y)\}_{i=1}^{120}$  generated as a result has been divided casually into training and forecasting samplings of 90 and 30 points, accordingly.

**3. Matching on tasks with discrete features.** Sampling of 200 3D objects according to a following rule is generated: -  $x_i^{(1)}, x_i^{(2)}$  implementations of a random variable  $\xi = 0, \dots, 10$  with probabilities  $\frac{1}{11}$ ,

$$x_i^{(3)} = i, i = 1, \dots, 200. \text{ It was considered the dependence } y = \begin{cases} x^{(3)}, & x^{(1)} \geq x^{(2)}, \\ -x^{(3)} & x^{(1)} < x^{(2)}. \end{cases} \text{ The generated sampling}$$

was considered divided casually on two on 100 objects (training and forecasting).

In Table 2 norms of vectors of errors of the algorithms, averaged on 10 independent experiments are resulted.

Table 2.

The task	Bayesian corrector	Linear regression	Nuclear smoothing
Data with noise	3.625	4.489	4.150
Discrete features	552.1	850.7	606.2

Results of experiments show a high accuracy of the Bayesian corrector, its ability successfully to handle data with noise and the different type data at rather small training samplings.

## 5. Conclusion

In a method of restoring of piecewise linear dependences it was supposed that the number of linear components is set. Clearly, that the criterion  $\Phi(z_1, z_2, \dots, z_{l-1})$  monotonously does not increase with growth of number of components and, therefore, cannot be used for definition of  $l$ . The true number of components can be found by exhaustive search of a small number  $l$  with calculation of  $\Phi(z_1, z_2, \dots, z_{l-1})$  in a mode of the cross-validation.

The Bayesian corrector for calculation of values of the dependent value is not the unique way grounded on problem solving of recognition. Here other approaches are possible which are developed now by authors.

---

## Aknowledgement

---

The study was supported by Russian Foundation for Basic Research (projects № 08-01-00636, 09-01-00409, 09-01-12060-ofi\_m), and by target program № 14 of Russian Academy of Sciences

---

## Bibliography

---

- [1] Drejper H, Smith G. Applied regression analysis. M.:Pabliishing house Williams, 2007.
- [2] Hardle B. Applied nonparametric regression. M, The World, 1993.
- [3] Zhuravlev Ju.I. Correct algebras over sets of not correct (heuristic) algorithms. I. Cybernetics. 1977. N4. pp. 5-17., II. Cybernetics, N6, 1977, III. Cybernetics. 1978. N2. pp. 35-43.
- [4] Zhuravlev Ju.I. About the algebraic approach to solving of recognition or classification problems. Cybernetics problems. M: The Science, 1978. 33. pp.5-68.
- [5] Dmitriev A.N., Zhuravlev Ju.I., Krendelev F.P., About mathematical principles of classification of subjects and the phenomena. The transactions "The Discrete analysis". Issue 7. Novosibirsk, IM SB AS USSR. 1966. pp. 3-11
- [6] Baskakova L.V., Zhuravlev Ju.I. Model of recognising algorithms with representative sets and systems of basic sets//Zhurn. vichisl. matem. and matem. phys. 1981. Vol.21, № 5. pp.1264-1275
- [7] Ryazanov V.V. Logical regularities in recognition tasks (the parametrical approach)// Zhurn. vichisl. matem. and matem. phys. 2007. Vol.47, № 10. pp.1793-1808
- [8] Zhuravlev Ju.I., Ryazanov V.V., Senko O.V. Pattern recognition. Mathematical methods. The program system. Applications. - Moscow: Phasys publisher, 2006, 176 p.
- [9] Sigal I.H., Ivanov A.P. Introduction in applied discrete programming: models and computing algorithms. The manual. M: PHYSMATLIT, 2002, 240 p.
- [10] P.Domingos and M.Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29:103-130, 1997.

---

## Authors information

---



**Vladimir Ryazanov** – Professor, Head of Department; Institution of Russian Academy of Sciences Dorodnicyn Computing Centre of RAS, Russia, 119991 Moscow, Vavilov's street, 40; e-mail: [rvv@ccas.ru](mailto:rvv@ccas.ru)

Major Fields of Scientific Research: Pattern recognition, Data mining, Artificial Intelligence



**Yury Tkachev** – The post-graduate student; Institution of Russian Academy of Sciences Dorodnicyn Computing Centre of RAS, Russia, 119991 Moscow, Vavilov's street, 40; e-mail: [tkachevy@gmail.com](mailto:tkachevy@gmail.com)

Major Fields of Scientific Research: Pattern recognition, Data mining, Artificial Intelligence



---

## COMPOSITE BLOCK OPTIMIZED CLASSIFICATION DATA STRUCTURES

Levon Aslanyan, Hasmik Sahakyan

**Abstract:** *There are different applications that require finding in a set all points most similar to the given query element. These are problems known as classification problems, or as best match, or the nearest neighbor search procedures. The main goal of this paper is to analyze the algorithmic constructions that appear in a model of recognizing the nearest neighbors in binary data sets. Our target is an earlier proposed algorithm [W,1971, R 1974, A,2000] where existence and effectiveness of structures that serve the search procedure is raised. These structures split the search area into the perfect codes and/or the discrete isoperimetry problem solutions, which coexist for a very limited area of parameters. We extend indirectly these parameters through the composite block constructions and estimate the resulting loss of effectiveness of the procedure under consideration.*

**Keywords:** *Pattern Recognition, Classification, Best Match, Perfect Code*

**ACM Classification Keywords:** *1.5. Pattern recognition, H.2.8 Database applications, Data mining*

---

### Introduction

---

Consider an application scenario.

Let an orthography dictionary and a text file for correction is given. Word by word correction scenario fulfills consecutive comparison of text words with words that are in dictionary. Suppose that a formalism for simple word 'mistake formats' is given – such as one or more wrong letters, single character transpositions, retyping mistakes, missing characters, etc. More detailed formalisms can be devised involving grammatical rules and relations but this simple model quiet well demonstrates the practical problem we consider. Suppose that we are able to define an appropriate measure (metric) between the words (correct and misspelled) – words of text and words of dictionary. Then, it is worthwhile to seek the correction word for the word from text among the closest words of the dictionary by the given metric [GM,2005].

Dramatically this simple scenario speaks about deep similarities between three very different research areas – coding for error correction theory, supervised classification and pattern recognition, and finally – the search for similarities - best matches and nearest neighbors. We do not have in our problem a single bit change like in basic coding theory but we have a larger change through the spelling error model. In terms of pattern recognition we work with too many classes – one for each dictionary word and the learning set (dictionary) is very large. Finally we deal with repeated searches in a single file (dictionary) which obviously is to be well structured to minimize the search time for these specific queries of word spelling. The problem is treating as is, without separating and allocating it to one of the mentioned research areas. Coding is used as the basic structuring tool, compactness hypothesis from pattern recognition is proving the optimality of structures, and search algorithm is composite to achieve the tradeoff between the structural validity and the functional optimality.

The initial structures and investigation is by earlier works of P. Elias and R. Rivest [W,1971, R,1974]. A special cellular block partitioning of basic word space is considered, and a special dynamic programming style mechanism of search of best match or nearest neighbor word sets is applied. Consecutively, there appeared in area different alternative search strategies such as: k-d trees [F,1975], vp-trees [Y,1993], Voronoi tessellation [L,1982], etc. Although new forms and approaches are more perspective, we consider the basic model [W,1971, R,1974] and our additional input is in applying our detailed study and results on discrete isoperimetry problem as

the known formalism for pattern recognition compactness hypothesis [A,1989, A,1981]. Then we construct composite blocks which split the basic set into the homogeneous blocks. Blocks are not isoperimetric so we estimate the resulting loss of search optimality.

There are several typical applications by the same scenario. A data file, for example [F,1975], might contain information on all cities with post offices in a region. Associated with each city is its longitude and latitude. If a letter is addressed to a town without a post office, the closest town that has a post office might be chosen as the destination. The solution to such problem is of use in many other applications too. Information retrieval might involve searching in a catalogue for those items most similar to a given query item; each item in the file would be catalogued by numerical attributes that describe its characteristics. Classification decisions can be made by selecting prototype features from each category and finding which of these prototypes is closest to the record to be classified. Multivariate density estimation can be performed by calculating the volume about a given point containing the closest neighbors.

---

### Basic Structures and Definitions

---

Let  $F$  be a finite set of some binary words of length  $n$ .  $x$  is an input binary  $n$ -word.  $F(x)$  denotes the set of all words from  $F$  having the (same) minimal possible distance from  $x$  (in the simplest case the Hamming distance is applied). Optimization and complexity issues of algorithms, working with  $F$  and  $x$  and composition of  $F(x)$  is considered, concerning with special constructions that map the basic set  $F$  onto the computer memory. The initial methods, based on error correcting perfect codes are defined in [W,1971, R,1974]. They are restricted by the very limited set of possible perfect codes that limits the special constructions mentioned above. It is well known that the only nontrivial classes of binary perfect codes are Hamming and Golay codes [TP,1971, ZL,1972]. We use the geometric interpretation, when the code centered and none intersecting sphere systems are given. Linear codes allow optimizing of addressing issues in considered models. Spheres play the role of blocks when overall algorithm optimality requires the discrete isoperimetry property (DIP) of blocks.

DIP problem is one of the typical issues of advanced discrete mathematics. DIP solutions are very close to the Hamming spheres geometrically. The complete coverage of basic binary set by such objects is highly important and provides more possibilities of construction of searching algorithms. Any positive steps on this direction require more knowledge on properties of the solutions of the DIP that are provided by authors in [A,1989 - A,2000].

So, let us suppose that the basic set  $\Xi^n$  of all binary words of length  $n$  is divided into the blocks  $B_1, B_2, \dots, B_m$ . Accordingly,  $L_i = F \cap B_i, i = 1, 2, \dots, m$  are the lists of elements of  $F$  belonging to these blocks. They are saved as separate lists by addresses  $h(B_i)$ , common for all elements of each such block of  $\Xi^n$ . The idea of this searching construction is transparent - using a dynamic programming style algorithm of class of branches and leaves and a partitioning the space  $\Xi^n$  into the geometrically compact blocks, the algorithm may apply to less input information to get the  $F(x)$  by an input  $x$ . To achieve this result we have to divide  $\Xi^n$  into the disjoint blocks - solutions of the DIP [R,1974]. For example, splitting into the spheres by a perfect code is ideal. And the problem is that neither the splitting of  $\Xi^n$  into the solutions of the isoperimetry problem nor the construction of a perfect code for an arbitrary  $n$  is possible. Alternatives are to be constructed and then evaluated.

Let us summarize the main points of structuring  $\Xi^n$  for search:

$\Xi^n$  is divided into the simple blocks  $B_1, B_2, \dots, B_m$ . Blocks are similar in their sizes, and the lists of elements in blocks  $L_i = F \cap B_i, i = 1, 2, \dots, m$  practically have comparable sizes.

Given an arbitrary  $x$ , blocks can be quickly and simply ranged by their distances to  $x$ .

The simplest partition as we mentioned is by a Hamming code. But these codes exist for dimensions  $n = 2^q - 1$  only. And the block size by Hamming codes is very much limited:  $n + 1$ . The same time ranging blocks by the input vector  $x$  is very quick and ideally simple. Other codes such as Golay codes and non linear codes or codes in other distances do minimal help. It is also to take into account the quasi-perfect and nearly perfect codes which provide more constructions for approximating the problem. Next issue is the search optimality that depends on block shapes. The main point of optimality is proposed in [R,1974]:

The block is optimal when its shape is the DIP solution.

Let us comment the proposed optimality of DIP solutions in [R,1974]. In a supposition that  $F$  is a random set of vertices of  $\Xi^n$  with membership probability  $p$ , probabilities of blocks that are analyzed for  $F(x)$  are evaluated. The following resolution is used:

$$\Psi(C) - \Psi(B) = 2^{-n} \left( |C^{(m)}| - |B^{(m)}| \right) \Theta(n, m - 1) + \quad (1)$$

$$2^{-n} \sum_{m < \delta \leq n} \left( |C^{(\delta)}| - |B^{(\delta)}| \right) \Theta(n, \delta - 1) \geq \quad (2)$$

$$2^{-n} \left( |C^{(m)}| - |B^{(m)}| \right) \left( \Theta(n, m - 1) - \Theta(n, m) \right) \geq 0 \quad (3)$$

Here  $C$  and  $B$  are blocks of the same size, and  $B$  is DIP optimal. In fact  $B$  is spherical [R,1974] or initial segment of standard placement  $L_n$  (see below) [A,1989].  $B^{(i)}$  is the  $i$  neighborhood of  $B$  - the vertices that are in distance  $i$  from  $B$ .  $\Theta(n, m)$  is probability that a sphere of radius  $m$  is empty (of  $F$ ).  $m$  is the position/index, that  $|C^{(i)}| = |B^{(i)}|$  for  $i = 0, \dots, m - 1$  (that probably may happen) but  $|C^{(m)}| > |B^{(m)}|$ . Because of DIP optimality of  $B$ :

$|C^{(i)}| > |B^{(i)}|$  for some number of indexes  $i$  after  $m$ , then this may become negative, and in case  $|C^{(i)}|$  (4) may become 0.

[R,1974] formulates the lexicographic minimality of  $|B^{(0)}|, \dots, |B^{(i)}|, \dots, |B^{(k)}|$ . This is not satisfactory for transfer from (2) to (3). The stronger and satisfactory is (4) that we brought above in accord to DIP postulations of [A,1989].

Consider a set  $A \subseteq \Xi^n$ . A point  $\alpha \in A$  is called the inner point of  $A$  if the unit sphere  $S_n(\alpha) \equiv \{\beta \in A \mid \rho(\alpha, \beta) \leq 1\}$  at  $\alpha$ , is a member of  $A$ . In opposite case we call  $\alpha$  the boundary point of  $A$ . Let  $\mathfrak{I}(A) \subseteq A$  is the collection of all inner points of  $A$ . Then  $\mathfrak{B}(A) = A \setminus \mathfrak{I}(A)$  is the set of all boundary vertices. A set  $A \subseteq \Xi^n$  is called DIP optimal (isoperimetric) if  $|\mathfrak{I}(A)| \geq |\mathfrak{I}(B)|$  for any  $B \subseteq \Xi^n$ ,  $|B| = |A|$ .

Consider the linear order  $L_n$  of vertices  $\Xi^n$  called standard.  $L_n$  is the order of vertex sets of layers of  $\Xi^n$  from 0 to  $n$ , and inside the layers - order is lexicographical. Let  $a$  is a nonnegative integer,  $a \leq 2^n$ . Denote by  $L_n(a)$  the initial  $a$ -segment of  $L_n$ , then the Main Isoperimetric Theorem proves that  $L_n(a)$  is a solution of DIP [A,1989]. This result was formulated independently by different authors.

Call the number  $m$   $k, \delta$ -spherical, if  $m = \sum_{i=0}^k C_n^i + \delta$ . As it has been proposed in [R,1974] the optimal best match search algorithms of hashing class correspond to the selection of blocks  $B_i$  which are the DIP solutions.

To apply this result one need to split  $\Xi^n$  into such blocks. We have:

The arbitrary solution of the DIP contains a Hamming sphere of the maximal possible sphere by the given  $m$  (the radius of these sphere equals  $k$ ) and only the additional  $\delta$  vertices or the part of these might be arranged differently.

At least for  $2^{n-1}$  cases by all different  $m$  the more precise description of solutions of DIP is given. These DIP solutions are included in a sphere of radius  $k + 2$ . Such numbers  $m$  we call critical. The quantity of vertices between some two closest critical numbers is distributed randomly and they can't create additional inner vertices.

The number of subsets of  $\Xi^n$  with  $\xi$  interior vertices is distributed by the Poisson's distribution with the main value  $\frac{1}{2}$ . This is in case of random membership of vertices of  $\Xi^n$  into the considered sets by the probability  $\frac{1}{2}$ . For probabilities other than  $\frac{1}{2}$  the picture is similar but more complicated, which is described in a separate paper.

All the above mentioned descriptions of DIP solutions are rather complicated to construct the precise splitting of  $\Xi^n$  by these objects. So the geometrical shape of the DIP solutions give us the ideal partitioning objects exemplifying.

Coming back to the formula (3) let  $i_0$  be the minimal index  $i$  with  $|C^{(i)}| = 0$ . We take  $\Theta(n, i_0)$  instead of  $\Theta(n, m)$  in (3) for indexes  $i_0$  and above. For parts where  $|C^{(i)}| - |B^{(i)}|$  are positive and negative we may take some evaluating values  $\Theta(n, i_1)$  and  $\Theta(n, i_2)$  which is also possible. Then it is easy to see that  $\Theta(n, i_1)$  satisfies (3) because of  $\sum_{m \leq \delta \leq n} |C^{(\delta)}| = \sum_{m \leq \delta \leq n} |B^{(\delta)}|$ .

Recall the main requirements to block structures: partitioning of  $\Xi^n$ ; computation of distances; and DIP optimality. The Hamming sphere is the simplest DIP solution. For other block sizes DIP solutions are similar but different. First question arises is about the approximate block partitioning of  $\Xi^n$ . One approach is in partitioning into the DIP solutions with intersections. If diversity of DIP solutions is high and/or intersections can be minimized then the optimality loss is related to repetitions of elements in different lists which may be small. The second approach is in splitting the space into the subspaces for which block partitioning is effective. Even small blocks in subspaces become large in a Cartesian product. The optimality loss is related to non DIP optimality of product which is estimated below. Before that we mention in short some similar structures from the coding theory. Here accent is exactly on reduction of repetitions of elements in intersections and not to the DIP optimality.

An  $(n, t)$ -quasi-perfect code is a code for which the spheres of radius  $t$  around the code words are disjoint, and every vector is at most  $t + 1$  from some code word. A subclass of quasi-perfect codes is nearly perfect codes. A few nearly perfect code sets are known. For  $t = 1$  there exists a nearly perfect code for any  $n = 2^q - 2$ . For  $t = 2$  there exist a nearly perfect code for  $n = 4^q - 1$ . It is also known that no other nearly perfect codes exist.

For any nearly perfect code: vectors with a greater than  $t$  distance from any code word is at distance  $t + 1$  from exactly  $\lfloor n/(t + 1) \rfloor$  other code words; and vectors of distance  $t$  from some code word are at distance  $t + 1$  from exactly  $\lfloor (n - t)/(t + 1) \rfloor$  other code words.

### Composite Block Structures

Second approach concerned to effective structuring for best match search is to analyze the possibility of structuring the so called composite blocks. This is when we split the basic space  $\Xi^n$  into the Cartesian product of smaller size spaces. Then we use different constructions, based on the Hamming spheres of different sizes. Given an  $n$  we first choose a number of form  $2^q - 1$  to be not greater than  $n$ . As we know there exists a perfect Hamming code for this case, so, the exact partitioning of  $\Xi^n$  into the spheres of radius 1 is given and may be used. Considering partitioning of  $\Xi^n$  according with the Cartesian product of  $\Xi^{2^q-1}$  and  $\Xi^{n-2^q+1}$ , we can choose the splitting of the first subspace as a Hamming code while the second part might be reminded as is or further split by itself. The main advantage is that the blocks are Hamming spheres and that the values of corresponding  $h(x)$  functions as well as the distances of these blocks from the arbitrary points  $x \in \Xi^n$  are simply computable.

Generalization of this idea is related to the special representation of arbitrary numbers  $n$  by the sums of numbers of form  $2^q - 1$ . This is for decomposition of  $\Xi^n$  into the subcubes, which can be covered by the sets of disjoint Hamming spheres. In parallel additional compounds may be used such as Golay codes and the simple constructions which partition arbitrary cubes into the two subspheres, etc.

To be compact we can formulate the following properties:

Let us consider a binary vector  $\tilde{\alpha} = (\alpha_n, \alpha_{n-1}, \dots, \alpha_1)$ . The corresponding sum  $s(\tilde{\alpha}) = \sum_{i=1}^n \alpha_i (2^{i-1} - 1)$  is limited by numbers 0 and  $2^{n-1} - n$ . Moreover, all sums of form  $s(\tilde{\alpha})$  don't cover this interval completely.

Sums  $s(\tilde{\alpha})$  achieve the  $2^{n-1}$  different values. The last coordinate -  $\alpha_1$  is not essential for the value of  $s(\tilde{\alpha})$ .

Let us consider the vector  $\tilde{1}_i$  with the all 1 coordinates on positions  $j, j \geq i$  and 0 elsewhere. We'll double the last sum term -  $2^{i-1} - 1$ , which corresponds to the  $i$ -th 1 of  $\tilde{1}_i$ . Then we get the numbers from  $2^n - 2$  to  $2^n - n$  continuously. For the numbers, starting from  $2^n - n - 1$  and smaller we can prove by induction on  $n$ , that doubling only one sum term we can receive an arbitrary number from the remainder part.

Consider a simple case of composite blocks which are in 2 parts. Let  $\Xi^n = \Xi^{2^{q_1}-1} \times \Xi^{2^{q_2}-1}$ . Consider Hamming codes in  $\Xi^{2^{q_1}-1}$  and in  $\Xi^{2^{q_2}-1}$  and define the blocks as Cartesian product of unit spheres defined by these codes. Such block is included in a sphere of radius 2. Similarly in the case of Cartesian product of two arbitrary spheres - one of radius  $k_1$  and second of radius  $k_2$ , the result is included in sphere of radius  $k_1 + k_2$ . The points of these spheres that are not a block member are in different layers. In second layer for example there are  $k_1 \cdot k_2$  points. This formula can be extended for other layers but we prefer to compute the missing points from another point of view.

First claim is that constructed blocks consist of  $b = \sum_{i_1=0}^{k_1} C_{2^{q_1}-1}^{i_1} \cdot \sum_{i_2=0}^{k_2} C_{2^{q_2}-1}^{i_2}$  points each. Number of points of

$k_1 + k_2$  sphere is  $s = \sum_{i=0}^{k_1+k_2} C_{2^{q_1}+2^{q_2}-2}^i$ . The number of missing point of block to the complete sphere equals  $s - b$ . Compare  $s$  and  $b$ . In a simplified structure when  $q_1 = q_2 = q$  and  $k_1 = k_2 = k$  we apply the following known inequality [PW,1972]:

$$C_n^{\lambda n} < \sum_{i=0}^{\lambda n} C_n^i < \frac{1-\lambda}{1-2\lambda} C_n^{\lambda n}, \lambda < 1/2.$$

Here  $\lambda n$  supposed to take integer values. Blocks in general are not very large so that we suppose the sphere radius is some constant number. Let  $\lambda = o(1)$  with  $n \rightarrow \infty$ . Then  $b \sim (C_{2^{q-1}}^k)^2$  and  $s \sim C_{2(2^q-1)}^{2k}$ . For small  $k$ , the Hamming case included, these values are comparable.

---

## Conclusion

Finding nearest neighbors is a regular procedure in experimental data analysis. Pattern recognition is the closest model where the nearest elements of the learning set is a question. To quick up the search for similarities it is common to use divide and conquer approach through the partition of unit cube into the blocks. Three requirements are the main: partitioning; computation of distances; and optimality. The tradeoff between these concurring requirements can be resolved partially. DIP solutions and standard placement in particular, perfect and nearly perfect codes that exist for exceptional dimensions, and space partitioning into the Cartesian products are the main algorithmic resource. Complete solutions are linked to perfect codes and all other cases are accompanied with losses and approximations. Future work will describe similar structures in other metrics, e.g. Lee metrics.

---

## Bibliography

- [W, 1971] T.A. Welch. Bounds on the information retrieval efficiency of static file structures, Project MAC Rep. MAC-TR-88, Mass. Inst. of Tech., Cambridge, Mass., 164 p., 1971, Ph.D. thesis.
- [R, 1974] R.L. Rivest. On The Optimality of Elias's Algorithm for Performing Best-Match Searches, Information Processing 74, Nort-Holland Publishing Company, pp. 678-681, 1974.
- [F, 1975] J. H. Friedman, F. Baskett and L.J. Shustek, An algorithm for finding nearest neighbors, IEEE Trans. Comput., vol. C-24, pp. 1001-1006, Oct. 1975.
- [Y, 1993] Yianilos, Peter N., Data structures and algorithms for nearest neighbor search in general metric spaces, Proceedings of the fourth annual ACM-SIAM Symposium on Discrete algorithms, Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 311-321, 1993.
- [L, 1982] D.T. Lee, On k-Nearest Neighbor Voronoi Diagrams in the Plane, IEEE Transactiona on Computers, Vol. C-31, N. 6, June, pp. 478-487, 1982.
- [A, 1989] L.H. Aslanyan, The discrete isoperimetric problem and related extremal problems for discrete spaces, Problemy Kibernetiki, Moscow, v. 36, pp. 85-128, 1989.
- [A, 1981] L. Aslanyan and I. Akopova, On the distribution of the number of interior points in subsets of the n-dimensional unit cube, Colloquia Mathematica Societatis Yanos Bolyai, 37, Finite and Infinite Sets, (Eger) Hungary, pp. 47-58, 1981.
- [A 2000] L. Aslanyan, Metric decompositions and the Discrete Isoperimetry, IFAC Symposium on Manufacturing, Modelling, Management and Control, July 12-14, Patras, Greece, pp. 433-438, 2000.
- [AC, 2007] L. Aslanyan and J. Castellanos, Logic based Pattern Recognition - Ontology content (1), Information Theories and Applications, ISSN 1310-0513, Sofia, Vol. 14, N. 3, pp. 206-210, 2007.
- [AR, 2008] L. Aslanyan and V. Ryazanov, Logic Based Pattern Recognition - Ontology Content (2), Information Theories and Applications, ISSN 1310-0513, Sofia, Vol. 15, N. 4, pp. 314-318, 2008.
- [TP, 1971] A. Tietäväinen and A. Perko, There are no unknown perfect binary codes, Ann. Univ. Turku, Ser. AI, 148, pp. 3-10, 1971.
- [ZL, 1972] V. Zinoviev and V. Leontiev, On the Perfect Codes, Problems of Information Transferring, Moscow, Vol. 8, N. 1, pp. 26-35, 1972.

---

[PW, 1972] W. W. Peterson and E. J. Weldon, Error-Correcting codes, second edition, The MIT Press, Cambridge, 1972.

[GM, 2005] H. Ghazaryan and K. Margaryan, Armenian spell checker, "Computer Science & Information Technologies" Conference, Yerevan, September 19-23, pp. 632-634, 2005.

---

### Authors' Information

---



**Levon Aslanyan** – Head of Department, Institute for Informatics and Automation Problems, NAS RA, P. Sevak St. 1, Yerevan 14, Armenia, e-mail: [lasl@sci.am](mailto:lasl@sci.am)



**Hasmik Sahakyan** – Leading Researcher, Institute for Informatics and Automation Problems, NAS RA, P. Sevak St. 1, Yerevan 14, Armenia, e-mail: [hasmik@jia.sci.am](mailto:hasmik@jia.sci.am)

## SYNTHESIS OF CORRECTOR FAMILY WITH HIGH RECOGNITION ABILITY\*

Elena Djukova, Yurii Zhuravlev, Roman Sotnezov

**Abstract:** *The model of recognizing procedures based on construction of family of logic correctors is proposed. For these purposes the genetic approach is used. This method allows, firstly, to reduce calculation costs and, secondly, to construct correctors with high recognition ability. The proposed model is tested on real problems.*

**Key words:** *logical recognition procedures, covering of the Boolean matrix, logical corrector, algebra-logic data analysis, genetic algorithm.*

**ACM Classification Keywords:** *I.5.1 Computing Methodologies - Pattern Recognition - Models*

---

### Introduction

In the paper questions of logic and algebra-logic data analysis are considered. The most important problem in this direction is concerned with generation of informative fragments from feature description of objects. These fragments play the role of elementary classifiers and allow to differ objects from different classes. As a rule, the correctness of recognition algorithm (the ability to classify the objects from a training sample correctly) is provided by the correctness of each used elementary classifier [Djukova, Zhuravlev, 1997], [Djukova, Zhuravlev, 2000].

*For generalization of this approach the correct recognition procedures can be constructed from the arbitrary sets of valid feature values. As a corrective function a monotone Boolean function can be used. In this case the constructing a corrector of minimal complexity can be reduced to the search of minimal cover of a Boolean matrix, which can be constructed from the training sample and has large size even in the simplest case. The idea of constructing the logic corrector was proposed in the work [Djukova, Zhuravlev, Rudakov, 1996]. Unfortunately, the problem has not been given sufficient attention, in particular, because of the large computational complexity.*

In the paper a new model of recognition procedures based on constructing the logic corrector family is proposed. For these purposes the genetic approach is used. This method allows, firstly, to reduce calculation costs and, secondly, to construct correctors with high recognition ability. Proposed model is tested on real problems.

---

### Description of voting model on logic correctors

*The problem of recognition by precedents is considered in standard formulation. [Zhuravlev, 1978]. The set of objects  $M$  that can be represented as a union of disjoint subsets (classes)  $K_1, \dots, K_l$  is investigated. Objects of set  $M$  are described by the set of integer attributes  $x_1, \dots, x_n$ . Each attribute has a finite number of valid values.*

*As the initial information the set of object descriptions  $T$  from  $M$  is given (training sample). For these objects it is known which classes they belong. It is required on training sample and the description in system of attributes  $x_1, \dots, x_n$  of some object  $S$  to determine to what class it belongs.*

Let  $H = \{x_{j_1}, \dots, x_{j_r}\}$  be a set of  $r$  attributes and  $\sigma = (\sigma_1, \dots, \sigma_r)$ , where  $\sigma_i$  - one of valid values of feature  $x_{j_i}$  under  $i = 1, 2, \dots, r$ . The pair  $(H, \sigma)$  is called an elementary classifier (e.c.). The e.c.  $(H, \sigma)$  generates the predicate  $P_{(H, \sigma)}(S)$  which is defined on objects  $S \in M$ ,  $S = (a_1, \dots, a_n)$ , and such that



$$P_{(H_i, \sigma)}(S) = \begin{cases} 1, & \text{if } a_{j_i} = \sigma_i, \\ 0, & \text{otherwise.} \end{cases}$$

The set of e.c.  $U = \{(H_1, \sigma_1), \dots, (H_q, \sigma_q)\}$  is called a (monotone) correct set for the class  $K$ ,  $K \in \{K_1, \dots, K_j\}$ , if there exists the (monotone) function of the logic algebra  $F_K$ , dependent on  $q$  variables, such that

$$F_K(P_{(H_1, \sigma_1)}(S), P_{(H_2, \sigma_2)}(S), \dots, P_{(H_q, \sigma_q)}(S)) = \begin{cases} 1, & \text{if } S \in K \cap T, \\ 0, & \text{if } S \in \bar{K} \cap T. \end{cases}$$

(here and in the following text  $\bar{K} = \{K_1, \dots, K_j\} \setminus \{K\}$ ).

The function  $F_K$  is called a (monotone) corrector for the class  $K$ , denote as  $\omega_U(S)$  the binary set  $(P_1(S), P_2(S), \dots, P_q(S))$ .

Let  $a' = (a'_1, \dots, a'_n)$ ,  $a'' = (a''_1, \dots, a''_n)$  be binary sets. The note  $a' \succ a''$  means that  $a'_i \geq a''_i$  for all  $i = 1, 2, \dots, n$ .

Let  $S', S'' \in M$  and  $U$  be the correct set of e.c. Let's put

$$\delta(S, S', U) = \begin{cases} 1, & \text{if } \omega_U(S) \succ \omega_U(S'), \\ 0, & \text{otherwise} \end{cases}$$

in the case when the  $U$  is the monotone correct and

$$\delta(S, S', U) = \begin{cases} 1, & \text{if } \omega_U(S) = \omega_U(S'), \\ 0, & \text{otherwise} \end{cases}$$

in the case when the  $U$  is the correct set that is not monotone.

Let  $W_K = \{U_1, U_2, \dots, U_t\}$  be the set of (monotone) correct sets of e.c. for the class  $K$ , then the score for the class  $K$  for recognizing object  $S$  has a form

$$\Gamma(S, K) = \frac{1}{|T \cap K|} \sum_{U \in W_K} \sum_{S' \in T \cap K} \delta(S, S', U)$$

We obviously have the next

**Statement 1.** The set of e.c.  $U = \{P_{(H_1, \sigma_1)}(S), P_{(H_2, \sigma_2)}(S), \dots, P_{(H_q, \sigma_q)}(S)\}$  is monotone correct set for the class  $K$  if and only if for any two objects  $S'$  and  $S''$  from training sample such that  $S' \in K$ ,  $S'' \notin K$  there exists  $i \in \{1, \dots, q\}$  that

$$P_{(H_i, \sigma_i)}(S') = 1 \text{ and } P_{(H_i, \sigma_i)}(S'') = 0 \quad (1)$$

The condition of monotony in the last statement can be removed if replace (1) on

$$P_{(H_i, \sigma_i)}(S') \neq P_{(H_i, \sigma_i)}(S'') \quad (2)$$

Let  $U$  be the correct set of e.c. for the class  $K$ . The set  $U$  is called irreducible if from the condition  $U' \subset U$  it follows that the set of e.c.  $U'$  is not correct set for the class  $K$ . The set  $U$  is called minimal if there is not exist smaller capacity correct set of e. c. for the class  $K$ .

Let  $L$  be an arbitrary Boolean matrix. The set of columns  $H$  of matrix  $L$  is called a covering if each row of the matrix  $L$  in crossing even with one of the columns in  $H$  gives 1. The covering is called irreducible if any its own

subset is not covering. Let  $c = (c_1, \dots, c_n)$  be the vector of weights of columns of the matrix  $L$ . The sum of columns weights of the covering is called a weight of covering. The covering, minimal on weight, is called a minimal covering. Note that in the case of unit column weights vector the minimal covering is the covering with minimum columns number. In the further it is considered, that the Boolean matrix has a unit vector of weights if it is not told opposite.

The training object  $S' = (a'_1, \dots, a'_n)$  generates the e.c.  $(H, \sigma)$ ,  $H = \{x_{j_1}, \dots, x_{j_r}\}$ ,  $\sigma = (\sigma_1, \dots, \sigma_r)$ , if  $\sigma_i = a_{j_i}$  under  $i = 1, 2, \dots, r$ . The set  $U_K = \{(H_1, \sigma_1), \dots, (H_{N_K}, \sigma_{N_K})\}$  of all e.c. of the class  $K$  is considered. The difference between the number of training objects from the class  $K$  that generate the e.c. and the number of training objects from the  $\bar{K}$  that also generate the same e.c. is called a weight of this e.c..

For pair of objects  $S'$  and  $S''$  we shall construct the binary vector  $B(S', S'') = (b_1, \dots, b_{N_K})$  where

$$b_j = \begin{cases} 1, & \text{if } P_{(H_j, \sigma_j)}(S') = 1 \text{ and } P_{(H_j, \sigma_j)}(S'') = 0 \\ 0, & \text{otherwise,} \end{cases}$$

$j = 1, 2, \dots, N_K$ . For the class  $K$  we shall construct the Boolean matrix  $L_K$  from all rows  $B(S', S'')$  such that  $S' \in K$  and  $S'' \notin K$ .

By constructing each column in matrix  $L_K$  corresponds to some elementary classifier from the set  $U_K$ . Let  $R$  be the of e.c. that corresponds to the column set  $H$  of matrix  $L_K$ . Following two statements are true.

**Statement 2.** The set of e.c.  $R$  is the monotone correct set for the class  $K$  if and only if  $H$  is the covering of the matrix  $L_K$ .

**Statement 3.** The set of e.c.  $R$  is the monotone irreducible (minimal) correct set for the class  $K$  if and only if  $H$  is the irreducible (minimal) covering of the matrix  $L_K$ .

In case of correct sets that is not monotone the set of all e.c.  $U'_K = \{(H'_1, \sigma'_1), \dots, (H'_N, \sigma'_N)\}$  is formed from parts of object descriptions from all classes. In this case the Boolean matrix  $L'_K$  is constructed from rows  $D(S', S'') = (d_1, \dots, d_N)$  where

$$d_j = \begin{cases} 1, & \text{if } P_{(H'_j, \sigma'_j)}(S') \neq P_{(H'_j, \sigma'_j)}(S'') \\ 0, & \text{otherwise,} \end{cases}$$

$j = 1, 2, \dots, N$  for all pairs of objects  $S' \in K$  and  $S'' \notin K$ .

Each column of the matrixes  $L_K$  and  $L'_K$  is associated with a weight of the according elementary classifier.

In the present work two models of recognizing procedures are constructed and investigated. The first model is founded on constructing the one correct set of e.c. which is close to minimal on complexity. The second model is founded on constructing the family of the most informative corrects sets of e.c.

As a rule, even for problems of small dimension the number of elementary classifiers is great and procedure of construction of the minimal correct set with use matrixes  $L_K$  and  $L'_K$  demands significant computing resources, therefore a question on development of effective methods of the decision of problems of algebra-logic correction of elementary classifiers. Because of NP-completeness of the set-covering problem the exact algorithms for the search of solution are practically inapplicable. For problems with big dimensions the approximate algorithms are used. The algorithms using the genetic approach concern to such algorithms.

In the present it is shown that the voting procedure on the one minimal correct set of e.c. cannot provide comprehensible quality of recognition. Therefore in the present work the model which builds family of logic proof-readers is offered. For constructing this model the genetic algorithm from [Sotnezov, 2008] is used. Thus in a population correct set of e.c. with good recognizing ability are selected.

### The construction of family of logic correctors on the basis of the genetic approach

The training sample  $T$  is divided on two subsamples: base ( $T_0$ ) and tuning ( $T_1$ ) according to a technique described in [Djukova, Peskov, 2005]. The sample  $T_0$  is used for construction matrixes  $L_K$  and  $L'_K$ , the sample  $T_1$  is used for an estimation of quality of recognition of correctors found by genetic algorithm.

The quality of recognition  $\tau(U)$  for the correct set of e.c.  $U$  of the class  $K$  is estimated under the formula

$$\tau(U) = \frac{1}{|T_1 \cap K|} \sum_{S \in T_0 \cap K} \sum_{S' \in T_1 \cap K} \delta(S, S', U) - \frac{1}{|T_1 \cap \bar{K}|} \sum_{S \in T_0 \cap \bar{K}} \sum_{S' \in T_1 \cap \bar{K}} \delta(S, S', U)$$

Thus, quality of recognition of the corrector  $U$  is equal to difference between the number of objects from the class  $K$  which are correctly recognized by the corrector  $U$ , and numbers of objects from other classes which also are recognized by the corrector  $U$  to the class  $K$ .

Work of genetic algorithm reminds development of a biological population in which each object, is characterized by a set of genes. Updating of such population eventually happens according to the law "survives the most adapted". Thus there is an opportunity of reception of new objects by means of operators of crossing and a mutation who in special way combine genes of parents.

Let  $L = (a_{ij})_{m \times n}$  be the Boolean matrix constructed on sample  $T_0$  and  $c = (c_1, \dots, c_n)$  is a vector of its column weights. The covering of the matrix  $L$  we shall represent as the integer vector  $Q = (q_1, \dots, q_m)$  where  $q_i$  is the number of one of columns which cover the  $i$ -th row.

By the greedy heuristic from [Sotnezov, 2009] the initial family of decisions  $P = (Q_1, \dots, Q_N)$  that is called a population is formed. Elements of the set  $P$  are called individuals.

For the individual  $Q_j$  the function of the fitness  $f(Q_j)$  describing quality of the found decision is defined. The individual  $Q_j$  describes the covering of the matrix which corresponds to some correct set of e.c.  $U_j$ . As a function of fitness  $f(Q_j)$  we shall use value

$$\tau(U_j) - \min_{i \in \{1, 2, \dots, N\}} \tau(U_i) + 1$$

For each individual  $Q_j$  from the population  $P$  the probability  $p_j$  is calculated by the formula

$$p_j = \frac{1/f_j}{\sum_{i=1}^N 1/f_i} \quad (5)$$

where  $f_j$  - the fitness of the individual  $Q_j$ .

On the next step of genetic algorithm with the set of probabilities  $\{p_i\}$   $i = 1, 2, \dots, N$ , two parental individuals  $Q^{(1)} = (q_1^{(1)}, \dots, q_m^{(1)})$  и  $Q^{(2)} = (q_1^{(2)}, \dots, q_m^{(2)})$  are selected. These individuals are used for generating child individual  $Q = (q_1, \dots, q_m)$  with the crossing operator by following rules.

Let  $f_1$  and  $f_2$  be fitnesses of individuals  $Q^{(1)}$  and  $Q^{(2)}$  accordingly, then the  $i$ -th component of the child  $Q$  is equal to

$$q_i = \begin{cases} q_i^{(1)}, & \text{with probability } \frac{c_{q_i^{(2)}} \cdot f_2}{c_{q_i^{(1)}} \cdot f_2 + c_{q_i^{(2)}} \cdot f_2} \\ q_i^{(2)}, & \text{with probability } \frac{c_{q_i^{(1)}} \cdot f_1}{c_{q_i^{(1)}} \cdot f_1 + c_{q_i^{(2)}} \cdot f_2} \end{cases}, \quad i = 1, 2, \dots, m$$

In difference from most often used one-point and two-point crossovers proposed operator of crossing considers the structure of parental individuals and their relative fitness. The more relative fitness of the parental individual has the bigger probability, that its gene will be copied in the descendant.

The using only the operator of crossing for updating the population can lead to formation of individuals with approximately identical set of columns. It means, that the algorithm converges in some local minimum in which neighbourhood there will be all new descendants. For overcoming local minima the operator of a mutation is used. This operator in random way changes (mutates) the given number of genes in the description of the child. As especially strong influence of the operator of a mutation on the individual should occur at convergence of the search process of the optimum decision it is offered to increase the number of mutating genes  $k(t)$  with growth of the number of algorithm steps according to the formula

$$k(t) = K \left( 1 - \frac{1}{C \cdot t + 1} \right),$$

where  $t$  - the number of the algorithm step,  $K$  and  $C$  are variable parameters which characterize the number of mutating genes on the last step of genetic algorithm and the speed of change of the mutating genes number accordingly.

After using of the operator of crossing and a mutation we receive the integer vector  $Q$  that corresponds to some covering  $H$  of the matrix  $L$ . If  $H$  is not the irreducible covering the procedure of feasibility restoration of the decision is applied. The procedure of feasibility restoration of the decision can be described as follows. Let  $M_j$  be the set of rows of the matrix  $L$ , covered by the column  $j$ , then

1. for each  $i \in \{1, 2, \dots, m\}$  the value  $w_i$  as number of columns from  $H$  covering  $i$ -th row is defined;
2. for each  $j \in H$  in decreasing order of weights the set  $w_i, i \in M_j$  is analyzed. If all such  $w_i$  is more than one the column  $j$  leaves the set  $H'$  and for each  $i \in M_j$  the value  $w_i$  is reduced by one;

As a result of the feasibility restoration procedure application the individual  $Q$  that corresponding to irreducible covering of the matrix  $L$  is received. The individual  $Q$  replaces one of individuals of the population  $P$  if 1) in a population there is not the individual identical to  $Q$  2) the set  $R_Q = \{Q' \in P \mid f(Q') \geq f(Q)\}$  is not empty.

For replacement in random way the individual from  $R_Q$  is chosen. The first condition is necessary for prevention of occurrence of identical individuals and, as consequence, degeneration of a population. The second condition means, that in a population the most adapted individuals (that is individuals with the least weights of coverings) is occurred.

The genetic algorithm stops the work if the population has been updated  $N_{\max}$  time. That means it has been received  $N_{\max}$  unique, more adapted individuals. The number  $N_{\max}$

The population received as a result contains the description of family of correctors with high recognizing ability. Thus the individual with the greatest fitness corresponds to the logic corrector with the maximal recognizing ability. The changing fitness function of the population individuals, it is possible to construct logic correctors of various type. For example, if as fitness function use weight of the corresponding covering of the expanded matrix of comparison, the genetic algorithm will give out the logic correctors close to minimal.

### The model testing on real problems

The model was tested for two cases. In the first case one minimal corrector, and in the second family of correctors was used. Problems for testing have been taken from repository UCI [Asuncion, Newman, 2007]. Characteristics of problems are resulted in Table 1.

Problem	The number of attributes	The number of objects in the first class	The number of objects in the second class
A	24	51	237
B	19	51	218
C	35	38	107
D	9	626	332
E	16	168	265

Table 1. Characteristics of test problems

The algorithm of voting on the corrector close to minimal (algorithm A2) and algorithm of voting on family of correctors (algorithm A1) were compared to the algorithms realized in system Recognition [Zhuravlev, Rjazanov, Senko, 2006] on the general percent of recognition  $R_1$  - the percent of number of correctly recognized control objects to the number of all control objects and weighed percent of recognition  $R_2$ , which are defined as follows. Let  $r_i$  and  $n_i$  be accordingly the number of correctly recognized objects and the total number of objects in the class  $K_i, i = 1, 2, \dots, l$ , then

$$R_1 = \frac{\sum_{i=1}^l r_i}{\sum_{i=1}^l n_i},$$

$$R_2 = \sum_{i=1}^l \frac{r_i}{n_i}.$$

In Table 2 for each compared algorithm on each problem there are specified: the general percent of recognition  $R_1$  (the first line of each cell), general percent of recognition for each of classes (the second line of each cell) and the weighed percent of recognition  $R_2$  (the third construction of each cell).

Recognizing algorithm	Problem A	Problem B	Problem C	Problem D	Problem E
A1	<b>84.91%</b> (71.43, 88.23) <b>79.83%</b>	<b>85.56%</b> (83.33, 86.08) <b>84.71%</b>	85.45% (75.0, 91.43) <b>83.21%</b>	99.2% (100, 97.87) 98.94%	<b>96.04%</b> (94.6, 97.0) <b>95.84%</b>
A2	67.20% (16.0, 80.0) 48.0%	63.64% (30.0, 82.86) 56.43%	54.64% (22.22, 62.03) 42.12	98.96% (100, 97.16) 98.58	93.87% (92.7, 95.4) 94.05%
Algorithm of calculation of estimations	80.2% (0.0, 100.0) 50.0%	72.7% (50.0, 85.7) 67.85%	81.4% (0.0, 100.0) 50.0%	72.7% (94.7, 34.8) 64.75	59.7% (0, 100) 50%
Binary decision tree	80.2% (0.0, 100) 50.0%	80.0% (45.0, 100.0) 72.5%	81.4% (0, 100) 50.0	63.4% (100, 0) 50.0	59.7% (0, 100) 50%
Logical patterns of classes	80.2% (42.9, 89.4) 62.65%	54.5% (20.0, 74.3) 47.15%	75.3% (50, 81) 65.5%	<b>99.5%</b> (100, 98.6) <b>99.3%</b>	50.3% (36.7%, 59.6%) 48.15%
SVM	81.1% (52.2, 89.2) 70.7%	80.0% (50.0, 97.1) 73.55%	<b>89.7%</b> (50.0, 98.7) 74.35%	63.4% (100, 0) 50%	59.7% (0, 100) 50%
Voting on irreducible tests	80.2% (47.6, 88.2) 67.9%	76.4% (40, 97.1) 68.55%	87.6% (72.2, 91.1) 81.65%	73.5% (77.5, 66.7) 72.1%	56.4% (71.7, 46.1) 58.9%

Table 2. Results of testing

---

## Conclusion

The algorithm of voting on family of correctors A1 considerably overcomes other considered algorithms on problems with small number of values in object attributes (problems D and E) on parameters  $R_1$  and  $R_2$ . On other problems the algorithm A1 due to consideration of each of classes separately also overcomes other algorithms on parameter  $R_2$ .

---

## Bibliography

- [Djukova, Zhuravlev, 1997] E.V. Djukova, Yu.I. Zhuravlev. Discrete Methods of Information Analysis in Recognition and Algorithm Synthesis // Pattern Recognition and Image Analysis. 1997. Vol.7. No.2. P. 192-207.
- [Djukova, Zhuravlev, 2000] Djukova E.V., Zhuravlev Yu. I. Discrete Analysis of Feature Descriptions in Recognition Problems of High Dimensionality. // Zh. Vychisl. Mat. Mat. Fiz. 2000, 40(8), 1264-1278 (2000) [Comput. Math. Math. Phys. 40(8), 1214-1227 (2000)]
- [Djukova, Zhuravlev, Rudakov, 1996] E.V. Djukova, Yu.I. Zhuravlev, K.V. Rudakov. On an algebra-logic synthesis of correct recognition procedures on base of elementary classifiers // Zh. Vychisl. Mat. Mat. Fiz. 1996 36(8), 215-223
- [Zhuravlev, 1978] Zhuravlev Yu.I. On an Algebraic Approach to Recognition or Classification Problems. // Problemy Kibernetiki, Moscow: Nauka, 1978, no.33, pp. 5-68 [in Russian]
- [Sotnezov, 2008] Sotnezov R.M. Genetic Algorithms in problems of discrete optimization and recognition, International Conference on "Pattern Recognition and Image Analysis: new Information Technologies",

Nizhni Novgorod, Russian Federation, 2008. V.2, P 173-175

[Sotnezov, 2009] Sotnezov R.M. Genetic Algorithms for Problems of Logical Data Analysis in Discrete Optimization and Image Recognition // Pattern Recognition and Image Analysis, 2009, Vol. 19, No 3, pp. 469-477

[Djukova, Peskov, 2005] Djukova E.V., Peskov N.V. Constructing recognition procedures on base of elementary classifiers// Problemy Kibernetiki, Moscow: Fizmatlit, 2005. no 14 p. 57-92 [in Russian]

[Asuncion, Newman, 2007] Asuncion A., Newman D.J. UCI Machine Learning Repository, University of California, Irvine. - 2007. [www.ics.uci.edu/~mllearn/MLRepository.html](http://www.ics.uci.edu/~mllearn/MLRepository.html)

[Zhuravlev, Rjazanov, Senko, 2006] Zhuravlev Yu.I., Rjazanov V.V., Senko O.V. «Recognition». Mathematic Methods. Programm System. Practical Applications // Moscow FAZIS, 2006. 176 p.

---

## Author's Information

---



**Djukova Elena Vsevolodovna** – postdoctoral (Doktor Nauk) degree in the physical and mathematical sciences, Major scientist, Dorodnycyn Computing Centre of the Russian Academy of Sciences, Vavilova st., 40, Moscow, 119333, Russia; e-mail: [edjukova@mail.ru](mailto:edjukova@mail.ru)

*Major Fields of Scientific Research: logical data analysis, pattern recognition, discrete mathmeatics, logical recognition procedures, computational complexity of discrete problems, synthesis of asymptotically optimal algorithms for solving discrete problems.*



**Zhuravlev Yuri Ivanovich** – academician of the Russian Academy of Sciences, postdoctoral (Doktor Nauk) degree in the physical and mathematical sciences, vice-president of Dorodnycyn Computing Centre of the Russian Academy of Sciences in scientific research, Vavilova st., 40, Moscow, 119333, Russia, e-mail: [zhur@ccas.ru](mailto:zhur@ccas.ru)

*Major Fields of Scientific Research: mathematical cybernetic and theoretical informatics; discrete analysis; theory of local algorithms of information processing; forecasting and recognition methods; development of mathematical methods of decision-making based on incomplete, contradictory, heterogeneous information*



**Sotnezov Roman Mihailovich** – post graduate student in Moscow State University, faculty of Computational Mathematic and Cybernetic, Leninskie gory, 1, Moscow, 119234, Russia; e-mail: [rom.sot@gmail.com](mailto:rom.sot@gmail.com)

*Major Fields of Scientific Research: pattern recognition, discrete mathematics, optimization problems*

## METHODS FOR EVALUATING OF REGULARITIES SYSTEMS STRUCTURE

Irina Kostomarova, Anna Kuznetsova, Natalia Malygina, Oleg Senko

**Abstract:** *The new method for analysis of regularities systems is discussed. Regularities are related to effect of explanatory variables on outcome. At that it is supposed that different levels of outcome correspond to different subregions of explanatory variables space. Such regularities may be effectively uncovered with the help of optimal valid partitioning technique. The OVP approach is based on searching partitions of explanatory variables space that in the best way separate observations with different levels of outcomes. Partitions of single variables ranges or two-dimensional admissible areas for pairs of variables are searched inside corresponding families. Output system of regularities is formed with the help of statistical validity estimates by two types of permutation tests. One of the problems associated with OVP procedure is great number of regularities in output system in high-dimensional tasks. The new approach for output system structure evaluating is suggested that is based on searching subsystem of small size with possibly better forecasting ability of convex combination of associated predictors. Mean error of convex combination becomes smaller when average forecasting ability of ensemble members becomes better and deviations between prognoses associated with different regularities increase. So minimization of convex combination mean error allows to receive subsystem of regularities with strong forecasting abilities that significantly differ from each other. Each regularity of output system may be characterized by distances to regularities in subsystem.*

**Keywords:** *Optimal partitioning, statistical validity, permutation test, regularities, explanatory variables effect, complexity*

**ACM Classification Keywords:** *H.2.8 Database Applications - Data mining, G.3 Probability and Statistics - Nonparametric statistics, Probabilistic algorithms*

---

### Introduction

In many applied forecasting or data analysis possibility of exact prognoses is connected with existing of subregions in explanatory variables  $X_1, \dots, X_n$  space where distributions of outcome variable  $Y$  significantly differ from its distributions in neighboring subregions or in whole data set. Variety of techniques exists now for revealing of such subregions: linear discriminant functions, classification or regression trees. It must be noted that these techniques are focused mainly on constructing single optimal forecasting algorithm. However in majority of applied tasks it is important not only to construct optimal predicting algorithm but also to receive most complete and valid empirical description of dependences of  $Y$  on  $X_1, \dots, X_n$ . Some approaches that allow to receive empirical descriptions of dependencies were developed during last decade. Method for searching systems of complete or partial logical regularities in pattern recognition tasks may be considered as an example[Ryazanov,2003]. The another approach is constructing of optimal valid partitions (OVP) of  $X$  variables space, that allow to achieve optimal separation of observation with different levels of  $Y$ .

The OVP models were previously developed in [Senko,1998], [Kuznetsova,2000], [Senko,2003]. The OVP procedures allow to calculate the sets of optimal partitions of one-dimensional admissible intervals of single variables or two-dimensional admissible areas of pairs of variables and to estimate statistical validity of regularities associated with these partitions. At that statistical validity is evaluated using two types of permutation



tests. Unlike traditional statistical criteria (Chi-square or ANOVA for example) permutation tests allow to evaluate statistical significance by the same dataset which was previously used for boundaries searching. One more advantage of permutation tests before alternative statistical technique is absence of necessity for any suppositions about variables distribution or any restrictions on groups sizes. OVP models may be applied for different types of outcome variables.

One of the problems associated with OVP procedure is great number of regularities in output system in high-dimensional tasks. In typical biomedical tasks with several tens independent variables and several hundreds cases number of discovered statistically valid regularities may achieve several hundreds. In such situations manual analysis of regularities systems is difficult. So development of new computer methods that would allow to evaluate interrelations between regularities, to reveal groups of similar regularities and recover internal structure of regularities system. The approach that is discussed in paper is based on calculating of mutual distances between regularities. Let  $r_1$  and  $r_2$  is pair of regularities that were found with the help of OVP method. Then two forecasting functions  $Z_1$  and  $Z_2$  may be put in correspondence to regularities as it is described below. Distance function  $P(r_1, r_2)$  between regularities  $r_1$  and  $r_2$  is defined as mathematical mean of squared difference between  $Z_1$  and  $Z_2$  or  $P(r_1, r_2) = P(Z_1, Z_2) = E_{\Omega}(Z_1 - Z_2)^2$ . Distance functions between regularities may be easily estimated by training information.

Various cluster analysis methods may be used for revealing clusters of similar regularities in case distance function  $P$  is defined. However clusterization methods are based only on distances and do not take into account prognostic ability of regularities. Besides previous experiments demonstrated that clusterization techniques tend to form classes that strongly differ from each other by size. So an alternative method was suggested that allow to select from system small number of regularities with possibly better forecasting ability of collective convex predictor. This subset that will be further referred to as basic subset or  $\tilde{Q}_B$ . It was shown previously ([A. Krogh and J. Vedelsby, 1995], [O.V.Senko,2009]) that forecasting ability of collective convex predictor depends both on forecasting ability of single ensemble members and mutual distances  $P(Z_i, Z_j)$  between predictors. At that mean error of convex combination decrease when average forecasting ability of ensemble improves and deviations between prognoses associated with different regularities grows. So minimization of convex combination mean error allows to receive a basic subset  $\tilde{Q}_B$  with following properties: a)  $\tilde{Q}_B$  consists of regularities with relatively strong forecasting abilities, b) regularities in  $\tilde{Q}_B$  significantly differ from one another in terms of distance  $P$ . Structure of output system may be characterized by distances to regularities from  $\tilde{Q}_B$ . In other words distances may be considered as new "coordinate" of regularities from initial set of regularities that have been found by OVP method. Thus structure of initial system of regularities may be evaluated.

---

### Optimal Partitioning

---

Let  $Y$  belongs to some set  $M_y$ . It is supposed that distance function  $\rho$  defined on Cartesian product  $M_y \times M_y$  satisfies following conditions:

- a)  $\rho(y', y'') \geq 0$ , b)  $\rho(y', y'') = \rho(y'', y')$ , c)  $\rho(y', y') = 0 \quad \forall y', y'' \in M_y$ .

The OVP methods are based on optimal partitioning of independent variables admissible regions. The partitions that provide for best separation of observations from dataset  $\tilde{S}_0$  with different levels of dependent variable are searched inside apriori defined families by optimizing of quality functional.

**Partitions families.** The partition family is defined as the set of partitions with limited number of elements that are constructed by the same procedure. The unidimensional and two-dimensional families are considered. The unidimensional families includes partitions of admissible intervals of single variables. The simplest Family I includes all partitions with two elements that are divided by one boundary point. The more complex Family II includes all partitions with no more than three elements that are divided by two boundary points. The two-dimensional Family III includes all partitions of two-dimensional admissible areas with no more than four elements that are separated by two boundary lines parallel to coordinate axes. Family IV includes all partitions of two-dimensional admissible areas with no more than two elements that are separated by linear boundary with arbitrary orientation relatively coordinate axes.

**Quality functionals.** Let consider at first standard OVP. Let  $\tilde{Q}$  is partition of admissible region of independent variables with elements  $q_1, \dots, q_r$ . The partition  $\tilde{Q}$  produces partition of dataset  $\tilde{S}_0$  on subsets  $\tilde{S}_1, \dots, \tilde{S}_r$ , where  $\tilde{S}_j$  ( $j = 1, \dots, r$ ) is subset of observations with independent variables vectors belonging to  $q_j$ . The evaluated  $Y$  mean value of subsets  $\tilde{S}_j$  is denoted as  $\hat{y}(\tilde{S}_j)$ . The integral quality functional  $F_I(\tilde{Q}, \tilde{S}_0)$  is

defined as the sum:  $F_I(\tilde{Q}, \tilde{S}_0) = \sum_{j=1}^r \rho[\hat{y}(\tilde{S}_0), \hat{y}(\tilde{S}_j)] m_j$ , where  $m_j$  - is number of observations in subset

$\tilde{S}_j$  .. The optimal value of quality functional in dataset  $\tilde{S}$  will be further referred to as  $F_I^o(\tilde{S})$ .

**Regularities validation.** For validation of found optimal partitions two types of permutation test is used. The first variant PT1 is used to test null hypothesis about independence of outcome on explanatory variables related to considered regularity. PT1 is used in cases when: a) significance of simplest regularities associated with partitions from family I is evaluated, b) significance of more complicated regularities is evaluated and no simplest valid regularities were previously discovered for related variables. The second variant PT2 is used to evaluate if more complicated partitions models are needed instead of simplest one to describe existing regularities. It tests null hypothesis about independence of outcome on explanatory variables inside suregions of explanatory variables space that are elements of partitions associated with simplest regularities.

---

### Forecasting associated with regularities

---

Examples of partitions describing regularities are given at figures 1 and 2. Difference between distributions of various biomedical parameters was evaluated in groups of patients in light and severe stage of encephalopathy Sparse diagram 1 represent regularity associated with relationship between encephalopathy stage on age group number (axe X). Statistics for each quadrant are give at the left part of sparse diagram. It is seen that fraction of light forms decreases as group number increases. Then probability of light form for patients from quadrant j is

calculated by forecasting function  $Z(s, j) = \frac{n^+(j)}{n^+(j) + n^o(j)}$ . It is seen from sparse diagram 1 that for

quadrant I number of patients in light stage  $n^+$  is equal 32 and number of patients in severe stage  $n^o$  is equal 11 and  $Z(s, I) = 0.744$

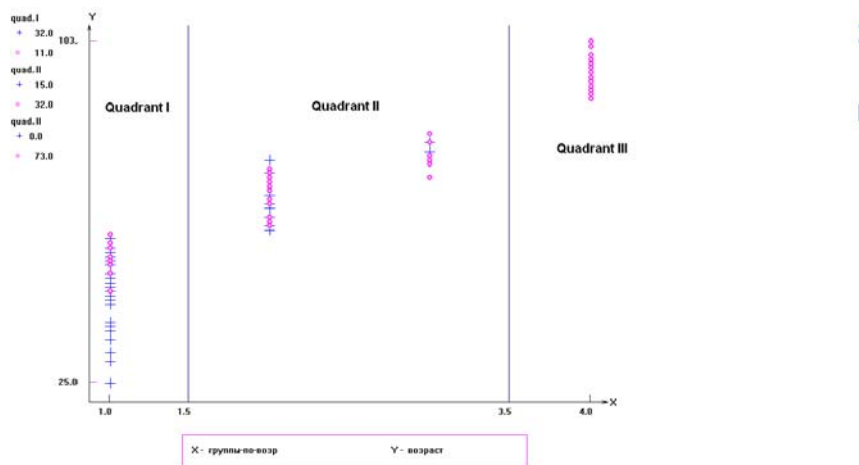


Fig. 1 .Optimal partition describes dependence of encephalopathy form on age (axe X) of patients and existing of severe disorder of cerebral blood circulation (axe Y). Partition belongs to the family III. Patients with severe form are denoted by It is seen that

Sparse diagram 2 represent regularity associated with relationship between encephalopathy stage and pair of input variables: age (axe X) and existing of severe disorder of cerebral blood circulation (axe Y). It is seen that light stage of encephalopathy predominates over severe stage only in quadrant IV corresponding to patients younger 70.5 years without severe disorder of cerebral blood circulation: number of patients in light stage  $n^+$  is equal 43 , number of patients in severe stage  $n^o$  is equal 2 and  $Z(s, IV) = 0.955$  .

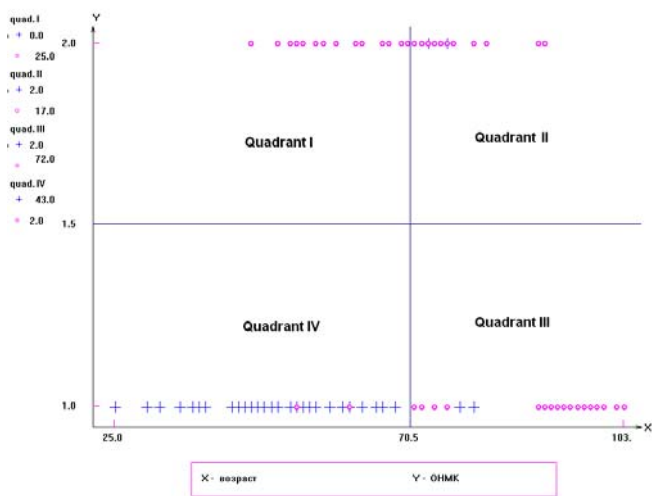


Fig. 2 .Optimal partition describes relationship between encephalopathy stage and pair of variables: age (axe X) of patients and existing of severe disorder of cerebral blood circulation (axe Y). Partition belongs to the family III. Patients with severe form are denoted by o and patients with light form are denoted by +.

---

### Generalized Error Functional

---

Let  $\tilde{Z} = \{Z_1, \dots, Z_l\}$  is set of prognostic variables (predictors) forecasting outcome variable  $Y$  for objects that are elements of some probability space  $\Omega$ . The convex corrector  $\hat{Z}_{ccp}(\mathbf{c})$  at  $\tilde{Z}$  is defined as  $\hat{Z}_{ccp}(\mathbf{c}) = \sum_{i=1}^l c_i Z_i$ , where  $\sum_{i=1}^l c_i = 1$  and  $c_i \geq 0, i = \overline{1, l}$ . The squared error of forecasting at general set will be denoted as  $\delta$ . Let distance between two predictors  $Z_{i'}$  and  $Z_{i''}$ . The squared error for  $\hat{Z}_{ccp}(\mathbf{c})$  may be represented as

$$\delta[\hat{Z}_{ccp}(\mathbf{c})] = \sum_{i=1}^l c_i \delta(Z_i) - \frac{1}{2} \sum_{i'=1}^l \sum_{i''=1}^l c_{i'} c_{i''} P(Z_{i'}, Z_{i''}) \quad (1)$$

where  $P(Z_{i'}, Z_{i''}) = E_{\Omega}(Z_{i'} - Z_{i''})^2$ .

It is seen from (2) that  $\delta[\hat{Z}_{ccp}(\mathbf{c})]$  is always lower than  $\sum_{i=1}^l c_i \delta(Z_i)$  and difference between them increases as increase distances between prediction. In case when contributions of all predictors (1) take form

$$\delta[\hat{Z}_{ccp}(\mathbf{u}_l)] = \frac{1}{l} \sum_{i=1}^l \delta(Z_i) - \frac{1}{2} \frac{1}{l^2} \sum_{i'=1}^l \sum_{i''=1}^l P(Z_{i'}, Z_{i''}) \quad (2)$$

where  $\mathbf{u}_l = (\frac{1}{l}, \dots, \frac{1}{l})$  is  $l$ -dimensional vector.

---

### Optimal Subset Selecting

---

At the initial stage optimal system of regularities  $\tilde{Q}_o$  is searched using described previously optimal partitioning method. Then optimal subset selecting (OSS) procedure may be used to find  $\tilde{Q}_B$ .

Step 1. At initial step squared error  $\delta$  is evaluated for each predictor associated with regularity from system  $\tilde{Q}_o$ . At that leave-one-out cross validation technique is used. Regularity  $q_1^{best} \in \tilde{Q}_o$  corresponding to predictor with minimal error  $\delta_1^{\min}$  is added to optimal subset.

Step 2. The squared error functional (1) is calculated for pairs of predictors associated with pairs of regularities from set  $\{(q_1^{best}, q) \mid q \in \tilde{Q}_o \setminus q_1^{best}\}$ . Let minimal value of functional (1)  $\delta_2^{\min}$  is achieved for pair  $(q_1^{best}, q_2^{best})$ . In case  $\delta_2^{\min} < \delta_1^{\min}$  regularity  $q_2^{best}$  is added to optimal subset.

Step k. The squared error functional (1) is calculated for sets of predictors associated with series of regularities from  $\{(q_1^{best}, \dots, q_{k-1}^{best}, q) \mid q \in \tilde{Q}_o \setminus \{q_1^{best}, \dots, q_{k-1}^{best}\}\}$ . Let minimal value of functional (1)  $\delta_k^{\min}$  is achieved for series  $(q_1^{best}, \dots, q_k^{best})$ . In case  $\delta_k^{\min} < \delta_{k-1}^{\min}$  regularity  $q_k^{best}$  is added to optimal subset. Otherwise forming of optimal regularities subset is finished and series  $(q_1^{best}, \dots, q_k^{best})$  is fixed as basic subset.

$\tilde{Q}_B$ .

---

### Experiment

---

Performance of developed method was evaluated in task of computer diagnostics of encephalopathy severity. Group of 47 patients in early (first) stage was compared with group of 116 patients in severe stage (third) by 122 input clinical or biomedical indicators. At initial stage of research OVP method was used to find statistically valid correlations between severity and levels of input variables. As a result 300 regularities belonging to families

I, II, III was revealed at statistical significance level  $p < 0.05$ . Previously described procedure of optimal regularities subsystem was used. Basic subset  $\tilde{Q}_B$  consisting of 3 regularities was found:

two-dimensional regularity associated with relationship between encephalopathy stage and pair of input variables: age (axe X) and existing of severe disorder of cerebral blood circulation that is diagnosed by magnetic resonance tomography (axe Y).

regularity associated with relationship between encephalopathy stage on age group number (axe X)

- a) two-dimensional regularity associated with relationship between encephalopathy stage and pair of input variables: existing of severe disorder of cerebral blood circulation that is diagnosed by magnetic resonance tomography and MPT LO.

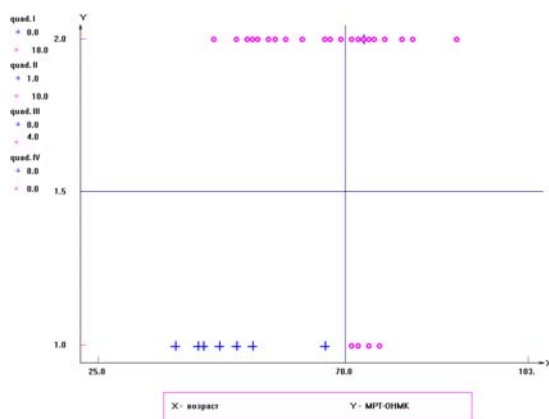


Fig. 3 .Optimal partition describes relationship between encephalopathy stage and pair of variables: age (axe X) of patients and existing of severe disorder of cerebral blood circulation (axe Y) diagnosed by magnetic resonance tomography. Partition belongs to the family III. Patients with severe form are denoted by o and patients with light form are denoted by +.

Forecasting errors associated with these 3 regularities and distances between them are given in next table 1.

Table 1

	$\delta$	1	2	3
1	0.026	0	0.315	0.029
2	0.148	0.315	0	0.296
3	0.053	0.029	0.296	0

It is seen from table that prognostic abilities of regularities 1 and 3 from  $\tilde{Q}_B$  are much better than prognostic ability of regularity 2. But distance of regularity 2 from regularities 1 and 3 is great. Distances to regularities from  $\tilde{Q}_B$  were calculated for each regularity from all regularities from system  $\tilde{Q}_o$ , and 20 nearest neighbors were found for each element of  $\tilde{Q}_B$ . It is appeared that regularities close to (a) and (c) are mainly related to existing severe disorder of cerebral blood circulation.

### Conclusion

The new method for analysis of regularities systems was represented. Method is based on searching of regularities subsystems  $\tilde{Q}_B$  in initial system  $\tilde{Q}_o$  with the best collective forecasting ability according

expression (2). Algorithm for regularities subsystems  $\tilde{Q}_B$  searching in initial system  $\tilde{Q}_o$  is represented. The developed technique was tested at the high-dimensional task of computer diagnostics of encephalopathy severity where using OVP procedure discovered great number of regularities at significance level  $p < 0.05$ .

---

### Acknowledgment

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA ( [www.ithea.org](http://www.ithea.org) ) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine ( [www.aduis.com.ua](http://www.aduis.com.ua) ).

---

### Bibliography

- [V.V.Ryazanov ,2003] Ryazanov V.V. About some approach for automatic knowledge extraction from precedent data // Proceedings of the 7th international conference "Pattern recognition and image processing", Minsk, May 21-23, 2003, vol. 2, pp.35-40.
- [Gorman, 2001] T.W. O'Gorman An adaptive permutation test procedure for several common test of significance. Computational Statistics & Data Analysis. 35(2001) 265-281.
- [Senko, 2003] Senko O.V., Kuznetsova A.V., Kropotov D.A. (2003). The Methods of Dependencies Description with the Help of Optimal Multistage Partitioning. Proceedings of the 18-th International Workshop on Statistical Modelling Leuven, Belgium, 2003, pp. 397-401.
- [Sen'ko, 1998] Sen'ko O.V., Kuznetsova A.V. (1998). The use of partitions constructions for stochastic dependencies approximation. Proceedings of the International conference on systems and signals in intelligent technologies. Minsk (Belarus), pp. 291-297.
- [Kuznetsova, 2000] Kuznetsova A.V., Sen'ko O.V., Matchak G.N., Vakhotsky V.V., Zabolina T.N., Korotkova O.V. The Prognosis of Survivance in Solid Tumor Patients Based on Optimal Partitions of Immunological Parameters Ranges //J. Theor. Med., 2000, Vol. 2, pp.317-327.
- [Sen'ko, 2006] Oleg V.Senko and Anna V. Kuznetsova, The Optimal Valid Partitioning Procedures . Statistics on the Internet <http://statjournals.net/>, April, 2006
- [O.V.Senko,2009]. An Optimal Ensemble of Predictors in Convex Correcting Procedures// Pattern Recognition and Image Analysis, MAIK Nauka/Interperiodica. 2009, Vol. 19, No. 3, pp. 465-468.
- [A. Krogh and J. Vedelsby, 1995] Neural network ensembles, cross validation, and active learning. NIPS, 7:231–238, 1995.

---

### Authors' Information

**Senko Oleg Valentinovich** – Leading researcher in Dorodnicyn Computer Center of Russian Academy of Sciences, Russia, 119991, Moscow, Vavilova, 40, [senkoov@mail.ru](mailto:senkoov@mail.ru)

**Kuznetsova Anna** – senior researcher in Institute of Biochemical Physics of Russian Academy of Sciences, 117997, Kosygina, 4, Moscow, Russia, [azfor@narod.ru](mailto:azfor@narod.ru)

**Malygina Natalia Aleksandrovna** – chief of laboratory in Russian Clinical-Research Center of Gerontology, principal scientific officer of laboratory of population ageing genetic, 129226, Leonova str. 16 build. 2, Moscow, Russia

**Kostomarov Irina Victorovna** - Russian Clinical-Research Center of Gerontology, chief of laboratory of population ageing genetic, 129226, Leonova str. 16 build. 2, Moscow, Russia [erri06@rambler.ru](mailto:erri06@rambler.ru)

---

## GROWING SUPPORT SET SYSTEMS IN ANALYSIS OF HIGH-THROUGHPUT GENE EXPRESSION DATA

**Arsen Arakelyan, Anna Boyajian, Hasmik Sahakyan,  
Levon Aslanyan, Krassimira Ivanova, Iliya Mitov**

**Abstract:** *Genome wide expression analysis with DNA microarrays has become a mainstay of genomics research. The problem with microarrays is that there is no “gold standard” for microarray-generated data analysis. In most cases, traditional statistical and pattern recognition approaches are not productive in analysis of high dimension low sample size data analysis, and improvements, alternative approaches are currently being developed in order to effectively analyze and interpret microarray data. In this study we design and extend logic-combinatorial scheme that can be used in high-throughput gene expression data. Applied results show that proposed algorithm is able accurately discriminate different biological phenotypes, although some improvements should be further made.*

**Keywords:** *Pattern Recognition, Functional Pathways, gene expression, post traumatic stress disorder*

**ACM Classification Keywords:** *1.5 Pattern Recognition, 1.5.2 Design Methodology*

---

### Introduction

Genome wide expression analysis with DNA microarrays has become a mainstay of genomics research. It is obvious, that monitoring expression levels for thousands of genes at a time provides insights into cellular processes and responses that cannot be obtained by looking at one or a few genes. Traditional methods for gene expression measurements such as Northern blots can be time-consuming and labor-intensive and are not practical for application on a very large scale. The more global view and increased throughput made possible by the advent of parallel expression measurements with DNA microarrays has therefore opened a new window on cellular activity [Hill et al 2000; Shoemaker et al 2001]. Being introduced in early 1990th, the rapidly increasing popularity of these techniques is evidenced by the number of publications involving microarrays. Gene expression differences between two samples have been applied to disease diagnosis and classification [Bittner et al 2000]. Effects of reagents or drugs on gene expression patterns have been used to test drug efficacies and to determine pharmacological mechanisms [Namba et al 2001; Dan et al 2002]. Correlated mRNA expression profiles under different cellular conditions have been used to predict gene functions [Hughes et al 2000]. DNA and RNA quantifications have also been used for detecting bacterial and viral pathogens, as well as host-pathogen interactions [Grayson et al 2002]. However having undoubted advantages, DNA arrays also has disadvantages that at the moment limit their usage. Currently, both oligonucleotide and spotted cDNA arrays are hybridized and read one at a time and significant time and effort is required to process even a modest number of samples. In many cases, a small number of experiments that cover thousands of genes is not sufficient. It has become increasingly clear that large collections of expression results are much more than the sum of their parts. The analysis of multidimensional expression patterns can reveal new insights that may not be apparent when looking at the results from small numbers of samples [Hughes et al 2000; Ross et al 2000].

Second problem with microarrays are that there is no “gold standard” for microarray-generated data analysis. From the very beginning the primary interest was to identify differentially expressed genes and elucidate related biological processes. The most widely used approach is individual gene analysis which evaluates the significance of individual genes between two groups of samples compared. This type of analysis typically yields a list of

altered genes from a cutoff threshold. Generally the basic approach of analysis is mathematical statistics (MS). Having satisfactory amount of experimental data (statistics) it helps to form conclusions that some properties and postulations take place in some probabilistic level. Simple correlation, regression and hypothesis estimation algorithms are components of the statistical approach. However, strong normality, and independence assumptions make them impractical and not powerful enough. A different situation appears in area of pattern recognition (PR). There is no satisfactory statistics in this case. These heuristics are more responsible and conditional. Learning set is given as a limited number of known classifications but it has to be large enough to describe the class properties in application area. A number of basic approaches are known in PR - Metric Algorithms, Logic Separation (LS), Neural Networks, etc. One of the well-known classes of metric algorithms is the voting (or estimation calculation) model [Zhuravlev 1998]. This is an algorithmic model with a number of additional parameters, requiring optimization during the learning stage. Several improvements and alternative approaches were developed, based on clustering techniques [Divina and Aguilar-Ruiz 2006; Von Borries 2008], support vector machines [Brown et al 2000; Benito et al 2004], "cut-off free" gene ranking [Subramanian et al 2005], dimension reduction [Nicolau et al 2007], and so on, but there is still a room for improvements of existing and development of new algorithms for analysis of high-throughput gene expression data. In this study we design and extend logic-combinatorial scheme for microarray data analysis.

Several assumptions related to the biological nature in terms of classes, features and learning sets were used defining the algorithmic model:

1. Given a dataset of features (gene expression values) and objects characterized by them, and objects are classified in two or more groups (healthy vs. diseased, different treatments, etc.). It is supposed that the biological process exists that is responsible for the differences of classes in terms of phenotype (blue vs. brown eyes, high vs. normal biological activity).
2. If such process exists, features (groups of features) related to that process should be able discriminate given classes of objects.
3. A feature set has higher probability to accurately classify objects, if its subsets, basically, are known to separate these classes.
4. As more features sets representing same biological process is found, as easier and accurate to identify the process itself.

---

## Algorithm

---

Formal mathematical model considered in this article refers to the natural classification of objects characterized by the sets of features. Objects are characterized by numerical values of features. We suppose a learning set is given which is a list of example objects with their feature assignments, and their properties in terms of membership to some nonintersecting classes. The global problem in area is how to learn classification by limited learning set. Our primary target is in identifying the features sets of most effective functionality in terms of considered classification. In some point of view we consider a typical problem of applied mathematical statistics or a basic pattern recognition postulation – the supervised learning.

This is true in part and our model gains additional specifics, coming with gene expression data. First of all it is to take into account the extremely large number of features appears here - genes and their expressions. Formally with limited and poor learning set large number of groups of these features may appear that function similarly in object classification. Redundancy appears as a consequence of the learning set limitations vs. very large feature set. If learning set size  $\pi$  is given, then a corresponding number  $\varphi$  of features can effectively participate in process of partitioning the object set into the subsets.  $\log \pi$  plays an important role in this. This is minimal power to split the set to separate objects, and this is upper estimate of learning set size because of groups are



supposed not separate objects. Determining such feature  $\varphi$ -sets is the primarily goal of the algorithm we construct. This part is a typical task in pattern recognition area. The technique we apply is the support set systems generation. This technique is introduced first by [Zhuravlev 1998] and now is known in data mining area as frequent subsets [Aslanyan and Sahakyan 2009]. Our algorithm applies a supervised search procedure to construct the set of all support  $\varphi$ -sets of features. Besides this we have to deal with another problem which is more specific, and nonstandard. The applied problem of classification with gene expression data deals with very large sets of objects. The real support sets on these data are large and due to limitations of practical/experimental learning sets we can see only the  $\varphi$ -subsets of them. The same time due to redundancy the quality  $\varphi$ -subsets are mixed with the noisy features and no separation of features is possible due to low learning and large feature sets. Fortunately the application problem helps. The concept of functional pathways is applied. We code a pathway as the corresponding set of features in that pathway. Among several hundreds of pathways considered in a problem only several may correspond to the given classification. We take a hypothesis that such pathways have higher intersections with found support  $\varphi$ -subsets. This is because of real supportive  $\varphi$ -subsets always appear although the noisy ones appear spuriously.

Finally, the situation we are with gene expression data doesn't obey the regular requirement neither statistics nor the pattern recognition. Extremely small learning set and extremely large feature set creates a situation when there is not enough statistical information to treat the hypotheses and when there is a set of features that doesn't allow applying analytic tools to recognize the valid feature subsets. This situation is known in high-throughput gene expression data analysis. Microarray data is known as high dimension low sample size data (HDLSS) [Von Borries 2008] and this specific class of experimental data sets initializes a new algorithmic research area. Current study applies theoretical and algorithmic issue of set systems to solve the problems with HDLSS. In general description we grow the support sets with quality estimations, then apply this estimates to restrict the sets to computability sizes, and after recurrent  $\varphi$ -subsets generation continue through evaluation stage of subsets given by functional pathways.

Consider gene expression data  $\mathfrak{S}$  which is a numeric array of  $n$  columns representing samples and  $D$  rows corresponding to genes.  $\mathfrak{S} = \{S_1, S_2, S_3, \dots, S_n\}$ , where  $S_i$  is a  $1 \times D$  column containing gene expression data for sample  $i$ . HDLSS means  $n \ll D$ . Classes which are 2 w.l.o.g. consist of objects:  $S_1, S_2 \dots S_k$  - belonging to the first class (e.g. disease class), and  $S_k, S_{k+1} \dots S_n$  - samples belonging to second class (e.g. healthy class). It is evident that almost any unique row  $S_1(i), S_2(i), \dots, S_n(i)$  of  $\mathfrak{S}$  can correctly classify the two classes by linear hyperplane or by some other similar classification mean. The number of such rows might be very large among the  $D$ . The same time, it is realistic that different sets of rows are classifying the classes differently. Formally, a collection of subsets of the set  $\{1, 2, \dots, n\}$  is known as a set of support systems  $\Omega$  [Zhuravlev 1998]. In our model support system is the "unit" of comparison of object description pairs. This is when a set of distances, - each by a member of  $\Omega$  is considered, which collectively describe the differences among the objects of different classes. The application counterpart is that support set is a minimal set of features - so that no any smaller and larger feature set is describing particular classification in a higher approximation. This brings to the problem of determining the proper row subsets (support systems), which provide maximal differences between classes (quality vs. accuracy of classification). In doing this we will eliminate the equivalent (in some sense) rows from one side; and will compose the sets of rows representing different equivalency subsets as approximations to the proper support systems.

## 1. Classifiers

At first we define **Elementary classifiers**.

These are hyperplane (or neural or any similar) classifiers by small number of rows. **1\_classifier** is defined through one single row and its expression values  $S_1(i), S_2(i), \dots, S_n(i)$ . Denote by  $t(i)$  the classification level/accuracy of this attribute. 1\_classifiers  $c_1(i)$ ,  $i \in \overline{1, D}$  ranked by classification levels are truncated by a version of sigma's rule. We leave out low classification genes from any further consideration. **2\_classifiers** consider the pairs of genes and expression values. Logically 2\_classifiers are to be composed by pairs of genes higher ranked by corresponding 1\_classifiers. 2\_classifiers and in general **k\_classifiers** consider any  $k$  rows, construct "hyperplanes" and define structures  $c_k(i_1, \dots, i_k)$  and  $t_k(i_1, \dots, i_k)$ . They construct convex hulls in areas of two considered classes; consider their partition hyperplane, rank by the classification level and narrows if necessary the set of rows used in future algorithmic steps.

## 2. Growing Support Systems

Among  $2^n$  elementary classifiers mentioned above, we look for those, which correspond to subsets of genes that are more differently expressed in classes. The simplest way is to start by a 1\_classifiers, and grow them step by step to  $k$ \_classifiers so that the classification level is strictly increasing. Any  $k$ \_classifier may be considered as a composition of one  $k - 1$ -classifier together with one new row. Concepts  $c_k(i_1, \dots, i_k)$  and  $t_k(i_1, \dots, i_k)$  in this way introduce monotony relation between gene sets, putting them into 1-1 correspondence to the vertices of an  $n$  dimensional unit cube. However, practical implementation might be rather hard because of for large  $k$  it will become impossible to consider all  $2^k$  sub-classifiers. The search area for these subsets is very large, and appropriate heuristics to combat this complexity is necessary.

We consider several heuristics:

- Sorting 1\_classifiers by decreasing forces  $t_1(i)$ , and eliminating from the further treatment rows with forces lower than the threshold selected. Let the rows in sorted sequence are as  $i_1, i_2, \dots, i_k, \dots, i_D$ . An important point of this sequence is the first index  $j$  where  $t_k(i_1, \dots, i_k)$  is increasing in area  $i_k < j$  and this increase interrupts at the point  $j$ . Sorting is applied to the composite multi-feature classifiers because of the comparability of their classification measures.
- Consider an arbitrary elementary classifier  $c_k(i_1, \dots, i_k)$ . Compose an  $n$  dimensional binary vector, assigning its coordinates  $i_1, \dots, i_k$  as 1. Completing by 0 all reminder coordinates, we create 1-1 correspondence between classifiers and  $n$ -cube vertices. Applying hierarchical clustering in  $n$ -cube layers we split  $k$ -classifiers into the groups by the equivalency relation (after some cut of dendrogram). Similarity measure used in clustering is some correlation between the hyperplanes (their coefficient vectors). We consider the representative sets of clusters. Some of them may give the same force of classifying objects by gene expressions as the whole descriptive table does. In this way we reduce the dimensionality combating the exponential explosion for large  $n$ .
- As it was mentioned, 1\_classifiers might be directly sorted by their forces. Any  $k$ \_classifier may be considered as a composition of one  $k - 1$ -classifier  $c_k(i_1, \dots, i_{k-1})$  together with one new row  $i_k$ . In terms of class vectors this change means concatenation of a new dimension in direction  $i_k$ . Concepts  $c_k(i_1, \dots, i_k)$  and  $f_k(i_1, \dots, i_k)$  in this way introduce monotony relation between gene sets in the same way as the vertices of  $n$  dimensional unit cube which are in 1-1 correspondence to elementary classifiers. Considering subsets of different  $n$ -cube layers and taking into account monotony we may apply the chain split technology [Aslanyan and Sahakyan 2009] in finding the best separating gene sets.

It is important to note that chain split (and other known frequent subsets growing algorithms of association rule mining) work on systematized structure of all objects which is a hard computationally. Instead, the representative set mentioned above are a valuable heuristic that may help in reducing the computational complexity in growing.

### 3. Approximating Support Systems

Mathematical model and algorithms we consider have to implement the application area hypotheses that we defined above. The formal picture that satisfies the hypotheses is a monotone Boolean function over the set of gene expressions. Upper zeros of this function corresponds to the real support sets. Hypotheses say that having very large and satisfactory data the correct growing procedure will achieve these support sets. Having smaller learning sets procedures will deduce some partial of approximate collections of support sets. It is a hard question how to evaluate the quality of these sets. Formal postulation might be an approximation scheme for monotone Boolean functions. Two schemes were considered in this regard:

- Frequency based approach. For an arbitrary binary vector  $\alpha$  (a unit cube vertex) vertices of layers of subcube between 0 and that vertex are considered. Frequencies of support sets after the growing stage are computed. Vertex  $\alpha$  is determined as a recognized support set if the frequencies are over a given threshold.
- Chain split and interior/exterior sets approach. The chain split recognition of monotone Boolean function is considered [Aslanyan and Sahakyan 2009]. The usual procedure intends to find the exact function and Hansel chains are ideal for this. Consider the extension of chain split with rare chains. Chains are with edges that connect comparable but distance vertex pairs and their relative completions similarly defined, or the same Hansel chains are considered but recognized areas are extended in iterations to the interior and exterior vertices of monotone function by the given threshold. Growing support sets means one way recognition that is recognition through consideration of only (or basically) zero values of function. The same time this is constrained recognition in a sense that hypothetical pathways involve limited number of genes.

---

### Algorithm Realization

Algorithm was realized using Matlab R2008a (Mathworks, Inc). At each iteration selection the list of classifiers and corresponding errors of classification are generated. In order to decrease computational intensity we used very strict cut-off for classifier selection  $M(err) - 2 * SD(err)$ . Only classifiers with error less than specified threshold were considered for future growing. Search of functional pathways were performed KEGG (Kyoto Encyclopedia of Genes and Genomes, [www.genome.jp/kegg/](http://www.genome.jp/kegg/)) database using the SOAP/WSDL based web service.

---

### Dataset Description and Preprocessing

To test the functionality of proposed algorithm dataset GDS1020 on gene expression patterns related to post-traumatic stress disorder (PTSD), publicly available in Gene Expression Omnibus repository (<http://www.ncbi.nlm.nih.gov/geo/>) was used. Description of samples was retrieved from dataset summary. GDS1020 dataset contains gene expression profiles of peripheral blood mononuclear cells (PBMC) isolated from the blood of trauma survivors at admission to emergency room and at 4 month follow-up. Overall, 33 (16 control and 17 PTSD subjects) gene expression profiles were available in dataset. PTSD affected subjects persistently manifested full criteria for PTSD.

From human dataset probes that had at least 50% of present calls were chosen. The missing values were replaced by group mean. Normalization was performed by division of gene expression in individual samples to geometric mean of gene expression of the all samples. The probe identifiers from each platform were converted to the HUGO-approved gene symbols, averaging log transformed expression values of multiple probes targeting the same gene.

## Results and Discussion

During data preprocessing 12568 probes were collapsed in 8742 genes that were used in future analyses. Search for the 1-classifier identified 405 genes with classification error less than specified threshold. From selected 405 genes only 149 were included in 77 KEGG functional pathways. These 405 genes were further used for identification of k-classifiers. The iterations stop at 4<sup>th</sup> step (group of 4 genes) with final classification error 0. During the analysis 605 2-classifiers, 449 3-classifiers and 2270 4-classifiers were formed. Figure 1 shows the increase of classification accuracy in parallel with classifier order increase. When we performed search of KEGG functional pathways using formed classifiers, we have identified only 6 pathways that exactly matched with 2-classifiers and 2 pathways exactly matched with 3-classifiers. None of pathways contain all genes from any of 4-classifiers.

The results showed that the group of 4 genes is enough for precise separation of classes in our case. However, we could not identify any KEGG functional pathways that include these 4-classifiers. Here, several suggestions can be made. First of all, this approach identifies “best genes”, i.e. genes with maximal classification force. If the classification force of other genes from the same pathway is less profound they will not be included in 1-classifiers. Thus, this is very useful tool for the search of biomarker combinations that are able to discriminate classes with maximal accuracy, which is not possible with the use of single biomarker. Second, this can be the result of limited number of pathways annotated in KEGG database. At the moment KEGG database contain information only about 206 pathways. Although all 405 genes were annotated in KEGG, only 149 (36.7 %) of them were associated with one or more functional pathways. In order to check this suggestion, we re-evaluated our data using gene ontology (GO, [www.geneontology.org/](http://www.geneontology.org/)) database to search terms covered by 4-classifiers. In contrast to KEGG, we identified 841 gene ontology terms that exactly matched 4-classifiers. And finally, it is possible that dataset used in study is not valid input for this approach. The question of validity of initial data are of great importance, however, at the moment we do not have methodology to evaluate their adequacy to the considered problem of functional pathways analysis. In this regard, interpretations given above are conditional and larger and valid datasets needed to be considered.

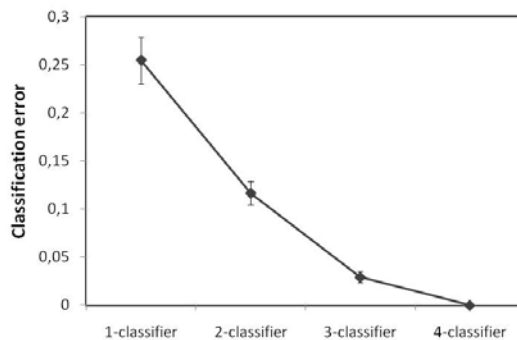


Figure 1. Classification accuracy of growing support sets.

---

## Conclusion

---

Genome wide expression analysis models and algorithms were considered. It is recognized that in most cases, traditional statistical and pattern recognition approaches suffer to analyze these information due to disbalance between the numbers of features and objects. Traditional in these area high dimension low sample size data (HDLSS) were considered and logic-combinatorial pattern recognition is applied to analysis of this information. The structured algorithm is similar to frequent sets algorithm in data mining. After growing support sets approximation of maximal support sets is considered in a model of monotone Boolean recognition. Applied results show that proposed algorithm is able accurately discriminate different biological phenotypes, although some improvements should be further made.

---

## Bibliography

---

- [Hill et al 2000] A.A. Hill, et al. Genomic analysis of gene expression in *C. elegans*. *Science*, 2000, 290: 809–812.
- [Shoemaker et al 2001] D.D. Shoemaker, et al. Experimental annotation of the human genome using microarray technology. *Nature*, 2001, 409: 922–927.
- [Bittner et al 2000] M. Bittner, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 2000, 406(6795): 536-540.
- [Namba et al 2001] H. Namba, et al. Efficacy of the bystander effect in the herpes simplex virus thymidine kinase-mediated gene therapy is influenced by the expression of connexin43 in the target cells. *Cancer Gene Ther*, 2001, 8(6): 414-420.
- [Dan et al 2002] S. Dan, et al. An integrated database of chemosensitivity to 55 anticancer drugs and gene expression profiles of 39 human cancer cell lines. *Cancer Res*, 2002, 62(4): 1139-1147.
- [Hughes et al 2000] T.R. Hughes, et al. Functional discovery via a compendium of expression profiles. *Cell*, 2000, 102(1): 109-126.
- [Grayson et al 2002] T.H. Grayson, et al. Host responses to *Renibacterium salmoninarum* and specific components of the pathogen reveal the mechanisms of immune suppression and activation. *Immunology*, 2002, 106(2): 273-283.
- [Ross et al 2000] D.T. Ross, et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet*, 2000, 24(3): 227-235.
- [Zhuravlev 1998] Yu. Zhuravlev. Selected research publications. Magistr, Moscow, 1998, 420p (in russian).
- [Von Borries 2008], G.F. Von Borries, Partition clustering of high dimensional low sampling size data base on p-values, PhD dissertation, Kansas State University, 2008, p. 139.
- [Divina and Aguilar-Ruiz 2006] F. Divina and J. S. Aguilar-Ruiz. Biclustering of expression data with evolutionary computation. *IEEE Transactions on Knowledge and Data Engineering*, 2006, 18: 590–602.
- [Benito et al 2004] M. Benito, et al. Adjustment of systematic microarray data biases. *Bioinformatics*, 2004, 20: 105-114.
- [Brown et al 2000] M.P.S. Brown, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *PNAS*, 2000, 97(1): 262–267.
- [Subramanian et al 2005], A. Subramanian, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 2005, 102(43):15545-15550.
- [Nicolau et al 2007], M. Nicolau, et al. Disease-Specific Genomic Analysis: Identifying the Signature of Pathologic Biology. *Bioinformatics*, 2007, 23(8): 957-965.
- [Aslanyan and Sahakyan 2009] L. Aslanyan, H. Sahakyan. Chain split and computation in practical rule mining, *Information Science and Computing*, International book series no. 8., Classification, forecasting, data mining, 2009:132-135.

---

## Authors' Information

---

**Arsen Arakelyan, Anna Boyajian** – "Laboratory of Information Biology" Project of the Institute of Molecular Biology and Institute for Informatics and Automation Problems NAS RA, Yerevan, Armenia,  
e-mail: [arakelyan@sci.am](mailto:arakelyan@sci.am)

**Hasmik Sahakyan, Levon Aslanyan** – Institute for Informatics and Automation Problems, NAS RA, P. Sevak St. 1, Yerevan 14, Armenia, e-mail: [hasmik@ipia.sci.am](mailto:hasmik@ipia.sci.am), [lasl@sci.am](mailto:lasl@sci.am)

**Krassimira Ivanova, Ilija Mitov** – Institute of Mathematics and Informatics – Bulgarian Academy of Sciences, Sofia, Bulgaria, e-mail: [kivanova@math.bas.bg](mailto:kivanova@math.bas.bg), [mitov@foibg.com](mailto:mitov@foibg.com)

## ON THE COMPLEXITY OF SEARCH FOR CONJUNCTIVE RULES IN RECOGNITION PROBLEMS

Elena Djukova, Vladimir Nefedov

**Abstract:** *New results are obtained in logical data analysis and in the development of recognition procedures based on constructing conjunctive rules. Asymptotically optimal methods of constructing normal forms for characteristic functions of classes are developed. The technique of effective search is improved for the conjunctive rules in the case of incomplete data. Theoretical results are confirmed by experimental estimations.*

**Keywords:** *logical recognition procedures, enumeration problem complexity, dualization problem, algorithm with a polynomial delay, asymptotically optimal algorithm, normal form of a logical function, maximal conjunction, transformation of normal forms.*

**ACM Classification Keywords:** *F.2.1 Theory of Computation – Analysis of Algorithms and Problem Complexity – Numerical Algorithms and Problems – Computations on matrices, G.2.1. Mathematics of Computing – Discrete Mathematics – Combinatorics – Counting problems.*

---

### Introduction

Logical analysis of Integer data in recognition is based on constructing normal forms of double-valued logical functions which are characteristic functions of classes [Zhuravlev, 1978], [Djukova, 1977], [Djukova, 1982], [Djukova, 1987], [Djukova, 1989], [Djukova, Zhuravlev, 1997], [Djukova, Zhuravlev, 2000], [Djukova, 2004]. The characteristic function of a class is completely defined by the training set and the concerned recognition algorithm model. If the descriptions of objects are complete (for each precedent all feature values are defined) then the problem appears concerned with the transformation of the perfect conjunctive normal form (CNF) of the function  $F_K$  into the disjunctive normal form (DNF), i.e. the problem of multiplication of logical terms with exactly  $n$  variables in each term, where  $n$  is the number of features. In the case of incomplete data some variables may be absent in some terms and thus the initial CNF would not be perfect. Constructing the required DNF brings us to constructing the set of the conjunctive rules of the class  $K$ .

The search for conjunctive rules is a difficult computational problem. Constructing the reduced DNF of the function  $F_K$  is of a particular complexity (its length as a rule grows exponentially with the growth of the initial DNF length). This problem is one of difficult generation problems. The effectiveness of algorithms for enumeration problems is usually estimated by the complexity of a step (the time delay of a step). As the major enumeration problem the dualization is regarded. The dualization is the problem of constructing the reduced DNF of a monotone Boolean function  $f(x_1, \dots, x_n)$  define by the CNF of the following type

$$D_1 \& \dots \& D_u. \quad (1)$$

The search for effective algorithm for this problem is carried since the middle of the last century, starting from the classical works by Yablonskiy [Chegis, Yablonskii, 1958]. In [Gurvich, Khachiyan, 1999] an “incremental quasi-polynomial algorithm” is suggested. The number of elementary operations performed by that algorithm on each step is bounded by a quasi-polynomial over  $u$ ,  $n$  and the number of the maximal conjunctions found on the previous steps. The existence of an algorithm working with a (quasi)-polynomial delay depending only on  $u$  and  $n$  is not ascertained by the time.

In [Djukova, 1982], [Djukova, 1987] an approach is proposed, which allows to solve the considered problem of the transformation of the normal forms of  $f$  efficiently not always but almost always (for almost all CNFs of type (1) as  $n \rightarrow \infty$ ). In this case an algorithm is allowed to perform “unnecessary” (empty) steps, but the number of such

steps should have “almost always” lower growth degree than the number of all algorithm steps does. For this purpose the conception of an asymptotically optimal algorithm was introduced. This conception has been defined more exactly in a series of followed works [Djukova, 2003], [Djukova, 2004], [Djukova, 2007]. Let us adduce this definition.

Let  $B(f)$  be a set of all maximal conjunctions of the function  $f$ , let  $Q(f)$  be a finite sequence of elementary conjunctions which contains  $B(f)$ . It is supposed that some conjunctions may be found in  $Q(f)$  more than once.

We will say that an algorithm  $A$  constructs  $Q(f)$  with a *polynomial delay* if exactly one conjunction from  $Q(f)$  is built on each step with no more than  $d(u, n)$  elementary operations performed,  $d(u, n)$  is bounded above by a polynomial over  $u$  and  $n$ . The elementary operation is an examination of one symbol of a variable in the CNF.

Let an algorithm  $A$  construct the sequence  $Q(f)$  with a polynomial delay checking each built conjunction from  $Q(f)$  to be in  $B(f)$  (such check can be performed at a polynomial time according to statements 2 and 3 formulated below). The number of steps of the algorithm  $A$  equal to the length of the sequence  $Q(f)$  is denoted by  $N_A(f)$ .

The algorithm  $A$  is called *asymptotically optimal*, if for almost all CNFs of type (1) as  $n \rightarrow \infty$  it is true that

$$N_A(f) \approx |B(f)|$$

(the number of steps of the algorithm  $A$  is asymptotically equal to the power of  $B(f)$ ).

Later the concept of an asymptotically optimal algorithm was spread for the case of a double-valued logical function  $F$  defined on  $k$ -ary  $n$ -dimensional arrays and determined by a CNF of  $u$  elementary disjunctions [Djukova, 1987], [Djukova, 1989], [Djukova, Zhuravlev, 1997] [Djukova, Zhuravlev, 2000], [Djukova, Inyakin, 2008].

By the moment the series of asymptotically optimal algorithms has been constructed for the condition  $\log_d u \leq (1 - \varepsilon) \log_d n$ ,  $\varepsilon > 0$ , where  $d = 2$  for the case of a monotone Boolean function,  $d = k$  for the case when  $F$  is determined by a perfect CNF and  $d = (k + 1) / k$  for the case when  $F$  is determined by a CNF which is not perfect. In these algorithms there are unnecessary steps of two types. There are unnecessary steps of two types in these algorithms. Each step of the first type yields a conjunction which is not maximal whereas each step of the second type yields a maximal conjunction constructed before. Among the constructed algorithms there are both the algorithms that make only one type unnecessary steps and the algorithms containing the unnecessary steps of both types.

The experimental comparison of the constructed algorithm has been carried out. The best result has been shown by the algorithm OPT and by its modification, the algorithm OPT+, assigned for constructing the maximal conjunctions of a monotone Boolean function and for constructing the maximal conjunctions of a double-valued logical function defined by the perfect CNF, respectively.

In this paper an algorithm OPT++ is built which is a modification of the algorithm OPT for the case when a double-valued logical function  $F$  is defined by an arbitrary CNF. The algorithm OPT++ is asymptotically optimal as well, for the case  $\log u \leq (1 - \varepsilon) \log n$ ,  $\varepsilon > 0$ ,  $d = (k + 1) / k$  (this fact follows from works [Djukova, Zhuravlev, 1997], [Djukova, Zhuravlev, 2000]). In the following section it is shown that for the case  $n \leq u$  the logarithm of the number of the algorithm steps to the base  $d$  is equal to the logarithm of the number of all maximal conjunctions of the function  $F$  “almost always” as  $n \rightarrow \infty$ . The justification of the algorithm effectiveness is based upon the analysis of metric (quantitative) characteristics of the set of all maximal

conjunction and of the so-called irreducible conjunctions of the function  $F$ . Показано, что эти асимптотики совпадают. Asymptotic estimates of typical values of the logarithm to the base  $d$  of the number of the maximal conjunctions of  $F$  and the logarithm to the base  $d$  of the number of the irreducible conjunctions of  $F$  are obtained. These asymptotic estimates are shown to be equal. When obtaining these estimates we used the technique from the papers [Dem'yanov, Djukova, 2007], [Djukova, 2007]. The effectiveness of the algorithm OPT++ is approved experimentally, particularly the experimental estimates of the number of "unnecessary" steps are obtained.

---

### Major results

---

Let  $E_k^n$  be the set arrays  $(\sigma_1, \dots, \sigma_n)$ , where  $\sigma_i \in \{0, 1, \dots, k-1\}$ ,  $i \in \{1, \dots, n\}$ , and let a function  $F(x_1, \dots, x_n)$  be determined on  $E_k^n$  and possesses values 1 and 0 on subsets  $N_F$  and  $N_{-F}$ , respectively.

Let

$$x^\sigma = \begin{cases} 1, & \text{if } x = \sigma, \\ \text{otherwise,} \end{cases}$$

$x, \sigma \in \{0, 1, \dots, k-1\}$ . The notions of an elementary conjunction and elementary disjunction are defined routinely. An elementary conjunction (EC) over variables  $x_1, \dots, x_n$  is the formula of the type  $x_{j_1}^{\sigma_1} \cdot \dots \cdot x_{j_r}^{\sigma_r}$ , where  $x_{j_i} \in \{x_1, \dots, x_n\}$ , when  $i = 1, 2, \dots, r$ , and  $x_{j_q} \neq x_{j_t}$  when  $t, q \in \{1, 2, \dots, r\}$ ,  $t \neq q$ . The value of the EC is equal to 1 if and only if all its factors are equal to 1. The number  $r$  is called the rank of the elementary conjunction.

An elementary disjunction (ED) over variables  $x_1, \dots, x_n$  is the formula of the type  $x_{j_1}^{\sigma_1} \vee \dots \vee x_{j_r}^{\sigma_r}$ , where  $x_{j_i} \in \{x_1, \dots, x_n\}$ , when  $i = 1, 2, \dots, r$ , and  $x_{j_q} \neq x_{j_t}$ , when  $t, q \in \{1, 2, \dots, r\}$ ,  $t \neq q$ .

Let  $B$  be an elementary conjunction over variables  $x_1, \dots, x_n$  and let  $M(B, R)$  be the number of the disjunctions not containing variables from  $B$  in the CNF  $R$ . Let  $N_B$  be in interval of verity of the EC  $B$ . The EC  $B$  is called *admissible* for  $F$  if  $N_B \cap N_{-F} = \emptyset$ ; i.e., when  $M(B, R) = 0$ . The EC  $B$  is called *irreducible* for  $F$  if there does not exist an admissible EC  $B'$  such that  $N_{B'} \supset N_B$  and  $M(B', R) = M(B, R)$ . The EC  $B$  is called *maximal* for  $F$  if it is both admissible and irreducible. The EC  $B$  is called *irredundant* for  $F$  if it is irreducible for  $F$  and there does not exist an irreducible for  $F$  conjunction  $B'$  such that  $N_{B'} \supset N_B$ .

**Statement 1.** If an EC  $B$  is maximal for  $F$  then it is irredundant for  $F$ .

The converse statement is false (an irredundant conjunction may not be admissible).

Let the function  $F$  be determined by a CNF  $R$  of the type  $D_1 \& \dots \& D_u$ , where  $D_i$ ,  $i = 1, 2, \dots, u$ , - ED over variables  $x_1, \dots, x_n$ . Denote by  $D_C(F)$  the reduced DNF of the function  $F$ , i.e. the DNF consisting of all maximal conjunctions of the function  $F$ .

Let us consider the problem of the transformation of  $R$  into  $D_C(F)$ . The required statements 2 and 3 are presented below.

**Statement 2** [Djukova, Nefedov, 2009]. An EC  $B$  is admissible for  $F$  if and only if each disjunction  $D_i$ ,  $i = \{1, 2, \dots, u\}$ , contains at least one factor of  $B$ .



**Statement 3** [Djukova, Nefedov, 2009]. An EC  $B = x_{j_1}^{\sigma_1} \cdot \dots \cdot x_{j_r}^{\sigma_r}$  is irreducible for  $F$  if and only if there exist  $r$  disjunctions in the CNF  $K$ , each disjunction containing exactly one factor of  $B$  and, if  $r > 1$ ,  $p, q \in \{j_1, \dots, j_r\}$ ,  $p \neq q$ , the disjunctions  $D_p$  and  $D_q$  contain different factors of  $B$ .

Statements 2 and 3 are respectively the criteria of admissibility and irreducibility of the EC of a function determined by a CNF.

Let us describe the algorithm OPT++ constructing DNF  $D_C(F)$ . The algorithm is based on constructing an irredundant for  $F$  conjunction  $B$  and checking the condition  $N_B \cap N_{-F} = \emptyset$  on each step. If this condition is true then according to statement 1 the constructed on this step conjunction is maximal. Else the performed step is “empty” (unnecessary).

It is convenient to present the work of the algorithm OPT++ as the process of constructing a decision tree  $\Delta_F$ . Each inner vertex of this tree corresponds to the pair  $(B, R')$ , where  $B$  is the irreducible for  $F$  conjunction and  $R'$  is the CNF derived from  $R$  by deleting some disjunctions and variables. A dangling vertex corresponds to an irredundant conjunction. The set of the dangling vertices corresponds to a subset of all irredundant conjunctions. On each step the algorithm constructs one dangling vertex of the tree  $\Delta_F$  at the time  $O(kunq(u + q))$ , where  $q = \min(u, kn)$ . The important characteristic of the algorithm is that different vertices of the decision tree  $\Delta_F$  correspond to different irreducible conjunctions. Thus the algorithm does not perform “repeated” steps.

Denote by  $S_r(F)$  the set of all irreducible conjunctions of the rank  $r$  of the function  $F$ ; denote by  $B_r(F)$  the set of all maximal conjunctions of the rank  $r$  of the function  $F$ ;

$$S(F) = \bigcup_{r=1}^n S_r(F), \quad B(F) = \bigcup_{r=1}^n B_r(F).$$

Let  $n \leq u \leq k^{n^\beta}$ ,  $\beta < \frac{1}{2}$ ,  $r_1 = \lceil \log_d u - \log_d \ln \log_d u - 1 \rceil$ ,  $d = (k + 1)/k$ . For the considered case the asymptotic form of the logarithm of the typical number of the conjunctions in  $S(F)$  with ranks no lower than  $r_1$  is obtained (when  $n \rightarrow \infty$ ). It is shown that this asymptotic form is congruent with the asymptotic form of the logarithm of the typical number of the conjunctions in  $B(F)$  and is also congruent with the asymptotic form of the typical number of the conjunctions in  $B(F)$  with ranks no lower than  $r_1$ . The estimate of the typical rank value of a conjunction in  $B(F)$  is obtained. The mention estimates are stated in Theorems 1-3 formulated below.

**Theorem 1.** If  $n \leq u \leq k^{n^\beta}$ ,  $\beta < 1/2$ , then as  $n \rightarrow \infty$  for almost all CNFs of type (1) the logarithm to the base  $d$  of the number of all conjunctions in  $S(F)$  with ranks no lower than  $r_1$  is asymptotically equal to the logarithm to the base  $d$  of the number of all conjunctions in  $B(F)$  with ranks no lower than  $r_1$  and is asymptotically equal to  $\log_d C_n^{r_1} + r_1$ .

**Theorem 2.** If  $n \leq u \leq k^{n^\beta}$ ,  $\beta < 1/2$ , then as  $n \rightarrow \infty$  for almost all CNFs of type (1) the logarithm to the base  $d$  of the number of all conjunctions in  $B(F)$  with ranks no lower than  $r_1$  is asymptotically equal to the logarithm to the base  $d$  of the number of all conjunctions in  $B(F)$  and is asymptotically equal to  $\log_d C_n^{r_1} + r_1$ .

**Theorem 3.** If  $n \leq u \leq k^{n^\beta}$ ,  $\beta < 1/2$ , then as  $n \rightarrow \infty$  for almost all functions determined by a CNF of type (1) the ranks of almost all conjunctions from  $B(F)$  are in  $[r_1, \log_d un]$ .

The proofs of Theorems 1-3 are based on Lemmas 1-6 stated below. In the proofs of Lemmas 1-6 the results from [Djukova, Peskov, 2002] and [Djukova, Inyakin, 2003] are used.

Let  $N_{un}^k = \{F\}$  be the space of simple events in which each event  $F$  happens with a probability of  $1/|N_{un}^k|$ . The mathematical expectation of a random variable  $X(F)$  defined on  $N_{un}^k$  is denoted by  $\mathbf{M}X(F)$ .

The following result is easy to prove

**Lemma 1.** Let  $X(F) \geq 0$ ,  $\theta > 0$ , and  $\nu_\theta(n)$  be the fraction of functions  $F$  from  $N_{un}^k$  for which  $X(F) \geq \theta \mathbf{M}X(F)$ . Then  $\nu_\theta(n) \leq 1/\theta$ .

In what follows, the notation  $a_n \approx b_n$  as  $n \rightarrow \infty$  and  $a_n \leq_n b_n$  as  $n \rightarrow \infty$  means  $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = 1$  as  $n \rightarrow \infty$  and  $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} \leq 1$  as  $n \rightarrow \infty$ , respectively.

Let  $a_r = C_n^r C_u^r r! (k-1)^{r^2-r} k^{r-r^2}$ . Since  $a_{r+1} = o(a_r)$  for  $n \leq u$  and  $r \geq r_1$ , we have the following result.

**Lemma 2.** For  $n \leq u$

$$\sum_{r \geq r_1} C_n^r C_u^r r! d^{r-r^2} \approx C_n^{r_1} C_u^{r_1} r_1! d^{r_1-r_1^2}, \quad n \rightarrow \infty.$$

**Lemma 3.** For  $u \leq k^{n^\beta}$  and  $\beta < 1/2$

$$\log_d (C_u^{r_1} r_1! d^{-r_1^2}) = \bar{o}(\log_d (C_n^{r_1} d^{r_1})), \quad n \rightarrow \infty.$$

**Proof.** The lemma can be proved by direct verification. Indeed, the obvious inequality  $C_n^r \geq \left(\frac{n-r}{r}\right)^r$  implies that

$$\log_d C_n^{r_1} \geq_n (1-\beta)r_1 \log_d n, \quad n \rightarrow \infty. \quad (2)$$

On the other hand,

$$b = C_u^{r_1} r_1! d^{-r_1^2} \leq (k^2 \ln \log_d u)^{r_1}.$$

Consequently, we have

$$\log_d b \leq_n r_1 \log_d \ln n, \quad n \rightarrow \infty. \quad (3)$$

The assertion of the lemma follows from (2) and (3).

On  $N_{un}^k = \{F\}$  consider the random variables  $\eta_r(F) = |B_r(F)|$  and  $\xi_r(F) = |S_r(F)|$ . It is easy to calculate

$$\mathbf{M}\xi_r(F) \leq C_n^r C_u^r r! k^{r^2-r} (k+1)^{r-r^2}$$

(see [Djukova, Peskov, 2002]).

**Lemma 4.** For  $n \leq u \leq k^{n^\beta}$  and  $\beta < 1/2$ , for almost all  $F$  from  $N_{un}^k$

$$\log_d \sum_{r \geq r_1} \eta_r(F) \leq \log_d \sum_{r \geq r_1} \xi_r(F) \leq_n \log_d C_n^{r_1} + r_1, \quad n \rightarrow \infty.$$

**Proof.** Lemmas 2 and 3 straightforwardly imply that

$$\log_d \sum_{r \geq r_1} \mathbf{M} \eta_r(F) \leq \log_d \sum_{r \geq r_1} \mathbf{M} \xi_r(F) \approx \log_d C_n^{r_1} + r_1, \quad n \rightarrow \infty.$$

Applying Lemma 1 with  $\theta = \log_d \log_d n$ , we derive the assertion of Lemma 4.

**Lemma 5.** For  $u \leq k^{n^\beta}$  and  $\beta < 1/2$ , for almost all  $F$  from  $N_{un}^k$

$$\log_d \sum_{r \geq r_1} \xi_r(F) \geq \log_d \sum_{r \geq r_1} \eta_r(F) \geq_n \log_d C_n^{r_1} + r_1, \quad n \rightarrow \infty.$$

**Proof.** It was shown in [Djukova, Inyakin, 2003] that if  $F$  is determined by a perfect CNF and the condition of the lemma is fulfilled then

$$\eta_{r_1}(F) \approx C_n^{r_1} k^{r_1} (1 - k^{-r_1})^u, \quad n \rightarrow \infty.$$

Using the same technique we derive a similar estimate for the case when  $F$  is determined by an arbitrary CNF. This estimate is

$$\eta_{r_1}(F) \approx C_n^{r_1} k^{r_1} (1 - d^{-r_1})^u, \quad n \rightarrow \infty.$$

Estimating  $(1 - d^{-r_1})^u$ , we have

$$(1 - d^{-r_1})^u \approx e^{\frac{-u}{d^{r_1}}} \geq (\log_d u)^{-2d}.$$

Combining this with  $\log_d \log_d u = o(\log_d C_n^{r_1})$ , we derive the assertion of the lemma.

Theorem 1 follows from Lemmas 4 and 5.

**Lemma 6.** For  $n \leq u \leq k^{n^\beta}$  and  $\beta < 1/2$ , for almost all  $F$  from  $N_{un}^k$

$$\sum_{r < r_1} \eta_r(F) = o\left(\sum_{r \geq r_1} \eta_r(F)\right), \quad n \rightarrow \infty.$$

**Proof.** It was shown in [Djukova, Inyakin, 2003] that if  $F$  is determined by a perfect CNF and the condition of the lemma is fulfilled then

$$\sum_{r < r_1} \eta_r(F) = o(C_n^{r_1} k^{r_1}), \quad n \rightarrow \infty.$$

Using the same technique we derive a similar estimate for the case when  $F$  is determined by an arbitrary CNF.

This estimate for  $u \leq k^{n^\beta}$  and  $\beta < 1/2$  is

$$\sum_{r < r_1} \eta_r(F) = o(C_n^{r_1} d^{r_1}), \quad n \rightarrow \infty.$$

On the other hand, according to Theorem 1, we have

$$\sum_{r \geq r_1} \eta_r(F) = (C_n^{r_1} d^{r_1})^{1+\delta(n)}, \quad \text{where } \delta(n) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Consequently,

$$\sum_{r < r_1} \eta_r(F) = o\left(\sum_{r \geq r_1} \eta_r(F)\right), \quad n \rightarrow \infty,$$

which proves the lemma.

Theorem 2 follows from Theorem 1 and Lemma 6. Theorem 3 follows from Lemma 6.

**Remark 1.** If a CNF has two identical conjunctions, then the set  $B(F)$  obviously does not change after removing one of them. Consequently, when the number of maximal conjunctions of  $F$  is calculated, it is reasonable to consider the case when the CNF does not have identical conjunctions. For  $u \leq k^{n^\beta}$  and  $\beta < 1/2$ , this property is possessed by almost all CNFs of type (1).

**Remark 2.** The algorithm OPT++ is tested on random data. Experimental estimates of the number of maximal conjunctions and of the number of the “unnecessary” steps of the algorithm are obtained. The experimental results show that OPT++ works faster and performs considerably less “unnecessary” steps than other asymptotically optimal algorithms developed earlier. The part of “unnecessary” steps of the algorithm OPT++ increases as the part of missed feature values in the learning data increases; and it decreases as the fraction  $u/n$  increases.

---

## Conclusion

New results are obtained in logical data analysis and in constructing recognition procedures based on the transformation of the normal forms of the characteristic functions of classes. An approach is developed, which allows to solve the problem of the transformation of the normal forms almost always with as asymptotical accuracy both for complete data and incomplete data. An algorithm OPT++ is developed for solving the problem of constructing the maximal conjunctions of a double-valued function  $F$  defined on  $k$ -ary  $n$ -dimensional arrays and determined by a CNF of  $u$  elementary disjunctions. The algorithm is justified both theoretically and experimentally. The theoretical justification is based on obtaining new asymptotic estimates of the typical values of the number of the maximal conjunctions and irreducible conjunctions of the function  $F$ . It is shown that OPT++ on experimental characteristics surpasses other similar algorithms.

---

## Bibliography

- [Chegis, Yablonskii, 1958] I. A. Chegis and S. V. Yablonskii, “Logical Methods for Control for Electric Circuits,” Tr. Mat. Inst. im. V. A. Steklova, Akad. Nauk SSSR, 1958, Vol. 51, pp. 270-360 [in Russian].
- [Dem'yanov, Djukova, 2007] E. A. Dem'yanov and E. V. Djukova, “On the Construction of Irredundant Coverings of an Integer Matrix,” Zh. Vychisl. Mat. Mat. Fiz. 47 (3), 539-547 (2007) [Comput. Math. Math. Phys. 47 (3), 518-526 (2007)].
- [Djukova, 1977] E. V. Djukova, “About Asymptotically Optimal Algorithm for Building Irreducible Tests,” Dokl. Akad. Nauk SSSR 233 (4), 527-530 (1977) [in Russian].
- [Djukova, 1982] E. V. Djukova, “Asymptotically Optimal Test Algorithms in Problems of Recognition,” in Problems of Cybernetics, (Nauka, Moscow, 1982), Issue 39, pp. 165-199 [in Russian].
- [Djukova, 1987] E. V. Djukova, “The Complexity of Realization of Some Recognition Procedures,” Zh. Vychisl. Mat. Mat. Fiz. 27 (1), 114-127 (1987) [in Russian].
- [Djukova, 1989] E. V. Djukova, “Recognition Algorithms of the Kora Type: Complexity of Implementation and Metric Properties,” in Recognition, Classification, Forecast: Mathematical Methods and Their Applications (Nauka, Moscow, 1989), Vol. 2, pp. 99-125 [in Russian].
- [Djukova, 2003] E.V. Djukova. Discrete (Logical) Recognition Procedures: Principles of Construction, Complexity of Realization and Basic Models // Pattern Recognition and Image Analysis. 2003. Vol. 13. No. 3. P. 417-425.
- [Djukova, 2004] E. V. Djukova, “On the Implementation Complexity of Discrete (Logical) Recognition Procedures,” Zh. Vychisl. Mat. Mat. Fiz. 44 (3), 562-572 (2004) [Comput. Math. Math. Phys. 44 (3), 532-541 (2004)].
- [Djukova, 2005] E. V. Djukova, “On the Number of Irreducible Coverings of an Integer Matrix,” Zh. Vychisl. Mat. Mat. Fiz. 45 (5), 935-940 (2005) [Comput. Math. Math. Phys. 45 (5), 903-908 (2005)].

- [Djukova, 2007] E. V. Djukova, "On Constructing the Irredundant Coverings of a Boolean Matrix," *Doklady Akademii Nauk.* 412 (1), 15-17 (2007) [in Russian].
- [Djukova, Inyakin, 2003] E. V. Djukova and A. S. Inyakin, "Classification Procedures Based on the Construction of Class Coverings," *Zh. Vychisl. Mat. Mat. Fiz.* 43 (12), 1884-1895 (2003) [*Comput. Math. Math. Phys.* 43 (12), 1812-1822 (2003)].
- [Djukova, Inyakin, 2008] E. V. Djukova and A. S. Inyakin, "On Asymptotically Optimal Constructing the Irredundant Coverings of an Integer Matrix," in *Matematicheskie Voprosy Kibernetiki (Fizmatlit, Moscow, 2008)*, Issue 17, pp. 135-146 [in Russian].
- [Djukova, Nefedov, 2009] E.V. Djukova, V.Y. Nefedov. The Complexity of Transformation of Normal Forms for Characteristic Functions of Classes // *Pattern Recognition and Image Analysis.* 2009. Vol. 19. No. 3. P. 435-440.
- [Djukova, Peskov, 2002] E. V. Djukova and N. V. Peskov, "Search for Informative Fragments in Descriptions of Objects in Discrete Recognition Procedures," *Zh. Vychisl. Mat. Mat. Fiz.* 42, 741-753 (2002) [*Comput. Math. Math. Phys.* 42, 711-723 (2002)].
- [Djukova, Zhuravlev, 1997] E.V. Djukova, Yu.I. Zhuravlev. Discrete Methods of Information Analysis in Recognition and Algorithm Synthesis // *Pattern Recognition and Image Analysis.* 1997. Vol.7. No.2. P. 192-207.
- [Djukova, Zhuravlev, 2000] E. V. Djukova and Yu. I. Zhuravlev, "Discrete Analysis of Feature Descriptions in Recognition Problems of High Dimensionality," *Zh. Vychisl. Mat. Mat. Fiz.* 40 (8), 1264-1278 (2000) [*Comput. Math. Math. Phys.* 40 (8), 1214-1227 (2000)].
- [Gurvich, Khachiyan, 1999] Gurvich V., Khachiyan L. On generating the irredundant conjunctive and disjunctive normal forms of monotone Boolean functions // *Discrete Appl. Math.* 1999. V. 96-97. № 1-3. P. 363-373.
- [Jonson, 1988] Jonson D.S., Yannakakis M., Papadimitriou C.H. On General All Maximal Independent Sets // *Information processing Letters.* 1988. V. 27. P. 119-123.
- [Yablonskii, 1986] S. V. Yablonskii, *An Introduction to Discrete Mathematics* (Nauka, Moscow, 1986) [in Russian].
- [Zhuravlev, 1978] Zhuravlev Yu.I., On an Algebraic Approach to Recognition or Classification Problems. *Problemy Kibernetiki*, Moscow: Nauka, 1978, no. 33, pp. 5-68 [in Russian].

---

### Authors' Information

---



**Elena Djukova** – Doktor Nauk, Major scientist, Dorodnicyn Computing Centre of RAS, Vavilova st. 40, 119333 Moscow, Russia; e-mail: edjukova@mail.ru

*Major Fields of Scientific Research: Pattern recognition, Discrete mathematics, Logical recognition procedures, Computational complexity of discrete problems*



**Vladimir Nefedov** – post-graduate student, Dorodnicyn Computing Centre of RAS, Vavilova st. 40, 119333 Moscow, Russia; e-mail: nefedov85@mail.ru

*Major Fields of Scientific Research: Discrete mathematics, Mathematical logic, Pattern Recognition*

---

## OPTIMAL FORECASTING BASED ON CONVEXCORRECTING PROCEDURES

**Oleg Senko, Alexander Dokukin**

**Abstract:** Properties of convex correcting procedures (CCP) on ensembles of predictors are studied. CCP calculates integral solution as convex linear combination of predictors' prognoses. Structure of forecasting squared error and generalized error are analyzed. At that generalized error is defined as mean of squared error at Cartesian product of forecasted objects space and space of training sets. It is shown that forecasting squared error, bias and variance component of generalized error have similar structure. Search of optimal CCP coefficients is reduced to quadratic programming task which is solved in terms of ensemble superfluity. Ensemble is considered superfluous if some members can be removed without loss of forecasting ability. Necessary and sufficient conditions of superfluity absence are proven. A regression method based on the described principles has been developed. Its concepts as well as testing results are shown revealing CCP's significant superiority over stepwise regression.

**Keywords:** forecasting, bias-variance decomposition, convex combinations, variables selection

**ACM Classification Keywords:** G.3 Probability and Statistics - Correlation and regression analysis, Statistical computing

---

### Introduction

Goal of this work is study of correcting procedures for sets of forecasting algorithms calculating integral solution as convex linear combination of prognoses, given by each algorithm from the set. Let's suppose that we have set of  $L$  algorithms, forecasting some variable  $Y$  by set of explanatory variables  $X_1, \dots, X_n$  at objects that are elements of probability space  $\Omega$ . Prognosis that is calculated by  $i$ -th algorithm for some  $\omega$  will be further denoted as  $z_i(\omega)$ . Let  $\mathbf{c} = (c_1, \dots, c_L)$  is vector of real nonnegative coefficients satisfying  $\sum_{i=1}^L c_i = 1$ .

Convex correcting procedures (CCP) are discussed in the paper that calculate collective solution  $Z(\omega, \mathbf{c})$  as

$$Z_{ccp}(\omega, \mathbf{c}) = \sum_{i=1}^L c_i z_i(\omega).$$

Using of average by set of prognoses is special case of CCP.

Convex correcting procedures are rather often used in theory of pattern recognition or forecasting by empirical data. Neural networks ensembles, methods of weighed combining, boosting and bagging methods, pattern recognition methods based on voting by systems of regularities [1, 2] are well known examples of optimal convex solutions. Last years some new techniques were suggested that are based on searching balance between accuracy of data approximation and diversity of ensembles [5]. Our approach is based on analysis of generalized error structure.

---

### Forecasting error for CCP

We begin with discussing of mean squared error of CCP forecasting. Mean squared error of  $Y$  forecasted by some predictor  $Z$  will be denoted as  $\delta(Z)$ . So,  $\delta(Z) = E_{\Omega}(Y - Z)^2$ . It is easily to show that

$$\begin{aligned}
\sum c_i [Y(\omega) - z_i(\omega)]^2 &= \sum_{i=1}^L c_i [Y(\omega) - Z_{ccp}(\omega, \mathbf{c}) + Z_{ccp}(\omega, \mathbf{c}) - z_i(\omega)]^2 = \\
&= \sum_{i=1}^L c_i [Y(\omega) - Z_{ccp}(\omega, \mathbf{c})]^2 + \sum_{i=1}^L c_i [Z_{ccp}(\omega, \mathbf{c}) - z_i(\omega)]^2 = \\
&= [Y(\omega) - Z(\omega, \mathbf{c})]^2 + \sum_{i=1}^L c_i [Z_{ccp}(\omega, \mathbf{c}) - z_i(\omega)]^2
\end{aligned}$$

So CCP error in case of forecasting of  $Y$  for object  $\omega$  that is equal  $[Y(\omega) - Z_{ccp}(\omega, \mathbf{c})]^2$  can be presented as difference

$$\sum_{i=1}^L c_i [Y(\omega) - z_i(\omega)]^2 - \sum_{i=1}^L c_i [Z_{ccp}(\omega, \mathbf{c}) - z_i]^2 \quad (1)$$

Decomposition (1) was received in [4].

Task of optimal CCP search may be discussed as task of minimization of mathematical mean of error  $[Y(\omega) - Z(\omega, \mathbf{c})]^2$  in space of forecasted objects. It is evident that

$$\begin{aligned}
\delta(Z_{ccp}) &= E_{\Omega} \left\{ \sum_{i=1}^L c_i [Y(\omega) - z_i(\omega)]^2 - \sum_{i=1}^L c_i [Z_{ccp}(\omega, \mathbf{c}) - z_i]^2 \right\} = \sum_{i=1}^L c_i \delta(z_i) - \\
&\quad - \sum_{i=1}^L c_i [Z_{ccp}(\omega, \mathbf{c}) - z_i]^2 \}.
\end{aligned}$$

It follows from non-negativeness of variation component  $E_{\Omega} \left\{ \sum_{i=1}^L c_i [Z_{ccp}(\omega, \mathbf{c}) - z_i(\omega)]^2 \right\}$  that error  $\delta(Z_{ccp})$  never exceed weighed with  $c_i > 0$  mean of individual prognostic algorithms errors. Mathematical mean  $E_{\Omega} [z_i - z_j]^2$ , characterizing discrepancy of  $i'$ -th and  $j$ -th forecasting algorithms will be denoted as  $E_{\Omega} [z_i - z_j]^2 = \rho_{ij}^e$ . Let note that

$$\begin{aligned}
-\sum_{i=1}^L c_i [Z_{ccp}(\omega, \mathbf{c}) - z_i]^2 &= E_{\Omega} [Z(\omega, \mathbf{c})]^2 - \sum_{i=1}^L c_i E_{\Omega} [z_i^2(\omega)] = \\
&= \sum_{i'=1}^L \sum_{i''=1}^L c_i c_{i''} E_{\Omega} [z_{i'}(\omega) z_{i''}(\omega)] - \sum_{i=1}^L c_i E_{\Omega} [z_i^2(\omega)].
\end{aligned}$$

Taking into account that  $z_{i'} z_{i''} = \frac{1}{2} \{ -(z_{i'} - z_{i''})^2 + (z_{i'})^2 + (z_{i''})^2 \}$ , we receive that

$$\begin{aligned}
\sum_{i=1}^L \sum_{i=1}^L c_i c_{i''} E_{\Omega} [z_{i'}(\omega) z_{i''}(\omega)] - \sum_{i=1}^L c_i E_{\Omega} [z_i^2(\omega)] &= -\frac{1}{2} \sum_{i=1}^L \sum_{i=1}^L c_i c_{i''} E_{\Omega} [z_{i'}(\omega) - z_{i''}(\omega)]^2 + \\
&+ \frac{1}{2} \left\{ \sum_{i=1}^L c_i E_{\Omega} [z_i(\omega)]^2 \sum_{i=1}^L c_{i''} + \sum_{i=1}^L c_{i''} E_{\Omega} [z_{i''}(\omega)]^2 \sum_{i=1}^L c_i \right\} - \sum_{i=1}^L c_i E_{\Omega} [z_i^2(\omega)] =
\end{aligned}$$

$$\begin{aligned}
& + \sum_{i''=1}^L c_{i''} E_{\Omega} [z_{i''}(\omega)]^2 \sum_{i'=1}^L c_{i'} \} - \sum_{i=1}^L c_i E_{\Omega} [z_i^2(\omega)] = - \sum_{i'=1}^L \sum_{i''=1}^L c_{i'} c_{i''} E_{\Omega} [z_{i'} - z_{i''}]^2 = \\
& = - \sum_{i'=1}^L \sum_{i''=1}^L c_{i'} c_{i''} \rho_{i' i''}^e
\end{aligned} \tag{2}$$

So, error of CCP forecasting may be written as

$$\delta(Z_{ccp}) = \sum_{i=1}^L c_i \delta(z_i) - \frac{1}{2} \sum_{i'=1}^L \sum_{i''=1}^L c_{i'} c_{i''} \rho_{i' i''}^e \tag{3}$$

---

### Generalized forecasting error of CCP

---

**Generalized error.** Forecasting error  $\delta(Z)$  describes exactness of forecasting algorithm (predictor)  $Z$  that was previously trained by some fixed training set  $\omega_t$ . But training set often may change during process of training. In these cases predictor  $Z$  is function of  $\omega$  and  $\omega_t$ . Effectiveness of training procedure is better characterized with help of generalized error that is mathematical mean of error  $\delta(Z)$  by space of various training sets  $\Omega_t$ . The generalized error for predictor  $Z$  will be denoted as  $\Delta(Z)$ . The mean value  $Z$  at point of  $\mathbf{x} \in \mathbf{R}^n$  by space  $\Omega_t$  will be denoted as  $\hat{Z}(\mathbf{x})$ . The following decomposition is true for generalized error  $\Delta(Z)$ :

$$\Delta(Z) = \Delta_{noise} + \Delta_{bias}(Z) + \Delta_{var}(Z),$$

where  $\Delta_{noise} = E_{\Omega} \{Y - E[Y | \mathbf{x}(\omega)]\}^2$  is irreducible noise component that characterize only random process associated with each specific forecasting task and is not related to forecasting algorithm, component  $\Delta_{bias}(Z) = E_{\Omega} \{\hat{Z}[\mathbf{x}(\omega)] - E_{\Omega}[Y | \mathbf{x}(\omega)]\}^2$  describes deviation of  $\hat{Z}(\mathbf{x})$  from conditional means  $E_{\Omega}[Y | \mathbf{x}(\omega)]$ ,  $\Delta_{var} = E_{\Omega} E_{\Omega_t} \{\hat{Z}[\mathbf{x}(\omega)] - Z[\mathbf{x}(\omega, \omega_t)]\}^2$  describes variation of  $Z[\mathbf{x}(\omega, \omega_t)]$  at Cartesian product  $\Omega_t \times \Omega$ . At that bias component is related to inconsistency between type of used model and dependency that really exists in training data set, while variance component is related to inconsistency between complexity and dimension of model and training data set size. Variance component describes variation of forecasting function at relatively small and statistically admissible changes in training data. So, it may be also referred to as instability component. The bias component may be improved by using more complicate families of functions that are tried for data approximation. While high complexity of models often leads to increase of variance component. Such contradiction between two components is known as bias/variance dilemma [6].

**Bias component structure.** Let's consider structure of CCP generalized errors components. Calculation of  $\Delta_{bias}$  mostly repeats calculation of  $\delta(Z_{ccp})$  structure. It is evident that  $\hat{Z}_{ccp}[\mathbf{x}(\omega), \mathbf{c}] = \sum_{i=1}^L c_i \hat{z}_i[\mathbf{x}(\omega)]$  and it

is easy to show that  $\{E_{\Omega}(Y | \mathbf{x}) - \hat{Z}_{ccp}[\mathbf{x}(\omega), \mathbf{c}]\}^2$  may be written as difference

$$\sum_{i=1}^L c_i \{E_{\Omega}(Y | \mathbf{x}) - \hat{z}_i[\mathbf{x}(\omega)]\}^2 - \sum_{i=1}^L c_i \{\hat{Z}_{ccp}(\omega, \mathbf{c}) - \hat{z}_i[\mathbf{x}(\omega)]\}^2.$$



So,

$$\begin{aligned}\Delta_{bias}(Z_{ccp}) &= \sum_{i=1}^L c_i E_{\Omega} \{E_{\Omega}(Y | \mathbf{x}) - \widehat{z}_i[\mathbf{x}(\omega)]\}^2 - \sum_{i=1}^L c_i E_{\Omega} \{\widehat{Z}_{ccp}[\mathbf{x}(\omega), \mathbf{c}] - \widehat{z}_i[\mathbf{x}(\omega)]\}^2 = \\ &= \sum_{i=1}^L c_i \Delta_{bias}(z_i) - \sum_{i=1}^L c_i E_{\Omega} \{\widehat{Z}_{ccp}[\mathbf{x}(\omega), \mathbf{c}] - \widehat{z}_i[\mathbf{x}(\omega)]\}^2.\end{aligned}$$

It is easy to get similar calculations from (2), showing that

$$\sum_{i=1}^L c_i E_{\Omega} \{\widehat{Z}_{ccp}[\mathbf{x}(\omega), \mathbf{c}] - \widehat{z}_i[\mathbf{x}(\omega)]\}^2 = \sum_{i'=1}^L \sum_{i''=1}^L c_i c_{i''} \rho_{i'i''}^{bc},$$

where  $\rho_{i'i''}^{bc} = E_{\Omega} \{\widehat{z}_{i'}[\mathbf{x}(\omega)] - \widehat{z}_{i''}[\mathbf{x}(\omega)]\}^2$ . Thus,

$$\Delta_{bias}(Z_{ccp}) = \sum_{i=1}^L c_i \Delta_{bias}(z_i) - \frac{1}{2} \sum_{i'=1}^L \sum_{i''=1}^L c_i c_{i''} \rho_{i'i''}^b \quad (4)$$

Variance component structure.

$$\begin{aligned}\Delta_{var}(Z_{ccp}) &= E_{\Omega} E_{\Omega_t} \{\widehat{Z}_{ccp}[\mathbf{x}(\omega)] - Z_{ccp}[\mathbf{x}(\omega, \omega_t)]\}^2 = \\ &= \sum_{i'=1}^L \sum_{i''=1}^L c_i c_{i''} E_{\Omega} E_{\Omega_t} \{\widehat{z}_{i'}[\mathbf{x}(\omega)] - z_{i'}[\mathbf{x}(\omega, \omega_t)]\} \{\widehat{z}_{i''}[\mathbf{x}(\omega)] - z_{i''}[\mathbf{x}(\omega, \omega_t)]\}\end{aligned}$$

It is evident that

$$\begin{aligned}&\{\widehat{z}_{i'}[\mathbf{x}(\omega)] - z_{i'}[\mathbf{x}(\omega, \omega_t)]\} \{\widehat{z}_{i''}[\mathbf{x}(\omega)] - z_{i''}[\mathbf{x}(\omega, \omega_t)]\} = \\ &= -\frac{1}{2} \{\widehat{z}_{i'}[\mathbf{x}(\omega)] - z_{i'}[\mathbf{x}(\omega, \omega_t)] - \widehat{z}_{i''}[\mathbf{x}(\omega)] + z_{i''}[\mathbf{x}(\omega, \omega_t)]\}^2 + \frac{1}{2} \{\widehat{z}_{i'}[\mathbf{x}(\omega)] - z_{i'}[\mathbf{x}(\omega, \omega_t)]\}^2 + \\ &\quad \frac{1}{2} \{\widehat{z}_{i''}[\mathbf{x}(\omega)] - z_{i''}[\mathbf{x}(\omega, \omega_t)]\}^2.\end{aligned}$$

However,

$$E_{\Omega} E_{\Omega_t} \{\widehat{z}_{i'}[\mathbf{x}(\omega)] - z_{i'}[\mathbf{x}(\omega, \omega_t)]\}^2 = \Delta_{var}(z_{i'}),$$

$$E_{\Omega} E_{\Omega_t} \{\widehat{z}_{i''}[\mathbf{x}(\omega)] - z_{i''}[\mathbf{x}(\omega, \omega_t)]\}^2 = \Delta_{var}(z_{i''}).$$

Let's denote

$$z_{i'}^{vc}[\mathbf{x}(\omega, \omega_t)] = \widehat{z}_{i'}[\mathbf{x}(\omega)] - z_{i'}[\mathbf{x}(\omega, \omega_t)], \quad z_{i''}^{vc}[\mathbf{x}(\omega, \omega_t)] = \widehat{z}_{i''}[\mathbf{x}(\omega)] - z_{i''}[\mathbf{x}(\omega, \omega_t)],$$

$$\rho_{i'i''}^{vc} = E_{\Omega} E_{\Omega_t} \{z_{i'}^{vc}[\mathbf{x}(\omega, \omega_t)] - z_{i''}^{vc}[\mathbf{x}(\omega, \omega_t)]\}^2.$$

Then

$$\begin{aligned}&\sum_{i'=1}^L \sum_{i''=1}^L c_i c_{i''} E_{\Omega} E_{\Omega_t} \{\widehat{z}_{i'}[\mathbf{x}(\omega)] - z_{i'}[\mathbf{x}(\omega, \omega_t)]\} \{\widehat{z}_{i''}[\mathbf{x}(\omega)] - z_{i''}[\mathbf{x}(\omega, \omega_t)]\} = \\ &= \sum_{i'=1}^L \sum_{i''=1}^L c_i c_{i''} \left\{ \frac{1}{2} [\Delta_{var}(z_{i'}) + \Delta_{var}(z_{i''})] - \rho_{i'i''}^{vc} \right\}.\end{aligned}$$

Thus,

$$\Delta_{\text{var}}(Z_{\text{ccp}}) = \sum_{i=1}^L c_i \Delta_{\text{var}}(z_i) - \frac{1}{2} \sum_{i'=1}^L \sum_{i''=1}^L c_{i'} c_{i''} \rho_{i'i''}^{\text{vc}} \quad (5)$$

So, structure of generalized error components  $\Delta_{\text{bias}}$  and  $\Delta_{\text{var}}$  for CCP forecasting practically coincides with the structure of mean squared error. At that both components are always lower then corresponding components for single predictors. In other words convex combining allows to improve both contradictory constituents of bias-variance dilemma.

---

### Variance of CCP

---

Let's consider structure of CCP squared variance. Let  $\hat{Z}_{\text{ccp}} = E_{\Omega}(Z_{\text{ccp}})$  and  $V_{\text{ccp}} = E_{\Omega}(Z_{\text{ccp}} - \hat{Z}_{\text{ccp}})^2$ .

Variance  $V_{\text{ccp}}$  may be written as  $\sum_{i'=1}^L \sum_{i''=1}^L c_{i'} c_{i''} E_{\Omega}[\hat{z}_{i'} - z_{i'}(\omega)][\hat{z}_{i''} - z_{i''}(\omega)]$ . Further calculations repeat calculations made for variance component structure evaluating:

$$\begin{aligned} & [\hat{z}_{i'} - z_{i'}(\omega)][\hat{z}_{i''} - z_{i''}(\omega)] = \\ & = -\frac{1}{2}[\hat{z}_{i'} - z_{i'}(\omega) - \hat{z}_{i''} + z_{i''}(\omega)]^2 + \frac{1}{2}[\hat{z}_{i'} - z_{i'}(\omega)]^2 + \frac{1}{2}[\hat{z}_{i''} - z_{i''}(\omega)]^2. \end{aligned}$$

Let's denote  $z_{i'}^v(\omega) = \hat{z}_{i'} - z_{i'}(\omega)$  and  $\rho_{i'i''}^v = E_{\Omega}[z_{i'}^v(\omega) - z_{i''}^v(\omega)]^2$ . Then

$$V(Z_{\text{ccp}}) = \sum_{i=1}^L c_i V(z_i) - \frac{1}{2} \sum_{i'=1}^L \sum_{i''=1}^L c_{i'} c_{i''} \rho_{i'i''}^v.$$

Thus, it is shown that variance of CCP forecasting is always lower than the same convex combination of forecasting variances related to single predictors. It must be noted that decrease of prediction variance leads to loss of forecasting ability. So, additional transformation of convex forecasting function must be done with the help of uni-dimensional linear regression:  $Z_{\text{ccp}}^t = \alpha_{\text{ccp}} Z_{\text{ccp}} + \beta_{\text{ccp}}$ , where  $Z_{\text{ccp}}^t$  is transformed forecasting,  $\alpha_i$  and  $\beta_i$  are real regression coefficients that may be found by training information with the help of least squares method. It is evident that any linear combination of predictors  $\sum_{i=1}^L \lambda_i z_i + \gamma_0$  with  $\lambda_i \geq 0$  ( $i = \overline{1, L}$ ) may be constructed by successive execution of convex correcting and uni-dimensional linear transformation.

---

### CCP optimization

---

Optimal CCP may be found by minimization of forecasting error. As it seen from expression (3) task of  $\delta(Z_{\text{ccp}})$  minimization may be reduced to quadratic programming task:

$$\sum_{i=1}^L c_i \delta(z_i) - \frac{1}{2} \sum_{i'=1}^L \sum_{i''=1}^L c_{i'} c_{i''} \rho_{i'i''}^v \rightarrow \min \quad (6)$$

$$\sum_{i=1}^L c_i = 1,$$

$$c_i \geq 0, \quad i = 1, \dots, L.$$

The quadratic programming task (6) is difficult NP complete problem. But solving (6) may be facilitated with the help of procedure evaluating whether all predictors from initial set give optimal CCP or some predictors are nuisance variables and may be removed. The problem will be discussed further in terms of predictors superfluity.

Subsets  $\mathbf{R}^L$ ,  $\bar{\mathbf{D}}_L$  and  $\mathbf{D}_L$  are defined as

$$\bar{\mathbf{D}}_L = \{ \mathbf{c} \mid \sum_{i=1}^L c_i = 1, c_i \geq 0, i = \overline{1, L} \},$$

$$\mathbf{D}_L = \{ \mathbf{c} \mid \sum_{i=1}^L c_i = 1, c_i > 0, i = \overline{1, L} \}.$$

A set of predictors will be called not superfluous or satisfying conditions of superfluity absence (CSA) if there exists a point  $\mathbf{c} \in \mathbf{D}_L$  such that  $\delta[Z_{ccp}(\mathbf{c})] < \delta[Z_{ccp}(\mathbf{c}')] , \forall \mathbf{c}' \in \bar{\mathbf{D}}_L \setminus \mathbf{D}_L$ .

CSA actually mean existing of CCP that uses all predictors and has forecasting error that is lower than error of any CCP that does not use all predictors. The necessary and sufficient conditions for CSA correctness that are formulated by theorem 1.

**Theorem 1** . Let matrix of mutual distances between predictors  $\|\rho_{i'i''}\|_{L \times L}$  is not singular and  $\|\rho_{i'i''}^-\|_{L \times L}$  is matrix inverse to  $\|\rho_{i'i''}\|_{L \times L}$ . Then simultaneous correctness of inequalities

$$\sum_{i'=1}^L \{ \delta(z_{i'}) \rho_{i'i}^- + \frac{\frac{1}{2} - \sum_{j'=1}^L \sum_{j''=1}^L \delta(z_{j'}) \rho_{j'j''}^-}{\sum_{j'=1}^L \sum_{j''=1}^L \rho_{j'j''}^-} \rho_{i'i}^- \} > 0$$

for  $i = \overline{1, L}$  and positiveness of quadratic form  $-\frac{1}{2} \sum_{j'=1}^L \sum_{j''=1}^L \rho_{j'j''} \varepsilon_{j'} \varepsilon_{j''}$  for each real vector  $\varepsilon_1, \dots, \varepsilon_L$ , such that

$$\sum_{j=1}^L \varepsilon_j = 0$$

is necessary and sufficient condition for CSA correctness.

**Method of CCP optimization based on CSA conditions.** A method for solving quadratic programming task (6) is proposed. It is based on gradual raising of predicates set meeting superfluity condition. First, a set of all possible predictor pairs  $P_2^{irr}$  is considered. A set of all irreducible pairs  $P_2^{irr}$  is then extracted using Theorem 1 results. Subsequently, a set of triplets  $P_3^{irr}$  is formed using  $P_2^{irr}$ . The process is going on until step i in which  $P_i^{irr}$  becomes empty. After that an optimal aggregate is chosen, which is one from  $P_{i-1}^{irr}$  with minimum error estimate.

---

### Experiments with CCP over uni-variate linear regressions

---

**CCP multiple linear model.** The goal of studies is performance evaluation of multiple linear regression model that is convex combination of simple regressions. At the initial stage parameters of simple linear regression models  $Y = \alpha_i + \beta_i X_i + \varepsilon_i$  are evaluated by training set with least squares method (LS) for each independent variable from initial set  $X$ . So, a set of  $L - |X|$  predictors is received:

$$\{Z_i(\omega) = \alpha_i + \beta_i X_i(\omega) \mid i = \overline{1, l}\}.$$

After that generalized errors of single predictors and discrepancies between predictors are estimated using leave-one-out technique. Then optimal CCP is searched as solution of quadratic programming task (1). Let  $\mathbf{c}^0 = \{c_1, \dots, c_l\}$  are optimal CCP coefficients. That gives a solution

$$Z(\omega, \mathbf{c}^0) = \sum_{i=1}^l c_i \alpha_i + \sum_{i=1}^l c_i \beta_i X_i(\omega).$$

Usually a majority of coefficients  $\mathbf{c}^0$  in high-dimensional tasks is equal to zero. So, task of CCP optimization also naturally incorporates another important task of regression analysis — significant variables selection.

CCP prognoses  $Z(\omega)$  may strongly correlate with  $Y$  but at the same time forecasting errors may be great due to low variance of  $Z(\omega)$ . So, additional linear transformation of  $Z(\omega)$  is necessary. Parameters of linear regression models  $Y = \alpha_{cpp} + \beta_{cpp} X_{cpp} + \varepsilon_{cpp}$  are evaluated by training information with LS. As a result the final CCP multiple linear model is received:

$$Z(\omega, \mathbf{c}^0) = \alpha_{cpp} + \sum_{i=1}^l c_i \beta_{cpp} \alpha_i + \sum_{i=1}^l c_i \beta_{cpp} \beta_i X_i(\omega) + \varepsilon_{cpp}.$$

It must be noted that methods of regression models optimization based on quadratic programming became rather popular last years. The known Lasso technique [3] may be mentioned thereupon.

**Scenarios for experiments**. In all studies dependent variable  $Y$  and regression variables  $X$  are stochastic functions of 3 latent variables  $U_1, U_2, U_3$ . The vector levels of variables  $U$  are independently distributed multivariate normal with mean 0 and standard deviation 1. The value of dependent variable  $y_j$  in j-th case is

generated by formula  $y_j = \sum_{k=1}^3 u_{jk} + e_j^y$  where  $u_{jk}$  is value of latent variable  $U_k$  and  $e_j^y$  is random error term distributed  $N(0, 1)$ .

The values of relevant variable  $X_i$  were generated by binary vector  $\beta^i = (\beta_1^i, \beta_2^i, \beta_3^i)$  In j-th case

$x_{ij} = \sum_{k=1}^3 \beta_k^i u_{jk} + e_j^{ix}$  where  $u_{jk}$  is value of latent variable  $U_k$ ,  $\sum_{k=1}^3 u_{jk} = 2$ ,  $e_j^{ix}$  is random error term

distributed  $N(0, 0.05)$  The levels of irrelevant variable  $X_i$  in j-th case is generated by formula  $x_{ij} = v_j^{ix}$  where  $v_j^{ix}$  is random error term distributed  $N(0, d_{ix})$ .

In each experiment 100 pairs of data sets were calculated by the random numbers generator according to the same scenario. The only exclusions are simulated tasks with size 50 and dimension 100. In these experiments too great amount of calculations was necessary for SR method. So only 50 pairs of data sets were generated (these results are marked asterisk in tables). Variables were selected and optimal regressions were calculated on one set from a pair and forecasting ability was evaluated on another.

*First scenario*. In all experiments number of relevant variables  $n_{rel}$  was fixed and equal 5: 2 were generated at  $\beta = (1, 1, 0)$ , 2 at  $\beta = (1, 0, 1)$ , 1 at  $\beta = (0, 1, 1)$ . Number of irrelevant variables  $n_{irrel}$  varied and was equal 5, 20, 45, 95.

Second scenario. In experiments by this scenario number of relevant variables  $n_{rel}$  was proportional to full number of variables  $n_{full}$ . Numbers  $n_{full}$ ,  $n_{rel}$  and numbers of relevant variables generated by different  $\beta$  levels are given in Table 3. Tables 4 and 5 has similar structure as Tables 1 and 2 respectively.

Table 1. Simulations parameters.

	$n_{irrel}$	$\beta = (1, 1, 0)$	$\beta = (1, 0, 1)$	$\beta = (0, 1, 1)$
$n_{full} = 25$	13	5	5	2
nfull = 50	25	10	10	5
nfull = 100	50	20	20	10

**Results.** Results of experiments of the first scenario are given in Tables 1-2. In Table 1 for each pair of sample size  $m$  and full number of variables  $n_{full}$  3 values are represented in corresponding cells: mean values of correlation coefficients between forecasted and true values of  $Y$  for CCP (upper left) and SR (upper right); fractions of tables where prognostic ability estimates for CCP regression was better than estimates for SR (bottom). In Table 2 numbers of correctly (top) and mistakenly (bottom) selected variables are represented both for CCP and SR.

Table 2. Results for the first scenario. Comparison of CCP and SR prognostic abilities.

	m = 20		m = 30		m = 50	
	CCP	SR	CCP	SR	CCP	SR
Nfull = 10	0.75	0.75	0.77	0.79	0.80	0.82
	0.43		0.30		0.36	
nfull = 25	0.78	0.64	0.78	0.72	0.79	0.77
	0.76		0.65		0.57	
nfull = 50	0.73	0.5	0.77	0.57	0.80	0.69
	0.83		0.90		0.84	
nfull = 100	0.75	0.5	0.76	0.53	0.79*	0.57*
	0.92		0.95		0.98*	

Table 3. Results of the second scenario experiments. Numbers of correctly and mistakenly selected variables

	m = 20		m = 30		m = 50	
	CCP	SR	CCP	SR	CCP	SR
nfull = 10	235	246	258	275	290	303
	3	60	1	47	0	52
nfull = 25	236	233	255	272	287	300
	11	272	3	239	0	197
nfull = 50	227	211	259	265	279	303
	28	603	4	719	0	565
nfull = 100	218	172	244	230	139*	153*
	37	725	6	1185	0*	946*

Table 4. Results of the second scenario experiments. Comparison of CCP and SR prognostic abilities

	m = 20		m = 30		m = 50	
	CCP	SR	CCP	SR	CCP	SR
nfull = 25	0.78	0.68	0.79	0.74	0.80	0.79
	0.79		0.61		0.51	
nfull = 50	0.75	0.6	0.78	0.62	0.80	0.73
	0.82		0.87		0.78	
nfull = 100	0.75	0.58	0.77	0.59	0.80*	0.57*
	0.86		0.95		0.98*	

Table 5. Results of the second scenario experiments. Numbers of correctly and mistakenly selected variables.

	m = 20		m = 30		m = 50	
	CCP	SR	CCP	SR	CCP	SR
nfull = 25	253	294	288	3111	332	348
	3	171	2	156	0	120

nfull = 50	283	391	326	498	368	451
	9	335	1	397	0	307
nfull = 100	281	448	319	670	196*	529*
	11	440	2	666	0*	510*

It is seen from tables that effectiveness of SR decrease dramatically when full number of regressor variables significantly exceeds number of cases in datasets. Prognostic ability of SR decreases from 0.75-0.82 for  $n_{full} = 10$  to 0.50-0.56 for  $n_{full} = 100$  in first scenario experiments and from 0.78-0.79 for  $n_{full} = 25$  to 0.57-0.59 for  $n_{full} = 100$  in second scenario experiments. Fraction of irrelevant variables in selected set exceed 50% in all first scenario experiments with  $n_{full} > 50$ . At the same time CCP regression keeps efficiency in all datasets. There is only slight decrease of prognostic ability for both scenarios: from 0.75-0.80 for  $n_{full} = 10$  to 0.75-0.795 for  $n_{full} = 100$  in first scenario experiments and from 0.78-0.80 for  $n_{full} = 25$  to 0.75-0.8 for  $n_{full} = 100$  in second scenario experiments. Fraction of irrelevant variables in selected set is small in all experiments.

---

## Conclusion

So it is shown that squared error of forecasting for CCP, CCP variance, bias and variance components of generalized error have the same structure:  $\sum_{i=1}^L c_i t_i - \frac{1}{2} \sum_{i=1}^L \sum_{i'=1}^L c_i c_{i'} \rho_{ii'}^*$ , where  $t_i$  is corresponding term for  $i$ -th single predictor,  $\rho_{ii'}^*$  is non-negative distance function between predictors  $Z_{i'}$  and  $Z_i$ , that is equal 0 when predictors coincide and increase when correlation between predictors at spaces  $\Omega$  or  $\Omega_i$  diminishes. Thus CCP procedures allows to improve both components of bias variance decomposition. On the other hand CCP decrease also full variance of predicting functions. So additional linear transformation of CCP collective solutions is necessary.

Problem of CCP optimization was discussed. It was shown that search of optimal CCP coefficients is reduced to quadratic programming task which is solved in terms of superfluity. Concept of ensemble superfluity in CCP was discussed in details. An ensemble of predictors is called superfluous if at least one of them may be removed without loss of prediction accuracy. Necessary and sufficient conditions of superfluity absence are given in Theorem 1. A method for solving quadratic programming task using Theorem 3 has been developed. A linear regression method based on CPP optimization was considered that inherently incorporates variables selection. Testing results reveal CCP's significant superiority over stepwise regression in high-dimensional task. Method preserves effectiveness of variables selection and prognostic ability in tasks where number of potential regressor variables is several times greater than number of cases in datasets. The described results can be used in different tasks of regression analysis, forecasting or recognition.

---

## Acknowledgment

The work was supported by Russian Foundation for Basic Research grants 08-07-00437-a, 08-01-00636-a and President's grant Scientific Schools 7950.2010.1.

---

**Bibliography**

---

- [1] Zhuravlev Yu.I., Kuznetsova A.V., Ryazanov V.V., Senko O.V., Botvin M.A. The Use of Pattern Recognition Methods in Tasks of Biomedical Diagnostics and Forecasting // Pattern Recognition and Image Analysis, MAIK Nauka/Interperiodica. 2008, Vol. 18, No. 2, pp. 195-200.
- [2] Zhuravlev Yu.I., Ryazanov V.V., Senko O.V. RECOGNITION. Mathematical methods. Program System. Applications. - Moscow: Phasiz, 2006, (in Russian).
- [4] Tibshirani R. Regression shrinkage and selection via the lasso // J.Roy.Stat.Soc..1996. Vol. 58,p.267–288.
- [5] A. Krogh, J. Vedelsby. Neural network ensembles, cross validation, and active learning. NIPS, 7:231–238, 1995.
- [6] Gavin Brown, Jeremy L. Wyatt, Peter Tino. Managing Diversity in Regression Ensembles. Journal of Machine Learning Research 6: 1621–1650.
- [7] S. Geman, E. Bienenstock, R. Doursat. Neural networks and the bias/variance dilemma. NeuralComputation, 4(1):1–58, 1992.

---

**Authors' Information**

---

**Oleg Senko** – Leading researcher in Dorodnicyn Computer Center of Russian Academy of Sciences, Russia, 119333, Moscow, Vavilova, 40, [senkoov@mail.ru](mailto:senkoov@mail.ru)

**Alexander Dokukin**– researcher in Dorodnicyn Computer Center of Russian Academy of Sciences, Russia, 119333, Moscow, Vavilova, 40, [dalex@ccas.ru](mailto:dalex@ccas.ru)



---

## REFERENCE-NEIGHBOURHOOD SCALARIZATION FOR MULTIOBJECTIVE INTEGER LINEAR PROGRAMMING PROBLEMS

Krassimira Genova, Mariana Vassileva

**Abstract:** Various real problems can be modeled as multicriteria optimization problems (MOP). In the general case, there is no single solution that optimizes all the criteria, but there is a set of solutions where improvement in the value of one criterion leads to deterioration in the value of at least another criterion. This set is known as a Pareto optimal set and any element of this set could be the final solution of the MOP. In order to select the final solution, additional information is necessary and it is supplied by the so-called decision maker. The quality of the interactive algorithms for solving MOP depends mainly on the scalarizing problems they are designed and based on. The scalarizing problems of the reference neighborhood, which are presented in the paper, are especially appropriate for solving multiobjective linear integer programming problems.

**Keywords:** *Multicriteria Linear Integer Optimization, Scalarizing problem.*

**ACM Classification Keywords:** *G.1.6. Optimization – Integer Programming*

---

### Introduction

Several criteria (or objective functions) are simultaneously optimized in the feasible set of solutions (or alternatives) in the problems of multicriteria optimization (MO). In the general case, there is no single solution that optimizes all the criteria. Instead, there is a set of solutions where improvement in the value of one criterion leads to deterioration in the value of at least another criterion. This set is known as a Pareto optimal set. Any element of this set could be the final solution of the multicriteria optimization problem. In order to select the final solution, additional information is necessary and it is supplied by the so-called decision maker (DM). The information that the DM gives reflects his/her global preferences with respect to the quality of the solution obtained. Generally, multicriteria optimization has to combine two aspects: optimization and decision support.

There are two main approaches in solving MO problems: a scalarizing approach and an approximation approach. The major representatives of the scalarizing approach are the interactive algorithms. Scalarization means transformation of the multicriteria optimization problem into one or several single-criterion optimization problems, called scalarizing problems. The quality of the interactive algorithms depends mainly on the scalarizing problems they are designed and based on. The main property of every scalarizing problem is that each optimal solution generated is a Pareto (or weakly Pareto) optimal solution of the original MO problem. The properties of the scalarizing problems of the reference neighborhood, presented in the paper, make them especially appropriate in realizing a “continuous-integer” approach in interactive algorithms for solving a general multiobjective linear integer programming problem /MOILP/. Instead of generating integer solutions for evaluation at each iteration, the DM may evaluate linear continuous solutions at most of the iterations. In the criteria space the PO integer solutions are placed relatively close to PO continuous solutions. Thus, in the learning phase, the DM could be trained on the basis of linear continuous solutions, instead on the basis of integer solutions, which is particularly important for MOILP problems with large dimensions. In addition, the DM may also learn on the basis of approximate weak PO solutions, found relatively near to a weak PO surface.

The present paper describes scalarizing problems of the reference neighborhood, called RNP1, RNP1e, RNP1-L, RNP1-Le, RNP1-L and RNP3. The rest of rest of the paper is organized as follows: the second section describes

the scalarizing problems of the reference neighborhood, called RNP1, RNP1e, RNP1-L, RNP1-Le, RNP1-L and RNP3. The third section presented the basic properties of the scalarizing problems of the reference neighborhood. Finally, the conclusions are given in the last section.

---

### Description of the scalarizing problems

---

Let us consider the multiobjective integer linear programming (MOILP) problems:

$$\text{"max"} \{f_k(x) = c^k x\}, k \in K \quad (1)$$

$$\text{s.t. } Ax \leq b \quad (2)$$

$$0 \leq x \leq d \quad (3)$$

$$x - \text{integer} \quad (4)$$

where  $x$  is an  $n$ -dimensional vector of variables,  $d$  is an  $n$ -dimensional vector of variables upper bounds,  $A$  is an  $m \times n$  matrix,  $b$  is the RHS vector and the vector  $c^i$  ( $i=1, \dots, k$ ) represents the coefficients of the objective functions. Constraints (2)-(4) define the feasible set of the variables (solutions)  $X_1$ . Problem (1)-(3) is a multiobjective linear programming (MOLP) problem, which is the weakened problem of a MOILP problem. The feasible set of the continuous solutions will be denoted by  $X_2$ .

Let  $Z$  denotes the feasible region in the criteria space, i.e. the set of points  $z \in \mathfrak{R}^k$  such that  $z_i = f_i(x), i=1, \dots, k, x \in X_1$ .  $x^* \in X_1$  is an *efficient* or *Pareto optimal* solution if there is no  $x \in X_1$  such that  $c^i x \geq c^i x^*$  for all  $i$  and  $c^i x > c^i x^*$  for at least one  $j$ .  $x^{**} \in X_1$  is said to be *weakly efficient* / *Pareto optimal* solution if there is no  $x \in X_1$  such that  $c^i x > c^i x^{**}$  for all  $i$ . Although the term "*efficient*" is more often used for points  $x$  and the term *Pareto optimal* (PO) for points  $z$ , they can be used interchangeably.

*Reference neighborhood* is the area around the currently preferred solution in the feasible criteria space of MOLP/MOILP problems, determined by the local preferences of the DM, in which the next currently preferred solution will be sought.

When solving a MOILP problem, the DM evaluates and compares the currently found (weak) PO solutions. In case he wants to look for a better solution, he/she sets his/her preferences for desired or feasible alterations of the values of a part or of all the criteria. Depending on these preferences, the criteria set can be implicitly divided into seven or less than seven classes:  $K^>$ ,  $K^{\geq}$ ,  $K^=$ ,  $K^<$ ,  $K^{\leq}$ ,  $K^{<>}$ , and  $K^0$ . Every criterion  $f_k(x)$ ,  $k \in K$  may belong to one of these classes:  $K^>$  - improvement as a desired direction of change;  $K^{\geq}$  - improvement by desired (aspiration) values  $\Delta_{hj}$ ;  $K^=$  - to either preserve or improve the current value of the criteria;  $K^<$  - acceptable deterioration as a desired direction of change;  $K^{\leq}$  - acceptable deterioration by no more than  $\delta_{hj}$ ;  $K^{<>}$  - the criteria value to lie within an interval,  $(a_{hj} - t_{hj}^- \leq a_{hj} \leq a_{hj} + t_{hj}^+)$  around the current value  $a_{hj}$ ;  $K^0$  - the DM is indifferent about the value of these criteria and as such they may be altered freely.

On the basis of this criteria division by the DM, the following scalarizing problem, denoted as RNP1, is suggested for finding a PO solution of MOILP problem:

$$\min_{x \in X_1} T(x) = \min_{x \in X_1} \max \left( \begin{array}{l} \max_{k \in K^{\geq}} \left( \frac{\tilde{f}_k - f_k(x)}{|\tilde{f}_k - f_k|} \right), \\ \max_{k \in K^{\leq}} \left( \frac{\tilde{f}_k - f_k(x)}{|\tilde{f}_k - f_k|} \right) \end{array} \right) + \max \left( \begin{array}{l} \max_{k \in K^{<}} \left( \frac{f_k - f_k(x)}{|f_k^* - f_k|} \right), \\ \max_{k \in K^{>}} \left( \frac{f_k - f_k(x)}{|f_k^* - f_k|} \right) \end{array} \right) + \delta \left( \sum_{k \in K^{\geq}} (\tilde{f}_k - f_k(x)) + \right. \quad (5)$$

$$\left. + \sum_{k \in K^{\leq}} (\tilde{f}_k - f_k(x)) + \sum_{k \in K^{<} \cup K^{>}} (f_k - f_k(x)) - \sum_{k \in K^{\leq} \cup K^{>} \cup K^0} f_k(x) \right)$$

s.t.  $f_k(x) \geq \tilde{f}_k, \quad k \in K^{>} \cup K^{>>} \cup K^{\leq} \cup K^{\leq}$  (6)

$$f_k(x) \leq \tilde{f}_k, \quad k \in K^{>>} \quad (7)$$

$$x \in X_1, \quad (8)$$

where  $\delta$  is an arbitrary small number,  $f_k$  is the value of the criterion with an index  $k \in K$  for the current preferred solution,  $\overline{f}_k = f_k + \Delta_k$  is the desired (aspiration) level of the criterion with an index  $k \in K^{\geq}$ ,

$$\tilde{f}_k = \begin{cases} f_k, & \text{if } k \in K^{\leq} \cup K^{>}, \\ f_k - t_k^-, & \text{if } k \in K^{>>}, \\ f_k - \delta_k, & \text{if } k \in K^{\leq}. \end{cases}$$

$$\tilde{f}_k = f_k + t_k^+, \quad \text{if } k \in K^{>>}$$

The current preferred solution of MOILP problem is a feasible solution of the current scalarizing problem RNP1, i.e., the scalarizing problem RNP1 has an initial feasible solution. This is a very important feature, because finding a feasible solution of integer problems is also a NP-problem. In addition, the feasible solutions of the scalarizing problem RNP1 are located near to the non-dominated surface of the multicriteria problem in the criteria space Z. They belong to the reference area, defined by DM's preferences.

**Theorem 1:** The optimal solution of the scalarizing problem RNP1 is an efficient/PO solution of MOILP problem.

**Proof:** The scalarizing problem RNP1 is solved, when the DM wants improvement with respect to one criterion at least, or when  $K^{\geq} \neq \emptyset$  or  $K^{>} \neq \emptyset$ .

Let  $x^*$  be the optimal solution of the scalarizing problem RNP1. Then the following conditions are satisfied:

$$T(x^*) \leq T(x), \quad x \in X_1, \quad (9)$$

and constraints (6)-(7).

Let us assume that  $x^* \in X_1$  is not an efficient/PO solution of the initial MOILP problem. In this case there must exist another  $x' \in X_1$ , that is an efficient/PO solution of MOILP problem, for which:

$$f_k(x') \geq f_k(x^*), \text{ for } k \in K, \quad f_k(x') > f_k(x^*) \text{ for at least one } k \in K \quad (10)$$

and constraints (6)-(7) are satisfied:

After transformation of the objective function  $T(x)$  of the scalarizing problem RNP1, using inequalities (10), the following relation is obtained:

$$\begin{aligned}
(11) \quad T(x') &= \max \left( \begin{array}{l} \max_{k \in K^{\geq}} \left( \frac{\bar{f}_k - f_k(x')}{|\bar{f}_k - f_k|} \right) \\ \max_{k \in K^{\leq}} \left( \frac{\tilde{f}_k - f_k(x')}{|\tilde{f}_k - f_k|} \right) \end{array} \right) + \max \left( \begin{array}{l} \max_{k \in K^{<}} \left( \frac{f_k - f_k(x')}{|f_k^* - f_k|} \right) \\ \max_{k \in K^{>}} \left( \frac{f_k - f_k(x')}{|f_k^* - f_k|} \right) \end{array} \right) + \delta \left( \sum_{k \in K^{\geq}} (\bar{f}_k - f_k(x')) + \right. \\
&\quad \left. + \sum_{k \in K^{\leq}} (\tilde{f}_k - f_k(x')) + \sum_{k \in K^{<} \cup K^{>}} (f_k - f_k(x')) - \sum_{k \in K^{\leq} \cup K^{>} \cup K^0} f_k(x') \right) = \\
&= \max \left( \begin{array}{l} \max_{k \in K^{\geq}} \left( \frac{\bar{f}_k - f_k(x^*)}{|\bar{f}_k - f_k|} + \frac{f_k(x^*) - f_k(x')}{|\bar{f}_k - f_k|} \right) \\ \max_{k \in K^{\leq}} \left( \frac{\tilde{f}_k - f_k(x^*)}{|\tilde{f}_k - f_k|} + \frac{f_k(x^*) - f_k(x')}{|\tilde{f}_k - f_k|} \right) \end{array} \right) + \\
&\quad + \max \left( \begin{array}{l} \max_{k \in K^{<}} \left( \frac{f_k - f_k(x^*)}{|f_k^* - f_k|} + \frac{f_k(x^*) - f_k(x')}{|f_k^* - f_k|} \right) \\ \max_{k \in K^{>}} \left( \frac{f_k - f_k(x^*)}{|f_k^* - f_k|} + \frac{f_k(x^*) - f_k(x')}{|f_k^* - f_k|} \right) \end{array} \right) + \\
&\quad + \delta \left( \sum_{k \in K^{\geq}} ((\bar{f}_k - f_k(x^*)) + (f_k(x^*) - f_k(x'))) + \sum_{k \in K^{\leq}} ((\tilde{f}_k - f_k(x^*)) + (f_k(x^*) - f_k(x'))) + \right. \\
&\quad \left. + \sum_{k \in K^{<} \cup K^{>}} ((f_k - f_k(x^*)) + (f_k(x^*) - f_k(x'))) - \sum_{k \in K^{\leq} \cup K^{>} \cup K^0} f_k(x^*) + (f_k(x') - f_k(x^*)) \right) < \\
&< \max \left( \begin{array}{l} \max_{k \in K^{\geq}} \left( \frac{\bar{f}_k - f_k(x^*)}{|\bar{f}_k - f_k|} \right) \\ \max_{k \in K^{\leq}} \left( \frac{\tilde{f}_k - f_k(x^*)}{|\tilde{f}_k - f_k|} \right) \end{array} \right) + \max \left( \begin{array}{l} \max_{k \in K^{<}} \left( \frac{f_k - f_k(x^*)}{|f_k^* - f_k|} \right) \\ \max_{k \in K^{>}} \left( \frac{f_k - f_k(x^*)}{|f_k^* - f_k|} \right) \end{array} \right) + \delta \left( \sum_{k \in K^{\geq}} (\bar{f}_k - f_k(x^*)) + \right. \\
&\quad \left. + \sum_{k \in K^{\leq}} (\tilde{f}_k - f_k(x^*)) + \sum_{k \in K^{<} \cup K^{>}} (f_k - f_k(x^*)) - \sum_{k \in K^{\leq} \cup K^{>} \cup K^0} f_k(x^*) \right) = T(x^*)
\end{aligned} \tag{11}$$

It follows from (11) that  $T(x') < T(x^*)$  and constraints (6)-(7), that contradicts to (9). Hence  $x^*$  is an efficient solution, and  $f(x^*)$  is a PO solution in the criteria space of MOILP problem.

In order to find a PO solution of MOLP problem, scalarizing problem RNP1 may be used, in which the constraint requiring integers (8) is removed. The relaxing problem obtained is denoted as RNP1-L.

Since the objective functions of the scalarizing problems RNP1 and RNP1-L are non-differentiable, each one of them may be reduced to the equivalent optimization problem, on the account of additional variables and constraints, but with a differentiable objective function [7]. The equivalent linear integer problem of RNP1 problem, denoted as RNP1e, can be presented as follows:

$$\min_{x \in X_1} \left( \alpha + \beta + \delta \sum_{k \in K} y_k \right) \tag{12}$$

$$\text{s.t. } \alpha \geq (\bar{f}_k - f_k(x)) / |\bar{f}_k - f_k|, \quad k \in K^{\geq} \tag{13}$$

$$\alpha \geq (\tilde{f}_k - f_k(x)) / |\tilde{f}_k - f_k|, \quad k \in K^{\leq} \tag{14}$$

$$\beta \geq (f_k - f_k(x)) / |f_k^* - f_k|, \quad k \in K^{<} \tag{15}$$

$$\beta \geq (f_k - f_k(x)) / |f_k^* - f_k|, \quad k \in K^{>} \tag{16}$$

$$\bar{f}_k - f_k(x) = y_k, \quad k \in K^{\geq} \tag{17}$$

$$\tilde{f}_k - f_k(x) = y_k, \quad k \in K^{\leq} \quad (18)$$

$$f_k - f_k(x) = y_k, \quad k \in K^{<} \cup K^{>} \quad (19)$$

$$f_k(x) = y_k, \quad k \in K^= \cup K^0 \cup K^{>>} \quad (20)$$

$$f_k(x) \geq \tilde{f}_k, \quad k \in K^{>} \cup K^{>>} \cup K^= \cup K^{\leq} \quad (21)$$

$$f_k(x) \leq \tilde{f}_k, \quad k \in K^{>>} \quad (22)$$

$$x \in X_1 \quad (23)$$

$$\alpha, \beta, y_k, k \in K - \text{arbitrary.} \quad (24)$$

Problems RNP1 and RNP1e have one and the same feasible set of solutions, and the values of their objective functions are equal. This comes from the following affirmation:

**Theorem 2:** The optimal values of scalarizing problems RNP1 and RNP1e are equal, i.e.

$$\min_{x \in X_1} (\alpha + \beta + \delta \sum_{k \in K} y_k) = \min_{x \in X_1} \left( \max \left( \max_{k \in K^{\geq}} \left\{ \frac{\tilde{f}_k - f_k(x)}{|\tilde{f}_k - f_k|} \right\}, \max_{k \in K^{\leq}} \left\{ \frac{\tilde{f}_k - f_k(x)}{|\tilde{f}_k - f_k|} \right\} \right) + \max \left( \max_{k \in K^{<}} \left\{ \frac{f_k - f_k(x)}{|f_k^* - f_k|} \right\}, \max_{k \in K^{>}} \left\{ \frac{f_k - f_k(x)}{|f_k^* - f_k|} \right\} \right) + \delta \left( \sum_{k \in K^{\geq}} (\tilde{f}_k - f_k(x)) + \sum_{k \in K^{\leq}} (\tilde{f}_k - f_k(x)) + \sum_{k \in K^{<} \cup K^{>}} (f_k - f_k(x)) - \sum_{k \in K^= \cup K^{>>} \cup K^0} f_k(x) \right) \right)$$

**Proof:** It follows from (13) that  $\alpha \geq (\tilde{f}_k - f_k(x)) / |\tilde{f}_k - f_k|, k \in K^{\geq}$ . Since this inequality is true for every  $k \in K^{\geq}$ , then it is also true that

$$\alpha \geq \max_{k \in K^{\geq}} (\tilde{f}_k - f_k(x)) / |\tilde{f}_k - f_k| \quad (25)$$

From (14) it could be noticed that  $\alpha \geq (\tilde{f}_k - f_k(x)) / |\tilde{f}_k - f_k|, k \in K^{\leq}$ . Since this inequality is true for every  $k \in K^{\leq}$ , then it is also true that

$$\alpha \geq \max_{k \in K^{\leq}} (\tilde{f}_k - f_k(x)) / |\tilde{f}_k - f_k| \quad (26)$$

Using (25) and (26) the following is derived:

$$\alpha \geq \max \left( \max_{k \in K^{\geq}} (\tilde{f}_k - f_k(x)) / |\tilde{f}_k - f_k|, \max_{k \in K^{\leq}} (\tilde{f}_k - f_k(x)) / |\tilde{f}_k - f_k| \right) \quad (27)$$

In the same way it follows from (15):

$$\beta \geq \max_{k \in K^{<}} ((f_k - f_k(x)) / |f_k^* - f_k|) \quad (28)$$

and from (16):

$$\beta \geq \max_{k \in K^{>}} ((f_k - f_k(x)) / |f_k^* - f_k|) \quad (29)$$

From (28) and (29) it can be written for  $\beta$ :

$$\beta \geq \max \left( \max_{k \in K^{<}} ((f_k - f_k(x)) / |f_k^* - f_k|), \max_{k \in K^{>}} ((f_k - f_k(x)) / |f_k^* - f_k|) \right) \quad (30)$$

In case the left and right sides of inequalities (27) and (30) are summed up, the following relation is obtained:

$$(\alpha + \beta) \geq \max \left( \max_{k \in K^{\geq}} (\bar{f}_k - f_k(x)) / |\bar{f}_k - f_k|, \max_{k \in K^{\leq}} (\tilde{f}_k - f_k(x)) / |\tilde{f}_k - f_k| \right) + \max \left( \max_{k \in K^{<}} (f_k - f_k(x)) / |f_k^* - f_k|, \max_{k \in K^{>}} (f_k - f_k(x)) / |f_k^* - f_k| \right) \quad (31)$$

If the term  $\delta \sum_{k \in K} y_k$  is added to both sides of (31), the following inequality is obtained:

$$(\alpha + \beta + \delta \sum_{k \in K} y_k) \geq \max \left( \max_{k \in K^{\geq}} (\bar{f}_k - f_k(x)) / |\bar{f}_k - f_k|, \max_{k \in K^{\leq}} (\tilde{f}_k - f_k(x)) / |\tilde{f}_k - f_k| \right) + \max \left( \max_{k \in K^{<}} (f_k - f_k(x)) / |f_k^* - f_k|, \max_{k \in K^{>}} (f_k - f_k(x)) / |f_k^* - f_k| \right) + \delta \sum_{k \in K} y_k \quad (32)$$

Let  $x^*$  be the optimal solution of RNP1e problem. Then

$$\min_{x \in X_1} (\alpha + \beta + \delta \sum_{k \in K} y_k) = \min_{x \in X_1} \left\{ \max \left( \max_{k \in K^{\geq}} \left( \frac{\bar{f}_k - f_k(x^*)}{|\bar{f}_k - f_k|} \right), \max_{k \in K^{\leq}} \left( \frac{\tilde{f}_k - f_k(x^*)}{|\tilde{f}_k - f_k|} \right) \right) + \max \left( \max_{k \in K^{<}} \left( \frac{f_k - f_k(x^*)}{|f_k^* - f_k|} \right), \max_{k \in K^{>}} \left( \frac{f_k - f_k(x^*)}{|f_k^* - f_k|} \right) \right) + \delta \sum_{k \in K} y_k \right\} \quad (33)$$

for in the opposite case  $(\alpha + \beta + \delta \sum_{k \in K} y_k)$  could be still decreased. Since after summing up inequalities (17-20) the relation given below is obtained

$$\left( \sum_{k \in K^{\geq}} (\bar{f}_k - f_k(x)) + \sum_{k \in K^{\leq}} (\tilde{f}_k - f_k(x)) + \sum_{k \in K^{<} \cup K^{>}} (f_k - f_k(x)) - \sum_{k \in K^{\cup K^{>} \cup K^0}} f_k(x) \right) = \sum_{k \in K} y_k,$$

then the right side of (33) can be written down also as:

$$\min_{x \in X_1} (\alpha + \beta + \delta \sum_{k \in K} y_k) = \min_{x \in X_1} \left( \max \left( \max_{k \in K^{\geq}} \left( \frac{\bar{f}_k - f_k(x)}{|\bar{f}_k - f_k|} \right), \max_{k \in K^{\leq}} \left( \frac{\tilde{f}_k - f_k(x)}{|\tilde{f}_k - f_k|} \right) \right) + \max \left( \max_{k \in K^{<}} \left( \frac{f_k - f_k(x)}{|f_k^* - f_k|} \right), \max_{k \in K^{>}} \left( \frac{f_k - f_k(x)}{|f_k^* - f_k|} \right) \right) + \delta \left( \sum_{k \in K^{\geq}} (\bar{f}_k - f_k(x)) + \sum_{k \in K^{\leq}} (\tilde{f}_k - f_k(x)) + \sum_{k \in K^{<} \cup K^{>}} (f_k - f_k(x)) - \sum_{k \in K^{\cup K^{>} \cup K^0}} f_k(x) \right) \right),$$

which proves the theorem.

The continuous scalarizing problem RNP1-L has a similar equivalent problem of linear programming RNP1-Le, in which the feasible set of the variables  $x \in X_2$  is expanded. In order to find more than one weak PO solutions of the continuous scalarizing problem RNP1-Le, its parametric extension, called RNP1-Lp, may be applied:

$$\min (\alpha + \beta)$$

s.t.

$$f_k(x) + |\bar{f}_k - f_k| \alpha \geq f_k + (\bar{f}_k - f_k)t, \quad k \in K^{\geq},$$

$$f_k(x) + |\tilde{f}_k - f_k| \alpha \geq f_k + (\tilde{f}_k - f_k)t, \quad k \in K^{\leq}$$

$$\begin{aligned}
 f_k(x) + |f_k^* - f_k| \beta &\geq f_k - t, \quad k \in K^<, \\
 f_k(x) + |f_k^* - f_k| \beta &\geq f_k + t, \quad k \in K^>, \\
 f_k(x) &\geq \tilde{f}_k, \quad k \in K^> \cup K^{><} \cup K^= \cup K^{\leq}, \\
 f_k(x) &\leq \tilde{f}_k, \quad k \in K^{><} \\
 x &\in X_2, \quad t \geq 0, \quad \alpha, \beta \text{ - arbitrary.}
 \end{aligned}$$

With the help of this parametric problem more new (weak) PO solutions of MOLP problem are sought, shifting the reference neighbourhood to direction of the preferences, set by the DM for desired improvement and acceptable deterioration of some of the criteria. In thus way it will not be necessary for the DM to make one more iteration in direction of improvement of the selected criteria, in order to acquire knowledge about the compromises that he has to accept with respect to the rest of the criteria. Let us assume that a (weak) PO solution of MOLP problem has been found with the help of scalarizing problems RNP1-Le or RNP1-Lp, evaluated by the DM as satisfactory. Let it be denoted by  $\hat{f} = (\hat{f}_1, \dots, \hat{f}_p)^T$ . In order to find a (weak) Pareto optimal solution of MOILP problem, close to the solution  $\hat{f}_k$ , the following Tchebycheff's problem RNP3 [7] may be used:

$$\min_{x \in X_1} S(x) = \min_{x \in X_1} \max_{k \in K} (|\hat{f}_k - f_k(x)| / |\hat{f}_k'|),$$

where

$$\hat{f}_k' = \begin{cases} \hat{f}_k, & \text{if } |\hat{f}_k| > \varepsilon, \\ \varepsilon, & \text{if } |\hat{f}_k| \leq \varepsilon. \end{cases} \quad \varepsilon \text{ is a small positive number.}$$

This problem is equivalent to the following problem of mixed integer programming RNP3e:

$$\min \alpha$$

s.t.

$$\begin{aligned}
 \alpha &\geq (|\hat{f}_k - f_k(x)| / |\hat{f}_k'|), \\
 x &\in X_1,
 \end{aligned}$$

MOILP problems belong to the class of NP-problems. The computational efforts, connected with optimal solution finding in them depend considerably on the form of the feasible solutions area. That is why the search for new scalarizing problems, formulated in a way that could direct faster to the optimal solution, close to DM's preferences, is a constant task which all the researchers face.

---

### Basic properties of the scalarizing problems of the reference neighborhood

---

The scalarizing problems of the reference neighborhood (RNP) are formulated on the basis of inexplicit classification of the criteria in accordance with DM's preferences, aimed at improvement of the current preferred solution. Similar scalarizing problems have been offered in [3], [4], [5], [6], which also use inexplicit criteria division into groups. An open communication protocol to interact with the DM, which enables free exploration of the problem and progressive learning of the non-dominated solution set [2], is applied in the interactive algorithms, based on classification-oriented scalarizing problems. Four features of RNP problems facilitate the dialogue with the DM with respect to the information required about his/her local preferences, with respect to decrease of the time the DM has to wait for finding a new solution and also with respect to easing DM's efforts in the evaluation of more than one solutions:

- the DM has variable possibilities to express his/her preferences, concerning the alteration of the values of some or all the criteria with respect to the current solution found. He/she may apply the most appropriate and comprehensible approach for every criterion, setting levels of attainability or a feasible trade-off, define the direction of improvement only, or specify the range of a feasible alteration;
- the current preferred solution of MOILP problems, found at the previous iteration, is a feasible solution of the integer scalarizing problems RNP1e, which are to be solved at the next iteration. This reduces considerably the computational effort, since the discovery of a feasible solution of a single-criterion integer problem is an NP-problem;
- the feasible solutions of the scalarizing problems of the reference neighborhood are located near to PO surface of MOILP problems. The feasible area of RNP scalarizing problems is a part of the feasible area of MOILP problems, unlike the feasible area of the scalarizing problems of the reference point, which coincides with it. Depending on the way, in which the DM sets his/her preferences (by values of desired improvement and acceptable deterioration), this feasible area may be comparatively narrow and the feasible solutions in the criteria space, found with the help of approximate algorithms of integer programming, might be located close to the non-dominated surface of MOILP problems. The use of approximate weak PO solutions leads to considerable decrease of the time duration, when the DM is expecting a new solution for evaluation and choice. In this way, on the account of slight worsening in the quality of the solutions obtained, the dialogue with the DM can be improved;
- the strategy “not high profits - small losses” is realized. This is achieved, since the optimal solution of RNP scalarizing problems aspires to minimize the maximal Tchebycheff distance to the current preferred solution, both in direction of an improvement and also in direction of a compromise deterioration of the criteria, determined by DM’s preferences. The solutions from the reference neighborhood, that are obtained, are comparatively close, which facilitates the DM in his/her comparing of several solutions and also in selecting of the new currently preferred solution;
- the application of the interactive algorithm in RNP1Lp scalarizing problem will result in finding more than one (weak) PO solutions of MOLP problems. It is obvious, that in case the DM evaluates at each iteration more (weak) PO solutions, then he/she will learn faster in problem specifics and will find quicker the most preferred solution to MOILP problems.

---

## Conclusion

In the area of multicriteria decision making the interactive approach has found wide application in a specific and well defined class of algorithms, which aid the DM in the study of a set of solutions with the purpose to select one or a limited set from them. These algorithms, built on the basis of open communication protocols with the DM, prove to be some of the most promising ways of research to develop adequate MOILP tools for decision aiding in many complex practical situations [1]. Undoubtedly, in recent years the interest towards explicit multiple objectives inclusion in different real life application areas of integer programming models, has grown significantly.

The quality of the interactive algorithms depends mainly on the scalarizing problems they are designed and based on. The properties of the scalarizing problems of the reference neighborhood herein proposed make them especially appropriate in realizing a “continuous-integer” approach in interactive algorithms for solving a general MOILP problem. Instead of generating integer solutions for evaluation at each iteration, the DM may evaluate linear continuous solutions at most of the iterations. In the criteria space the PO integer solutions are placed relatively close to PO continuous solutions. Thus, in the learning phase, the DM could be trained on the basis of linear continuous solutions, instead on the basis of integer solutions, which is particularly important for MOILP problems with large dimensions. In addition, the DM may also learn on the basis of approximate weak PO



solutions, found relatively near to a weak PO surface. Another advantage of this class of scalarizing problems is that in them the DM operates in the criteria space. In real life problems the criteria have their quantitative and financial aspects, so that the DM is able through their definition to express easier his/her preferences in the search for a trade-off solution.

---

## Bibliography

---

- [1] Alves M.J. and Climaco J. A review of interactive methods for multiobjective integer and mixed-integer programming. *European Journal of Operational Research*, 180, 99-115, 2007.
- [2] Alves M.J. and Climaco J. A note on a decision support system for multiobjective integer and mixed-integer programming problems. *European Journal of Operational Research*, 155, 258-265, 2004.
- [3] Narula, S. C. and Vassilev, V. An Interactive Algorithm for Solving Multiple Objective Integer Linear Programming Problems. *European Journal of Operational Research*, 79, 443-450, 1994.
- [4] Vassilev V., Genova K., Vassileva M., Narula S. Classification-Based Method of Linear Multicriteria Optimization. *International Journal on Information Theories and Applications*, 10, 3, 266-270, 2003.
- [5] Vassilev V., Narula S. and Gouljaski V. An Interactive Reference Directions Algorithm for Solving Multiobjective Convex Nonlinear Integer Programming Problems, *International Transactions in Operational Research*, 8, 367-380, 2001.
- [6] Vassileva, M. A Learning –oriented Method of Linear Mixed Integer Multicriteria Optimization, *Cybernetic and Information Technologies*, 1,13-25, 2004.
- [7] Wolsey, L. A.. *Integer Programming*, Wiley-InterScience, 1998.

---

## Acknowledgements

---

The paper is partially supported by the Bulgarian National Science Fund, Grant No DTK02/71 “Web-Based Interactive System, Supporting the Building Models and Solving Optimization and Decision Making Problems” and by IIT-BAS by the project “Research, development and application of effective methods for solving NP-hard single and multiobjective programming problems”.

---

## Authors' Information

---

**Mariana Vassileva** – Assoc.Prof., PhD, Institute of Information Technologies, BAS, Acad. G. Bonchev St., bl. 29A, Sofia 1113, Bulgaria; e-mail: [mvassileva@iinf.bas.bg](mailto:mvassileva@iinf.bas.bg)

**Krasimira Genova** - Assoc.Prof., PhD, Institute of Information Technologies, BAS, Acad. G. Bonchev St., bl. 29A, Sofia 1113, Bulgaria; e-mail: [kgenova@iinf.bas.bg](mailto:kgenova@iinf.bas.bg)

## MULTIAGENT APPLICATIONS IN SECURITY SYSTEMS: NEW PROPOSALS AND PERSPECTIVES

Vladimir Jotsov

**Abstract:** *The topic of the presented investigation is the contemporary threats, that will lead to big problems in the nearest future. The prevention of such threats is impossible without applications of intelligent agents. Even more, the multi-agent system should possess some features of knowledge discovery, web mining, collective evolutionary systems, and other advanced features which are impossible to be applied in only one agent. Advantages and disadvantages of synthetic data mining methods are investigated, and obstacles are revealed to their distribution in information security systems. Original results for juxtaposing statistical vs. logical data mining methods aiming at possible evolutionary fusions are described, and recommendations are made on how to build more effective applications of classical and/or presented novel methods: kaleidoscope, funnel, puzzle, frontal and contradiction. The usage of ontologies is investigated with the purpose of information transfer by sense in security agent environments or to reduce the computational complexity of practical applications. It is shown that human-centered methods are very suitable for resolutions in case, and often they are based on the usage of dynamic ontologies. Practical aspects of agent applications are discussed at the information security and/or the national security levels. Other cryptography applications, multiple software and e-learning research results are mentioned aiming to show that intelligent and classical technologies should be carefully combined in one software/hardware complex to achieve the goals of the security. It is shown that all the demonstrated advantages may be successfully combined with other known methods and information security technologies.*

**Key words:** *agent, knowledge discovery, data mining, web mining, ontology, information security systems, national security, human-centered systems, knowledge management, automation of creative processes, human-machine brainstorming methods.*

---

### 1. Introduction

Contemporary Information Security Systems (ISS) and especially the web-based systems are a wide field for applications of modern methods and technologies. The need to create sufficiently effective and universal tools to protect computer resources grows every year in systems for detection and prevention from intrusions (Intrusion Detection Systems IDS, Intrusion Prevention Systems IPS). For this reason different applications of intelligent data processing are initiated based on a combination of methods from statistical and logical information processing [1,2]. Other elaborations with growing influence in this domain are artificial immune systems and multiagent systems [3,4]. The unifying factor is the longer life cycle, elaborations require bigger teams and time for introduction. Due to the complex structure the prevention from direct attacks against these systems is very challenging.

Modern applications of statistical methods are effective and convenient to use at the expense of information encapsulation. In other words, it is impossible to construct tools to acquire new knowledge or to solve other problems of logical nature in this area. If we split methods in two groups (quantitative and qualitative) then statistical methods belong to the first group and logical ones belong to the second group. For this reason, their mechanical union is of no perspective. We do not attempt to propose any isolated solutions, instead we offer a combination of novel methods that is well adjustable to the existing ones. Our research includes a new evolutionary metamethod for joint control of statistical and logical methods where the statistical approach is

widely applied on the initial stage of the research when the information about the problem is scanty and it is possible to choose the solution arbitrarily [5]. The accumulation of knowledge makes logical applications more and more effective and more universal than the probabilistic ones, as well as fuzzy estimates and similar applications. The paper introduces the SMM (Synthetic MetaMethod) metamethod to control the process of consecutive replacement of applications by other ones and is synthetic by nature. The difference from the classical analytic methods is in the fact that the design of systems controlled by synthetic (meta)methods is not just science, sometimes it is an art. If we make an analogy with the set of traditional methods and fashion clothes then the synthetic method will apply the design of the display window with the fashion clothes. In the common case during intelligent data processing, there is no convergence of the results but this does not hamper practical applications of these systems. In other words, bad and good designers will arrange the display window in quite different ways and there is no guarantee that every user understands the technology and that his access to the system will have positive results.

The cited innovations are made and demonstrated here for the following reason. The problems presented above show that there is a need to introduce elements of machine creative work in ISS. It is demonstrated in the paper that this goal is accessible, if the usage of possibilities for human-machine contact and a set of comparatively simple intelligent technologies are done in the right way. On the other hand, the innovations can hardly be described in a single paper using the traditional academic style. For this reason, in the paper we avoided when possible the technical details and formalizations, and for the sake of the contents reduction descriptions by analogy are used, illustration visual aids and other nonstandard approaches.

How can security agents operate autonomously? In the first place this is because of the usage of neural networks where the agent most conscientiously copies the acts of the teacher. At that the agent itself does not understand the sense of teacher's transfer of knowledge, it operates as an universal approximator instead. It is shown schematically at Fig. 1 in the following way.

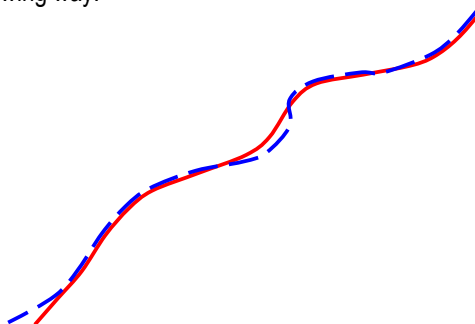


Fig. 1. Approximation of teacher's activities

Teacher's acts are principally presented as a dotted curve. The neural network approximates this curve via a continuous line on the same figure ; the deviation between the two lines must not exceed 3%. One of the main disadvantages of this approach is related to the necessity the agents in ISS to apply the learned knowledge in a rather creative way because frequently they operate in unexpected by the teacher situations. And as they are poorly trained or they are not at all trained how to act in unexpected situations, this approach as a whole is not very effective in its classical appearance. The presented approaches in anomaly ISS are combined with statistical applications which count for the normal traffic and other mean statistical values related to the operation of the guarded place. In Fig. 2 this process is schematically shown in the following way.

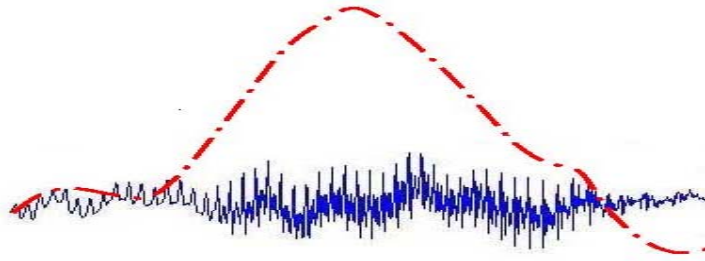


Fig. 2. Anomaly detection

The normal traffic is presented via a continuous line and the dotted line shows a case with drastic deviation from the normal traffic which is considered as an anomaly and it is analyzed whether it is a consequence from an intrusion. One of the disadvantages of this approach lies in the fact that intruder(s) know the principles of operation of such systems and they can 'fit' in the constraints of the normal traffic.

There is a variety of other applied approaches when the system is overloaded by heuristic information but they are not discussed here due to their evident weakness for ISS.

Here we introduce a new way for agents' operation. Best of all is the usage of ontologies to model the domain but this is not obligatory. On the other hand, as it is shown in the next section, the multiagent system functions more effectively if a system of ontologies is included.

But ontologies also do not contribute to a great extent in order to understand the sense of matters by the agents or with respect to transmit the sense of matters during communications between the agents. On the way to produce an analogy to how agents think we offer evolutionary methods to process data or knowledge because thinking (and understanding) is an evolutionary process. The problem is how to direct evolutionary methods with no exaggerations of heuristics, statistical information and other relative methods. Our approach is rather untraditional. We have elaborated for more than twenty years methods to detect and to resolve contradictions. A method for machine learning is built based on them. Searching and solving conflicts and contradictions the agent improves its knowledge and at the same time it may solve other problems. Detection of contradictions is based on using models of contradictions that can also be improved gradually. The agent may request an external help to solve the conflicts but this takes place only in extraordinary situations. At the same time it is shown how to change the reasoning component of security agents. Different logical methods are used that are rather analogous to means-ends analysis, constraint satisfaction, variable fitness function, brainstorming, and cognitive graphics. The combination of new methods to a great extent mechanizes creative efforts and also it serves agents' operation to improve the abilities of security experts working with similar types of systems.



Fig. 3. Example of supervised learning by using critical analysis

Suggested innovations serve the more effective application of data mining, Web mining, collective evolutionary components in multiagent systems. They are very well combined especially for applications of evolutionary approaches with classical neuro-fuzzy, statistical applications, genetic algorithms, etc. methods for the domain. For example such systems critically accept teacher's acts in cases of supervised learning (Fig. 3): they may precise or argue teacher's acts and in this way they can learn more effectively and deep.

A wide application of intelligent agents is forecasted in the field of information security systems. This will lead to the situations when the agent has no possibility to learn from the expert (teacher) but should swiftly learn from other agents or should self-learn without teacher. Then the role of the above considered critical learning will significantly increase.

---

## 2. Knowledge Discovery Methods

---

A synthetic metamethod (SMM) is elaborated and applied aiming at application of a set of 'creative' elements in agent environments. They work most effectively as a system, but even particular elements of them, let's say, applied in semantic reasoners, are proved to be very useful. Their principles are easy-to-be explained: bind the unknown knowledge from the goal with the knowledge from the knowledge base, apply a flexible constraint system to manage a system of variable fitness/goal functions, make the goals automatically via self-improvement of the existing knowledge, etc. The concise goals are better applied in the intelligent agents.

---

### 2.1 Puzzle Method

---

The basic methods of the suggested metamethod SMM are presented below. Let the constraints connected to the defined problem form a line in the space described by the equation (1).

$$\frac{x - x_1}{x_2 - x_1} = \frac{y - y_1}{y_2 - y_1} = \frac{z - z_1}{z_2 - z_1} \quad (1)$$

For example if a bachelor who has graduated SALSIT lives in Sofia and he/she does not want to work anywhere else, then the line restricts the search space and in this way a lot of unnecessary work is avoided. It is also possible to inspect a case when the constraint is defined as a type of surface but as a result a more general solution is obtained where a special interest is provoked by the boundary case of the crossing of two or more surfaces. When the common case is inspected in details, then in the majority of cases the problem is reduced to exploring the lines of type (1) or to curves with complicated forms obtained as a result of crossing surfaces by constraints. Therefore, below we investigate the usage in systems of constraints by lines of first or higher orders.

If the mentioned curves have common points of intersection and if they lie in a common plain so that a closed figure (triangle, tetragon, etc.) is formed then the search space is significantly reduced and it is searched much easier. The practical usage of the classical technology, as well as the constraint satisfaction, is complicated by the following. The viewed plains are not only nonlinear in the common case but they also include fuzzy estimates. The usage of fuzzy logic significantly raises the algorithmic complexity of the problem and it can make the application ineffective. Even when the usage of constraints significantly reduces the number of the inspected solutions, for example up to 10, this does not mean that the problem is solved and that all that must be done is to explore the possibilities one by one.

The following example below shows how the search process is reduced via using ontologies. Let's admit that the search space is presented on fig. 4 where statistical data about ISS are generalized about the regions depending on their price and quality. It is necessary to select an acceptable ISS to our project.

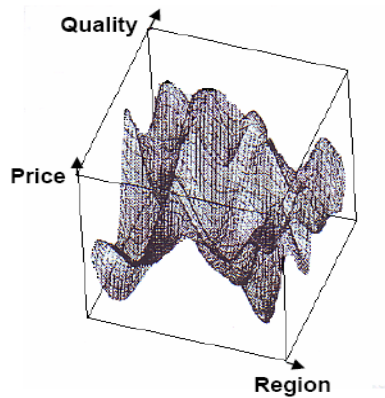


Fig. 4. Example of a space of solutions

In fig. 5 a subset of feasible solutions is chosen without ISS designed outside Europe. The space of feasible solutions is to the left of the separating surface that is depicted on the figure in blue color.

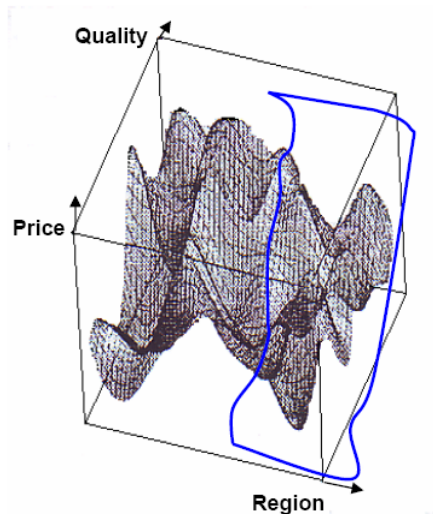


Fig. 5. Nonlinear space division(s) of the region

In fig. 6 another surface in green is shown delimiting the search space of the solutions. In our case it is 'systems with unknown principles of operation'. It is accepted that in the data bases there is no clear distinction related to the presented criteria so the search of the feasible solutions is nonlinear of high dimensionality and practically it cannot be solved using traditional methods. Nevertheless, by applying ontologies analogous to the ones from the previous section the problem is solvable via the PUZZLE method. There are two red dots on the same fig. 6 in its left corner. Each of them is a kind of constraint but of another type which we name a binding constraint and it is introduced by us. Its semantics is the following: it is not a solution but it resides close to the searched solution. For example we have the information that Fensel's elaborations are a good solution to the problem and that they define the left dot; the right dot has semantics of some other type. By introducing new constraints, our goal is to show that it is possible to use causal links that are different from implications.

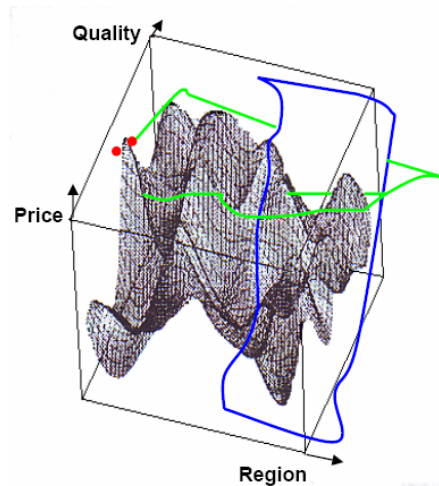


Fig. 6. Binding and other constraints

The same situation is presented on fig. 7 but some of the solutions are absent and this is evident in comparison to the images from the previous figures. It will be demonstrated below that the pointed incompletenesses are often met often and, even in this situation which is an obstacle for other existing methods, we offer an effective solution.

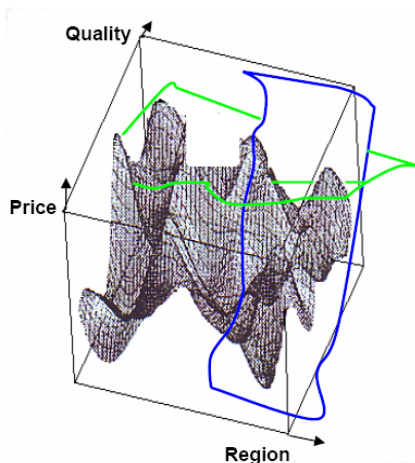


Fig. 7. Two nonlinear intersections best executed by using ontologies

---

## 2.2. Funnel Method

---

Below, we discuss in brief the next proposed FUNNEL method. Fig. 8 presents the main elements of the method: a system of constraints in the form of a funnel around a center which is a goal (fitness) function which points to the desired direction for information output or to search for new knowledge. As it is evident, the goal of this method is the gradual narrowing of the space of the feasible solutions, together with the progress of the dynamic information processes. Usually the FUNNEL method operates properly when combined with the other methods introduced here and that is why its peculiarities are viewed in detail in the next section where the interactions between the methods are examined. For example, it is convenient to concentrate on fig. 7 shown above over one of the peak values of the diagram by fixing a funnel above it.

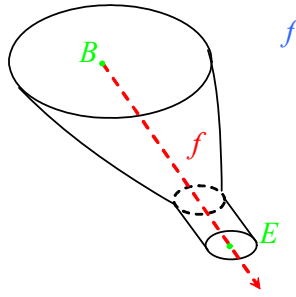


Fig. 8. Funnel system of constraints

### 2.3. Conflict Resolution Method and its Machine Learning Applications

Any lack of collaboration in a group of agents or intrusion could be found as an information conflict with existing models. Many methods exist where a model is given and every non-matching it knowledge is assumed as contradictory. Let's say, in an anomaly intrusion detection system, if the traffic has been increased, it is a contradiction to the existing statistical data and an intrusion alert has been issued. The considered approach is to discover and trace different logical connections to reveal and resolve conflict information. The constant inconsistency resolution process gradually improves the system DB and KB, and leads to better intrusion detection and prevention. Models for conflicts are introduced and used, and they represent different forms of ontologies.

Let the strong (classical) negation be denoted by '¬' and the weak (conditional, paraconsistent) negation [6,7,8] be '∼'. In the case of an evident conflict (inconsistency) between the knowledge and its ultimate form—the contradiction—the conflict situation is determined by the direct comparison of the two statements (the *conflicting sides*) that differ one from another by just a definite number of symbols '¬' or '∼'. For example: A and ¬A; B and not B (using ¬ equivalent to 'not'), etc.

In the case of implicit (or hidden) negation between two statements, A and B can be recognized only by an analysis of preset models of the type of (2).

$$\{U\}[\eta: A, B] \quad (2)$$

where  $\eta$  is a type of negation, U is a statement with a validity including the validities of the concepts A and B, and it is possible that more than two conflicting sides may be present. It is accepted below that the contents in the figure in brackets U is called *an unifying feature*. In this way, it is possible to formalize not only the features that separate the conflicting sides but also the unifying concepts joining the sides. For example, the intelligent detection may be either automated or of a human-machine type but the conflict cannot be recognized without the investigation of the following model.

$$\{\text{detection procedures}\}[\neg: \text{automatic, interactive}].$$

The formula (1) formalizes a model of the conflict the sides of which unconditionally negate each another. In the majority of the situations, the sides participate in the conflict only under definite conditions:  $\chi_1, \chi_2, \dots, \chi_z$ .

$$\{U\}[\eta: A_1, A_2, \dots, A_p] \langle \chi_1 \sim^* \chi_2 \sim^* \dots \sim^* \chi_z \rangle. \quad (3)$$

where  $\chi_1 \sim$  is a literal of  $\chi$ , i.e.  $\chi_1 \sim \equiv \chi$  or  $\chi_1 \sim \equiv \neg \chi$ , \* is the logical operation of conjunction, disjunction or implication.



The present research allows a transition from models of contradictions to ontologies [9] in order to develop new methods for revealing and resolving contradictions, and also to expand the basis for cooperation with the Semantic Web community and with other research groups. This is the way to consider the suggested models from (2) or (3) as one of the forms of static ontologies.

The following factors have been investigated:

T – time factor: non-simultaneous events do not bear a contradiction.

M – place factor: events that have taken place not at the same place, do not bear a contradiction. In this case, the concept of place may be expanded up to a coincidence or to differences in possible worlds.

N – a disproportion of concepts emits a contradiction. For example, if one of the parts of the contradiction is a small object and the investigated object is very large, then and only then it is the case of a contradiction.

O – identical object. If the parts of the contradiction are referred to different objects, then there is no contradiction.

P – the feature should be the same. If the parts of the contradiction are referred to different features, then there is no contradiction.

S – simplification factor. If the logic of user actions is executed in a sophisticated manner, then there is a contradiction.

W – mode factor. For example, if the algorithms are applied in different modes, then there is no contradiction.

MO – contradiction to the model. The contradiction exists if and only if (*iff*) at least one of the measured parameters does not correspond to the meaning from the model. For example, the traffic is bigger than the maximal value from the model.

Example. We must isolate errors that are done due to lack of attention from tendentious faults. In this case we introduce the following model (4):

$$\{ \text{user : faults } \} [ \sim : \text{accidental, tendentious} ] \langle T, \neg M, O; \neg S \rangle \quad (4)$$

It is possible that the same person does sometimes accidental errors and in other cases tendentious faults; these failures must not be simultaneous on different places and must not be done by same person. On the other hand, if there are multiple errors (e.g. more than three) in short intervals of time (e.g. 10 minutes), for example, during authentications or in various subprograms of the security software, then we have a case of a violation, nor a series of accidental errors. In this way, it is possible to apply comparisons, juxtapositions and other logical operations to form security policies thereof.

Recently we shifted conflict or contradiction models with ontologies that give us the possibility to apply new resolution methods. For pity, the common game theoretic form of conflict detection and resolution is usually heuristic-driven and too complex. We concentrate on the ultimate conflict resolution forms using contradictions. For the sake of brevity, the resolution groups of methods are described schematically.

The conflict recognition is followed by its resolution. The schemes of different groups of resolution methods have been presented in Fig. 9 to Fig. 12.

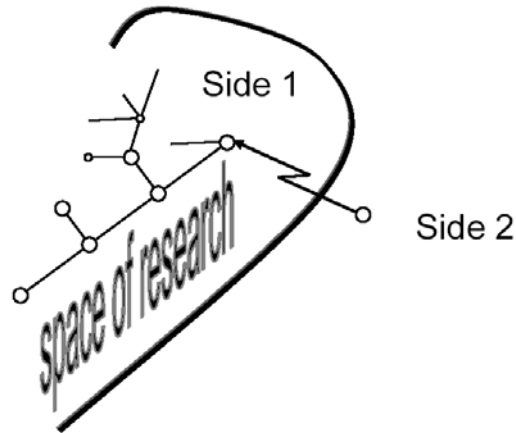


Fig. 9. Avoidable (postponed) conflicts when Side 2 is outside of the research space.

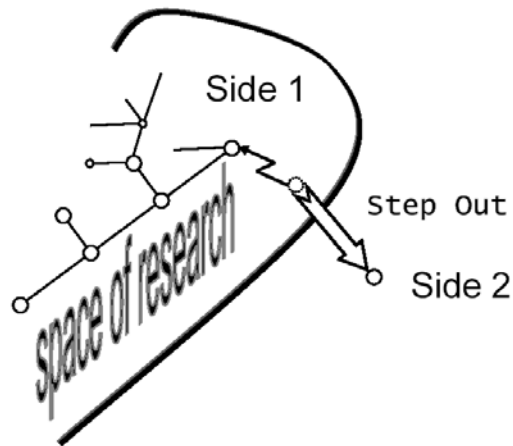


Fig. 10. Conflict resolution by stepping out of the research space (postponed or resolved conflicts).

In situations from Fig. 9, one of the conflicting sides does not belong to the considered research space. Hence, the conflict may not be immediately resolved, only a conflict warning is to be issued in the future. Let's say, if we are looking for an intrusion attack, and side 2 matches printing problems, then the system could avoid the resolution of this problem. This conflict is not necessary to be resolved automatically, experts may resolve it later using the saved information. In Fig. 10, a situation is depicted where the conflict is resolvable by stepping out from the conflict area. This type of resolution is frequently used in multi-agent systems where conflicting sides step back to the pre-conflict positions and one or both try to avoid the conflict area. In this case a warning on the conflict situation has been issued.

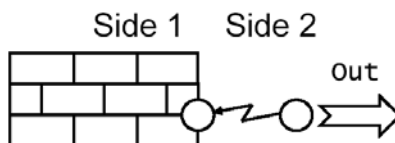


Fig. 11. Automatically resolvable conflicts.

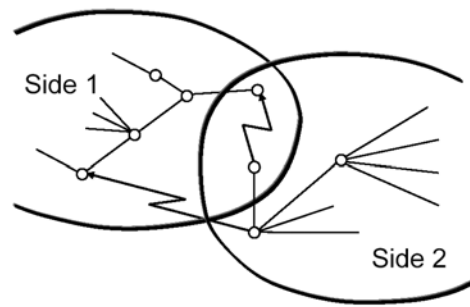


Fig. 12. Conflicts resolvable using human-machine interaction.

The situation from Fig. 11 is automatically resolvable without issuing a warning message. Both sides have different priorities, say side 1 is introduced by a security expert, and side 2 is introduced by a non-specialist. In this case, side 2 has been removed immediately. A situation is depicted on Fig. 12 where both sides have been derived by an inference machine, say by using deduction. In this case, the origin for the conflict could be traced, and the process is using different human-machine interaction methods.

Knowledge bases (KBs) are improved after isolating and resolving contradictions in the following way. One set is replaced by another while other knowledge is supplemented or specified. The indicated processes are not directed by the elaborator or by the user. The system functions autonomously and it requires only a preliminary input of models and the periodical updates of strategies for resolving contradictions. Competitions to the stated method may be methods for machine supervised – or unsupervised – learning. During supervised learning, for example by using artificial neural networks, training is a long, complicated, and expensive process, and the results from the applications outside the investigated matter are unreliable. The 'blind' reproduction of teacher's actions is not effective and it has no good prospects except in cases when it is combined with other unsupervised methods. In cases of unsupervised training via artificial neural networks the system is overloaded by heuristic information and algorithms for processing heuristics, and it cannot be treated as autonomous. The presented method contains autonomous unsupervised learning based on the doubt-about-everything principle or on the doubt-about-a-subset-of-knowledge principle. The contradiction-detecting procedure can be resident; it is convenient to use computer resources except for peak hours of operation.

The unsupervised procedure consists of three basic steps. During the first step, the contradiction is detected using models from (2) to (4). During the second step, the contradiction is resolved using one of the resolution schemes presented above, depending on the type of conflict situation. As a result from the undertaken actions, after the second stage the set  $K$  is transformed into  $K'$  where it is possible to eliminate from  $K$  the subset of incorrect knowledge  $W \subseteq K$ , to correct the subset of knowledge with an incomplete description of the object domain  $I \subseteq K$ , to add a subset of new knowledge for specification  $U \subseteq K$ . The latter of cited subsets includes postponed problems, knowledge with a possible discrepancy of the expert estimates (problematic knowledge), and other knowledge for future research which is detected based on the heuristic information.

In cases of ontologies, metaknowledge or other sophisticated forms of management strategies, the elimination of knowledge and the completion of KBs becomes a non-trivial problem. For this reason the concepts of orchestration and choreography of ontologies are introduced in the Semantic Web and especially for WSMO [10,11]. The elimination of at least one of the relations inside the knowledge can lead to discrepancies in one or in several subsets of knowledge in  $K$ . That is why after the presented second stage, and on the third stage, a check-up of relations is performed including elimination of modified knowledge and the new knowledge from subsets  $W$ ,  $N$ ,  $I$ ,  $U$  are tested for non-discrepancies via an above described procedure. After the successful finish of the process a new set of knowledge  $K'$  is formed that is more qualitative than that in  $K$ ; according to this criterion it is

a result from a machine unsupervised learning managed by models of contradictions defined a priori and by the managing strategies with or without the use of metaknowledge.

---

### 3. Applications

---

The presented system source codes are written in different languages: C++, VB, and Prolog. It is convenient to use the applications in freeware like RDF, OWL, Ontoclean or Protégé. Many of the described procedures rely on the usage of different models/ ontologies in addition to the domain knowledge thus the latter are metaknowledge forms. In knowledge-poor environment the human-machine interactions have a great role, and the metaknowledge helps make the dialog more effective and less boring to the human. The dialog forms are divided in 5 categories from 1='informative' to 5='silent' system. Knowledge and metaknowledge fusions are always documented: where the knowledge comes from, etc. This is the main presented principle: every part of knowledge is useful and if the system is well organized, it will help us resolve some difficult situations.

We rely on nonsymmetrical reply 'surprise and win',

---

### 4. Conclusion

---

The main conclusion is that to overcome the shortcomings, methods and applications are considered concerning the logical parts of knowledge discovery and data mining. Special attention is paid to methods for identification and resolution of conflicts, and to machine (self-) learning based on them. The role of the above methods for the security bots and agents is discussed.

Analysis is represented for technologies used for machine learning in intelligent agents, and for sending information by sense, and for understanding the semantics of the information. Common disadvantages for different existing groups of contemporary applications are revealed.

Same methods in different combinations are effectively used to enhance security administrator possibilities or in contemporary e-learning systems in the field of Information/National Security [12]. Applications outside the field of information security have been made since a long time, but their explanation goes out of the field of the considered research.

---

### Acknowledgement

---

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA ( [www.ithea.org](http://www.ithea.org) ) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine ( [www.aduis.com.ua](http://www.aduis.com.ua) ).

---

### References

---

- B. Йоцов. Сигурност и защита на информацията. - София: За буквите - О писменехъ, 2006 (V. Jotsov. Information Security Systems. Sofia: Za bukвите-o pismeneh, 2006, 156 p).
- B. Thuraisingham. Data Mining Technologies, Techniques, Instruments and Trends. NY etc.: CRC Press, 1999.
- D. Dasgupta. Artificial Immune Systems. Moscow: Fizmatlit, 2006.
- V. Jotsov. Novel Intrusion Prevention and Detection Systems. In: Proc. 4th International IEEE Conference on Intelligent Systems, Yager R., Sgurev V and Jotsov V. (Eds.), Vol. II, Varna, Bulgaria, September 6-8, 2008, pp. 14.20-14.27.
- V. Jotsov. "Evolutionary parallels," In: Proc. First Int. IEEE Symp. 'Intelligent Systems', T. Samad and V. Sgurev (Eds.), Varna, Bulgaria, vol. 1, pp. 194-201, 2002.

- 
- A. Arruda, "A survey on paraconsistent logic," in: Math. Logic in Latin America, A. Arruda, C. Chiaqui, N. Da Costa, Eds. North-Holland, Berlin NY, pp. 1-41, 1982.
- V. Jotsov. Semantic Conflict Resolution Using Ontologies, Proc. 2nd Intl. Conference on System Analysis and Information Technologies, SAIT 2007, RAS, Obninsk, September 11-14, 2007, vol. 1, pp. 83-88.
- V. Jotsov. "Knowledge acquisition during the integer models investigation," Proc. XXXV Int.Conf. "Communication, Electronic and Computer Systems", Technical University of Sofia, pp. 125-130, 2000.
- T. Gruber. A translation approach to portable ontologies. Knowledge Acquisition, Vol. 5, No. 2, pp.199-220, 1993.
- Web Service Modeling Ontology (WSMO). <http://www.wsmo.org/TR> to date.
- D. Fensel. Ontologies: A Silver Bullet for Knowledge Management and Electronic Systems. Berlin Heidelberg New York: Springer-Verlag, 2004.
- V. Jotsov. Emotion-Aware Education and Research Systems. J. Issues in Informing Science and Information Technologies, USA, vol. 6, pp. 779-794, 2009 (online version available).

---

### Author information

---

**Vladimir S. Jotsov (В.С. Йоцов):** e-mail: [bgimcssmc@gmail.com](mailto:bgimcssmc@gmail.com)

State University of Library Studies and Information Technologies (SULSIT);

Institute of Information Technologies of the Bulgarian Academy of Sciences (IIT-BAS); P. O. Box 161, Sofia 1113, BULGARIA, EU;

## NUMERIC-LINGUAL DISTINGUISHING FEATURES OF SCIENTIFIC DOCUMENTS

Vladimir Lovitskii, Ina Markova, Krassimir Markov, Ilia Mitov

**Abstract:** *The classification of scientific papers is based on the ability of the artificial system (let's call such a system ARSA i.e. Automated Review of Scientific Articles) to reflect the similarity of different scientific papers and differential of similar papers. To identify the text as similar to and different from other texts a set of characteristics needs to be used. In this paper the approach of the extraction of "linguistic items" from scientific paper that provides representative information about the document content is considered.*

**Keywords:** *text mining, word's properties, text pattern.*

**ACM Classification Keywords:** *1.2 Artificial intelligence: 1.2.7 Natural Language Processing: Text analysis.*

---

### Introduction

A large number of texts can be retrieved from the Internet for research purposes through the use of search engines. This overload of textual materials poses new methodological challenges in text analysis. How can one automate the analysis of large amounts of texts that can no longer be analyzed qualitatively or coded manually, and still obtain conceptually meaningful and valid results? Several research traditions, such as computer-aided content analysis, corpus-based linguistics, and the so-called 'sociology of translation' [Callon et al., 1986; Stegman & Grohmann, 2003; Lovitskii et al., 2007] have developed tools for the automated analysis of texts. The main general task of these tools is the extraction of the "linguistic items" from unconstrained text that provides representative information about the document content. However, none of these guarantee the "best solution" for this task. Despite the different disciplinary backgrounds and research agendas of these traditions, they have all faced similar problems with the ambiguity of language. Words and the relations among words mean different things in other contexts, and the meaning of words can be expected to change, particularly in science.

Natural Language Environment (NLE) is so complicated that at the present time (from our point of view) it is simply impossible to create an artificial system which provides proper natural language processing of any kind of text. Traditional search mechanisms focusing on statistics (i.e., the frequency of keywords) provide imperfect results: the keyword may be misspelled in some target documents; it may appear in a plural or conjugated form; it may be replaced by a synonym; it may have different meanings according to context. In such cases traditional searches will typically return results that prove either too voluminous or too restricted to be helpful. That is why we restrict the NLE to quite short scientific papers (SP).

Natural language has a very rich expressive power and even at the lexical level, the large variability due to synonymy and homonymy causes serious problems to retrieval methods based on keyword matching. Synonym means that the same concept can be expressed using different sets of terms (terms mean the lexical items and may consist of words as well as expressions). Below some example of synonyms are shown:

- *Get rid of a cursor;*
- *Delete a cursor;*
- *Remove a cursor from the screen;*
- *Eliminate a cursor;*
- *Erase a cursor;*
- *Makes a cursor hidden;*
- *Set the cursor size to 0;*
- *Take away a cursor from the screen.*

Homonym means that identical terms can be used in very different semantic contexts. For example,

- *The season of growth;*
- *A natural flow of ground water;*
- *Jump: move forward by leaps and bounds;*
- *A metal elastic device that returns to its shape or position when pushed or pulled or pressed.*

Overall, synonymy and homonymy lead to a complex relation between terms and concepts that cannot be captured through simple matching. Restriction of the NLE by SP allows us to minimize this problem.

In the NLE the mathematical symbolic can be used to describe some ideas but to prove that these ideas are working as expected they need to be computerized. That is why ARSA has been developed. In this paper we will discuss in some detail techniques for analyzing the textual content of SP. Although implementation of ARSA as a Web application and using SP represented as a PDF files appears to be a straightforward problem, in many practical situations the task can actually be quite challenging. The processing of PDF files can be difficult to handle because these are not data formats but algorithmic descriptions of how the document should be rendered. The general objective of this paper is to describe the steps of SP processing and discuss the results of such processing. The initial step of SP processing is obvious: The SP is downloaded from the Internet and saved in PDF format. Then the PDF file was parsed to be represented as a separate text file. This text content of SP was then broken down into sentences and words. The next steps of SP processing will be considered by details:

- **Initial text conversion to a “skeleton”;**
- **The properties of initial text calculation;**
- **Keywords (KW) extraction from the whole text.** Very often (e.g. in a neural network) the huge indexes are not appropriate but only a few KWs of the document should be stored in a database. How many KWs should be extracted from the document? What is the criterion for it? We will answer these questions;
- The properties of KWs calculation;
- **The pattern of SP creation.** Text properties and KWs properties are used to create the SP pattern (SPP). SPP will be used to measure a *similarity* between the different SP and *differential* between similar SP. The result of SP processing is shown on Figure 1.
- **Comparison analysis of SP.** The measurement of *differential* between SP from the same scientific domain is discussed.

---

## Text to Skeleton Conversion

---

Text to skeleton conversion is a part of syntactic simplification of SP. Syntactic simplification is the process of reducing the grammatical complexity of a text, while retaining its information content and meaning. The aim of syntactic simplification is to make text easier to process by programs. ARSA takes the SP as a character sequence, locates the sentence boundaries, and converts the original SP to a *skeleton*. Such conversion will require several steps:

- Noisy (non-searchable) word elimination;
- Irregular verb normalisation. Once the word has been identified then it should be changed back to its simplest form for efficient word recognition. For example, *writes, writing, wrote, written* will be changed to *write* and the corresponding attributes of the original form will be saved;
- Initial word to root form conversion.

The first step is the removal of “noisy words” (or stop words), i.e. common words such as articles, prepositions, and adverbs that are not informative about the semantic content of a document [Fox 1992]. Since noisy words are very common, removing them from the text can also significantly help in reducing the size of the initial text. In

practice, noisy words may account for a large percentage of text, up to 20–30%. Naturally, removal of noisy words also improves computational efficiency during retrieval.

To reduce all the morphological variants of a give word to a single term. For example, a SP might contain several occurrences of words like *fish*, *fishes*, and *fishers* but would not be retrieved by a query with the keyword *fishing* if the term *fishing* never occurs in the text. That is why all words should be converted to their root form (such as *fish* in our example).

The screenshot shows the ARSA web application interface. The main content area displays the results of SP processing for a document. The document title is "http://www.foibg.com/jita/vol15/jita15.4-p04.pdf". The interface is divided into several sections:

- Document Statistics:**
  - Size of Initial Document (Dsz) = 27,421 characters
  - Number of Sentences in a Document (Sd) = 180
  - Size of Skeleton (Ssz) = 10,428 characters
  - Percentage of Document "Noisy":  $(Dsz - Ssz) / Dsz * 100 = 61.97\%$
  - Number of Words in a Skeleton (Sw) = 1,687
  - Number of Distinct Words in a Skeleton (Dw) = 327
  - % of Words Repeatability:  $(1 - Dw / Sw) * 100 + 1 = 81.61\%$
  - Number of Words for 100% of Doc Covering (KW<sub>100</sub>) = 20 (100%) - 3 (75%) - 2 (50%)
  - Centering of text content (C<sub>cnt</sub>) = 55.0%
  - Number of Sentences Covered Independently by KW (Si) = 403
  - Text Cohesion = 1.90
- Skeleton:** ontology application category knowledge management system control vocabulary web site document organization navigation browse search semantic generalization specialization sense disambiguation consistency check restriction auto completion interoperability informatiprocess integration validation verification configuration structured comparative
- Pattern:** Cohesion=1.90 Centering=55.0% Recurrence=81.61% KW No: 20-3-2 KWontology (75-75), domain (29-50), set (35-50), knowledge (9-42), property (3-23), devalue (8), property (7), section (5), answer (3), meaning (2), concept (2), element (1), object (1), process (1)
- 20 Keywords Statistics:**

Words	Frequency	Percentage	Sntucs %	Cohesion
ontology	75	22.93%	41.66%	1.52
domain	50	15.29%	27.77%	1.70
set	50	15.29%	27.77%	1.92
knowledge	42	12.84%	23.33%	1.90
property	23	7.03%	12.77%	2.47
devalue	22	6.72%	12.22%	2.59
information	20	6.11%	11.11%	1.85
element	20	6.11%	11.11%	2.40
object	19	5.81%	10.55%	2.21
process	18	5.50%	10.00%	1.61
system	15	4.58%	8.33%	2.00

Figure 1. The result of SP processing

### Properties of SP

Properties of the SP are used for the documents initial filtering. We will distinguish eleven properties:

- Size of initial document (Dsz) = 27,421 characters;
- Number of sentences in a Document (Sd) = 180;
- Size of the Skeleton (Ssz) = 10,428 characters;
- Percentage of document "noisy":  $(Dsz - Ssz) / Dsz * 100 = 61.97\%$ ;
- Number of words in the Skeleton (Sw) = 1,687;
- Number of distinct words in the Skeleton (Dw) = 327;
- Percentage of words repeatability:  $(1 - Dw / Sw) * 100 + 1 = 81.61\%$ ;
- Number of Words for 100% of Doc Covering (KW<sub>100</sub>) = 20 (100%) - 3 (75%) - 2 (50%);
- Centering of text content (C<sub>cnt</sub>) = 55%;
- Number of Sentences Covered Independently by KW (Si) = 403;
- Text Cohesion = 1.9.



Let us clarify and explain the meaning of some terminology used to describe SP properties:

- **Distinct word** is readily distinguishable from all others;
- Counting repeated word usage allows us to identify important sentences and then use this information for meta-analysis of the document;
- From a linguistic point of view a keyword is a word which occurs in a text more often than we would expect to occur by chance alone. We offer some criterion to restrict number of KW. **This is the criterion for sufficient amount of KW to cover 100% (KW<sub>100</sub>) sentences of SP (SSP)**. The algorithm of KW extracting is quite simple:
- All distinct words are sorted in descending order in accordance with their frequencies. The frequency of each word equal number of sentences where word occurs at least one time.
- In SSP the sentences which include the first word from the sorted list are selected.
- The selected sentences are excluded from the SSP.
- For the next words the same procedure is repeated until the SSP is empty.
- Number of sentences covered independently by each KW equals the KW frequency. The result of KW extraction is that we have not only the list of KW but three very important numbers: **20(100%)-3(75%)-2(50%)** (i.e. KW<sub>100</sub>=20, KW<sub>75</sub>=3 and KW<sub>50</sub>=2) which allows us to describe the **centering** of SP content (**C<sub>cnt</sub>**) [Grosz et al., 1995]. The **C<sub>cnt</sub>** is the need to formalise a notion of connectedness in text in order to explain why one SP appears intuitively to be more connected and coherent than another despite both SPs being from the same scientific area. The heuristic rule to define the **C<sub>cnt</sub>** is:

$$C_{cnt} = (Sd / KW_{100} + 0.75 * Sd / KW_{75} + 0.5 * Sd / KW_{50}) / Sd * 100\% = 67.5\%,$$

where Sd = 180 (see Figure 1). For comparison let's consider another SP where Sd = 227, KW<sub>100</sub>=81, KW<sub>75</sub>=28 and KW<sub>50</sub>=10 (<http://www.foibg.com/ijita/vol15/ijita15-2-p07.pdf>). For this SP **C<sub>cnt</sub> = 8.91%**.

- Cohesiveness is an essential requirement for SP to be useful. Document cohesion (**D<sub>chsn</sub>**) is defined by Halliday and Hasan [Halliday and Hasan, 1976] as the phenomenon where the interpretation of some element of text depends on the interpretation of another element and the presupposing element cannot be effectively decoded without recourse to the presupposed element. Let us consider some artificial SP to explain the idea of cohesion calculation. Suppose in this SP there are 100 sentences i.e. Sd=100. The first KW with the maximum frequencies covers 80 sentences and the other 20 sentences are covered by second KW i.e. KW<sub>100</sub>=2. Now we have to define how many sentences can cover each KW from the full set of sentences which equals 100. If the second KW will cover the same number of sentences i.e. 20 then total number of sentences covered by KWs independently will equal 100 i.e. S<sub>ind</sub> = 100 and **D<sub>chsn</sub> = (S<sub>ind</sub> - Sd) / KW<sub>100</sub> = 0**. But if the second KW independently covers 60 sentences (i.e. S<sub>ind</sub> = 140) then **D<sub>chsn</sub> = 20**. This calculation of text cohesion is given just to demonstrate the idea. The proper calculation of KW and SP cohesion will be considered below in detail.

It is also appropriate to note that so far there is not much data on the features of any body of text that might serve as a **standard for comparison**, much less the detailed studies of the characteristics of a variety of texts that will be essential to ensure continuing progress in this field. A systematic study of different types of texts, and of the purposes for which they can be analyzed, would provide useful guidelines for research at this stage of our understanding. We hope that the some SP properties might be considered as standard characteristics of any text.

### Properties of Keywords

The method for the identification of words as keywords is based on the technique of word properties calculation. In the previous section of SP properties the definition of KW has been given and importance of KW associations and their co-occurrence has been considered implicitly as SP cohesion. Here we want to discuss in detail some

KW properties such as KW co-occurrence neighbourhoods which are critical to define SPs similarity. KW properties are shown in Figure 2.

Words	Frequency	Percentage	Sntucs %	Cohesion
ontology	75	22.93%	41.66%	1.52
domain	50	15.29%	27.77%	1.70
set	50	15.29%	27.77%	1.92
knowledge	42	12.84%	23.33%	1.90
property	23	7.03%	12.77%	2.47
devalue	22	6.72%	12.22%	2.59
information	20	6.11%	11.11%	1.85
element	20	6.11%	11.11%	2.40
object	19	5.81%	10.55%	2.21
process	18	5.50%	10.00%	1.61
system	15	4.58%	8.33%	2.00

Figure 2. Keywords properties

The set of KW created during SP processing is ordered by frequency. The percentage of KW frequencies allows us to compare different SP with different KW distribution, e.g. in SP (<http://www.foibg.com/ijita/vol11/ijita11-2-p02.pdf>) the frequency of KW "system" equals 24 and percentage = 4.61% (compare with KW "system" from Figure 2)..

### Keywords Neighbourhood

The idea of adjacent words is based on the assumption that with the successive presentation of a number of words the strongest relation is the relation between the nearest neighbour words. *"Their succeeding one after another presents evidently an important condition of structuring"* [Hoffmann, 1982, p.231].

The study of word co-occurrence in a text is based on the cliché that *"one (a word) is known by the company one keeps"*. We hold that it also makes a difference *where* that company is kept: since a word may occur with different sets of words in different contexts, we construct word neighbourhoods for each SP and also word neighbourhoods which depend on the text of enquiry (or abstract of SP). Word associations have been studied for some time in the fields of psycholinguistics (by testing human subjects on words) [Lovitsky, 1983; Lovitskii et al, 1997], linguistics (where meaning is often based on how words co-occur with each other), and more recently, by researchers in natural language processing using statistical measures to identify sets of associated words for use in various natural language processing tasks. One of the tasks where the statistical data on associated words has been used with some success is lexical disambiguation.

Words which co-occur frequently with the given word may be thought of as forming a "neighbourhood" of that word. SPs may contain words that are associated with many different senses. For example, in one SP the word "bank" could co-occur frequently with such words as "money", "loan", and "robber", while in another SP the word "bank" would be more frequently associated with "river", "bridge", and "earth". Despite the fact that the word "bank" has the highest frequency in both SPs it is obvious that these two SPs do not have any similarity.

Individual words in different SPs have more or less differing contexts around them. Semantic similarity of words depends on similarity of their contexts. Words found in similar contexts tend to be semantically similar. Such measures have traditionally been referred to as measures of distributional similarity. If two words have many co-occurring words, then similar things are being said about both of them and therefore they are likely to be semantically similar. Therefore if two words are semantically similar then they are likely to be used in a similar fashion in text and thus end up with many common co-occurrences. For example, the semantically similar “car” and “vehicle” are expected to have a number of common co-occurring words such as “parking”, “garage”, “accident”, “traffic”, and so on.

### Keywords Cohesion

For considered SP (see Figures 1 and 2) 20 KW have been extracted, namely: “ontology”(75-75), “domain”(29-50), “set”(35-50), “knowledge”(9-42), “property”(3-23), “devalue”(3-22), “information”(3-20), “element”(2-20), “object”(1-19), “process”(6-18), “system”(1-15), “section”(3-15), “constraint”(1-14), “meaning”(1-6), “answer”(3-6), “concept”(1-4), “webont”(1-1), “ibm”(1-1), “difference”(1-1), “catalyst”(1-1). After each KW a pair of numbers is placed. The right number means frequencies of KW, or number of sentences where the KW occurs at least one time. The left number was used to select KW from the list of distinct words and also means number of sentences where the word occurs at least one time, but occurrence is calculated in the set of sentences which are left after the subset of sentences covered by the previous word has been excluded. For example, word “ontology” occurred in the 75 sentences from 180, word “domain” occurred in the 29 sentences from 105 (because 75 sentences have been excluded), word “set” occurred in the 35 sentences from 76 etc.

**It is important to mention that we can not maintain that our algorithm, when KW have been selected from the descending ordered list of distinct words, gives us the best solution i.e. the minimum number of KW. Moreover, we understand that the minimum number of KW does not represent the best solution.** Let’s consider some examples for explanation. Suppose we have SP with 100 sentences i.e.  $S_d=100$  and four distinct words (dw) ordered by frequency i.e.  $F(dw_1)=52$ ,  $F(dw_2)=48$ ,  $F(dw_3)=46$  and  $F(dw_4)=33$ . Suppose that the first two distinct words cover 100% sentences of SP. In accordance with our algorithm we have that  $KW_1(52-52)=dw_1$  and  $KW_2(48-48)=dw_2$  but co-occurrence of pair  $KW_1$ - $KW_2$  equals 0 i.e.  $Co(KW_1, KW_2)=0$  and indicate that the current SP does not have any cohesion. But, at the same time, if the three KW have been represented by the sequence of distinct words:  $dw_1$ ,  $dw_3$  and  $dw_4$  i.e.  $KW_1(52-52)=dw_1$ ,  $KW_2(27-48)=dw_3$  and  $KW_3(21-33)=dw_4$ , even  $Co(KW_1, KW_2)=21$  shows quite a high level of cohesion of the same SP.

We can offer a very challenging mathematical problem: **“In what sequence should the distinct words be used to provide the minimum number of KW?”** and **“What is the criteria of the best solution?”**. It is easy to become convinced that there are several different solutions e.g. if the word “set” (but not the word “domain”) is used immediately after the word “ontology” the number of covered sentences will be equal 39 and not 35 despite the words “domain” and “set” having the same frequencies. There is an even more convincing example. The word “property” covers just 3 sentences from  $32=180 - (75+29+35+9)$ , at the same time when the word “process” covers 6 sentences from 20 despite this the frequency is 18 in comparison with frequency of 23 of “property”.

**Let us call  $KW_{CW}$  the context word (or neighbour) of KW focus ( $KW_{FCS}$ ) if they occur together within the same sentence boundaries.** Algorithm of  $KW_{CW}$  searching is quite simple. Each KW is considered consecutively as a  $KW_{FCS}$ . The occurrence of the rest of the KW among sentences covered by  $KW_{FCS}$  is checked. The number of sentences where the KW occurred is counted. For example, for  $KW_{FCS} = \text{“ontology”}$ , among 75 sentences ARSA found 15  $KW_{CW}$ , namely: “knowledge”(22), “domain”(21), “set”(11), “system”(11), “constraint”(10), “information”(9), “devalue”(8), “property”(7), “section”(5), “answer”(3), “meaning”(2), “concept”(2), “element”(1), “object”(1), “process”(1). The number in brackets shows number of sentences from 75 where current KW co-

occurs together with  $KW_{FCS}$ . It is easy to explain why, for example,  $KW_{CW}$  "domain" co-occurred in 21 sentences.  $KW$  "domain"(29-50) has been selected immediately after the first  $KW$  "ontology"(75-75). 29 sentences from 105=180-75 have been covered by word "domain". Why was it 29 but not 50? Because the word "domain" occurred in 21 = 50 - 29 sentences from 75 which have been excluded from the initial set of sentences.

The total number of sentences where  $KW_{CW}$  of  $KW_{FCS}$  = "ontology" co-occurred equals 114. **The cohesion of  $KW$  ( $KW_{chsn}$ ) "ontology" equals  $114/75=1.52$**  (see Figure 2). If  $KW_{chsn} = 1$  it means that on average in each sentence covered by  $KW_{FCS}$  at least one  $KW$  co-occurs, or two  $KWs$  if  $KW_{chsn} = 2$ . The  $D_{chsn}$  depends on  $KW_{chsn}$  and calculated as a sum of  $KW_{chsn}$  divided by number of  $KW$  for which  $KW_{chsn} > 0$ . For considered SP  $D_{chsn} = 1.9$ . For  $KWs$  "webont"(1-1), "ibm"(1-1), "difference"(1-1), "catalyst"(1-1)  $KW_{chsn} = 0$ .

---

### Pattern of the document

---

The pattern of SP is used to provide a similarity measurement between the different documents or between the enquiry (extract of a SP might be considered as a kind of enquiry) and a SP. We will distinguish two parts of the pattern: document properties (DP) and  $KW$  properties (KWP). DP will be used for meta-analysis of SP and KWP – for the direct measurement of document similarity. Information science research has focused on how the measurement of meaning can be operationalised using words and their co-occurrences. It was proposed [Salton and McGill, 1983] to use the cosine between word vectors as providing a spatial representation of how words are positioned in relation to other words. But our belief is that before starting to use traditional methods to measure the similarity of two SP they should first be classified. We think that the automation of the preliminary selection and classification of the SP might improve the similarity measurement.

Therefore meta-analysis of SP will be classified. **Our idea is that for SP, which belong to different classes, a different algorithm for measurement of similarity should be used.** We will discuss in detail the process of SP meta-analysis in our next paper. Here we simply explained the general idea. In the considered SP just 3  $KW$  are required to cover 75% of sentences and the centering of this SP equals 55% whereas the SP (<http://www.foibg.com/ijita/vol15/ijita15-2-p07.pdf>) of a similar size requires 28  $KW$  to cover 75% of sentences and its centering equals 8.91%. There should be a completely different algorithm to measure their similarity in comparison with the SP (<http://www.foibg.com/ijita/vol10/ijita10-1-p09.pdf>) where 4  $KWs$  are used to cover 75% of SP and the centering equals 49.3%.

---

### Conclusion

---

In this paper we described our vision of SP analysis. Implementation of our ideas as ARSA allows us to provide instant analysis of hundreds of SP; and due to this we can evaluate our original ideas. In the result of SP analysis both document and  $KW$  properties have been extracted. Our next step is to provide the measurement of the semantic similarity of two SP. Preliminary analysis of traditional methods for computing similarity measures allows us conclude that they should be modified in accordance with our ideas to provide more adequate similarity measures. The full potential of the automatic SP analysis presented here will be deployed when ARSA will be enlarged to incorporate automatic calculation of two SP semantic similarities.

---

### Acknowledgement

---

This work is partially financed by Bulgarian National Science Fund under the project D002 308/19.12.08.

---

## Bibliography

---

- Callon, M., J. Law, & A. Rip (Eds.), 1986. Mapping the Dynamics of Science and Technology. London: Macmillan.
- Christopher Fox, 1992. "Lexical Analysis and Stop Lists." *Information Retrieval: Data Structures and Algorithms*. William Frakes and R. Baeza-Yates (eds.), Prentice-Hall, 1992, 102-130.
- Barbara Grosz, Aravind Joshi, and Scott Weinstein, 1995. "Centering: A framework for modelling the local coherence of discourse". *Computational Linguistics*, 21(2):203–226.
- Michael A.K. Halliday and Ruqaiya Hasan, 1976. Cohesion in English. Longman Group Ltd, London, U.K.
- J.Hoffmann, 1982. Das Aktive Gedachtnis. Psychologische Experimente und Theorien zur Menschlichen Gedachtnistatigkeit, *VEB Deutscher Verlag der Wissenschaften*, Berlin.
- Vladimir Lovitskii, Michael Thrasher, David Traynor, 2007. "Automated Response To Query System", *Proc. of the XIII-th International Conference on Knowledge-Dialogue-Solution: KDS-2007*, Varna (Bulgaria), 534 – 543.
- V.A.Lovitsky, 1983. "A bionic approach to the realization of analytic and synthetic grammars in computational structures", *Proc. of the Symposium on Grammars of Analysis and Synthesis and their Representation in Computational Structures*, Tallin, 58-60.
- Salton, G., & M. J. McGill. 1983. Introduction to Modern Information Retrieval. Auckland, etc.: McGraw-Hill.
- Stegmann, J. & Grohmann, G., 2003. Hypothesis generation guided by co-word clustering. *Scientometrics* 56(1).

---

## Authors information

---



**Vladimir Lovitskii** – University of Plymouth, Plymouth, Devon, PL4 6DX, UK,

e-mail: [vladimir.lovitskii@fsmail.net](mailto:vladimir.lovitskii@fsmail.net).

Major Fields of Scientific Research: Artificial Intelligence



**Ina Markova** – Tester; Institute of Information Theories and Applications FOI ITHEA, P.O. Box: 775, Sofia-1090, Bulgaria; e-mail: [ina@foibg.com](mailto:ina@foibg.com)

Major Fields of Scientific Research: Information systems



**Krassimir Markov** – ITHEA ISS IJ, IBS and IRJ Editor in chief, P.O. Box: 775, Sofia-1090, Bulgaria; e-mail: [markov@foibg.com](mailto:markov@foibg.com)

Major Fields of Scientific Research: General theoretical information research, Multi-dimensional information systems



**Ilia Mitov** – Vice-president, Institute of Information Theories and Applications FOI ITHEA, P.O. Box: 775, Sofia-1090, Bulgaria; e-mail: [mitov@foibg.com](mailto:mitov@foibg.com)

Major Fields of Scientific Research: Business informatics, Software technologies, Multi-dimensional information systems

## DATA AND METADATA EXCHANGE REPOSITORY USING AGENTS IMPLEMENTATION

**Tetyana Shatovska, Iryna Kamenieva**

**Abstract:** *For implementation of an intelligent data and metadata exchange repository the intelligent agent oriented approach has been selected. In this work the conceptual structure and interaction principles of intelligent agents and ontological models in the intelligent data and metadata exchange repository will be offered. The main attention in this work will be paid to the development of intelligent search agent model realizing information extraction on ontological model of Data mining (DM) methods. In a client part of system there is considered the building of the intelligent agent of the repository user, the coordinator (manager) agent, which controls the common state of the system, and also fulfils the registration and authorizations of users, the resource (dataset) agent with partial usage of files structure with SDMX standard data. The model uses the service oriented architecture. Here is used the cross platform programming language Java, multi-agent platform Jadex, database server Oracle Spatial 10g, and also the development environment for ontological models - Protégé Version 3.4.*

**Keywords:** *repository, SDMX standart, data mining, semantic web, ontology, multiagent system, search algorithms, agent-oriented systems, intelligent agent, jadex, sdk, java, rdf, protégé, sparql, oracle splatlat.*

**ACM Classification Keywords:** *H.3.3 Information Search and Retrieval*

---

### Introduction

Digital repositories are networked software applications primarily used for storing, managing and disseminating data (e.g. digital publications, theses, data sets and so on). The Repositories differ from conventional content management systems because they include technologies to ensure that data are preserved for long-term access and use.

We are focused on developing multi-agent system for processing and storage of any statistical data. Research of existing repositories allowed identifying the main bottlenecks of the similar statistical repositories that were taken into account. Operating with UCI repository the user is able to filter, according to subsection of data mining area, the data files to view the brief characteristic of a file, to download a file. Using DEA Dataset Repository the user is able to search in any of criteria, view the brief characteristic of a file and to download a file, but only after .XML registration. In Data Repository the user can download any file of subjects without registration. Operating with Frequent Itemset Mining Dataset Repository the user does not need to be registered, he can obtain the information about researches made on samplings and the contact information of researchers, to download a file.

A key feature of the developed system via above mentioned typical statistical repositories is implementation of the datasets metadescription using the European standard SDMX 2.0 and ontological models that are stored in the system.

The advantage and novelty of the work is implementation of an ontological models set of the Data mining methods, which is used for the selection of a proper method under the sample source of the user datasets. To work with set of the ontological models have been developed a set of search algorithms that implement simple and advanced search supporting, account search, which takes individual user interests, direction of scientific activities, previous search queries of the user, as well as architecture of search module based on these search algorithms. Also our system (intelligent data and metadata exchange repository) has a taxonomy of DM

methods that allows to establish connection between DM methods and problem domain data on which they could be applied. The user ontological model, resources ontological model have been developed in protégé version 3.4. For ontological models interaction and implementation of search algorithms it was developed a set of general intelligent agents models. They can be used as a mechanism for displaying information on the ontological models, as well as a mechanism for user interaction with the system. This set of general models include model for integrating intelligent agents with web systems, a model of intelligent search agent, and model for relationship between agents. The user of the developed intelligent data and metadata exchange repository is able to make formal description of the user's problem domain (filling in the necessary fields in the ontology model) and formal description of the dataset which is need for specific tasks. All this kind of activities is a part of the search agent. The search agent, having processed the received information, transfers it to the coordinator agent and via the search agent the necessary connection with a data file is made. The user intelligent agent (user agent (profile agent)) allows to personalize the answer to the following questions: what is the user name; what is e-mail address; what language the user prefers; what are current goals of the user; whether the user is beginner or advanced one; what academic institution the user belongs to; what are localization preference of the user; whether the interests of user coincide with other users interests in the system; what are recent inquiries of the user. The result of applying the multi-agent approach for creating such system is the ability to perform a simple search for users regardless of user type; to search by different criteria for authorized users; to provide popular data sets; to perform a search taking into account the personal needs of the user; to provide user relevant queries information; to keep statistics of requests and, if necessary, provide this information; to remember the successful search results.

Here is used the cross platform programming language Java, multi-agent platform Jadex, database server Oracle Spatial 10g, and also the development environment for ontological models – Protégé Version 3.4. Database management system Oracle Spatial 10g which allows to work with ontologies in RDF format was chosen as a method of resource ontological models storage. Development environment of ontological models is Protégé.

---

### **Intelligent search agent design and development**

---

As one of basic concept of DMDR system is search agent.

For searching in the data and metadata exchange repository we have to develop a search module. It would consider the current state of system and different searching criteria to adopt any strategy of search [Ratushin, 2001]. One of the most suitable solutions to this problem is the intelligent agents based on goal. This intelligent agent will act not just in reflective way when a request came, but would decide what actions are needed to achieve its goals in terms of the current state of environment. This agent is not able to supervise the environment, where it's executed, in full [Wooldridge, 1995].

In the search module of "data and metadata exchange repository" is set problems such as simple and advanced search or personal search. On the other hand, the search agent is used in the multi-agent environment and agent needs to communicate between it-self and other agents as well as to exert the medium, where it's executed. From these two points of view the functionality of search agent may be divided into functionality in terms of user and functionality in terms of other agents and execution environment [Russell, 2006].

Functionality in terms of user should include the following basic set: to execute a simple search only for non-authorized user, to execute the various searches for authorized user and to provide useful services for the search (Figure 1).

Both authorized and non-authorized users may execute the simple search, in both cases there will be shown the most popular data sets in the repository or the most popular queries (queries most often made by users in the



repository) and the results may also be hints as content search queries (queries correlated with the current ones). But still the authorized user has more privileges in comparison with non-authorized one. The following is available for authorized user: advanced search (search by various data set criteria). When agent uses the search there is displayed some of recent queries. Information about user's requests and their results are stored using a personal agent and will be used further to provide user with more relevant results considering his previous requests.

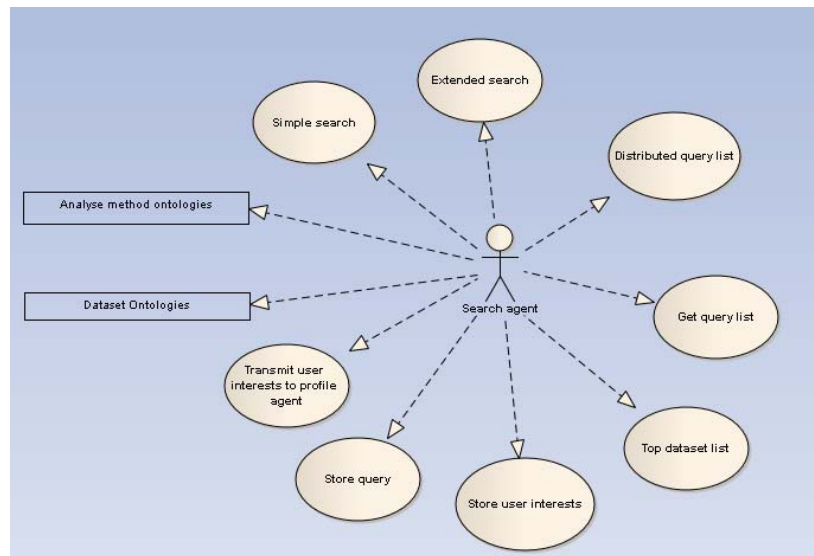


Figure 1 The use case diagram in terms of user

On the other hand, the search agent must interact with other agents to successfully achieve the goals. To render the useful information to them and to request the necessary information from them or to request them to provide a service. Personal Agent of user obtains the results from the search agent and the request itself, as well as to take the transition sequence of user until the user will find the necessary dataset. This information will be stored by the private agent in order that the search agent could further use it.

### User agent Scenario

The base information unit of the personal agent is ontology model of user . At the level of agent conception about the user is object model of user ontology. The main objective of the personal agent is to transfer information about users to other agents and to transfer the necessary information to user from the other system agents [Zaborovski, 2005]. So, personal agent should be able to form answers to queries from other agents of "data and metadata exchange repository" system and to modify the user profile during his work with the system. In accordance with the information and ontology model of user the personal agent should be able to form answers to questions related to user. We can allocate the following two partitions of information about user: personal information about user, information about current goals of user. In general case the personal agent should be able to respond the following questions: what is user name, what is his e-mail address, what language user prefers, what are the current goals of user; is user advanced or beginner (naïve, simple), what academic institutions the user belongs to; what are localization preference of the user; are the interests of user coincide with other users interests in the system; what are the recent requests of the user. Here the personal agent applies the developed information and ontology model of the user for questions, which can be requested by other agents while interaction with personal program agent during the work of user with the system [Xacken, 2005].



The User Agent is created after the user authentication. Figure 2 shows the User Agent functionality. When authentication and authorization are completed the user agent retrieves its knowledge information about users that later allows the user and other agents to access this information quickly. The user agent stores this information in its knowledge when active user is in the system and before the work cessation; the agent unloads this information into the database. The User Agents in the system as much as users have passed authentication at the current period. If the user does not operate with agent over 30 minutes, the user agent removes the search agent and itself from the system.

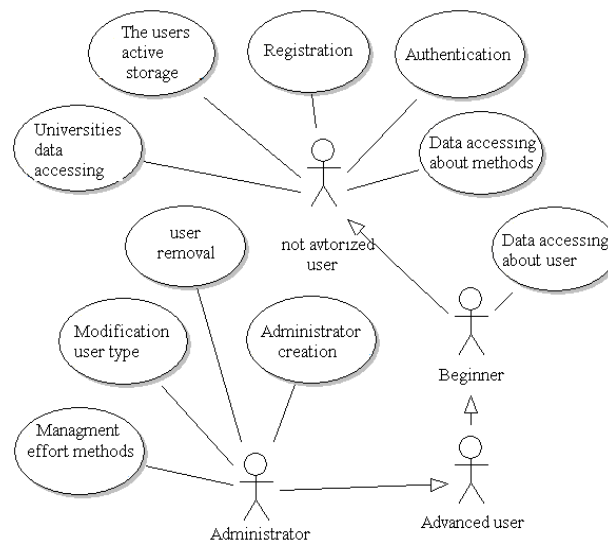


Figure 2 Use case for user agent

The user sets the following tasks before the user-agent:

- user personal data storing;
- changing of user data;
- preserve a user's search query to the system;
- conservation of user activity in the system;
- tracking the status of the user in the system;
- retention of the data sets loaded into the system;
- retention of the data sets unloaded from the system;
- User communication with other users of the system.

### Manager agent Scenario

To manage the overall system, registration and authorization of users in a “data and metadata exchange repository” operates the manager agent [Zimmermann, 2006]. The manager agent always suspends user and other agent's queries. Agent Manager exists in the system as a single copy. Agent Manager is parallelized by agent platform. Manager Agent provides functionality from the standpoint of the user schematically shown in Figure 3.

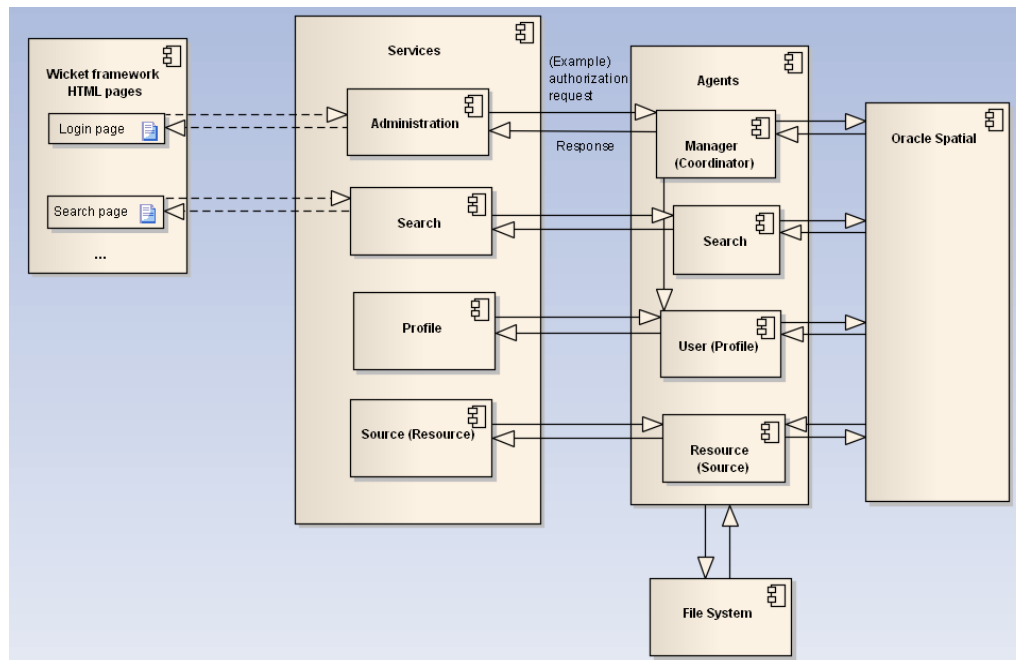


Figure 3 The explanation of overall system

Manager Agent stores the following internal information about the current state of the system:

- the number of users, who use the system;
- the number of beginners, who use the system;
- the number of advanced users, who use the system;
- research methods;
- general information about the universities of the system.
- Manager Agent receives information from the end users and from the environment. Input information from end-user of the system:
  - account and password;
  - user registration data;
  - user account that needs to be transferred into the status of the administrator;
  - user account that to be deleted.

End users interact with the agent manager. They invoke a Web service manager from the user interface.

System Administrators remove the users from the system by sending a request to the manager agent. For user to obtain the administrator rights, the other user-administrator should send the request to create a new administrator account on the basis of an existing one. Input information from the user agent: user account whose status should be changed. The transition from the one status to another is carried out by manager agent at the request of user agent of specific user. The Algorithm for the transition as follows: the user agent monitors the user activity in the system, and after getting some experience in the system, the agent prompts the user to raise his status and to receive additional options. If the user agrees the agent sends a request to the manager agent to change the type of user. Also the user invites to enter additional information about himself to obtain additional options. Manager

---

Agent sends the information to other agents and transmits the information to the end user through a Web service [Fatudimu, 2008].

---

### Source agent Scenario

---

The main functions of the source agent are:

- scientific data sets addition;
- interaction with the user agent to display the newly added samples to the user depending on the user's interests. The Source agent informs the user agent about adding of scientific datasets to show users information about it after adding a new set to the storage;
- Metadata of datasets edit. The users, who create system or administrator have the possibility to edit ;
- metadata dataset extract from repository;
- selection of entire information about a specific dataset and detailed information may be viewed only by registered users;
- to establish the dataset status numbers of downloads depending on estimates. The rating can be mark to each dataset. The rating assigned using the professional coefficient of a user, who makes it. At the moment of assess its assessment multiplied by a coefficient. This function performs source agent. The source agent should request the user agent ratio, calculate the result and save it in the database. Status of sampling can also increase depending on the number of downloads;
- interaction with the user agent to modify the coefficient of user professionalism depending on the status of scientific data sets, which he has added to the assessment or in storage;
- datasets filtering of metadata datasets by a specific parameter;
- new datasets adding to the repository that were found by search agent in the Internet.

The system must know the properties of the agent to create and run the agent. The state of the agent is determined by beliefs, goals, current plans, as well as libraries of known plans. Jadex uses the declarative and procedural approaches for implementing the components of the agent. The body of the plan is executed as ordinary Java classes. All other notions (beliefs, goals, filters, and conditions) are defined by language. They are allowed to create Jadex objects in a declarative manner. The program developer can refer to the Java code, for example, to define methods. Full identification of the agent is reflected in the so-called agent definition file (ADF). In the ADF file the developer defines the initial beliefs and goals, announcing Java facilities. Announce plans to show the necessary classes from Java code. In addition to the BDI components in ADF file can be stored, some other information, for example, the default arguments for starting the agent or service descriptions for the registration of the agent in the facilitator directory. The structure consists of Jadex API, performed by the model, reusable common features. API provides access to the concept Jadex during programming plans. Plans are obvious classes Java. It is extend a special abstract class which provides a useful method of sending messages, the organization of secondary objectives or expectations of the events. Plans are able to read and modify the agent's thoughts. It uses the API framework agreement. Special function Jadex is that, in addition to the direct extraction of the remaining facts, intuitive OQL - like query language is allow to formulate a random complex expressions using the facilities which are contained in the database views. In addition to plans, coded in Java, provides the developer based on the XML agent definition files (ADF). It establishes the initial thoughts, objectives and plans of the agent. The Jadex mechanism reads file and starts the agent. It tracks its goals during a continuous selection of steps and launches a plan based on internal events and messages from other agents. Jadex is equipped with some advance features - such as access to the directory facilitator service. Feature encoded in the individual plans, linked agent used in many modules which are called abilities. Ability is described in a format similar to the ADF. It can be easily incorporated into existing agents. So summarize, in Jadex agents

is thought, can be any type JAVA-site and stored in the database views. Objectives - explicit or imply descriptions of conditions that must be achieved. The agent executes the plans to achieve their goals. They are JAVA code procedural means.

Currently, there are many repositories of scientific datasets [Bresciani, 2004]. The main disadvantages occurred in these systems are: text-only format is not convenient to use and to change the format of files, not user-friendly interface, and the search is only by one of many criteria, i.e. not allowed to combine the search for a number of conditions, poor search.

In many systems, there is no any understanding for what tasks you can use this dataset, there is also insufficient information on the data. Currently, the agent technologies are widespread, where the main part is the agent - a software entity capable of such qualities as autonomy, activity, commitment, mobility, sociability. The creation of ontologies is a prospective direction of up-to-date research in processing of information provided in natural language. One of the advantages of using ontologies as a tool for learning is a systematic approach to the study of the subject area. Meanwhile achieved: regularity - Ontology provides a holistic view of the subject area, uniformity - the material presented in a unified format is much better perceived and reproduced; scientific - Building the ontology allows to restore the missing logical link in their entirety. Also, ontologies allow the use the great volumes of data from different systems, due to the fact they creating the semantic description of data. a) studied the main stages of work with the repository of scientific research data sets; b) reviewed the existing repositories of scientific data sets, to identify their strengths and weaknesses; c) studied the technology Semantic Web; d) investigated the possibility of agent technology; e) analyzed the ways to develop a web-oriented multi-applications; f) developed the architecture of multi-repository of scientific data sets; g) developed the ontological model of the user; h) developed and realized as a software BDI agent model of the user; i) developed and realized as a software BDI agent model.

---

## Conclusion

---

The results of the research is developed multi-agent system for processing and storage of any statistical data. A key feature of the developed system via others typical statistical repositories is implementation of the datasets metadescription using the European standard SDMX 2.0 and ontological models that are stored in the system.

The advantage and novelty of the work is implementation a set of the ontological models of Data mining methods, which is used for the selection of a proper method under the sample source of the user datasets. To work with set of the ontological models have been developed a set of search algorithms that implement simple and advanced search supporting, account search, which takes individual interests, orientation of activities, previous search queries of the user, as well as architecture of search module based on these search algorithms. Also this system (intelligent data and metadata exchange repository) has a taxonomy of DM methods that allows to establish connection between DM methods and data on which they can be applied, that for the user of "beginner" class represents itself as the expert system. The user ontological model, resources ontological model have been developed in protégé version 3.4, which allows working fast with ontologies.

For ontological models interaction and implementation of search algorithms it was developed a set of general intelligent agents models. They can be used as a mechanism for displaying information on the ontological models, as well as a mechanism for user interaction with the system. This set of general models include model for integrating intelligent agents with web systems, a model of intelligent search agent, and model for relationship between agents. The user of the developed intelligent data and metadata exchange repository is able to make formal description of the user's problem domain (filling in the necessary fields in the ontology model) and formal description of the dataset which is need for specific tasks. All this kind of activities is a part of the search agent. The search agent, having processed the received information, transfers it to the coordinator agent and via the

search agent the necessary connection with a data file is made. The user intelligent agent (user agent (profile agent)) allows to personalize the answer to the following questions: what is the user name; what is e-mail address; what language the user prefers; what are current goals of the user; whether the user is beginner or advanced one; what academic institution the user belongs to; what are localization preference of the user; whether the interests of user coincide with other users interests in the system; what are recent inquiries of the user. The result of applying the multi-agent approach for creating such system is the ability to perform a simple search for users regardless of user type; to search by different criteria for authorized users; to provide popular data sets; to perform a search taking into account the personal needs of the user; to provide user relevant queries information; to keep statistics of requests and, if necessary, provide this information; to remember the successful search results.

---

### Acknowledgement

---

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA ( [www.ithea.org](http://www.ithea.org) ) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine ( [www.aduis.com.ua](http://www.aduis.com.ua) ).

---

### Bibliography

---

- [Ratushin, 2001] Ratushin U., Polenok, S., Tkachenko, S. Information society ontology at the network. In: University book.
- [Wooldridge, 1995] Wooldridge, M., Jennings, N. (1995). Intelligent agents: Theory and practice. In: The Knowledge Engineering Review 10(2), 115-152.
- [Russell, 2006] Russell, S., Norvig, P. Russian translation of Artificial Intelligence: A Modern Approach, 2nd Edition, Translated by Pitstyn K. Ed: Moscow: Williams Publishing, ISBN Press, 356.
- [Zaborovski, 2005] Zaborovski, V. Intelligent technologies, 324.
- [Xacken, 2005] Xacken, G. Information and self-organization. Macroscopic approach to Complex system, 248.
- [Gennari, 2002] Gennari, J. The Evolution of Protégé. An Environment for Knowledge-Based Systems Development.
- [Xie, 2006] Xie T., Pei, J. MAPO: mining API usages from open source repositories. In: Proceedings of the International Workshop on Mining Software Repositories (MSR '06), Shanghai, China, ED: ACM Press, New York, 54-57.
- [Zimmermann, 2006] Zimmermann, T. Knowledge Collaboration by Mining Software Repositories. In: Saarland University, Saarbrücken, Germany
- [Fatudimu, 2008] Fatudimu I.T., Musa, A.G., Ayo, C.K, Sofoluwe, A. B. Knowledge Discovery in Online Repositories: A Text Mining Approach. In: European Journal of Scientific Research, 22 (2), 241-250. ED: EuroJournals Publishing
- [Bresciani, 2004] Bresciani, P., Perini, A., Giorgini, P., Giunchiglia, F., Mylopoulos, J. Tropos: An agent-oriented software development methodology. In: Journal of Autonomous Agents and Multi-Agent Systems 8 (3), 203–236

---

### Authors' Information

---

**Shatovska Tetyana** – Ass.Prof, Kharkiv National University of Radioelectronics, Kharkiv-166, av.Lenina 14, Ukraine; e-mail: [shatovska@gmail.com](mailto:shatovska@gmail.com)

Major Fields of Scientific Research: Data and Web mining, Artificial Intelligence

**Irana Kamenieva** –PhD student, Kharkiv National University of Radioelectronics, Kharkiv-166, av.Lenina 14, Ukraine; e-mail: [irina.kamenieva@gmail.com](mailto:irina.kamenieva@gmail.com)

Major Fields of Scientific Research: Data and Web mining, Artificial Intelligence

## LSPL-PATTERNS AS A TOOL FOR INFORMATION EXTRACTION FROM NATURAL LANGUAGE TEXTS

Elena Bolshakova, Natalia Efremova, Alexey Noskov

**Abstract:** *The paper describes main features of formal lexico-syntactic pattern language (LSPL) proposed for specification of linguistics information about NL expressions automatically extracted from Russian texts. A fully-implemented procedure for matching LSPL-patterns with text are presented, as well as developed programming tools for extraction of phrases specified by the patterns. Two applications of the language and the tools are discussed: terminological analysis of scientific texts and processing of NL sentences for question answering. LSPL-patterns developed for these applications are briefly characterized.*

**Keywords:** *information extraction from NL texts, lexico-syntactic patterns, matching procedure, automatic terms recognition and extraction, analysis of NL phrases for question answering.*

**ACM Classification Keywords:** *1.2.7 [Artificial Intelligence]: Natural language processing – Text analysis*

---

### Introduction

---

Information extraction (IE) from natural language (NL) texts is one of the most important problems of modern computer linguistics and artificial intelligence [Grishman R., 2003]. Traditionally, IE aims at identification in texts of selected types of entities (names and titles), relations, or events [Hearst, 1998; Boudin F. et al., 2008]. As a rule, IE applications are based on shallow syntactic analysis of the text and exploits both heuristics and linguistics information about items to be automatically recognized in it.

Among programming tools commonly used to create IE applications we should point out well-known system for text engineering GATE [Bontcheva K. et al., 2002], and analogous systems, such as Ellogon [Petasis G. et al., 2002]. They are rather universal and propose special formal languages for annotating text segments and describing annotation transformations. As a consequence of their universality, the annotation languages require skilled users to develop application for a new problem domain and different natural language. The languages have no built-in devices for describing specific linguistics properties, in particular, grammatical agreement, which are typical for such flexional natural languages, as Russian.

So far, a more specific language called LSPL (Lexico-Syntactic Pattern Language) was proposed for formal specification of linguistics information about NL expressions to be automatically recognized within Russian texts [Bolshakova et al., 2007]. The key language structure is lexico-syntactic pattern that describes certain NL phrase - its words and other constituent elements, as well as their morphologic and syntactic properties. An example of pattern for the English phrase for topic actualization is *"let us consider" NP* with *NP* denoting a noun phrase. In general case, LSPL-pattern combines both lexical and syntactic information about the described phrase, and thereby LSPL language is convenient to formally specify a wide range of common scientific expressions used for automatic discourse analysis of scientific and technical texts [Bolshakova, 2008].

In contrast to the annotation languages, LSPL was created as a linguistically-oriented and purely declarative formal language that is easy to use for:

- formal specification of a wide variety of NL phrases (noun-noun and verb-noun combinations, adverbial and participle phrases, etc.) within information extraction systems based on surface syntactical analysis of texts;

- user's queries for text browsers performing search of NL phrases and expressions specified by their lexico-syntactic patterns.

Elaborated LSPL language includes devices for specifying within patterns both particular word forms, lexemes and arbitrary words of particular part of speech (POS), as well as their morphological attributes and conditions of grammatical agreement. The latter presents an important language feature proposed specially for description of Russian noun phrases.

For proposed LSPL language a procedure for matching a pattern with a given Russian text was developed, as well as corresponding programming tools for recognition and extraction of phrases specified by LSPL-patterns. Making use of the tools, we investigated two different applications: automatic extraction of terms in scientific texts and analysis of NL sentences for question answering. The former is relatively well studied for English and French texts [Jacquemin C., 2003]. Aiming at terminological analysis of Russian scientific and technical texts, we created a representative set of LSPL-patterns that describes linguistics properties of multi-word term occurrences in texts and then we experimentally studied these patterns. Another developed application of LSPL language (and its supporting tools) is analysis of NL queries in question-answering system. The language proved to be convenient for both applications.

The paper starts with an overview of LSPL language; basic principles of the matching procedure for LSPL-patterns are overviewed as well. Their applications for automatic terms extraction and NL query analysis in question-answering systems are then discussed and conclusions are drawn. Since LSPL language was primary proposed and used for formalizing Russian phrases, illustrative examples are given mainly for Russian.

---

### LSPL language and LSPL-patterns

---

Lexico-syntactic pattern formalizes structure and properties of some NL phrase (noun phrase or verb-noun phrase, etc.). The pattern has a name and a body, the latter is separated by symbol of equality. Pattern body includes elements describing constituents of the phrase to be formalized. The order of the elements corresponds to the order of constituents in the phrase. As a rule, the pattern also specifies conditions of grammatical agreement for its elements. For example, the pattern  $NP = AN \langle A=N \rangle$  has the name AN and the body that consists of elements A, N and agreement conditions  $\langle A=N \rangle$ . This pattern describes a simple noun phrase: adjective (A) and noun (N) that are fully grammatically agreed (i.e. in case, number, and gender).

Basic pattern elements are elements-strings and elements-words. *Element-string* describes either a particular word form (e.g. Rus. "задаче́й" – word *problem* in instrumental case of singular), or particular symbols (for example, abbreviations or punctuation marks: ";"). *Element-word* describes a word, for which it may be specified:

- part of speech (POS: *N* – noun, *V* – verb, *A* – adjective, *Pr* – preposition, *Pn* – pronoun and so on);
- particular lexeme (i.e. all possible word forms of this word);
- particular values of morphologic attributes (they diminish the set of allowable word forms).

Morphologic attributes are written in angle brackets after the lexeme, with letter *t* denoting time, letter *p* denoting person, *c* – case, *n* – number, *g* – gender, etc.). For example, element-word

$V \langle \text{пониматься}, t=\text{pres}, p=3, m=\text{ind} \rangle$  describes Russian verb with lexeme *пониматься* taken in all forms of third person, present indicative (two word forms: *понимается* and *понимаются*). While describing an element-word, its morphologic attributes or its particular lexeme may be omitted, which makes it possible to allow within the corresponding phrase any word form of the given lexeme (e.g.  $N \langle \text{файл} \rangle$ ), or any word of the particular part of speech with needed values of morphologic attributes (e.g.  $A \langle ;c=\text{ins}, n=\text{sing} \rangle$  specifies an arbitrary adjective in instrumental case of singular).

Since LSPL-pattern often includes either several elements-words of different part of speech, or several different words of the same part of the speech, indices are used to distinguish the words. For example, the pattern  $NN = N1 N2 <c=gen>$  includes two different nouns  $N1$  and  $N2$ , the second is taken in gender case.

Agreement conditions describe relation of grammatical agreement for elements-words within the pattern. The conditions are written in angle brackets at the end of LSPL-pattern, similar to specification of values of morphologic attributes. They express the equality of values of morphologic attributes to be agreed. For instance, the pattern  $PnV = Pn V <Pn.n=V.n, Pn.g=V.g>$  specifies an arbitrary pair of pronoun and verb, which are agreed in number and gender (*Rus.: мы предположим; Eng.: we suppose*).

If some element of the phrase may occur in it successively several times, such a sequence is specified in corresponding pattern as repetition of elements, which is written in figure brackets. For example, repetition  $\{N <c=gen>\}$  describes sequence of nouns, each taken in genitive case. If the number of elements in the repetition is limited, it is specified in the pattern, for instance,  $\{A\} <1,3> N$  describes sequence including one, two or three adjective and a noun.

LSPL language also provides such useful device as optional element, which are written in square brackets, for example, the element ["не"] means, that particle *не* optionally enters the NL phrase under description. Another convenient device is alternative variants of the phrase – they should be written in the pattern through sign |. For instance, the pattern  $AP = A|Pa$  specifies Russian concept of adjective, i.e. adjective (A) or participle (Pa).

In order to describe patterns of complex phrases, one can use yet defined LSPL-patterns as auxiliary patterns within the main pattern. Let us consider the pattern  $NG = \{A\} N1 [N2 <c=gen>] <A1=N1>$  that includes the element-word  $N1$  (principal word of the phrase), sequence of adjectives  $\{A\}$ , which are agreed with the principle word ( $<A1=N1>$ ), and also optional noun in genitive case  $[N2 <c=gen>]$ . Phrases with such structure are frequently used as terms in Russian texts (e.g., *восходящий процесс порождения, удаленный банковский терминал*). Based on the auxiliary pattern NG, the pattern  $S = NP V <t=past>$  specifies any phrase including a noun phrase NP and a verb in the past (e.g., *опорная точка уточнялась*).

Lexico-syntactic pattern may also have parameters, they are written in brackets, after all pattern elements and agreement conditions. The parameters fix some unvalued morphological attributes of pattern elements. For the LSPL-pattern  $AAN = A1 A2 N <A1=A2=N>$  (N), morphological parameters of the element-word N are specified as pattern parameters (the pattern describes noun phrase with elements-adjectives A1 and A2 fully agreed with noun N).

Pattern parameters are especially useful when the pattern is used as an element within another pattern. Suppose the pattern NG considered above has the parameter N1 (i.e. morphological attributes of the noun N1):

$NG = \{A\} N1 <A=N1> [N2 <c=gen>] (N1)$

Then one can use the parameter for agreement. For instance, the pattern  $NG V <NG=V>$  describes phrase consisting of noun phrase NG and verb V grammatically agreed with it (e.g., Russian word combination *внутренний файл проверялся* is allowable, but *внутренний файл проверялись* is not, since the noun is not agreed with the verb).

Pattern parameters are also useful for specifying values of morphological attributes of the pattern used within the outer pattern; the specification is written in angle brackets, similar to specification of attributes of elements-words, for instance, in the pattern  $NG <c=gen> V$  the noun phrase NG is specified in gender case.

In overall, LSPL language is a flexible and powerful tool for describing lexical and grammatical properties of NL phrases to be recognized in texts.



## Matching LSPL-Pattern with Text

For recognizing within a given NL text all phrases described by the particular LSPL-pattern, a matching procedure was developed. We call recognized phrases and corresponding text segments *variants of matching* of the pattern with the text. Each matching variant presents a text segment together with particular morphologic attributes of all its constituent words; the set of the particular values of morphologic attributes we call *syntactic interpretation* of the segment. When the segment consists of a single word, its syntactic interpretation is simply all morphological attributes of the word. In general case, for a given LSPL-pattern and text there exists several variants of matching, they correspond to different occurrences of the phrase described by the pattern.

Our matching procedure is based on special inner representation of the text – *graph of the text*. Nodes of the graph corresponds to space symbols, punctuation marks and all the other symbols that are not significant for matching; to be more precise, any segment of all such adjacent symbols constitute a node. Edges of the graph correspond to syntactic interpretations of text segments between the nodes.

While constructing the graph, first, segmentation of the text is done (words are delimited, as well as sequences of symbols that are not significant), and nodes of the graph are constructed and numbered from the beginning of the end of the text. Then morphologic analysis of all words is performed, and neighbor nodes are connected with edges represented morphologic interpretations of the words between them. If there exist several different morphologic interpretations of the same word, the corresponding nodes are connected with several edges. An example of graph representation for Russian sentence is presented on Figure 1 (the segmentation is also shown above the graph). One can notice that the number of morphologic interpretations for the same word form may be quite great. For example, word form *большой* has six interpretations, while the word *нечеткий* has only two.



Figure 1. Graph of text with variants of matching the pattern  $NP = A N \langle A=N \rangle (N)$

Various ways in the graph of text corresponds to various possible combinations of morphologic interpretations of words. Therefore we consider the task of matching a pattern with the text as the task of searching a way (or a subway) within the graph that conforms to the pattern (i.e. pattern elements and agreement conditions).

Intermediate results of matching are also saved within the graph of text: any phrase recognized by matching with the pattern (main or auxiliary) is represented in the graph as a new edge connecting nodes pointing to beginning and end of the phrase (i.e. its text segment). This new edge presents matching variant for the pattern, and if the pattern has parameters, their values are additional attributes of corresponding syntactic interpretation.

In Figure 1 edge A represents the variant of matching of the pattern  $NP = A N \langle A=N \rangle (N)$  with text segment *большой проблемой*, while edges B and C represent two different variants of matching of the pattern with segment *нечеткий поиск* (they differ in syntactic interpretation: B corresponds to nominative case and C to accusative). Thus, more than one variant of matching may be detected for the same text segment.

When LSPL-pattern includes repetition of elements, its matching also gives a new edge connecting all elements of repetition recognized in the text.

Therefore, the proposed graph of text is a convenient way to represent various syntactic interpretations and their combinations, as well as to uniformly process both elements-words and auxiliary patterns. It also allows optimizing of matching with the aid of indices constructed simultaneously with the graph. For this purpose, three types of indices are used: index of particular words, indices of parts of speech, and index of patterns yet matched. Another optimization method also used by matching procedure is grouping of various syntactic interpretations, which diminishes the number of matching variants to be considered while searching way within the graph of text. The described matching procedure is a core of programming tools developed to support LSPL language. These tools include console utilities for integration the core with various scripts, API for Java programming language, and graphic user interface. All the tools were first used to develop and to test automatic term extraction procedures for Russian scientific and technical texts.

### LSPL-Patterns for Terminological Analysis of Scientific Texts

In order to formalize heterogeneous linguistics information needed to automatically extract terms and term definitions, an empirical study of terminology dictionaries and texts in several scientific fields (approx. 330 texts in computer science and physics) was performed. Based on the study, the formalization was done with the aid of LSPL language, resulting in a set of LSPL-patterns. The set comprises 6 groups of patterns that take into account various properties of term occurrences within Russian scientific texts. These groups, corresponding examples of patterns and examples of recognized term occurrences are presented in Table 1.

Table 1. LSPL-patterns for terminological analysis

N	Pattern Groups	Examples of Patterns	Examples of Terms and Term Occurrences
1	Morphosyntactic patterns of terms	$A1 N1 <A1=N1> (N1)$ $N1 A2 N2<c=gen> <A2=N2> (N1)$	активные долги технология двойной накачки
2	Definitions of authors' terms	$Defin<c=acc> "называют" ["также"] Term<c=ins> \# Term<c=nom>$ $"нод" Term<c=ins> "понимается" Defin<c=nom> \# Term<c=nom>$	Эту проблему называют также проблемой скрытого состояния Под прерыванием понимается сигнал...
3	Contexts of introduction of terms' synonyms	$Term1 ("Term2") <Term1.c=Term2.c>$ $\# Term1<c=nom>, Term2<c=nom>$	<u>автокорреляционной функции (АКФ)</u> , зоны анализа (сегменты)
4	Dictionary terms	$N1<вектор> [N2<намагниченности,c=gen>  N2<состояния,c=gen> "Умова"]$	вектор, вектор намагниченности, вектор состояния, вектор Умова
5	Lexico-syntactic variants of terms	$N1 N2<c=gen> \# N1,$ $N1 N4<c=gen> <Syn(N2,N4)>,$ $N3 N2<c=gen> <Syn(N1,N3)>,$ $A1 N1 <A1.st=N2.st>$	коллекция текстов – коллекция (N1), корпус текстов (N3 N2), текстовая коллекция (A1 N1)

6	Combinations of several terms	$N1 \ N2<c=gen> \ " \ N3<c=gen> \ \{ "u"   "илу" \}$ $N4<c=gen>$ $\# N1 \ N2<c=gen>, N1 \ N3<c=gen>, N1 \ N4<c=gen>$	шина адреса, данных и управления – шина адреса, шина данных, шина управления
		$N1 \ A2 \ N2<c=gen> \ <A2=N2>$ $\# N1 \ N2, A2 \ N2$	разрядность внутреннего регистра – разрядность регистра, внутренний регистр

The first group describes morphosyntactic structure of one-, two- and tree-word terms frequently used in texts. Each pattern fixes part of speech of its element-words and morphological attributes of words (if necessary).

The second group formalizes typical one-sentence definitions of new terms introduced by authors of texts (so called *author's terms*); an example of such English definition is *Light quanta came is called photons*. Each LSPL-patterns of the group uses special auxiliary patterns Term and Defin. The former comprises all allowable morphosyntactic patterns of terms (i.e. patterns of the first group), the latter describes syntactic structure of phrases explicating meaning of new terms. Each pattern of the group also includes special element  $\# \text{Term} \ <c=nom>$ , which specifies a constituent part of the recognized term definition to be extracted, as well as its lemmatization conditions (nominative case is specified for extracted term Term).

The third group includes LSPL-patterns of contexts typically used in Russian scientific texts to introduce synonymous terms (in particular, acronyms, such as CPU for term *central processing unit*).

As LSPL language proved to be convenient for describing entries of terminology dictionaries, the fourth group of patterns was constructed to specify particular terminological words and word combinations in two scientific fields – computer science and physics.

The last two groups of LSPL-patterns describe general derivation rules for text variants of terms. Term variation and methods of term variants recognition was well investigated for English and French text [Nenadic G. et al, 2003], and we conducted analogous research for Russian texts. Besides variation of single term (cf. the group of lexico-syntactic variants in Table 1), we additionally considered typical combinations of several terminological word combinations and formalized their properties.

Each pattern of the fifth group fixes particular morphosyntactic structure of the term and specifies as the extracted element (i.e., after special sign #) morphosyntactic structure of its possible lexico-syntactic variants. In particular, if the structure of term is  $N1 \ N2<c=gen> \ \#N1$ , the following lexico-syntactic variants are described:

- i) insert (or deletion) of word (Rus. ввод данных – ввод, Eng. data input – input);
- ii) substitution of a synonym (in the given problem domain) for constituent part of the term (Rus. *фрейм активации - запись активации*; Eng. *activation frame - activation record*);
- iii) substitution of a word with the same root but another part of speech (Rus. шина адреса – адресная шина, Eng. address bus – bus of address).

The last group of LSPL-patterns describes (in a similar manner) derivation rules of text variants combining several terms. The rules take into account two different cases:

- combinations with coordinating conjunctions (Rus. шина адреса, шина данных, шина управления – шина адреса, данных и управления; Eng. address bus, data bus, control bus – address, data, and control bus);
- conjunctionless combinations (Rus. разрядность регистра, внутренний регистр – разрядность внутреннего регистра; Eng. capacity of register, internal register – capacity of internal register).

In both cases within described combinations one or more multi-word terms are discontinuous or truncated, and this is the real problem of their automatic recognition.

For each group of patterns, an automatic recognition procedure was developed and experimentally studied (in particular, a procedure for extracting new terms and their definitions). The recall of automatic recognition proved to be from 57% (for synonymous term recognition) to 85% (for dictionary terms), while precision varies from 32% (for synonymous term recognition) to 97 % (for authors' terms). In order to accomplish more accurate and full term detection and extraction, we then elaborated a strategy of consequent call of the procedures, which gives 7-14 % increase of F-measure (the combined measure of precision and recall).

### LSPL-Patterns for Processing NL Sentences in Question-Answering System

LSPL language was also used to formally specify input NL phrases in a prototype question-answering system based on logical inference. The system is domain-independent, it gives answers to questions about existence of entities (animals, humans, games, cars, etc.) with particular properties (high, quick, black, difficult, etc.) or about properties of particular entities. Some properties of the entities are to be previously described by Russian sentences (these are initial statements). The system translates them to first-order logical formulas, which serves as axioms. In general case, axioms include universally and existentially quantified formulas-sentences (e.g. *all women like talking, some cats are black*), as well as formulas with implication (*if book is large, it is high-priced*).

The system uses axioms to infer answers to questions formulated in Russian (e.g. *Are all cats grey?*). For this purpose, the given question is translated to a logical formula and the resolution method is applied to prove it. Since questions are to be closed first-order formulas, the system gives either positive or negative answer to the given question.

LSPL-patterns developed for the question-answering system describes lexicon and syntax of input Russian sentences: either statements and questions. The patterns are divided into 5 groups presented in Table 2 together with corresponding examples (terms of computer games are mainly used in them).

Table 2. LSPL-patterns for question answering

N	Pattern Groups	Examples of Patterns	Examples of Phrases
1	Auxiliary patterns	$SubjDelim = ", " ["a" "также"]   "u"   "илу"   "а" "также"$	маги, колдуны, а также волшебники
		$Exists = V <существовать> (V)   V <быть> (V)   V <бывают> (V)$	эльфы <u>бывают</u> светлые
2	Patterns of entities	$EntityBase = \{Adjective\} N$ $\{SubjDelim Entity\} <Adjective=N> (N)$	красный рыцарь
		$Entity = EntityBase [ Which Predicate \{Delim Predicate\} ["", "]]$ $<EntityBase=Predicate> (EntityBase)$	герой, который хорошо колдует
3	Patterns of properties	$Predicate = \{Av\} A (A)$	Очень сложный уровень
		$Predicate = \{Av\} V (V)$	Маг легко обучается

4	Patterns of statements	$Statement = Pn\langle\text{некоторый}\rangle Entity Predicate \{Delim Predicate\} \langle Entity=Predicate \rangle$	Некоторые феи хорошо поют
		$Statement = [Pn\langle\text{весь}\rangle] Entity ["/"] Predicate \{Delim Predicate\} \langle Entity=Predicate \rangle$	Все маги – бессмертные
		$Statement = \text{"если"} Entity ["/"] Predicate \{Delim Predicate\} ["/"] \text{"мо"} ["/\text{он}/\text{она}"] Predicate \{Delim Predicate\} \langle Entity=Predicate \rangle$	Если рыцарь быстро бегают, то он неуязвим
5	Patterns of questions	$Question = Predicate \text{"лу"} Entity \{Av\} \langle Av=Predicate=Entity \rangle$	Бессмертны ли эльфы?
		$Question = \{Av\} \langle 1 \rangle \text{"лу"} Predicate Entity \langle Av=Predicate=Entity \rangle$	Долго ли живут орки?

The first group specifies general-purpose words (such as Rus. *быть*) and structure of auxiliary language constructs (in particular, enumeration phrases). The second and the third groups formalize respectively syntax of various phrases denoting entities (noun and participle phrases) and syntax of phrases denoting their properties (noun and verb phrases). The forth group comprises patterns of various statements to be translated to logical axioms: universal sentences, existential sentences, and sentences expressing implications. The last group includes patterns of user questions (similar to the previous group, universal and existential questions are allowable and specified). Most patterns of two last groups take into account various order of constituent words in the phrases (in Russian, the order is quite free), and as a consequence, these patterns are complicated.

We should note that the expressive power of LSPL language has made it possible very quick development of a procedure for translation NL sentences to logic formulas.

---

## Conclusion

In the paper we have overviewed formal declarative language LSPL proposed to specify lexico-syntactic patterns of NL phrases to be recognized in Russian texts and extracted from them. We also described main features of matching procedure intended to recognize the phrases within a given text based on a particular set of LSPL-patterns and shallow syntactic analysis.

The programming tools (with the matching procedure as a core) developed to support the language were experimentally tested while investigating two quite different applications, the first is terminology analysis of scientific and technical texts, and the second is processing of NL phrases for question answering. The programming tools (including the user interface similar to a text browser driven by patterns) have demonstrated their working efficiency.

Our experience shows that LSPL language is well-suited for quite different NL processing tasks. Another potential applications of the language to be further investigated include text summarization, computer-aided editing of scientific and technical texts, and intra-document browsing and retrieval.

---

## Bibliography

- Bolshakova, E.I., Baeva N.V., Bordachenkova E.A., Vasilieva N.E., Morozov S.S. Lexicosyntactic Patterns for Automatic Processing of Scientific and Technical Texts. In: Proc. of 10th National Conference on Artificial Intelligence with International Participation 2006. Moscow, Fizmatlit, Vol 2, 2006, p. 506-514 (in Russian).
- Bolshakova E. I. Common Scientific Lexicon for Automatic Discourse Analysis of Scientific and Technical Texts // International Journal on Information Theories and Applications. Vol. 15, 2008, No 2, p. 189-195.

- Bontcheva K. et al. Developing Reusable and Robust Language Processing Components for Information Systems using GATE. In: Proceedings of the 13th Int. Workshop on Database and Expert Systems Applications, DEXA. Washington, 2002, p. 223-227.
- Boudin F. et al. Mixing Statistical and Symbolic Approaches for Chemical Names Recognition. In: Computational Linguistics and Intelligent Text Processing. A. Gelbukh (Ed.). LNCS, No. 4919. Springer, 2008, p. 334-343.
- Hearst, M.A. Automated Discovery of WordNet Relations. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database. MIT Press, Cambridge, 1998, p.131-151.
- Grishman R. Information extraction. In: Mitkov R. (ed.): Handbook of Computational Linguistics. Oxford University Press, 2003. p. 545-59.
- Jacquemin C., Bourigault D. Term extraction and automatic indexing. In: Mitkov R. (ed.): Handbook of Computational Linguistics. Oxford University Press, 2003. p. 599-615.
- Nenadic G., Ananiadou S., McNaught J. Enhancing Automatic Term Recognition through Variation. In: Proceedings of 20th Int. Conference on Computational Linguistics COLING'04, 2004, p. 604-610.
- Petasis G., et al. Ellogon: A New Text Engineering Platform. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002). Las Palmas, 2002, p. 72-78.

---

### Authors' Information

---

**Elena I. Bolshakova** – Moscow State Lomonossov University, Faculty of Computational Mathematics and Cybernetic, Algorithmic Language Department; Leninskie Gory, Moscow State University, VMK, Moscow 119899, Russia; e-mail: bolsh@cs.msu.su

*Major Fields of Scientific Research: Artificial Intelligence, Natural Language Processing*

**Natalia E. Efremova** – Moscow State Lomonossov University, Faculty of Computational Mathematics and Cybernetic, Algorithmic Language Department; Leninskie Gory, Moscow State University, VMK, Moscow 119899, Russia; e-mail: nvasil@list.ru

*Major Fields of Scientific Research: Automatic Extraction of Terms and Relations, Sentiment Analysis*

**Alexey A. Noskov** – Moscow State Lomonossov University, Faculty of Computational Mathematics and Cybernetic, Algorithmic Language Department; Leninskie Gory, Moscow State University, VMK, Moscow 119899, Russia; e-mail: alexey.noskov@gmail.com

*Major Fields of Scientific Research: Artificial Intelligence, Software Engineering, Natural Language Processing*

---

## COMPUTER SUPPORT OF SEMANTIC TEXT ANALYSIS OF A TECHNICAL SPECIFICATION ON DESIGNING SOFTWARE

Alla V. Zaboлева-Zotova, Yulia A. Orlova

**Abstract:** *The given work is devoted to development of the computer-aided system of semantic text analysis of a technical specification. The purpose of this work is to increase efficiency of software engineering based on automation of semantic text analysis of a technical specification. In work it is offered and investigated a technique of the text analysis of a technical specification is submitted, the expanded fuzzy attribute grammar of a technical specification, intended for formalization of limited Russian language is constructed with the purpose of analysis of offers of text of a technical specification, style features of the technical specification as class of documents, , algorithmic support of semantic text analysis of a technical specification and construction of software models are considered, recommendations on preparation of text of a technical specification for the automated processing are formulated. The computer-aided system of semantic text analysis of a technical specification is considered. This system consist of the following subsystems: preliminary text processing, the syntactic and semantic analysis and construction of software models, storage of documents and interface.*

**Keywords:** *natural language, semantic text analysis, technical specification.*

**ACM Classification Keywords:** *I.2.7 Natural Language Processing*

---

### Introduction

---

Most known of the commercial software products used at designing of the software, basically are intended for visualization intermediate and end results of process of designing. Some of them allow to fully automate last design stages: generation of a code, creation of the accounting and accompanying documentation, etc. Thus the problem of automation of the initial stage of designing - formations and the analysis of the text of the technical project remains open. It is connected to extraordinary complexity of a problem of synthesis and the analysis of semantics of the technical text for which decision it is necessary to use methods of an artificial intellect, applied linguistics, psychology, etc. However, it is possible to come nearer to achievement of the given purpose, having allocated some small subtasks quite accessible to the decision by known methods of translation.

Proceeding from the aforesaid, it is possible to draw a conclusion, that the problem of creation of means for automation of process of designing is actual [1]. Ideas of a developed direction realization of the unified procedures of the designing equally answering to requirements of the expert - designer and requirements to technology to modelling of software products is main.

Area knowledge of software design consists of the following topics: basic concepts of design software, the key issues of designing software structure and software architecture, analysis and evaluation of the quality of design software, notation software design, strategy and methods of designing software.

In the modern information technology has an important place tools, systems development and maintenance of software. These technologies and the environment form a CASE-systems. The well-known CASE-systems, such as BPWin, ERWin, OOWin, Design / IDEF, CASE-Analyst, Silverrun, Rational Rose, Vantage Team Builder, S-Designer, etc., allow partially automate the process of designing software. However, as shown by the analysis, these systems automate the final stages of design software, such as the creation of the balance sheet and

accompanying documentation, code generation, etc. The initial phase of the design - text analysis specification and construct a model of software - runs an analyst and automate of this phase remains open.

On CAD-department of the Volgograd state technical university questions of automation of designing of software products with use of natural - language support for a number of years are investigated.

Ideas of a developed direction realization of the unified procedures of the designing equally answering to requirements of the expert - designer and requirements to technology to modelling of software products is main.

Designing of the software at the initial stages with use of a natural language is based on the following main principles:

1. Performance of all design procedures is modelled in language of internal representation of system. Internal representation is the unified model of designing of the software, based on methodology of the theory of systems and technologies of natural language processing.
2. A number of representations of the project is generated. Translation of a condition of the project into the certain language which is distinct from language of internal representation refers to as representation. Programming languages, natural languages or artificial formal languages of modelling of processes of designing can be attributed to such languages (UML, IDEF-diagrams, model of diagrams of streams of the data). Different representations reflect only separate aspects of the project.
3. Thus due to use of uniform internal model consistency of representations is provided.
4. The software of process of the designing, guaranteeing an opportunity of conducting the project on any of languages of representations is developed.
- 5 The basic language of representation of the project for the person - the customer and the designer - is the natural language. Dialogue between the customer and the designer is traditionally conducted in a natural language - language of human dialogue, but, as a rule, are entered new formalism- diagrams, circuits, schedules. According to the developed concept, natural - language representation of the project supplements formal and serves as the tool facilitating understanding of process of designing.

As illustration of process of designing ON with use of the offered concept the diagram «to be», resulted on figure 1 serves.

The given work is devoted to development of the computer-aided system of semantic text analysis of a technical specification.

The purpose of this work is to increase efficiency of software engineering based on automation of semantic text analysis of a technical specification (TS). To achieve this purpose it is necessary to solve the following tasks:

1. To carry out the analysis of software engineering process and models of semantic text analysis;
2. To develop a technique of the text analysis of a technical specification;
3. To develop and investigate semantic model of the text of a technical specification;
4. To develop algorithmic maintenance of analysis of text of a technical specification and automatic construction of the software models;
5. To realize developed formalisms, a technique and algorithms as system of automation of the initial stage of designing software.



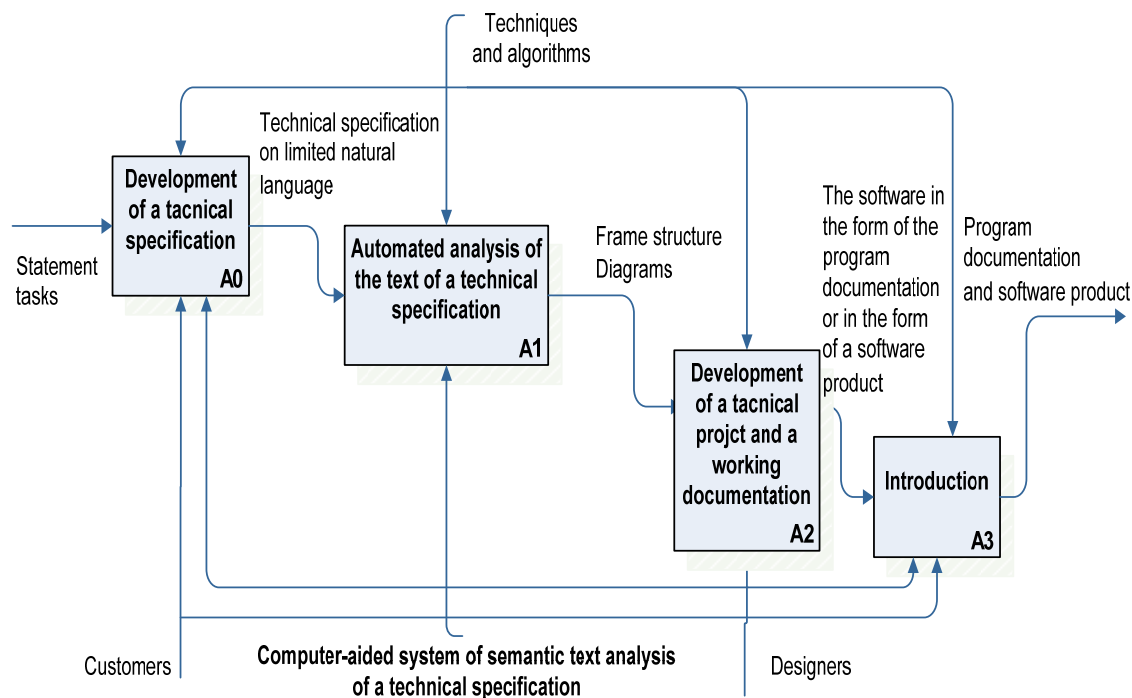


Figure 1: Diagram of process of designing «TO BE»

## A Technique Of The Text Analysis Of A Technical Specification

In work it is offered and investigated a technique of the analysis of the text of a technical specification is submitted, the fuzzy attribute grammar of a technical specification, intended for formalization of limited Russian is constructed with the purpose of analysis of offers of text of a technical specification, style features of the technical specification as class of documents are considered, recommendations on preparation of text of a technical specification for the automated processing are formulated.

A technique of the analysis of the text of a technical specification consist of three stages: semantic text processing, creation of frame structure and creation of data flow diagrams of system described in the technical specification. (see Figure 2).

For realization of the first stage of a technique the semantic model of the text of a technical specification, including the requirements formulated as the document in the limited natural language has been developed; the second stage - the frame structure being internal representation of requirements; the third stage - model of software as the description of requirements in graphic language Data Flow Diagrams.

The semantic model of the text of a technical specification contains the developed expanded fuzzy attribute grammar above frame structure of the formal document "Technical specification" which allows to display contents TS most full.

The expanded fuzzy attribute grammar, necessary for the automated analysis of the text of a technical specification, is determined as:

$$AG = \langle N, T, P, S, B, F, A, D(A) \rangle,$$

where  $N$  - final set of non-terminal symbols;  $T$  - not crossed with  $N$  set of terminal symbols;  $P$  - final set of rules;  $S$  - the allocated symbol from  $N$ , named an initial symbol;  $B$  - set of linguistic variables  $\beta_{k,i}$ , corresponding to terminal symbols  $T$  (a variable  $i$  on  $k$  level);  $F$  - set of functions of a belonging  $f_{k,i}$ , determining a degree of belonging  $m_{k,i}$

linguistic variables  $\beta_{k,i}$ ; A - set of attributes,  $A = A_{sin} \cup A_{sem}$ , where  $A_{sin}$  - syntactic attributes,  $A_{sem}$  - semantic attributes; D (A) - final set of semantic actions. The fragment of grammar is submitted in table 1.

Linguistic variables from set  $B = \{\beta_{k,i}\}_{k,i}$  used for the analysis of the text of a technical specification is described by the following five:

$$\beta_{k,i} = \langle \beta, T(\beta), U, G, M \rangle,$$

$\beta$  - name of linguistic variable (basis for development, purpose of development, technical requirements to a program product, a stage and development cycles, etc.);

$T(\beta)$  - language expressions. For linguistic variables of the top level they are the linguistic variables corresponding to terminals of the right part of a rule. For linguistic variables of the bottom level – fuzzy variables, that is expressions of a natural language.

U - Set of all probable values,  $T(\beta) \subset U$ ;

G - rules of the morphological and syntactic description of language expressions which determine syntactic attributes  $A_{sin}$ ;

M - a semantic rule for linguistic variables which is induced by morphological and syntactic rules as the sense of a term in T is in part determined by its syntactic tree, and semantic attributes  $A_{sem}$ .

Methods of representation connections between rules are broadcast on language of fuzzy mathematics. Thus connections are represented by fuzzy relations, predicates and rules, and sequence of transformations of these relations - as process of an fuzzy conclusion.

Linguistic variables of the top level are compound, that is include linguistic variables of the bottom level. Due to this it is possible to construct a tree of linguistic variables and to establish dependence between them.

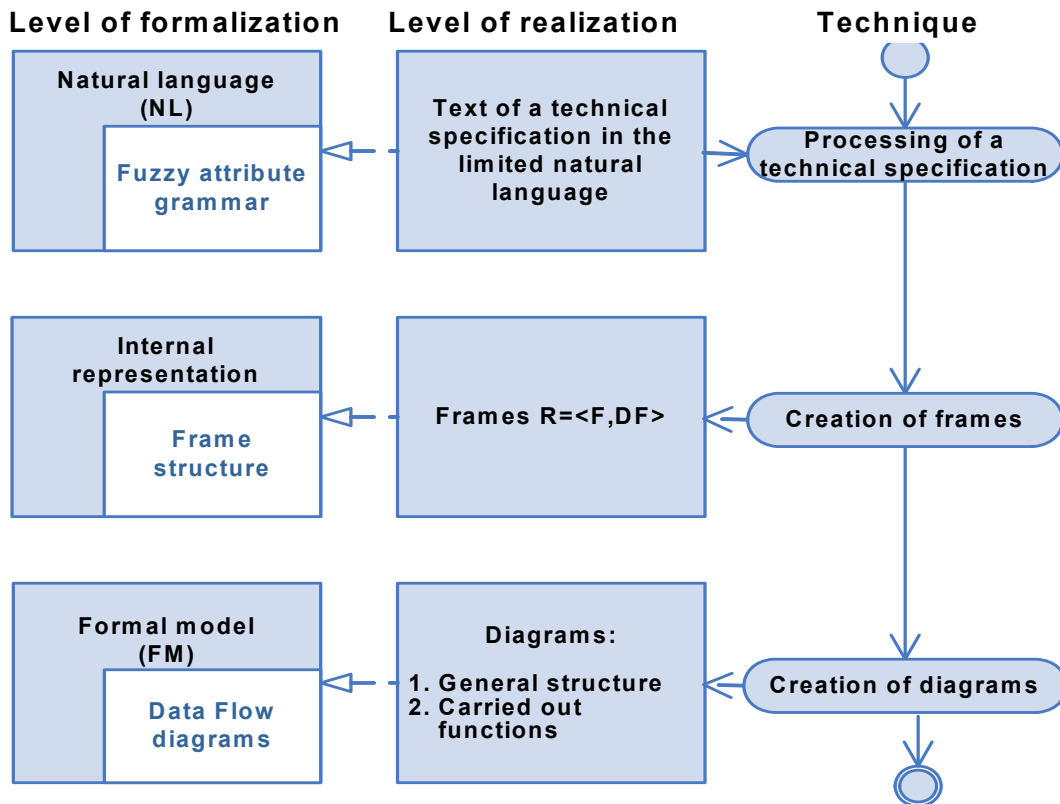


Figure 2: Technique Of The Text Analysis Of A Technical Specification

Table 1: Fragment of the developed fuzzy attribute grammar above frame structure of a technical specification

$\beta_1$	<i>&lt;list of incoming data flows &gt;</i>	<i>&lt;incoming data flow name &gt; :: 'Name' &lt;incoming data flow description&gt; :: 'Contents' &lt; list of incoming data flows &gt;   ε</i>
	<i>&lt;incoming data flow description&gt;</i>	The text containing "entrance" or "entrance data " :: 'Clause' <incoming data flow>::"Frame Data Flow=Creation ", "Input=Giving"
$\beta_{1,2}$	<i>&lt;incoming data flow&gt;</i>	[<Number of data units>]:: "Slot AMOUNT OF DATA = Giving" [<Type of data>]:: "SLOT TYPE OF DATA = Giving " <the Name of incoming data flow >:: " Slot NAME OF INCOMING DATA FLOW= Giving"
$\beta_2$	<i>&lt;function specification &gt;</i>	< function type < name of the functions liss > :: 'Name'< function description >:: "Frame FUNCTION = Creation "; < List of functions>   ε
$\beta_{2,1}$	<i>&lt; function type &gt;</i>	«main»   «basic»   «additional»
$\beta_{2,2}$	<i>&lt;function description &gt;</i>	< Name of function>:: 'Name', " Slot NAME OF FUNCTION = Giving " < List of incoming data flow > <List of out coming data flow>

Functions of an a belonging from set  $F = \{f_{k,i}\}_{k,i}$  linguistic variables  $\{\beta_{k,i}\}_{k,i}$ , are necessary for construction of an fuzzy conclusion. In particular, to each rule of grammar from set P function of a belonging  $f_{k,i}$  is put in conformity. This dual system of substitutions is used for calculation of sense of a linguistic variable.

Actually grammar of a technical specification is used for splitting the initial text of the document into sections and processings of most important of them for our problem. It needs precise observance of structure of the document. Technical specification represents the structured text consisting of sequence of preset sections.

The frame structure of the technical specification is submitted as:

$$R = \langle N_R, \overline{F}_R, \overline{I}_R, \overline{O}_R \rangle$$

where  $N_R$  is a name of system,  $F_R$  is system functions vector,  $I_R$  is incoming data flows vector,  $O_R$  is outgoing data flows vector.

$$\overline{F}_R = \langle F_R^1, F_R^2, \dots, F_R^k \rangle, \text{ then } F_R^i = \langle N_F^i, \overline{I}_F^i, D_F^i, G_F^i, H_F^i, \overline{O}_F^i \rangle,$$

Where  $N_F^i$  - a name of function  $F_R^i$ ,  $\overline{I}_F^i$  - incoming data flows vector of F function,  $D_F^i$  - the name of the action which are carried out by function,  $G_F^i$  - subject of the function action,  $H_F^i$  - restrictions on function,  $\overline{O}_F^i$  - a outgoing data flows vector of F function.

Let's denote the data flow by DF (Data Flow), then  $I_R, O_R, I_F, O_F$  are denoted by:

$$DF = \langle N_{DF}, D_{DF}, T_{DF}, C_{DF} \rangle$$

Where  $N_{DF}$  - data flow name,  $D_{DF}$  - data flow direction,  $T_{DF}$  - data type in flow,  $C_{DF}$  - data units per frame.

The model proposed is represented as a frame model with "a-kind-of" links (see Figure 3).

### Algorithmic support Of Semantic Text Analysis Of A Technical Specification And Construction Of Software Models

General algorithm of semantic text analysis of a technical specification consists of the following blocks: preliminary text processing, syntactic and semantic analysis and construction of software models. Preliminary text processing is carried out using the apparatus of finite state machine, one of which is shown in Figure 4.

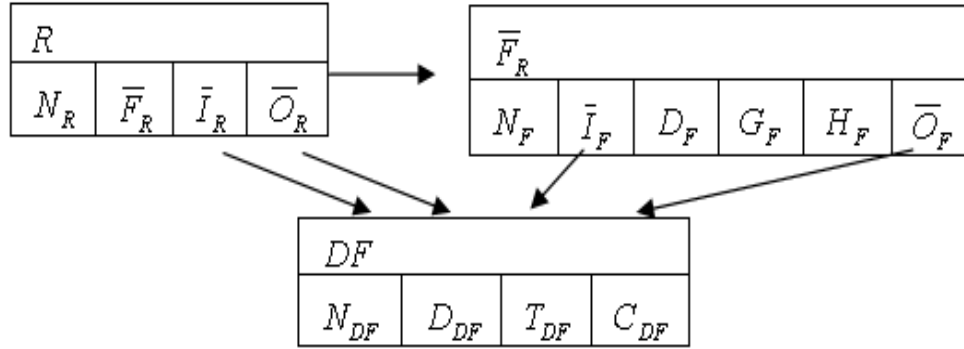


Figure 3: Frame network

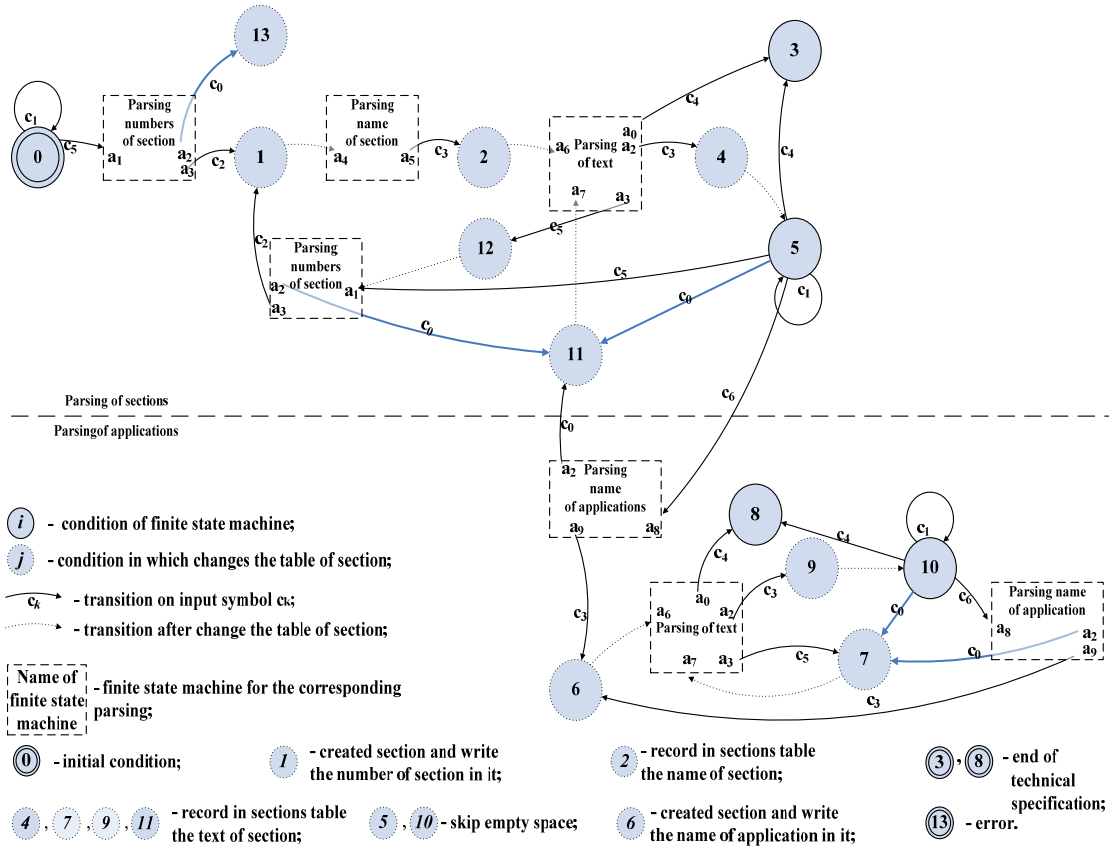


Figure 4: Finite state machine parsing of the text top-level technical specification

Preliminary text processing is necessary to share of a technical specification on separate lexemes. The incoming information of a subsystem is the text of a technical specification in the limited natural language, the target information - tables of sections, sentences and lexemes of a considered technical specification. Results can be submitted both as corresponding tables, and as a tree of sections.

Already after the first stage work not with the text of a technical specification, but with its parts submitted on sections is made. On a course of work of a technical specification shares all over again on more and more fine sections, then on separate sentences (with preservation of sections structure) and lexemes with the instruction of an accessory to sentences.

The input symbols of a finite state machine:  $c_1$  - empty space,  $c_2$  - space,  $c_3$  - a new line,  $c_4$  - the end of the text,  $c_5$  - '1'..'9',  $c_6$  - 'Π',  $c_0$  - any other lexemes. Intermediate condition of finite state machine:  $a_1$  - start parsing

section number,  $a_2$  - a sequence of lexemes - text,  $a_3$  - a sequence of lexemes - numbering,  $a_4$  - start parsing the section name,  $a_5$  - a sequence of lexemes - the section name,  $a_6$  - start parsing the text of section or an application,  $a_7$  - a sequence of lexemes - the continuation of the text section or application,  $a_8$  - start parsing the application name,  $a_9$  - a sequence of lexemes - the name of the application,  $a_0$  - the end of technical specification.

In the course of a finite state machine lexemes acting on its entrance, collect in the buffer. In certain conditions, finite state machine the recording of the current contents of the buffer in one of the tables, after which the buffer is emptied. Work of finite state machine proceeds up to achievement of a final condition.

The output of the algorithm preliminary text processing of the text formed a set of tables: sections, sentences and lexemes. After this table obtained are fed to the algorithm of semantic analysis (Fig. 5).

**General algorithm of semantic analysis Algorithm for constructing a tree of linguistic variables  $\beta_{k,i}$**

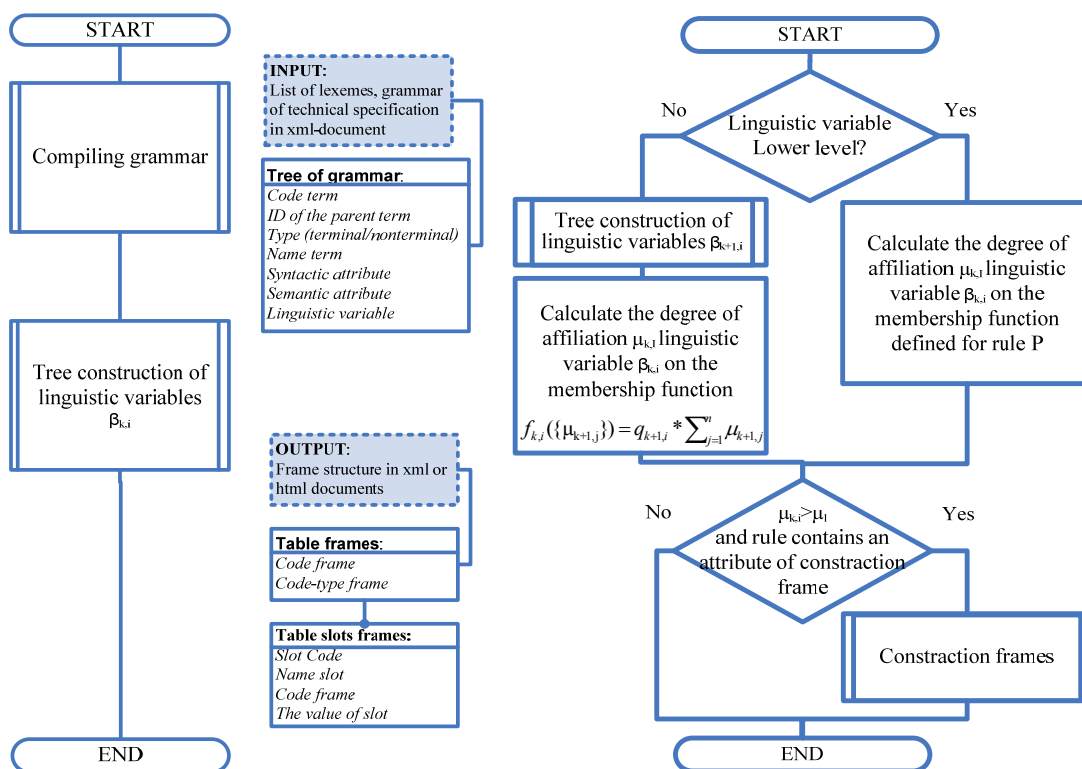


Fig. 5. Algorithm for semantic text analysis of a technical specification

The semantic analysis of a text is made on the basis of the developed grammar of text of a technical specification.

Rules of top level serve for analysis of sections of top level. Rules for analysis of sections consist of two parts: the first part serves for analysis of a section name; the second part serves for analysis of a text contents in section. Symbols of the given grammar possess syntactic attributes. In attributes of non-terminal symbols names of frames or names of slots in which the information received during the further analysis should be placed are specified. Syntactic attributes of text can be in addition specified in attributes of terminal symbols. Comparison of words at analysis is made in view of their morphology. During analysis the syntactic and morphological analysis are made only in the event that there is such necessity that time of performance of semantic analysis is considerably reduced.

Let's consider a fragment of the developed attribute grammar submitted in a xml-format:

```

... <global-rule id="Section42" comment = "Section 4.2. Requirements to functional characteristics">
<rule><ruleref uri="#Section42Name"/><ruleref uri="#Section42x"/></rule></global-rule>
<global-rule id="Section42Name" sectionPart="Name" comment= "Heading of the unit 4.2."><rule><clause
clauseType="UNCERTAIN"/><rule type="or"><words contains="Functions"/> <words contains= " functional
characteristics "/> </rule></rule></global-rule>
<global-rule id="Section42x" frame= "FunctionFrame" frameSlot="Function" comment="Function"><rule> <ruleref
uri="#Section42xName" /><ruleref uri="#Section42xContent" /> </rule></global-rule>
<global-rule id="Section42xContent" sectionPart="Content" comment="Inputs and outputs of
function"><rule><ruleref uri= "#Section42xInputs" minOccurs="0"/><ruleref uri="#Section42xOutputs"
minOccurs="0"/></rule></global-rule>
<global-rule id="Section42xInputs" comment="Inputs of function">
<rule><sentence/><clause/><rule type="or"><words contains="Inputs"/> <words contains="entrance
data"/></rule><ruleref uri="#Input" maxOccurs="unbounded"/></rule></global-rule> ...

```

Semantic analysis is based on the developed fuzzy attribute grammar over the frame structure of text of a TS:

1. Each linguistic variable of a technical specification being reviewed, to result in the linguistic tree, end-vertices are fuzzy variables.
2. Fuzzy variables in the final vertices of the tree is assigned their meaning and then using a system of rules P and the corresponding membership functions  $f_{k,i}$  is determined by the meaning of the linguistic variable corresponding to the left side of the rule.

Rules of the upper levels are used to parse sections of the upper level. The rules for parsing section consists of two parts: the first part is to parse the title of the section, the second part is to parse the text content section.

For some linguistic variable  $\beta_{k,i}$  value of membership function:  $\mu_{k,i} = f_{k,i}(\mu_{k+1,1}, \mu_{k+1,2}, \dots, \mu_{k+1,n})$ , where the specific value  $\mu_{k,i}$  - degree of linguistic affiliation variable  $\beta_{k,i}$ . Initially, we say that all the linguistic variables of the lower level make the same contribution to the value of membership function, so you can say that the membership function of linguistic variable  $\beta_{k,i}$ :

$$f_{k,i}(\{\mu_{k+1,j}\}) = q_{k+1,i} * \sum_{j=1}^n \mu_{k+1,j}$$

where  $\mu_{k+1,j}$  - degree of linguistic affiliation variable  $\beta_{k,i}$ ;  $q_{k+1,i} = 1/n$  - contribution of degrees of linguistic affiliation variables in the value of membership function. At the lower level membership function are defined.

Calculated  $\mu_{k,i}$  compared with  $\mu_i$ , which is the limiting value of the degree of affiliation. If  $\mu_{k,i} > \mu_i$ , and the rules specified syntactic or semantic attributes, then creates frames and slots, which can hold the text of the linguistic variable.

3. Then the tree of linguistic variables cutting back so calculated linguistic variables were end-vertices of the remaining subtree.

This process is repeated until there is no sense to calculate the linguistic variable corresponding to the root of the source tree. The main purpose of this procedure is to associate the meaning of a linguistic variable with the meaning of its fuzzy variables by fuzzy attribute grammar above frame structure of a technical specification.

During parsing syntactic and morphological analysis is only done if there is a need, which significantly reduces the run-time semantic analysis. If the rules of grammar meets terminal with syntactic attribute, then run the parsing mechanism of semantic analysis for the current sentences [2]. After creating a tree of linguistic variables begin construction framing descriptions of a technical specification. It uses information about the frames and the names of slots, which is contained in the attributes of grammar symbols.

The resulting frame structure contains significant information about the system: information about the inputs and outputs of the system, functions and limitations. For each function is also provided inputs and outputs. This allows on the basis of frame structure to obtain a Data Flow Diagrams, which is described in the technical specifications. Algorithms for creating frames carries out construction and ordering the column of data flows, and also creation the figures of data flow diagrams in Microsoft Office Visio (Fig. 6).

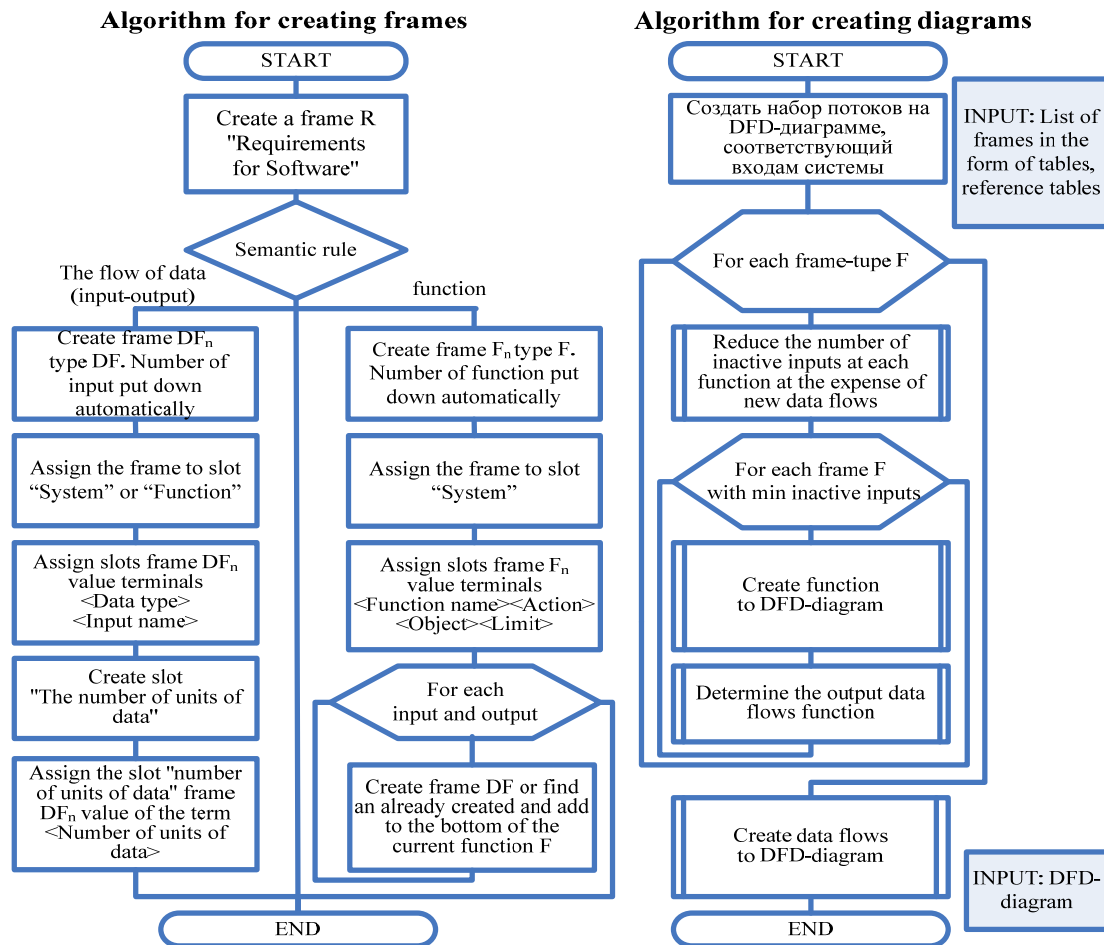


Fig. 6. Algorithms for creating frames and construction of Data Flow Diagrams

For construction of data flows it is prospected of functions inputs conterminous to system inputs. Then functions on which all inputs data act, are located on the one level of diagram. Their inputs incorporate to system inputs. Further it is prospected functions which inputs coincide with outputs of functions received on the previous step. They are located on the following level, their inputs incorporate to outputs of the previous levels functions and with system inputs.

Work of algorithm proceeds until all functions will not be placed on the diagram. After that connection of function outputs with necessary system outputs is made.

**Computer-Aided System Of Semantic Text Analysis Of A Technical Specification**

The computer-aided system of semantic text analysis of a technical specification consist of the following subsystems: preliminary text processing, the syntactic and semantic analysis and construction of software models, storage of documents and interface (see Figure 7).

The computer-aided system of semantic text analysis of a technical specification is developed on Microsoft .NET Framework 2.0 platform (language of development C#) using integrated development environment Visual Studio 2005. Tables are stored in XML, and their visual representation is possible using XSL-transformation. Obtained in the semantic analysis of the framing description is also stored in XML. Building a data flow diagram by means of interaction with the program MS Visio.

**Scientific Novelty**

Scientific novelty consists in the following: a technique of text analysis of a technical specification at the initial stages of software engineering, including semantic model of text of a technical specification, transformation matter of text into the frame structure and construction of model of the software on its basis are developed.

1. New semantic model of text of a technical specification is represented as a fuzzy extended attribute grammar over frame structure, containing syntactic, semantic attributes and linguistic variables.

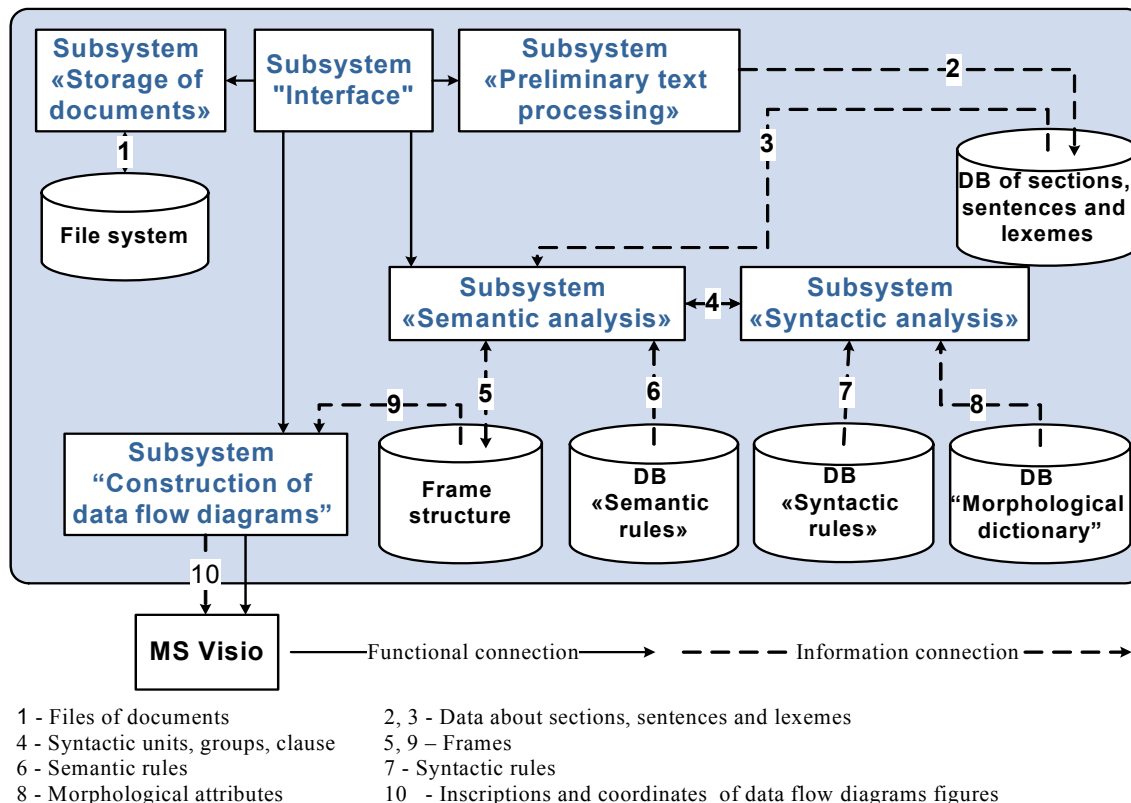


Figure 7: Architecture of computer-aided system of semantic text analysis of a technical specification

2. Proposed and developed methods and algorithms designed to transform the source text of a technical specification developed in a frame structure, which is a description of the requirements to the software.

3. A method of constructing models of the software described in the technical specification in the form of data flow diagram is developed.



---

## Practical Value

---

Practical value of work is that as a result of development and introduction of a suggested technique quality of software engineering raises due to automation of routine work of the person on extraction of helpful information from standard documents and to displaying it as software models. Thus, developed an automated system improves to increase efficiency of software design at the initial stage by reducing the time working on technical specifications and increase the quality of the result. Software designing differs from designing in other areas of a science and technics a little, therefore it is possible to expand results of the given work for application in other areas of human knowledge. Thus, opening prospects raise a urgency of the given work.

---

## Conclusions and Future Work

---

The result of this work the following results:

1. The analysis of the process of designing software, the existing models and methods for word processing is carried out, show the importance of the specification of projects and the importance of the analysis stage in the process of software development. Justified necessity of automating the initial stages of software design and, in particular, the semantic analysis of text specification to identify the functional structure of the system described in the specifications.
2. Developed technique of text analysis specification, designed for text processing in the early stages of software design and containing the formalisms necessary for representing the semantics of the software requirements at the early stages: the semantic model of the text specification, frame structure and a formal model. The technique of analysis involves three stages: semantic processing of text, creating frame structure and the creation of a model of software as a data flow diagram of the system described in the specifications.
3. Analyzed the stylistic features of the text specification, based on which developed a semantic model of the text specification, is an extended fuzzy attribute grammar over a frame structure containing syntactic, semantic attributes and linguistic variables. A frame structure, which is an internal representation requirements to the software and allows the automated system to be universal with respect to the natural language user-designer.
4. Proposed algorithms of semantic analysis of text and technical specification: a preliminary text processing, syntactic and semantic analysis and modeling software. preliminary text processing by using the apparatus of finite state mashine, which results are generated tables of section, sentences and lexemes. Semantic analysis of text is based on the developed fuzzy attribute grammar. Developed algorithmic operations to calculate the meaning of linguistic variable and the construction of fuzzy inference. Based on the tree of linguistic variables and semantic attributes created frame description of the system and implemented the construction of a model of software as a data flow diagram.
5. Developed formalisms, methods and algorithms are implemented in the form of automation systems initial stage of software design "SemantikaTS" platform Microsoft. NET Framework 2.0 (development language C #) using a visual programming environment Visual Studio 2005. Building a data flow diagram implemented in MS Visio.

---

## Bibliography

---

Tools Development For Computer Aided Software Engineering Based On Technical Specification`s Text Analysis / A.Zaboleeva-Zotova, Y.Orlova // Interactive Systems And Technologies: The Problems Of Human-Computer Interaction: Proc. of the Int. Conf., Ulyanovsk.

- Zaboleeva-Zotova, A. Computer Support of Symantic Text Analysis of a Technical Specification on Designing Software / A.Zaboleeva-Zotova, Y.Orlova // Intelligent Processing : suppl. to Int. Journal "Information Technologies and Knowledge" Vol. 3. - 2009. - Int. Book Series "Information Science & Computing", № 9. - C. 29-35.
- Zaboleeva-Zotova, A. Formalization of text analysis of a technical specification / A.Zaboleeva-Zotova, Y.Orlova // Congress on intelligent systems and information technologies (AIS-IT'09), Divnomorskoe, Russia, September, 3-10 : proc. - M., 2009. - Vol. 4. - C. 62
- Zaboleeva-Zotova, A. Automation of procedures of the semantic text analysis of a technical specification / A.Zaboleeva-Zotova, Y.Orlova // Congress on intelligent systems (INTELS'2008). - M., 2008. - P. 245-248.
- Zaboleeva-Zotova, A. Computer-aided System of Semantic Text Analysis of a Technical Specification / A.Zaboleeva-Zotova, Y.Orlova // Advanced Research in Artificial Intelligence: suppl. to Int. Journal "Information Technologies and Knowledge". - 2008. - Vol. 2, [Int. Book Series "Inform. Science & Comput."; № 2]. - P. 139-145.
- Orlova, Y Computer-aided system of the semantic text analysis of a technical specification / Y.Orlova // Distance Learning - Learning Environment of the XXI century: VI Intern. science method. Conf., Minsk, 22-23 November 2007 / Belorus. state. Univ of Informatics and Radioelectronics. - Minsk, 2007. - P.127-129.
- Panchenko, D. Implementation of Genetic Algorithms for Transit Points Arrangement / D. Panchenko, M. Scherbakov / Information Technologies & Knowledge: suppl. "Information Science and Computing". No. 9. - 2009. - Vol. 3. - C. 129-131. - Engl.
- Kamaev V., Shcherbakov, M., Skorobogatchenko D. Automated prediction of transport operating conditions of the roads. Journal of Computer and Information Technologies, № 4, 2004, Moscow: "Engineering", pp 2-6.
- Kamaev V., Shcherbakov M. A neural approach, identification of complex systems, Journal of Computer and Information Technologies, № 3, 2004, Moscow: Mashinostroenie, S. 20-24.
- Rosaliev, V. The model of emotion in human speech / V/Rosalia // Proceedings VolgGTU. A series of "Actual problems of management, computing and informatics in technical systems": Hi. Sat scientific. Art. / VolgGTU. - Volgograd, 2007. - Vol.3, № 9. - P.62-65.
- Rosalie, V. Background, and prospects of creating an automated pattern recognition system emotional speech / V. Rosalia // Proceedings VolgGTU. A series of "Actual problems of management, computing and informatics in technical systems": Hi. Sat scientific. Art. / VolgGTU. - Volgograd, 2008. - Issue 4, № 2. - P.58-61.

---

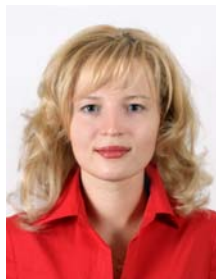
### Authors' Information

---



**Alla V. Zaboleeva-Zotova** – PhD, professor; CAD department, Volgograd State Technical University, Lenin av., 28, Volgograd, Russia; e-mail: zabzot@vstu.ru

*Major Fields of Scientific Research:*



**Yulia A. Orlova** – PhD; CAD department, Volgograd State Technical University, Lenin av., 28, Volgograd, Russia; e-mail: yulia.orlova@gmail.com

*Major Fields of Scientific Research:*

---

## LINGUISTICS RESEARCH AND ANALYSIS OF THE BULGARIAN FOLKLORE. EXPERIMENTAL IMPLEMENTATION OF LINGUISTIC COMPONENTS IN BULGARIAN FOLKLORE DIGITAL LIBRARY

**Konstantin Rangochev, Maxim Goynov,  
Desislava Paneva-Marinova, Detelin Luchev**

**Abstract:** *The observation of the lexical structure of the Bulgarian folklore is very important task for different science domains such as folkloristic, ethnology, linguistics, computational linguistics, Bulgarian language history, etc. Until today, such a linguistic analysis hasn't been made; it is unclear what is the lexical structure of Bulgarian folklore works. First attempt for computational lexical analysis of the Bulgarian folklore and its constituents is made during the "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage" 1. During the project the Bulgarian folklore digital library (BFDL) is designed and developed. In its structure it is implemented linguistic components, whose aim is the realization of different types of analysis of folk objects from a text media type. Thus, we lay the foundation of the linguistic analysis services in digital libraries aiding the research of kinds, number and frequency of the lexical units that constitute various folk objects. This paper presents basic types of dictionaries needed to carry out such linguistic analysis. It describes the BFDL Linguistics Search in sets of folklore objects of text media type and a linguistic component for frequency analysis of the folklore vocabulary. Finally, a project for implementation of a dictionary - concordances of songs, prose, interviews, etc. is outlined.*

**Keywords:** *multimedia digital libraries, systems issues, user issues, online information services*

**ACM Classification Keywords:** *H.3.5 Online Information Services – Web-based services, H.3.7 Digital Libraries – Collection, Dissemination, System issues.*

---

### Introduction

The main component of the linguistic research of the Bulgarian folklore is the analysis of its lexical structure: How many and what token it contains? Is there and what is the domination or the lack of some groups of tokens, etc. Until today, such a linguistic analysis hasn't been made; it is unclear what is the lexical structure of Bulgarian folklore works. With a few exception (for Bulgarian heroic epic [Rangochev, 1994] and for "Veda Slovena", <http://www.bultreebank.org/veda/index.html>) lexical analysis for the Bulgarian folklore and its constituents is missing, the regional characteristics of the folklore lexical structure is unknown. Unfortunately, in 2010 the Bulgarian linguistics, folklore, ethnology, etc. cannot answer the question what are the lexical components of Bulgarian folklore (number, frequency, word forms, etc.) and so far, this type of research is carried out systematically and with a purpose.

---

<sup>1</sup> The "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage" is a national research project of the Institute of Mathematics and Informatics, supported by National Science Fund of the Bulgarian Ministry of Education and Science under grant No IO-03/2006. Its main goal is to build a multimedia digital library with a set of various objects/collections (homogeneous and heterogeneous), selected from the fund of the Institute for Folklore of the Bulgarian Academy of Science. This research aims to correspond to the European and world requirements for such activities, and to be consistent with the specifics of the presented artefacts.

In the project, named "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage" (FolkKnow) [Paneva-Marnova et al., 2009] [Luhev et al., '08b] the attention was directed to these researches in order to enrich both the content and functionality of the developed multimedia digital library of Bulgarian folklore (also called Bulgarian Folklore Digital Library or BFDL, FDL) [Rangochev et al., '07a] [Rangochev et al., '08]. Thus we aim to expand the target group of potential users of the library, covering not only those who are interested in Bulgarian folk music, but also narrow specialists in different fields of humanities (folklore, ethnology, linguistics, text linguistics, structural linguistics, *etc.*). Digital library with similar services are presented at [Pavlov and Paneva, 2007] [Pavlova-Draganova et al., 2007a] [Pavlov et al., 2006].

The Bulgarian folklore digital library has a flexible structure that involves the addition of linguistic components, whose main task is the realization of different types of analysis of folk objects from a text media type. This article presents the basic types of dictionaries needed to carry out such linguistic analysis. It describes the BDFL Linguistics Search in sets of folklore objects of text media type and a linguistic component for frequency analysis of the folklore vocabulary. Finally, a project for implementation of a dictionary - concordances of songs, prose, interviews, *etc.* is outlined.

---

### Frequency Dictionaries and Concordance Dictionaries

---

The frequency dictionary presents the frequency of the lexemes in a definite corpus of texts. It is considered that the facts in one frequency dictionary are reliable enough if there are minimum 20 000 lexical units in it. The frequency dictionaries gave versatile information: presence/ absence of definite lexemes or group of lexemes in comparison with a standard frequency dictionary of the Bulgarian speech [Radovanova, 1968]; frequency of verbs (the so called "verb temperature" [Gerganov et al., 1978] (for the Bulgarian speech at least 21 % verbs in the examined corpus of texts); investigating of the paradigmatic relations in the vocabulary of the text corpus (river-stream- brook- rill...). The domination of group lexemes and respectively small number or absence of other group reveals the constituent characteristics of the text type and its originators.

- A general frequency dictionary – it contains the all lexical units which are in the BFDL (songs, proverb and descriptions of the rites...);
- A regional frequency dictionary – it contains all the text units which come of a definite folklore region or of a concrete settlement (if there are enough texts). Practically, this is a dialect dictionary of the region/ settlement as far as the folklore regions coincides with the dialect areas.
- A functional frequency dictionary – it contains all the text units which have identical functions: descriptions of the rites, various types of songs, narratives *etc.* This kind of dictionary would describe some genre specifics of the different parts of the Bulgarian folklore;
- Another dictionary – by user's wish.

The advantage of creating of frequency dictionaries is the possibility to make comparisons between the different types of texts and it can be also followed the tendencies in the dynamics of the lexis – presence/ absence of various group of lexemes, *etc.*

The following table illustrates the comparison of the Bulgarian folklore and spoken languages based on data available in frequency dictionaries.

Rank list	
Bulgarian spoken language <sup>1</sup>	Bulgarian heroic ерос <sup>2</sup>
1. съм – 4 041	1. съм – 1342
2. и – 3764	2. да – 1 247
3. да – 3 148	3. си – 548
4. аз – 2 433	4. Марко – 1 036
5. той – 2 288	5. се – 828
6. не – 1 956	6. на – 801
7. се – 1 928	7. и – 796
8. този – 1 701	8. па – 657
9. на – 1 669	9. у – 582
10. ти – 1 249	10. я – 553
11. ще – 1 183	11. та – 526
12. един – 1 131	12. не – 412
13. в – 1 099	13. юнак – 396
14. си – 1065	14. го – 338
15. казвам – 1 045	15. му – 320
16. тя – 1044	16. че – 318
17. викам – 1 031	17. а – 286
18. те – 1014	18. кон – 276
19. какъв – 938	19. от – 272
20. за – 913	20. ми – 233
21. че – 874	21. ти – 225
22. с – 809	22. що – 222
23. имам – 768	23. по – 218
24. така – 742	24. добър – 201
25. от – 731	25. три – 201

Table 1: Comparison of the Bulgarian folklore and spoken languages.

Concordance dictionaries are these which show the lexeme with/ in her context – it is present the previous one (or more than one) lexeme and the following lexeme according to the examined lexeme. Example: “Fifty heroes are drinking wine” – the underlined lexeme is the examined and the lexemes in italic are her context. Of course, about the songs this could be concordance dictionary of their verses, about the narrative texts (descriptions of the rituals, etc.) – sentences in which they are contained (from point to point...). The creating and using of

<sup>1</sup> The frequency dictionary is made of texts of the Bulgarian spoken language and the corpus contains 100000 lexemes [Nikolova, 1987].

<sup>2</sup> The frequency dictionary is made of 100 song from [Romanska, 1971] and the texts of the songs contains 7871 verses while there are in it 40042 lexemes.

concordance dictionaries of the texts from BFDL would give good possibilities for folklorists and ethnologists to solve a series of problematic areas as presence/ absence of formulas in the folklore songs and epics, the structure of the folklore text, *etc.*

---

### **Linguistic Search in Set of Folklore Objects of Text Media Type**

---

This type of a search has for aim to supply the needs of linguists and explorers of the Bulgarian dialects for researching the language of the “folklore song”, the “folklore prose”, *etc.* The variants for searching of folklore objects of text media type are the following:

- Search of a word in the different types of dictionaries;
- Search of two or more words – searching of verbal formulas in the folklore lexis: “Drinking wine”, “Marko seated”.
- Search of a group of words – this has for aim to investigate the paradigmatic relations in the folklore lexis (river- stream- brook- rill...) – for example, the frequency of the lexemes, verses/ sentences in which they are, number, numbering in the song, *etc.* of the verses/ sentences.
- Search for a root of a word for studying the folklore word-formation: ‘drink’ (I am drinking, I have drunk, they have drunk...).

---

### **A Frequency Dictionary in BFDL. A Project for Concordance Dictionary about Songs, Prose, Interviews, etc. in BFDL**

---

In the process of the primary testing of BFDL come into being the necessity of insurance of resources for linguistic analysis of the folklore knowledge. For this aim it was projected and worked out a frequency dictionary with the following functional specification:

- Linguistic analysis of the available multitude of folklore objects of text media type in BFDB;
- Determination of the frequency of meeting the lexemes in text folklore objects;
- Creating of lists of the lexemes,
  - in frequency order
  - in alphabetical order.
- Taking the number of the lexical units;
- Taking the number of the repeats of the lexical units.

Figure 1 depicts the sequence of actions that has to be executed in order to be generated a frequency dictionary. Standard step is the passing through BFDL search service and its sub-functions: 1) user searches by some criteria; 1.1) service performs search in metadata repository, 1.1.1) service gets media data for the found objects, 1.1.1.1) service returns all found media objects by the search criteria, and 1.1.1.1.1) result sent to user. When the result set is generated the user could choose to generate a functional dictionary (step 2). Dictionary generation is performed and the result is shown by frequency or alphabetically.

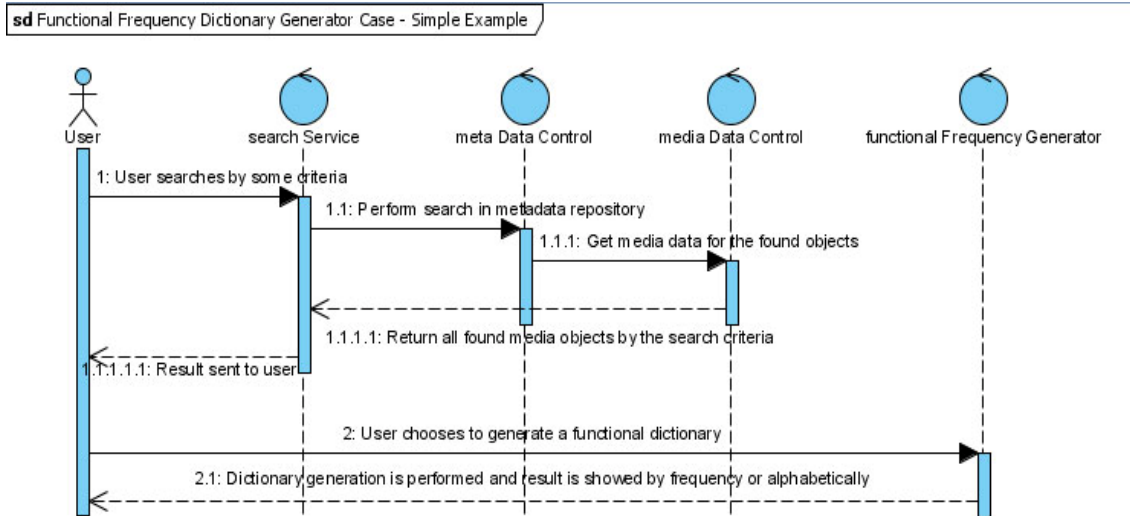


Figure 1: Sequence Diagram

Figure 2 depicts analysis class diagram for the BFDL linguistic component.

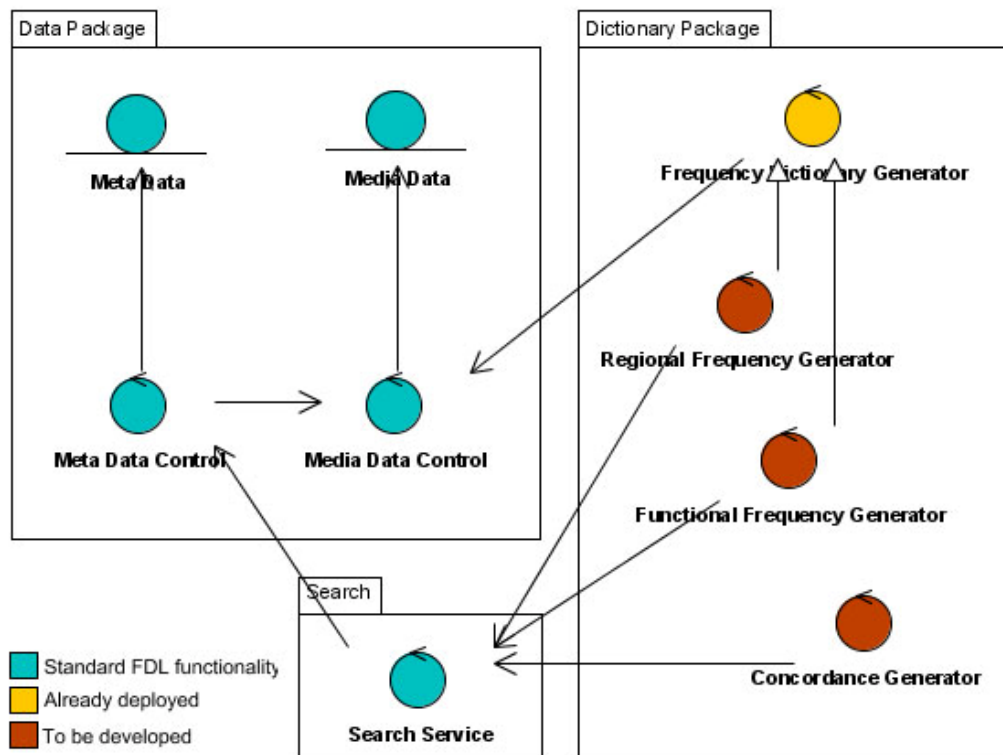


Figure 2: Analysis class diagram

The diagram shows the relations between the data package, the dictionary package and the search service. In the dictionary package there are clearly illustrated different types of generators for frequency dictionary, regional frequency dictionary, functional frequency dictionary and dictionary-concordance.

---

### Acknowledgements

This work is supported by National Science Fund of the Bulgarian Ministry of Education and Science under grant No IO-03-03/2006 "Development of Digital Libraries and Information Portal with Virtual Exposition "Bulgarian Folklore Heritage"" from the project "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage".

---

### Bibliography

- [Rangochev et al., '07a] Rangochev K., D. Paneva, D. Luchev. Bulgarian Folklore Digital Library, In the Proceedings of the International Conference on Mathematical and Computational Linguistics „30 years Department of Mathematical Linguistics”, 6 July, 2007, Sofia, Bulgaria, pp. 119-124.
- [Rangochev et al., '08] Rangochev, K., D. Paneva, D. Luchev, Data and Functionality Management in a Folklore Digital Library, In the Proceedings of the International Conference - Slovo: Towards a Digital Library of South Slavic Manuscripts, 21-26 February, 2008, Sofia, "Boian Penev" Publishing Centre, pp. 246 – 250.
- [Paneva-Marnova et al., 2009] Paneva-Marnova, D., R. Pavlov, K. Rangochev, D. Luchev, M. Goynov (2009), Toward an Innovative Presentation and Creative Usage of the Bulgarian Folklore Wealth, International Journal „Information Technologies & Knowledge”, vol. 3, 2009 (in print)
- [Luchev et al., '08b] Luchev D., D. Paneva, K. Rangochev, Approaches for Utilization of the Semantic Web Technologies for Semantic Presentation of the Bulgarian Folklore Heritage. In the Proceedings of the national conference "Bulgarian Museums in the circumstances of the country membership in the European Union". Sliven, 2008, pp.271-281.
- [Gerganov et al., 1978] Gerganov, E., A. Mateeva, Experimental Research of the Frequency of the Bulgarian Language, In the proceedings "Contemporary problems of the native language education, Sofia, 1978
- [Rangochev, 1994] Rangochev, K., "Structural particularities of the epic text (using material of the Bulgarian heroic epos)", Канд. дис., СУ „Св. Кл. Охридски”, София.
- [Radovanova, 1968] Radovanova, V., „Representative frequency dictionary of text with length 500 000 tokens”, Master thesis, University of Sofia 'St. Kl. Ohridski", Sofia, 1968.
- [Nikolova, 1987] Nikolova, Cv., A frequency dictionary of the Bulgarian spoken language. Sofia.
- [Romanska, 1971] Romanska Cv. (Ed.), Сборник за народни умотворения, col.53, "Bulgarian heroic epos", Sofia, 1971
- [Pavlov and Paneva, 2007] Pavlov R., D. Paneva. Toward Ubiquitous Learning Application of Digital Libraries with Multimedia Content, International Journal "Cybernetics and Information Technologies", vol. 6 (2007), № 3, pp. 51-62.
- [Pavlova-Draganova et al., 2007a] Pavlova-Draganova L., V. Georgiev, L. Draganov. Virtual Encyclopaedia of Bulgarian Iconography, Information Technologies and Knowledge, vol.1 (2007), №3, pp. 267-271.
- [Pavlov et al., 2006] Pavlov R., L. Pavlova-Draganova, L. Draganov, D. Paneva, e-Presentation of East-Christian Icon Art, In the Proceedings of the Open Workshop "Semantic Web and Knowledge Technologies Applications", Varna, Bulgaria, 12 September, 2006, pp. 42-48.



---

**Authors' Information**

---



**Konstantin Rangochev** – PhD in Philology, Assistant Professor, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: [krangochev@yahoo.com](mailto:krangochev@yahoo.com)

Major Fields of Scientific Research: Ethnology, Folklore studies, Culture Anthropology, Linguistics, Computational Linguistics, Digital Libraries.



**Maxim Goynov** – Programmer, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: [maxfm@abv.bg](mailto:maxfm@abv.bg)

Major Fields of Scientific Research: Multimedia Digital Libraries and Applications.



**Desislava Paneva-Marinova** – PhD in Informatics, Assistant Professor, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: [dessi@cc.bas.bg](mailto:dessi@cc.bas.bg)

Major Fields of Scientific Research: Multimedia Digital Libraries, Personalization and Content Adaptivity, eLearning Systems and Standards, Knowledge Technologies and Applications.



**Detelin Luchev** – PhD in Ethnology, MA in Informatics, MA in History; Research Fellow at Ethnographic Institute and Museum, BAS; 6A, Moskovska Str., Sofia 1000, Bulgaria; e-mail: [luchev\\_detelin@abv.bg](mailto:luchev_detelin@abv.bg)

Major Fields of Scientific Research: Communities and Identities, Ethno-statistics, Museums and Archives, Digital Libraries.

## NATURAL INTERFACE TO ELECTION DATA

Elena Long, Vladimir Lovitskii, Michael Thrasher

**Abstract:** *Modern technology has the facility to empower citizens by providing easy access to vital electoral information. The majority of such users simply want to use the information; they do not wish to become embroiled in technological details that provide that access; the technology is a means to an end and if allows to obscure the real purpose (to access information) it represents a cost not a benefit. Much of the potential benefit is therefore lost unless a simple and consistent interface can be provided which shields the user from the complexity of the underlying data system retrieval and should be natural enough to be used without training. Currently there are limited tools and information available online where end users can view and interrogate electoral data. The main purpose of our paper is to report upon developments that seek to provide an easy to use interface for users to obtain information regarding the results of general elections within the United Kingdom (UK).*

**Keywords:** *natural interface, natural language processing, database accessing, SQL-query, production rules*

**ACM Classification Keywords:** *I.2 Artificial intelligence: I.2.7 Natural Language Processing: Text analysis.*

---

### Introduction

---

Following the rapid development of both computer and communications technologies, our society now has the potential to access vast amounts of information almost instantaneously on a world-wide basis. One of the major obstacles to achieve it is to realising the potential for wealth and knowledge creation that information represents is the means of simple access by naïve users to relevant information locked in possibly complex data structures. Individuals are not expected to know in detail what information is required, or where it might be found and certainly does not know about data structures. In respect of electoral data, for example, the citizen simply requires information that is relevant to his or her particular area of interest - and no more.

This paper represents results of our further research in the natural language interface creation to database (DB) [V.A.Lovitskii and K.Wittamore, 1997; Guy Francis *et al.*, 2007; Elena Long *et al.*, 2009]. The data source addressed here is the DB with the results of 2005 UK General Election. The result is our current vision of “*natural interface*” has been implemented (<http://141.163.170.152:8080/NITED/NITEDJSP.jsp>) as a Web application named **NITED (Natural Interface To Election Data)** where a user can see essential election data online. The aim of design is that the application must offer simple, intuitive and responsive user interfaces that allows users to achieve their objectives regarding information retrieval with minimum effort and time..

Despite the intuitive appeal of a natural language interface, some researchers have argued that a language like English has too many ambiguities to be useful for communicating with computers. Indeed, there is little experimental data supporting the efficacy of a natural language interface, and the few studies that have compared natural language interfaces to other styles of interface have been generally negative towards the former.

Indeed, two major obstacles lie in the way of achieving the ultimate goal of support for arbitrary natural language queries. First, automatically understanding natural language (both syntactically and semantically) remains an open research problem. Second, even if there were a perfect parser that could fully understand any arbitrary natural language query, translating the parsed natural language query into a correct formal query still remains an issue since this translation requires mapping the understanding of intent into a specific database schema.

---

Natural language is not only very often ambiguous but is dependent on a great deal of world knowledge. In order to implement a working natural language system one must usually restrict it to cover only a limited subset of the vocabulary and syntax of a full natural language. This allows ambiguity to be reduced and processing time to be kept within reasonable bounds. In order for it still to be considered a natural language interface, most of the positive traits of a general natural language interface would have to be maintained. To retain the properties of ease of use and ease of remembering, the limitations of the system must somehow be conveyed to the user without requiring them to learn the rules explicitly.

Natural language interfaces, if they are the only form of interaction, do not take advantage of the capabilities of the computer -- those strategies that work in human-human communication are probably not best suited to human-computer interactions, where the computer can display information many times faster than people can enter commands

The principal purpose of our paper is to offer the natural (versus natural language) user interface which makes it easy, efficient, and enjoyable to operate NITED in a way which produces the desired result. This generally means that the user is required to provide minimal input to achieve the desired output, and also that NITED minimizes undesired outputs or data clutter.

Reading this paper will tell you the following:

- Natural user interface.
- Natural user enquiry.
- Help instructions.
- Production rules.
- Natural enquiry to SQL query conversion.

---

## Natural User Interface

---

The natural user interface (NUI) is a key to application usability. NUI is needed when interaction between users and NITED occurs. The goal of interaction between the user and the NITED at the NUI is effective operation and control of the NITED, and feedback from the NITED in desirable for the user format i.e. NUI provides a means of input, allowing the users to ask question, and output, allowing the NITED to reply on user's question.

The design of a NUI affects the amount of effort the user must expend to provide input and to interpret the output of the system, and how much effort is required to learn this. Usability is mainly a characteristic of the NUI, but is also associated with the functionalities of the product and the process to design it. It describes how well the NITED can be used for its intended purpose by its target users with efficiency, effectiveness, and satisfaction, also taking into account the requirements from its context of use. A key property of a good user interface is consistency.

There are three important aspects [[http://en.wikipedia.org/wiki/User\\_interface](http://en.wikipedia.org/wiki/User_interface)] to be taken into account. First, the controls for different features should be presented in a consistent manner so that users can find the controls easily. For example, users find it very difficult to use software when some commands are available through menus, some through icons, and some through right-clicks. A good user interface might provide shortcuts or "synonyms" that provide parallel access to a feature, but users do not have to search multiple sources to find what they're looking for.

Second, the "principle of minimum astonishment" is crucial. Various features should work in similar ways. For example, some features in Adobe Acrobat are "select tool, then select text to which apply." Others are "select text, then apply action to selection.

Third, user interfaces should strive for minimum change version-to-version -- user interfaces must remain upward compatible. For example, the change from the menu bars of Microsoft Office 2003 to the "ribbon" of Microsoft Office 2007 is universally hated by established users, many of whom found it difficult to achieve what had become routinized tasks. The "ribbon" could easily have been "better" in the mid-1990's than the menu interface if writing on a blank slate, but once hundreds of millions of users are familiar with the old interface, the costs of change and adaptation far exceed the benefit of improvement. The vast majority of users viewed this forced change, without a backward-compatibility mode, as unfavorable; more than a few viewed it as verging on malevolence. Re-design should introduce change incrementally such that existing users are not alienated by a revised product.

Good user interface design is about setting and meeting user expectations because the best NUI from a programmer's point of view is not, as a rule, the best from a user's point of view.

We have tried to create a NUI to improve the efficiency, effectiveness, and naturalness of user-NITED interaction by representing, reasoning, and acting on models of the user, domain and tasks. The main part of NUI is a graphical interface, which accepts input via computer keyboard and mouse. The actions are usually performed through direct manipulation of the graphical control elements. The natural way to represent the output for election application domain (EAD) is a table. In the next section we will discuss in detail the input enquiry presentation.

---

### Natural User Enquiry

---

- Over a number of years [Guy Francis *et al.*, 2007; Elena Long *et al.*, 2009] users' natural language enquiries (NLE) have been collected by us in a series of research programmes. Direct observation of users' NLE shows, unsurprisingly, that all users are lazy i.e. they want to achieve the desired result whilst expending minimum effort. They do not want to type in the long NLE such as "*How many votes did the Demanding Honesty in Politics and Whitehall candidate obtain in Dumfriesshire, Clydesdale and Tweeddale?*" This is the natural behaviour of human being in accordance with the **principle of simplicity**, or **Occam's razor principle** (*Occam's (or Ockham's) razor is a principle attributed to the 14th century logician and Franciscan friar; William of Occam. Ockham was the village in the English county of Surrey where he was born*). The principle states that "Everything should be made as simple as possible, but not simpler". Finding a balance between simplicity and sophistication at the input side has been discussed elsewhere [L.Huang *et al.*, 2001].

On the one hand, NLE provides end users with the ability to retrieve data from a DB by asking questions using plain English. But, on the other hand, there are several problems of using NLE:

- The end users are generally unable to describe completely and unambiguously what it is they are looking for at the start of a search. They need to refine their enquiry by giving feedback on the results of initial search e.g. "*I'm looking for a nice city in France for holiday*" (where *Nice* is a city in France but also an adjective in English). Similar ambiguities exist for the UK general election database. For example, the words *Angus, Bath, Corby, ..., Wells* are values of fields *Constituency* and *Surname* in the *General Election data 2005 DB* but are also common nouns and place names. Parsing of such simple NLE is quite complicated and requires powerful knowledge base from system [V.A.Lovitskii and K.Wittamore, 1997].

- It is simply impossible to require that users know the exact values in DB (e.g. name of constituency). For example, if user makes the enquiry: “*Who won the election in Suffolk Central & Ipswich North*”? but instead of using the symbol ‘&’ types in “**and**” NITED will not find the constituency in DB.
- In the case when user simply made a mistake and instead of typing in the desirable constituency *Hereford* in the NLE: “*Who won the election in Hereford*” user entered *Hertford* (it’s **wrong** but at the same time it’s **right** from the NITED point of view because it has the right part of an existing constituency *Hertford & Stortford*), NITED will find the answer for the constituency *Hertford & Stortford*. When user sees the response, he/she realises that constituency was wrong and simply corrects it.
- As a rule a user’s NLE cannot be interpreted by NITED without additional knowledge because the concepts involved in NLE are outside of the EAD. For example, in NLE “*How did the Conservative perform in South West?*” NITED should know the meaning of word “*perform*” regarding the election data, and in the NLE “*Which party won the Aberdeenshire West and Kincardine constituency?*” correctly interprets word “*won*”.
- In conclusion it would be sensible to underline the main problem which hinders the use of NLE the cognitive process of “*understanding*” is itself not understood. First, we must ask: “*What it means to understand a NLE?*” The usual answer to that question is to model its meaning. But this answer just generates another question: “*What does meaning mean?*” The meaning of a NLE depends not only on the things it describes, explicitly and implicitly, but also on both aspects of its causality: “*What caused it to be said*” and “*What result is intended by saying it*”. In other words, the meaning of a NLE depends not only on the sentence itself, but also on the context: **Who** is asking the question, and **How** the question is phrased.

In the result of NLE analysis we decided to distinguish two different types of NUE: (1) NLE Template (**NLET**) and (2) Natural Descriptors Enquiry (**NDE**). Such enquiries permit users to communicate with a DB in a natural way rather than through the medium of formal query languages. Obviously issues in these two NUE are related, and the knowledge needed to deal with them is represented as a set of Production Rules (PR). Let us consider these two types of NUE.

**Natural Language Enquiry Template** combines a list of values to be selected when required and generalization of users’ NLETs. Examples of some Frequently Asked Questions (**FAQ**) are shown below:

- What was the **result** in [constituency]?
- In which **constituency** did [party] achieve its **highest vote**?
- **Who won** the [constituency]?

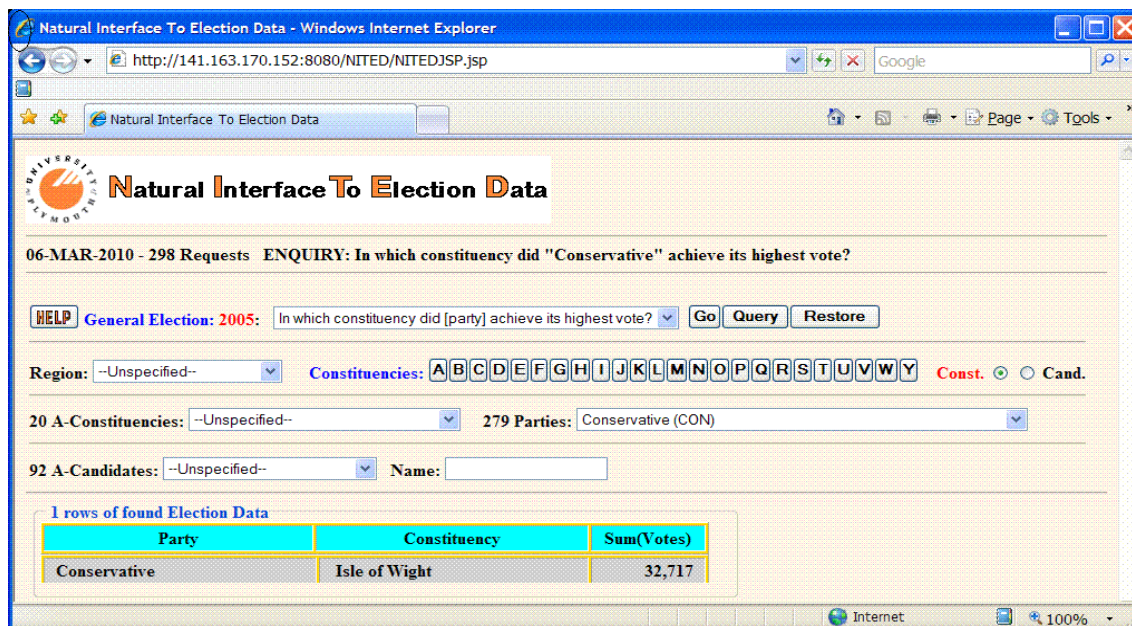


Figure 1. Natural Language Enquiry Template

The initial set of FAQ has been created by export in *EAD* but in the result of activities new NUE have been collected by NITED, analysed, generalized, converted to the NLET. These have then either been added to FAQ, or substituted for the under-used NLET. When the user selects an appropriate NLET with some descriptor in square brackets, selects the corresponding values from the list and click button **Go** the result will be displayed instantly (see Figure 1).

The user can build his/her own enquiries using any combination of the descriptors, each of them represents the corresponding meaningful field of the Election DB (see Figure 2). The definition of “meaningful fields” depends on AD objectives. For the considered EAD is a list of descriptors: {*region, constituency, party, etc.*}. Between descriptors and meaningful fields exist one-to-one attitude. Such attitudes are represented by the production rules (see section below).

Let’s call enquiries using descriptors as a **Natural Descriptors Enquiries (NDE)**. For example, if user wants a list of all the women elected in the South West region simply click the following check boxes: "Party", "Candidate", "Votes", "Sits in Parliament" and radio button "Female". Then select the South West region from the drop down menu of regions. When NDE is ready the user should simply click "Go" button and NITED instantly displays the result (see Figure 2). As user clicks the check boxes and selects the radio buttons NDE appears in the space next to the date above. If user clicks a check box but then change his/her mind the check box should simply be clicked again.

09-MAR-2010 - 302 Requests ENQUIRY: FIND: Party,Candidate,Votes FOR: Region='South West',Sits='Y',Gender='F'

HELP General Election: 2005: --Frequently Asked Questions-- Go Query Restore

Region:  Constituency:  Party:  Candidate:  Sits:  Votes:  SUM:  Max  Min

Region: South West Constituencies: A B C D E F G H I J K L M N O P Q R S T U V W Y Const.  Cand.

47 H-Constituencies: --Unspecified-- 279 Parties: --Unspecified--

92 A-Candidates: --Unspecified-- Name:  Sits in Parliament:  Male  Female

6 rows of found Election Data

Gender	Region	Sits	Party	First Name	Surname	Vote
Female	South West	Yes	Conservative	Angela	Browning	27,838
			Liberal Democrat	Annette	Brooke	22,000
			Labour	Dawn	Primarolo	20,778
			Labour	Valerie	Davey	16,859
			Labour	Linda	Gilroy	15,497
			Labour	Candy	Atherton	14,861

Figure 2. Natural Descriptors Enquiry

## Help Instructions

Help Instructions (HI) in a Web application context means on-screen help. HI are needed for system efficiency and users' satisfaction. Clear HI can significantly reduce the number of disappointed users. Producing clear instructions that really help people is difficult as evidenced by the low-quality instructions encountered in many web applications. If designing HI were easy, there would not be so many poor examples!

Good HI have to take into account the type of users who will presumably use the NITED:

- Users' computer literacy is the basic IT literacy.
- Users should not require a conceptual background before they can use the NITED.
- Users might be absolute beginners or moderately familiar with the subject but they should not be subject matter expert.

Requirements to Help Instructions:

- HI should be short enough but provide sufficient information about the screen function.
- Good HI does not mean that all options should be explained in detail.
- HI should include brief information that is at least sufficient to get started.
- The most frequently used features should be explained.
- Top-level tasks, without much detail about particular fields, should be described.
- Step-by-step worked examples that users can follow should be represented in the .HI.

We tried to meet all of these requirements in the HI for NITED (see Figure 3).



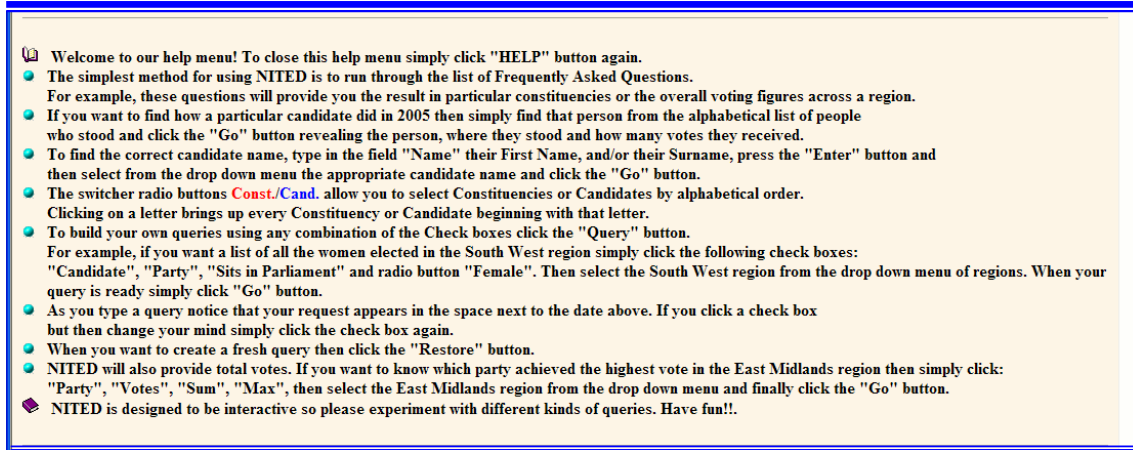


Figure 3. Help Instructions

## Production Rules

At first glance, the NLET is an ideal way to communicate with EAD but in reality there are some problems, which need to be solved to provide lightness of communication. To highlight such problems is enough to consider quite a simple NLET: "Who won an election in [constituency]?" or "How did the [party] perform in [region]?". Without knowing "who is who" and meaning of "won" and "perform" NITED cannot answer such questions. To explain it to NITED the **Production Rules (PR)** need to be involved. Many researchers are investigating what information is needed and how the information needs to be represented in the PR. From our point of view the **Preconditioned PR (PPR)** should be used. The PPR is a quite powerful approach to solve this problem. The subset of PPR in format:

$$\langle \text{Precondition} \rangle \mapsto \langle \text{Antecedent} \rangle \Rightarrow \langle \text{Consequent} \rangle$$

is shown below.

- AD:Election2005  $\mapsto$  who  $\Rightarrow$  candidate;
- AD:Election2005  $\mapsto$  [candidate]:<win $\oplus$ won $\oplus$ highest >  $\Rightarrow$  [SQL]:<MAX(votes)>;
- AD:Athletics  $\mapsto$  [runner]:<win $\oplus$ won>  $\Rightarrow$  [SQL]:<MIN(time)>;
- AD:Athletics  $\mapsto$  [shooter]:<win $\oplus$ won>  $\Rightarrow$  [SQL]:<MAX(distance)>;
- AD:Election2005 & DB:MS Access  $\mapsto$  votes  $\Rightarrow$  [Field]:<gcr\_post\_election\_votes>;
- AD:Election2005 & DB:MS Access  $\mapsto$  candidate  $\Rightarrow$  [Field]:<can\_first\_name, can\_last\_name>;
- AD:Election2005 & DB:Oracle  $\mapsto$  [party]:<win $\oplus$ won $\oplus$ highest>  $\Rightarrow$  [SQL]:<MAX(SUM(votes))>;
- AD:Election2005 & DB:MS Access  $\mapsto$  [party]:<win $\oplus$ won $\oplus$ highest >  $\Rightarrow$  [SQL]:<TOP1, SUM(votes),DESC>;
- AD:Election2005 & DB:MS Access  $\mapsto$  perform  $\Rightarrow$  candidate,votes;

where  $\oplus$  - denotes "exclusive OR". **Precondition** consist of **class<sub>1</sub>:value<sub>1</sub> {& class<sub>2</sub>:value<sub>2</sub>}**. **Antecedent** might be represented by: (i) **single word** (e.g. *who, won, perform, etc.*), (ii) **sequence of words** (e.g. *as soon as,*



create KB, How are you doing, etc.), or (iii) **pair - [context]:<value>**. Context allows one to avoid word ambiguity and thereby distinguish difference between “Candidate won an election” and “Party won an election”. Presentation of **Consequent** is similar to Antecedent structure except (iii). For Consequent pair represents **[descriptor]:<value>**.

Region	Party	First Name	Surname	Vote
South West	Conservative	Christopher	Chope	28,208
		Angela	Browning	27,838
		Michael	Ancram	27,253
		James	Gray	26,282
		Robert	Key	25,961
		Adrian	Flook	25,191
		Charles	Cox	25,013
		Oliver	Letwin	24,763
		Andrew	Murrison	24,749
		Robert	Walter	23,714
		Geoffrey	Clifton-Brown	23,326

Figure 4. Reply to NLET after describing the word “perform” in the PPR

For EAD subset {1, 2, 5, 6, 8, 9} of PPR is used. PPR 3 and 4, in fact, show another meaning of the same word “won” but for a different AD. The PPR 7 shows the simplest way to cover the difference in SQL for different DB. Result of using selected PPR to reply to NLET “How did the [party] perform in [region]?” is shown on Figure 4.

Thus, NLET allows the user to “be lazy” but requires some effort to create the proper set of PPR.

## Natural Enquiry to SQL Query Conversion

Two types of NUE have been considered. The NDE does not require great effort to be converted to the corresponding SQL query. Only NLET need some parsing. The mechanism of NLET parsing is very simple: “eliminating the unnecessary until only the necessary remains”. Several steps involved in NLET processing.

- NITED takes the NLET as a character sequence and converts the original NLET to a *skeleton* by noisy (non-searchable) words elimination. As a result of such conversion the NLET will contain only **meaningful** words: let’s call the word meaningful if it represents DB field descriptor or DB field value.
- EAD is represented by DB. DB **meaningful** fields (i.e. they don’t represent primary or foreign keys) contain election data. Each meaningful fields has a list of descriptors. Between descriptors and meaningful fields exists an one-to-one attitude.

- The purpose of NLET processing is to match NLET meaningful words against the DB fields descriptors.
- The final step of NLET to SQL query conversion is rather complicated because it is necessary to access data from many different tables within an EAD and join those tables together in SQL query.

---

## Conclusion

NITED is designed through the Internet to make nationwide election results available to any user. We hope that NITED has the potential to change certain aspects of political behaviour, including people's desire to engage with the political process. Like any technology, systems like NITED can have a wide variety of effects on political behaviour and practices, but it is too soon yet to make general conclusions about its impact. Nevertheless, we intend that, following the 2010 UK General Elections in the NITED will play an important role, helping to make nationwide election results available to Web users.

---

## Bibliography

- Guy Francis, Mark Lishman, Vladimir Lovitskii, Michael Thrasher, David Traynor, 2007. "Instantaneous Database Access", *International Journal "Information Theories & Applications"*, Vol 14(2), 161-168.
- L.Huang, T.Ulrich, M.Hemmje, E.Neuhold, 2001. "Adaptively Constructing the Query Interface for Meta Search Engines", Proc. of the Intelligent User Interface Conf.
- Elena Long, Vladimir Lovitskii, Michael Thrasher, David Traynor, 2009. "Mobile Election", International Book Series "Information Science and Computing", Book 9, Intelligent Processing, 19-28.
- V.A.Lovitskii and K.Wittamore, 1997. "DANIL: Databases Access using a Natural Interface Language", *Proc. of the International Joint Conference on Knowledge-Dialogue-Solution: KDS-97, Yalta (Ukraine)*, 282-288.

---

## Authors information



*Elena Long – University of Plymouth, Plymouth, Devon, PL4 6DX, UK,  
e-mail: [elena.long@plymouth.ac.uk](mailto:elena.long@plymouth.ac.uk)  
Major Fields of Scientific Research: Political science*



*Vladimir Lovitskii – University of Plymouth, Plymouth, Devon, PL4 6DX, UK,  
e-mail: [vladimir.lovitskii@fsmail.net](mailto:vladimir.lovitskii@fsmail.net)  
Major Fields of Scientific Research: Artificial Intelligence*



*Michael Thrasher – University of Plymouth, Plymouth, Devon, PL4 6DX, UK  
e-mail: [mthrasher@plymouth.ac.uk](mailto:mthrasher@plymouth.ac.uk)  
Major Fields of Scientific Research: Political science*

---

## ANALYSIS OF NATURAL LANGUAGE OBJECTS

Oleksii Vasylenko

**Abstract:** *This paper describes technology of computer processing of knowledge contained in natural language. Formulated topical areas of applied research related to the recovery and processing of knowledge in the texts of the Internet, technical specifications, etc.*

**Keywords:** *knowledge acquisition, knowledge processing, automated knowledge management system, a computer analysis of the text, social networks, the Internet, technical task.*

---

### Introduction

The most common form of knowledge representation are natural-language texts. Text only form of knowledge is human, such knowledge is easily treated and are generated, replicated and modified. However, the rapid growth of text areas is a cause of difficult accessibility of target knowledge when they are needed. An additional problem is the complexity of the validation text array that consists in finding and correcting errors, removing duplicates and inconsistencies. Information retrieval systems are not designed to address this problem, since uses such words of text, not the knowledge contained therein. In this connection, get the relevance of knowledge extraction from texts. As a result of extraction of knowledge become explicit form and are suitable for automated processing, for example, associating systems analysis, performing a comparison with the extraction of a reference model domain for the purpose of validation. The problem of extracting devoted a lot of foreign works, united in a single class of problems in extracting information from texts. Retrievable information is data structures whose fields are filled with text fragments. The disadvantage of foreign developments is the strong dependence on the particular grammar. Among the domestic works are known only two complete systems of companies RCO and Yandex, with very limited application, since there is no simple way to adapt them to an arbitrary domain. Moreover, in today's papers there is no information about the system correlates the analysis, in use. Thus, the development of mathematical models of extraction applicable for text without reference to a specific language and is easily adaptable to the needs of a particular subject area, represents a major scientific challenge, and develop a model of knowledge representation, which is formed by the extraction, convenient to carry associating analysis has significant practical importance. Extracting information from texts is a subtask of a larger problem - namely, extraction of knowledge. To identify in the texts of the data structure necessary to have two sets of rules: the rules of morphological analysis and rules extraction. First identify the linguistic properties of words of text, whereas the second, using these properties, impose conditions on the composition and structure of the context of the task information. The rules of both types on a par with extractable data structures are the domain knowledge. Formation of such rules in the existing domestic designs carried out manually, that is the cause of the complexity of the system setup of extraction. In this regard, the development of automated drafting rules and regulations extraction of morphological analysis is an important problem whose solution in general terms without reference to a particular language is currently absent.

---

## The essence of work

---

**Goals and objectives of my work:** The aim is to develop a model of knowledge extraction from texts and their methods of training for associating systems analysis of texts in natural language. To achieve this goal within the thesis addressed the following objectives:

1. study of contemporary models of extracting information from texts and teaching methods of such models;
2. develop a model of knowledge representation, which allows you to effectively assume its associates analysis of texts;
3. creating a model of knowledge extraction from object-oriented texts;
4. Develop a method for learning models of knowledge extraction from texts;
5. creating a model of morphological analysis of words and the method of teaching;
6. Pilot testing of proposed models and methods. The object and subject of study. The object of the study are natural-language texts as a form of knowledge representation domain. The subject of the study are the processes of automated identification and formalization of knowledge presented in the form of natural language texts. When developing models and methods has been applied apparatus of algebraic systems, including algebraic lattices and graphs, as well as the apparatus of formal grammars and automata theory.

### Sphere of application:

1. Information intelligence. By applying technology is Dana compile dossiers on the interesting, the subject matter on which information is available in open sources. For example, the object of interest can be a politician, the dossier which may include: name, age, origin, education, relationship to the parties and other political leaders, opinion on, regarding the events of interest, etc. Likewise, exploration is carried out for commercial purposes, as some companies' interest in the activity of a competitor, whose actions are covered in the media. In this case, extraction of products are advertised competitor's transactions with other market participants, changes in leadership positions, as well as acquisitions and mergers.
2. Automated compilation of directories and dictionaries. Retrieval methods can also be used for filling the domain-specific ontologies, thesauri and dictionaries. In this case, extraction of subject concepts and relationships between them, reflected in the texts of the subject area. Next, this knowledge can be used in the conceptual text indexing to improve the quality of full-text searching and classification.
3. Revealing the contradictions in the texts of documents. Extracted from the texts of knowledge can be used for further comparison with the standard base domain. For example, from internal documents can be extracted: Name employees, their positions and names of units in which they operate. Further, with reference staff base organization, you can compare with it the extracted information. In case of discrepancies drafter may be issued a semantic error message indicating a text fragment, where the discrepancy was discovered.
4. Validation and recovery of the text. The texts of some domains may contain incomplete or erroneous information that is necessary to check and repair. An example of such texts are, addresses clients of the organization, recorded by the operator as a continuous line. Typically, the operator enters an incomplete address information, omitting the index, the name of the region, etc. The proposed methods allow to extract the values of specific fields (names of cities, streets, regions much more await you.) From a continuous line of the address. Further, having a reference database of postal addresses of the country

an opportunity to compare the extracted field with it, correct errors in individual fields and restore the missing values of address fields.

5. Monitoring the flow of texts: the greatest interest to the systems of this kind are often by U.S. intelligence, for which the most common themes extracted facts are attacks and riots. These facts are revealed by the analysis of Web, a similar process may be, and e-mail to identify and prevent the crimes being prepared. Extracted facts are composed of structural elements, describing, for example, participants of the event, their objectives and means, and the place of event, its causes and consequences

**Scientific novelty:** The model extraction of ontological components of object-oriented texts. Easy to structure the rules of extraction provides practical feasibility of machine learning, as well as the implementation of the method of extraction on the basis of a finite automaton, independent of the grammar of natural language. The method for learning extraction models, which proposed a new strategy of compressing a group of generalization of training examples, as well as a new approach to the coupled synthesis of the rules based on an assessment of total error of generalization of their individual elements. A modification of the principle of analogy, the morphological analysis of texts, thus reducing the volume of the morphological dictionary and reduce the computational complexity of algorithm analysis. A model of morphological analysis, acting in accordance with the modified principle and propose a method of learning, allowing, without human intervention to build a morphological analyzer, which has better quality of analysis in comparison with the dictionary methods.

**Practical value:** The developed model extraction can be used in systems analysis associating performing search spelling errors, restoring missing data in the text, as well as identify the contradictions between the contents of text and reference knowledge base. The model is applicable in systems extract information from texts in the following areas: automated content of relational databases, directories, dictionaries and ontologies, information exploration, monitoring text streams. The modified principle of analogy, morphological analysis and built on its base model allows an order to reduce the amount of morphological dictionary. The proposed method of teaching this model completely eliminates the expert from the preparation of training examples.

**Implementation approaches:** The first point of view, the focus in the pragmatic aspects and adopted under the direction of knowledge management, knowledge represents the data obtained in the right place at the right time to solve practical problems, usually for a decision, including the implementation of actions that a person or a technical system. Moreover, by its structure and method of storing knowledge can in no way different from other data - any piece of the database or the full-text archive of documents converted into knowledge as soon as it is drawn to look interested consumers. That is the position of this view - the focus of utilitarian interest - determines which piece of data is currently interpreted as knowledge.

The second point of view, the focus in the content aspects, and adopted under the direction of artificial intelligence, believes that knowledge is different from conventional data primarily by its structure. It is a set of specially structured data applies the concept of the knowledge base, implying:

- A logical ordering of data based on certain criteria established by the domain model (ontology);
- Representation of data in accordance with certain formal model (the semantic network, a framed set of products)
- The possibility of obtaining new data from the old on the basis of some formal mechanism;
- Data storage in special structures that provide high efficiency of typical operations on them (search on graphs, analysis hierarchy, the logical conclusion, etc.)

### **Applied automatic warning systems of knowledge management (AWSK) today:**

The main consumers of knowledge today, the following groups:

1. Executives, management decisions;
2. Analysts, surveys, forecasts and recommendations for (1), including the security services;
3. Narrow community of professionals in certain areas, which are developed specialized systems - expert systems in medicine, geology, extraction system of formulas of organic compounds from scientific publications in chemistry, etc.
4. Other officials of scientific information and areas in need of timely and complete receipt of information for the production of intellectual products (eg, structured news on interesting sections of the science and technology, socio-political life);
5. Non-professional users - people who want to use the knowledge for everyday needs with a view to deciding, for example, choosing the model of the goods when buying, selecting a service provider or mode of action in certain situations (legal and medical questions, troubleshooting techniques).

As the results of a literature review, including online submissions from producers of software solutions for knowledge management, to date, all the attention of theoreticians and practitioners address the needs of groups 1 and 2, which is apparently associated with the greatest expected return on this investment. For example, in automated information systems, positioned in the market as a decision support system, competitive intelligence, business intelligence, already has a set of subsystems that implement certain functions of extraction, storage, retrieval and generation of new knowledge. Also, some attention has attracted groups (3), which is usually at the expense of budget financing, develop specialized applications of information systems, which include data mining tools and text mining, expert systems. Groups (4) and especially the group (5), which applies to every person deprived of the attention of developers and, in spite of free access to a potential source of knowledge - the Internet is limited in the tools of knowledge extraction simplest (in terms of consumer functions) by search engines like Google or Yandex. This is partly motivated by the immense breadth of interests of these groups, the lack of a limited subject area.

Finally, I was not able to detect not only full CPSE that combines the phase of knowledge extraction from text with the phase of their treatment, but even a convincing example of practical use of such a system. Application programs using artificial intelligence methods that can convert non-trivially extracted from the text of the elements of knowledge (interpret, synthesize, identify dependencies, predict, etc.), today there is even the English language. This situation is caused, apparently, for two reasons. First, a weak distribution systems of linguistic analysis, the ability to interpret the relationship between words and therefore really extract knowledge as certain elements that have internal structure and is suitable for a non-trivial semantic processing of the artificial brain - such a system understanding of the text on Russian and international markets have only recently begun to appear and have not yet had time to acquire applications: Net Owl ([www.netowl.com](http://www.netowl.com)), Attensity ([www.attensity.com](http://www.attensity.com)), RCO Fact Extractor ([www.rco.ru](http://www.rco.ru)). Secondly, the potentially low reliability of automatically extracted from the text of the allegations and facts that due to both imperfect interpretation algorithms for text and low-quality sources of information, since virtually no interesting extract knowledge from the scientific literature, and from all kinds of text "pomoe" to what are the social Internet, modern media, and even the archives of scientific and technical reports. As a result, despite the boom around the need for knowledge extraction from text processing and recycling, raised today by developers and sellers of CPSE, it seems that in practice such systems are still useless, at least, outside of highly specialized areas, which, however, not true.

**The proposed method:** After carefully considering all the existing methods, capabilities and operating time, We can interpretate our hike to understand the meaning of natural language by computer.

The concepts can replace each other on the principle of consistency of meaning. It is above all that are essential to accelerate the process of text recognition as well - reducing the load on the knowledge base. (Kuzemin A., thesis work). That is – replaceability of the concepts significantly reduces the volume of the knowledge base by reducing the number of own concepts as its components. Create the new theorem replaceability of the concepts:

**Theorem 1.** If the concept  $PO_2$  inherits the concept  $PO_1 - G(PO_2, PO_1)$  in a certain category of concepts  $C$ , the notion of  $PO_2$  can substitute for a mikrosituations concept  $PO_1$  without losing the semantic load and informative of this mikrosituation.

**Proof.** In accordance with the structure and strategy categories identify the concept to identify the concept  $PO_2$  must affirmatively respond to the decision rules  $P_1, P_{k_1}, \dots, P_{k_n}, P_2$  that relevant concepts  $PO_1, PO_{k_1}, \dots, PO_{k_n}, PO_2$ . This means that identified the problematic concept as we have passed these decision rules, including  $p_1$  which refers to the notion  $PO_1$ . Hence, the notion  $PO_2$  may be perceived as a concept  $PO_1$ , possessing its characteristic features.

**It is possible to obtain the final simplified formula.** To begin to define a mathematical model for this area, that is, a preliminary algorithm of the program domain.

Thus, we set the field, which we consider to find items that need to be put in the text as a priority. Items that are behind them before the end of the line, will be the ones most desired predicates that become the nodes of a semantic network specification.

So, try to express this simple mathematical formula:

Consider the above described concept  $\mathcal{E}_{i_m/j}$  - the word in a sentence  $\mathcal{E}_j$  determinants in appliance has  $\mathcal{A}_i$  and

$\mathcal{T}_i$ :  $\mathcal{A}_{i_m}, \mathcal{T}_{i_m}$  when  $m=1$  (definite value)  $\Leftrightarrow \mathcal{A}_{i_1}, \mathcal{T}_{i_1}$  - given, = const

(The above mentioned node predicates, which is searched, that is - nodes ontology similarity Product RCO). If the nodes of ontologies are not specified clearly (for thematic units characterized by peculiar and persistent concept), then to search for meaning in a sentence erants, first highlight the main predicate, then place it in a network node.

1. **Consider the 1-st case** with known known predicates:

Suppose there is  $\mathcal{E}_{i_m/j}$ , that

$$(\mathcal{A}_{i_1}, \mathcal{T}_{i_1} \in \mathcal{E}_{i_m/j}, \{\mathcal{E}_{i_m/j_1}, \mathcal{E}_{i_m/j_2}, \mathcal{E}_{i_m/j_3}, \dots, \mathcal{E}_{i_m/j_n}\}) \in \mathcal{E}_j \Rightarrow$$

$$\Rightarrow [\Pi_n] = \{\mathcal{E}_{i_m/j_2}, \dots, \mathcal{E}_{i_m/j_n}\},$$

Where  $\Pi_n$  - predicate found meeting the criteria  $\mathcal{A}_{i_1}$  and  $\mathcal{T}_{i_1}$

2. **Consider Case 2**, where the proposal is a set of elements without an explicit predicate. In this case, I have proposed for the allocation of nodal predicates analyze the proposal for the parts of speech, then find the subject and the predicate method for counting the frequency of occurrence of noun and a verb. Mathematically, it is possible to express this:

We have a set of words  $\{t_{i_m/j_1}, t_{i_m/j_2}, t_{i_m/j_3}, \dots, t_{i_m/j_n}\} \in t_j$ ,

Let each of them has an additional parameter  $\alpha$  - the frequency for each element, as well as the parameter responsible for the definition of the speech clearly designed combinations of endings – p (look application "A", list of terminals), where  $(1...n) \in p$ , we have a structure:

$$\{t_{i_m/p(1...n)}^\alpha, t_{i_m/p(2...n)}^\alpha, t_{i_m/p(3...n)}^\alpha, \dots, t_{i_m/p(n)}^\alpha\} \in t_j,$$

Moreover, if found by the predicate coincides with the already existing sites - it is not analyzed in the future, and put in its place. If the predicate is unique - it is analyzed as part of speech and related items - similar to the ongoing analysis of combinations. It is defined in the predicate belonging to the p terminal number (1, 2 ... 20), then determined the meaning of a combination of Grammar small model of the Russian language (application "A"), currently developed 37 combinations. (<http://neurotechnica.info>)

After receiving the new value  $\Pi_n$ , determine its meaning - it is entered in the knowledge base, which compares to the value of suschestvubschimi concepts  $Po_{1...n}$ . Further processing of meaning is on a similar principle, but with reference to a specific predicate.

---

## Conclusion

---

In this work the method of analysis of natural language objects, which accelerates the work with large volumes of the analyzed verbal text. By itself, the semantic network is a self-learning, as accumulating predicates, new to them in their meaning. Existing predicates contained in the semantic network to the new analysis may serve as benchmarks. Thus, the proposed recognition model of natural language objects can be faster and more efficiently than the existing ones.

---

## Application "A"

---

### List of terminals:

1. [мо] - модификаторы прилагательных и наречий
2. [пи] - прилагательные
3. [кф] - краткие формы прилагательных или причастий
4. [ср] - степени сравнения прилагательных
5. [нр] - наречия обстоятельственные
6. [сщ] - существительные
7. [гл] - глаголы в личной форме
8. [нф] - инфинитивы глаголов
9. [дч] - деепричастия
10. [пч] - причастия
11. [пе] - предлоги
12. [юо] - союзы обстоятельственные, так\_как, поскольку, поэтому ...
13. [юл] - союзы типа либо
14. [юи] - союзы типа и, а, но, однако...
15. [ча] - частицы не, даже, даже\_не ...
16. [ко] - формы который, которого, ...
17. [как] - союзы как, как будто, ...
18. [тч] - то\_зпт\_что,
19. [быть] - быть, был, была, будет ...



**Grammar of a small model of the Russian language**

1. дк<\*\*-дк\_1, [если], пп\_лог, [зпт\_то], пп\_лог.
2. дк<\*\*-дк\_2, пп\_лог.
3. пп\_лог<\*\*-пп\_лог\_1, пп\_и.
4. пп\_лог<\*\*-пп\_лог\_2, пп\_или.
5. пп\_или<\*\*-пп\_или\_1, [либо], пп, пп2.
6. пп\_и <\*\*-пп\_и\_1, пп, пп3.
7. пп2<\*\*-пп2\_1, [юл], пп, пп2.
8. пп2<\*\*-пп2\_2, [].
9. пп3<\*\*-пп3\_1, [юи], пп, пп3.
10. пп3<\*\*-пп3\_2, [].
11. пп<\*\*-пп\_1, го, гс, гг, ва, го. % Простое предложение
12. го<\*\*-го\_1, ча0,[пе], гс, го. % Предложная конструкция
13. го<\*\*-го\_2, гм, ча0,[нр], го. % Группа наречий обст.
14. го<\*\*-го\_3, [как], гс, [зп], го. % Выражения подобия с "как"
15. го<\*\*-го\_4, [юо], пп, [зп], го. % Союзные обст.оборот
16. го<\*\*-го\_5, гн, ча0,[дч],ва,го,[зп], го. % Деепричастный оборот
17. го<\*\*-го\_6, гм, ча0,[ср],ва, [зп], го. % Сравнительная степень
18. го<\*\*-го\_7, []. % Необязательность
19. гс<\*\*-гс\_1, гп, ча0,[сщ], ва, по. % Группа существительного
20. гс<\*\*-гс\_2, [тч], пп, [зп]. % гипер-существительное
21. гг<\*\*-гг\_1, гн, ча0,[гл],бы0. % Группа глагола
22. гг<\*\*-гг\_2, ча0,быть0, гн, ча0,[кф]. % Группа глагола, кр. форма
23. гг<\*\*-гг\_3, ча0,[быть], гп, [зп]. % Констр. с тире и "быть"
24. гг<\*\*-гг\_4, гм, ча0,[ср]. % Сравнительный оборот
25. ва<\*\*-ва\_1, гс, ва. % Группа валентностей
26. ва<\*\*-ва\_2, гн, ча0,[нф], ва. % Инфинитив как валентность
27. ва<\*\*-ва\_3, [].
28. по<\*\*-по\_1, [зп],[ко], гг, ва,го,[зп]. % Правое определение-1
29. по<\*\*-по\_2, [зп],[ко],гс, гг, ва,го,[зп]. % Правое определение-2
30. по<\*\*-по\_3, [зп],гн,ча0,[пч],ва,го,[зп]. % Причастный оборот
31. по<\*\*-по\_4, [].
32. гп<\*\*-гп\_1, гм, ча0,[пи], гп. % Группа прилагательных
33. гп<\*\*-гп\_2, [].
34. гн<\*\*-гн\_1, гм, ча0,[нр], гн. % Группа наречий
35. гн<\*\*-гн\_2, [].
36. гм<\*\*-гм\_1, ча0,[мо],гм. % Группа модификаторов
37. гм<\*\*-гм\_2, [].

**Exception Handling:**

1. пе0 <\*\*- пе0\_1,[пе]. пе0 <\*\*- пе0\_2,[].
2. ча0 <\*\*- ча0\_1,[ча]. ча0 <\*\*- ча0\_2,[].
3. бы0 <\*\*- бы0\_1,[бы]. бы0 <\*\*- бы0\_2,[].
4. быть0<\*\*- быть0\_1,[быть]. быть0 <\*\*- быть0\_2,[].

---

**Bibliography**

---

1. Ландэ Д.В. Поиск знаний в Internet. – М.:Диалектика, 2005.
2. Гаврилова Т., Хорошевский В. Базы знаний интеллектуальных систем: Учебник для вузов. - СПб.: Питер, 2000. - 384 с.
3. Букович У., Уильямс Р. Управление знаниями: руководство к действию: Пер. с англ. – М.: ИНФРА-М, 2002. - 504 с.
4. W3C Semantic Web Activity. – <http://www.w3.org/2001/sw/>
5. Колесов А. А управлять – так знаниями! // Byte. - N.2 - М., 2002.
6. Голубев С.А., Толчеев Ю.К., Шаров Ю.Л. Опыт внедрения и использования информационно-поисковой системы ODB-Text в Совете Федерации Федерального Собрания РФ // Современные технологии в управлении и образовании - новые возможности и перспективы использования. Сборник научных трудов. ФГУП НИИ "Восход", МИРЭА. - М., 2001. – С. 58 – 61.
7. A.H.F. Laender, B. A. Ribeiro-Neto, Juliana S.Teixeria. A brief survey of web data extraction tools. ACM SIGMOD Record 31(2), pp 84-93. 2002.
8. И. Некрестьянов, Е. Павлова. Обнаружение структурного подобия HTML-документов. СПГУ, 2002. – С. 38 – 54. – <http://meta.math.spbu.ru>
9. B. Courcelle. the monadic second-order logic of graphs xvi: canonical graph decompositions//Logical Methods in Computer Science, Vol. 2 (2:2) 2006, pp. 1–46. [Электронный ресурс] – Режим доступа: [www.lmcs-online.org](http://www.lmcs-online.org), свободный.
10. Wei Han, David Buttler, Calton Pu. Wrapping Web data into XML, SIGMOD Record, vol. 30, №3, September 2001. – pp 33 – 38.
11. Андреев А.М., Березкин Д.В., Симаков К.В. Модель извлечения фактов из естественно-языковых текстов и метод ее обучения //Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Восьмой Всероссийской научной конференции RCDL'2006 (г. Суздаль, 17 - 19 октября 2006 г.). – Ярославль: Ярославский гос. унив.-т им. П.Г. Демидова, 2006. – С.252 – 261.

---

**Authors' Information**

---



**Oleksii Vasylenko**– *Kharkiv National University of Radioelectronics; Kharkiv, Ukraine;*

*e-mail:* [ichbierste@gmail.com](mailto:ichbierste@gmail.com)

*tel.:* +380 63 841 66 23

*Major Fields of Scientific Research: General theoretical information research, Knowledge Discovery and Engineering, Business Informatics.*

## БАЗОВЫЕ СТРУКТУРЫ ЕВКЛИДОВЫХ ПРОСТРАНСТВ: КОНСТРУКТИВНЫЕ МЕТОДЫ ОПИСАНИЯ И ИСПОЛЬЗОВАНИЯ

Владимир Донченко, Юрий Кривонос, Виктория Омардибирова

**Аннотация:** Предложена и детально рассмотрена концепция базовых структур линейного пространства, включающих основные линейные и основные нелинейные объекты. Развита конструктивные методы их построения, описания, перехода и использования на основе систематического развития и применения аппарата псевдообращения по Муру - Пенроузу. Важность приведённых результатов и возможности их применения проиллюстрирована на примере широкого спектра прикладных задач.

**Ключевые слова:** Псевдообращение по Муру - Пенроузу, сингулярное представление матрицы, метод наименьших квадратов, линейная регрессия, системы оптимального управления, прогноз, кластеризация, искусственные нейронные сети.

**ACM Classification Keywords:** G.3 Probability and statistics, G.1.6. Numerical analysis: Optimization; G.2.m. Discrete mathematics: miscellaneous.

---

### Вступление

В работе предложена и обоснована концепция базовых структур евклидова пространства, к которым предлагается отнести основные линейные структуры, а также – основные нелинейные. И те и другие структуры евклидова пространства проявляются либо во множественной форме: подпространства и гиперплоскости, либо в единичной: линейные операторы. То же относится к нелинейным структурам. К таким нелинейным структурам относятся с одной стороны: матрицы квадратичных форм (в работе – неотрицательно определённых). С другой – поверхности уровня, отвечающие единичному значению соответствующей квадратичной формы. Такими поверхностями уровня являются эллипсоиды, точнее: цилиндрические эллипсоиды. Описаны конструктивные способы задания взаимного перехода от одних типов структур к другим: от линейных подпространств и гиперплоскостей к матрицам и наоборот, а также от набора векторов к матрицам квадратичных форм и эллипсам группировки. В числе других рассмотрены конструктивные способы порождения подпространств и гиперплоскостей, а также ортогональных проекторов, связанных с указанными объектами. В том же русле лежит рассмотрение группирующих операторов. Упомянутая конструктивность обеспечивается применением псевдообращения по Муру – Пенроузу (ПДО), а также новыми результатами в этой области, берущими своё начало и опирающимися на фундаментальную работу Н.Ф Кириченко [Кириченко, 1997]. Важность и эффективность использования приведённых результатов проиллюстрирована на широком спектре задач, включающем линейную регрессию, в том числе векторную, теорию оптимального управления, кластеризацию, прогноз и функциональные сети, являющиеся обобщением искусственных нейронных сетей.

Как отмечалось в работе [Донченко, 2009], «структура объекта» ассоциируется с представлением об объекте, как чём-то едином, составленном из взаимодействующих, связанных между собой частей. Математическое описание - моделирование объекта связано с передачей представления о «структуре» объекта математически: средствами математического описания «связей». Это означает, что структура объекта должна быть передана, отражена в математической модели средствами математического

структурирования. К последним относятся четыре базовых структуры, а также их комбинации. К четырём базовым следует отнести отношения, функции, операции и наборы подмножеств. К комбинациям относятся, например, структуры линейного и евклидова пространства.

«Линейные структуры»: структуры линейного пространства, занимают ведущее место среди важнейших математических структур. Именно в рамках линейных пространств осуществляется уточнение понятия линейной структуры. В рамках рассмотрения линейных пространств к ним относят линейные подпространства и гиперплоскости а также – линейные операторы и функционалы. Евклидово пространство: конечномерное линейное плюс скалярное произведение, занимает ведущее место среди линейных структур по богатству возможностей использования связей. Богатство свойств линейных структур: и в варианте линейных пространств и подпространств, и в варианте операторов соответствующего вида, – трудно переоценить в математическом моделировании объектов. Это касается как абстрактных математических исследований, так и исследований прикладного характера. В полной мере сказанное выше относится, в частности, к алгебре, регрессионному анализу, теории случайных процессов, теории дифференциальных и интегральных уравнений, систем оптимального управления, прикладным задачам классификации, прогноза и т.д. Важную роль в прикладных исследованиях играют конструктивные методы описания соответствующих объектов. В том, что касается линейных операторов и линейных функционалов, вопрос конструктивности решается построением матриц соответствующих объектов, а для операций – использованием операций матричной алгебры. В том, что касается подпространств, порождённых теми или иными совокупностями векторов, дело обстоит сложнее. Их конструктивное описание можно получить, связав указанные объекты с пространством значений подходящей матрицы, которое в свою очередь описывают подходящим ортогональным проектором. Именно этот подход развивается ниже. Отметим, что ортогональные проекторы играют важную роль в исчерпывающем исследовании систем линейных алгебраических уравнений (СЛАУ). Принципиально важны они также в постановке и решении важных оптимизационных задач с квадратичными функционалами качества в евклидовых пространствах. Это в полной мере относится к задаче построения наилучших квадратичных приближений правой части СЛАУ значениями левой, когда СЛАУ не имеет точных решений. Такие наилучшие приближения называют также псевдорешениями. Конструктивное описание ортогональных проекторов в связи с естественными подпространствами линейного оператора прямо определяется псевдообращением по Муру – Пенроузу [Moore, 1920], [Penrose, 1955] (см. также, например, [Алберт, 1977]). Отметим, также, что важную роль в конструктивном решении прикладных задач с использованием линейных структур играет сингулярное представление (его называют также сингулярным разложением или SVD - представлением) матрицы в специфической записи в виде взвешенной суммы тензорных произведений специального набора пар векторов. Ниже рассматриваются основные свойства линейных структур, основные особенности и возможности их конструктивного описания, а также – использования для конструктивного решения важнейших прикладных задач прогноза, кластеризации и классификации, в других областях.

В заключение отметим, что основные идеи, дух и результаты предлагаемой работы восходят и используют результаты развития теории псевдообращения в работах нашего безвременно ушедшего коллеги и друга, профессора Н.Ф. Кириченко.

---

### **Евклидовы пространства: базовые линейные структуры и связи между ними**

---

В дальнейшем, говоря о евклидовом пространстве  $R^n$ , будем иметь в виду множество конечных числовых последовательностей одной и той же длины  $n$ , записанных в столбик с покоординатными операциями сложения и умножения на скаляр и суммой покоординатных произведений в качестве

скалярного произведения. Именно такой вариант евклидового пространства будем стандартным образом

обозначать через  $R^n$ , а его элементы – через  $a = \begin{pmatrix} a_1 \\ \dots \\ a_n \end{pmatrix}$ . Стандартные ортонормированные базисы,

составленные из векторов с единственной единичной компонентой (остальные – нули) на месте с соответствующим номером будут обозначаться для  $R^m$  и  $R^n$  соответственно через  $e(j) \in R^m, j = \overline{1, m}, e_{(i)} \in R^n, i = \overline{1, n}$ . Оператор  $A$  из  $R^n$ , в  $R^m: A: R^n \rightarrow R^m$ , в ортонормированных базисах  $e(j) \in R^m, j = \overline{1, m}, e_{(i)} \in R^n, i = \overline{1, n}$  будем отождествлять с  $m \times n$ -матрицей  $A = (a_{ij})$  этого оператора. Для матрицы  $A = (a_{ij})$  будем использовать также блочное представление по столбцам (столбцовое) и строкам (строчное):

$$A = \begin{pmatrix} a_{(1)}^T \\ \dots \\ a_{(m)}^T \end{pmatrix} = (a(1) : \dots : a(n)), a_{(i)}^T \in R^n, i = \overline{1, m}, a(j) \in R^m, j = \overline{1, n}.$$

Линейное пространство всех  $m \times n$  матриц будем обозначать  $R^{m \times n}$ .

Линейное подпространство, порождённое системой векторов,  $c_k \in R^p, k = \overline{1, K}$  будет обозначаться через  $L(c_k, k = \overline{1, K}) \equiv L(c_1, \dots, c_K)$ , а линейное подпространство значений линейного оператора  $A: R^n \rightarrow R^m$  – через  $L_A$ .

Первым из набора базовых свойств является утверждение о том, что

1.

$$L_A = L(a(1), \dots, a(n)).$$

Таким образом, линейное подпространство, порождённое набором векторов, совпадает с подпространством значений матрицы, составленной из векторов набора, как из столбцов.

2. “Для элементов столбцового и строчного представления матрицы  $A \in R^{m \times n}$  справедливы соотношения

$$a(j) = A e_{(j)}, j = \overline{1, n},$$

$$a_{(i)}^T = e^T(i) A, i = \overline{1, m}.$$

3. Для произведения произвольных матриц  $B, C$  со столбцовым и строчным представлением

$$B = (b(1) : \dots : b(r)), b(j) \in R^m, j = \overline{1, r}, C = \begin{pmatrix} c_{(1)}^T \\ \dots \\ c_{(r)}^T \end{pmatrix}, c_{(i)} \in R^n, i = \overline{1, r}$$

соответственно и диагональной матрицы  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_r)$  справедливо соотношение

$$B \Lambda C = \sum_{i=1}^r \lambda_i b(i) c_{(i)}^T.$$

Важной составляющей аппарата конструктивного описания и использования линейных структур является понятие ортогонального проектора, которое полностью отвечает стандартному геометрическому представлению об ортогональном проектировании. Общей, основой эффективного использования

ортогональных проекторов является наличие двух эквивалентных определений таких проекторов и возможности их конструктивного построения в связи с линейными подпространствами через псевдообращение.

4.«Геометрическое определение ортогонального проектора»: для разложения  $R^p = L + L^\perp$  в прямую сумму ортогональных подпространств ортогональным проектором  $P_L$  на линейное подпространство  $L \subseteq R^p$  называется оператор, определяемый соотношением

$$\begin{aligned} P_L x &= P_L(x_L + x_{L^\perp}) = x_L, \text{ где} \\ x &= x_L + x_{L^\perp}, x_L \in L, x_{L^\perp} \in L^\perp \end{aligned} \quad (1)$$

– однозначное разложение произвольного вектора  $x \in R^p$  по двум составляющим ортогональной суммы. Очевидным образом оператор ортогонального проектирования является линейным оператором.

5.Разложение (1) произвольного вектора  $x \in R^p$  в силу симметричности относительно ортогональных слагаемых определяет одновременно два ортогональных проектора:  $P_L, P_{L^\perp}$  с очевидным соотношением

$$P_L + P_{L^\perp} = E_p$$

где  $E_p$  – единичная матрица соответствующей размерности.

6.Для ортогонального проектора  $P_L$  на подпространство  $L$  оператор  $Z_L \equiv E_p - P_L$  является ортогональным проектором на ортогональное дополнение  $L^\perp$  к  $L$ :  $Z_L \equiv E_p - P_L = P_{L^\perp}$ .

7.Абстрактное определение ортогонального проектора: для того, чтобы линейный оператор  $P: R^p \rightarrow R^p$ , был оператором ортогонального проектирования необходимо и достаточно, чтобы он был идемпотентным симметричным оператором. Линейное пространство  $L_p$ , на которое совершается ортогональное проектирование в соответствии с «геометрическим определением» описывается одним из двух соотношений:

$$L_p = \{x: x = Pu, u \in R^p\} = \{x: x = Px, x \in R^p\}.$$

8.Сингулярное или SVD- представление произвольной  $m \times n$  матрицы: для произвольной  $A \in R^{m \times n}$  ранга  $r \leq \min(m, n)$  справедливо следующее представление матрицы в виде взвешенной суммы тензорных произведений

$$A = \sum_{i=1}^r \lambda_i u_i v_i^T \quad (2)$$

где

- $\lambda_1^2 \geq \dots \geq \lambda_r^2 > 0$  общий набор ненулевых собственных чисел матриц  $AA^T, A^T A$ .
- $u_i \in R^m, i = \overline{1, r}$  – ортонормированный набор собственных векторов матрицы  $AA^T$ , отвечающих ненулевым собственным числам:  $AA^T u_i = \lambda_i^2 u_i, \lambda_i^2 > 0, i = \overline{1, r}, u_i^T u_j = \delta_{ij}, i \neq j$ ;
- $v_i \in R^n, i = \overline{1, r}$  – ортонормированный набор собственных векторов матрицы  $A^T A$ , отвечающих ненулевым собственным числам:  $A^T A v_i = \lambda_i^2 v_i, \lambda_i^2 > 0, i = \overline{1, r}, v_i^T v_j = \delta_{ij}, i \neq j$ .

Принципиальную роль в описании базовых структур евклидовых пространств играет псевдообращение по Муру - Пенроузу: [Moore,1920], [Penrose, 1955] как одноместной операции  $A^+$  над прямоугольными матрицами произвольной размерности: для произвольных  $A \in R^{m \times n}$ . В дальнейшем термин «псевдообращение» будем сокращать до ПдО.

### Евклидовы пространства, базовые линейные структуры и ПдО

9.Определение псевдообращения через SVD - представление матрицы: для произвольной  $\square$  - матрицы  $A$  её ПдО  $A^+$  определяется соотношением

$$A^+ = \sum_{i=1}^r \lambda_i^{-1} v_i u_i^T : R^m \rightarrow R^n, \quad (3)$$

где  $u_i, v_i, \lambda_i, i = \overline{1, r}$  – элементы сингулярного разложения матрицы из соотношения (2).

Заметим, что SVD - определение ПдО (соотношение (3)) позволяет легко установить, что ПдО коммутирует с транспонированием, а также ряд других полезных соотношений, в частности, что

$$A^T (A^T)^+ = A^+ A.$$

10.Ортогональные проекторы на подпространства  $L_A, L_{A^T}$  в евклидовых пространства  $R^m, R^n$  соответственно, обозначим их  $P(A^T), P(A)$  соответственно, определяются соотношениями

$$P(A^T) = AA^+ = \sum_{i=1}^r u_i u_i^T,$$

$$P(A) = P((A^T)^T) = A^T (A^T)^+ = A^+ A = \sum_{i=1}^r v_i v_i^T.$$

11.“Операторы  $Z(A), Z(A^T)$ , определяемые соотношениями

$$Z(A) = E_n - P(A) = E_n - A^+ A,$$

$$Z(A^T) = E_m - P(A^T) = E_m - A^T A^+ = E_m - AA^+,$$

являются операторами ортогонального проектирования на подпространства  $L_A^\perp, L_{A^T}^\perp$

Важность последних соотношений определяются тем, что  $L_{A^T}^\perp$  является множеством нулей оператора  $A$ .

12.Подпространство  $L_{A^T}^\perp$  является ядром  $\text{Ker}A$  (множеством нулей) оператора  $A$ :

$$L_{A^T}^\perp = \text{Ker}A = Z(A)R^n.$$

13.Для совместности СЛАУ  $Ax = y$  необходимо и достаточно, чтобы  $y^T Z(A^T)y = 0$ . В этом случае  $A^+y$  является наименьшим по норме решением. Оно ортогонально к  $\text{Ker}A$ , а множество всех решений  $\Omega_y$  описывается соотношением

$$\Omega_y = A^+y + Z(A)R^n = \{x : x = A^+y + Z(A)v, v \in R^n\}. \quad (4)$$

14.Если СЛАУ  $Ax = y$  несовместна, т. е  $y^T Z(A^T)y > 0$  множество, определяемое соотношением (4) описывает совокупность всей наилучших квадратичных приближений правой части значениями левой:

$$\Omega_y = A^+ y + Z(A)R^n = \underset{x \in R^n}{\text{Arg min}} \|Ax - y\|^2. \quad (5)$$

Значение невязки для любого наилучшего квадратичного приближения составляет  $y^T Z(A^T)y$ .

15. Для того, чтобы матричное уравнение  $AX = Y, A \in R^{m \times n}, X \in R^{n \times N}, Y \in R^{m \times N}$  имело корни необходимо и достаточно, чтобы  $\text{tr} Y^T Z(A^T)Y = 0$ . В этом случае множество  $\Omega_Y$  определяется соотношением

$$\Omega_Y = \{X : X = A^+ Y + Z(A)V, V \in R^{n \times N}\}. \quad (6)$$

16. Для линейной зависимости вектора  $d \in R^m$  от столбцов матрицы  $A \in R^{m \times n}$  необходимо и достаточно, чтобы выполнялось соотношение

$$d^T Z(A^T)d = 0. \quad (7)$$

Ссылка на векторы—элементы столбцового представления матрицы  $A$  несколько не ограничивают возможности применения утверждения этого пункта для определения линейной зависимости того или иного вектора от фиксированного набора векторов. Для использования результата в общем случае необходимо и достаточно составить матрицу, в которой векторы набора являются столбцами, и дополнительно использовать п.1 базовых свойств.

Отметим также, что условием линейной независимости строки  $a^T, a \in R^n$  от строк матрицы  $A \in R^{m \times n}$  является условие

$$a^T Z(A)a = 0. \quad (8)$$

17. Прямые формулы Гревия: ПдО произвольной матрицы  $A \in R^{m \times n}$ , дополненной строкой  $a^T \in R^n$ , определяется элементами  $P \in R^{n \times m}, q \in R^n$  блочного представления ПдО  $\begin{pmatrix} A \\ a^T \end{pmatrix}^+ = (P : q)$  расширенной матрицы в соответствии с формулами

$$q = \begin{cases} \frac{Z(A)a}{a^T Z(A)a}, a^T Z(A)a > 0 (\text{нез.}) \\ \frac{R(A)a}{1 + a^T R(A)a}, a^T Z(A)a = 0 (\text{зав.}) \end{cases}, \quad (9)$$

$$P = (E_m - qa^T)A^+,$$

где

$$R(A) = A^+ A^{+T}.$$

Первая строка в (9) отвечает случаю линейной независимости строки - расширения от строк матрицы  $A$ , второй – линейной зависимости.

С учётом коммутирования транспонирования с ПдО, прямые формулы Гревия очевидным образом переписываются для варианта расширения матрицы столбцом.

18. Обратные формулы Гревия: для блочного представления ПдО расширенной матрицы в виде

$$\begin{pmatrix} A \\ a^T \end{pmatrix}^+ = (P : q) \text{ ПдО матрицы } A \text{ определяется соотношениями}$$



$$A^+ = \begin{cases} (E - \frac{qq^T}{\|q\|^2})P, a^T q = 1(\text{незав.}) \\ (E + \frac{qa^T}{1-a^T q})P, a^T q < 1(\text{зав.}) \end{cases} \quad (10)$$

Условие в первой строке в соотношениях (10) отвечает линейной независимости строки, которая удаляется, а второй – зависимости от остальных строк матрицы  $A$ . Справедливость этих условий непосредственно вытекает из п.17 “, и используется, когда дополнительно известна ПДО для расширенной матрицы.

19. Квадрат расстояния  $\rho^2(a, L_A)$  вектора  $a \in R^m$  от подпространства  $L_A$  определяется соотношением

$$\rho^2(a, L_A) = \min_{y \in R^m} \|a - y\|^2 = a^T Z(A^T) a.$$

20. Квадрат расстояния  $\rho^2(a, \Gamma(b, L_A))$  вектора  $a \in R^m$  от гиперплоскости  $\Gamma(b, L_A) = b + L_A$  определяется соотношением

$$\rho^2(a, \Gamma(b, L_A)) = \min_{y \in \Gamma(b, L_A)} \|a - y\|^2 = (a - b)^T Z(A^T)(a - b).$$

Отметим, что привязка подпространств в пп.19, 20 к множеству значений оператора  $A$  не ограничивает сферу применимости результатов. С помощью п.1 они очевидным образом распространяются на ситуацию, когда подпространство порождается заданной конечной совокупностью векторов.

Формулы аналитического возмущения описывают ПДО изменённой матрицы, когда возмущение имеет вид аддитивной добавки  $ab^T, a \in R^m, b \in R^n$ . В работе [Кириченко, 1997] исчерпывающим образом исследованы варианты представления ПДО возмущённой матрицы  $(A + ab^T)^+$ . Как оказывается, вид соответствующих формул определяется линейной зависимостью или независимости векторов  $a, b^T$  от, соответственно, столбцов и строк матрицы  $A$ . Кроме того, на вид соответствующих формул влияет сохранение или падение ранга возмущённой матрицей. Последнее касается случая, когда одновременно оба вектора  $a, b^T$  линейно зависимы с, соответственно, столбцами и строками матрицы  $A$ . И условия линейной независимости и условие падения ранга носят аналитический характер. В п.16 представлены условия линейной зависимости. Условие сохранения ранга представлено следующим пунктом.

21. Ранги матриц  $A$  и  $A + ab^T$  одинаковы ( $a, b^T$  одновременно зависимы от, соответственно, столбцов и строк матрицы  $A$ ):  $\text{rank}(A + ab^T) = \text{rank} A$ , тогда и только тогда, когда  $b^T A^+ a \neq -1$ . Ранг возмущённой матрицы падает, когда  $b^T A^+ a = -1$ .

Принимая во внимание громоздкость соответствующих формулировок, ниже приведен один из вариантов утверждения о виде ПДО для возмущённой матрицы. С полным вариантом утверждения можно ознакомиться в уже упомянутой работе [Кириченко, 1997].

22. Аналитические формулы возмущения ПДО матриц, фрагмент: если компоненты возмущения:  $a, b^T$  линейно не зависимы от, соответственно, столбцов и строк матрицы  $A$ , т.е.  $a^T Z(A^T) a > 0, b^T Z(A) b > 0$ , то

$$(A + ab^T)^+ = A^+ - \frac{A^+ a a^T Z(A^T)}{a^T Z(A^T) a} - \frac{Z(A) b b^T A^+}{b^T Z(A) b} + Z(A) b a^T Z(A^T) \frac{1 + b^T A^+ a}{a^T Z(A^T) a b^T Z(A) b}.$$

### Евклидовы пространства, базовая нелинейная структура: группирующие операторы

Важнейшими нелинейными структурами евклидова пространства являются квадратичные формы, точнее: неотрицательно определённые квадратичные формы, – и отвечающие им эллипсы или эллипсоидальные цилиндры. Среди таких нелинейных структур принципиальными являются матрицы квадратичных форм, которые, как, оказывается, естественным образом связаны с групповыми свойствами набора векторов и которые поэтому естественно называть группирующими операторами. Группирующие операторы возникают в связи с набором векторов  $a_j \in R^m, j = \overline{1, n}$  и отвечающей ему матрицей  $A = (a_1 : \dots : a_n), a_j \in R^m, j = \overline{1, n}$ . В паре набор-матрица первый будем называть столбцовым представлением матрицы, вторую – матричным представлением набора. Как и ортогональные проекторы, группирующие операторы являются парными. Будем обозначать их, соответственно  $R(A), R(A^T)$ .

22. Определение группирующих операторов:

$$R(A) = A^+ A^{+T}$$

$$R(A^T) = A^{+T} A^+.$$

23. Проектирование на нормированный вектор  $u \in R^m : \|u\| = 1$  элементов набора векторов  $a_j \in R^m, j = \overline{1, n}$ . Основным результатом этого пункта представлен леммой 1 ниже.

Лемма 1. Для произвольного набора  $a_i \in R^m, i = \overline{1, n}$  с матричным представлением  $A = (a_1, \dots, a_n)$  и произвольного нормированного вектора  $u \in R^m : \|u\| = 1$ , справедливо равенство:

$$\sum_{j=1}^n a_j^T u u^T a_j = u^T A A^T u \quad (11)$$

Доказательство. Действительно, принимая во внимание связь векторов набора  $a_i \in R^m, i = \overline{1, n}$  со своим матричным представлением в виде

$$a_j = A e_{(j)}, j = \overline{1, n},$$

где  $e_{(j)} \in R^n, j = \overline{1, n}$  - стандартный ортонормированный базис в  $R^n : e_{(j)}^T = (0, \dots, 1, 0, \dots, 0), j = \overline{1, n}$ , имеем:

$$\sum_{j=1}^n a_j^T u u^T a_j = \sum_{j=1}^n e_{(j)}^T A^T u u^T A e_{(j)} = \sum_{j=1}^n u^T A e_{(j)} e_{(j)}^T A^T u = u^T A \left[ \sum_{j=1}^n e_{(j)} e_{(j)}^T \right] A^T u.$$

Остаётся только заметить, что

$$\sum_{j=1}^n e_{(j)} e_{(j)}^T = E_n,$$

где  $E_n$  - единичная матрица (оператор) в  $R^n$ , и доказательство леммы завершено.

Замечание 1. Левая часть соотношения (11) леммы 1, собственно, представляет собою сумму квадратов проекций векторов набора  $a_i \in R^m, i = \overline{1, n}$  на нормированный вектор  $u \in R^m : \|u\| = 1$ .

24. Проектирование на элементы  $u_i \in R^m, i = \overline{1, r}$  SVD –разложения матричного представления  $A$  набора векторов  $a_j \in R^m, j = \overline{1, n}$ . Основным результатом этого пункта представлен леммой 2, приведённой ниже.

Лемма 2. Для произвольного набора  $a_i \in R^m, i = \overline{1, n}$  с матричным представлением  $A = (a_1, \dots, a_n)$  имеет место соотношение:

$$\sum_{j=1}^n a_j^T u_i u_i^T a_j = u_i^T A A^T u_i = \lambda_i^2, i = \overline{1, r}$$

Доказательство вытекает из леммы предыдущего пункта и из п.8, в котором в рамках сингулярного представления матрицы наборы  $u_i \in R^m, v_i \in R^n, r = \text{rank} A$  определяются как ортонормированные наборы собственных векторов матриц  $A A^T, A^T A$ , отвечающих общему набору ненулевых собственных чисел  $\lambda_i^2 > 0, i = \overline{1, r}$ .

25. Группирующие операторы: эллипсы группировки набора векторов  $a_i \in R^m, i = \overline{1, n}$ . Основное утверждение пункта – теорема 1 ниже.

Теорема 1. Пусть  $a_i \in R^m, i = \overline{1, n}$  произвольный набор векторов из  $R^m$  с матричным представлением  $A = (a_1 : \dots : a_n)$ . Тогда все векторы набора принадлежат внутренности эллипса, точнее: эллипсоидального цилиндра, определяемого уравнением

$$x^T R(A^T) x = r, x \in R^m,$$

где,  $R(A^T)$ , «группирующий» оператор, определяемый стандартным образом

$$R(A^T) = A^{+T} A^+.$$

Доказательство. Рассмотрим квадраты проекций векторов набора  $a_j, j = \overline{1, n}$  на каждый из векторов  $u_i, i = \overline{1, r}$  SVD-представления (2) матрицы  $A$ . Принимая во внимание ортонормированность набора  $u_i, i = \overline{1, r}$ :  $u_i^T u_j = \delta_{ij}, i \neq j, i, j = \overline{1, r}$ , а также сингулярное представление (3), и обозначая квадраты проекций через  $\|Pr_{u_i} a_j\|^2, i, j = \overline{1, r}$ , очевидным образом имеем

$$\|Pr_{u_i} a_j\|^2 = a_j^T u_i u_i^T a_j, j = \overline{1, n}, i = \overline{1, r}.$$

Суммирование по всем векторам набора  $a_j, j = \overline{1, n}$  и применение леммы 2 даёт

$$\sum_{j=1}^n a_j^T u_i u_i^T a_j = u_i^T A A^T u_i = \lambda_i^2, i = \overline{1, r},$$

т.е.

$$\sum_{j=1}^n a_j^T u_i u_i^T a_j = \lambda_i^2, i = \overline{1, r}.$$

Таким образом, после деления обеих частей последнего соотношения соответственно на  $\lambda_i^2, i = \overline{1, r}$ , имеем

$$\sum_{j=1}^n \frac{a_j^T u_i u_i^T a_j}{\lambda_i^2} = \frac{u_i^T A A^T u_i}{\lambda_i^2} = 1, i = \overline{1, r},$$

т.е.

$$\sum_{j=1}^n \frac{a_j^T u_i u_i^T a_j}{\lambda_i^2} = 1, i = \overline{1, r}.$$

Свернув (просуммировав) последнее равенство по  $i = \overline{1, r}$ , получаем

$$\sum_{i=1}^m \sum_{j=1}^n \frac{a_j^T u_i u_i^T a_j}{\lambda_i^2} = \sum_{i=1}^m \frac{u_i^T A A^T u_i}{\lambda_i^2} = r.$$

Поменяв порядок суммирования в двойной сумме, получаем

$$\sum_{j=1}^n \sum_{i=1}^m \frac{a_j^T u_i u_i^T a_j}{\lambda_i^2} = \sum_{j=1}^n a_j^T \sum_{i=1}^m \frac{u_i u_i^T}{\lambda_i^2} a_j = r.$$

Приняв во внимание, что

$$\sum_{i=1}^r \frac{u_i u_i^T}{\lambda_i^2} = A^+ A^+ = R(A^T),$$

получаем окончательно

$$\sum_{j=1}^n \frac{a_j^T u_i u_i^T a_j}{\lambda_i^2} = \sum_{j=1}^n a_j^T \sum_{i=1}^m \frac{u_i u_i^T}{\lambda_i^2} a_j = \sum_{j=1}^n a_j^T A^+ A^+ a_j = \sum_{j=1}^n a_j^T = \sum_{j=1}^n a_j^T R(A^T) a_j = r,$$

т.е.

$$\sum_{j=1}^n a_j^T R(A^T) a_j = r. \quad (12)$$

Поскольку  $R(A^T)$  – симметричная, неотрицательно определённая матрица, то следствием соотношения (12) является одновременное выполнение неравенств

$$a_j^T R(A^T) a_j \leq r, j = \overline{1, n}. \quad (13)$$

Принимая во внимание уже упомянутую выше симметричность и неотрицательную определённую квадратичной формы  $x^T R(A^T) x, x \in R^m$ , уравнение

$$x^T R(A^T) x = r, x \in R^m, \quad (14)$$

определяет эллипс, точнее: эллипсоидальный цилиндр, в  $R^m$  с длинами  $\frac{1}{\lambda_i \sqrt{r}}, i = \overline{1, r}$  нетривиальных полуосей. Напомним, что  $r = \text{rank} A \leq \min(m, n)$ .

Таким образом, выполнение неравенства (13) для всех векторов набора  $a_j \in R^m, j = \overline{1, n}$  означает их одновременную принадлежность внутренности эллипсоидального цилиндра с уравнением (14), и доказательство теоремы завершено.

Замечание 2. В действительности неравенство (13) может давать существенное закругление «радиуса» эллипса. Так, при векторах  $a_j \in R^m, j = \overline{1, n}$ , близких к ортогональным, очевидным образом, константа в

правой части в правой части (13) можно выбрать близкой к 1. Так что справедлив более жёсткий вариант теоремы 1, являющийся предметом следующего утверждения.

26. Усиление результата об эллипсах группировки.

Теорема 2. Векторы произвольного набора  $a_i \in R^m, i = \overline{1, n}$  с матричным представлением  $A = (a_1 : \dots : a_n)$  принадлежат внутренности эллипсоидального цилиндра

$$x^T R(A^T) x = r_{\max}, r_{\max}^2 \leq r = \text{rank} A, x \in R^m,$$

$$r_{\max}^2 = \max_{j=1, n} a_j^T R(A^T) a_j.$$

---

### Применения: линейная регрессия, скалярные наблюдения

---

Применение ПдО в линейной регрессии определяется тем, что МНК – оценка  $\hat{\beta}$  (оценка метода наименьших квадратов) неизвестного параметра  $\beta \in R^p$  линейной регрессии  $y = \sum_{j=0}^{p-1} \beta_j f_j(x) + \varepsilon$  на основе

наблюдений  $(x_i, y_i), x_i \in R^m, y_i \in R^1, i = \overline{1, n}$  определяется решением оптимизационной задачи

$$\hat{\beta} \in \underset{\beta \in R^p}{\text{Arg min}} \|X\beta - Y\|^2, \quad (15)$$

в которой матрица  $X$  – матрица плана, а  $Y$  – вектор – столбец с компонентами  $y_i \in R^1, i = \overline{1, n}$  – вектор наблюдений регрессанта. В соответствии с соотношением (4) п.13 общее решение задачи (15) определяется соотношением

$$\hat{\beta} \in X^+ Y + Z(X)v, v \in R^p \quad (16)$$

со свободным параметром  $v \in R^p$ .

Решение задачи МНК - оценивания в виде (16) полностью согласуется с классическим решением уравнения Гаусса – Маркова в виде

$$\hat{\beta} = (X^T X)^{-1} X^T Y,$$

поскольку в случае полного столбцового ранга матрицы плана, её ПдО  $X^+$  и ортогональный проектор  $Z(X)$  определяются соотношениями

$$X^+ = (X^T X)^{-1} X^T, Z(X) = 0.$$

В этом случае классическом случае множество МНК – оценок в (16) является одноэлементным.

---

### Применения: задача терминального управления

---

Под задачей терминального управления для линейной динамической системы с дискретным временем

$$\mathbf{x}(k+1) = \mathbf{A}(k)\mathbf{x}(k) + \mathbf{b}(k)u(k),$$

$$\mathbf{x}(0) = \mathbf{x}_{(0)},$$

где

$$\mathbf{x}(k) \in R^n, u(k) \in R^1, \mathbf{A}(k) \in R^{n \times n}, \mathbf{b}(k) \in R^n, k = \overline{0, N},$$

имеют в виду задачу выбора такого управления  $u(k)$ ,  $k = \overline{0, N}$ , которое позволяет вывести фазовую траекторию в момент  $N+1$  на уровень  $x_{(1)}$  или, если это невозможно, выбором того же управления минимизировать отклонение  $\|x(N+1) - x_{(1)}\|^2$ .

Принципиальным результатом для исследования задачи терминального управления является теорема редукции, позволяющая свести задачу терминального управления к СЛАУ.

Теорема 3 (теорема редукции). Задача терминального управления является эквивалентной СЛАУ

$$\mathbf{W}(N+1)\mathbf{u} = \mathbf{x}_{(1)} - \mathbf{A}(N-1)\mathbf{A}(N-2)\dots\mathbf{A}(0)x_{(0)},$$

в которой вектор  $u \in R^{N+1}$  – объединенный вектор управления, а матрица  $\mathbf{W}(N+1)$  является блочной:

$$\mathbf{W}(N+1) = (\mathbf{W}(N+1,0) : \mathbf{W}(N+1,1) : \dots : \mathbf{W}(N+1,N)),$$

с блоками  $\mathbf{W}(N+1,k)$ ,  $k = \overline{0, N}$ , определяемыми соотношениями

$$\mathbf{W}(N+1,k) = \mathbf{A}(N)\mathbf{A}(N-1)\dots\mathbf{A}(k+1)b(k), k = \overline{0, N-1},$$

$$\mathbf{W}(N+1,N) = b(N).$$

Теорема редукции позволяет исчерпывающим образом исследовать задачу терминального управления с помощью пп.13,14. Следует добавить, что аналогичным образом, с помощью ПдО удаётся исчерпывающим образом исследовать задачу терминального наблюдения, в том числе в случае ошибок и шумов (см., например, [Кириченко, Донченко, 2005])

### Применения: кластеризация

ПдО расширяет возможности кластеризации, позволяя эффективно погружать классифицируемые объекты в подходящие подпространства или гиперплоскости. П.1 дает возможность связывать подпространство, порождённое набором векторов, с подходящей матрицей. Если объект связывается с гиперплоскостью, то её смещение – это, как правило, среднее по векторам порождающей совокупности, а подпространство – это подпространство значений матрицы, построенной из центрированных средним векторов порождающей совокупности, как из столбцов. Результаты пп.19, 20 обеспечивают возможность конструктивного вычисления расстояний от объектов (подпространств или гиперплоскостей), ассоциируемых с порождающей совокупностью. Применение стандартных рекуррентных последовательно уточняемых разбиений с расстояниями соответствия из п.19 или п. 20 придаёт процедуре кластеризации необходимой завершенности. С подробностями можно ознакомиться, например, в работах: [Кириченко, Донченко, 2007], [Кириченко, Донченко, 2008]).

Группирующие операторы дают возможность построить расстояния соответствия в связи с использованием базой нелинейной структуры: эллипса группирования и соответствующей квадратичной формы, описываемой группирующим оператором. В сущности, речь идёт о том, что результат теоремы 2 можно использовать для определения расстояния соответствия, точнее его квадрата, обозначим его  $\rho^2(x, Kl)$ , между вектором  $x \in R^m$  и кластером  $Kl$ , порожденным обучающей выборкой  $a_j \in R^m, j = \overline{1, n}$ .

Обозначим через  $\bar{a}$  среднее обучающей выборки  $a_j \in R^m, j = \overline{1, n}$ , а через  $\tilde{A}$  матричное представление для векторов  $\tilde{a}_j \in R^m, j = \overline{1, n}$  первоначального набора, центрированных средним:

$$\bar{a} = \frac{1}{n} \sum_{j=1}^n a_j,$$

$$\tilde{a}_j = a_j - \bar{a},$$

$$\tilde{A} = (\tilde{a}_1 \dots \tilde{a}_n).$$

Тогда квадрат расстояния, определяемый соотношением

$$\rho^2(x, Kl) = \frac{1}{r_{\max}^2} (x - \bar{a})^T R(\tilde{A}^T) (x - \bar{a}),$$

является расстоянием, определяемой стандартной поверхностью уровня, т. е. поверхностью, определяемой уравнением

$$(x - \bar{a})^T R(\tilde{A}^T) (x - \bar{a}) = r_{\max}^2.$$

Такое определение расстояние определяется, собственно минимальным эллипсоидом группировки векторов  $\tilde{a}_j \in R^m, j = \overline{1, n}$ , тогда как среднее  $\bar{a}$  по элементам обучающей выборки задаёт центр соответствующего эллипса.

При наличии набора кластеров  $Kl, l = \overline{1, L}$ , расстояния соответствия определяются, соответственно, соотношениями

$$\rho^2(x, Kl_l) = \frac{1}{r_{l, \max}^2} (x - \bar{a}_l)^T R(\tilde{A}_l^T) (x - \bar{a}_l), l = \overline{1, L}, \quad (17)$$

с очевидной детализацией обозначений в связи с употреблением соответствующего индекса.

Вычисление расстояний до кластеров в соответствии с (17) отвечает расстояниям до представителей кластеров  $\bar{a}_l, l = \overline{1, L}$  в соответствии с минимальными эллипсами группировки центрированных элементов обучающих выборок.

Примеры эффективного применения расстояний соответствия вида (17) можно найти, например, в работах [Кириченко, Донченко, 2008], [Кириченко, Донченко, 2007], [Донченко, Омардибирова 2005].

---

### Применения базовых свойств линейных структур: RFT- функциональные сети

---

Искусственные нейронные сети являются стандартным технологическим инструментом исследования в задачах прогноза, классификации и кластеризации. В сущности, такие сети представляют собой графические изображения: графы суперпозиций, стандартизованных функциональных элементов. В этом смысле их с полным правом можно назвать функциональными сетями. Стандартизация функциональных элементов (нейронов) проявляется в том, что они реализуют скалярную функцию векторного аргумента как суперпозицию линейного функционала и скалярной функции скалярного аргумента:

$$y = F(w^T x), w, x \in R^m, F: R^1 \rightarrow R^1.$$

Упомянутая стандартизация - унификация может проявляться и в выборе внешней функции:  $F$ , которая называется функцией инициализации нейрона. Она может быть фиксированной или принимать значения из конечного набора функций.

Заметим, что, как правило, фиксированной является и структура сети: количество стандартных функциональных элементов нейронов и способ их соединения: топология сети.

Возможности функциональных сетей можно значительно расширить, если 1) придать большую функциональную универсальность и обеспечить адаптивность в построении каждого из стандартизованных элементов; 2) обеспечить более гибкие возможности в соединении элементов: большую свободу в формировании топологии сети; 3) гарантировать адаптивное построение структуры всей сети в целом. Последнее может быть конструктивно реализовано в ходе выполнения последовательных шагов наращивания сети, имеющих рекуррентный характер.

Реализация такой программы для задачи прогноза - восстановления функции, представленной своими значениями  $(x_i, y_i), x_i \in R^n, y_i \in R^m, i = \overline{1, n}$ , состоит в построении стандартного функционального элемента (RFT - преобразователя) в виде

$$y = A_+ \Psi(Cx) , \quad (18)$$

в котором матрица  $C$  – матрица предварительного преобразования вектора признаков  $x$ , а  $\Psi$  – нелинейное по координатное преобразование измененного вектора признаков,  $A_+$  – матрица МНК оценивания на выборке  $(x_i, y_i), x_i \in R^n, y_i \in R^m, i = \overline{1, n}$ .

Эффективность преобразователя (18) в прогнозе зависимости может контролироваться по невязке и по тестовой выборке.

В общем, адаптивная процедура построения функциональной сети развивает идею МГУА А.Г. Ивахненко

---

## Заключение

В работе предложена и обоснована концепция базовых структур евклидового пространства, включая линейные: подпространства, гиперплоскости, линейные операторы, – и нелинейные: неотрицательно определённые квадратичные формы и отвечающие им эллипсоиды группировки. Изложены конструктивные способы описания и взаимного перехода от одних типов структур к другим: от линейных подпространств и гиперплоскостей к матрицам и наоборот, а также от набора векторов к матрицам квадратичных форм и эллипсам группировки и наоборот. В числе других рассмотрены конструктивные способы порождения подпространств и гиперплоскостей, а также ортогональных проекторов, связанных с указанными объектами. В том же русле лежит рассмотрение группирующих операторов. Упомянутая конструктивность обеспечивается применением псевдообращения по Муру – Пенроузу (ПДО), а также новыми результатами в этой области. Важность и эффективность использования приведённых результатов проиллюстрирована на широком спектре задач, включающем линейную регрессию, в том числе векторную, теорию оптимального управления, кластеризацию, прогноз и функциональные сети, являющиеся обобщением искусственных нейронных сетей.

---

## Литература

- [Донченко. 2009] Донченко В.С. Неопределённость и математические структуры в прикладных исследованиях/ Human aspects of Artificial Intelligence International Book Series Information science & Computing.– Number 12. – Supplement to International Journal “Information technologies and Knowledge”. –Volume 3.–2009. – P. 9-18.
- [Донченко, Омардибирова 2005] Донченко В.С., Омардибирова В.Н. Технология классификации электронных документов с использованием теории возмущения псевдообратных матриц// Proceedings of the XI-th International Conference “Knowledge-Dialogue-Solution”. – June 20-30, Varna, 2005. – Volume 1. – С.223-226.



- [Кириченко, Донченко, 2008] В.С Кириченко Н.Ф. Донченко В.С. Гиперплоскости в «множествах и расстояниях соответствия»: кластеризация / Artificial Intelligence and Decision Making.– International book series “INFORMATION SCIENCE&COMPUTING”, Number 7.–Supplement to the International Journal “INFORMATION TECHNOLOGES&COMPUTING” V.2/2008 FOI ITHEA, Sofia 2008.– P. 25-36.
- [Moore, 1920] Moore E.H. On the reciprocal of the general algebraic matrix // Bulletin of the American Mathematical Society. – 26, 1920. – P.394 -395.
- [Penrose, 1955] Penrose R. A generalized inverse for matrices // Proceedings of the Cambridge Philosophical Society 51, 1955. – P.406-413.
- [Алберт, 1977] Алберт А. Регрессия, псевдоинверсия, рекуррентное оценивание. – М.: Наука. – 1977.– 305 с.
- [Кириченко, 1997] Кириченко Н.Ф. Аналитическое представление псевдообратных матриц //Киб. и СА.- №2. –1997.– С.98-122.
- [Кириченко, Донченко,2005] Кириченко М.Ф., Донченко В.С. Задача термінального спостереження динамічної системи: множинність розв’язків та оптимізація//Журнал обчислювальної та прикладної математики. – 2005. –№5– С.63-78.
- [Кириченко, Донченко, 2007] Кириченко Н.Ф., Донченко В.С. Псевдообращение в задачах кластеризации// Киб. и СА.- №4, 2007– С.98-122.
- [Кириченко, Донченко, 2008] Кириченко Н.Ф. Донченко В.С. Гиперплоскости в «множествах и расстояниях соответствия»: кластеризация / Artificial Intelligence and Decision Making.– International book series “INFORMATION SCIENCE&COMPUTING”, Number 7.–Supplement to the International Journal “INFORMATION TECHNOLOGES&COMPUTING” V.2/2008 FOI ITHEA, Sofia 2008. – P. 25-36.

---

### **Информация об авторах**

---

**Владимир С. Донченко** – профессор; Киевский национальный университет имени Тараса Шевченко, факультет кибернетики, Украина, e-mail: voldon@unicyb.kiev.ua

**Юрий Г. Кривонос** – академик НАНУ; зам. директора Института кибернетики НАНУ, Украина

**Виктория Н.Омардибирова** – аспирантка; Киевский национальный университет имени Тараса Шевченко, факультет кибернетики, Украина

---

## НЕЙРОСЕТЕВАЯ АРХИТЕКТУРА НА ЧАСТИЧНЫХ ОБУЧЕНИЯХ

Николай Мурга

**Аннотация:** рассматривается нейросетевая архитектура, обучение в которой происходит не с целью минимизации единого критерия качества, а с разбиением выборки данных на подмножества, для каждого из которых происходит обучение с целью минимизации своего критерия. Рассматривается применение сети к анализу и прогнозированию валютных пар EUR/GBP, EUR/USD, USD/CHF, USD/JPY.

**Ключевые слова:** нечёткая логика, вывод Такаги-Сугено, обучение нейронных сетей, кластеризация, прогноз значений валютных пар.

**ACM Classification Keywords:** G.1.0 Mathematics of Computing – NUMERICAL ANALYSIS – General – Error analysis; G.1.2 Mathematics of Computing – NUMERICAL ANALYSIS – Approximation – Least squares approximation; G.1.6 Mathematics of Computing – NUMERICAL ANALYSIS – Optimization - Gradient methods, Least squares methods; I.2.3 Computing Methodologies - ARTIFICIAL INTELLIGENCE - Deduction and Theorem Proving - Uncertainty, “fuzzy”, and probabilistic reasoning; I.2.6 Computing Methodologies - ARTIFICIAL INTELLIGENCE – Learning - Connectionism and neural nets; I.5.3 - Computing Methodologies - PATTERN RECOGNITION - Clustering.

---

### Вступление

Данная работа посвящена модификации классического метода обучения нейросетевых архитектур, который базируется на минимизации критерия (критериев) качества для всей обучающей выборки данных. В противовес этому подходу, предлагается разбиение всей выборки на непересекающиеся подмножества, на которых и происходит минимизация критерия (критериев), зависящих лишь от значений точек данных подмножеств. В работе рассматривается нечёткая система, с механизмом нечёткого логического вывода Такаги-Сугено с тригонометрическими полиномами в «то»-части нечётких правил. Поставлены эксперименты для анализа свойств рассмотренной архитектуры на периодических зависимостях и случайных процессах. Произведён анализ приложения сети к дневным котировкам валютных пар EUR/GBP, EUR/USD, USD/CHF, USD/JPY за период с 25.03.2009 по 24.03.2010, которые были взяты из [1]. Анализировалась прогностическая способность сети на основании анализа значений критериев RMSE и MAPE.

---

### 1 Архитектура, метод обучения и анализ качества работы предлагаемой нейронной сети

Данный раздел является теоретическим. Он состоит из двух подразделов. В первом подразделе описываются: используемая в работе нейросетевая архитектура и метод её обучения; второй подраздел посвящён описанию и анализу критериев качества работы сети.

---

#### 1.1 Описание архитектуры и метода обучения сети

Предлагаемая нечёткая (гибридная) нейронная сеть использует в качестве механизма нечёткого вывода механизм нечёткого вывода Такаги-Сугено (TS).

База нечётких правил TS выглядит следующим образом:

$$\begin{aligned}
 R_1: & \text{Если } x_1 \in A_1^{(1)}, x_2 \in A_2^{(1)}, \dots, x_n \in A_n^{(1)}, \text{ то } y_1 = \sum_{j=1}^n f_j^{(1)}(x_j) \\
 & \dots \\
 R_K: & \text{Если } x_1 \in A_1^{(K)}, x_2 \in A_2^{(K)}, \dots, x_n \in A_n^{(K)}, \text{ то } y_K = \sum_{j=1}^n f_j^{(K)}(x_j)
 \end{aligned}
 \tag{1}$$

где  $R_i$  - это  $i$ -е нечёткое правило ( $i = \overline{1, K}$ );  $K$  - это количество нечётких правил;  $x_j$  -  $j$ -я компонента входного вектора;  $n$  - размерность входного пространства;  $y_i$  - выход  $i$ -го правила;  $A_j^i$  - значение лингвистической переменной  $x_j$  для правила  $R_i$  с симметричной функцией принадлежности.

Про зависимости  $f_j^{(i)}(x_j)$  следует сказать, что для классической сети TSK ([2], [3]) они имеют фиксированный порядок и обычно либо линейные функции, либо - функции-константы. В данной работе данные зависимости имеют нефиксированный порядок и представляют собой тригонометрические полиномы.

Как отмечается в работе [3], тригонометрическим полиномом порядка  $M$  называется следующее выражение:

$$T_M(x) = \frac{a_0}{2} + \sum_{k=1}^M (a_k \cos kx + b_k \sin kx) \tag{2}$$

С учётом новых обозначений, вышеуказанные зависимости  $f_j^{(i)}(x_j)$  можно записать в виде:

$$f_j^{(i)}(x_j) = A_j^{(i)} \cdot T_{M(i,j)}^{(i)}(x_j) \tag{3}$$

Коэффициенты  $a_{k,j}^{(i)}$ ,  $b_{k,j}^{(i)}$  и  $A_j^{(i)}$  находятся при помощи метода наименьших квадратов (используется метод скорейшего спуска [4]). Порядок  $M(i, j)$ , как уже не раз отмечалось, не фиксирован, в начале работы алгоритма он принимается равным 1 и начинает расти, пока не будет достигнута заданная точность либо заданное максимальное значение.

Однако новизна работы не в этом. В отличие от классической сети TSK, нахождение  $a_{k,j}^{(i)}$ ,  $b_{k,j}^{(i)}$  и  $A_j^{(i)}$  выполняется не на всех  $x_l = (x_{1l} \dots x_{nl}), l = \overline{1, L}$ , а лишь на основании соответствующего подмножества. Это требует детального пояснения. Цель обучения классической сети TSK - минимизация функционала  $E$ :

$$E = \frac{1}{2} \sum_{l=1}^L (y_l - d_l)^2 \tag{4}$$

где  $y_l$  - реальный выход сети, а  $d_l$  - желаемое значение (стоит отметить, что обучающая выборка данных представляет собой набор пар  $(x_l, d_l), l = \overline{1, L}$ ). Однако функционал (4) задаёт поиск некоторой «усреднённой» закономерности данных, что не всегда адекватно делать на реальных данных. В противовес такому подходу предлагается следующий. Производится кластеризация данных обучающей выборки. Вся обучающая выборка разбивается на  $K$  (где  $K$  - это количество кластеров) подмножеств по признаку принадлежности точек выборки определённому кластеру (наибольшему значению принадлежности). Таким образом, каждая  $(x_l, d_l)$  принадлежит некоторому множеству точек  $S_K$  - каждому отдельному кластеру  $i$  соответствует множество  $S_i$ . А из того, что каждый кластер  $i$  задаёт каждое правило  $i$  (компоненты центра кластера - центры функций принадлежности соответствующих входов для

данного правила), то каждому правилу  $i$  соответствует множество  $S_i$ . Задача обучения сети сводится к минимизации  $K$  функционалов вида:

$$E_{S_i} = \frac{1}{2} \sum_{l=1}^{L_i} (y_l^{(i)} - d_l^{(i)})^2 \quad (5)$$

где  $L_i$  - количество точек обучающей выборки, принадлежащих  $i$ -му кластеру, а индекс  $(i)$  над компонентами  $y_l$  и  $d_l$  обозначает, что  $x_l$  принадлежит кластеру  $i$ .

Необходимо отметить тот факт, что в общем случае верно следующее неравенство:

$$\min E \neq \sum_{i=1}^K \min E_{S_i} \quad (6)$$

Схематически метод формирования описанной в данном разделе системы нечёткого логического вывода можно представить в следующем виде (см. Рис. 1).

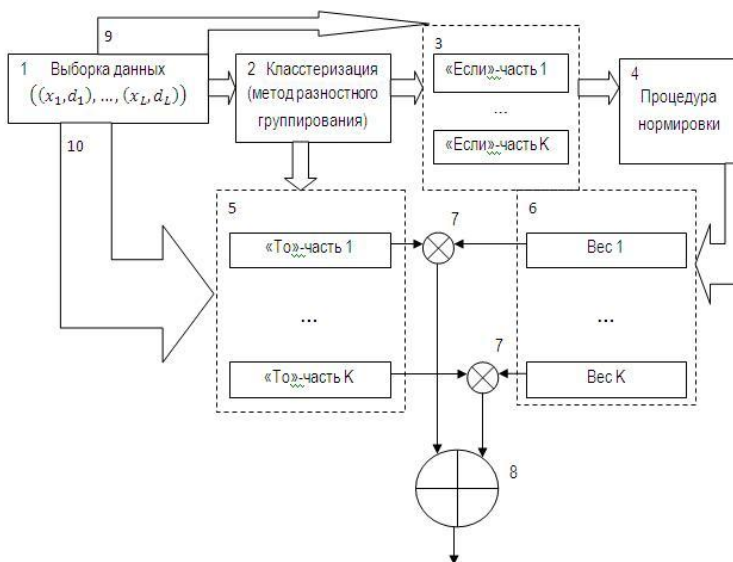


Рис. 1. Схематическое изображение предложенного метода

Описание схемы Рис. 1 следующее.

Этап обучения сети. Имеется выборка данных, состоящая из пар  $(x_l, d_l)$  (блок 1), компоненты  $x_l$  которой подаются на вход блока кластеризации (блок 2). В качестве метода кластеризации выбран метод разностного группирования со сферой влияния равной 0,5 ([2], [3]). В результате кластеризации получается  $K$  кластеров, центры которых становятся центрами функций принадлежности (выбраны функции принадлежности гауссовского вида)  $\mu_{A_j^i}(x_j), i = \overline{1, K}, j = \overline{1, n}$ . Параметры  $\sigma$  ([2], [3]) этих функций принадлежности выбираются одинаковыми для каждого отдельного  $x_j$  для всех правил. Так строится «Если»-часть нечётких правил (Блок 3). Для построения «То»-части нечётких правил, прежде всего, производится разбиение данных обучающей выборки на описанные ранее непересекающиеся подмножества  $S_i$  - разбиение происходит на основании максимальной принадлежности определённой точки выборки определённому кластеру. Далее, для каждого подмножества  $S_i$ , а следовательно, для

каждого правила  $i$  происходит обучение, целью которого является минимизация функционала  $E_{S_i}$ , который вычисляется по формуле (5) (Этот этап символизирует переход к блоку 5 из блока 2 и блок-переход 10). Этап обучения завершён.

Этап использования сети. Компоненты  $x_i$  пар  $(x_i, d_i)$  подаются при помощи перехода 9 на блок 3. Это символизирует расчёт пересечения значения термов для каждого правила. В реализации метода, использованного в работе, применяется пересечение в форме произведения значений функций принадлежности термов. Так получаются веса нечётких правил. Однако для того, чтобы сумма весов правил была всегда равна 1, для каждого  $x_i$  выполняется процедура пересчёта весов (нормировки весов, что, де-факто, является просто делением каждого отдельного веса на сумму всех весов правил), что символизируют блоки 4 и 6. Рассчитываются значения  $y_i$  в «То»-части нечётких правил по полученным в результате обучения сети формулам (эти зависимости представлены символически в выражении (1)). Это символизирует блок 5, в который подаются  $x_i$  из блока 1 по блоку-переходу 10. Далее полученные  $y_i$  умножаются на полученные ранее нормированные веса нечётких правил и суммируются, что даёт выходы сети  $y_i$ . Символически это обозначено операциями-блоками 7 и 8. Далее, применяя анализ значений определённых функционалов качества, делается заключение о том, насколько значения  $y_i$  удалены от  $d_i$ , то есть, проводится анализ качества работы сети, что и является предметом экспериментальных исследований предложенных ниже в данной работе.

## 1.2 Анализ качества работы нейронной сети

Анализ качества работы сети в данной работе будет производиться на основе значений критериев RMSE и MAPE.

Критерий RMSE, при условии, что  $d_l$  - желаемый выход сети, а  $y_l$  - реальный выход сети, где  $l = \overline{1, L}$ ;  $L$  - объём выборки данных вычисляется по формуле:

$$RMSE = \sqrt{\frac{1}{L} \sum_{l=1}^L (y_l - d_l)^2} \quad (7)$$

Данный критерий является оценкой абсолютного отклонения выдаваемых сетью значений от значений, которые она бы должна выдавать в идеале. Очевидным является тот факт, что критерий RMSE является чувствительным к масштабу данных, как и все критерии, оценивающие абсолютное отклонение. Для

объяснения необходимо лишь привести следующих два примера:  $\sqrt{\frac{1}{1} \sum_{l=1}^1 (0,5 - 0,4)^2} = 0,1$ , в то время,

когда  $\sqrt{\frac{1}{1} \sum_{l=1}^1 (95 - 94)^2} = 1$ . То есть, одно и тоже значение данного критерия может говорить как о высоком качестве работы сети, так и очень плохом, так как масштаб задаёт вес каждой отдельной цифры в числе.

Критерий MAPE при условии, что  $d_l \geq 0$  - желаемый выход сети, а  $y_l > 0$  - реальный выход сети, где  $l = \overline{1, L}$ ;  $L$  - объём выборки данных вычисляется по формуле:

$$MAPE = \frac{1}{L} \sum_{l=1}^L \frac{|y_l - d_l|}{d_l} \quad (8)$$

Данный критерий является оценкой относительного отклонения выдаваемых сетью значений от желаемых значений. Критерий, в отличие от рассмотренного ранее, не чувствителен к масштабу чисел,

хотя и определяет косвенно количество верно распознанных цифр в числе в порядке убывания их значимости. Если значение критерия умножить на 100%, то будет получен средний процент значений отклонений значений выдаваемых сетью от желаемых значений.

Аббревиатуры критериев расшифровываются следующим образом. RMSE – root mean square error – корень квадратный из среднеквадратического отклонения. MAPE – mean absolute percentage error – средняя абсолютная процентная ошибка. Слово «абсолютная» вызывает недоумение – ведь критерий относительный, однако данное слово здесь обозначает, что берётся сумма модулей отклонений поделенных на соответствующие реальные значения, на что указывает слово «процентная». В работе будут использованы английские аббревиатуры из-за того, что они гораздо чаще используются в научной литературе, чем русские.

Ещё следует отметить такой момент, что, вследствие наличия у обоих критериев компоненты  $y_i - d_i$ , в экспериментах можно будет наблюдать некоторое подобие динамики критериев в зависимости от изменения параметров экспериментов. Однако необходимо изучение обоих критериев, так как они выполняют разные функции. Функция RMSE – описать то, насколько сеть хорошо обучена, а цель MAPE – описать то, насколько велико отклонение реальных значений, выдаваемых сетью от желаемых значений.

---

## 2 Экспериментальные исследования

---

Данный раздел посвящён экспериментальным исследованиям сети и состоит из двух подразделов. Первый подраздел является теоретическим исследованием свойств сети и свойств экспериментальной среды, которые влияют на качество функционирования сети. Второй подраздел посвящён приложению рассматриваемой нейросетевой архитектуры и метода обучения к задаче прогноза значений котировок валютных пар EUR/GBP, EUR/USD, USD/CHF, USD/JPY.

---

### 2.1 Теоретические исследования

---

Данный подраздел состоит из двух подразделов. Первый подраздел посвящён выявлению свойств рассматриваемой нейросети и среды эксперимента, которые влияют на качество работы сети путём приложения её к случайно сгенерированным данным. Второй раздел посвящён анализу сети на основании приложения её к периодическим зависимостям.

---

#### 2.1.1 Экспериментальное исследование сети на случайных входных и выходных данных

---

Для выявления и анализа скрытых свойств функционирования сети проводилось её применение для анализа и поиска зависимостей в данных, в которых входы и выходы сети были случайными равномерно распределёнными величинами. Следует сразу заметить, что в данном случае, в отличие от последующих в данной работе исследований, значения обоих критериев (RMSE и MAPE) будут очень большими. И это обоснованно – нельзя искать закономерность там, где её нет. Отличие данного исследования от последующих в том, что для него не важны конкретные значения критериев, а важно – поведения данных значений в зависимости от изменений параметров экспериментальной модели.

Эксперимент построен следующим образом. Количество входов сети является параметром экспериментальной модели и изменяется от 2 до 4. Количество точек выборки – параметр эксперимента и принимает значения 10, 20, 30, ..., 80, 90, 100. Значения точек выборок данных генерируются с помощью генератора случайных чисел (равномерное распределение); значения принадлежат интервалу – [0;1]. Делать выводы, проведя лишь одно исследование для конкретных значениях параметров эксперимента – неадекватно, в силу случайной природы исследуемых данных; в то же время, проведение огромного

числа опытов при конкретных значениях параметров эксперимента и расчёт их значений и дальнейший расчёт средних по значениям критериев – тоже неадекватно, так как с ростом количества рассмотренных вариантов падает влияние каждого отдельного варианта на среднее значений критериев. Из этих рассуждений для количества обучений и проверки сети при конкретных значениях параметров было выбрано число 10 – не один опыт и, в то же время, каждый отдельный опыт имеет значительное влияние на среднее критериев. Таким образом, как уже стало понятно из предшествующих объяснений, для каждого конкретного параметров экспериментов случайно генерируются данные выборки и количество таких выборок данных равно 10. Далее сеть обучается на данных выборках, считаются значения критериев RMSE и MAPE для каждой отдельной выборки, а после – происходит усреднение данных критериев по всем 10 реализациям. Деление выборок на обучающие и проверочные не проводилось из-за бессмысленности подобного деления для данного эксперимента. Параметры самой сети: максимальный порядок частичного описания – 3, допустимая погрешность – 0,00001. Выбор столь малых значений был обусловлен желанием изучить свойства сети на данном эксперименте так, чтобы минимизировать влияние самих параметров сети на результаты.

Результаты опытов представлены в следующей таблице.

Таблица 1. Результаты экспериментальных исследований на выборках данных, состоящих из случайных чисел

Количество точек в выборке	Количество входов сети					
	2		3		4	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
10	0,141	0,987	0,104	1	0,0586	0,49
20	0,326	5,36	0,154	2,24	0,216	2,53
30	0,22	4,2	0,174	3,27	0,158	3,67
40	0,166	3,78	0,126	2,93	0,117	2,35
50	0,147	4,04	0,141	3,19	0,157	3,12
60	0,145	3,8	0,115	2,77	0,111	3,37
70	0,11	2,98	0,116	3,72	0,116	3,85
80	0,117	3,46	0,113	3,46	0,097	4,21
90	0,0943	3,65	0,12	3,48	0,0854	3,6
100	0,109	4,26	0,0809	3,62	0,0926	3,64

Прежде всего, необходимо перечислить факты, которые следуют из приведённой выше таблицы данных. Значения критериев RMSE и MAPE говорят, что в то время, когда сеть достаточно хорошо настроена (следует напомнить, что данные из диапазона [0;1]) – об этом говорят значения критерия RMSE - сеть не делает ни одного точного прогноза и даёт очень большие ошибки – об этом говорят значения критерия MAPE. Вторым важным фактом является то, что при количестве точек выборки равном 10, значения критериев значительно отличаются от значений критериев при прочих значениях данного параметра

эксперимента – сеть относительно точно распознаёт некоторые ситуации. Третий факт - MAPE не проявляет никаких закономерностей при изменениях значений параметров эксперимента, кроме рассмотренного выше варианта. Четвёртый – значения критерия RMSE падают с ростом объёма выборки и ростом числа входов сети.

Объяснения первого, третьего и четвёртого факта лежат в следующих рассуждениях. Сеть обучается на критерии RMSE, а данный критерий характеризует отклонение точек от «некоторого тренда» в данных – и применение данного критерия адекватно для анализа случайных чисел. С ростом объёма выборки получаемый «тренд» становится ближе к «идеальному тренду» и, следовательно, описательное качество сети по данному критерию улучшается. С ростом числа входов сети – ситуация аналогична. Однако критерий MAPE для анализа работы сети на случайных числах применять нельзя, так как он является мерой отклонения реальных результатов работы сети от предложенной сетью нестатистической закономерности, а как можно требовать от сети такую закономерность, когда она не существует а priori. Второй факт говорит о том, что сеть при данных конфигурациях и количестве входов сети лишь запоминает все точки выборки – «зубрит» их, а не находит в них скрытые закономерности; однако и «зубрёжка» эта не всегда эффективна.

### 2.1.2 Экспериментальное исследование сети на периодических зависимостях выхода от входов

Главная цель проведения данного экспериментального исследования – определить: способна ли сеть описывать периодические зависимости.

Для данного эксперимента, по сравнению с экспериментом со случайными числами, на первый план выходят значения критериев RMSE и MAPE. Однако поведения критериев на периодических зависимостях тоже важны.

Эксперимент был построен следующим образом. Параметры эксперимента идентичны параметрам эксперимента со случайными числами и, наверное, нет смысла их детально описывать. Следует напомнить, что это – количество входов сети и число точек выборки данных. Деление выборки данных на обучающую и проверочную не производилось, поскольку для данного эксперимента важно исследование описательного свойства сети.

Отдельно следует рассмотреть то, как строились выборки данных. Прежде всего, определялось количество входов сети для исследования. Далее задавался шаг для построения сетки точек пространства. Следует отметить, что значение каждого входа сети принадлежало интервалу [0;1]. После этого, с учётом выбранного шага, строилась сетка, покрывающая единичный гиперкуб входных данных. Данная сетка дополнялась случайным количеством дублей случайно выбранных точек сетки. Из построенного дискретного пространства входных векторов случайным образом выбирались точки в количестве, задаваемым вторым параметром эксперимента – количеством точек выборки. Выход сети строился по формуле:

$$y_j = \sum_i f_i(x_{ij}) \quad (9)$$

где

$$f_i(x_{ij}) = rand_0 + \sum_{k=1}^2 (rand_{2k-1} \cdot \cos kx_{ij} + rand_{2k} \cdot \sin kx_{ij}) \quad (10)$$

В данном эксперименте  $rand_i$  - были случайными целыми числами, выбираемыми из диапазона [-5;5].



Аналогично эксперименту со случайными числами, опираясь на те же рассуждения, для конкретных значений параметров эксперимента вышеуказанные выборки строились по 10 раз, рассчитывались значения критериев качества сети RMSE и MAPE и в нижеприведённую таблицу записывались усреднённые значения критериев. Параметры сети аналогичны параметрам из эксперимента со случайными числами.

Следующая таблица содержит результаты проведения опытов.

Таблица 2. Результаты экспериментальных исследований на выборках данных периодических зависимостей

Количество точек в выборке	Количество входов сети					
	2		3		4	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
10	0,263	0,582	0,367	0,326	0,338	0,168
20	0,355	0,265	0,364	0,375	0,283	0,174
30	0,207	0,425	0,27	0,897	0,243	0,901
40	0,153	0,233	0,228	0,608	0,249	0,335
50	0,153	0,607	0,208	0,623	0,265	0,441
60	0,121	0,393	0,233	0,929	0,394	33,7
70	0,0888	0,43	0,197	0,498	0,211	0,804
80	0,103	0,189	0,143	0,711	0,255	0,635
90	0,0795	0,205	0,163	0,705	0,204	0,614
100	0,0682	0,88	0,154	0,472	0,19	1,19

Полученные результаты очень интересны и они ни в коем случае не говорят о неспособности сети описывать периодические зависимости. В постановке эксперимента был заложен один очень важный момент, который дал такие большие значения MAPE и о котором будет сказано чуть позже. Сейчас следует сразу сказать, что с устранением данного момента сеть давала значения критерия MAPE равное  $10^{-2}$ , когда использовался вывод TS и  $10^{-5}$ , когда использовался нечёткий вывод Белмана-Задэ. А момент этот – в покрытии точками выборки всего входного пространства. Прежде всего, введение в пространство дублей точек сетки придавало отдельным точкам выборки, в случае попадания в выборку двух одинаковых векторов – больший вес, а значит, сеть больше обучалась под эти точки, или вернее сказать точку выборки, чем под другие точки. И попадание в выборку нескольких таких точек и приводит к среднему значению MAPE равному, например, 33,7. Также, следует отметить, что при построении сетки для 2 входов, 3 и 4 входов использовался один и тот же шаг – 0,01, а количество точек выборки – тоже было одно и то же, и являлось параметром эксперимента. Таким образом, с ростом числа входов сети росла вероятность того, что будут набраны практически все точки из одного подпространства входного пространства и несколько точек относительно удалённых от данного пространства. А это обуславливает рост средних значений параметра MAPE в эксперименте с ростом числа входов сети. Этот факт является

очень поучительным – он говорит о том, что для корректного функционирования сети требуется как можно больше обучающих данных, которые как можно полнее описывают входное пространство. Ещё один нюанс, который следует объяснить, почему при 100 точках выборки с двумя входами средний MAPE получился равным 0,88, что означает, де факто, 88% ошибок, а проблема эта в наибольшей концентрации в выборке из 100 точек дублей, которые мешают верной работе сети.

В окончании подраздела следует ещё раз подчеркнуть тот факт, что подобная ситуация смоделирована искусственно и при правильном построении выборки, которая очень хорошо представляет определённую периодическую зависимость, были получены порядки точности  $10^{-5}$ , при условии, что ограничение для алгоритма обучения сети на точность было равно  $10^{-5}$ . Также следует отметить, что полагаться лишь на критерий RMSE как на единственный критерий качества работы сети нельзя, так как он описывает лишь то, насколько хорошо обучена сеть и насколько значения, выдаваемые сетью, близки к реальным значениям в среднеквадратическом смысле. Анализ качества работы сети всегда должен подкрепляться анализом критерия MAPE.

---

## 2.2 Применение сети к анализу данных котировок валют

---

Для проверки эффективности работы рассматриваемой нейронной сети относительно практических задач была выбрана задача прогноза котировок валют на основании их предыстории.

Постановка конкретной задачи, которая решалась в данной главе следующая. Имеется выборка данных дневных котировок валютных пар EUR/GBP, EUR/USD, USD/JPY, USD/CHF за период с 25.03.2009 по 24.03.2010. Необходимо произвести прогноз значения котировки на шаг вперёд на основании её предыстории.

Эффективность работы рассмотренной ранее нейросетевой архитектуры применительно к решению данной задачи оценивалась по двум критериям: RMSE и MAPE. Следует отдельно отметить, что, согласно алгоритму обучения сети, сеть настраивалась с целью минимизации, де-факто, критерия RMSE. Таким образом, критерий MAPE выступает в роли «независимого эксперта», то есть оценивает качество работы сети, при этом не участвуя в её обучении. Выходит, RMSE является критерием эффективности настройки сети, в то время, когда MAPE оценивает эффективность её приложения.

Эксперимент был поставлен следующим образом. Из указанных выше выборках данных котировок валют составлялись новые выборки данных для применения рассматриваемой нейросетевой архитектуры. Выбиралось количество периодов в предыстории, на основании которых необходимо сделать прогноз – в эксперименте это: 2, 3, 4 дня – это входы сети. Выходом сети являлось значение котировки в последующий момент времени. Указанные выборки данных котировок валют делились на обучающие и проверочные в соотношениях: 10:90, 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, 80:20, 90:10 соответственно. Это означает, что, например, для котировки EUR/USD соотношение 10:90: 10% выборки – обучающая, а 90% выборки – проверочная. Сеть обучалась на обучающих выборках и фиксировались значения вышеуказанных критериев на обучающих и проверочных выборках при данных их соотношениях и при различном числе входов сети.

Результаты эксперимента для котировок EUR/GBP, EUR/USD, USD/JPY, USD/CHF представлены ниже.

Таблица 3. Значения критериев для валютной пары EUR/GBP

обуч.:пров	Количество входов											
	2				3				4			
	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.
10:90	0,002608	0,013772	0,006047	0,071061	0,002109	0,010416	0,007302	0,084142	0,001473	0,007201	0,006759	0,079349
20:80	0,000641	0,004802	0,000449	0,006671	0,002954	0,02282	0,001722	0,026742	0,000654	0,004785	0,000593	0,0081
30:70	0,000545	0,005095	0,000351	0,004864	0,000463	0,004413	0,000343	0,004695	0,000461	0,004444	0,000362	0,005079
40:60	0,000372	0,003944	0,000341	0,004363	0,000389	0,004317	0,00034	0,00437	0,000403	0,004289	0,000413	0,004969
50:50	0,000353	0,004268	0,000429	0,004994	0,000339	0,003984	0,000343	0,003943	0,000337	0,003889	0,000318	0,003563
60:40	0,000345	0,004362	0,000342	0,003619	0,000325	0,004067	0,000333	0,003463	0,000338	0,004351	0,000384	0,004155
70:30	0,000318	0,004367	0,000372	0,003392	0,000287	0,003816	0,000367	0,003277	0,000331	0,004526	0,000482	0,004493
80:20	0,000287	0,004209	0,00051	0,003865	0,000265	0,003797	0,000475	0,003548	0,000262	0,003702	0,000469	0,003473
90:10	0,000245	0,003676	0,000681	0,003759	0,000251	0,00384	0,000615	0,003481	0,000312	0,005037	0,001098	0,00593

Таблица 4. Графическое представление результатов экспериментов для валютной пары EUR/GBP

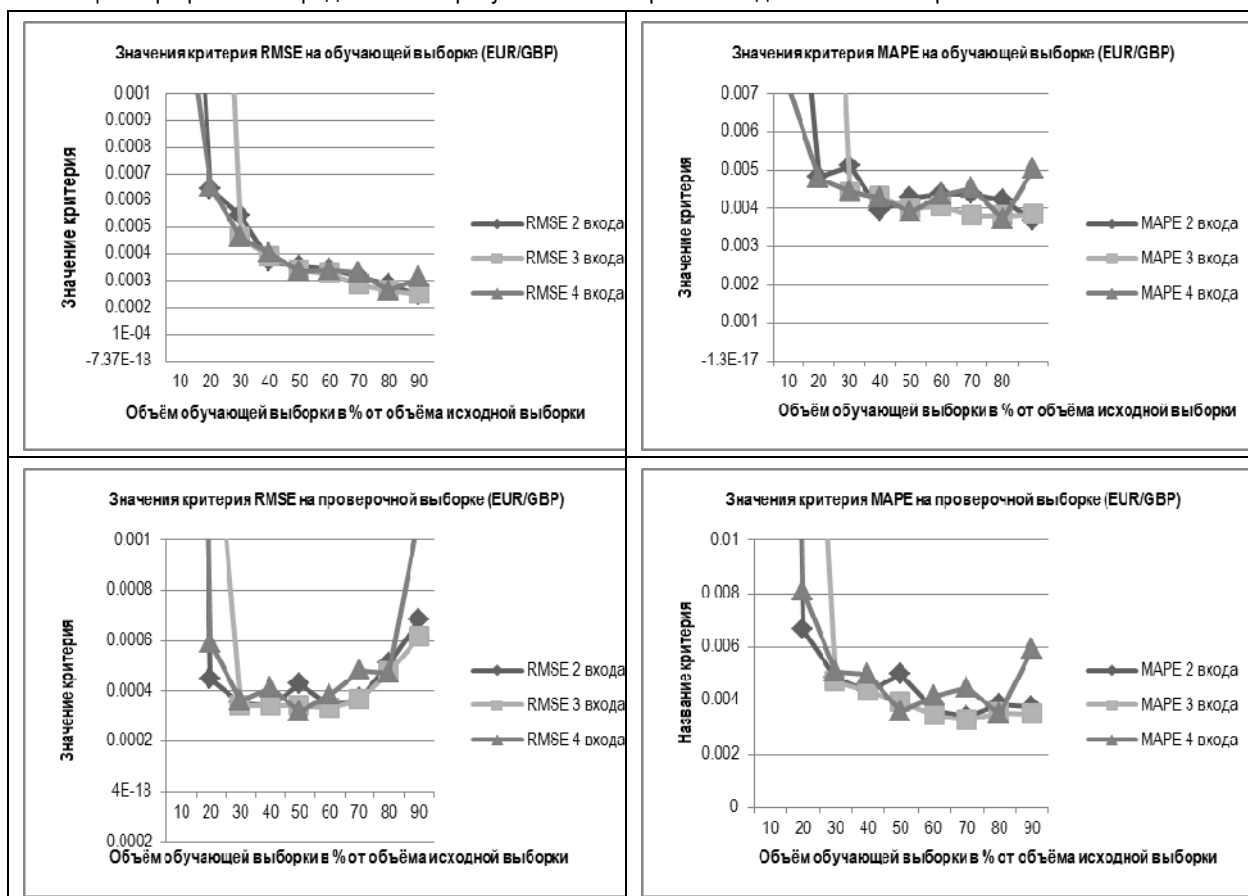


Таблица 5. Значения критериев для валютной пары EUR/USD

обуч.:пров	Количество входов											
	2				3				4			
	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.
10:90	0,002596	0,008119	0,02675	0,22567	0,001805	0,006454	0,03573	0,33629	0,001937	0,007164	0,038888	0,39704
20:80	0,001114	0,00528	0,004944	0,034266	0,001904	0,008783	0,013127	0,095096	0,001113	0,005631	0,003617	0,026909
30:70	0,001036	0,006263	0,003192	0,024584	0,000977	0,005755	0,002153	0,016265	0,000884	0,005047	0,002987	0,021669
40:60	0,000811	0,005485	0,002073	0,013896	0,000723	0,004796	0,001985	0,013074	0,000727	0,004817	0,003997	0,025057
50:50	0,000862	0,006316	0,00123	0,007417	0,001077	0,008568	0,00163	0,010054	0,001065	0,0083	0,002203	0,014389
60:40	0,000743	0,006218	0,000734	0,004903	0,000584	0,004721	0,000883	0,006112	0,000547	0,004388	0,000778	0,0052
70:30	0,000543	0,004711	0,000687	0,003757	0,000506	0,004413	0,000798	0,004429	0,000611	0,005103	0,000882	0,005125
80:20	0,000485	0,004515	0,00082	0,003721	0,00047	0,00436	0,001017	0,004823	0,00055	0,005184	0,001069	0,005254
90:10	0,000534	0,005175	0,001559	0,006103	0,00042	0,003925	0,00121	0,003785	0,000425	0,004014	0,00113	0,003547

Таблица 6. Графическое представление результатов экспериментов для валютной пары EUR/USD

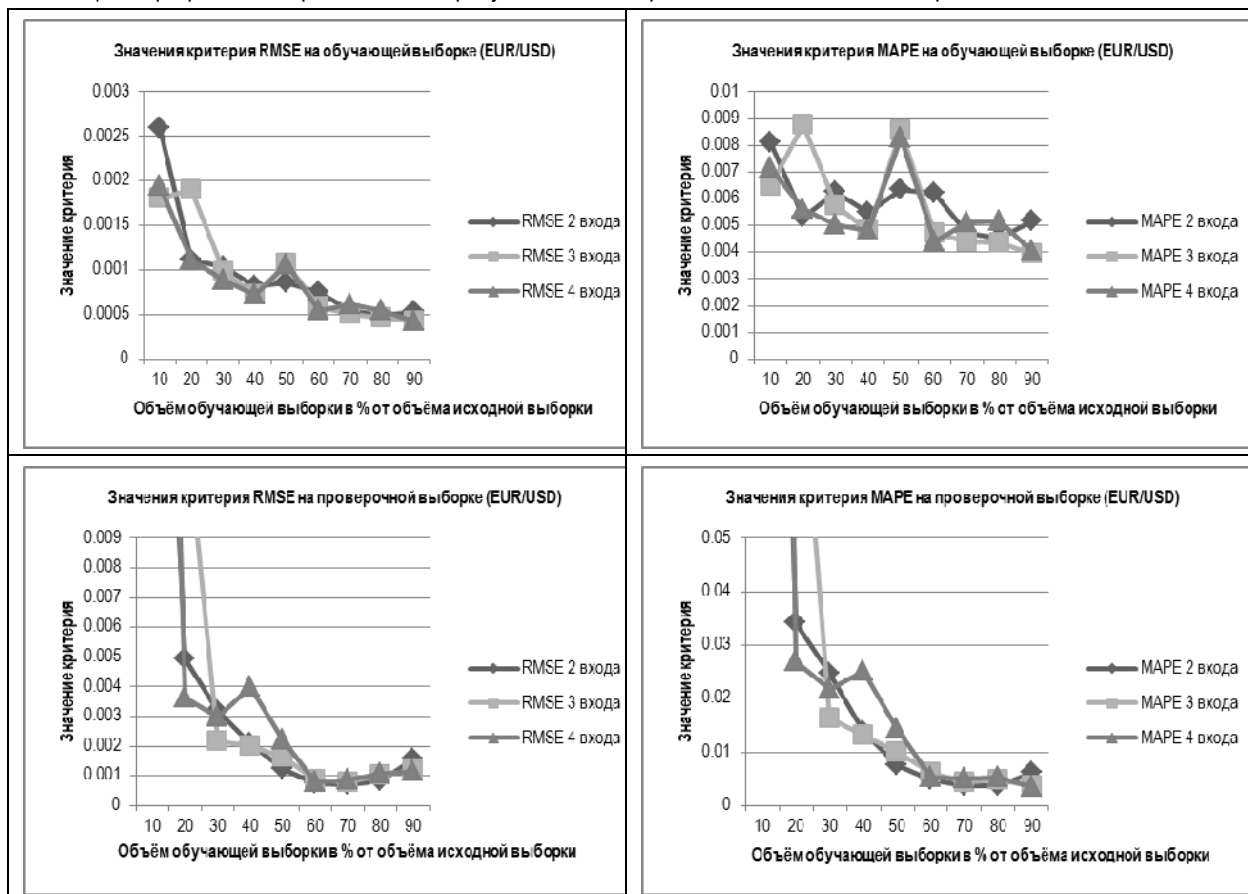


Таблица 7. Значения критериев для валютной пары USD/JPY

обуч:пров	Количество входов											
	2				3				4			
	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.
10:90	0,10798	0,005251	0,45759	0,078326	0,11526	0,005468	1,1285	0,19462	0,15089	0,007299	1,8272	0,31217
20:80	0,08949	0,005902	0,32706	0,042931	0,094153	0,00644	0,11509	0,016595	0,11621	0,007547	0,96241	0,1484
30:70	0,071606	0,005814	0,052713	0,005479	0,099422	0,008136	0,14912	0,021071	0,078494	0,00636	0,16827	0,020172
40:60	0,067385	0,006201	0,42312	0,051221	0,077731	0,007054	0,65208	0,080762	0,061552	0,00561	0,72667	0,089607
50:50	0,051304	0,005456	0,063959	0,007649	0,050244	0,004964	0,14253	0,013533	0,049009	0,005052	0,056701	0,005995
60:40	0,052345	0,006029	0,06225	0,005453	0,045704	0,005369	0,051846	0,005141	0,04887	0,005757	0,068554	0,00615
70:30	0,04346	0,005542	0,054179	0,004602	0,041365	0,005181	0,054236	0,004717	0,050213	0,00674	0,085693	0,008241
80:20	0,04996	0,007199	0,069813	0,004987	0,044928	0,006461	0,069315	0,004746	0,049953	0,007028	0,059017	0,003899
90:10	0,042724	0,006117	0,089295	0,004181	0,039285	0,005608	0,08961	0,004133	0,12461	0,019378	0,53685	0,033844

Таблица 8. Графическое представление результатов экспериментов для валютной пары USD/JPY

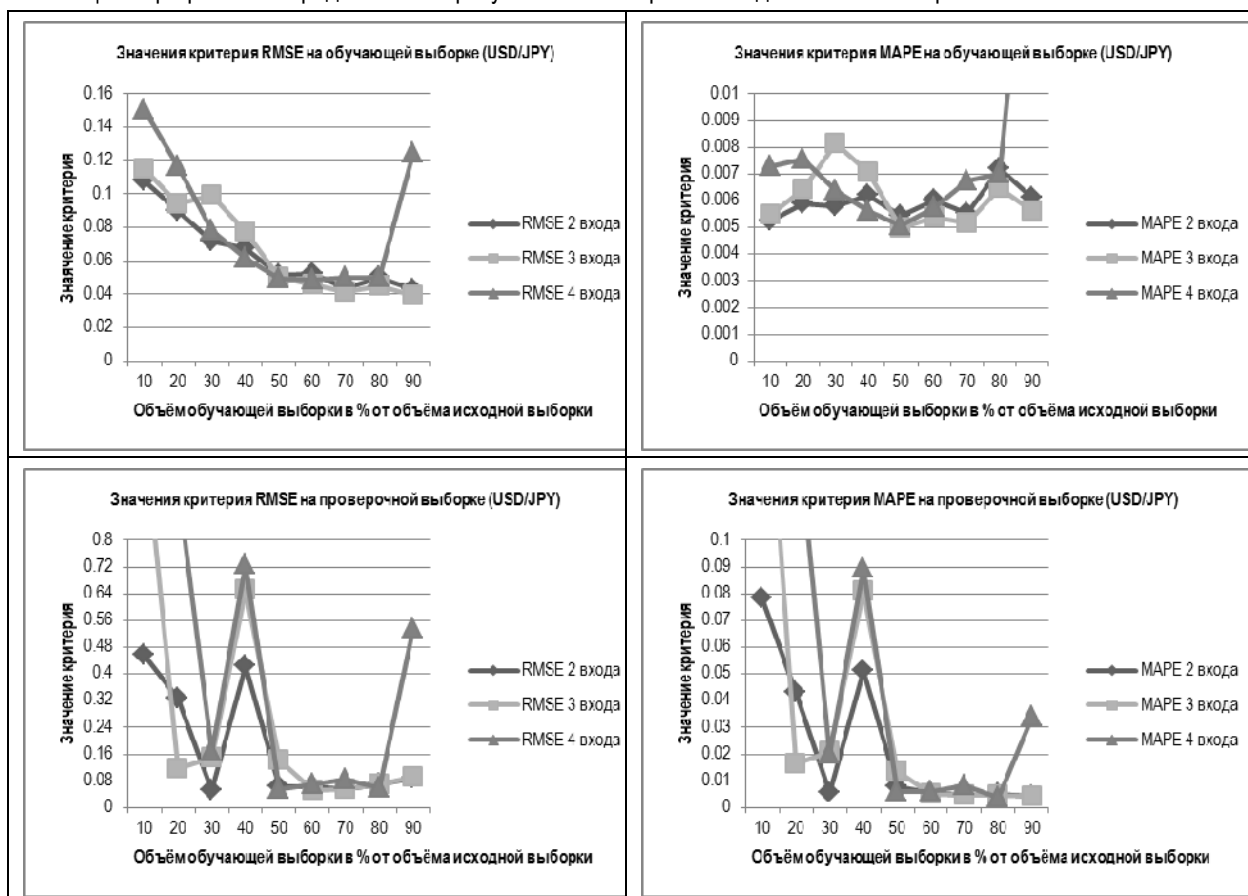
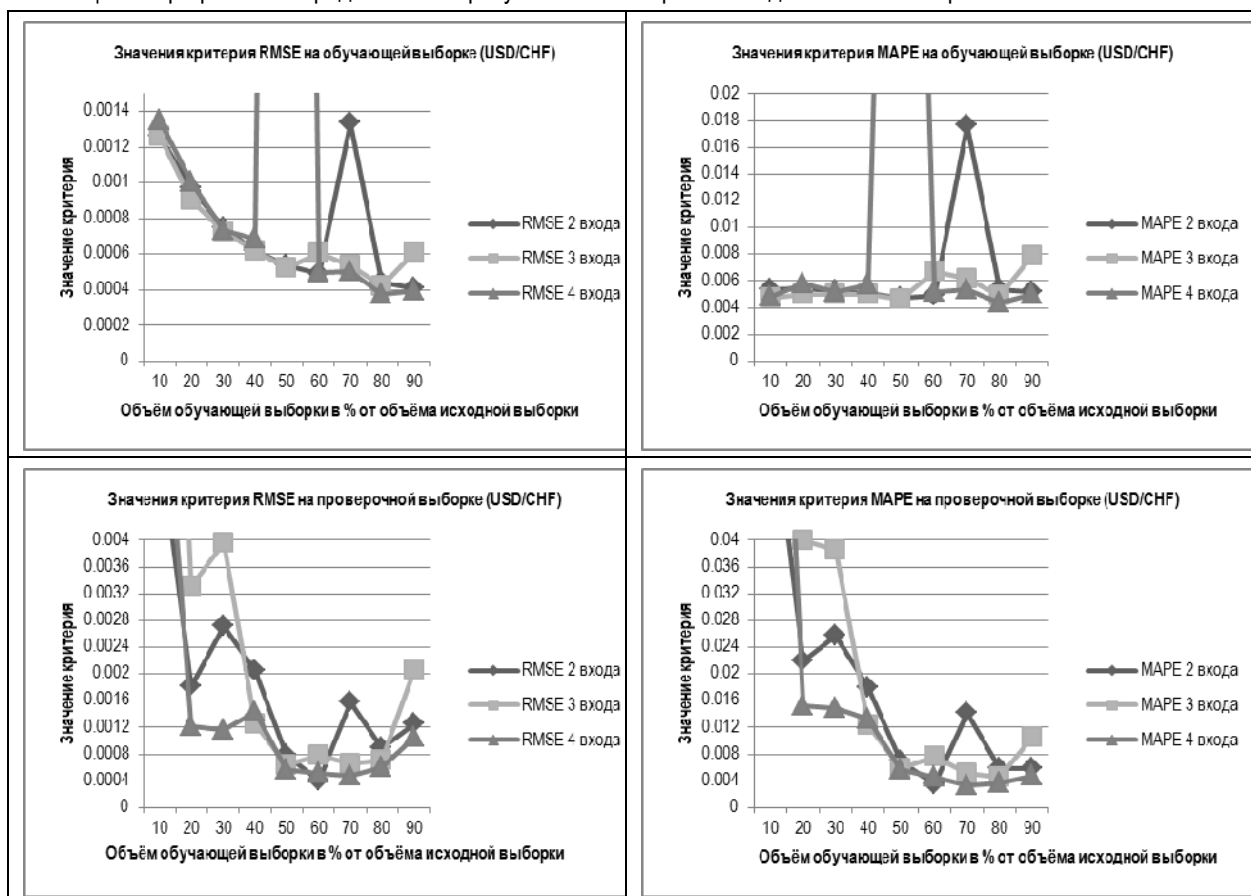


Таблица 9. Значения критериев для валютной пары USD/CHF

обуч.:пров	Количество входов											
	2				3				4			
	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.
10:90	0,00126	0,005325	0,00552	0,062745	0,001263	0,004744	0,01377	0,17941	0,00135	0,004777	0,007384	0,11671
20:80	0,000971	0,005424	0,00181	0,021933	0,000907	0,00501	0,003308	0,03995	0,001003	0,005767	0,001212	0,015209
30:70	0,000748	0,005233	0,002721	0,025682	0,000727	0,00496	0,003964	0,038628	0,000736	0,005187	0,001167	0,014883
40:60	0,000621	0,005054	0,002051	0,017899	0,000618	0,00504	0,00126	0,012279	0,000687	0,005728	0,001437	0,013268
50:50	0,000537	0,004729	0,000776	0,006943	0,000527	0,00464	0,000639	0,00588	0,014786	0,090375	0,000573	0,005672
60:40	0,000491	0,004889	0,000412	0,003576	0,000604	0,006677	0,000782	0,007651	0,000499	0,005168	0,000509	0,004593
70:30	0,001337	0,017609	0,001571	0,014083	0,000534	0,006209	0,000648	0,005141	0,000503	0,005384	0,000467	0,003291
80:20	0,000443	0,005329	0,000882	0,005893	0,000421	0,004925	0,000708	0,00454	0,000376	0,004403	0,000603	0,00368
90:10	0,000411	0,005218	0,001256	0,00584	0,00061	0,007912	0,002062	0,010478	0,000397	0,005001	0,001056	0,004748

Таблица 10. Графическое представление результатов экспериментов для валютной пары USD/CHF



Прежде чем переходить к анализу полученных результатов, следует описать факты, которые были получены в результате экспериментов. Во-первых, очевидно, что для обоих критериев на обучающих выборках происходит постепенное падение их значений. Во-вторых, значения критериев на проверочных выборках сначала падают, наступает «период относительной стабильности», а после – начинают расти. Как показывают эксперименты, период стабильности в среднем лежит в пределах соотношений обучающей и проверочной выборки 50:50 – 70:30 соответственно. В-третьих, для обоих критериев на обучающей и проверочной выборке при объёме обучающей выборки 10-20% от исходной выборки их значения заметно отличаются от значений при других объёмах обучающей выборки – они значительно больше. В-четвёртых, для критерия MAPE, при объёмах обучающих выборок больше 50%, для всех котировок на обучающих выборках порядок значений остаётся один и тот же и аналогичная ситуация наблюдается, если сравнивать между собой значения данного критерия для котировок на проверочных выборках, в то время, когда для RMSE для разных котировок, сравнивая его значения на соответствующих выборках и при соответствующих объёмах обучающих выборок он разный. В-пятых, оба критерия на одних и тех же выборках демонстрируют одинаковое поведение. В-шестых, отмечается незначительный рост качества работы сети с ростом числа входов.

Анализ результатов будет, фактически, объяснением выделенных фактов. Для объяснения всех последующих фактов необходимо, прежде всего, объяснить: почему при определённых соотношениях обучающей и проверочной выборок поведение и значения критериев значительно отличаются от поведения и значений при других соотношениях. Ответ лежит в следующем: с падением числа данных, участвующих в расчёте критерия, растёт вес каждой отдельной точки выборки этих данных. Таким образом, происходит некоторая переоценка значимости (именно при указанных соотношениях) отдельно взятой точки для обучения сети, или для оценки качества её работы. Это – объяснение второго и третьего факта. Первый факт, с учётом предыдущего вывода, говорит, что с ростом объёма обучающей выборки растёт эффективность функционирования сети на обеих выборках. Однако незначительный рост качества свидетельствует о существовании «периода насыщения», когда рост объёма обучающей выборки даёт прирост показателей качества, которыми можно пренебречь. Шестой факт является отчасти следствием ограничения на максимальное число итераций при обучении сети – сеть с большим числом входов требует большего времени на обучение, – а отчасти констатацией аналогичного факта существования «периода насыщения» для количества входов сети, так как отмечается незначительный прирост качества. Объяснение четвёртого факта уже приводилось в подразделе 1.2: RMSE является чувствительным к масштабу исследуемых величин, в то время, когда MAPE является практически нечувствительным. Пятый факт – соответствие поведения обоих рассматриваемых критериев на одних и тех же выборках – свидетельствует об адекватности выбора критерия обучения сети для конкретных данных и об отсутствии необходимости в его модификации. Главный вывод, с учётом всех вышеуказанных рассуждений и выводов предыдущих глав – сеть достаточно точно распознаёт поведение рынка и с приемлемой точностью угадывает котировки валют.

---

### **Выводы и перспективы дальнейших исследований**

---

Прежде всего, необходимо отметить, что приложение сети к реальным котировкам валютных пар показало эффективность сети применительно к задаче прогноза значений валютных котировок на основании их предыстории. Данное заключение было сделано на основании полученных значений критериев качества, которые приведены в работе. Однако сразу же следует отметить, что необходимы дальнейшие исследования, направленные на усовершенствования нейросетевой архитектуры и методики её применения. Это следует из того, что лучшее значение критерия MAPE было равно где-то 0,003, в то время, когда практически полным распознаванием ситуации является его значение меньше чем 0,0001.

Это улучшение может быть получено путём предварительного анализа и преобразования данных, которые поступают на вход сети. Также значения параметра  $\sigma$  в функциях принадлежности системы ([2], [3]) были для каждого отдельного входа сети фиксированными. Однако возможно устранение этого недостатка путём дообучения сети путём оптимизации градиентным методом для каждого нечёткого правила сети этих параметров. Оптимизация должна происходить с целью максимизации значений итоговых весов правил на точках соответствующего подмножества обучающей выборки. Кроме проведения модификации среды использования сети необходимо также произвести сравнительный анализ результатов полученных в данной работе с результатами применения к вышеуказанной задаче прочих нейросетевых архитектур, в особенности, архитектур, реализующих нечёткий вывод Такаги-Сугено (TSK, радиальные базисные нейронные сети, сети на нео-фази нейронах и пр.). Всё это будет предметом дальнейших исследований. Ещё необходимо отметить, что выбор критериев RMSE и MAPE для анализа сети в данной работе был продиктован целью исследования того, насколько точно сеть распознаёт значение конкретной котировки. То есть, ошибкой считаются и значения большие и значения меньше реального значения котировки. Однако необходимо учитывать тот факт, для чего производится анализ – с целью дальнейшей покупки либо продажи валюты. В этом случае понятие «ошибка» значительно меняет свой смысл.

В заключение следует ещё раз повториться, что, как показывают эксперименты, сеть при предложенных модификациях способна относительно точно угадывать котировки валют и вполне применима для практического использования.

---

### Благодарности

Статья частично финансирована из проекта ITHEA XXI Института Информационных теорий и Приложений FOI ITHEA и консорциума FOI Bulgaria ([www.ithea.org](http://www.ithea.org), [www.foibg.com](http://www.foibg.com))

---

### Литература

1. Дневные котировки валют за период с 25.03.2009 по 24.03.2010 взяты со странички <http://www.finam.ru/analysis/export/default.asp>
2. Зайченко Ю.П. Основы проектирования интеллектуальных систем. Навчальний посібник. – К.: Видавничий Дім «Слово», 2004. – С. 352
3. Зайченко Ю.П. Нечёткие модели и методы в интеллектуальных системах. Учебное пособие для студентов высших учебных заведений. – К.: «Издательский Дом «Слово»», 2008. – С. 344
4. Зайченко Ю.П. Дослідження операцій. Підручник. Сьоме видання, перероблене та доповнене. – К.: Видавничий Дім «Слово», 2006. – С. 816

---

### Информация про автора



**Мурга Николай Алексеевич** – аспирант Национального технического университета Украины «КПИ», адрес электронной почты: [murga.nicholas@gmail.com](mailto:murga.nicholas@gmail.com)

**Основные сферы научных исследований автора:** применение теории нечёткой логики и теории детерминированного хаоса к анализу финансовых рынков.



## РАЗРАБОТКА АВТОМАТИЗИРОВАННОЙ ПРОЦЕДУРЫ СОВМЕЩЕНИЯ ИЗОБРАЖЕНИЙ ПРОИЗВЕДЕНИЙ ЖИВОПИСИ В ВИДИМОМ И РЕНТГЕНОВСКОМ СПЕКТРАЛЬНЫХ ДИАПАЗОНАХ

Дмитрий Мурашов

**Аннотация:** Разработана процедура автоматизированного совмещения фотографий и рентгенограмм произведений живописи. В качестве контрольных точек используются локальные экстремумы яркости, найденные на размытых гауссовым ядром изображениях. Для сопоставления найденных характерных точек изображений используется алгоритм на основе метода SVD-сопоставления, дополненный итерационной процедурой исключения ложных соответствий. Для совмещения использована модель проективных преобразований. Разработанная процедура позволяет обеспечить точность совмещения изображений в пределах одного пиксела в контрольных точках.

**Ключевые слова:** анализ и обработка изображений, совмещение многоспектральных изображений, изображения произведений живописи.

**ACM Classification Keywords:** Computing Methodologies - Image Processing and Computer Vision - Applications

---

### Введение

Рассматривается задача, связанная с анализом изображений произведений изобразительного искусства, полученных в разных спектральных диапазонах, и используемых при исследовании истории создания картин, атрибуции и реставрации. Одним из аспектов таких исследований является анализ информации, скрытой под верхними красочными слоями. Так, например, рентгенограмма позволяет увидеть детали сразу всех слоев картины: красочные слои, сделанные в разное время автором и реставраторами, фактуру и дефекты холста, элементы конструкции подрамника и др. [Kirsh, 2000]. Тяжелые металлы, не пропускают рентгеновские лучи, что позволяет видеть красочные слои, выполненные, например, свинцовыми белилами. Исследования с инфракрасным излучением проявляют углеродосодержащие красители (наброски, сделанные углем, чернилами). Ультрафиолетовое излучение позволяет увидеть участки, ранее подвергавшиеся реставрации, и ряд дефектов красочного слоя. Для эффективного выявления информации, скрытой под верхними слоями краски, необходимо автоматизировать операции совмещения, сравнения и анализа совмещенных изображений. В мировой музейной практике широко применяются компьютерные технологии анализа цифровых многоспектральных изображений произведений искусства [Martinez, 2002], [Maitre, 2001], [Stork, 2009], [Kirsh, 2000]. Так в работе [Heitz, 1990] представлен метод автоматизированного поиска скрытой информации по фотографии картины и ее рентгенограмме. Совмещение изображений производится по контрольным точкам, выбираемым вручную. В работе [Kammerer, 2004] представлена программная система для сравнения изображений видимых и скрытых слоев живописи. Система предназначена для исследования истории создания произведений на основе анализа первичных авторских эскизов и сравнения с окончательными вариантами картин. Входной информацией системы являются цифровые цветные изображения в видимом диапазоне и изображения, полученные в инфракрасном (700-1100 нм) диапазонах спектра. Изображения формируются одной и той же CCD камерой. Системой выполняются операции (а) совмещения изображений и компенсации искажений, возникающих при съемке, и которые моделируются аффинным преобразованием; (б)

комбинирование изображений. В большинстве разработанных систем автоматизированы операции совмещения изображений в видимом, инфракрасном и ультрафиолетовом диапазонах. Рентгеновские изображения имеют ряд особенностей, осложняющих автоматизацию их совмещения с другими изображениями. Задача усложняется тем, что для рентгеновских лучей часть красочного слоя оказывается прозрачной, и на рентгенограммах могут быть видны фактура холста, скрытые красочные слои и карандашные наброски, элементы рамы и другие объекты. Указанные обстоятельства осложняют поиск контрольных точек и совмещение.

В данной работе рассматривается задача построения автоматизированной процедуры совмещения фотографий и рентгеновских изображений произведений живописи (см. Рис.1.). Совмещаемые цифровые изображения в формате JPEG имеют размеры около 2800x4200 точек и глубину 8 или 24 бита на пиксел. Размеры изображений и поля зрения различны и обусловлены параметрами рентгеновской установки и областью интереса художников-реставраторов. Фотосъемка картин в видимом диапазоне спектра производилась CCD камерой. К особенностям изображений, обуславливающих трудности совмещения, относятся различия в полях зрения, ориентации изображений, содержании изображений: на Рис. 2 показан один и тот же фрагмент на цифровой фотографии и рентгенограмме.

Формально решаемая задача соответствует традиционной задаче привязки изображений, однако ее решение усложняется указанными особенностями. Предполагается, что имеются модель  $u(x, y): R^2 \rightarrow R^1$  и изображение  $v(x', y'): R^2 \rightarrow R^1$  (изображения, полученные в видимом и рентгеновском диапазонах). Требуется найти преобразование  $T: R^2 \rightarrow R^2$ , переводящее точки изображения  $v(x', y')$  в точки модели:

$$X = F(X'), \quad (1)$$

где  $X = (x, y)^T \in R^2$  и  $X' = (x', y')^T \in R^2$  - векторы координат изображения и модели, и минимизирующее среднеквадратичную ошибку совмещения.



(a)



(б)

Рис. 1. Изображения, полученные в видимом (а) и рентгеновском (б) спектральных диапазонах

Создаваемая процедура совмещения включает этапы: (а) предварительная обработка совмещаемых изображений, на котором производится фильтрация изображений и выделение областей интереса для поиска контрольных точек; (б) поиск контрольных точек на совмещаемых изображениях, необходимых для построения преобразования изображений; (в) сопоставление найденных контрольных точек с целью

установления попарного соответствия; (г) вычисление коэффициентов преобразования по координатам контрольных точек и совмещение изображений.

В последующих разделах будут предложены решения перечисленных задач.

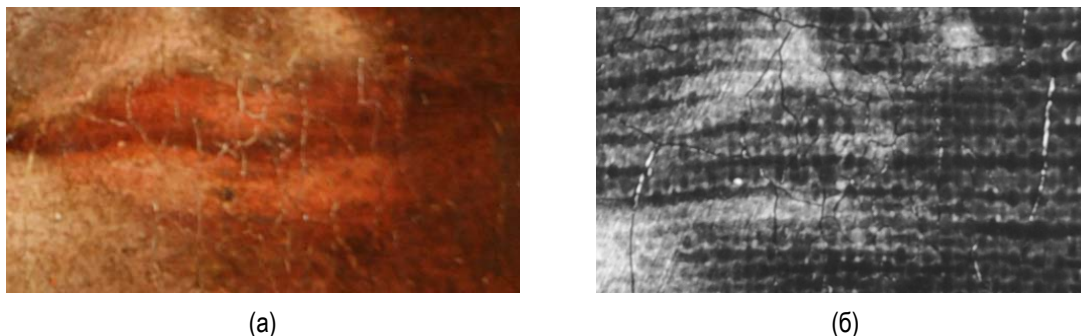


Рис. 2. Фрагменты изображений Рис.1: (а) – в видимом и (б) – в рентгеновском спектральных диапазонах

---

### Известные подходы к решению

---

Задача совмещения многомодальных изображений произведений живописи во многом схожа с задачами совмещения многомодальных изображений в медицине, аэрофотосъемке и других прикладных областях. Традиционно задача совмещения изображений, полученных в различных спектральных диапазонах, включает четыре этапа. (а) выделение характерных признаков (в частности, контрольных точек) на совмещаемых изображениях; (б) установление соответствия между найденными признаками с использованием выбранной меры соответствия; (в) построение модели преобразований; (г) преобразование изображений [Zitova, 2003].

Признаки, выделяемые на первом шаге, разделяются по типу объектов, выделяемых на совмещаемых изображениях: региональные, линейные, точечные. Признаки связаны с хорошо различимыми объектами на изображениях. Они должны быть инвариантными относительно выбранной модели преобразований. Если изображение не содержит высококонтрастных деталей, могут быть использованы признаки, не связанные непосредственно с объектами, а являющиеся информационными характеристиками изображений. В простейшем случае контрольные точки отмечаются вручную в интерактивном режиме. В работах [Schmid, 1997], [Delponte, 2006] и ряде других контрольные точки находятся с помощью детектора Харриса [Harris, 1988]. Дополнительно могут быть использованы такие признаки, как дифференциально-геометрические инварианты, моментные инварианты и др. Так в [Delponte, 2006] используются признаки, инвариантные к масштабированию (SIFT-признаки [Lowe, 1999]) В [Kammerer, 2004] при совмещении изображений, зафиксированных в инфракрасном и видимом диапазонах, используется модель аффинных преобразований, а контрольные точки выбираются вручную и уточняются процедурой на основе кросс-корреляции. В [Carpellini, 2005] алгоритм автоматического совмещения фотографий картин в видимом и ультрафиолетовом диапазонах основан на региональных признаках и реализует метод максимизации меры статистической зависимости пары изображений (maximization of the mutual information). Для поиска максимума разработана эвристическая итерационная поисковая процедура. Поиск осуществляется в пространстве четырех параметров: смещение по двум направлениям, коэффициент масштаба (в пределах 1%), угол поворота.

Следующей задачей, которую надо решить при совмещении изображений, является поиск соответствия между контрольными точками на изображениях. При этом количество найденных точек на изображениях может различаться, и часть из них может не иметь прообраза на другом изображении.

В работе [Schmid, 1997] предложены описания контрольных точек, найденных с помощью модифицированного детектора Харриса [Harris, 1988], в виде наборов значений дифференциальных инвариантов. Кроме того, в качестве дополнительных геометрических признаков используются значения углов между направлениями на соседние  $p$  точек. Применяемые описания обладают инвариантностью относительно вращения, изменения масштаба, изменения точки наблюдения и частичной окклюзии объектов.

Одним из распространенных алгоритмов сопоставления является алгоритм итерационного совмещения ближайших точек пары изображений (Iteration closest point registration - ICP) [Chen, 1992]. ICP алгоритм предназначен для нахождения жесткого преобразования  $T$ , которое позволяет наилучшим образом совместить облако точек на изображении сцены с ее геометрической моделью. На каждой итерации алгоритм находит преобразование из условия минимума среднеквадратичной ошибки между точками преобразованного изображения и соответствующими точками модели. Алгоритм обладает монотонной сходимостью к локальному минимуму. Ограничениями алгоритма являются (a) необходимость достаточно хорошего начального приближения преобразования  $T$ ; (b) каждой точке сцены должна соответствовать точка в модели. В работах [Sharp, 2002], [Rusinkiewicz, 2001] предложены модификации ICP алгоритма. Для улучшения работы алгоритма и сокращения числа итераций используется описание точек сцены вектором признаков, включающим пространственные координаты и значения инвариантных признаков [Sharp, 2002]. В [Rusinkiewicz, 2001] сравниваются различные варианты ICP алгоритма и исследуются комбинации алгоритмов для повышения быстродействия.

Разработан ряд методов сопоставления на основе анализа спектральных характеристик матрицы взвешенных попарных расстояний между контрольными точками. В работе [Scott, 1991] предложен алгоритм, устанавливающий соответствие между признаками двух изображений, одно из которых является результатом преобразования другого, на основе свойств сингулярного разложения матрицы взвешенных попарных расстояний между векторами признакового описания. Матрица соответствия признаков (в частности, координат характерных точек изображений) определяется из условия максимума следа произведения взвешенной матрицы попарных расстояний и самой матрицы соответствия. Показано, что условием максимума следа является ортогональность векторов, составленных из строк матрицы соответствия. Допустимые преобразования – смещение, сдвиг, изменение масштаба. Предложен способ формирования матрицы соответствия. Данный алгоритм не требует равного количества опорных точек на совмещаемых изображениях и прост с точки зрения реализации. В работах [Pilu, 1997] и [Zhao, 2004] предложены модификации данного метода, использующие локальные признаки. В работе [Delpronte, 2006] предложена модификация метода [Scott, 1991], позволяющая расширить диапазон его применения за счет уменьшения чувствительности к изменению масштаба и точки наблюдения. Эффективность метода повышается за счет применения вместо координат контрольных точек, используемых в прототипе, SIFT-признаков (Scale Invariant Feature Transform), вычисленных в точках, найденных детектором Харриса. Проведено сравнение результатов экспериментов, полученных для различных весовых функций, применяемых для построения матрицы попарных расстояний.

В [Shapiro, 1992] сопоставление двух множеств точек на плоскости производилось на основе сравнения собственных векторов матриц взвешенных попарных расстояний между точками каждого из сопоставляемых множеств. В работе [Carcassoni, 2003] для анализа спектральных характеристик матриц попарных расстояний сопоставляемых множеств предложено использовать двухшаговый EM алгоритм.

В качестве моделей преобразования обычно используются аффинные [Kamnerer, 2004], [Cappellini, 2005] или проективные преобразования [Hartley, 2004], которые достаточно хорошо описывают искажения,

возникающие при получении изображений. В ряде задач авторами рассматриваются нежесткие преобразования.

Проведенный анализ публикаций показал, что: (а) задача автоматизированного совмещения рентгеновских изображений с другими типами изображений произведений живописи в литературе освещена недостаточно полно; (б) особенности рентгеновских изображений затрудняют использование признаков, обычно используемых при решении задач совмещения; (в) достаточно эффективные и несложные в реализации SVD-основанные методы сопоставления контрольных точек соответствуют особенностям решаемой задачи и могут быть использованы в разрабатываемой процедуре; (г) модель проективных преобразований адекватна решаемой задаче совмещения и позволяет компенсировать искажения, характерные для используемых изображений.

Ниже рассматривается общая схема построения процедуры и ее отдельные этапы.

---

### **Общая схема построения процедуры**

---

Нахождение контрольных точек при совмещении рентгенограммы и фотографии в видимом свете осложняется тем, что содержание изображений может существенно отличаться. Найденные характерные точки в одном из изображений могут не быть найдены в другом. Кроме того, на изображениях произведений живописи не всегда удается выделить геометрические примитивы и признаки, которые можно использовать при совмещении.

На рентгеновских изображениях картин хорошо видны объекты, написанные красками, содержащими тяжелые металлы (в частности, свинец). На изображении таким участки, непрозрачные для рентгеновского излучения, выглядят светлыми. Обычно светлым участкам на рентгенограммах соответствуют светлые участки на картине. Именно это свойство будет использоваться для локализации контрольных точек при совмещении. В качестве контрольных точек предлагается использовать точки локальных экстремумов яркости на сглаженных изображениях. Максимумы яркости будут соответствовать наиболее светлым участкам картины, где в красочном слое присутствуют свинцовые белила. Такие контрольные точки обладают свойствами инвариантности относительно смещения, вращения, изменения масштаба, а также вариаций яркости. Необходимо будет выбрать величину сглаживания изображений, исходя из требуемой точности совмещения.

В качестве алгоритма сопоставления для разрабатываемой процедуры подходит алгоритм, основанный на сингулярном разложении матрицы попарных расстояний между контрольными точками двух изображений. Алгоритм не требует одинакового количества элементов сопоставляемых множеств контрольных точек. Для исключения ложных соответствий, которые возникают на реальных изображениях, алгоритм необходимо дополнить процедурой проверки найденных соответствий. По множеству найденных пар точек будет вычисляться матрица преобразования, вычисляться функционал ошибки совмещения и оцениваться вклад каждой пары в функционал ошибки совмещения.

Задача поиска оптимального преобразования изображений решается традиционными методами. Наиболее общей моделью преобразований, описывающей искажения при съемке фотокамерой, является проективное преобразование [Hartley, 2004].

В следующих разделах более подробно рассмотрены этапы разрабатываемой процедуры.

---

### **Предварительная обработка**

---

При поиске точек локальных максимумов яркости как на рентгеновских снимках, так и на фотографиях картин, фактура холста вносит искажения. Для ослабления периодических составляющих применяется

операция фильтрации в частотной области  $v_f = \Phi^{-1}(\Phi(v) \cdot I_m)$ , где  $v$  и  $v_f$  - исходное и отфильтрованное изображения рентгеновского снимка,  $\Phi$  и  $\Phi^{-1}$  обозначают операции прямого и обратного преобразования Фурье,  $I_m$  - изображение маски фильтра,  $(\cdot)$  - операция поэлементного умножения.

Для упрощения алгоритмов, используемых в процедуре совмещения, цветные изображения преобразуются к полутоновым.

С целью снижения вычислительных затрат используются пирамидальные представления изображений. Основные последующие этапы процедуры выполняются для уровня гауссовой пирамиды, соответствующего уменьшенным в 8 раз изображениям, а на заключительном этапе происходит обратное масштабирование координат контрольных точек и возврат к полноразмерным изображениям.

Поскольку контрольные точки связаны с областями, где в красочном слое присутствуют свинцовые белила, непрозрачные для рентгеновского излучения и образующие светлые области на картине в видимом спектральном диапазоне, то целесообразно предварительно выделить такие области на совмещаемых изображениях. Для выделения областей интереса применяется операция пороговой бинаризации с автоматическим определением порога [Otsu, 1979], [Kittler, 1986], [Niblack, 1986] и др. Полученные бинарные маски позволяют исключить локальные экстремумы, не связанные с областями интереса. Бинарные маски для изображений Рис. 1, полученные методом [Niblack, 1986] и обработанные с помощью операций морфологического размыкания, замыкания и удаления периферийных объектов, показаны на Рис. 3.

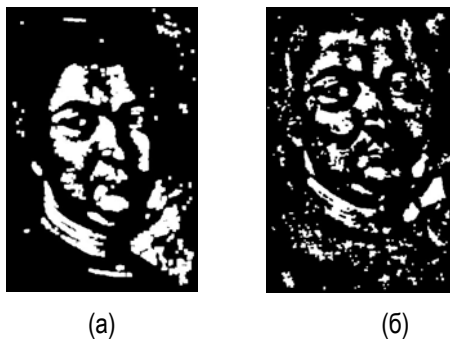


Рис. 3. (а), (б) - бинарные маски изображений, показанных на Рис. 1

### Поиск контрольных точек и вычисление матрицы преобразований

В качестве кандидатов в контрольные точки, необходимые для совмещения изображений предлагается использовать локальные максимумы яркости на рентгенограмме и полутоновом изображении в видимой части спектра, соответствующие светлым связным компонентам, присутствующим на обоих изображениях. Для исключения влияния оставшегося после фильтрации шума, порождаемого фактурой холста, изображения сглаживаются сверткой изображения с гауссовым ядром

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2},$$

где  $x, y$  - пространственные координаты,  $\sigma$  - параметр.

Параметр  $\sigma$  гауссова ядра выбирается таким образом, чтобы с одной связной светлой областью было связано небольшое количество локальных максимумов. Нахождение локальных максимумов

осуществляется алгоритмом, предложенным в [Kuijper, 2002]. Найденные точки максимумов показаны на Рис. 4(а, б).

При размывании объектов изображений экстремальные точки дрейфуют. Траектории максимумов яркости для полутонового изображения на Рис. 1(а) и рентгенограммы Рис. 1(б) показаны на Рис. 4 (в, г). Поэтому для уменьшения ошибки совмещения, обусловленной смещением экстремальных точек, нужно выбрать оптимальные значения параметра  $\sigma$ . Значения  $\sigma$  выбираются, исходя из обеспечения наилучшей точности совмещения контрольных точек:

$$I(\sigma) = \sum_{j=1}^p d_j^2(\sigma) \rightarrow \min_{\sigma},$$

$$d_j^2 = (x_j - x_j^H)^2 + (y_j - y_j^H)^2;$$

или

$$\max(|d_j|) \rightarrow \min, 1 \leq j \leq p,$$

где  $x_j, y_j$  - координаты  $j$ -ой контрольной точки в первом изображении;  $x_j^H, y_j^H$  - координаты преобразованной контрольной точки из второго изображения к координатам первого;  $d_j$  - евклидово расстояние между точками  $j$ -ой пары соответствующих контрольных точек;  $p$  - число пар контрольных точек на совмещаемых изображениях.

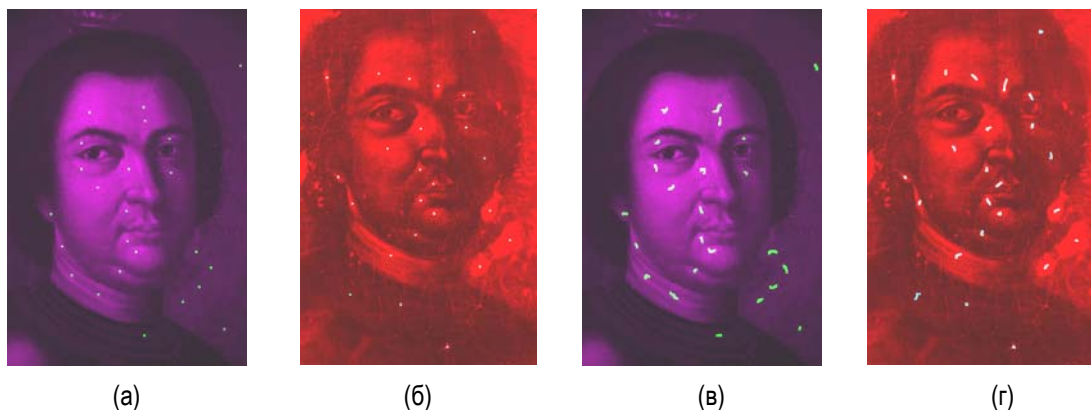


Рис. 4. (а), (б) – локальные максимумы яркости, из которых отбираются контрольные точки; (в), (г) – траектории максимумов яркости при изменении параметра гауссова ядра  $\sigma$  в интервале 2 - 10

Для сопоставления найденных характерных точек совмещаемых изображений применяется метод [Scott, 1991]. Как упоминалось ранее, метод основан на сингулярном разложении взвешенной матрицы попарных взвешенных расстояний между контрольными точками двух изображений:

$$W(k, l) = \exp(-d^2(x_k, x_l)/2r^2),$$

где  $d(x_k, x_l)$  - расстояние между точкой  $k$  изображения  $u$  и точкой  $l$  изображения  $v$ ;  $r$  - параметр, характеризующий допустимое расстояние между точками. В данной работе применяется следующая весовая функция, позволяющая получить лучшие результаты сопоставления [Delpon, 2006]:

$$W(k, l) = \exp(-|d(x_k, x_l)|/|r|),$$

значение параметра выбрано при настройке процедуры  $r = 25$ .

Результатом является матрица соответствия, в которой максимальный элемент в строке  $i$  и столбце  $j$  указывает на соответствие элемента  $i$  одного изображения элементу  $j$  другого. Метод достаточно простой в реализации, но при работе на реальных изображениях появляются ошибки. При решении практических задач для уменьшения количества ошибок при сопоставлении наборов контрольных точек совмещаемых изображений обычно используются различные локальные признаки, вычисляемые в окрестности контрольных точек. Различия рассматриваемых изображений не позволяют воспользоваться дополнительными локальными признаками. С целью исключения ошибочно выбранных пар точек предложена следующая итерационная процедура, которая позволяет исключить ложные соответствия. Соответствие считается ложным, если ошибка совмещения этой пары точек вносит максимальный вклад в квадратичный функционал качества. Пусть качество совмещения контрольных точек изображений на итерации  $i$  выражается функционалом

$$I_i = \sum_{j=0}^{p-i} d_{ij}^2,$$

$$d_{ij}^2 = (x_{ij} - x_{ij}^H)^2 + (y_{ij} - y_{ij}^H)^2,$$

где  $i$  – номер итерации,  $j$  – номер точки,  $p$  – начальное количество контрольных точек,  $x_j^H, y_j^H$  – координаты контрольной точки, преобразованной с помощью найденной на итерации  $i$  матрицы  $H$ . Пара точек с номером  $k$  отбрасывается, если выполняется условие

$$\Delta I_{i+1} = \max(I_i - I_{i+1}),$$

где

$$I_{i+1} = \sum_{j=0}^{p-i} d_{ij}^2, \quad j \neq k,$$

$\Delta I_{i+1}$  – вклад пары точек с номером  $k$  в функционал  $I$ .

Процесс заканчивается при  $\max_j (|d_{ij}|) < d_{\max}$ , где  $d_{\max}$  – допустимая абсолютная ошибка совмещения или  $I_i < I_{\max}$ ,  $I_{\max}$  определяется допустимой среднеквадратичной ошибкой.

Точность совмещения оценивается по среднеквадратичной ошибке в контрольных точках, максимальным абсолютным значением ошибки в контрольных точках и визуально. Графики зависимостей среднеквадратичной и абсолютной ошибок (в пикселах) от параметра гауссова ядра  $\sigma$  и количества используемых контрольных точек представлены на Рис. 5. Минимальные ошибки получены при  $\sigma=4$ , однако визуальный контроль показал, что наилучший результат соответствует  $\sigma=6$  и  $p=6$ . Данный эффект объясняется влиянием ярких мелких деталей на положение локальных максимумов яркости при относительно небольшом сглаживании изображений.

Как было отмечено в предыдущих разделах, подходящей моделью преобразований в решаемой задаче является проективное преобразование. Задача нахождения преобразования изображений формулируется следующим образом. Предполагается, что имеются модель  $u(x, y)$  и изображение  $v(x', y')$  (в данном случае это изображения, полученные в видимом и рентгеновском диапазонах). Требуется найти преобразование  $H$ , переводящее точки изображения  $v(x', y')$  в точки модели  $u(x, y)$ :

$$\tilde{X} = H\tilde{X}', \quad (1)$$



где  $\tilde{X} = (x, y, 1)^T$  и  $\tilde{X}' = (x', y', 1)^T$  - однородные координаты модели и изображения,  $H$  - однородная  $3 \times 3$  матрица. При этом требуется минимизировать среднеквадратичную ошибку совмещения. Для вычисления матрицы преобразования (1)  $H$  необходимо решить систему  $2n$  линейных алгебраических уравнений, где  $n$  - число пар соответствий контрольных точек. Требуется иметь как минимум четыре пары соответствий [Hartley, 2004]. При  $n > 4$  система решается методом наименьших квадратов, реализованным алгоритмом Левенберга-Марквардта [Madsen, 2004]. Этот метод обладает хорошей сходимостью и широко применяется в подобных задачах. Полученное преобразование обеспечивает точность совмещения в пределах одного пиксела, что соответствует аналогичным известным процедурам.

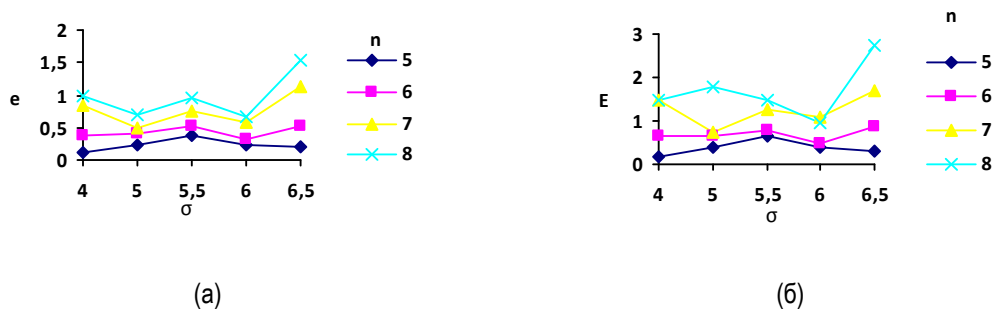


Рис. 5. Зависимости (а) среднеквадратичной  $e$  и (б) абсолютной  $E$  ошибок совмещения контрольных точек от параметра гауссова ядра  $\sigma$  и количества используемых при совмещении контрольных точек  $n$

На заключительном этапе процедуры производится масштабирование координат контрольных точек и преобразование полноразмерных изображений.

Результат совмещения цифровой фотографии портрета и его рентгенограммы (см. Рис. 1) для 6 контрольных точек, полученных при  $\sigma=6$ , показан на Рис. 6.



Рис. 6. Совмещенные изображения, показанные на Рис.1

## Заключение

Разработана процедура автоматизированного совмещения фотографий и рентгенограмм произведений живописи. В качестве контрольных точек используются локальные экстремумы яркости, найденные на

размытых гауссовым ядром изображениях. Выбран параметр гауссова ядра. Полученные контрольные точки связаны с деталями красочного слоя, видимыми на обоих совмещаемых изображениях. Для сопоставления найденных характерных точек изображений используется алгоритм [Scott, 1991] на основе сингулярного разложения матрицы взвешенных попарных расстояний между контрольными точками двух изображений, дополненный итерационной процедурой выявления ложных соответствий. Модификация алгоритма позволила получить приемлемую точность сопоставления на реальных изображениях без использования дополнительных признаков. Для совмещения использована модель проективных преобразований, соответствующая процессу получения изображений. Разработанная процедура позволяет обеспечить точность совмещения изображений в пределах одного пиксела в контрольных точках. Процедура реализована в виде программного модуля и используется при решении практических задач.

В дальнейшем предполагается исследовать возможность использования в процедуре других методов локализации инвариантных контрольных точек и применить комбинированные методы сопоставления для повышения точности.

---

### Благодарности

---

Работа выполнена при частичной поддержке РФФИ, проект № 09-07-00368.

---

### Библиография

---

- [Cappellini, 2005] Cappellini V., et.al. An automatic registration algorithm for cultural heritage images. In Proc.of ICIP2005, Genoa, Italy. V.2, P. 566–569, 2005.
- [Carcassoni, 2003] M. Carcassoni, E.R. Hancock, Correspondence Matching with Modal Clusters, IEEE Transactions on Pattern Analysis and Machine Intelligence, V. 25 N 12, P.1609-1615, December 2003
- [Chen, 1992] Y. Chen and G. Medioni, "Object Modelling by Registration of Multiple Range Images", International Journal of Computer Vision and Image Understanding, V. 10, N. 3, P. 145-155, April 1992.
- [Delponete, 2006] E. Delponete, F. Isgrò, F. Odone, A. Verri. SVD-matching using SIFT features, Graphical Models, V.68 N.5, P.415-431, 2006.
- [Harris, 1988] C. Harris and M. Stephens. A combined corner and edge detector, in Alvey Vision Conf, 1988, P. 147-151.
- [Hartley, 2004] R. Hartley, A. Zisserman. Multiple View Geometry in Computer Vision, Cambridge University Press, 2004.
- [Heitz, 1990] F. Heitz, H. Maitre, C. de Couessin. Event Detection in Multisource Imaging: Application to Fine Arts Painting Analysis. IEEE transactions on acoustics, speech. and signal processing. V. 38, N. 1. April 1990. P. 695-704.
- [Kammerer, 2004] P. Kammerer, A. Hanbury, E. Zolda. A Visualization Tool for Comparing Paintings and Their Underdrawings. In Proc. of the Conference on Electronic Imaging & the Visual Arts (EVA 2004), - 2004. - P. 148–153.
- [Kirsh, 2000] A. Kirsh and R. S. Levenson, Seeing through paintings: Physical examination in art historical studies, Yale U. Press, New Haven, CT, 2000.
- [Kittler, 1986] J. Kittler and J. Illingworth. Minimum error thresholding. Pattern Recogn. Vol.19., 1986. P. 41–47.
- [Kuijper, 2002] A. Kuijper, The Deep Structure of Gaussian Scale Space Images. Ph.D. Thesis, Utrecht University. ISBN 90-393-3061-1, 2002.
- [Lowe, 1999] D.G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of ICCV, 1999, P. 1150–1157.
- [Madsen, 2004] K. Madsen, H.B. Nielsen, O. Tingleff. Methods for Non-Linear Least Squares Problems. Technical University of Denmark, 2004.
- [Maintz, 1998] J.B.A. Maintz, M.A. Viergever. An Overview of Medical Image Registration Methods. - URN:NBN:NL:UI:10-1874-18921, Utrecht University, 1998.

- [Maitre , 2001] H. Maitre, F. Schmitt et C. Lahanier, 15 years of image processing and the fine arts, IEEE-ICIP'2001, Salonique (Greece), Vol. 1, 2001, P. 557-561.
- [Martinez, 2002] K. Martinez, J. Cupitt, D. Saunders, R. Pillay. Ten Years of Art Imaging Research. Proceedings of the IEEE, V. 90, N. 1, 2002, P. 28-41.
- [Niblack, 1986] W. Niblack. An Introduction to Digital Image Processing. Prentice Hall, Englewood Cliffs, NJ, 1986
- [Otsu, 1979] N. Otsu. A threshold selection method from gray level histograms// IEEE Trans. Syst. Man Cybern. SMC-9. 1979, P. 62–66.
- [Pilu, 1997] P. P. P. A direct method for stereo correspondence based on singular value decomposition. IEEE Proceedings of CVPR. 1997, P. 261-266.
- [Rusinkiewicz, 2001] S. Rusinkiewicz, M. Levoy, "Efficient Variants of the ICP Algorithm," 3dim, Third International Conference on 3-D Digital Imaging and Modeling (3DIM '01), 2001, P.145
- [Schmid, 1997] C. Schmid, R. Mohr, Local Greyvalue Invariants for Image Retrieval, PAMI V.19, N. 5, P. 872-877, 1997.
- [Scott, 1991] G. Scott, and H. C. Longuet-Higgins, "An Algorithm for Associating the Features of Two Images," in Proceedings of the Royal Society London, 1991, V. B244, P. 21-26.
- [Shapiro, 1992] L.S. Shapiro, J. M. Brady. Feature-based correspondence: an eigenvector approach. Image and Vision Computing, V. 10, Issue 5, P. 283 – 288, 1992.
- [Sharp, 2002] G. Sharp, S. Lee, and D. Wehe. ICP registration using invariant features. IEEE. T. Pattern Anal., V. 24 N. 1, P. 90–102, 2002.
- [Stork, 2009] D.G. Stork. Computer Vision and Computer Graphics Analysis of Paintings and Drawings: An Introduction to the Literature, LNCS 5702, Springer-Verlag Berlin Heidelberg. 2009, P. 9–24.
- [Zhao, 2004] F. Zhao. Image matching based on singular value decomposition. PCM 2004, LNCS, V. 3333, 2004, pp. 119–126.
- [Zitova, 2003] B. Zitova, J. Flusser. Image registration methods: a survey. Image and Vision Computing, V. 21, N. 11, P.977–1000, 2003.

---

### Authors' Information

---



**Dmitry Murashov** – *Computing Centre of the Russian Academy of Sciences, Senior researcher, 40, Vavilov str., Moscow, 119333, Russian Federation; e-mail: d\_murashov@mail.ru.*

*Major Fields of Scientific Research: Image processing, Image analysis, Pattern recognition.*

## РАСПОЗНАВАНИЕ ОБЕКТОВ С НЕПОЛНОЙ ИНФОРМАЦИЕЙ И ИСКАЖЕННЫХ ПРЕОБРАЗОВАНИЯМИ ИЗ ЗАДАННОЙ ГРУППЫ В РАМКАХ ЛОГИКО- ПРЕДМЕТНОЙ РАСПОЗНАЮЩЕЙ СИСТЕМЫ

Татьяна Косовская

**Abstract:** В рамках логико-аксиоматической распознающей рассмотрены задача распознавания объектов с неполной информацией, в частности, частично заслоненных объектов, и задача распознавания объектов, подвергнутых искажениям из заданной группы преобразований при условии, что классы объектов замкнуты относительно этой группы. Приведены алгоритмы решения этих задач. Доказаны оценки числа шагов этих алгоритмов при различных способах решения стандартной задачи распознавания.

**Keywords:** распознавание образов, неполная информация, инвариантность к группе преобразований, сложность алгоритмов.

**ACM Classification Keywords:** I.2.4 Knowledge Representation Formalisms and Methods – Predicate logic, I.5.1 PATTERN RECOGNITION Models – Deterministic, F.2.2 Nonnumerical Algorithms and Problems – Complexity of proof procedures.

---

### Введение

---

Под логико-предметной распознающей системой понимается следующая [1]. Пусть имеется множество  $\Omega$  конечных множеств  $\omega = \{ \omega_1, \dots, \omega_n \}$ , которые в дальнейшем будут называться распознаваемыми объектами. Частью  $\tau$  объекта  $\omega$  называется любое его подмножество. Пусть также на частях  $\tau$  задан набор предикатов  $p_1, \dots, p_n$ , характеризующих свойства и отношения между элементами распознаваемого объекта  $\omega$ . Пусть задано разбиение множества  $\Omega$  на  $K$  классов  $\Omega = \cup_{k=1}^K \Omega_k$ .

**Логическим описанием**  $S(\omega)$  **расознаваемого объекта**  $\omega$  называется набор всех истинных постоянных формул вида  $p_i(\tau)$  или  $\neg p_i(\tau)$ , выписанных для всех возможных частей  $\tau$  объекта  $\omega$ .

**Описанием класса** называется множество условий, задающих необходимые и достаточные условия принадлежности этому классу.

Здесь и далее через  $x$  будем обозначать список элементов конечного множества  $x$ , соответствующий некоторой перестановке номеров его элементов. То, что элементами списка  $x$  являются элементы множества  $y$ , будем записывать в виде  $x \subseteq y$ .

Для того, чтобы записать, что значения для переменных списка  $x$ , удовлетворяющие формуле  $A(x)$ , различны, будет использоваться обозначение  $\exists x_{\neq} A(x)$ .

**Логическим описанием класса**  $\Omega_k$  называется такая формула  $A_k(x)$ , что  $A_k(x)$  содержит в качестве атомарных только формулы вида  $p_i(y)$  (при  $y \subseteq x$ ;  $A_k(x)$ ), не содержит кванторов и если истинна формула  $A_k(\omega)$ , то  $\omega \in \Omega_k$ .

Отметим, что логическое описание класса всегда может быть записано в виде дизъюнкции элементарных конъюнкций атомарных формул. С помощью построенных описаний предлагается решать следующие задачи распознавания образов.

**Задача идентификации.** Проверить, принадлежит ли  $\omega$  или его часть классу  $\Omega_k$ .

**Задача классификации.** Найти все такие номера классов  $k$ , что  $\omega \in \Omega_k$ .

**Задача анализа сложного объекта.** Найти и классифицировать все части  $\tau$  объекта  $\omega$ , для которых  $\tau \in \Omega$ .

Решение задач идентификации, классификации и анализа сложного объекта в [1] сведено к доказательству соответственно формул

$$S(\omega) \Rightarrow \exists \mathbf{x} \neq A_k(\mathbf{x}), \quad (1)$$

$$S(\omega) \Rightarrow \bigvee_{k=1}^K A_k(\omega), \quad (2)$$

$$S(\omega) \Rightarrow \bigvee_{k=1}^K \exists \mathbf{x} \neq A_k(\mathbf{x}). \quad (3)$$

### Распознавание объекта в условиях неполной информации

При распознавании объекта в условиях неполной информации задано не полное описание объекта  $S(\omega)$ , содержащее все истинные на  $\omega$  атомарные формулы или их отрицания, а лишь некоторое его подмножество  $S^-(\omega) \subseteq S(\omega)$ .

Так как описание класса является дизъюнкцией элементарных конъюнкций, то введем обозначение  $A_k(\mathbf{x}) = \bigvee_{j=1}^{J_k} A_k^j(\mathbf{y}_k^j)$ , где для каждого  $j$  ( $1 \leq j \leq J_k$ )  $\mathbf{y}_k^j$  является подстрокой списка переменных  $\mathbf{x}$ . При решении задач идентификации, классификации и анализа сложного объекта с неполным описанием объекта вместо проверки справедливости следствий (1), (2) и (3) соответственно имеется возможность проверки лишь того, что  $S^-(\omega) \Rightarrow \exists \mathbf{x} \neq A_k(\mathbf{x})$ ,  $S^-(\omega) \Rightarrow \bigvee_{k=1}^K A_k(\omega)$ ,  $S^-(\omega) \Rightarrow \bigvee_{k=1}^K \exists \mathbf{x} \neq A_k(\mathbf{x})$ . Это равносильно тому, что хоть при одном значении  $j$  ( $1 \leq j \leq J_k$ ) (и хоть при одном значении для  $k$  ( $1 \leq k \leq K$ )) для задач классификации и анализа сложного объекта справедливо  $S^-(\omega) \Rightarrow \exists \mathbf{y}_k^j \neq A_k^j(\mathbf{y}_k^j)$ ,  $S^-(\omega) \Rightarrow A_k(\omega)$ ,  $S^-(\omega) \Rightarrow \exists \mathbf{y}_k^j \neq A_k^j(\mathbf{y}_k^j)$ .

Пусть  $A(\mathbf{x})$  – элементарная конъюнкция атомарных формул,  $A^-(\mathbf{x}^-)$  – некоторая ее подформула ( $\mathbf{x}^-$  – подсписок списка переменных  $\mathbf{x}$ ),  $a$  и  $a^-$  – количество атомарных формул в  $A(\mathbf{x})$  и в  $A^-(\mathbf{x}^-)$  соответственно,  $m$  и  $m^-$  – количество предметных переменных в  $A(\mathbf{x})$  и в  $A^-(\mathbf{x}^-)$  соответственно.

Числа  $q$  и  $r$  вычисляются по формулам  $q = a^-/a$ ,  $r = m^-/m$  и характеризуют степень совпадения формул  $A(\mathbf{x})$  и  $A^-(\mathbf{x}^-)$ . При этом  $0 < q \leq 1$ ,  $0 < r \leq 1$ . Кроме того,  $q = r = 1$  тогда и только тогда, когда  $A^-(\mathbf{x}^-)$  совпадает с  $A(\mathbf{x})$ .

При таких обозначениях формулу  $A^-(\mathbf{x}^-)$  будем называть **(q,r)-фрагментом формулы**  $A(\mathbf{x})$ .

**Замечание.** Возможен следующий вариант определения чисел  $q$  и  $r$ . Каждому предикату и каждой предметной переменной формулы  $A(\mathbf{x})$  можно приписать "вес", определяемый либо экспертами, либо из вероятностных соображений. Тогда  $q = w^-/w$ ,  $r = v^-/v$ , где  $w$  и  $w^-$  – сумма "весов" предикатных формул в  $A(\mathbf{x})$  и  $A^-(\mathbf{x}^-)$ ,  $v$  и  $v^-$  – сумма "весов" предметных переменных в  $A(\mathbf{x})$  и  $A^-(\mathbf{x}^-)$  соответственно.

Если следствие  $S^-(\omega) \Rightarrow \exists \mathbf{x} \neq A(\mathbf{x})$  не имеет места, но для некоторого (q,r)-фрагмента  $A^-(\mathbf{x}^-)$  (при  $q \neq 1$ ) имеет место следствие  $S^-(\omega) \Rightarrow \exists \mathbf{x}^- \neq A^-(\mathbf{x}^-)$ , то будем говорить, что  $S^-(\omega) \Rightarrow \exists \mathbf{x} \neq A(\mathbf{x})$  является **частично (q,r)-выводимой**.

Формула  $\{DA\}(\mathbf{x})$  называется **негативным дополнением** формулы  $A(\mathbf{x})$  до ее фрагмента  $A^-(\mathbf{x}^-)$ , если она является элементарной дизъюнкцией, состоящей из отрицаний конъюнктивных членов формулы  $A(\mathbf{x})$ , не вошедших во фрагмент  $A^-(\mathbf{x}^-)$ , то есть  $A^-(\mathbf{x}^-) \& \neg\{DA\}(\mathbf{x}) \Leftrightarrow A(\mathbf{x})$ . Ниже будет

использоваться обозначение  $\{DA\}(x)|(x^-, \tau)$  для результата замены переменных списка  $x^-$  на список констант  $\tau$ .

Если  $S(\omega) \Rightarrow \exists x A(x)$  частично  $(q,r)$ -выводима и ни для каких наборов различных констант  $\tau$ , для которых из истинности  $S(\omega)$  следует истинность  $A(\tau)$ , нет следствия  $S(\omega) \Rightarrow \exists x \{DA\}(x)|(x^-, \tau)$ , то  $S(\omega) \Rightarrow \exists x A(x)$  называется  $(q,r)$ -выводимой.

Если формула  $(q,r)$ -выводима при некоторых  $q$  и  $r$ , то будем говорить, что у нее имеется **неполный вывод**.

По сути дела, понятие  $(q,r)$ -выводимости для  $S(\omega) \Rightarrow \exists x A(x)$  означает, что имеется набор различных констант  $\tau = (\omega_{i,1}, \dots, \omega_{i,a})$ , количество которых составляет долю  $r$  от общего количества переменных формулы  $A(x)$ , для которого истинна формула  $A(\tau)$ , количество атомарных формул которой составляет долю  $r$  от общего количества атомарных формул формулы  $A(x)$ , а также нет информации о том, что формула  $A(x)$  не выполнима на  $\omega$ .

Алгоритм проверки  $(q,r)$ -выводимости для  $S(\omega) \Rightarrow \exists x A(x)$  подробно описан в [3] и состоит в последовательном выделении подформулы  $A(x^-)$  элементарной конъюнкции  $A(x)$  и нахождении таких списков значений  $\tau$  для списка переменных  $x^-$ , что  $S(\omega) \Rightarrow A(\tau)$ , но неверно  $S(\omega) \Rightarrow \exists x \{DA\}(x)|(x^-, \tau)$ . При этом нахождение списков значений  $\tau$  и проверка следствия  $S(\omega) \Rightarrow \exists x \{DA\}(x)|(x^-, \tau)$  может осуществляться как полным перебором всех различных значений наборов с  $a^-$  различными значениями из  $\omega$ , так и построением вывода в исчислении предикатов.

Пусть  $t$  – число элементов в множестве  $\omega$ ,  $m$  – число аргументов в  $A(x)$ ,  $a$  – максимальное число атомарных формул (вхождений признаков) в  $A(x)$ ,  $|A|$  – число вхождений предметных переменных в  $A(x)$ ,  $|S|$  – число вхождений предметных констант в  $S(\omega)$ .

**Теорема 1.** Число шагов решения задачи идентификации объекта с неполным описанием при использовании переборного алгоритма проверки неполной выводимости составляет  $O(t^m 2^a |S| |A|)$ .

При этом для выделенных частей распознаваемого объекта  $\omega$ , будут вычислены значения параметров  $q$  и  $r$ , определяющие степень уверенности  $q$  того, что эта часть объекта составляет  $r$ -ую долю объекта заданного класса.

Пусть  $s$  – максимальное число атомарных формул в  $S(\omega)$  с одним и тем же предикатом,  $a$  – число атомарных формул в  $A(x)$ ,  $J_k$  – число дизъюнктивных членов в описании класса.

**Теорема 2.** Число шагов решения задачи идентификации объекта с неполным описанием при использовании построения вывода в исчислении предикатов для проверки неполной выводимости составляет  $O(J_k s^a)$ .

При этом для выделенных частей распознаваемого объекта  $\omega$ , будут вычислены значения параметров  $q$  и  $r$ , определяющие степень уверенности  $q$  того, что эта часть объекта составляет  $r$ -ую долю объекта заданного класса.

Доказательства теорем основаны на оценках числа шагов работы алгоритмов распознавания объектов логико-аксиоматической распознающей системой [2].

---

### Пример распознавания частично заслоненного объекта

---

Пусть имеется множество контурных изображений, составленных из отрезков прямых, задаваемых своими концами. Заданы два предиката  $V$  и  $L$ , определяемые следующим образом:  $V(x,y,z) \Leftrightarrow \angle yxz < \pi$ ,  $L(x,y,z) \Leftrightarrow$  "x между y и z".

Заданы два класса контурных изображений, эталоны которых имеют вид, представленный на рис. 1.

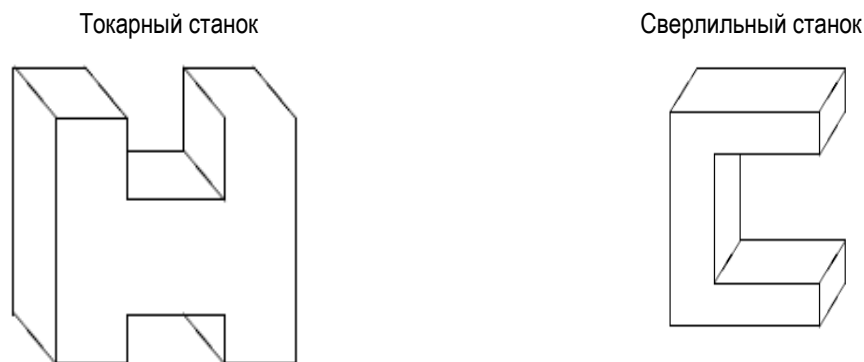


Рис. 1. Эталонные изображения объектов

Описания классов, составленные по этим эталонам, имеют следующие параметры:  $m_1 = 10$ ,  $m_2 = 22$ ,  $m_3 = 15$ ;  $a_1 = 22$ ,  $a_2 = 52$ ,  $a_3 = 35$ ;  $|A_1| = 228$ ,  $|A_2| = 537$ ,  $|A_3| = 357$ .

Для распознавания представлена сцена, изображенная на рис. 2, и поставлен вопрос: "Имеется ли на сцене сверлильный станок?".

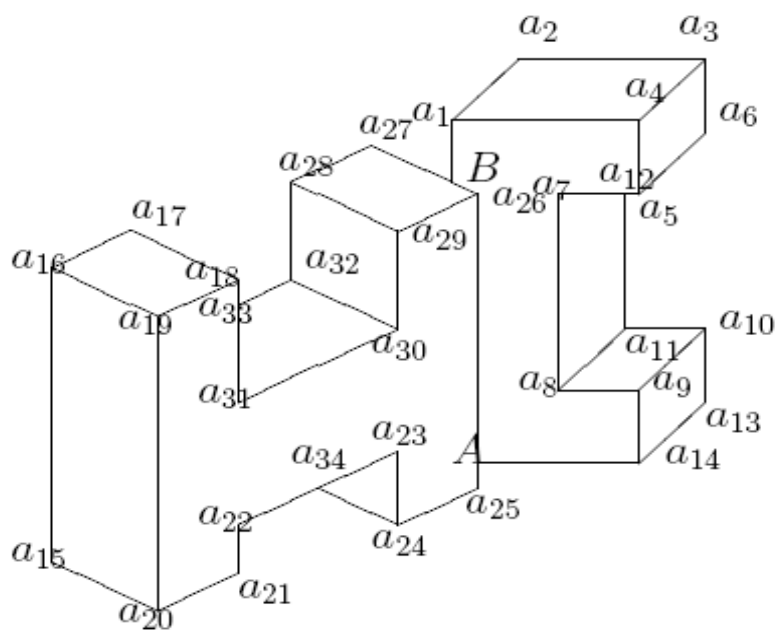


Рис. 2. Сцена с частично заслоненным объектом.

Распознаваемый объект имеет 32 элемента ( $t=32$ ). Описание сцены содержит 79 атомарных формул, каждая из которых имеет по 3 аргумента ( $s=76$ ,  $|S|=79 \cdot 3=237$ )

При попытке доказать выводимость  $S(\omega) \Rightarrow \exists x_{\neq} A_3(x)$  получим частичную выводимость этой секвенции, а именно выводимость  $S(\omega) \Rightarrow \exists x_{\neq} A_3^1(x)$ , причем в качестве значений для  $x$  выступают константы  $(a_1, \dots, a_{14}, A)$ , либо  $(a_1, \dots, a_{14}, B)$ . При этом в обоих случаях  $q = 24/26$ ,  $r = 1$ .

Негативное дополнение формулы  $A_3^{1-}(x)$  до  $A_3^1(x)$  на этих значениях имеет один из следующих видов  $\neg V(a_1, a_4, A) \vee \neg V(A, a_{14}, a_1)$  или  $\neg V(a_{14}, a_9, B) \vee \neg V(B, a_{14}, a_1)$ . В обоих случаях негативное дополнение не выводимо из  $S(\omega)$ . Следовательно, со степенью уверенности 12/13 можно утверждать, что часть сцены  $(a_1, \dots, a_{14}, A)$ , либо  $(a_1, \dots, a_{14}, B)$  представляет из себя часть сверлильного станка.

### Инвариантность системы к заданной группе преобразований

Пусть на множестве  $\Omega$  задана совокупность преобразований  $G$ , отображающих это множество на себя. Обозначим посредством  $G(\omega)$  множество термов вида  $g(\omega)$ , где  $g \in G, \omega \in \Omega$ .

Логико-предметная распознающая система называется **инвариантной относительно совокупности  $G$** , если она одинаково идентифицирует любые два объекта, отличающиеся только преобразованиями из совокупности  $G$ .

Класс объектов  $\Omega_k$  называется замкнутым относительно совокупности преобразований  $G$ , если любые два объекта, отличающиеся только преобразованиями из совокупности  $G$ , одновременно принадлежат (или не принадлежат) этому классу.

Наиболее простым случаем построения инвариантной логико-аксиоматической распознающей системы является система, построенная на основании инвариантного набора исходных признаков. Однако, зачастую признаки, адекватно описывающие классы объектов, не обладают этим свойством.

Далее будем рассматривать совокупность преобразований, являющуюся группой с конечным числом образующих. При этом множество образующих группы будем обозначать посредством  $G = \{g_1, \dots, g_T\}$ , а саму группу  $G^*$ . Образующие группы  $G^*$  будем называть элементарными преобразованиями.

Пусть для каждого элементарного преобразования  $g_j$  ( $j = 1, \dots, T$ ) можно указать, как изменяются значения отдельных признаков или их совокупностей при воздействии преобразования  $g_j$  на распознаваемый объект. Для каждого  $g_j$  таких изменений может быть несколько (обозначим количество таких изменений посредством  $l_j$ ). Эти изменения определяются эквивалентностями вида

$$V_l^j(x) \Leftrightarrow C_l^j(g_j(x)), \quad (4)$$

где  $V_l^j(x)$  и  $C_l^j(g_j(x))$  – элементарные конъюнкции атомарных формул,  $l=1, \dots, l_j$ . Равносильности вида (4) будем называть описаниями преобразования  $g_j$ . Множество описаний для всех преобразований будем обозначать посредством  $\Gamma(x)$ .

**Теорема 3.** Пусть на  $\Omega$  задана группа  $G^*$  с конечным числом образующих  $G = \{g_1, \dots, g_T\}$ , для каждого преобразования  $g_j$  которой справедливы  $l_j$  описаний преобразования вида (4).

Если для каждого  $j$  описание  $k$ -го класса  $A_k(x)$  вместе с каждым дизъюнктивным членом, в который входит  $V_l^{j1}(x) \& \dots \& V_l^{jl}(x)$  содержит дизъюнктивный член с элементарной конъюнкцией  $C_l^{j1}(x) \& \dots \& C_l^{jl}(x)$ , причем все остальные конъюнктивные члены этих дизъюнктов одинаковы и инвариантны относительно  $g_j$ , то  $A_k(x)$  инвариантно относительно группы преобразований  $G^*$ .

Описания преобразований позволяют расширить понятие логико-аксиоматической распознающей системы введением в неё равносильностей вида (4). При этом задачи инвариантного распознавания могут быть сведены к следующим задачам.

**Задача инвариантной идентификации:** Проверить, принадлежит ли объект  $\omega$  или его часть классу  $\Omega_k$ , если класс  $\Omega_k$  замкнут относительно группы преобразований  $G^*$  с конечным числом образующих  $G = \{g_1, \dots, g_T\}$ .

Эта задача сводится к доказательству формулы



$$S(\omega) \& \Gamma(x) \Rightarrow \exists x_{\neq} A_k(x)$$

**Задача инвариантной классификации.** Найти все такие номера классов  $k$ , что  $\omega \in \Omega_k$ , если класс  $\Omega_k$  замкнут относительно группы преобразований  $G^*$  с конечным числом образующих  $G = \{g_1, \dots, g_T\}$ .

Эта задача сводится к доказательству формулы

$$S(\omega) \& \Gamma(x) \Rightarrow \bigvee_{k=1}^K A_k(\omega)$$

с указанием всех таких номеров  $k$ , для которых соответствующий дизъюнктивный член истинен на  $\omega$ .

**Задача инвариантного анализа сложного объекта.** Найти и классифицировать все части  $\tau$  распознаваемого объекта  $\omega$ , для которых  $\tau \in \Omega$ , если класс  $\Omega_k$  замкнут относительно группы преобразований  $G^*$  с конечным числом образующих  $G = \{g_1, \dots, g_T\}$ .

Эта задача сводится к доказательству формулы

$$S(\omega) \& \Gamma(x) \Rightarrow \bigvee_{k=1}^K \exists x_{\neq} A_k(x)$$

с указанием всех частей объекта  $\omega$ , поддающихся классификации, и идентифицировать их.

Для произвольной группы  $G^*$  с конечным числом образующих эти задачи алгоритмически неразрешимы. Но если глубина вложенности терма, задающего преобразования из  $G^*$ , не превосходит заданного числа, то можно предложить следующий алгоритм решения задачи инвариантной идентификации.

Проверяем справедливость  $S(\omega) \Rightarrow \exists x_{\neq} A_k(x)$ . Если формула верна, то подмножества  $\omega$ , выполняющие формулу  $A_k(x)$ , принадлежат классу  $\Omega_k$ .

Создаем очередь из описаний  $S^j(\omega)$  объекта  $\omega$ , подвергнутого преобразованиям с номерами из списка  $J$ . Первоначально в очереди находится исходное описание  $S(\omega)$  (т. е.  $J$  пуст). Результат приписывания номера  $j$  к списку  $J$  обозначим посредством  $J||j$ .

Если длина списка  $J$  меньше заданного числа  $R$ , то для каждого  $j$  ( $1 \leq j \leq T$ ) и каждой элементарной конъюнкции  $C_l^j(x)$  ( $l=1, \dots, l_j$ ), входящей в равносильность вида (4), то из  $S^j(\omega)$ , находящегося первым в очереди, выделяем все его подмножества  $\tau$  объекта  $g_j(\omega)$ , для которых  $S^j(\omega)$  содержит все конъюнктивные члены этой элементарной конъюнкции; в  $S^j(\omega)$  заменяем все конъюнктивные члены  $C_l^j(\tau)$  на конъюнктивные члены  $B_l^j(\tau)$ , полученное описание обозначаем  $S^{j||l}(\omega)$ ; проверяем справедливость  $S^{j||l}(\omega) \Rightarrow \exists x_{\neq} A_k(x)$ . Если формула верна, то подмножества  $g_{j||l}(\omega)$ , выполняющие формулу  $A_k(x)$ , принадлежат классу  $\Omega_k$ ; в противном случае заносим  $S^{j||l}(\omega)$  в очередь.

Если  $j = T$ , то берем следующее описание из очереди описаний и повторяем предыдущий шаг.

Алгоритм закончит работу, если для некоторого списка  $J$  верно  $S^j(\omega) \Rightarrow \exists x_{\neq} A_k(x)$ . В этом случае найдено преобразование  $g_j$ , отличающее распознаваемый объект от эталонного, и те части распознаваемого объекта, которые принадлежат классу  $\Omega_k$ , или ни для одного списка  $J$  длины не более  $R$  не выполняется  $S^j(\omega) \Rightarrow \exists x_{\neq} A_k(x)$  (в этом случае распознаваемый объект не является объектом, отличающимся от содержащего части из класса  $\Omega_k$  преобразованием с глубиной вложенности терма, не превосходящей  $R$ , из группы  $G^*$  с конечным числом образующих).

**Теорема 4.** Если для доказательства формулы вида  $S(\omega) \Rightarrow \exists x_{\neq} A(x)$  использован алгоритм полного перебора, то число шагов инвариантной идентификации для класса, замкнутого относительно группы  $G^*$  с конечным числом образующих  $G = \{g_1, \dots, g_T\}$  при ограничении, что глубина вложенности термов, задающих преобразования из группы  $G^*$ , не превосходит заданного числа  $R$ , составляет

$$O(T^R R |S| (t^m |A| + t^c |C| L) + \Delta (t^m |A| + t^c |C| L)),$$

где  $t$  – число элементов в множестве  $\omega$ ,  $m$  – максимальное число аргументов в дизъюнктивных членах описания класса,  $|A|$  – число вхождений предметных переменных в описание класса,  $|S|$  – число вхождений предметных констант в  $S(\omega)$ ,  $s$  – максимальное число аргументов в формулах  $C_j(\mathbf{x})$ ,  $|C|$  – число вхождений предметных переменных в формулы  $C_j(\mathbf{x})$ ,  $\Delta$  – максимальная разность количества различных вхождений предметных переменных в  $C_j(\mathbf{x})$  и  $B_j(\mathbf{x})$ .

**Теорема 5.** Если для доказательства формулы вида  $S(\omega) \Rightarrow \exists \mathbf{x} \neq A(\mathbf{x})$  использован алгоритм поиска вывода в исчислении предикатов, то число шагов инвариантной идентификации для класса, замкнутого относительно группы  $G^*$  с конечным числом образующих  $G = \{g_1, \dots, g_T\}$  при ограничении, что глубина вложенности термов, задающих преобразования из группы  $G^*$ , не превосходит заданного числа  $R$ , составляет


$$O(T^R (J_k (s+R\delta)^a + (s+R\delta)^c)),$$

где  $J_k$  – число дизъюнктивных членов в описании класса,  $s$  – максимальное число атомарных формул в  $S(\omega)$  с одним и тем же предикатом,  $a$  – число вхождений атомарных формул в  $A(\mathbf{x})$ ,  $c$  – максимальное количество вхождений атомарных формул в элементарные конъюнкции  $C_j(\mathbf{x})$ ,  $s$  – максимальное количество вхождений одного и того же предиката в  $S(\omega)$ ,  $\delta$  – максимальное изменение количества атомарных формул с одним и тем же предикатом в множестве  $S^{li}(\omega)$  по сравнению с  $S^j(\omega)$ .

Доказательства теорем основаны на оценках числа шагов алгоритмов распознавания объектов логико-аксиоматической распознающей системой [2].

### Примеры инвариантных распознающих систем с неинвариантными признаками

1. Пусть имеется множество контурных изображений, составленных из отрезков прямых, задаваемых своими концами. Заданы два предиката  $V$  и  $L$ , определяемых следующим образом.



$$V(x, y, z) \cdot \quad \underline{y \quad x \quad z} \quad L(x, y, z, )$$

Оба эти предиката инвариантны относительно таких аффинных преобразований как  $g^l$  – сдвиг на  $l$ ,  $g_r^\varphi$  – поворот на угол  $\varphi$  и  $g_c^k$  – растяжение в  $k$  раз. Предикат  $L$  инвариантен также относительно зеркального отображения  $g_m$ . Предикат  $V$  не инвариантен относительно  $g_m$ . Предикат  $g_m$  имеет описание преобразования

$$V(x, y, z) \Leftrightarrow V(g_m(x), g_m(z), g_m(y)).$$

В этом примере  $\Delta = \delta = 0$ . Кроме того, глубина вложенности терма с преобразованием  $g_m$  не превышает 1, так как  $g_m(g_m(x)) = x$ . Время распознавания изображения многогранника, отличающегося от эталонного аффинным преобразованием, увеличится разве лишь вдвое по сравнению с распознаванием эталонного изображения.

2. Пусть имеется множество изображений на экране дисплея, заданных матрицей яркости. Такие изображения могут быть описаны с помощью одного предиката  $p(x, i, j) \Leftrightarrow$  “пиксель с координатами  $(i, j)$  имеет яркость  $x$ ”. Этот предикат не инвариантен относительно аффинных преобразований, но для него можно выписать их описания преобразований. Приведем пример описания растяжения в два раза по оси  $Ox$

$$p(x, i, j) \Leftrightarrow p(x, 2i, j) \& p(x, 2i+1, j)$$

и пример описания сжатия в два раза по оси  $OX$

$$p(x_1, 2i, j) \& p(x_2, 2i+1, j) \Leftrightarrow p((x_1+x_2)/2, i, j).$$

В этих примерах  $\Delta = 0$ ,  $\delta = 1$ .

---

### Благодарности

Статья частично финансирована из проекта ИТНЕА XXI Института Информационных теорий и Приложений FOI ИТНЕА и консорциума FOI Bulgaria ([www.ithea.org](http://www.ithea.org), [www.foibg.com](http://www.foibg.com))

---

### Библиография

- [1] Т.М.Косовская, А.В.Тимофеев. Об одном новом подходе к формированию логических решающих правил // Вестник Ленинградского университета. 1985. № 8. С. 22 – 29.
- [2] Т.М.Косовская. Доказательства оценок числа шагов решения некоторых задач распознавания образов, имеющих логические описания // Вестн. С.-Петербур. ун-та. Сер. 1: Математика, механика, астрономия. 2007. Вып.4. С. 82 – 90.
- [3] Т.М.Косовская. Частичная выводимость предикатных формул как средство распознавания объектов с неполной информацией // Вестн. С.-Петербур.ун-та. Сер. 10. 2009. Вып. 1. С. 74 – 84.

---

### Информация об авторе



**Татьяна Косовская** – Докторант, СПИИРАН, 14 линия, д.39, Санкт-Петербург, 199178, Россия; e-mail: [kosov@NK1022.spb.edu](mailto:kosov@NK1022.spb.edu)

Основные направления научных исследований: Теория распознавания образов, теория сложности алгоритмов