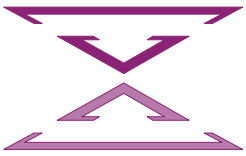


INFORMATION SCIENCE & COMPUTING

International Book Series Number 8

# Classification, Forecasting, Data Mining

Supplement to
International Journal "Information Technologies and Knowledge" Volume 3 / 2009

ITHEA
SOFIA, 2009

Krassimir Markov, Vladimir Ryazanov, Krassimira Ivanova, Ilia Mitov (ed.)

Classification, Forecasting, Data Mining

International Book Series "INFORMATION SCIENCE & COMPUTING", Number 8

Supplement to the International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" Volume 3 / 2009

Institute of Information Theories and Applications FOI ITHEA

Sofia, Bulgaria, 2009

This issue contains a collection of papers in the fields of Classification and Clustering, Pattern Recognition and Forecasting, Features Processing and Transformations, and Data Mining and Knowledge Discovery.

Papers in this issue are selected from the International Conferences of the Joint International Events of Informatics "ITA 2009", Summer Session, Varna, Bulgaria.

**International Book Series "INFORMATION SCIENCE & COMPUTING", Number 8**
**Supplement to the International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" Volume 3, 2009**

Edited by **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria,
in collaboration with
- **V.M.Glushkov Institute of Cybernetics of NAS**, Ukraine,
- **Institute of Mathematics and Informatics, BAS**, Bulgaria,
- **Institute of Information Technologies, BAS**, Bulgaria.

Publisher: Institute of Information Theories and Applications FOI ITHEA, Sofia, 1000, P.O.B. 775, Bulgaria.
Издател: Институт по информационни теории и приложения ФОИ ИТЕА, София, 1000, п.к. 775, България
www.ithea.org, www.foibg.com, e-mail: info@foibg.com

General Sponsor: **Consortium FOI Bulgaria** (www.foibg.com).

Printed in Bulgaria

**Copyright © 2009 All rights reserved**
© 2009 Institute of Information Theories and Applications FOI ITHEA - Publisher
© 2009 Krassimir Markov, Vladimir Ryazanov, Krassimira Ivanova, Ilia Mitov – Editors
© 2009 For all authors in the issue.

**ISSN 1313-0455 (printed)** **ISSN 1313-048X (online)** **ISSN 1313-0501 (CD/DVD)**

# PREFACE

The scope of the International Book Series “Information Science and Computing” (**IBS ISC**) covers the area of Informatics and Computer Science. It is aimed to support growing collaboration between scientists from all over the world. IBS ISC is official publisher of the works of the members of the ITHEA International Scientific Society.

The official languages of the IBS ISC are English and Russian.

IBS ISC welcomes scientific papers and books connected with any information theory or its application.

IBS ISC rules for preparing the manuscripts are compulsory.

The rules for the papers and books for IBS ISC are given on www.foibg.com/ibsisc .

The camera-ready copyes of the papers should be received by ITHEA Submition System http://ita.ithea.org .

The camera-ready copyes of the books should be received by e-mail: info@foibg.com .

Responsibility for papers and books published in IBS ISC belongs to authors.

The Number 8 of the IBS ISC contains collection of papers from the fields of Classification, Clustering, Pattern Recognition, Forecasting, Features Processing, Transformations, Data Mining, and Knowledge Discovery.

Papers are peer reviewed and are selected from the several International Conferences, which were part of the Joint International Events of Informatics "ITA 2009" – summer session, Varna, Bulgaria.

The book maintains articles on actual problems of classification, data mining and forecasting:

- New approaches, algorithms and methods of construction of steady and smooth logic algorithms of type of computation of the estimations, steady piece-wise linear algorithms of classification;
- The algebraic theory of algorithms - problems of complexity and resolvability of challenges of classification, construction of optimum algebraic proof-readers over sets of algorithms of computation of estimations;
- Methods of search of logic regularities of classes (knowledge) and their statistical verification, association rule mining, extract of knowledge by means of neural networks;
- Researches in area of neural network classifiers and self-organizing maps, principles of designing and results of use heterogeneous gene - neural networks;
- Questions of complexity of some discrete optimization tasks and corresponding tasks of data analysis and pattern recognition;
- Estimation of probability of erroneous classification, comparison of approaches and optimization of estimations, risk estimation in regression models;
- The specialized task-oriented algorithms for analysis and recognition of numerical and vector sequences, structures in DNA-sequences, methods of automatic classification and modeling of a genetic code;
- Logic and probabilistic models constructing for multivariate heterogeneous time series,
- Machine learning methods for variable aggregation and transformation.

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.

ITA 2009 has been organized by

ITHEA International Scientific Society

in collaboration with:

- Institute of Information Theories and Applications FOI ITHEA
- Dorodnicyn Computing Centre of the Russian Academy of Sciences
- International Journal “Information Theories and Applications”
- International Journal "Information Technologies and Knowledge"
- Association of Developers and Users of Intelligent Systems (Ukraine)
- Association for Development of the Information Society (Bulgaria)
- V.M.Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine
- Institute of Mathematics and Informatics, BAS (Bulgaria)
- Institute of Information Technologies, BAS (Bulgaria)
- Institute of Mathematics of SD RAN (Russia)
- Taras Shevchenko National University of Kiev (Ukraine)
- Universidad Politecnica de Madrid (Spain)
- BenGurion University (Israel)
- Rzeszow University of Technology (Poland)
- University of Calgary (Canada)
- University of Hasselt (Belgium)
- Kharkiv National University of Radio Electronics (Ukraine)
- Astrakhan State Technical University (Russia)
- Varna Free University “Chernorizets Hrabar” (Bulgaria)
- National Laboratory of Computer Virology, BAS (Bulgaria)
- Uzhgorod National University (Ukraine)

The main ITA 2009 events were:

| | |
|---|---|
| **KDS** | XVth International Conference "Knowledge - Dialogue – Solution" |
| **i.Tech** | Seventh International Conference "Information Research and Applications" |
| **MeL** | Fourth International Conference "Modern (e-) Learning" |
| **INFOS** | Second International Conference "Intelligent Information and Engineering Systems" |
| **CFDM** | International Conference "Classification, Forecasting, Data Mining" |
| **GIT** | Seventh International Workshop on General Information Theory |
| **ISSI** | Third International Summer School on Informatics |

More information about ITA 2009 International Conferences is given at the www.ithea.org .

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

We express our thanks to all authors, editors and collaborators who had developed and supported the International Book Series "Information Science and Computing".

General Sponsor of IBS ISC is the **Consortium FOI Bulgaria** (www.foibg.com).

*Sofia, June 2009* *Kr. Markov, Vl. Ryazanov, Kr. Ivanova, I. Mitov*

# TABLE OF CONTENTS

## Features Processing and Transformations



## Data Mining and Knowledge Discovery

# INDEX OF AUTHORS

# Classification and Clustering

## OPTIMAL DECISION RULES IN LOGICAL RECOGNITION MODELS

**Anatol Gupal, Vladimir Ryazanov**

***Abstract**: The task of smooth and stable decision rules construction in logical recognition models is considered. Logical regularities of classes are defined as conjunctions of one-place predicates that determine the membership of features values in an intervals of the real axis. The conjunctions are true on a special no extending subsets of reference objects of some class and are optimal. The standard approach of linear decision rules construction for given sets of logical regularities consists in realization of voting schemes. The weighting coefficients of voting procedures are done as heuristic ones or are as solutions of complex optimization task. The modifications of linear decision rules are proposed that are based on the search of maximal estimations of standard objects for their classes and use approximations of logical regularities by smooth sigmoid functions.*

***Keywords**: precedent-recognition recognition, logical regularities of classes, estimate calculation algorithms, integer programming, decision rules, sigmoid formatting rules*

***Conference**: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

### Introduction

The paper is dedicated to development of recognition algorithms that are based on partial-precedence principle (logical-combinatorial methods, estimate calculation algorithms). The first studies in this field were made by Yu.I.Zhuravlev (a test algorithm [Dmitriev, 1966] , estimate calculation algorithms [Zhuravlev, 1971]), recognition model based on voting over representative sets [Baskakova, 1981]. The well-known practical recognition algorithm Kora has been presented in [Vaintsvaig, 1973]. The basic principle of these algorithms is the search of irredundant fragments of objects descriptions in terms of features that are the incident ones to the classes. Such important fragments are used later for recognition of new objects. These models were elaborated for *k*-valued features. To work with real-valued features, the data quantization is made in advance that preserves the separability of classes on training sample [Zhuravlev, 1978], [Zhuravlev, 1998], [Zhuravlev, 2002], [Dyukova, 2000], [Dyukova, 1989]. Later, the term logical regularity (LR) will be used. The predicate $P(S) = A_1(S)\,\&\,A_2(S)\,\&\ldots\&\,A_k(S)$ will be understood as logical regularity, where $A_1, A_2, \ldots, A_k$ are one-placed predicates that depend on one of the features and determine the membership of the value of this feature in a certain interval of the real axis. The LR is true for all reference objects of some "no extending" subsets $\tilde{S}^*$ of training sample $\tilde{S}$ belonging to class $K_i$, moreover $P(S) = 0, \forall S \in CK_i \cap \tilde{S}$.

In [Kochetkov, 1989], recognition algorithms have been proposed that are invariant under some transformations of features, and some practical method for LR search was described [Bushmanov, 1988]. In [Ryazanov, 2007], [Kovshov, 2008], the parametrical approach was considered. The LR is described by vector of binary parameters and LR search is reduced to solution of special integer-valued mathematical programming task. It was proposed relaxation, combinatorial and genetic algorithms for LR search.

This paper is an extension of investigation [Ryazanov, 2007]. Let the sets of LR of all classes have been found by training sample. The LR of minimal complexity and equivalent to some one LR is calculated. To recognize any object $S$, the weighted sum of values of one-parametric sigmoid approximations of LR for $S$ is calculated. Some restrictions for weight coefficients in terms of equations are used. Finally, the task of construction of stable and smooth decision rule is reduced to linear programming problem. Coefficients of matrix of restrictions are the functions of smooth parameter. The algorithm for construction of stable smooth decision rules have been approved successfully by model and real data.

## Main Definitions

We consider the standard recognition task by precedents with $n$ numerical features $x_1, x_2, ..., x_n$, $l$ nonintersecting classes $K_1, K_2, ..., K_l$ and training sample $\tilde{S} = \{S_1, S_2, ..., S_m\}$. We use notation $\tilde{K}_i = \tilde{S} \cap K_i, i = 1, 2, ..., l,$ and suppose that $\tilde{K}_i \neq \varnothing,\ i = 1, 2, ..., l$.

Let $S = (x_1(S), x_2(S), ..., x_n(S))$, $S \in \bigcup_{i=1}^{l} K_i$, $S_t = (a_{t1}, a_{t2}, ..., a_{tn}), a_{tj} = x_j(S_t)$, $x_i \in R$.

The next parametric set of elementary predicates is considered

$$P^{1,c_j}(x) = \begin{cases} 1, & c_j \le x, \\ 0, & \text{otherwise}, \end{cases} \quad P^{2,c_j}(x) = \begin{cases} 1, & x \le d_j, \\ 0, & \text{otherwise} \end{cases}, \text{ where } c_j, d_j \in R, j = 1,2,...,n.$$

Let $\Omega \subseteq \{1,2,...,n\}$.

**Definition.** The predicate

$$P^{\Omega_1,\Omega_2,\mathbf{c},\mathbf{d}}(\mathbf{x}) = \mathop{\&}_{j\in\Omega_1} P^{1,c_j}(x_j) \mathop{\&}_{j\in\Omega_2} P^{2,d_j}(x_j) \qquad (1)$$

is called a logical regularity of the class $K_\lambda$, $\lambda = 1, 2, ..., l$, if it holds that

$\exists\ S_t \in \tilde{K}_\lambda :\ P^{\Omega_1,\Omega_2,\mathbf{c},\mathbf{d}}(S_t) = 1$,

$\forall\ S_t \notin \tilde{K}_\lambda,\ P^{\Omega_1,\Omega_2,\mathbf{c},\mathbf{d}}(S_t) = 1$,

$$\Phi(P^{\Omega_1,\Omega_2,\mathbf{c},\mathbf{d}}(\mathbf{x})) = \mathop{extr}_{P^{\Omega'_1,\Omega'_2,\mathbf{c}',\mathbf{d}'}(\mathbf{x})} \Phi(P^{\Omega'_1,\Omega'_2,\mathbf{c}',\mathbf{d}'}(\mathbf{x}))$$

Later, we consider the predicate objective function $\Phi(P^{\Omega_1,\Omega_2,\mathbf{c},\mathbf{d}}(\mathbf{x})) = \left|\{S_i : S_i \in \tilde{K}_\lambda, P^{\Omega_1,\Omega_2,\mathbf{c},\mathbf{d}}(S_i) = 1\}\right|$ to be maximized. The set $N(P^{\Omega_1,\Omega_2,\mathbf{c},\mathbf{d}}) = \{\mathbf{x} \in R^n : c_j \le x_j, j \in \Omega_1, x_j \le d_j, j \in \Omega_2\}$ is called the interval of the predicate $P^{\Omega_1,\Omega_2,\mathbf{c},\mathbf{d}}(\mathbf{x})$. The predicates $P^{\Omega_1,\Omega_2,\mathbf{c},\mathbf{d}}(\mathbf{x})$, $P^{\Omega'_1,\Omega'_2,\mathbf{c}',\mathbf{d}'}(\mathbf{x})$ are said to be equivalent if $P^{\Omega_1,\Omega_2,\mathbf{c},\mathbf{d}}(S_t) = P^{\Omega'_1,\Omega'_2,\mathbf{c}',\mathbf{d}'}(S_t), t = 1, 2, ..., m$. Two intervals $N(P^{\Omega_1,\Omega_2,\mathbf{c},\mathbf{d}})$, $N(P^{\Omega'_1,\Omega'_2,\mathbf{c}',\mathbf{d}'})$ are said to be equivalent if their predicates are equivalent ones.

The feasible predicate $P^{\Omega_1,\Omega_2,\mathbf{c},\mathbf{d}}(\mathbf{x})$ is local-optimal with respect to the criterion $\Phi(P^{\Omega_1,\Omega_2,\mathbf{c},\mathbf{d}}(\mathbf{x}))$ if there are no feasible predicates $P^{\Omega'_1,\Omega'_2,\mathbf{c}',\mathbf{d}'}(\mathbf{x})$ such that $N(P^{\Omega'_1,\Omega'_2,\mathbf{c}',\mathbf{d}'}) \supseteq N(P^{\Omega_1,\Omega_2,\mathbf{c},\mathbf{d}})$, $\Phi(P^{\Omega'_1,\Omega'_2,\mathbf{c}',\mathbf{d}'}) > \Phi(P^{\Omega_1,\Omega_2,\mathbf{c},\mathbf{d}})$.

## Optimization of Logical Decision Rules

Assume that we have the set of LRs $\mathrm{P}_\lambda = \{P^{\Omega_1,\Omega_2,\mathbf{c},\mathbf{d}}(\mathbf{x})\}$ for each class $K_\lambda$, and the set of intervals $\{N(P^{\Omega_1,\Omega_2,\mathbf{c},\mathbf{d}}) : P^{\Omega_1,\Omega_2,\mathbf{c},\mathbf{d}} \in \mathrm{P}_\lambda\}$ coves $\tilde{K}_\lambda$. The algorithms for finding $\mathrm{P}_\lambda$ have been proposed in [13]. We say that the LR from $\mathrm{P}_\lambda$ has the minimal complicity if there is not any equivalent one that has smaller value of $\Omega_1 + \Omega_2$. Let some LR $P^{\Omega_1,\Omega_2,\mathbf{c},\mathbf{d}}(\mathbf{x}) \in \mathrm{P}_\lambda$ is known. The equivalent LR of minimal complicity is founded as some solution of the following integer linear programming task:

$$\sum_{i\in\Omega_1} y_{1i} + \sum_{i\in\Omega_2} y_{2i} \to \min,$$

$$\sum_{i\in\Omega_1} (1 - P^{1,c_i}(a_{ti}))y_{1i} + \sum_{i\in\Omega_2} (1 - P^{2,d_i}(a_{ti}))y_{2i} \geq 1, \forall S_t \in \tilde{S} \setminus \tilde{K}_\lambda,$$

$$y_{1i} \in \{0,1\}, i \in \Omega_1, y_{2i} \in \{0,1\}, i \in \Omega_2.$$

The unities in $y_{1i}, y_{2i}$ define corresponding subsets $\Omega_1, \Omega_2$ for LR to be find. Later, we assume that the sets $\mathrm{P}_\lambda$ consist of LRs of minimal complicity.

The standard approach to recognizing of any object $S$ by estimate calculation algorithms is the following one.

1. The estimation $\Gamma_j(S) = \sum_{P_t \in \mathrm{P}_j} P_t(S)$ (2)

is calculated for any object $S$ and class $K_j$.

2. The standard decision rule $\alpha_j^A(S) = \begin{cases} 1, & \sum_{i=1}^{l} \delta_i^j \Gamma_i(S) \geq \delta_i^0, \\ 0, & \text{otherwise.} \end{cases}$ is used (or the simpler $\alpha_j^A(S) = \begin{cases} 1, & \Gamma_j(S) > \Gamma_i(S), i \neq j, \\ 0, & \text{otherwise.} \end{cases}$).

The notation $\alpha^A_j(S)=1$ $(\alpha^A_j(S)=0)$ denotes the solution $S \in K_j$ ($S \notin K_j$) of algorithm $A$. Parameters $\delta_i^j$ are founded in optimization process of recognition model with the use of control sample. The given scheme of recognition has obvious lacks.

1. An arbitrariness in calculation of estimations (2) as result of absence of weight factors of LRs.

2. Graduated character of estimations as functions of signs does not allow estimating stability of a solving rule.

3. Now there are no effective methods of optimization of standard criterion of quality of models of calculation of estimations with use of control sample.

Let's notice that as $\Gamma_j(S_t) > 0, S_t \in \tilde{K}_j$ and $\Gamma_j(S_t) = 0, S_t \in \tilde{S} \setminus \tilde{K}_j$, the algorithm is faultless on objects of the table of training at use of the elementary solving rule. Its extrapolating abilities thus to the user are not known. The following updating resulted above the general scheme of algorithms of calculation of estimations is offered.

Estimations for classes are calculated according to (3)

$$\Gamma_j(S) = \sum_{P_t \in \mathrm{P}_j} \gamma_t f_t\ (S)\ , \tag{3}$$

where $\gamma_t = \gamma_t(P_t^{\Omega_1^t, \Omega_2^t, \mathbf{c}_t, \mathbf{d}_t})$ - the non-negative parameters characterizing "weight" corresponding LT $P_t^{\Omega_1^t, \Omega_2^t, \mathbf{c}_t, \mathbf{d}_t} \in \mathrm{P}_j$, $f_t\ (S)$ - approximating LR $P_t^{\Omega_1^t, \Omega_2^t, \mathbf{c}_t, \mathbf{d}_t}$ sigmoid kind function $f_t\ (S) = \prod_{i \in \Omega_1^t} \frac{1}{1 + \exp(-\delta(x_i(S) - c_{ti}))} \prod_{i \in \Omega_2^t} \frac{1}{1 + \exp(\delta(x_i(S) - d_{ti}))}$.

Classification of $S$ is spent on a maximum of estimations (3). The parameter $\delta$ sets «smoothness degree» of LR approximations. Parameters $\gamma_t, t = 1,2,..., N$ ( $N$ - total number of logical regularities of all classes) are the solution of the following problem of linear programming:

$$\sigma \to \max\ , \tag{4}$$

$$\sum_{P_t \in \mathbf{P}_j} \gamma_t f_t\ (S_t) \geq \sigma,\ S_t \in K_j, t = 1, 2, ..., m, j = 1, 2, ..., l \tag{5}$$

$$\sum_{i=1}^{N} \gamma_i = N,\ \gamma_i \geq 0,\ i = 1, 2, ..., N, \tag{6}$$

In a problem (4) - (6) there are such weights factors for LRs of classes at which estimations of standards for classes will be maximum one. Thus, for the set degree of smoothness $\delta$ there are weight parameters $\gamma_i$, $i = 1, 2, ..., N$, providing steadiest solutions on the training data. The given approach is direct analogue search of the maximum gap in a support vector machine [Burges, 1998]. The algorithm of construction of steady smooth solving rules is successfully approved on the model and real data.

## Acknowledgements

The authors are glad to acknowledge support of the following organizations for execution of the described research: RFBR (projects 08-01-90427 ukr, 08-01-00636). The work has been also supported by the Presidium's program N2 of RAS.

## Bibliography

[Dmitriev, 1966] A.N.Dmitriev, Yu.I.Zhuravlev, and F.P.Krendelev, "On Mathematical Principles of Classification of Objects and Phenomena", in Discrete Analysis (Institute of Mathematics, Siberian Division, USSR Academy of Sciences, Novosibirsk, 1966), No. 7, pp.3-11 [in Russian].

[Zhuravlev, 1971] Yu.I.Zhuravlev and V.V.Nikiforov, "Recognition Algoithms based on Estimates Calculation", Kibernetika, No.3, pp. 1-11 (1971).

[Zhuravlev, 1978] Yu.I.Zhuravlev, "An Algebraic Approach to Recognition or Classification Problems ", in Problems of Cybernetics , Issue 33 (Nauka, Moscow, 1978; Hafner, 1986), pp.5-68.

[Vaintsvaig, 1973] M.N.Vaintsvaig, "Kora: A Learning Algorithm for Pattern Recognition", in Learning Algorithms for Pattern Recognition (Sovetskoe Radio, Moscow, 1973), pp. 8-12 [in Russian].

[Baskakova, 1981] L.V.Baskakova and Yu.I.Zhuravlev, "A Model of Recognition Algorithms with Representative Sets and Systems of Support Sets", Zh. Vychisl. Mat. Mat. Fiz. 21, pp.1264-1275 (1981).

[Zhuravlev, 1998] Yu.I.Zhuravlev, Selected Works (Magistr, Moscow, 1998) [in Russian].

[Zhuravlev, 2002] Yu.I.Zhuravlev. "Recognition Algorithms with Representative Sets (Logic Algorithms) Algorithms)" Zh. Vychisl. Mat. Mat. Fiz. **42**, 1425–1435 (2002) [Comput. Math. Math. Phys. **42**, 1372–1382 (2002)].

[Dyukova, 2000] E. V. Dyukova and Yu. I. Zhuravlev, "Discrete Analysis of Feature Descriptions in Recognition Problems of High Dimensionality," Zh. Vychisl. Mat. Mat. Fiz. **40**, 1264–1278 (2000) [Comput. Math. Math. Phys. **40**, 1214–1227 (2000)].

[Dyukova, 1989] E. V. Dyukova, "The Recognition Algorithms Kora: Complexity of Implementation and Metric Properties," in *Recognition, Classification, and Prediction: Mathematical Methods and Applications* (Nauka, Moscow, 1989), No. 2, pp. 99–125 [in Russian].

[Kochetkov, 1989] D. V. Kochetkov, "Recognition Algorithms that are Invariant under Transformations of the Space of Features," in *Recognition, Classification, and Prediction: Mathematical Methods and Applications* (Nauka, Moscow, 1989), No. 1, pp. 82–113; No. 2, pp. 178–206; No. 3, pp. 64–88 [in Russian].

[Bushmanov, 1988] O.N.Bushmanov, E. V. Dyukova, Yu. I. Zhuravlev, Дюкова E.B., Yu.I.Zhuravlev, D. V. Kochetkov, V.V.Ryazanov "Program system for data analysis and pattern recognition" in *Recognition, Classification, and Prediction: Mathematical Methods and Applications* (Nauka, Moscow, 1988), No. 2, pp. 250-273 [in Russian].

[Ryazanov, 2007] V.V.Ryazanov, "Logical Regularities in Pattern Recognition (Parametric Approach)" , Zh. Vychisl. Mat. Mat. Fiz. 2007. T.47, № 2.

[Kovshov, 2008] N. V. Kovshov, V. L. Moiseev, and V. V. Ryazanov, "Algorithms for Finding Logical Regularities in Pattern Recognition", Zh. Vychisl. Mat. Mat. Fiz.*, 2008, Vol. 48, No. 2, pp. 314–328.*

[Burges, 1998] Christopher J.C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition, Appeared in: Data Mining and Knowledge Discovery 2, 121-167, 1998.

## Authors' Information

***Gupal A.M.*** *– Head of Department, Glushkov Institute of Cybernetics NAS Ukraine, Akademision Glushkov st., 40, Kiev, 03680 MCP, Ukraina, e-mail: gupal_anatol@mail.ru*

***Ryazanov V.V.*** *– Head of Department, Computing Centre of the Russian Academy of Sciences, 40 Vavilova St., Moscow GSP-1, 119991, RUSSIAN FEDERATION, e-mail: rvvccas@mail.ru*

# EXACT DISCRIMINANT FUNCTION DESIGN USING SOME OPTIMIZATION TECHNIQUES

**Yury Laptin, Alexander Vinogradov**

***Abstract**: Some aspects of design of the discriminant functions that in the best way separate points of predefined final sets are considered. The concept is introduced of the nested discriminant functions which allow to separate correctly points of any of the final sets. It is proposed to apply some methods of non-smooth optimization to solve arising extremal problems efficiently.*

***Keywords**: cluster, solving 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*

***Conference**: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

## Introduction

Linear decision rule (LDR) keep relative simplicity at high computational efficiency. At use of the algorithms realizing LDR, the raised speeds of recognition can be reached that is important for the decision of various problems concerned to mass data processing. At the same time, construction of the best LDR quite often leads to posing complex optimization problems. Situation with strongly overlapped classes under condition of weakness of stochastic components in data can serve here as an example, when search exact LDR with a zero mistake on training sample is justified, but encounters difficulties of strictly combinatory character [1]. Similar difficulties arise also when each pair of classes is easily separable by means of LDR, but the number of classes is great. In such situations crucial importance gets a choice of an adequate method of solving the optimization problem. Researches on the given direction are carried out all over the world and continue to remain actual, since are based and supported from two parties, as by progress in the field of creation of new methods of optimization, as by successes of the theory of recognition [2-6]. In this work some applications of methods of non-smooth optimization are considered in problems of search of linear discriminant functions (linear classifiers) correctly separating clusters as final sets in $R^n$.

## 1. Simple discriminant functions

Let's consider as predefined some collection of final sets $\Omega_i = \left\{ p^t \in R^n,\ t \in T_i \right\},\ i = 1,..., m$, where $T_i$ is the set of point indices in $\Omega_i$. We use the term *discriminant function* for any function $\pi : R^n \to \{1,..., m\}$.

Let functions $f_i : R^n \to R,\ i = 1,..., m$, be set. In the further we consider discriminant functions of the following kind

$$\pi(x) = \arg\max_i \left\{ f_i(x) :\ i = 1,..., m \right\}. \qquad (1)$$

We say that discriminant function $\pi(x)$ correctly divides points from $\Omega_i,\ i = 1,..., m$, if $\pi(x) = i$, for all $x \in \Omega_i,\ i = 1,..., m$. Set $K_i = \left\{ x \in R^n : \pi(x) = i \right\}$ is referred to as *class* $K_i$ *generated by function* $\pi(x)$.

**Remark 1.** Function $\pi(x)$ is invariant concerning to multiplication of all functions $f_i$ by positive value, and to addition of any value to all of $f_i$.

Function $\pi(x)$ of a kind (1) is named *simple discriminant function* if all functions $f_i$ are linear. Let $m = 2$. It is easy to see, that if simple discriminant function correctly divides points of two final sets, a hyperplane defined by condition

$$(a^1, x) + b_1 = (a^2, x) + b_2, \quad (2)$$

separates sets $\Omega_1$, $\Omega_2$.

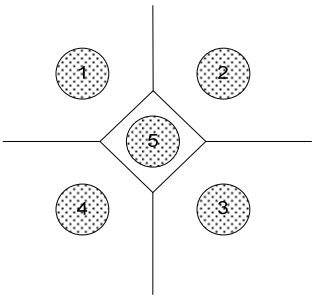



Fig. 1.

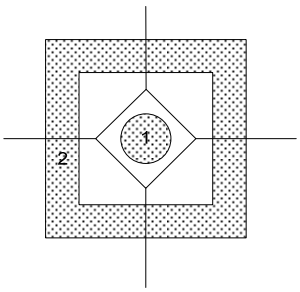



Fig. 2.

On Fig.1 an example of sets in $R^2$ and the division of a plane into classes by simple discriminant function is presented. Sets 1, 2, …, 5 are circles of radius 1 placed, accordingly, in points (-2,2), (2,2), (2,-2), (-2,-2), (0,0). Linear functions $l_i(x) = (a^i, x) + b_i$: $a^1 = (-1,1)$, $a^2 = (1,1)$, $a^3 = (1,-1)$, $a^4 = (-1,-1)$, $a^5 = (0,0)$; $b_i = 0$, $i = 1,..., 4$, $b_5 = 2$.

Generally (for any $m$) there is a question on existence of the discriminant function $\pi(x)$ correctly separating points from $\Omega_i$, $i = 1,..., m$.

**Theorem 1.** Let around of each set $\Omega_i$ the sphere $S_i$, $i = 1,..., m$, can be constructed, so that $S_i \cap S_j = \varnothing$, $i \neq j$. Then there is a simple discriminant function $\pi(x)$ separating points from $\Omega_i$, $i = 1,..., m$ correctly.

**Proof.** We shall consider all over again a case when each set $\Omega_i$ consists of one point. Let $F(x)$ be strictly convex smooth function such that all points from $\Omega_i$, $i = 1,..., m$ belong to domain of $F(x)$. To each set $\Omega_i = \{p^i\}$ we shall put in correspondence the function $f_i(x) = F(p^i) + (\nabla F(p^i), x - p^i)$, $i = 1,..., m$. By the strict convexity it is forced that $f_i(p^i) = F(p^i) > F(p^j) + (\nabla F(p^j), p^i - p^j) = f_j(p^i)$, $j \neq i$. Whence it follows, that discriminant function $\pi(x)$ correctly separates points from $\Omega_i$, $i = 1,..., m$.

Let's pass to the general case. As function $F(x)$ we shall choose a hemisphere of enough the big radius $r$ in space $R^{n+1}$ which center is located in a point $(x^0, r)$, where $x^0$ is fixed, and $r$ we shall vary (if necessary). For each set $\Omega_i$ we shall select linear function $f_i(x) = (a^i, x) + b_i$. We shall designate $E_i = \{x \in R^n : (a^i, x) + b_i \geq F(x)\}$. The set $E_i$ is a projection of crossing of a plane and a semicircle in $R^{n+1}$ on space $R^n$. We shall consider such linear functions $f_i(x) = (a^i, x) + b_i$, for which $E_i$ is an ellipsoid. It is easy to see, that if radius $r$ is big enough then always it is possible to choose function $f_i(x) = (a^i, x) + b_i$ so that $S_i \subseteq E_i$ will be valid. We shall choose functions $f_i(x)$ so that corresponded to

them ellipsoids $E_i$ had the minimal size (with the minimal small axis) and $S_i \subseteq E_i$ was still valid. It is easy to see, that increasing radius $r$ of a hemisphere it is possible always to achieve that ellipsoids $E_i$, $i = 1,..., m$, were not crossed.

Let such functions $f_i(x) = (a^i, x) + b_i$ are constructed, ellipsoids $E_i$ corresponding to them are not crossed and $S_i \subseteq E_i$ holds for all $i = 1,..., m$. It is easy to see, that at construction we have $F(x) > f_j(x),\ x \notin E_j$ and $f_i(x) \geq F(x) > f_j(x),\ x \in E_i,\ i \neq j$. Thus, $f_i(x) > f_j(x),\ x \in E_i,\ i \neq j$, and the discriminant function $\pi(x)$ does separate correctly points from $\Omega_i$, $i = 1,..., m$. Theorem is proved ■.

It should be noticed, that conditions of the Theorem 1 are rather rigid. It is possible to find many examples where these conditions don't hold, but the correct discriminant function for $\Omega_i$, $i = 1,..., m$ does exist.

Let's introduce a criterion of quality of function concerning to collection $\Omega_i \subset R^n$, $i = 1,..., m$

$$\delta(\pi) = \min\left\{ f_i(x) - f_j(x) : j \in \{1,..., m\} \setminus i,\ x \in \Omega_i,\ i = 1,..., m \right\}, \tag{3}$$

The criterion $\delta(\pi)$ characterizes how much values of functions $f_j(x),\ j \in \{1,..., m\} \setminus i$ differ from values $f_i(x)$ in points $x \in \Omega_i$. It is obvious, that if $\delta(\pi) > 0$ holds then the function $\pi(x)$ correctly separates points from $\Omega_i \subset R^n$, $i = 1,..., m$. Design of simple discriminantal function $\pi(x)$ is equivalent to a choice of values of vectors $a^i$ and parameters $b_i$, $i = 1,..., m$. In view of the Remark 1 the problem of choosing the best simple discriminant function for criterion $\delta(\pi)$ we shall present in the form of a problem of linear programming: to find

$$\delta^* = \max_{a,b,\delta} \delta, \tag{4}$$

at restrictions

$$(a^i - a^k, p^t) + b_i - b_k \geq \delta,\ t \in T_i,\ k \in \{1,..., m\} \setminus i,\ i = 1,..., m, \tag{5}$$

$$-1 \leq a_j^i \leq 1,\ i = 1,..., m,\ j = 1,..., n. \tag{6}$$

$$b_1 = 0. \tag{7}$$

Restriction (7) is added in view of invariance of functions $\pi(x)$ concerning addition of any number to all $f_i$. Restrictions (6) are the normalizing conditions. These conditions can be written as restrictions put on the norms:

$$\left\| a^i \right\|^2 \leq 1,\ i = 1,..., m. \tag{8}$$

In this case the problem (4), (5), (7), (8) will be a problem of quadratic programming.

It is easy to see, that if there exists the simple discriminant function $\pi(x)$ correctly separating points from $\Omega_i$, $i = 1,..., m$, then $\delta^* > 0$ and the decision of the problem (4) - (7) defines optimum discriminant function. Otherwise, any set for which $a^i = a^k$, $b_i = b_k$, $i, k \in \{1,..., m\}$, is optimum, $\delta^* = 0$, and the decision of problem (4) - (7) does not contain useful information.

Variables number of problem (4)-(7) is equal to $m(n+1)+1$, number of restrictions (5) – $N(m-1)+1$, where $N$ – total number of points in sets $\Omega_i$, $i = 1,..., m$.

For large $N$ it is advisable to consider the problem (4), (5), (7), (8) and to represent it in the form: find

$$\delta^* = \max_{a,b} \left\{ \min\left\{ (a^i - a^k, p^t) + b_i - b_k :\ t \in T_i,\ k \in \{1,..., m\} \setminus i,\ i = 1,..., m \right\} \right\}, \tag{9}$$

subject to (7), (8). Objective function of this problem is peace-wise linear, so, non-smooth optimization methods [**Error! Reference source not found.**] could be used to solve this problem.

In the case, when $\delta^* = 0$ for the problem (4)-(7), finding good simple discriminant function will be realized in two stages. Analogous approaches were considered in [7, 8]. At the first stage it is proposed to exclude some points from the sets $\Omega_i$, $i = 1,..., m$ in such a way that for other points inequality $\delta^* \geq \overline{\delta}$ be satisfied for the problem (4)-(7), where $\overline{\delta}$ is a parameter. On the second stage the values of $b_i$, $i = 1,..., m$ have to be chosen to improve the discriminant function.

Denote $T = \bigcup_{i=1}^{m} T_i$. Let associate with every point $p^t$, $t \in T$ a variable $y_t = 0 \vee 1$ such that $y_t = 1$, if a point $p^t$ should be considered while formulating the problem (4)-(7), and $y_t = 0$ otherwise. Let parameter $\overline{\delta} > 0$ and large positive number $M$ be given.

The problem of exclusion some points from the sets $\Omega_i$, $i = 1,..., m$ has the form: find

$$\max_{a,b,y} \left\{ \sum_{t \in T} y_t \right\}, \tag{10}$$

subject to

$$(a^i - a^k, p^t) + b_i - b_k + M(1 - y_t) \geq \overline{\delta}, \quad t \in T_i, \; k \in \{1,..., m\} \setminus i, \; i = 1,..., m, \tag{11}$$

$$-1 \leq a_j^i \leq 1, \; i = 1,..., m, \; j = 1,..., n, \tag{12}$$

$$\sum_{t \in T_i} y_t \geq 1, \quad i = 1,..., m, \tag{13}$$

$$0 \leq y_t \leq 1, \; t \in T, \tag{14}$$

$$b_1 = 0. \tag{15}$$

$$y_t = 0 \vee 1, \; t \in T, \tag{16}$$

It is evident that if $y_t = 0$, then for sufficiently large $M$ corresponding inequality of form (11) will be satisfied for any $a^i$, $b_i$, i.e. the point $p^t$ is excluded from the problem.

Constraints (13) specify the condition that at least one point from every set $\Omega_i$ must be included in the problem.

Let an approximate solution $\overline{a}^i$, $\overline{b}_i$, $i \in \{1,..., m\}$, $\overline{y}_t$, $t \in T$ of the problem (10)-(16) is found. At the second stage to improve the discriminant function we solve the problem (4)-(7) under fixed variables $a^i = \overline{a}^i$, $i \in \{1,..., m\}$.

It should be noted that the resulting discriminant function does not guarantee proper separating of points from sets $\Omega_i$, $i = 1,..., m$.

## 2. Nested discriminant functions

Partitioning the sets $\Omega_i$ into non-overlapping sets $\Omega_i = \bigcup_{j \in J_i} \Omega_i^j$ will be referred to be effective, if it is possible to build a simple discriminant function for the whole $\Omega_i^j$, $j \in J_i$, $i = 1,..., m$, properly separating the points of these sets. Such discriminant function may not exist for initial sets $\Omega_i$, $i = 1,..., m$.

Nevertheless, effective partitioning always exists, for example, when every set $\Omega_i^j$ consists from one point.

Let an effective partitioning $\Omega_i = \bigcup_{j \in J_i} \Omega_i^j$, $i = 1,..., m$ be given. Denote $\pi^*(x)$ an optimal simple discriminant function for the sets $\Omega_i^j$, $j \in J_i$, $i = 1,..., m$,

$$\pi^*(x) = \arg\max_{ij} \left\{ (a^{ij}, x) + b_{ij} : \ i = 1,..., m, \ j \in J_i \right\}. \tag{17}$$

The function $\pi^*(x)$ returns a pair $\left(i^*(x), j^*(x)\right)$, giving a maximum in (17). It is evident, that $i^*(x)$ is a discriminant function properly separating points from $\Omega_i \subset R^n$, $i = 1,..., m$.

Denote

$$\psi_i^*(x) = \max \left\{ (a^{ij}, x) + b_{ij} : \ j \in J_i \right\}, \ i = 1,..., m. \tag{18}$$

It is easy to see that

$$i^*(x) = \arg\max_i \left\{ \psi_i^*(x) : \ i = 1,..., m \right\}. \tag{19}$$

Functions (19) will be named *nested discriminant function*. The use of nested discriminant function allows us to improve the quality of the best approximation of sets $\Omega_i$, $i = 1,..., m$.

Let we consider two sets in Fig. 2. The nested discriminant function has a form $i^*(x) = \arg\max \left\{ \psi_i^*(x) : \ i = 1, 2 \right\}$, where $\psi_1^*(x) = l_5(x)$, $\psi_2^*(x) = \max \left\{ l_i(x) : i = 1,..., 4 \right\}$, functions $l_i(x), \ i = 1,..., 5$ are determined for Fig. 1.

Heuristic scheme for finding a nested discriminant function consists from finite number of steps of handling the current partitioning $\Omega_i = \bigcup_{j \in J_i} \Omega_i^j$, $i = 1,..., m$, and looks as follows:

1) On the first step $k = 1$, take $\Omega_i$, $i = 1,..., m$ as a current partition of $\Omega_i = \bigcup_{j \in J_i} \Omega_i^j$, $i = 1,..., m$.

2) On $k$ th step solve the problem (4)-(7) for the current partitioning $\Omega_i = \bigcup_{j \in J_i} \Omega_i^j$, $i = 1,..., m$. If optimal value $\delta^* > 0$, the process is finished. Otherwise find an approximate solution of (10)-(16). On the basis of this solution every set $\Omega_j^j$ is divided into two subsets: points with $y_t = 0$ and points with $y_t = 1$. Then define the current partition more precisely, put $k = k + 1$ and go to 2).

It is easy to see that the process is finite, and as a result we get the nested discriminant function, properly separating points from $\Omega_i$, $i = 1,..., m$.

## Conclusions

Approaches for finding discriminant function separating points from given sets $\Omega_i \subset R^n$, $i = 1,..., m$ are considered. The problem of finding an optimal discriminant function is formulated as a linear (4)-(7) or quadratic (4), (5), (7), (8) programming problems. However this problem has a sense only in the case when there exists simple discriminant function, properly separating points from $\Omega_i$, $i = 1,..., m$.

In the case, when proper separating points from $\Omega_i$, $i=1,...,m$ is impossible, a two-stage procedure for finding a simple discriminant function is proposed. At the first stage it is proposed to exclude some points from the sets $\Omega_i$, $i=1,...,m$, and at the second stage the resulting discriminant function can be improved.

The notion of nested discriminators allowing to make properly separating of points from any disjoint sets $\Omega_i \subset R^n$, $i=1,...,m$ is introduced. An heuristic scheme for finding nested discriminator is proposed.

Optimization problems arising in the considered approaches are large-scale problems and have a great number of constraints. These problems can be reduced to the problem of maximization a concave piece-wise linear function with a great number of pieces under simple constraints. To solve them it is advisory to use non-smooth optimization methods [6] – generalized subgradient descent methods for large number of variables or methods with space transformation, if the number of variables does no exceed 300.

## Acknowledgements

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

## Bibliography

1. Гупал А.М., Сергиенко И.В. Оптимальные процедуры распознавания. - Киев: Наук.думка, 2008. - 232 с.
2. Koel Das, Zoran Nenadic. An efficient discriminant-based solution for small sample size problem // Pattern Recognition – Volume 42, Issue 5, 2009, Pages 857-866.
3. Juliang Zhang, Yong Shi, Peng Zhang. Several multi-criteria programming methods for classification // Computers & Operations Research – Volume 36, Issue 3, 2009, Pages 823-836.
4. 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.
5. Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию. – К.: Наукова думка, 2004. – 545 с.
6. Shor N.Z. Nondifferentiable Optimization and Polynomial Problems. – Dordrecht, Kluwer, 1998. – 394 p.
7. Bennett K.P., Mangasarian O.L. Robust Linear Programming Discrimination of Two Linearly Inseparable Sets // Optimization Methods and Software. – 1996. –№5. – P. 23-34.
8. Журбенко Н.Г., Саимбетов Д.Х. К численному решению одного класса задач робастного разделения двух множеств // Методы исследования экстремальных задач. – К.: Ин-т кибернетики им. В.М. Глушкова НАН Украины, 1994. – С. 52–55.

## Authors' Information

***Yury Laptin*** *–senior researcher, V.M.Glushkov Institute of Cybernetics of the NASU, Prospekt Academika Glushkova, 40, 03650 Kyiv, Ukraine; e-mail: laptin_yu_p@mail.ru*

***Alexander Vinogradov*** *– senior researcher, Dorodnicyn Computing Centre of the RAS, Vavilova 40, 119333 Moscow, Russian Federation; e-mail: vngrccas@mail.ru*

# CLASSIFICATION OF DATA TO EXTRACT KNOWLEDGE FROM NEURAL NETWORKS

**Ana Martinez, Angel Castellanos, Rafael Gonzalo**

***Abstract:*** *A major drawback of artificial neural networks is their black-box character. Therefore, the rule extraction algorithm is becoming more and more important in explaining the extracted rules from the neural networks. In this paper, we use a method that can be used for symbolic knowledge extraction from neural networks, once they have been trained with desired function. The basis of this method is the weights of the neural network trained. This method allows knowledge extraction from neural networks with continuous inputs and output as well as rule extraction. An example of the application is showed. This example is based on the extraction of average load demand of a power plant.*

***Keywords:*** *Neural Network, Backpropagation, Control Feedback Methods.*

***ACM Classification Keywords:*** *F.1.1 Models of Computation: Self-modifying machines (neural networks); F.1.2 Modes of Computation: Alternation and nondeterminism.*

***Conference****: The paper is selected from Seventh International Conference on Information Research and Applications – i.Tech 2009, Varna, Bulgaria, June-July 2009*

## Introduction

The ability of artificial neural network to learn and generalize from examples makes them very suitable for use in numerous applications, where exact algorithmic approaches are unknown or too difficult to implement. The knowledge learned during the training process is distributed in the weights of the different neurons; it is very difficult to comprehend exactly what the neural network is computing. The problem of representing the knowledge learned by the network in a comprehensible form received a great deal of attention in the actual literature [Andrews, R., Diederich, J., Tickle,A. 1995], [Andrews, R., Diederich, J., Golea, M. 1998], [Cloete, I., Zurada, J.M. 2000].

Although both expert systems and neural networks are typical systems in the domain of artificial intelligence, the basic components of these two kinds of systems are different. The knowledge base of expert systems is a set of rules which are stored in symbolic form, while neural networks encode learned knowledge within an established structure with adjustable weights in numerical form. Hence, it is difficult to transfer the training results of a neural network to the knowledge base of an expert system.

In contrast, neural networks have excellent abilities for classifying data and learning inputs [Freeman J.A., Skapura D.M. 1992], but it is difficult to describe the decision process of a neural network or to merge more than one trained neural network [Krishnan R., Sivakumar G., Bhattacharya P. 1999].

This paper shows the importance of the knowledge stored in the weights of a neural network. A trained neural network stores the acquired knowledge in numeric values that weights define [Apolloni, B. et al 2004], [Garcez d'Avila, A. S., Broda, K. and Gabbay D. M. 2001], [Chang, B.L., Hirsch, M. 1991]. The interpretation and extraction of such knowledge is a difficult task due to the special configuration of neural network and to the wide domain of patterns.

## Method to Extract Knowledge

Tasks to follow in order to perform a study of the importance of input, variables over output variables are the following ones:

1. Normalization of the input and output variables into the interval $[-1,1]$.
2. A neural network with *n* inputs and one output. The training algorithm considered is the backpropagation. Defining the activation function as sigmoid function.
3. Division of the values associated to the variable to forecast into two intervals, the positive one with a positive output $[0,1]$ and the negative interval with a negative output $[-1,0)$. These way two independent neural networks are defined in order to be trained.
4. Established an error threshold for the forecasting process, each one of the two output classes of the variable to forecast (positive output values in the interval $[0,1]$ and negative output values in the interval $[-1,0)$ are divided into two new classes. For each one of the obtained classes (four classes), neural networks are trained and the value of the weights is observed. If in these new classes obtained, the values of weights that are fixed after the training process, is the same that the one obtained in the previous division, or is proportional, then go back to the previous division. If the value is not the same then this division is valid, therefore they will exist four neural networks associated to the output intervals. This iterative division must go ahead until the weights of a new division will be the same of the previous division. When the weights are similar, then the successive divisions end. This process achieves a better error ratio, getting more powerful classification properties than classical nets, and this way a set of neural networks with their corresponding weights the following information:
    a. The variable with the most influence over the variable to forecast will be the one with the highest absolute weight after the training process. These data must verify that the sign of the input variable multiplied by the sign of the weight must be equal to the sign of the variable to forecast.
    b. And if the relationship between the forecasting variable and the variable to forecast is a direct or inverse function, that is, if the sign of both variables are the same or not. If the output interval of the variable to forecast, is a subinterval of interval $[0,1]$ or a subinterval of interval $[-1,0]$ and, if the domain of the forecasting variable multiplied by the corresponding weight is positive for a subinterval of the variable to forecast of interval $[0,1]$, we will say that the relationship is a direct one, other way it will be an inverse one, taking into account that the absolute value of the highest weight shows the importance of the forecasting variable over the variable forecast. That is, the higher absolute value of the variable over the variable to forecast.

       That is, the higher absolute value of the variable, the deeper influence in the output. Different divisions of initial set of training data, obtained from study of weights in the training subset, make that each one of the obtained training subset defines a different neural network to train the whole subset. Each network, with is corresponding set of weights denotes the importance of the forecasting variables over the variable to forecast.
    c. Besides extracting the importance of each variable in each output interval, for each one of the input variables it exits a network and a weight set that define the forecasting equation.

Therefore, the method is divided into two steps in order to better understand the two main processes on it.

- The first step is used to classify using the bisection method the patterns of the initial set into several subsets, taking into account that this division is performed iteratively, studying the variation of the weights. When in a new division the weights do not change, then go back to the initial division.

- The second step is used once the initial pattern set is classify into several subsets and therefore into several neural networks. The importance of each input variable must be studied for each different network, taking into account the weight values, the variation domain of the input variable and the variation of the output; to study the influence over the variable to forecast.

  It must be considered:

  1. The variables with the highest absolute weight.
  2. Which of them verify that their variation domain for the input variable multiplied by its corresponding weight has the same sign of the variable to forecast according to the positive or negative interval $[0,1]$ or $[-1,0]$.

## Experimental data

The previous theoretical results described have been used in the construction of a rule-oriented knowledge base, applied on a system to predict the load demand for the next day in a power plant [1].

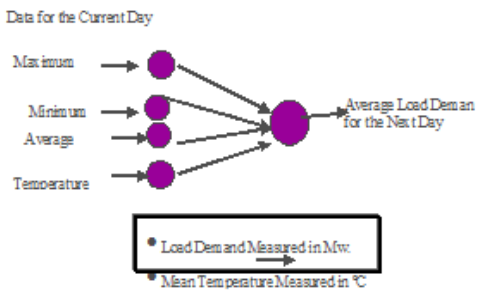
1. Obtaining the best classification: The proposed model takes into account the characteristics of forecasting variables could change from a different class to another, and that is the way it is necessary to use a division method, bisection method, studying the weights. This can be employed when dealing with a high number of patterns or to improve the error ratio.
2. Extract and study of the influence inputs variables: studying the weights decides which is the variable with more influence in the output using standardized weights and the bigger is the most important for the output.

### Example of application

The data used to design the training and test sets has been supplied by one of the most important spanish load suppliers on a specific format. That is featured by providing for each day the load demand data sampled for each hour measured in Mw., and the mean temperature of the day measured in Cº for two years. The input variables considered for the network were the maximum, minimum, average load demand and temperature for the current day. The output variable was the average load demand for the next day. The data was standardized in the range [-1,1]

There is a demand for making electrical charge per hour, taking data of 660 consecutive days. It also provides the average temperature each day. Taking 480 patterns for training and the rest for testing.

We have a number of input variables, which are defining the load curve for the next day. As we had shots of 24 hours a day and the average temperature of the day the variables used for forecasting were the maximum, minimum, average and the temperature of the previous day




### Obtaining the optimal classification

First level the output is ordered from lowest to highest. After the output range is standardized in [-1, 1], the output is divided in intervals by the middle of the range.

When the set is divided into different classes of patterns out of training improves, reaching a satisfactory ratio. At first you try to train the network with the entire set of patterns, to see what kind of predictions, and that knowledge was reflected in the weights, the data obtained were in table 1 and the ratio of Learning 0.2 is not good.

Table 1. Data weights

| Patterns | Error | Bias | Max | Min | Average | Temp | Output |
|---|---|---|---|---|---|---|---|
| All | 0.23969 | 0.5328 | 0.53414 | -0.2008 | 1.1656 | 0.1260 | [-1, 1] |

We try to train networks with different configurations, working with a hidden layer in which it was increasing the number of hidden neurons. But in any case learning improved, initially tested the whole set, with the values that are the table 1. The ratio of error should not be acceptable; the knowledge learned by the network is not good. The error is too large.

The bisection process begins by deciding the range of patterns that is obtained in each subclass and the values obtained for the weights associated with input variables after each division.

If the weights indicate the same importance for the variables, is no longer necessary to continue with the class divide. The network has found homogeneity in the patterns.

At this stage, the method of heuristic features, and was drawn to the rules. It is assumed that knowledge of the neural network must be stored in the weights.

The best classes were obtained testing with different division for classifications of the outputs.

The first branch was divided into two-out, or a class for the output, one class for positives outputs and other for negatives outputs. Obtaining two classes and then again divided in two new classes. For each one of the obtained classes (four classes) neural network is trained and the value of the weights is observed. If in this new obtained classes, the values of the weights are fixed after the training process is the same that the one obtained in the previous division, or is proportional then go back to the previous division.

If the values of the weights are similar or proportional we stop the division in classes, in this case, we obtained eight intervals or classes. It reached a suitable learning rate (average error 0.003) and is considered good to denormalize output.

**Final classification of all patterns**

Follows the evolution of weights in different classifications for all patterns.

The first division in positives and negatives outputs:

Table 2. Data weights with and without temperature

| Nº patterns | Bias | Max | Min | Average | Temp | Output |
|---|---|---|---|---|---|---|
| All | 0.3271 | 0.0958 | -0.6177 | 2.193 | 0.2442 | [-1, 1) |
| All | 0.4733 | 0.2993 | -0.2206 | 1.5576 | | [-1, 1] |
| positives | 0.2774 | 0.2255 | -0.5021 | 2.1254 | 0.2034 | [0, 1] |
| positives | 0.3888 | 0.3042 | -0.1897 | 1.6928 | | [0, 1] |
| negatives | -0.2542 | -0.5023 | -0.3518 | 0.7689 | 0.1996 | [-1, 0) |
| negatives | -0.1865 | -0.3413 | -0.1020 | 0.3613 | | [-1, 0) |

Study of the weight with different classes

Five networks trained for 459 patterns with the usual configuration

Table 3. Data weights for division in five output classes

| Patterns | Bias | Max | Min | Average | Temp | Output |
|---|---|---|---|---|---|---|
| 90 | -0.61717 | -0.4824 | -0.3528 | 0.4055 | 0.3523 | [-1, -0.13) |
| 87 | -0.1490 | -0.3317 | -0.1123 | 0.5451 | 0.0134 | [-0.13, 0) |
| 93 | 0.1726 | -0.1849 | -0.0986 | 0.4709 | 0.0365 | [0 , 0.2) |
| 114 | 0.4687 | 0.3752 | -0.1565 | 0.8179 | 0.1631 | [0.2, 0.5) |
| 75 | 0.7894 | 1.8697 | 0.0296 | -0.0226 | 0.1625 | [0.5 , 1] |

The error is less when you divide the total pattern set in subsets and is trained one RNA for each subset of pattern. In this example, finally we need construct 8 RNA: one for each Set of patterns S1, S3, … , S8 obtained, which outputs are I1,I3, … , I8, the subsets are obtained from de output division. One neural network is trained for each interval and different rule with the most important variable are obtained for each output interval, and one collection of rules R1, R3, …, R8 in the last step of the algorithm is obtained.

Table 4. Data weights of neural network training

| | Nº pattern | Interval | Bias | Max | Min | Average | Temp |
|---|---|---|---|---|---|---|---|
| (-) I1 | 19 | [-1 , - 0.5] | -2.0023 | -1.1458 | 0.1405 | 0.3903 | 0.7733 |
| (-) I2 | 55 | [-0.48 ,-0.31] | -0.7072 | 0.4279 | 0.1467 | -0.5488 | 0.014 |
| (-) I3 | 84 | [-0.30 , -0.2] | -0.4491 | -0.1087 | 0.0111 | 0.2206 | -0.0224 |
| (-) I4 | 116 | [-0.19 , 0] | -0.0967 | 0.1333 | -0.0437 | 0.1915 | 0.0135 |
| (+) I1 | 90 | [0 , 0.19] | 0.2021 | 0.3023 | -0.052 | 0.3942 | 0.0914 |
| (+) I2 | 58 | [0.2 , 0.35] | 0.5569 | 0.2408 | -0.0452 | 0.0681 | 0.0437 |
| (+) I3 | 40 | [0.35 , 0.59] | 0.6616 | 1.7795 | 0.0926 | -0.6939 | 0.1453 |
| (+) I4 | 18 | [0.62 , 1] | 0.0274 | -3.8252 | -1.7025 | 7.8928 | 0.403 |

The error is better than the first time with all patterns.

Table 5. Mean squared error of the trained ANN

| RNA i | Mean squared error |
|---|---|
| 1 | 0.006 |
| 3 | 0.042 |
| 4 | 0.049 |
| 5 | 0.038 |
| 6 | 0.043 |
| 7 | 0.042 |
| 8 | 0.04 |

We were looking for two things: a good learning and extracting a good knowledge in each class, it is, extract the most important input variables for each output interval.

- That knowledge stored by the network is reflected in the weights.
- Find the most important variables from the values of the weights.
- The rules that are obtained reflect what the network learned.

**Extracting rules for each intervals**

Once patterns have been divided in classes, we get eight subnets that give the best possible rating. We study the weights obtained from the network is trained and a characterization of the weights.

What is attempted is the order of importance of the variables and the degree of importance. The order and degree of importance of the variables is given by the value of the defined set of weights associated with the network.

The higher of the normalized weight, give the greater importance of the primary or principal input variable.

Table 6. Variables that can take part in the rules for each class of output

| Interval | Max | Min | Average | Temp | Output |
|---|---|---|---|---|---|
| I1 (-) | -1.145 | 0.14 | 0.39 | 0.773 | [-1 , -0.5] |
| I2 (-) | 0.4279 | 0.1467 | -0.5488 | 0.014 | [-0.48 , -0.31] |
| I3 (-) | -0.1087 | 0.0111 | 0.2206 | -0.0224 | [-0.3 ,-0.2] |
| I4 (-) | 0.1333 | -0.0437 | 0.1915 | 0.0135 | [-0.19 , 0] |
| I1 (+) | 0.3023 | -0.052 | 0.3942 | 0.0914 | [0 , 0.19] |
| I2 (+) | 0.2408 | -0.0452 | 0.0681 | 0.0437 | [0.2 , 0.35] |
| I3 (+) | 1.7795 | 0.0926 | -0.6939 | 0.1453 | [0.35 , 0.59] |
| I4 (+) | -3.8252 | -1.7025 | 7.8928 | 0.403 | [0.62 , 1] |

As shown in the table that follows, with the values obtained from the different networks once trained.

The values of the averages are almost identical and the standard deviations are not significant. What we succeeded in demonstrating that learning is good for every class.

Table 7. Pattern output and learning output

| | Nº pattern | Pattern output | Learning output | Average learning | Average output |
|---|---|---|---|---|---|
| I1 (-) | 19 | [-1 , -0.5] | [-0.79,-0.53] | -0.65 | -0.66 |
| I2 (-) | 55 | [-0.49,0.3] | [-0.38, 0.34] | -0.36 | -0.38 |
| I3 (-) | 84 | [-0.3 ,0.2] | [-0.27,-0.21] | -0.23 | -0.25 |
| I4 (-) | 117 | [-0.19, 0] | [-0.13,-0.01] | -0.07 | -0.11 |
| I1 (+) | 90 | [0 , 0.19] | [0.02 , 0.18] | 0.1 | 0.08 |
| I2 (+) | 58 | [0.2 ,0.35] | [0.27 , 0.33] | 0.29 | 0.27 |
| I3 (+) | 40 | [0.35,0.59] | [0.32 , 0.57] | 0.43 | 0.43 |
| I4 (+) | 18 | [0.62, 1] | [0.58 , 0.91] | 0.77 | 0.77 |

## Conclusion

In the algorithm proposed to extract knowledge from a neural network that has been trained, it is improved the learning of the RNA with a division of the output range while the weights are changing. In this way, we obtained the best division for getting the most important variables in the possible rule. It allows both antecedent (the most important variable in each interval together with the domain values for this variable) and consequent (the interval for the output obtained with iterative Method previously described in this article) obtain rules to take continuous values, and make them able to be applied to a greater number of cases.

In this way, the rules obtained will allow to complete the knowledge that could be extracted from an expert when building the knowledge base for an expert system. In the proposed method, the first task is to divide the problem in output ranges; then the most important variables are extracted from each interval, and finally the solution (set of rules) is globalize with all the output intervals. The proposed method also computes the forecasting value from the equation of weights.

The proposed model takes into account the fact that the characteristics of forecasting variables could change from a different class to another, and because of that it is necessary to use a division method or a bisection method. This can be used when dealing with a high number of patterns or to improve the error ratio.

The main advantage of this method is the simplicity of itself. The matrix of weight defines the most important forecasting variables as well as the equation to return a value. The only thing to do is to apply the bisection method to the data set and to train a neural network for each class identified by the algorithm.

## Bibliography

[Andrews, R., Diederich, J., Tickle,A. 1995] Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowsledge-Based Systems (1995)

[Andrews, R., Diederich, J., Golea, M. 1998] The truth will come to light directions and challenges in extracting the knowledge embedded within trained artificial neural networks. IEEE Trans. Neural Networks(1998).

[Apolloni, B. et al 2004] A general framework for learning rules from data," IEEE Trans. Neural Networks., vol. 15, no. 6, pp. 1333–1349, Nov. 2004.

[Chang, B.L., Hirsch, M. 1991] Knowledge Acquisition and Knowledge Representation in a Rule-Based Expert Systems. Computers in Nuring. Volume 9, Number5 Pp 174-178 (1991)

[Cloete, I., Zurada, J.M. 2000] Knowledge- Based Neurocomputing. MIT Press (2000).

[Freeman J.A., Skapura D.M. 1992] Neural Networks. Addison-Wesley, Reading.

[Garcez d'Avila, A. S., Broda, K. and Gabbay D. M. 2001] Symbolic knowledge from trained neural networks: A sound approach, Artif. Intell., vol. 125, no. 1, pp. 155–207, 2001.

[Krishnan R., Sivakumar G., Bhattacharya P. 1999] A search technique for rule extraction from trained neural networks. Patern Recognit Lett 20:273-280 (1999).

## Authors' Information

***Castellanos Angel*** *– Departamento de Ciencias Basicas aplicadas a la Ingeniería Forestal. Escuela de Ingeniería Técnica Forestal. Universidad Politécnica de Madrid, Avda. de Ramiro de Maeztu s/n 28040 Madrid, Spain. e-mail: angel.castellanos@upm.es*

***Gonzalo Rafael*** *– Natural Computing Group. Universidad Politécnica de Madrid, Spain. e-mail: rgonzalo@fi.upm.es*

***Martinez Ana*** *– Natural Computing Group. Universidad Politécnica de Madrid, Spain. e-mail: ana.martinez@upm.es*

# STRING MEASURE APPLIED TO STRING SELF-ORGANIZING MAPS AND NETWORKS OF EVOLUTIONARY PROCESSORS[1]

**Nuria Gómez Blas, Luis F. de Mingo, Francisco Gisbert, Juan M. Garitagoitia**

***Abstract:*** *This paper shows some ideas about how to incorporate a string learning stage in self-organizing algorithms. T. Kohonen and P. Somervuo have shown that self-organizing maps (SOM) are not restricted to numerical data. This paper proposes a symbolic measure that is used to implement a string self-organizing map based on SOM algorithm. Such measure between two strings is a new string. Computation over strings is performed using a priority relationship among symbols; in this case, symbolic measure is able to generate new symbols. A complementary operation is defined in order to apply such measure to DNA strands. Finally, an algorithm is proposed in order to be able to implement a string self-organizing map.*

***Keywords:*** *Neural Network, Self-organizing Maps, and Control Feedback Methods.*

***ACM Classification Keywords:*** *F.1.1 Models of Computation: Self-modifying machines (neural networks); F.1.2 Modes of Computation: Alternation and non-determinism.*

## Introduction

Most well known numeric models are Neural Networks that are able to approximate any function or classify any pattern set provided numeric information is injected into the net. Neural Nets usually have a supervised or unsupervised learning stage in order to perform desired response. Concerning symbolic information new research area has been developed, inspired by George Paun, called Membrane Systems. A step forward, in a similar Neural Network architecture, was done to obtain Networks of Evolutionary Processors (NEP), introduced by Victor Mitrana. A NEP is a set of processors connected by a graph, each processor only deals with symbolic information using rules. In short, objects in processors can evolve and pass through processors until a stable configuration is reach.

Self-Organizing maps are usually used for mapping complex, multidimensional numerical data onto a geometrical structure of lower dimensionality, like a rectangular or hexagonal two-dimensional lattice [2, 3]. The mappings are useful for visualization of data, since they reflect the similarities and vector distribution of the data in the input space. Each node in the map has a reference vector assigned to it. Its value is a weighted average of all the input vectors that are similar to it and to the reference vectors of the nodes from its topological neighbourhood. For numerical data, average and similarity are easily computed: for the average, one usually takes the arithmetical mean, and the similarity between two vectors can be defined as their inverse distance, which is most often the Euclidian one. However, for non-numerical data [4]– like symbol strings – both measures tend to be much more complicated to compute. Still, like their numerical counterparts, they rely on a distance measure. For symbol strings one can use the Levenshtein distance or feature distance.

For strings, one such measure is the Levenshtein distance [1], also known as edit distance, which is the minimum number of basic edit operations – insertions, deletions and replacements of a symbol – needed to transform one string into another. Edit operations can be given different costs, depending on the operation and the symbols involved. Such weighted Levenshtein distance can, depending on the chosen weighting, cease to be distance in the above sense of the word.

1 Supported by projects CCG08-UAM TIC-4425-2009 and TEC2007-68065-C03-02

Another measure for quantifying how much two strings differ is feature distance [2]. Each string is assigned a collection of its substrings of a fixed length. The substrings the features are typically two or three symbols long. The feature distance is then the number of features in which two strings differ. It should be noted that this measure is not really a distance, for different strings can have a zero distance. Nevertheless, feature distance has a practical advantage over the Levenshtein by being much easier to compute.

A similarity measure is simpler than distance. Any function $S : X^2 \rightarrow R$ can be declared similarity – the question is only if it reflects the natural relationship between data. In practice, such functions are often symmetrical and assign a higher value to two identical elements than to distinct ones, but this is not required.

## String Measure

Let *V* an alphabet over a set of symbols. A string *x* of length *m* belonging to an alphabet *V* is the sequence of symbols $a_1a_2\ldots a_m$ where the symbol $a_i \in V$ for all $1 \leq i \leq m$. The set of all strings over *V* is denoted by $V^*$, the empty symbol is $\lambda$ and the empty string is denoted by $\varepsilon = (\lambda)^*$.

Let $O{:}x \rightarrow n,\ x \in V,\ n \in N$ a mapping that establish a priority relationship among symbols belonging to *V* , $u \leq v$ iff $O(u) \leq O(v)$. Obviously $O(O(x)) = x,\ x \in V$ and $O(O^{-1}(n)) = n,\ n \in N$ , and $O(\lambda) = 0,\ O^{-1}(0) = \lambda$. This mapping can be extended over an string *w* in such a way that $O(w) = O(w_i),\ w_i \in w$. Usually, such mapping *O* covers a range of integer numbers, that is, the output is $0 \leq i \leq k$, where $k = card(S),\ S \subseteq V$ . It is important to note that new symbols can be generated provided that given two symbols $a, b \in V\ |O(a) - O(b)| > 1$, and there is no symbol *c* such that $O(a) < O(c) < O(b)$. That is,

$$\mathcal{O}^{-1}(k) = \begin{cases} x \in V & \text{iff } \mathcal{O}(x) = k \\ s_k & \text{i.o.c.} \end{cases}, \text{ with } k \in \mathcal{N}$$

Symbolic measure between two strings $u, v \in V^*$, denoted by $\Delta(u, v)$, with $|u| = |v| = n$ is another string defined as:

$$\Delta(u,v) = \bigcup_{i=1}^{n} \mathcal{O}^{-1}(|\mathcal{O}(u_i) - \mathcal{O}(v_i)|),\ \text{where } u_i/v_i \text{ is the } i\text{-th symbol} \in u/v \quad (1)$$

For example, let $u = (abcad),\ v = (abdac)$, and *O* the index of such symbol in the Latin alphabet, that is, $O(a) = 1$, $O(b) = 2,\ O(c) = 3,\ O(d) = 4$ then $\Delta(u, v) = \lambda\lambda a\lambda a$. If $u = (jonh),\ v = (mary)$ then $\Delta(u, v) = s_3njs_{11}$, two new symbols $s_3$, $s_{11}$ are generated (that correspond to $s_3 = c$ and $s_{11} = k$, usually such correspondence is unknown). A numeric value *D* can be define over a string *w*:

$$\mathcal{D}(w) = \sqrt{\sum_{i=0}^{|w|} \mathcal{O}(w_i)^2}, w_i \in w \quad (2)$$

It is clear to proof that:

$D(\Delta(u, v)) = D(\Delta(v, u)),\ D(\Delta(u, u)) = 0,\ D(\Delta(u, \varepsilon)) = D(u)$ and $D(\Delta(u, w)) \leq D(\Delta(u, v)) + D(\Delta(v, w))$.

Mappings *O/D* also define a priority relationship among string in $V^*$ is such a way that

$$u \leq v \text{ iff } \sqrt{\sum_{i=1}^{n=|u|} \mathcal{O}(u_i)^2} \leq \sqrt{\sum_{i=1}^{n=|v|} \mathcal{O}(v_i)^2}$$

$$u \leq v \text{ iff } \mathcal{D}(u) \leq \mathcal{D}(v)$$

In short, symbolic measure between two string *u, v* is obtained using $\Delta(u, v)$, see equation (2), and numeric measure is obtained using $D(\Delta(u, v))$, see equation (1). Let $x, y \in S \subseteq V$ two symbols belonging to alphabet, two symbols are complementary, denoted by $(x, y)^{-}$, iff $\Delta(x, y) = x$ or $\Delta(x, y) = y$. Such property can be extended over string, let $u, v \in S^* \subseteq V^*$, two strings are complementary, denoted by $(u, v)^{-}$, iff $\Delta(u, v) = u$ or $\Delta(u, v) = v$.

**Theorem 1**. - Let $u, v \in S^*$, $u$ and $\Delta(u, v)$ are complementary iff $O(ui) >= O(vi)$ for all $1 \leq i \leq n$.

*Proof*.

$$\Delta(u, v) = \bigcup_{i=1}^{n} \mathcal{O}^{-1}(|\mathcal{O}(u_i) - \mathcal{O}(v_i)|)$$

Hence:

$$\Delta(u, \Delta(u, v)) = \bigcup_{i=1}^{n} \mathcal{O}^{-1}(|\mathcal{O}(u_i) - \mathcal{O}(\Delta(u_i, v_i))|) =$$

$$= \bigcup_{i=1}^{n} \mathcal{O}^{-1}(|\mathcal{O}(u_i) - \mathcal{O}(\mathcal{O}^{-1}(|\mathcal{O}(u_i) - \mathcal{O}(v_i)|)|) =$$

$$= \bigcup_{i=1}^{n} \mathcal{O}^{-1}(|\mathcal{O}(u_i) - (|\mathcal{O}(u_i) - \mathcal{O}(v_i)|)|) =$$

$$= \bigcup_{i=1}^{n} \mathcal{O}^{-1}(\mathcal{O}(v_i)) = v$$

□

Two strings $u, v \in S^*$ are Watson-Crick complementary (WC complementary), denoted by $(u, v)^{-WC}$, iff $(u, v)^{-}$ for all $1 \leq i \leq |u|$.

**Theorem 2.** - Let $u, v \in S^*$, if $(u, v)^{-}$ then $(u, v)^{-WC}$.

Such duality in symbolic/numeric measures, see equations (1 and 2), is a good mechanism in order to implement algorithms on biological DNA strands [5, 6]. Like DNA or amino acid sequences, which are often, subject to research in computational molecular biology. There, a different measure – similarity – is usually used. It takes into account mutability of symbols, which is determined through complex observations on many biologically close sequences. To process such sequences with neural networks, it is preferable to use a measure, which is well empirically founded.

**Strings with different lengths**

Given two strings *u, v*, such that $|u| = n \geq |v| = m$, and $U(u)$ the set of all substring $w \subseteq u$ such that,

$$U(u)^m = \{w^{(j)} || w^{(j)}| = m, w = w_1 \cdots w_m, w_i = u_k, i = k + j\} \forall\, 0 \leq j \leq |u| - m$$

String measure between *u, v*, denoted by $\delta(u, v)$, is

$$\delta(u, v) = \{\Delta(s, v) | s \in U(u)^{|v|}, \mathcal{O}(\Delta(s, v)) \leq min_{x \in U(u)^{|v|}} \{\mathcal{O}(\Delta(x, v))\}\}$$

In this case, measure $\delta$ is a set of strings with the lower distance (see table below). Such distance can be read as the set of matching strings with lower distance. This $\delta$ can be used to identify cutting points (index $j$) over a DNA string when applying a restriction enzyme, from a biological point of view.

$u = abcdabcdab, v = cda$

| $U(u)^{\lvert v\rvert}$ | $u$ |
|---|---|
| a b c<br>b c d<br>c d a<br>d a b<br>a b c<br>b c d<br>c d a<br>d a b | $\mathcal{O}(\Delta(cda, v)) = 0$<br>$\mathcal{O}(\Delta(cda, v)) = 0$ |

$\delta(u, v) = \{\lambda\lambda\lambda, \lambda\lambda\lambda\}$

Let $|u| = |v|$, it is clear that $\delta(u, v) = \Delta(u, v)$ since $U(u) = u$.

## Future Work

Some results, in literature, that could be checked with this new measure can be: for an example application of the string SOM, Igor Fisher generated a set of 500 strings by introducing noise to 8 English words: always, certainly, deepest, excited, meaning, remains, safety, and touch, and initialized a quadratic map with the Sammon projection of a random sample from the set [1]. Another real world example is the mapping produced from 320 hemoglobin alpha and beta chain sequences of different species [2]. SOM and LVQ algorithms for symbol strings have been introduced by [5, 6] and applied to isolated word recognition, for the construction of an optimal pronunciation dictionary for a given speech recognizer.

Artificial Neural Networks (ANN) and Networks of Evolutionary Processors (NEP) [9, 10] can be considered as the present and the future of connectionist models. Both of them are based on the idea of simple processors that communicate in order to achieve a global objective. But there are two important facts that must be taken into account:

- ANN are numeric models while NEP are symbolic ones.
- There exists a learning algorithm that control the ANN behavior in order to achieve a desired result while NEP do not incorporate any kind of learning paradigm.

Some ideas of ANN can be translated into NEP architecture since ANNs are considered, in the literature, a good model to solve non-conventional problems. Following this point of view some kind of learning can be added to a NEP to obtain a more general model than simple NEP. Among all the neural networks architectures unsupervised neural networks, cal led Self Organizing Maps (SOM), are the most suitable.

## Conclusions

In some applications, like molecular biology, a similarity measure is more natural than distance and is preferred in comparing protein sequences. It is possible that self-organizing neural networks can successfully process such data. It can therefore be concluded that similarity-based neural networks are a promising tool for processing and analyzing non-metric data. This paper has proposed a string measure that can be applied to self-organizing maps or networks of evolutionary processors with the possibility of new symbols generation. Watson-Crick complementary concept was defined using such measure.

## Acknowledgements

This work is supported by projects CCG08-UAM TIC-4425-2009 and TEC2007-68065-C03-02.

## Bibliography

[1] LEVENSHTEIN L.I, Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics–Doklady 10, (1966) 707–710.

[2] TEUVO KOHONEN, Self-Organization and Associative Memory. Springer, Berlin Heidelberg, (1988).

[3] TEUVO KOHONEN, SOMERVUO P, Self-Organizing Maps of Symbol Strings with Application to Speech Recognition, (1997).

[4] TEUVO KOHONEN, SOMERVUO P, Self-organizing maps of symbol strings, Neurocomputing 21 (1998) 19–30.

[5] MARIA SANCHEZ, NURIA GOMEZ, LUIS MINGO, DNA Simulation of Genetic Algorithms: Fitness Function, International Journal on Information Theories and Applications, 14 (3). ISSN 1310-0513 (2007) 211–217.

[6] NURIA GOMEZ, EUGENIO SANTOS, MIGUEL ANGEL DIAZ, Symbolic Learning (Clustering) over DNA Strings, WSEAS Transactions on Information Science and Applications. 3 (4), ISSN: 1709-0832 (2007) 617–624.

[7] IGOR FISCHER, ANDREAS ZELL, String averages and self-organizing maps for strings, Proceeding of the ICSC Symposia on Neural Computation (NC'2000) May 23-26, 2000 in Berlin, Germany, (2000), 208–215.

[8] IGOR FISCHER, Similarity-based neural networks for applications in computational molecular biology, Lecture notes in computer science, 2779, ISSN 0302-9743, (2003) 208–218.

[9] JUAN CASTELLANOS, FLORIN MANEA, LUIS F. MINGO, VICTOR MITRANA, Accepting Networks of Splicing Processors with Filtered Connections, MCU (2007) 218–229.

[10] FLORIN MANEA, VICTOR MITRANA, Al l NP-problems can be solved in polynomial time by accepting hybrid networks of evolutionary processors of constant size, Inf. Process. Lett. 103(3), (2007), 112–118.

## Authors' Information

***Nuria Gómez Blas –*** *Dept. Organización y Estructura de la Información, Escuela Universitaria de Informática, Universidad Politécnica de Madrid, Crta. De Valencia km. 7, 28031 Madrid, Spain; e-mail: ngomez@eui.upm.es*

***Luis Fernando de Mingo –*** *Dept. Organización y Estructura de la Información, Escuela Universitaria de Informática, Universidad Politécnica de Madrid, Crta. De Valencia km. 7, 28031 Madrid, Spain; e-mail: lfmingo@eui.upm.es*

***Francisco Gisbert –*** *Dept. Lenguajes, Sistemas Informáticos e Ingeniería del Software, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, 28660 Madrid, Spain; e-mail: fgisbert@fi.upm.es*

***Juan M. Garitagoitia –*** *Dept. Organización y Estructura de la Información, Escuela Universitaria de Informática, Universidad Politécnica de Madrid, Crta. De Valencia km. 7, 28031 Madrid, Spain; e-mail: jmgmartin@eui.upm.es*

# МНОГОКРИТЕРИАЛЬНАЯ ОПТИМИЗАЦИЯ АРХИТЕКТУРЫ НЕЙРОСЕТЕВЫХ КЛАССИФИКАТОРОВ

**Альберт Воронин, Юрий Зиатдинов, Анна Антонюк**

***Аннотация.*** *Рассматривается постановка задачи и процедура векторной оптимизации архитектуры нейросетевого классификатора. В качестве целевой функции предложена скалярная свертка критериев по нелинейной схеме компромиссов. Используются поисковые методы оптимизации с дискретными аргументами. Приведен пример – нейросетевой классификатор текстов.*

***Ключевые слова****: многокритериальная оптимизация, нейронные сети, классификатор.*

***ACM Classification Keywords****: H.1 Models and Principles – H.1.1 – Systems and Information Theory; H.4.2 – Types of Systems; C.1.3 Other Architecture Styles – Neural nets*

***Conference****: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

## Содержание проблемы

Важной разновидностью искусственных нейронных сетей являются нейросетевые классификаторы. Они применяются для технической и медицинской диагностики, классификации различного рода информационных источников и пр. В достаточно общем случае структура $q$-слойного нейросетевого классификатора с прямыми связями представлена на Рис.1.

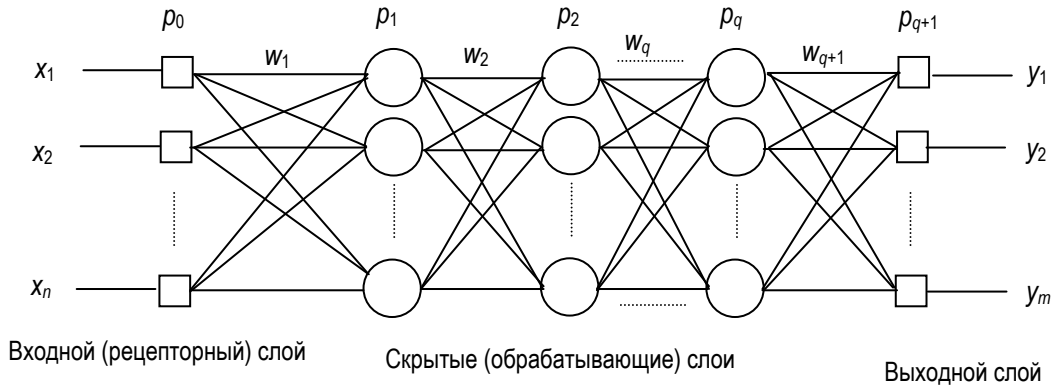



Рис.1

Здесь $x_1, x_2, ..., x_n$ – признаки объекта классификации, составляющие входной вектор $x = \{x_i\}_{i=1}^{n}$; $p_0 = n$ – число нейронных элементов в рецепторном слое; $p_1, p_2, ..., p_q$ – число нейронов в каждом из скрытых слоев; $p_{q+1} = m$ – число нейронов в выходном слое (количество классов);

$y = \{y_k\}_{k=1}^{m}$ – выходной вектор нейронной сети, определяющий принадлежность объекта классификации одному из $m$ классов; $w_1, w_2, ..., w_q, w_{q+1}$ – векторы синаптических весов нейронной сети.

Приведем необходимые сведения из теории нейронных сетей [1-3]. Искусственная нейронная сеть – это совокупность нейронных элементов и связей между ними. Каждый нейрон имеет группу синапсов – однонаправленных входных связей, соединенных с выходами других нейронов. Каждый синапс характеризуется величиной синаптической связи или ее весом $w_i$ (определяется при обучении нейронной сети). Нейрон имеет текущее состояние, определяемое как взвешенная сумма его входов: $s = \sum_{i=1}^{n} w_i x_i$.

Выход нейрона есть функция его состояния, которая называется функцией активации: $y = f(s)$. Сигнал возбуждения или торможения посредством аксона (выходная связь данного нейрона) поступает на синапсы следующих нейронов. Функции активации бывают пороговыми и непрерывными (биполярный сигмоид, гауссиан и пр). Множество всех нейронов искусственной нейронной сети разделяется на подмножества, называемые слоями. Слой – это множество нейронов, на которые в каждый такт времени параллельно поступают сигналы от других нейронов данной сети [2]. На выходе классификатора получается вектор функций активации $y = \{y_k\}_{k=1}^{m}$. Номер $j$, для которого выход $y_j$ имеет максимальную активность, т.е. $\max\limits_{k \in [1,m]} y_k = y_j$, соответствует номеру класса объекта классификации.

Количество нейронов входного слоя $p_0 = n$ определяется размерностью входного вектора признаков и не подлежит изменениям. Аналогично, количество нейронов выходного слоя $p_{q+1} = m$ определяется числом областей (классов), на которые делится пространство признаков и тоже является постоянным. Количество же обрабатывающих (скрытых) слоев $q$ и число нейронов в каждом из них $p_1, p_2, ..., p_q$ составляют понятие **архитектуры** [1] нейронной сети и могут служить аргументами (независимыми переменными) при ее оптимизации.

В настоящей работе ограничим исследование случаем, когда число $q$ является фиксированным и заданным. Тогда аргументами оптимизации архитектуры нейронного классификатора являются количества нейронов в каждом из обрабатывающих слоев, составляющие вектор независимых переменных $p = \{p_j\}_{j=1}^{q}$. От выбора архитектуры $p$ зависит качество функционирования нейронного классификатора.

Проблема заключается в таком выборе архитектуры, при котором нейронный классификатор в заданных условиях функционирования характеризуется наилучшими свойствами.

## Постановка задачи

В общем виде проблема может быть формально представлена задачей

$$p^* = \arg \underset{p \in P}{extr}\, Y(p), \tag{1}$$

где $Y(p)$ – целевая функция; $\underset{p \in P}{extr}$ – оператор экстремизации целевой функции по аргументам $p$; $P$ – допустимая область независимых переменных.

Для конструктивного решения задачи сделаем дополнительные частные предположения. Каждому свойству нейронного классификатора поставим в соответствие количественную характеристику $f(p)$,

имеющую смысл критерия качества его функционирования. Одним из таких критериев является вероятность ошибки классификации. Будем определять этот критерий экспериментально и приближенно представим его как количество ошибок классификации $e(p)$, отнесенное к общему, достаточно большому количеству испытаний $N$:

$$f_1(p) = \frac{e(p)}{N}. \quad (2)$$

Предполагается, что с ростом в некоторых разумных пределах числа нейронов в обрабатывающих слоях точность классификации повышается, и величина этого критерия уменьшается. Предельно допустимое значение ошибки сети должно быть известно из физических соображений и задано как ограничение $f_1(p) \le A_1$.

Второй критерий характеризует время, необходимое для обучения нейронной сети при данной архитектуре $p$. Наблюдается тесная корреляция между таким временем и суммарным количеством нейронов в скрытых слоях классификатора. Поэтому представим этот критерий в виде

$$f_2(p) = \sum_{k=1}^{q} p_k \ . \quad (3)$$

Отметим, что данным критерием характеризуется и время прохождения сигнала через нейронную сеть от входа к выходу. С ростом числа нейронов значение критерия увеличивается. Предельно допустимое значение второго критерия определяется допустимым временем обучения нейронной сети и задается как ограничение $f_2(p) \le A_2$.

Существуют и другие критерии для характеристики различных свойств нейронного классификатора. В данной работе мы ограничимся только приведенными двумя основными критериями, имея в виду, что излагаемая методика допускает включение в рассмотрение и других свойств классификатора.

Допустимая область аргументов оптимизации задается параллелепипедным ограничением $P = \{p \,|\, 0 \le p_k \le P_u, k \in [1, P_u], u \in [1, q]\}$, где $P_u$ – максимальное число нейронов в $u$-м слое.

Поскольку оба включенных в рассмотрение критерия подлежат минимизации (чем критерий меньше, тем лучше соответствующее свойство классификатора), то оператор экстремизации целевой функции приобретает вид: $\underset{p \in P}{extr} = \min\limits_{p \in P}$.

Таким образом, оба критерия являются противоречивыми, неотрицательными, минимизируемыми и ограниченными. Налицо все предпосылки для использования в качестве целевой функции скалярной свертки критериев по нелинейной схеме компромиссов [4]. Такая свертка в унифицированной версии выражается формулой

$$Y(p) = Y[f(p)] = \frac{A_1}{A_1 - f_1(p)} + \frac{A_2}{A_2 - f_2(p)}. \quad (4)$$

где $f(p) = \{f_r(p)\}_{r=1}^{r=2}$ – двумерный вектор частных критериев. Учитывая (2), (3) и (4), выражение (1) для задачи оптимизации архитектуры нейронного классификатора преобразуется к виду

$$p^* = \arg\min_{p \in P} \left[ \frac{A_1}{A_1 - e(p)/N} + \frac{A_2}{A_2 - \sum_{k=1}^{q} p_k} \right]. \quad (5)$$

Нетрудно видеть, что в формуле (5) зависимость $e(p)$ априори является неизвестной и подлежащей экспериментальному определению.

## Метод решения

Среди задач многокритериальной оптимизации имеются такие, аргументы которых по своей физической природе могут принимать только дискретные значения. Специальной нормировкой дискретные значения обычно всегда могут быть сведены к целочисленным. Такие задачи значительно сложнее непрерывных многокритериальных задач и для их решения должны применяться иные подходы [5].

Множество допустимых дискретных значений может быть бесконечным, конечным или даже состоящим всего из двух значений, например, 0 и 1. В первом случае задача вырождается в непрерывную задачу оптимизации. Для ее решения в [4] предложено эффективное и формализованное алгоритмическое и программное обеспечение. В последнем случае имеет место целочисленное программирование с булевыми переменными со своими специфическими методами (логический синтез конечных автоматов, функции Рвачева и пр.). С нашей точки зрения наиболее интересен и содержателен случай, когда множество допустимых дискретных значений не настолько велико, чтобы задача вырождалась в непрерывную, но и не настолько мало, чтобы ее можно было решить простым перебором. Именно такой является задача (5) – задача нелинейного дискретного (целочисленного) программирования.

Методы дискретного программирования не обладают таким единством, как методы вариационного исчисления, и в большинстве представляют собой набор частных приемов, пригодных для решения частных задач. Но их актуальность требует их развития и совершенствования, т.к. наиболее важные прикладные задачи сводятся, как правило, к задачам частично или полностью дискретного программирования. Сложность решения задач дискретного (целочисленного) программирования возрастает в том случае, когда задача является многокритериальной.

В том случае, когда компоненты возможных решений многокритериальных задач могут принимать только дискретные значения $p_k^{(P_u)}, k \in [1, P_u], u \in [1, q]$, скалярная свертка по нелинейной схеме компромиссов $Y(p)$ является *решетчатой* функцией, заданной на дискретном множестве $P$. Оптимизация решетчатой целевой функции, построенной по нелинейной схеме компромиссов, сводится к задаче нелинейного программирования с дискретными (целочисленными) аргументами, решение которой, как отмечено выше, достаточно сложно.

Для решения этой проблемы мы предполагаем, что при дискретном множестве $P$ существует вспомогательная область непрерывных аргументов $p_c \in P_c$, которая содержит все дискретные точки $p_k^{(P_u)}$ и всё непрерывное пространство между ними. В области $P_c$ определена непрерывная функция $Y(p_c)$, которая в точках $p_k^{(P_u)}$ совпадает с решетчатой функцией $Y(p)$.

Это предположение позволяет получить аналитическое решение, если в выражении (5) зависимость $e(p)$ задана, например, в виде регрессионной модели. Тогда можно воспользоваться необходимым

условием минимума функции: $\frac{\partial Y(p_c)}{\partial p_c} = 0$. Решение этой системы уравнений дает компромиссно-оптимальную непрерывную точку $p_c*$. Последний шаг алгоритма – поиск на $P$ *ближайшей* к $p_c*$ возможной *дискретной* точки, которая и будет искомым дискретным решением $p*$. В нашем случае, к сожалению, задание аналитических зависимостей весьма затруднительно или вообще невозможно.

Мы рассматриваем как основной случай, когда функции $e(p)$ и, следовательно $Y(p)$, неизвестны, но есть возможность определять значения функции $Y(p)$ в точках $p_k^{(P_u)}$ измерением или вычислением. Тогда можно организовать натурный или вычислительный эксперимент, в результате которого осуществляется поисковое движение к искомой дискретной компромиссно-оптимальной точке $p*$.

Возможны различные подходы к организации поисковой процедуры, которая должна давать последовательность улучшающихся решений. Один из них – дискретный аналог метода симплекс-планирования Нелдера-Мида (метод деформируемого многогранника) [4]. Это – разновидность градиентных методов, весьма часто и успешно применяющихся на практике. Второй – нелокальный (дуальный) подход [4], часто более эффективный, чем градиентные методы.

Так как в поисковых процедурах используются локальные или нелокальные модели непрерывной функции $Y(p_c)$, то общей для названных вариантов является необходимость поиска на $P$ возможной дискретной точки $p_d$, ближайшей к непрерывному решению $p_c$ на текущей или заключительной итерации. Если число скрытых слоев $q$ невелико, то решение этой задачи не вызывает затруднений (простое округление до целого). При многослойных классификаторах мы рекомендуем использовать следующий алгоритмический прием. В точке $p_c$ помещается центр гиперсферы, диаметр которой возрастает от нуля до тех пор, пока поверхность сферы не коснется ближайшей дискретной точки, которая тем самым идентифицируется как $p_d$. Возможны разные программные реализации этого алгоритма.

Известны нейросетевые классификаторы различного вида и назначения.

## Многокритериальная оптимизация нейросетевого классификатора текстов

В качестве примера рассмотрим в общих чертах задачу многокритериальной оптимизации архитектуры нейросетевого классификатора текстов. Система текстовой классификации [3] состоит из двух основных частей: частотный анализатор с системным словарем и собственно нейросетевой классификатор (Рис.2).


Рис.2

На вход системы поступает текст, на выходе получается номер темы, к которой относится этот текст (бизнес, политика, медицина, спорт, просто спам и т.п.).

Прежде чем приступить к оптимизации архитектуры нейросетевого классификатора, необходимо выполнить следующие этапы:

1. Определяются $m$ классов, с которыми будет работать система.

2. Подбираются соответствующие учебные тексты $t_k, k \in [1,m]$ и проверочные (тестовые) тексты $t_l, l \in [1,L], L \geq m$.

3. Из множества учебных текстов специальным образом выделяются слова $v_i, i \in [1,n]$ и формируется системный словарь $V$.

4. Частотный анализатор определяет для каждого слова $v_i$ из системного словаря $V$ его частоту вхождения $x_i$ в данный текст $t_k$. Частотная характеристика – это вектор $x = \{x_i\}_{i=1}^{n}$ признаков текста $t_k$, размерность которого равна количеству слов в системном словаре $v_i \in V$.

Получив результаты частотного анализа учебных текстов, можно приступать к обучению нейронного классификатора при некоторой архитектуре $p = \{p_j\}_{j=1}^{q}$. Процесс обучения нейронной сети заключается в установлении таких весовых коэффициентов ее связей $w_1, w_2, ..., w_q, w_{q+1}$, при которых максимальная ошибка сети на учебных текстах для данной архитектуры не превышает предельно допустимое значение. Конкретные алгоритмы обучения здесь не рассматриваются.

Теперь можно приступить непосредственно к процедуре векторной оптимизации. Для оптимизации архитектуры нейросетевого классификатора воспользуемся поисковым методом симплекс-планирования. Пусть для определенности число обрабатывающих слоев $q$ =2. Тогда идею метода в непрерывном варианте можно иллюстрировать при помощи Рис.3.

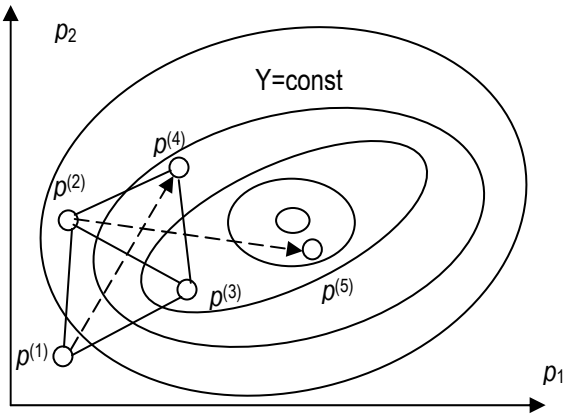



Рис.3

На плоскости аргументов $p_1 - p_2$ в некоторой стартовой области строим исходный регулярный симплекс, который в двумерном случае представляет собой равнобедренный треугольник с вершинами $p^{(1)}, p^{(2)}, p^{(3)}$. Для каждой из трех архитектур симплекса осуществляем процесс обучения классификатора и подаем на вход серию тестовых текстов $t_l$. В каждой вершине симплекса определяем количество ошибок классификации $e^{(1)}, e^{(2)}, e^{(3)}$ при общем количестве испытаний $N = L$. По формуле

(2) получаем критерии $f_1^{(1)}, f_1^{(2)}, f_1^{(3)}$. По формуле (3) определяются критерии $f_2^{(1)}, f_2^{(2)}, f_2^{(3)}$. Формула (4), выступающая в этом случае в роли не целевой, а оценочной функции, для нашего примера имеет вид

$$Y(p_1, p_2) = \frac{A_1}{A_1 - e(p_1, p_2)/L} + \frac{A_2}{A_2 - p_1 - p_2}. \quad (6)$$

Для архитектур стартового симплекса она дает значения скалярных сверток $Y^{(1)}, Y^{(2)}, Y^{(3)}$. Сравнивая между собой эти значения, находим, что одно из них, например, $Y^{(1)}$, оказалось больше (т.е. хуже), чем другие. С большой вероятностью можно утверждать, что архитектура $p^{(4)}$, полученная зеркальным отображением худшей в исходном симплексе точки $p^{(1)}$ относительно центра противоположной грани, окажется лучше. Осуществив все расчеты для архитектуры $p^{(4)}$, образуем новый симплекс с вершинами $p^{(2)}$, $p^{(3)}$ и $p^{(4)}$. Сравнив значения $Y^{(2)}, Y^{(3)}, Y^{(4)}$, обнаружим, что одна из точек, например, $p^{(2)}$, оказалась хуже других в смысле второго симплекса. Отобразив эту точку относительно центра противоположной грани второго симплекса, получим архитектуру $p^{(5)}$, и т.д. до тех пор, пока мы получим архитектуру $p*$, соответствующую минимуму целевой функции.

Это лишь иллюстрация идеи метода симплекс-планирования. На самом деле этот метод в модификации Нелдера-Мида предусматривает адаптацию симплексов к топографии целевой функции за счет деформации многогранников, он имеет хорошо разработанное алгоритмическое и программное обеспечение. Кроме того, нельзя забывать, что у нас имеет место случай оптимизации с целочисленными аргументами, что диктует необходимость для каждого полученного непрерывного решения $p_c$ искать ближайшее дискретное решение $p_d$.

Второй, нелокальный поисковый метод несколько сложнее в реализации, но он обычно более эффективен [4,5]. Метод основан на итерационном построении «плывущей» вместе с системой изменяющихся базисных точек уточняющейся по результатам эксперимента нелокальной модели $Y(p)$, причем совокупность опорных точек сжимается и стягивается к точке искомого экстремума («шагреневая кожа»). На каждой итерации одновременно и взаимозависимо осуществляется как уточнение наших представлений о целевой функции в области экстремума, так и определение такой оценки аргументов экстремума, которая адекватна уровню этих представлений на данной итерации. Поэтому нелокальный метод оптимизации относится к классу дуальных и может быть назван методом дуального программирования.

Оба поисковых метода предусматривают проведение серии экспериментов. Полученные при этом экспериментальные данные могут быть использованы для построения аналитических регрессионных моделей частного критерия $f_1(p) = e(p)/L$. С помощью таких моделей можно осуществлять не поисковую, а аналитическую векторную оптимизацию архитектуры других нейросетевых классификаторов такого же вида. Если это окажется сложным, то проводится поисковая процедура, но уже с применением не натурного, а *вычислительного* эксперимента, что существенно проще.

Решая задачу построения регрессионных моделей, мы должны задать вид аппроксимирующей зависимости, известной с точностью до коэффициентов регрессии. Анализ задачи приводит к предположению, что с достаточной для практики точностью можно ограничиться линейной регрессией:

$$f_1(p_1, p_2) \approx (a_1 p_1 + a_2 p_2)/L\,, \qquad (7)$$

где $a_1, a_2$ – коэффициенты регрессии, определяемые по экспериментальным данным методом наименьших квадратов. Линейная регрессионная модель проверяется на адекватность методами математической статистики. При необходимости модель может быть усложнена.

Рассмотренные методы предусматривают старт поисковой процедуры от архитектуры, которая, по мнению разработчика, находится достаточно близко к оптимальной точке. Если в процессе поиска имеет место возрастание числа нейронов в обрабатывающих слоях, то теория нейронных сетей [1] характеризует данный подход как *конструктивный*. При избыточном стартовом количестве нейронов подход именуется *деструктивным* (принцип Родена: чтобы изваять скульптуру, нужно взять целую глыбу мрамора и удалить из нее лишнее).

Осуществление изложенных в работе этапов векторной оптимизации позволяет получить архитектуру нейросетевого классификатора, при которой системно увязываются противоречивые критерии эффективности его функционирования.

## Благодарности

Статья частично финансированна из проекта **ITHEA XXI** Института Информационных теории и Приложений FOI ITHEA и Консорциума FOI Bulgaria (www.ithea.org, www.foibg.com).

## Библиография

1. Бодянский Е.В., Руденко О.Г. Искусственные нейронные сети: архитектуры, обучение и применение. – Харьков: ТЕЛЕТЕХ, 2004. – 372 с.
2. Головко В.А. Нейронные сети: обучение, организация и применение. – М.: ИПРЖР, 2001. – 256 с.
3. Борисов В.С. Самообучающийся классификатор текстов на естественном языке // Кибернетика и системный анализ – 2007. – №3. – С.169-176.
4. Воронин А.Н., Зиатдинов Ю.К., Козлов А.И. Векторная оптимизация динамических систем. – Киев: Техніка, 1999. – 284 с.
5. Воронин А.Н., Мосорин П.Д., Ясинский А.Г. Многокритериальные задачи оптимизации с дискретными аргументами // Автоматика-2000. Міжнародна конференція з автоматичного управління: Праці. –Т.1. – Львів: ДНДІІІ, 2000. – С. 75-78.

## Сведения об авторах

***Воронин Альберт Николаевич*** *– профессор, доктор технических наук, профессор кафедры компьютерных информационных технологий Национального авиационного университета, проспект Комарова, 1, Киев-58, 03058 Украина; e-mail: alnv@voliacable.com*

***Зиатдинов Юрий Кашафович*** *– профессор, доктор технических наук, заведующий кафедрой компьютерных информационных технологий Национального авиационного университета, проспект Комарова, 1, Киев-58, 03058 Украина; e-mail:* oberst@nau.edu.ua

***Антонюк Анна Александровна*** *– аспирант Национального авиационного университета, проспект Комарова, 1, Киев-58, 03058 Украина, e-mail niuriel@mail.ru*

# О НЕКОТОРЫХ ТРУДНОРЕШАЕМЫХ ЗАДАЧАХ ПОМЕХОУСТОЙЧИВОГО АНАЛИЗА СТРУКТУРИРОВАННЫХ ДАННЫХ[1]

**Александр Кельманов**

***Аннотация**: Рассматриваются дискретные экстремальные задачи, к которым сводятся некоторые варианты проблемы помехоустойчивого off-line обнаружения в числовой последовательности повторяющегося фрагмента, а также некоторые варианты проблемы поиска подмножеств векторов во множестве векторов евклидова пространства. Анализируется сложность редуцированных оптимизационных задач и соответствующих им задач анализа данных и распознавания образов. Дан обзор новых и известных алгоритмических результатов по решению этих задач.*

***Ключевые слова**: поиск подмножеств векторов, помехоустойчивое обнаружение повторяющегося фрагмента, кластерный анализ, дискретная оптимизация, NP-трудная задача, алгоритмы с гарантированными оценками точности.*

***ACM Classification Keywords**: F.2. Analysis of Algorithms and Problem Complexity, G.1.6. Optimization, G2. Discrete Mathematics, I.5.3. Pattern Recognition: Clustering.*

***Conference**: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

## Введение

Объект исследования работы – проблемы оптимизации в задачах анализа данных и распознавания образов. Предмет исследования – дискретные экстремальные задачи, к которым сводятся некоторые варианты проблемы помехоустойчивого off-line обнаружения повторяющегося фрагмента в числовой последовательности и некоторые варианты проблемы поиска подмножеств «похожих» векторов во множестве векторов евклидова пространства. Цель работы – обзор новых и известных результатов по изучению сложности, систематизации и исследованию алгоритмов решения этих задач. Данная работа дополняет сообщения [1-3].

Представленные в работе модели анализа данных типичны для широкого спектра приложений, в которых необходимым элементом является компьютерная обработка массивов зашумленных структурированных данных, включающих повторяющиеся, чередующиеся или перемежающиеся информационно значимые фрагменты в одномерном случае или векторы в многомерном случае. Формулировки анализируемых ниже задач являются результатом: 1) формализации соответствующих содержательных (прикладных) задач либо в виде задач максимизации функционала правдоподобия (в случае, когда помеха аддитивна и является последовательностью гауссовских независимых одинаково распределенных случайных величин), либо в виде задач среднеквадратического приближения (когда о помехе известно лишь то, что она аддитивна), 2) последующей редукции этих задач к задачам дискретной оптимизации.

[1] Работа поддержана грантами РФФИ 09-01-00032, 07-07-00022 и грантом АВЦП Рособразования 2.1.1/3235.

## Модели анализа структурированных данных

Пусть $\mathbf{x}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$, где $\mathcal{N} = \{1,2,\ldots,N\}$, – последовательность векторов евклидова пространства. Рассмотрим две возможные структуры этой последовательности.

**Структура 1**. Последовательность задается формулой

$$\mathbf{x}_n = \begin{cases} \mathbf{w}_1, & n \in \mathcal{M}_1, \\ \mathbf{w}_2, & n \in \mathcal{M}_2, \\ \ldots, & \ldots, \\ \mathbf{w}_J, & n \in \mathcal{M}_J, \\ \mathbf{0}, & n \in \mathcal{N} \setminus \bigcup_{j=1}^{J} \mathcal{M}_j, \end{cases} \tag{1}$$

где $\bigcup_{j=1}^{J} \mathcal{M}_j \subseteq \mathcal{N}$, причем $\mathcal{M}_i \cap \mathcal{M}_j = \varnothing$, если $i \neq j$.

**Структура 2**. Последовательность обладает свойством

$$\mathbf{x}_n = \begin{cases} \mathbf{w}_1, & n \in \mathcal{M}_1, \\ \mathbf{w}_2, & n \in \mathcal{M}_2, \\ \ldots, & \ldots, \\ \mathbf{w}_J, & n \in \mathcal{M}_J, \end{cases} \tag{2}$$

где $\bigcup_{j=1}^{J} \mathcal{M}_j = \mathcal{N}$, причем $\mathcal{M}_i \cap \mathcal{M}_j = \varnothing$, если $i \neq j$.

Положим $|\mathcal{M}_j| = M_j$, $j = 1,2,\ldots,J$, и $\{n_1,\ldots,n_M\} = \bigcup_{j=1}^{J} \mathcal{M}_j$, где $M = \sum_{j=1}^{J} M_j$. Векторы $\mathbf{w}_j$ будем интерпретировать как информационно значимые векторы, а $M_j$ – как число их повторов в последовательности $\mathbf{x}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$. Доступной для анализа будем считать последовательность

$$\mathbf{y}_n = \mathbf{x}_n + \mathbf{e}_n, \quad n \in \mathcal{N}, \tag{3}$$

где $\mathbf{e}_n$ – вектор помехи (ошибки измерения), независимый от вектора $\mathbf{x}_n$. Положим

$$S(\mathcal{M}_1,\mathcal{M}_2,\ldots,\mathcal{M}_J,\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_J) = \sum_{n\in\mathcal{N}} \| \mathbf{y}_n - \mathbf{x}_n \|^2. \tag{4}$$

Модели анализа данных сформулируем в форме задач среднеквадратического приближения.

Допустим сначала, что в отсутствие шума данные имеют структуру 1. Сформулируем следующие задачи.

**Задача 1.** *Дано*: совокупность $\{\mathbf{y}_1,\ldots,\mathbf{y}_N\}$ векторов из $\mathcal{R}^q$. *Найти*: семейство $\{\mathcal{M}_1,\mathcal{M}_2,\ldots,\mathcal{M}_J\}$ непустых непересекающихся подмножеств множества $\mathcal{N}$ и совокупность $\{\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_J\}$ векторов такие, что то целевая функция (4) минимальна.

Эту задачу можно трактовать как поиск семейства непересекающихся подмножеств векторов, похожих в среднеквадратическом.

Допустим, что в рамках структуры 1 компоненты набора $(n_1,\ldots,n_M)$, элементы которого соответствуют номерам ненулевых векторов в формуле (1), связаны дополнительными ограничениями

$$1 \le T_{\min} \le n_m - n_{m-1} \le T_{\max} \le N-1, \quad m = 2,\ldots,N, \tag{5}$$

где $T_{\min}$ и $T_{\max}$ – натуральные числа. Эти ограничения устанавливают допустимый интервал между двумя ближайшими номерами ненулевых информационно значимых векторов в последовательности (1).

**Задача 2.** *Дано*: последовательность $\mathbf{y}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$. *Найти*: семейство $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_J\}$ непустых непересекающихся подмножеств множества $\mathcal{N}$ и совокупность $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_J\}$ векторов такие, что целевая функция (4) минимальна, при ограничениях (5) на элементы упорядоченного набора $(n_1, \ldots, n_M)$, которые образуют совокупность $\{n_1, \ldots, n_M\} = \bigcup_{j=1}^{J} \mathcal{M}_j$.

Задачу 2 можно трактовать как совместное оптимальное обнаружение и оценивание по критерию минимума суммы квадратов уклонений ненулевых неизвестных информационно значимых векторов, повторяющихся и перемежающихся в ненаблюдаемой последовательности (1).

Для данных, имеющих структуру 2, сформулируем следующую задачу.

**Задача 3**. *Дано*: совокупность $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ векторов из $\mathcal{R}^q$. *Найти*: разбиение множества $\mathcal{N}$ на непустые подмножества $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_J$ и совокупность $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_J\}$ векторов такие, что то целевая функция (4) минимальна.

Эта задача отличается от задачи 1 тем, что в ней требуется найти разбиение множества $\mathcal{N}$, а не совокупность непересекающихся подмножеств этого множества. При этом предполагается, что структура данных описывается формулой (2).

## Редуцированные экстремальные задачи

Легко убедиться, что во всех сформулированных задачах для любого допустимого семейства $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_J\}$ подмножеств множества $\mathcal{N}$ минимум функционала (4) по переменным $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_J$, доставляется векторами $\overline{\mathbf{w}}_j = \sum_{n \in \mathcal{M}_j} \mathbf{y}_n / |\mathcal{M}_j|$, $j = 1,2,\ldots,J$. В задачах 1 и 2 в силу формулы (1) этот минимум равен

$$S_{\min} = \sum_{n \in \mathcal{N}} \| \mathbf{y}_n \|^2 - \sum_{j=1}^{J} \frac{1}{|\mathcal{M}_j|} \| \sum_{n \in \mathcal{M}_j} \mathbf{y}_n \|^2 . \quad (6)$$

Для задачи 3, учитывая (2), имеем

$$S_{\min} = \sum_{j=1}^{J} \sum_{n \in \mathcal{M}_j} \| \mathbf{y}_n - \overline{\mathbf{w}}_j \|^2 . \quad (7)$$

Таким образом, для отыскания решений сформулированных задач необходимо решить задачи на минимум функций (6) и (7). К идентичным оптимизационным задачам приводит статистический подход к проблеме анализа данных, если считать, что вектор $\mathbf{e}_n$ в формуле (4) есть выборка из $q$-мерного нормального распределения с параметрами $(\mathbf{0}, \sigma^2 \mathbf{I})$, где $\mathbf{I}$ единичная матрица, а в модели анализа данных в качестве критерия решения использовать максимум функционала правдоподобия.

Первый член в правой части равенства (6) является константой. Поэтому из задачи 1 получаем следующие редуцированные оптимизационные задачи.

**Задача $J$ -MSASVS-F** (максимум суммы средних значений квадратов длин сумм векторов из подмножеств фиксированной мощности). *Дано*: множество $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ векторов из $\mathcal{R}^q$ и натуральные числа

$M_1, M_2, \ldots, M_J$. *Найти*: семейство $\{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_J\}$ непустых непересекающихся подмножеств множества $\mathcal{Y}$ такое, что

$$\sum_{j=1}^{J} \frac{1}{|\mathcal{B}_j|} \| \sum_{y \in \mathcal{B}_j} \mathbf{y} \|^2 \to \max, \tag{8}$$

при ограничениях: $|\mathcal{B}_j| = M_j$, $j = 1, \ldots, J$.

**Задача $J$-MSASVS-NF** (максимум суммы средних значений квадратов длин сумм векторов из подмножеств, мощности которых не фиксированы). *Дано*: множество $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ векторов из $\mathcal{R}^q$. *Найти*: семейство $\{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_J\}$ непустых непересекающихся подмножеств множества $\mathcal{Y}$ такое, что имеет место (8).

Обе задачи можно трактовать как поиск подмножеств векторов «похожих» в среднеквадратическом смысле. Отличие задач состоит в том, что в первой из них мощности искомых подмножеств являются частью входа задачи, а во второй эти мощности – оптимизируемые величины. Аналогичным образом формулируются еще две задачи, которые следуют из задачи 2 и ориентированы на анализ последовательностей при наличии ограничений (5).

**Задача $J$-MSASVSO-F**. *Дано*: последовательность $\mathbf{y}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$, и натуральные числа $M_1, M_2, \ldots, M_J$, $T_{\min}$ и $T_{\max}$. *Найти*: семейство $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_J\}$ непустых непересекающихся подмножеств множества $\mathcal{N}$ такое, что

$$\sum_{j=1}^{J} \frac{1}{|\mathcal{M}_j|} \| \sum_{n \in \mathcal{M}_j} \mathbf{y}_n \|^2 \to \max, \tag{9}$$

при ограничениях $|\mathcal{M}_j| = M_j$, $j = 1, \ldots, J$, на мощности подмножеств и при дополнительных ограничениях (5) на элементы упорядоченного набора $(n_1, \ldots, n_M)$, которые образуют совокупность $\{n_1, \ldots, n_M\} = \bigcup_{j=1}^{J} \mathcal{M}_j$.

**Задача $J$-MSASVSO-NF**. *Дано*: последовательность $\mathbf{y}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$, и натуральные числа $T_{\min}$ и $T_{\max}$. *Найти*: семейство $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_J\}$ непустых непересекающихся подмножеств множества $\mathcal{N}$ такое, что имеет место (9), при ограничениях (5) на элементы упорядоченного набора $(n_1, \ldots, n_M)$, которые образуют совокупность $\{n_1, \ldots, n_M\} = \bigcup_{j=1}^{J} \mathcal{M}_j$.

Из задачи 3 и формулы (7) получаем хорошо известную задачу.

**Задача MSSC**. *Дано*: множество $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ векторов из $\mathcal{R}^q$ и натуральное число $J > 1$. *Найти*: разбиение множества $\mathcal{Y}$ на непустые подмножества (кластеры) $C_1, C_2, \ldots, C_J$ такое, что

$$\sum_{j=1}^{J} \sum_{\mathbf{y} \in C_j} \| \mathbf{y} - \overline{\mathbf{w}}_j \|^2 \to \min,$$

где $\overline{\mathbf{w}}_j = \sum_{\mathbf{y} \in C_j} \mathbf{y} / |C_j|$, $j = 1, 2, \ldots, J$, – центры кластеров.

Эта задача является классической задачей анализа данных и распознавания образов. Ниже сформулированы два важных специальных случая этой задачи.

**Задача $J$ -MSSC0-F**. *Дано*: множество $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ векторов из $\mathcal{R}^q$ и натуральные числа $M_1, M_2, \ldots, M_J$. *Найти*: разбиение множества $\mathcal{Y}$ на непустые подмножества $C_1, C_2, \ldots, C_J$ такое, что

$$\sum_{j=1}^{J-1} \sum_{\mathbf{y} \in C_j} \| \mathbf{y} - \overline{\mathbf{w}}_j \|^2 + \sum_{\mathbf{y} \in C_J} \| \mathbf{y} \|^2 \to \min, \tag{10}$$

где $\overline{\mathbf{w}}_j = \sum_{\mathbf{y} \in C_j} \mathbf{y} / |C_j|$, $j = 1,2,\ldots,J-1$, – центры кластеров, при ограничениях $|C_j| = M_j$, $j = 1,\ldots,J$.

**Задача $J$ -MSSC0-NF**. *Дано*: множество $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ векторов из $\mathcal{R}^q$. *Найти*: разбиение множества $\mathcal{Y}$ на непустые подмножества $C_1, C_2, \ldots, C_J$ такое, что имеет место (10).

Эти задачи можно трактовать как специальные случаи задачи MSSC, в которых центр одного из кластеров определять не требуется (считается, что центр этого кластера известен и равен нулю). В первой задаче предполагается, что мощности кластеров фиксированы, а во второй число кластеров и их мощности – оптимизируемые величины.

## Известные факты о сложности сформулированных задач и алгоритмах их решения

Прежде всего, заметим, что задача MSSC в силу своей широкой известности и давности постановки наиболее изучена в алгоритмическом плане. Имеется множество публикаций, ориентированных на построение эффективных алгоритмов с оценками точности для ее решения. Однако, лишь недавно в [4] дано корректное доказательство *NP*-трудности этой задачи для случая, когда $J = 2$. Все ранее опубликованные доказательства труднорешаемости этой задачи содержали ошибки [5]. Другие задачи, сформулированные в предыдущем параграфе, относятся к числу слабо изученных задач. Рассмотрим современное состояние исследований по их решению.

**Алгоритмическая сложность**. Относительно сложности задач поиска подмножеств векторов и специальных случаев задачи кластерного анализа получены следующие результаты. Статус *NP*-трудности задачи 1-MSASVS-F был установлен в [6, 7]. Из этого результата следует, что задача $J$ -MSASVS-F при $J > 1$ также *NP*-трудна, как обобщение задачи 1-MSASVS-F. *NP*-трудность задачи 1-MSASVS-NF доказана в [8, 9]. Этот результат позволил установить труднорешаемость задачи $J$ -MSASVS-NF при $J > 1$ в случае, когда число $J$ является частью входа. Позже в [10] была установлена труднорешаемость задачи $J$ -MSASVS-NF для случая, когда $J$ не является частью входа. В этой же работе было доказано, что задачи $J$ -MSSC0-F и $J$ -MSSC0-NF также *NP*-трудны.

О сложности задач с ограничением (5) на порядок выбора векторов известно следующее. Статус *NP*-трудности доказан [6, 7] лишь для задачи $J$ -MSASVSO-F. Статус сложности задачи $J$ -MSASVSO-NF пока не установлен. Скорее всего, она *NP*-трудна, как и задача $J$ -MSASVS-NF.

**Алгоритмы**. Какие-либо алгоритмы с доказуемыми оценками точности для решения задач $J$ -MSASVS-F и $J$ -MSASVS-NF поиска подмножеств векторов, задач $J$ -MSASVSO-F и $J$ -MSASVSO-NF поиска подпоследовательностей векторов в случае, когда $J > 1$, на сегодняшний день неизвестны. То же самое можно сказать про задачи $J$ -MSSC0-F и $J$ -MSSC0-NF , которые имеют смысл лишь при $J > 1$.

К числу задач, для которых удалось построить алгоритмы с доказуемыми оценками точности, относятся простейшие задачи 1-MSASVS-F, 1-MSASVS-NF и 1-MSASVSO-F, в которых требуется найти лишь одно ($J=1$) подмножество «похожих» векторов или один повторяющийся вектор в последовательности. В [7] обоснованы приближенные асимптотически точные алгоритмы решения задач 1-MSASVS-F и 1-MSASVSO-F, имеющие временную сложность $O[Nq^2(2l+1)^{q-1}]$ и $O[Nq(q+M)(2l+1)^{q-1}]$ соответственно, где $l$ – параметр алгоритма. Относительная погрешность у этих алгоритмов равна $(q-1)/(4l^2)$. В [6] предложен приближенный алгоритм решения задачи 1-MSASVSO-F. Его временная сложность есть величина $O(MN^2)$. К сожалению, для этого относительно «быстрого» алгоритма, хорошо зарекомендовавшего себя в численных экспериментах, гарантированная оценка точности пока не установлена.

Для решения задачи 1-MSASVS-NF в работе [10] предложен приближенный асимптотически точный алгоритм. Трудоемкость и относительная погрешность у этого алгоритма есть величины $O[Nq(q+\log N)(2l+1)^{q-1}]$ и $(q-1)/(4l^2)$, где $l$ – параметр алгоритма.

В [11] доказано, что задачи 1-MSASVS-F и 1-MSASVS-NF разрешимы за время $O(q^2N^{2q})$. Тем самым показано, что при фиксированной размерности $q$ пространства эти задачи могут быть точно решены за полиномиальное время.

Для вариантов задач 1-MSASVS-F и 1-MSASVSO-F с целочисленными координатами векторов в [12] обоснованы точные псевдополиномиальные алгоритмы. Трудоемкость этих алгоритмов есть величина $O[NqM^q(2b)^{q-1}]$, где $b$ - максимальная по абсолютной величине координата векторов из заданного множества.

## Заключение

К рассмотренным *NP*-трудным задачам сводятся простейшие проблемы из большого семейства (насчитывающего, по крайней мере, несколько сотен элементов [13]) проблем помехоустойчивого off-line анализа и распознавания структурированных последовательностей, включающих повторяющиеся, чередующиеся и перемежающиеся информационно значимые векторы (фрагменты) в качестве структурных элементов. Очевидно, что эти труднорешаемые задачи являются частными случаями для многих еще не изученных экстремальных задач, к которым сводятся проблемы анализа данных и распознавания образов, имеющих более сложную структуру над информационно значимыми векторами. Поэтому приведенные результаты могут служить в качестве базовых (при использовании известной [14] техники полиномиальной сводимости) для доказательства *NP*-трудности других более сложных проблем анализа структурированных данных и распознавания образов из упомянутого семейства.

Остается заметить, что для большинства из рассмотренных экстремальных задач какие-либо алгоритмы с оценками точности на сегодняшний день неизвестны. Высокая с практической точки зрения трудоемкость существующих приближенных алгоритмов решения некоторых из рассмотренных оптимизационных задач обуславливает продолжение исследований в направлении поиска новых алгоритмических решений, а также в направлении выделения подклассов задач, для которых возможно построение алгоритмов, имеющих меньшую временную сложность.

## Благодарности

Работа поддержана грантами РФФИ 09-01-00032, 07-07-00022 и грантом АВЦП Рособразования 2.1.1/3235.

## Литература

[1] Кельманов А.В. Полиномиально разрешимые и NP-трудные варианты задачи оптимального обнаружения в числовой последовательности повторяющегося фрагмента // Материалы Росс. конф. «Дискретная оптимизация и исследование операций» (Владивосток, 7-14 сентября 2007). – Новосибирск: Изд-во Института математики СО РАН, 2007.- http://math.nsc.ru/ conference/ door07/ DOOR_abstracts.pdf. С. 46-50.

[2] Кельманов А.В. О некоторых полиномиально разрешимых и NP-трудных задачах анализа и распознавания последовательностей с квазипериодической структурой // Сборник докладов 13-й Всеросс. конф. «Математические методы распознавания образов» (ММРО-13). Ленинградская обл., г. Зеленогорск, 30 сентября - 6 октября 2007 г. - М.: МАКС Пресс, 2007. - С. 261-264.

[3] Kel'manov A.V. Off-line Detection of a Quasi-Periodically Recurring Fragment in a Numerical Sequence // Proceedings of the Steklov Institute of Mathematics. 2008, Suppl. 2, pp. S84-S92.

[4] Aloise D., Deshpande A., Hansen P., Popat P. NP-Hardness of Euclidean Sum-of-Squares Clustering // Les Cahiers du GERAD, G-2008-33. 2008. 4 p.

[5] Aloise D., Hansen P. On the Complexity of Minimum Sum-of-Squares Clustering // Les Cahiers du GERAD, G-2007-50. 2007. 12 p.

[6] Gimadi E.Kh., Kel'manov A.V., Kel'manova M.A., Khamidullin S.A. A Posteriori Detecting a Quasiperiodic Fragment in a Numerical Sequence // Pattern Recognition and Image Analysis. 2008. Vol. 18, No.1. P. 30-42.

[7] Бабурин А.Е., Гимади Э.Х., Глебов Н.И., Пяткин А.В. Задача отыскания подмножества векторов с максимальным суммарным весом // Дискретный анализ и исследование операций. Серия 2. 2007. Т.14, №1. С. 32-42.

[8] Kel'manov A.V., Pyatkin A.V. On the Complexity of a Search for a Subset of "Similar" Vectors // Doklady Mathematics. 2008. Vol. 78, No. 1. P. 574-575.

[9] Кельманов А.В., Пяткин А.В. Об одном варианте задачи выбора подмножества векторов // Дискретный анализ и исследование операций. 2008. Т.15, №5. С. 25-40.

[10] Кельманов А.В., Пяткин А.В. О сложности некоторых задач поиска подмножеств векторов и кластерного анализа // Журнал вычислительной математики и математической физики. 2009 (принята в печать).

[11] Гимади Э.Х., Пяткин А.В., Рыков И.А. О полиномиальной разрешимости некоторых задач выбора подмножеств векторов в евклидовом пространстве фиксированной размерности // Дискретный анализ и исследование операций. 2008. Т.15, №6. С. 11-19.

[12] Гимади Э.Х., Глазков Ю.В., Рыков И.А. Задача выбора подмножества векторов с целочисленными координатами в евклидовом пространстве с максимальной нормой суммы // Дискретный анализ и исследование операций. 2008. Т.15, №4. С. 31-43.

[13] http://math.nsc.ru/~serge/qpsl/

[14] Garey M.R., Johnson D.S. Computers and Intractability: A Guide to the Theory of NP-Completeness, Freeman, San Francisco, CA, 1979.

## Информация об авторе

***Александр Кельманов*** *– д.ф.-м.н., ведущий научный сотрудник, Институт математики им. С.Л. Соболева Сибирского отделения РАН, проспект академика Коптюга, 4, Новосибирск, 630090, Россия; Новосибирский государственный университет, ул. Пирогова, 2, Новосибирск, 630090, Россия; e-mail: kelm@math.nsc.ru*

# ОПТИМИЗАЦИЯ ОЦЕНКИ ВЕРОЯТНОСТИ ОШИБОЧНОЙ КЛАССИФИКАЦИИ В ДИСКРЕТНОМ СЛУЧАЕ[1]

**Виктор Неделько**

***Abstract**: The goal of the paper is to investigate what training sample estimate of misclassification probability would be the best one for the histogram classifier. Certain quality criterion is suggested. The deviation for some estimates, such as resubstitution error (empirical risk), cross validation error (leave-one-out), bootstrap and for the best estimate obtained via some optimization procedure, is calculated and compared for some examples.*

***Keywords**: pattern recognition, classification, statistical robustness, deciding functions, complexity, capacity, overfitting, overtraining problem.*

***ACM Classification Keywords**: G.3 Probability and statistics, G.1.6. Numerical analysis: Optimization; G.2.m. Discrete mathematics: miscellaneous.*

***Conference**: The paper is selected from XV$^{th}$ International Conference "Knowledge-Dialogue-Solution" KDS 2009, Varna, Bulgaria, June-July 2009*

## Введение

Для оценивания качества решающих функций (одна из первых работ [Лбов, 1965]) в задачах распознавания образов (классификации с учителем) на практике обычно используются точечные оценки риска, т.е. вероятности ошибочной классификации. В роли таких оценок, как правило, выступают эмпирический риск (resubstitution error) оценка скользящего экзамена (cross validation) или оценка bootstrap. При этом эмпирический риск является смещенной оценкой риска. Для величины смещения в общем случае существуют лишь приближенные интервальные оценки в рамках подхода Вапника–Червоненкиса [Вапник, Червоненкис, 1974], хотя для частных случаев возможно точное оценивание смещения, например, для дискретного пространства [Неделько, 2003]. Также имеет смысл использование эмпирических интервальных оценок риска [Неделько, 2008].

Наилучшей с практической точки зрения из точечных оценок риска считается bootstrap, чье преимущество продемонстрировано на многочисленных примерах. Естественным образом напрашивается вопрос, насколько эта оценка близка к оптимальной, и в каком смысле можно вообще говорить об оптимальности такого рода оценки [Неделько, 2007].

Стандартной мерой качества точечной оценки является ее эффективность, которая характеризуется средним квадратом отклонения (deviation) от оцениваемой величины. Однако эта величина зависит от вероятностной модели, т.е. распределения, из которого взята выборка, и для разных распределений оптимальными будут разные оценочные функционалы. Получаем ситуацию многокритериального выбора. В этом случае можно рассматривать множества Парето-оптимальных оценок. Но в данной ситуации критерии сравнимы, поскольку являются фактически одним критерием при разных моделях. Это позволяет сравнивать оценки, считая, что один функционал лучше другого, если его выигрыш в лучшей ситуации превосходит проигрыш в худшей.

[1] Работа выполнена при поддержке РФФИ, гранты 07-01-00331-а и 08-01-00944-а.

Если бы на множестве всех распределений была задана некоторая мера [Лбов, Старцева, 1999], адекватно отражающая «важность» этих распределений, или их «встречаемость» в реальных задачах, то можно было бы просто использовать усредненный критерий. Но так как такой меры нет, разумным представляется использование различных вариаций минимаксного подхода.

Задача нахождения оптимального оценочного функционала в общем случае является сложной, поэтому в данной работе исследуется частный случай задачи классификации в дискретном пространстве (histogram classifier), при котором все требуемые статистики могут быть вычислены аналитически [Braga–Neto, Dougherty, 2005].

## Постановка задачи

Для введения основных понятий рассмотрим сначала общую постановку задачи построения решающих функций.

Пусть $X$ – пространство значений переменных, используемых для прогноза, а $Y$ – пространство значений прогнозируемых переменных, и пусть $C$ – множество всех вероятностных мер на $D = X \times Y$. Тогда элементом $c \in C$ будет $P_c[D]$. Здесь и далее квадратные скобки используются для указания множества, на $\sigma$-алгебре подмножеств которого задана мера.

Решающей функцией назовем соответствие $f : X \to Y$ и введем для нее функцию потерь: $L : Y^2 \to [0, \infty)$.

Под риском будем понимать средние потери:

$$R(c, f) = \int L(y, f(x)) dP_c[D].$$

Пусть $V = \left\{ \left(x^i, y^i\right) \in D \;\middle|\; i = \overline{1, N} \right\}$ – случайная независимая выборка из распределения $P_c[D]$, $V \in D^N$. Эмпирический риск определим как средние потери на выборке:

$$\widetilde{R}(v, f) = \frac{1}{N} \sum_{i=1}^{N} L\left(y^i, f\left(x^i\right)\right).$$

Пусть $Q : D^N \to \Phi$ – алгоритм построения решающих функций, а $f_{Q,V} \in \Phi$ – функция, построенная по выборке $V$ алгоритмом $Q$.

Оценкой скользящего экзамена называется величина

$$\breve{R}(V, Q) = \frac{1}{N} \sum_{i=1}^{N} L\left(y^i, f_{Q,V_i'}\left(x^i\right)\right),$$

где $V_i' = V \setminus \left\{ \left(x^i, y^i\right) \right\}$ – выборка, получаемая из $V$ удалением $i$-го наблюдения.

Также мы будем использовать оценку bootstrap

$$\widehat{R}(V, Q) = \frac{1}{\mathrm{E}|J_0|} \mathrm{E} \sum_{i \in J_0} L\left(y^i, f_{Q,\widehat{V}}\left(x^i\right)\right),$$

где $\widehat{V}$ – выборка, получаемая из $V$ путем $N$-кратного случайного (равновероятного) выбора ее значений с повторениями, $J_0$ – множество индексов объектов из $V$, ни разу не выбранных в $\widehat{V}$, математическое ожидание подразумевает усреднение по выборкам $\widehat{V}$. Легко показать, что $\mathrm{E}|J_0| = N\left(1 - \frac{1}{N}\right)^N \approx N e^{-1}$.

Ввиду того, что оценка bootstrap является смещенной, чаще используют ее в комбинации с эмпирическим риском

$$\ddot{R}(V,Q) = e^{-1} \cdot \tilde{R}(V,Q) + \left(1 - e^{-1}\right) \cdot \hat{R}(V,Q).$$

В общем случае оценочный функционал — это некоторая функция выборки (при фиксированном методе построения решающих функций).

Качество эмпирического функционала $\overline{R}\left(V, f_{Q,v}\right)$ как оценки риска естественно характеризовать средним квадратом уклонения, т.е.

$$\Delta = \mathrm{E}\left( \overline{R}\left(V, f_{Q,V}\right) - R\left(c, f_{Q,V}\right)\right)^2.$$

Существенная проблема заключается в том, что выражения зависят от $c$ – распределения, которое неизвестно. Решением может быть взятие супремума по всем распределениям и ориентирование таким образом на «наихудшее» распределение.

## Классификация в дискретном пространстве

Будем рассматривать задачу классификации двух образов.

Пусть $X$ дискретно, то есть $X = \{1, \ldots, k\}$, и решающая функция минимизирует эмпирический риск независимо в каждой точке $x$.

Тогда вероятностная мера $c \in C$ задается набором вероятностей

$$c = \left\{ \varsigma_j^{\omega} = \mathrm{P}\left(x = j, y = \omega\right) \middle| \; j = \overline{1,k},\ \omega = \overline{1,2} \right\}.$$

При этом $Y = \{1, 2\}$, функцией потерь будет: $L\left(y, y'\right) = \begin{cases} 0, & y = y' \\ 1, & y \neq y' \end{cases}$, а риском – вероятность ошибочной классификации.

Обозначим $\alpha_j = P\left(x = i\right) = \varsigma_j^1 + \varsigma_j^2$, $p_j = P\left(y = 1 / x = i\right)$, $q_j = 1 - p_j$, $c_j = \left(\alpha_j, p_j\right)$.

Для выборки $V$ объема $N$ пусть $n_j$ – число точек выборки, для которых $x = j$, и $m_j$ – число точек, для которых $x = j$ и $y = 1$. Таким образом, выборка в дискретном случае задается совокупностью пар $\nu_j = \left(m_j, n_j\right)$, т.е. $V = \left\{ \nu_j \middle| \; j = \overline{1,k} \right\}$. Описывая выборку, мы будем иногда для краткости говорить, что в «ячейке» $j$ находится $m_j$ точек первого и $n_j - m_j$ точек второго класса.

Будем рассматривать алгоритм $Q$, который минимизирует эмпирический риск независимо в каждой точке пространства $X$, т.е. $f_{Q,V}\left(j\right) = 2$, при $n_j - m_j > m_j$, $f_{Q,V}\left(j\right) = 1$, при $n_j - m_j < m_j$, и $f_{Q,V}\left(j\right)$ принимает равновероятно значения 1 и 2, при $n_j - m_j = m_j$.

На выборках имеет место полиномиальное распределение $P\left(V\right)$, суммируя по которому, можно вычислять в том числе моменты различных функций выборки. Однако осуществлять перебор всех выборок – трудоемкая в вычислительном плане процедура, поэтому непосредственное суммирование по выборкам осуществимо только для небольших $N$ и $k$.

При этом для аддитивных функций выборки вычисление моментов может быть произведено с полиномиальной трудоемкостью.

## Вычисление моментов для аддитивных функций

Пусть $f(V)=\sum_{j=1}^{k}\varphi(\nu_j,c_j)=\sum_{j=1}^{k}\varphi(m_j,n_j,\alpha_j,p_j)$ – аддитивная функция выборки и распределения.

Математическое ожидание $\mathrm{E}\,f(V)=\sum_{j=1}^{k}\mathrm{E}\varphi(\nu_j,c_j)$ также аддитивно.

Обозначим $B(m,n,p)=C_n^m p^m(1-p)^{n-m}$ – биномиальное распределение.

Введем функцию $\mu_\varphi(c)\equiv\mu_\varphi(\alpha,p)=\mathrm{E}\varphi(\nu,c)$. Легко получить, что

$$\mu_\varphi(\alpha,p)=\sum_{n=0}^{N}B(n,N,\alpha)\sum_{m=0}^{n}B(m,n,p)\varphi(m,n,\alpha,p)=\sum_{n=0}^{N}B(n,N,\alpha)\pi_\varphi(n,\alpha,p),$$

где $\pi_\varphi(n,\alpha,p)=\sum_{m=0}^{n}B(m,n,p)\varphi(m,n,\alpha,p)$.

Окончательно, математическое ожидание есть $\mathrm{E}\,f(V)=\sum_{j=1}^{k}\mu_\varphi(c_j)$.

Для вычисления дисперсии имеем $\mathrm{D}\,f(V)=\mathrm{E}\,f^2(V)-(\mathrm{E}\,f(V))^2$.

$$\mathrm{E}\,f^2(V)=\sum_{j=1}^{k}\mathrm{E}\varphi^2(\nu_j,c_j)+\sum_{i\neq j}\mathrm{E}\,\varphi(\nu_i,c_i)\varphi(\nu_j,c_j).$$

Введем функции

$$\sigma_\varphi(c)\equiv\sigma_\varphi(\alpha,p)=\mathrm{E}\varphi^2(\nu,c),\ \ \omega_\varphi(c_1,c_2)\equiv\omega_\varphi(\alpha_1,p_1,\alpha_2,p_2)=\mathrm{E}\varphi(\nu_1,c_1)\varphi(\nu_2,c_2).$$

Имеем

$$\sigma_\varphi(\alpha,p)=\sum_{n=0}^{N}B(n,N,\alpha)\pi_\varphi^2(n,\alpha,p),\ \text{где}\ \ \pi_\varphi^2(n,\alpha,p)=\sum_{m=0}^{n}B(m,n,p)\varphi^2(m,n,\alpha,p).$$

$$\omega_\varphi(\alpha_1,p_1,\alpha_2,p_2)=\sum_{n=0}^{N}B(n,N,\alpha_1+\alpha_2)\sum_{n_1+n_2=n}B(n_1,n,\alpha_1')\,\pi_\varphi(n_1,\alpha_1,p_1)\pi_\varphi(n_2,\alpha_2,p_2),$$

где $\alpha_1'=\frac{\alpha_1}{\alpha_1+\alpha_2}$.

Окончательно, второй момент есть $\mathrm{E}\,f^2(V)=\sum_{j=1}^{k}\sigma_\varphi(c_j)+\sum_{i\neq j}\omega_\varphi(c_i,c_j)$.

Пусть $g(V)=\sum_{j=1}^{k}\psi(\nu_j,c_j)$ – также аддитивная функция выборки и распределения.

Смешанный момент

$$\mathrm{E}\,f(V)g(V)=\sum_{j=1}^{k}\mathrm{E}\varphi(\nu_j,c_j)\psi(\nu_j,c_j)+\sum_{i\neq j}\mathrm{E}\,\varphi(\nu_i,c_i)\psi(\nu_j,c_j)$$

вычисляется аналогично рассмотренным.

## Оптимизация оценки риска

Пусть $f(V)$ – некоторая аддитивная оценка риска, а $g(V)$ – фактическое значение риска (вероятности ошибочной классификации), который в рассматриваемом дискретном случае также является аддитивной функцией.

Функция $f(V)$ полностью определяется функцией $\varphi(\nu, c)$, которая на самом деле не может зависеть от $c$, поскольку при построении оценки риска распределение неизвестно. Кроме того, данная функция дискретна и определяется счетным набором значений. Обозначим $\varphi(\nu) \equiv \varphi(m,n) = x_{mn}$.

Требуется подобрать $x_{mn}$ так, чтобы минимизировать погрешность оценивания риска, т.е. величину

$$\Delta_{fg} = \mathrm{E}(f-g)^2 = \mathrm{E}\, f^2 - 2\mathrm{E}\, fg + \mathrm{E}\, g^2 .$$

Пусть $\alpha_j = \alpha = \frac{1}{k}$ и $p_j = p$.

Вычислим частные производные

$$\frac{\partial \Delta_{fg}}{\partial x_{mn}} = 2k\, B(m,n,p)\Big((x_{mn} - \psi(m,n,\alpha,p))B(n,N,\alpha) + (k-1)c_{\varphi\psi}(n,N-n,\alpha,p)\Big), \text{ где}$$

$$c_{\varphi\psi}(n,N-n,\alpha,p) = \sum_{i=0}^{N-n} B(i+n,N,2\alpha)\; B(n,i+n,0{,}5)\big(\pi_{\varphi}(i,\alpha,p) - \pi_{\psi}(i,\alpha,p)\big).$$

Вторая производная

$$\frac{\partial^2 \Delta_{fg}}{\partial x_{mn}^2} = 2k\, B(m,n,p)\big(\, B(n,N,\alpha) + (k-1)B(2n,N,2\alpha)\; B(n,2n,0{,}5)\; B(m,n,p)\big).$$

Пусть $\delta^+(x_{mn}) = \max\limits_p \frac{\partial \Delta_{fg}}{\partial x_{mn}}$, $\delta^-(x_{mn}) = \min\limits_p \frac{\partial \Delta_{fg}}{\partial x_{mn}}$, а $p_{\max}$ и $p_{\min}$ – значения параметра $p$, при которых соответственно достигаются указанные максимум и минимум, и $\delta_2(x_{mn}) = \frac{\partial^2 \Delta_{fg}}{\partial x_{mn}^2}(p_{\max}) + \frac{\partial^2 \Delta_{fg}}{\partial x_{mn}^2}(p_{\min})$.

Наилучшей оценкой риска $x_{mn}^*$ будем считать значение, при котором $\delta^+(x_{mn}^*) = -\delta^-(x_{mn}^*)$. При изменении оценки в окрестности точки $x_{mn}^*$ максимальное по всем распределениям улучшение точности оценки будет равно максимальному ее ухудшению. Это значение представляется в определенном смысле оптимальным выбором, т.к. при других вариантах мы можем взять близкое значение, при котором максимальное уменьшение погрешности $\Delta_{fg}$ будет больше ее максимального увеличения. Оценку $x_{mn}^*$ будем называть *сбалансировано-оптимальной*.

Для решения уравнения и нахождения $x_{mn}^*$ использован аналог метода касательных, где начальным приближением взят эмпирический риск $x_{mn}^0 = \min(m, n-m)/N$, а последующие приближения вычислялись через предыдущие по формуле

$$x_{mn}^{i+1} = x_{mn}^i - \tau \frac{\delta^+(x_{mn}^i) + \delta^-(x_{mn}^i)}{\delta_2(x_{mn}^i)},$$

где $\tau \approx 0{,}1$ – параметр, введенный для обеспечения устойчивости (сходимости) метода. Заметим, что это не вполне метод касательных, поскольку $\delta_2(x_{mn})$ – не есть производная функции $\delta^+(x_{mn}) + \delta^-(x_{mn})$, но может выступать в роли эвристической оценки последней.

## Экспериментальное сравнение оценок

Было проведено численное сравнение точности перечисленных оценок риска при различных значениях параметров задачи: объема выборки $N$ и числа значений $k$.

Эмпирический риск и оценка скользящего экзамена являются аддитивными функциями и соответствующие им оценки выражаются соответственно

$$\widetilde{x}_{mn} = \tfrac{1}{N}\min(m, n-m),$$

$$\breve{x}_{mn} = \tfrac{1}{N}\Big(\min(m, n-m) + \max(m, n-m)\cdot\big(I(m = n-m) + \tfrac{1}{2}I(|n-2m| = 1)\big)\Big),$$

где $I(\cdot)$ – индикаторная функция (равна 1, если условие истинно, и 0 – иначе).

Оценка bootstrap вычисляется следующим образом

$$\widehat{x}_{mn} = \frac{\left(1-\frac{1}{N}\right)^{-N}}{N}\sum_{n'=0}^{N}\sum_{m'=0}^{n'}\sum_{n_0=0}^{n}\sum_{m_0=0}^{n_0}\widehat{r}(m', n'-m', m_0, n_0-m_0)\,p^N_{m,n-m}(m', n'-m', m_0, n_0-m_0),$$

где $\widehat{r}(i, j, i_0, j_0) = i_0\cdot I(j \ge i) + j_0\cdot I(i \ge j) + \tfrac{1}{2}\big(j_0\cdot I(j = i+1) + i_0\cdot I(i = j+1)\big)$,

а $p^N_{i,j}(i', j', i_0, j_0)$ – вероятность того, что в «ячейке», содержащей $i$ объектов первого и $j$ объектов второго класса, при генерировании bootstrap выборки окажется $i'$ и $j'$ точек первого и соответственно второго класса, и при этом $i_0$ и соответственно $j_0$ из исходных объектов не будут выбраны ни разу (по ним будет проводиться контроль). Данная вероятность может быть вычислена рекуррентно:

$$p^0_{i,j}(i', j', i_0, j_0) = I(i' = j' = i_0 = j_0 = 0),$$

$$p^{N+1}_{i,j}(i', j', i_0, j_0) = p^N_{i,j}(i'-1, j', i_0, j_0)\tfrac{i-i_0}{N} + p^N_{i,j}(i'-1, j', i_0-1, j_0)\tfrac{i_0}{N} +$$

$$+ p^N_{i,j}(i', j'-1, i_0, j_0)\tfrac{j-j_0}{N} + p^N_{i,j}(i', j'-1, i_0, j_0-1)\tfrac{j_0}{N} + p^N_{i,j}(i', j', i_0, j_0)\tfrac{N-i-j}{N}.$$

Комбинированная bootstrap оценка есть $\ddot{x}_{mn} = e^{-1}\cdot\widetilde{x}_{mn} + \left(1-e^{-1}\right)\cdot\widehat{x}_{mn}$.

Приведем численные результаты для $N = 50$, $k = 10$.

В таблице 1 приведены значения оценки $x^*_{mn}$, в таблице 2 — оценки $\ddot{x}_{mn}$. Видим, что при $n = 5$, что является наиболее вероятным числом выборочных точек в ячейке, оценки очень близки. При других значениях $n$ различие более существенно, любопытным представляется отрицательные значения $x^*_{mn}$ вклада в оценку вероятности ошибки для ячеек с большим числом точек и нулевым числом ошибок на обучении.

Таблица 1. Некоторые значения $x^*_{mn}$.

| $n$ | $m$ | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 |
| 0 | 2,21 | | | | | |
| 1 | 0,96 | 0,96 | | | | |
| 2 | 0,65 | 2,67 | 0,65 | | | |
| 3 | 0,41 | 1,89 | 1,89 | 0,41 | | |
| 4 | 0,21 | 1,59 | 3,35 | 1,59 | 0,21 | |
| 5 | 0,03 | 1,31 | 2,79 | 2,79 | 1,31 | 0,03 |
| 6 | –0,16 | 1,06 | 2,57 | 4,02 | 2,57 | 1,06 |
| 7 | –0,36 | 0,83 | 2,33 | 3,61 | 3,61 | 2,33 |
| 8 | –0,55 | 0,61 | 2,08 | 3,45 | 4,68 | 3,45 |

Таблица 2. Некоторые значения $\ddot{x}_{mn}$.

| $n$ | $m$ | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 |
| 0 | 0,00 | | | | | |
| 1 | 0,32 | 0,32 | | | | |
| 2 | 0,23 | 1,41 | 0,23 | | | |
| 3 | 0,12 | 1,59 | 1,59 | 0,12 | | |
| 4 | 0,054 | 1,53 | 2,54 | 1,53 | 0,05 | |
| 5 | 0,022 | 1,39 | 2,75 | 2,75 | 1,39 | 0,02 |
| 6 | 0,0087 | 1,26 | 2,73 | 3,65 | 2,73 | 1,26 |
| 7 | 0,0032 | 1,16 | 2,60 | 3,87 | 3,87 | 2,60 |
| 8 | 0,0011 | 1,09 | 2,44 | 3,88 | 4,74 | 3,88 |

Значения всех оценок при $n = 5$ приведены на рис. 1. Цифрами обозначены: 1 – эмпирический риск, 2 – скользящий экзамен, 3 – комбинированная bootstrap оценка, 4 – оптимизированная оценка $x^*_{mn}$.

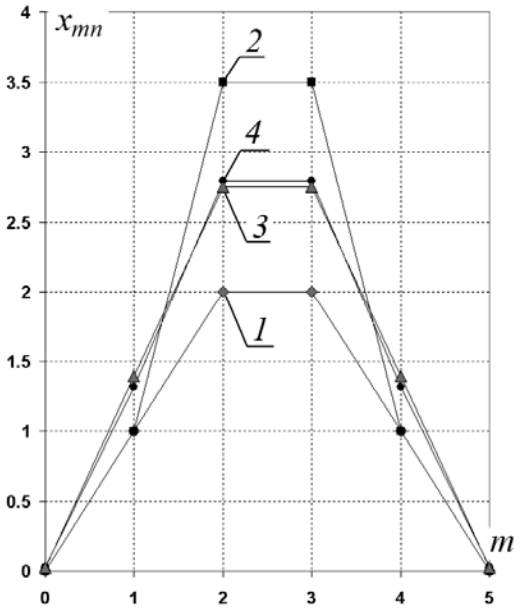



Рис. 1. Различные функции оценки риска.

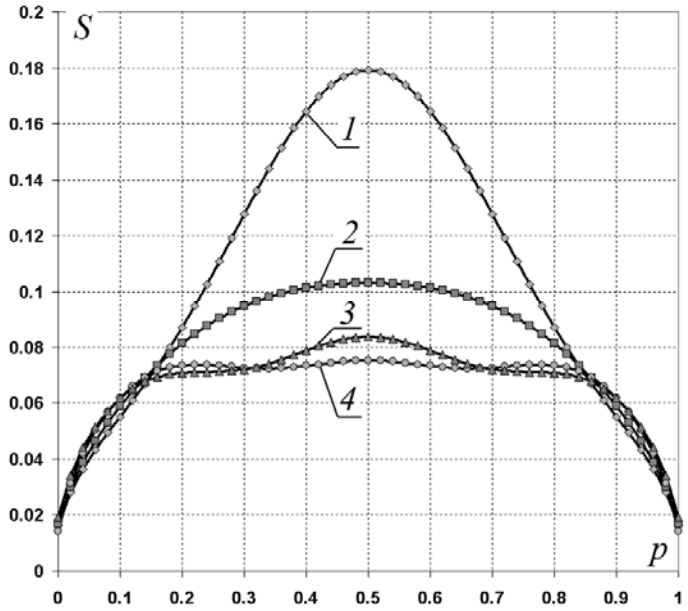



Рис. 2. Среднеквадратичная погрешность оценок.

На рис. 2 при различных значениях параметра $p$ приведены графики среднеквадратичной погрешности $S = \sqrt{\Delta_{fg}}$ для всех оценок, нумерация такая же, как на рис. 1.

Из рассмотренных оценок ни одна не доминирует другую, т.е. для каждой пары оценок существуют $p$, при которых лучше как одна, так и другая. Однако в количественном отношении различие качества при разных $p$ не равноценно. Так эмпирический риск имеет небольшое преимущество при малых $p$, но существенно проигрывает другим оценкам при $p$ в окрестности 0,5. Сбалансировано-оптимальная оценка $x^*_{mn}$ выглядит действительно наилучшей, при этом комбинированная оценка bootstrap очень близка к ней.

## Заключение

В работе рассмотрена задача построения оценки вероятности ошибочной классификации в дискретном пространстве переменных, которая была бы в каком-то смысле наилучшей при различных предположениях о распределениях. Предложен метод решения данной задачи, основанный на построении сбалансировано-оптимальной оценки.

Как показывают численные эксперименты, такая оценка оказывается близкой к оценке, получаемой методом bootstrap. Это позволяет сделать предположение о том, что метод bootstrap в некотором смысле близок к наилучшему способу оценивания вероятности ошибочной классификации. Для проверки данного предположения требуются дополнительные исследования, в частности, нужно построить оценку, оптимизированную по всем распределениям в дискретном пространстве, а не только по заданному их подклассу. Также открытым является вопрос о распространении выводов, полученных при анализе задачи классификации в дискретном пространстве, на непрерывный случай.

## Благодарности

Работа выполнена при поддержке РФФИ, гранты 07-01-00331-а и 08-01-00944-а.

## Литература

[Лбов, 1965] Лбов Г.С. Выбор эффективной системы зависимых признаков. // Выч. системы, вып. 19, Новосибирск, 1965, с. 21–34.

[Вапник, Червоненкис, 1974] Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. М.: Наука, 1974. 415 с.

[Лбов, Старцева, 1999] Г.С. Лбов, Н.Г. Старцева. Логические решающие функции и вопросы статистической устойчивости решений. Институт математики СО РАН, Новосибирск, 1999, 211 с.

[Неделько, 2003] V. M. Nedelko. Estimating a Quality of Decision Function by Empirical Risk // LNAI 2734. Machine Learning and Data Mining in Pattern Recognition. Third International Conference, MLDM 2003, Leipzig. Proceedings. Springer-Verlag. pp. 182–187.

[Неделько, 2007] Неделько В.М. Об эффективности функционалов эмпирического риска и скользящего экзамена как оценок вероятности ошибочной классификации // Proc. of int. conference, KDS'2007. Sofia. 2007. Vol. 1, P. 111–117.

[Неделько, 2008] V. M. Nedel'ko. Empirical bounds for misclassification probability // 9-th Int. Conf. "Pattern Recognition and Image Analysis: New Information Technologies" (PRIA–9–2008): Conference Proceedings. Vol. 2. – Nizhni Novgorod, 2008. P. 84–87.

[Braga–Neto, Dougherty, 2005] Braga–Neto. U. and Dougherty E.R. Exact performance of error estimators for discrete classifiers. // Pattern Recognition, Elsevier Ltd. 2005. V. 38, N 11. P. 1799-1814.

## Информация об авторе

***Виктор Михайлович Неделько*** *– с.н.с. лаборатории Анализа данных Института математики СО РАН, 630090, пр-т Коптюга, 4, Новосибирск, Россия, e-mail: nedelko@math.nsc.ru*

# КЛАССИФИКАЦИЯ И МОДЕЛИРОВАНИЕ ГЕНЕТИЧЕСКОГО КОДА И ГЕННО-НЕЙРОННЫХ СЕТЕЙ

**Адиль Тимофеев**

***Аннотация**: Предлагаются методы автоматической классификации и моделирования генетического кода. Излагаются принципы проектирования и результаты использования гетерогенных генно-нейронных сетей.*

***Ключевые слова**: модели генетического кода, генетическая информатика, гетерогенная генно-нейронная сеть.*

***ACM Classification Keywords**: E.4 Coding and Information Theory*

***Conference**: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

## Введение

Основным “строительным материалом” живых организмов являются белки, включающие в себя 20 основных аминокислот. При биохимическом синтезе белков организма используется генетическая информация, закодированная в главном “наследственном материале” – дезоксирибонуклеиновой кислоте (ДНК) [1].

В 1953 г. Дж.Уотт и Ф.Крик описали структуру ДНК и высказали гипотезу о генетическом коде и механизме самовоспроизведения ДНК [2]. За это открытие авторы были удостоены Нобелевской премии.

ДНК является полимером и представляет собой цепочки мономеров определенных типов, образующих “двойную спираль” [1–3]. В состав ДНК может входить только четыре типа оснований: адеин (A), тимин (T), гуанин (G), цитозин (C). Цепи оснований ДНК всегда соединены по принципу комплементарности (взаимодополнительности): A связано с T, а G – с C. Таким образом, водородные связи между основаниями A и T, G и C определяются “правилом комплементарности” [1–3].

Комплементарность оснований в двух цепях ДНК создает основу для репликации, т.е. самовоспроизведения ДНК. Действие этого механизма проявляется в раскручивании “двойной спирали” ДНК, после чего в точках разветвления пристраиваются соответствующие новые основания. В результате ДНК самоудваивается.

Наряду с процессом репликации ДНК происходит процесс транскрипции, т.е. перенос генетической информации с ДНК на рибонуклеиновую кислоту (РНК). Основное отличие “информационной” РНК от порождающей ее ДНК заключается в том, что вместо основания T включается основание U. В процессе транскрипции происходит “перекодировка” информации с преобразованием оснований $T \to U$.

Таким образом, транскрипция обеспечивает перенос генетической информации с ДНК на РНК. Размеры “информационной” РНК невелики по сравнению с размерами “родительской” ДНК [1,3].

## 1. Генетический язык: алфавиты, слова и семантика

Три рядом стоящих основания в ДНК соответствует только одной аминокислоте. Последовательность из трех оснований называется триплетом или кодоном. Поэтому любая цепь ДНК является последовательностью кодонов, начинающейся с определенного “стартового” участка.

Линейное расположение оснований в ДНК позволяет ввести простой “генетический язык” для кодирования и описания “наследственного материала”. Алфавитом этого языка является следующий набор букв

$$\alpha_D = \{A, C.G, T\}. \quad (1)$$

Будем называть этот набор букв (1) алфавитом оснований ДНК. Словами в этом алфавите являются «осмысленные» последовательности букв. Такими словами служат записи кодонов – упорядоченных троек (триплетов) из оснований ДНК, кодирующих некоторую аминокислоту. Очевидно, что в рассматриваемом алфавите (1) можно составить $4^3 = 64$ различных комбинаций слов из трех букв. Полезно также ввести «стартовый» участок цепей ДНК и «стоп-кордоны», обозначающие конец цепи.

Процесс записи последовательности слов, соответствующих рассматриваемой цепи ДНК, целесообразно начать с «пустого слова» (не пишется ничего), обозначающего “начало отсчета”, затем к нему справа приписывается первая буква, к ней приписывается вторая буква и т.д. до конца цепи, обозначенного одним из “стоп-кодонов”. При этом не используются никакие “знаки препинания”. В результате получается предложение вида

AGTCCATGGTAC (2)

Каждому предложению, описывающему одну цепь ДНК, однозначно соответствует комплементарное (взаимодополняющее) предложение, описывающее другую цепь ДНК. Например, для предложения (3) оно имеет вид

TCAGGTACCATG (3)

Генетическая информация, содержащаяся в кодонах ДНК, сначала “переписывается” в соответствующие кодоны “информационной” РНК. Эта РНК синтезируются в процессе транскрипции . В результате этого кодоны ДНК преобразуются в кодоны РНК. Алфавит оснований ДНК (2) порождает алфавит оснований РНК вида

$$\alpha_R = \{A, C.G, U\}. \quad (4)$$

Фрагменту цепи ДНК (2), записанной в алфавите (1), соответствует следующее описание синтезированной РНК

*AGUCCUGGUAC* . . . (5)

записанная в новом алфавите (4).

Слова в алфавитах (1) или (4), т.е. кодоны ДНК и РНК, могут быть графически равными (если они составлены из одинаковых букв, расположенных одинаковым образом) или графически различными (в противном случае). В каждом слове содержится информация о соответствующей аминокислоте, а в каждом предложении - информация о типе и последовательности аминокислот, синтезированных с помощью РНК.

Таким образом, каждое предложение в алфавитах оснований (1) или (4) содержит генетическую информацию, определяющую специфику организма с данной ДНК, т.е. его «унаследованную индивидуальность».

Белки организмов обычно состоят из 20 типов аминокислот. Поэтому любой белок можно закодировать словами, состоящими из последовательности букв алфавита аминокислот вида

$$\alpha_a = \{a_1, a_2, ..., a_{20}\}. \quad (6)$$

Буквы этого алфавита обозначают следующие аминокислоты:

$a_1$ – фенилаланин (Phe), $a_2$ – лейцин (Leu), $a_3$ – изолейцин (Ile), $a_4$ – метионин (Met), $a_5$ – валин (Val), $a_6$ – серин (Ser), $a_7$ – пролин (Pro), $a_8$ – треонин (Thr), $a_9$ – аланин (Ala), $a_{10}$ – тирозин (Tyr), $a_{11}$ – гистин (His), $a_{12}$ – глютамин (Giln), $a_{13}$ – аспарагин (Asn), $a_{14}$ – лицин (Lys), $a_{15}$ – аспарагиновая

кислота (Asp), $a_{16}$ – глютаминовая кислота (Giu), $a_{17}$ – цистеин (Cys), $a_{18}$ - триптофан (Trp), $a_{19}$ – аргинин (Arg), $a_{20}$ – глицин (Gly).

## 2. 3D-геометрическая и 2D-графовая модели генетического кода

В результате сложных биохимических исследований был установлен генетический код, т.е. соответствие между алфавитом аминокислот (6) и алфавитом оснований "информационной" РНК (4) [2,3]. Он состоит из 61 кодона, соответствующих 20 аминокислотам. Генетический код является вырожденным в том смысле, что одному типу аминокислоты может соответствовать несколько слов - синонимов (кодонов) в алфавите оснований РНК (4).

Наряду с классической табличной моделью генетического кода, полезна его трехмерная геометрическая модель типа «гиперкуб», предложенная автором в [4]. Каждому узлу этой 3D-модели соответствует аминокислота с соответствующим номером, а его проекции определяют кодон генетического кода.

Весьма удобной и полезной является также предложенная автором в [4] графовая модель представления генетического кода, Это новое 2D-представление генетического кода в виде графа (кодирующего дерева) имеет ряд общих черт с известной генетике "круговой диаграммой", описанной в [3].

Табличную, 3D-геометрическую и 2D-графовую модели генетического кода можно одинаково успешно применять для расшифровки ДНК и РНК растений, животных и человека.

## 3. Передача генетических сообщений

Рассмотрим алфавит $\alpha_X = \{x_1,...,x_4\}$, буквы которого совпадают с буквами алфавита оснований ДНК (1). Словом в этом алфавите будем называть последовательность из трех букв вида

$$X = x_{i1}x_{i2}x_{i3}\ ,\ x_{ij} \in \alpha_X\ . \qquad (7)$$

Обозначим через $S(\alpha_X)$ множество всех триплетных слов вида (7), а через $S'(\alpha_X)$ - подмножество слов из $S(\alpha_X)$, имеющих "генетический смысл", т.е. обозначающих соответствующие аминокислоты.

Объект, порождающий "осмысленные" слова из $S'$, называется в теории кодирования источником сообщения, а слова из $S'$ - сообщениями. В роли источника сообщения в генетике выступает ДНК. Описание дополнительной информации о ДНК как источнике сообщений может задаваться различными способами:

1) теоретико-множественное описание мощности, т.е. числа элементов, и других характеристик множеств $\alpha_X$, $S(\alpha_X)$, $S'(\alpha_X)$. Для ДНК мощность этих множеств определяется соотношениями

$$|\alpha_X| = 4,\ |S| = 4^3 = 64,\ |S'| = 61;$$

2) статистическое (частотное) описание осуществляется заданием вероятностей (частот) характеристик $\alpha_X$, $S(\alpha_X)$, $S'(\alpha_X)$. Для ДНК могут быть известны, например, вероятности (частоты) появления букв $p_1 = p(A),\ p_2 = p(C),\ p_3 = p(G),\ p_4 = p(T),$ соответствующих основаниям из алфавита (1);

3) логическое описание множеств с помощью языка исчисления двузначных или многозначных предикатов.

Пусть задан также алфавит $\alpha_Y = \{y_1,...y_4\}$, буквы которого совпадают с буквами алфавита оснований РНК (4). Через Y обозначим триплетное слово в алфавите $\alpha_Y$, а через $S(\alpha_Y)$ - множество всех слов в этом алфавите.

Генетическое преобразование (1) при транскрипции задает отображение F, которое каждому слову $X \in S'(\alpha_X)$, т.е. кодону ДНК, однозначно ставит в соответствие слово

$$Y = F(X) = y_{i1}y_{i2}y_{i3}, \quad Y \in S(\alpha_Y), \tag{8}$$

являющееся кодоном синтезированной “информационной” РНК. Слово (8) будем называть кодом сообщения X при синтезе РНК, а переход от слова X к слову Y кодированием ДНК в структуре РНК. Этот переход, происходящий в процессе транскрипции, можно интерпретировать, как передачу наследственной информации из “постоянной” памяти ДНК в “оперативную” память РНК.

Код РНК- сообщения Y подается в “рибосомный” канал связи и синтеза белков. Однако код сообщения Y’ на выходе канала связи может отличаться от входного кода Y. Источником искажения могут быть мутации генов, ошибки “считывания” кода и т.п.

В случае идеального канала связи передача генетической информации не искажается, т.е. $Y' = F(X) = Y$.

Поэтому возможно точное декодирование “генетического сообщения”, если существует обратное отображение $F^{-1}$ для (8). В случае, когда генетическая информация искажается в канале связи и при синтезе белков, включается система “репарации” (коррекции), обеспечивающая обнаружение и исправление ошибок.

Различные слова (8), являющиеся кодонами РНК, можно закодировать различными буквами $a_1, a_2, ..., a_{20}$ алфавита аминокислот (6). Этот натуральный генетический код является вырожденным, поскольку он не удовлетворяет требованию взаимной однозначности. Однако его избыточность значительно повышает надежность передачи генетической информации.

## 4. Мера и оценка генетической информации

В генетике важную роль играют дискретные формы кодирования, хранения и передачи “наследственной информации”. Поэтому естественно определить “количество” генетической информации в терминах двоичных знаков, т.е. в битах. При этом целесообразно исходить из комбинаторного подхода к теории информации, предложенного А.Н. Колмогоровым [5]. Этот подход обобщает вероятностный подход, развитый К.Шенноном.

Обозначим основание ДНК или РНК переменной *x*. Эта переменная может принимать значения, принадлежащие конечным алфавитам оснований ДНК (1) или РНК (4), которые состоят из четырех элементов. Поэтому “энтропия” H основания *x* в ДНК или РНК равна

$$H(x) = \log 4 = 2 . \tag{9}$$

В каждой ДНК или РНК основание *x* имеет определенное значение (например, *x* = A). Это означает, что каждое основание в цепи ДНК или РНК сообщает генетическую информацию, равную

$$I_x = H(x) = \log 4 , \tag{10}$$

и требует для своего описания 2 двоичных знака. При этом “снимается” энтропия, т.е. априорная неопределенность этих знаков.

Аналогично обозначим через Y произвольный кодон ДНК или РНК. Число различных кодонов, которые можно формально образовать в алфавите оснований ДНК или РНК, равно $M = 4^3 = 64$. Однако в природном генетическом коде обычно содержится $M_\Gamma = 61$ кодонов. Поэтому “количество генетической информации”, содержащейся в определенном кодоне Y ДНК или РНК, равно

$$I_Y = \log 61 < \log 64 . \tag{11}$$

Для записи любого кодона Y в двоичной системе требуется 6 двоичных знаков.

ДНК и РНК конкретных организмов имеет определенную длину L, равную числу оснований в цепи. Различные основания встречаются в этой цепи с различной частотой. Обозначим через $n_1, n_2, n_3, n_4$ число вхождений соответствующих оснований (например, U,C,A,G в алфавите оснований РНК) в цепь

длины L. Тогда, учитывая, что $L = n_1 + n_2 + n_3 + n_4$, легко подсчитать общее число возможных цепей длины L по формуле

$$R = \frac{L!}{n_1! n_2! n_3! n_4!} . \quad (12)$$

Количество генетической информации в цепи ДНК или РНК длины L, закодированной в соответствующем алфавите оснований, будет равно

$$I_L = \log R . \quad (13)$$

При больших длинах L, характерных для ДНК и РНК, при вычислении (14) можно воспользоваться формулой Стирлинга:

$\log(L!) \sim L \log L$.

Тогда получим следующую приближенную формулу

$$I_L \sim -L \sum_{i=1}^{4} p_i \log p_i , \quad (p_i = \frac{n_i}{L}) \quad (14)$$

Отсюда следует, что, если в цепи ДНК или РНК основания встречаются с частотами $p_i$, то количество генетической информации, приходящейся на одно основание, равно

$$H = -\sum_{i=1}^{4} p_i \log p_i \quad (15)$$

В случае равных частот $p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$ из (15) вновь получим формулу (9). При любых других соотношениях частот встречаемости оснований в цепи ДНК или РНК справедливо неравенство

$H < \log 4$.

Следовательно, для передачи “генетического сообщения” длины L достаточно употребить примерно L H двоичных знаков, не превышающее 2L.

Интересно также оценить количество “генетической информации”, содержащейся в переменной a из алфавита аминокислот $\alpha_a$ относительно связанных с ней кодонов Y генетического кода. Связь между переменными a и Y заключается в том, что генетический код допускает не все формально возможные пары (a,Y). Однако для любой аминокислоты $a \in \alpha_\alpha$ можно найти все кодоны, допускаемые генетическим кодом.

Тогда генетическую информацию в a относительно Y можно определить по формуле

$$I_a = H(Y) - H(Y/a) , \quad (16)$$

где $H(Y/a) = \log M_a, M_a$ - число кодонов генетического кода для a. Например, если $a = a_2$, то $M_a = 6$, и, следовательно, $I_a = \log 61 - \log 6$. Если же $a = a_4$, то $M_a = 1$ и $I_a = \log 61$.

## 5. Информационная сложность РНК

Генетическая информация тесно связана со “сложностью” ее носителя. Если этот носитель устроен “просто” (например, кодон), то для его описания достаточно небольшого количества информации. Для сложных “носителей” (например, для РНК) требуется много информации для его описания и передачи.

Стандартным способом описания “информационной” РНК является последовательность оснований Z в четырехбуквенном алфавите оснований (5). Поставим в соответствие рассматриваемой РНК некоторое

число n = *f* (Z). Например, это может быть десятичное число, определенное по двоичному представлению Z. Обозначим через *l* (Z) наименьшую длину цепочки оснований, определяющей данную (неизбыточную) РНК.

"Сложностью" РНК при способе ее задания с помощью f будем называть величину

$$K_f(\omega) = \min l(Z) \text{ при } f(Z) = n(\omega). \quad (17)$$

На генетическом языке это определение "сложности" РНК можно проинтерпретировать следующим образом. Конкретная цепочка оснований Z определяет "генетическую программу" синтеза белков, а оператор *f* – способ задания РНК. Тогда естественно считать, что $K_f(\omega)$ есть наименьшая длина "генетической программы", с помощью которой можно синтезировать объект $\omega$ при способе задания *f*.

Задание какого-либо носителя "генетической информации" (например, РНК) можно упростить, если уже задан какой-то другой объект (например, кодон). Для этого введем показатель "условной сложности" объекта $\omega$ при заданном объекте Z. Следуя А.Н. Колмогорову [5], определим этот показатель в виде

$$K_f(\omega / Z) = \min l(Z) \text{ при } f(n(\omega), Z) = n(\omega). \quad (18)$$

Здесь способ задания f является функцией от номера объекта Z и номера "генетической программы" вычисления $n(\omega)$ при заданном объекте Z.

Если "условная сложность" значительно меньше, чем безусловная, т.е.

$$K_f(\omega / Z) << K_f(\omega),$$

то естественно считать, что в объекте Z содержится значительная "генетическая информация" об объекте $\omega$. Количество этой условной информации зададим формулой

$$I_f(\omega / Z) = K_f(\omega) - K(\omega / Z). \quad (19)$$

В частном случае, когда $K_f(\omega / \omega) = 0$, получаем $I_f(\omega / \omega) = K_f(\omega)$.

В этом случае "информационная сложность" объекта $\omega$ совпадает с его "генетической информацией" о себе самом.

Важными достоинствами предложенных определений меры "генетической информацией" и "информационной сложности", является то, что они относятся к индивидуальным объектам, т.е. к конкретным кодонам, хромосомам, РНК и т.п. Однако их можно с одинаковым успехом использовать и в тех случаях, когда заданы вероятностные или частотные характеристики рассматриваемых объектов.

## 6. Генетические базы данных и знаний

Каждый ген, управляя синтезом белка, определяет некоторый элементарный признак организма. Множество признаков, характеризующих различные виды организмов, удобно представить в виде реляционной базы данных (БД) табличного типа. При формировании генетической БД каждому признаку ставится в соответствие "домен", т.е. множество дискретных значений признака. Ген, порождающий признак, может находиться в одном из возможных альтернативных состояний, определяемых аллелями. Например, у каждого кролика имеется ген, определяющий признак окраса его меха. Принято подразделять окрасы на "шиншиловый", "дикий тип", "альбинос" и "гималайский", что соответствует четырем аллелям.

Сложные признаки определяются хромосомой, состоящей из набора генов $x_1,...,x_n$. Число хромосом у каждого вида организмов фиксировано и равно 2n, где n - гаплоидное число, являющееся инвариантом данного вида. Например, у человека n=23, а у краба n=127. Поэтому в генетическую БД человека включается 23 отношения, а в БД краба - 127 отношений.

Процессу мейоза в генетической БД соответствует процесс соединения всех отношений, т.е. образуется их прямое произведение.

Генетическая БД является хранилищем индивидуальной информации, передаваемой от родителей к потомкам. Однако эта информация допускает обобщенное представление в виде “генетических знаний”.

В отличие от “индивидуальных данных”, закодированных в ДНК и РНК в алфавите оснований или аминокислот в форме длинных последовательностей слов (предложений), “обобщенные” знания представляют собой “высказывания” в терминах многозначных предикатов, которые являются истинными по крайне мере на всех “предложениях” БД. Совокупность этих “высказываний” образует генетическую базу знаний (БЗ).

Для автоматического синтеза БЗ по заданной генетической БД и минимизации ее сложности (без потери “генетической информации”) можно использовать логико-аксиоматический и логико-вероятностный методы синтеза решающих правил, предложенные автором в [6-8]. Совокупность этих правил (“генетических высказываний”) записывается в терминах логических или многозначных предикатов, связанных с алфавитом оснований или аминокислот, и обладает необходимыми свойствами полноты и непротиворечивости при описании генетической БД.

## 7. Когнитивные модели генетического кода и генетические алгоритмы

Применение логико-вероятностного метода оптимального синтеза генетических БЗ к генетической БД, представляющую собой классическую табличную модель генетического кода [2], позволяет автоматически построить когнитивную модель генетического кода [8,11]. Эта модель в виде классифицирующего дерева аминокислот минимальной сложности представлена автором в [4]. Каждый путь на этом дереве с вероятностью 1 описывает соответствующую аминокислоту в виде логического “высказывания” определённого типа.

Примером могут служить следующие “генетические правила” классификации вида

1) ЕСЛИ 2 основание = A И 1 основание = C И 3 основание = A
   ИЛИ 2 основание = A И 1 основание = C И 3 основание = G
   ТО аминокислота $a_{12}$ [Giln]
2) ЕСЛИ 2 основание = U И 1 основание = A И 3 основание = G
   ТО аминокислота $a_4$ [Met].

Методы математического моделирования и вычислительного эксперимента играют важную роль в генетических исследованиях. Они позволяют формализовать генетические механизмы в виде математических и информационных моделей, генетических БД и БЗ и т.п. Учет биологических принципов обработки информации позволяет создавать генетические алгоритмы и развивать теорию клеточных автоматов, нейронных сетей и т.п.

В последние годы сформировались новые разделы генетики - математическая генетика и генетическое программирование [6,7]. В их основе лежит оригинальный математический аппарат и програмное обеспечение. Этот новый инструментарий ориентирован не только на собственно генетические исследования, но и на решение широкого класса задач дискретной оптимизации, эволюционного моделирования и т.п.

Сегодня генетические алгоритмы успешно используются для оптимизации расписаний, планирования поведения, оптимальной трассировки компьютерных плат, автоматического управления нелинейными процессами и т.п. [6,7]. Они особенно эффективны в многоэкстремальных задачах, связанных с поиском глобального экстремума. Весьма перспективно использование таких алгоритмов для управления генетическими БД и БЗ и обучения нейронных сетей на основе принципов самоорганизации и естественного отбора наилучших архитектур [8–13].

Отличительными чертами генетических алгоритмов является их разветвленность и параллелизм, связанные с использованием “вычислительных популяций”, целенаправленная “селекция” с

“наследованием” наиболее важных признаков или фрагментов промежуточных результатов, многовариантное сравнение, “естественный отбор” наилучших решений и т.п. В этих алгоритмах используются принципиально новые вычислительные операторы. Примерами таких операторов, не имеющих аналогов в традиционных вычислительных моделях, являются нелинейные преобразования типа “мутации”, “инверсии” и “кроссинговера” [6,7].

## 8. Модели гетерогенных генно-нейронных сетей

Нейронные сети (НС) и нейросетевые технологии являются одним из наиболее эффективных средств массового распараллеливания и ускорения процессов обработки знаний и передачи потоков данных в задачах распознавания образов, классификации данных и диагностики состояний. Естественным прототипом искусственных НС является биологический мозг и центральная нервная система человека и животных как сложная гетерогенная нейронная сеть, обеспечивающая за счёт естественных био- и нанотехнологий высокую степень параллелизма, адаптации, самоорганизации и робастности при решении различных интеллектуальных задач (представление знаний, распознавание образов, классификация данных, поиск закономерностей, анализ изображений, диагностика состояний, прогнозирование явлений и т.п.). Возможности искусственных и биологических НС могут значительно расшириться при коллективном (мульти-агентном) решении сложных интеллектуальных задач.

Высокая сложность и размерность многих задач распознавания образов, классификации данных, анализа изображений и диагностики состояний, а также часто возникающая необходимость их решения в реальном времени требуют массового параллелизма и самоорганизации распределённых вычислений на базе НС. С этой точки зрения особый интерес и дополнительные возможности представляют гетерогенные полиномиальные нейронные сети (ПНС) с самоорганизующейся архитектурой и генно-нейронные сети (ГНС) [8-13].

Основные идеи, математические модели, методы оптимизации, алгоритмы обучения и принципы самоорганизации ПНС и ГНС были предложены автором в работах [8-13]. Они заключаются в следующем:

- архитектура НС гетерогенна и многослойна;

- наличие слоя полиномиальных нейронных элементов (П-нейронов);

- возможность обучения и адаптации НС к обучающим базам данных (БД);

- целесообразность самоорганизации и минимизации сложности архитектуры НС различных типов в процессе обучения;

- детерминированные, логические и вероятностные методы обучения и самоорганизации гетерогенных НС с самоорганизующейся архитектурой;

- принцип высокой экстраполяции (экстраполирующей силы) гетерогенных НС ;

- алгебраическое требование диофантовости (целочисленности синаптических весов) гетерогенных НС.

В процессе дальнейшего развития теории гетерогенных ПНС и ГНС были предложены модели многозначных нейронных элементов (М-нейронов) и связанных с ними конъюнктивных, полиномиальных, дизъюнктивных и суммирующих нейронных элементов (МК-, МП-, МД- и М$\Sigma$-нейронов), а также новые разновидности гетерогенных ПНС (генно-нейронные сети, квантовые нейронные сети, мульти-агентные ПНС и т.п.).

Предложенные гетерогенные модели и быстрые алгоритмы обучения ПНС и ГНС разных типов обеспечивают высокий параллелизм и самоорганизацию нейровычислений в процессе решения многих интеллектуальных задач. Они успешно применялись для решения ряда прикладных задач распознавания образов (распознавание кораблей по отраженным радиолокационным сигналам, распознавание команд и дикторов по видеограммам речи, распознавание и адресация деталей на конвейере, классификация дорожных ситуаций и т.д.), медицинской диагностики (диагностика и оценка эффективности лечения

артритов, векторная диагностика и расшифровка гастритов и т.д.), прогнозирования явлений (прогнозирование градоопасности облаков и исхода черепно-мозговых травм и т.д.) и нейросетевого представления генетического кода [4,8–13].

## Заключение

Бурное развитие генетики и теории биологической эволюции привело к созданию новых научных направлений, связанных с разработкой “генетических алгоритмов”, “генетического программирования”, “эволюционного моделирования” и “генной инженерии”. Генетические принципы и механизмы породили новые подходы в теории кодирования и передачи информации, теории алгоритмов и теории автоматов. Они оказали глубокое влияние на компьютерную информатику и программирование.

Значительный интерес представляет использование принципов генетики и нейрофизиологии в теории нейронных сетей и нейрокомпьютеров, а также моделирование генно-нейронных сетей и их реализация на базе нанотехнологий.

## Благодарности

Работа выполнена при частичной поддержке **грантов РФФИ № 08–08–12183-офи** и **№ 09–08–00767-а** и **Программы № 1** Президиума РАН.

## Литература

[1]. Айала Ф., Кайгер Дж. Современная генетика (М.: Мир, 1968).

[2]. Уотсон Дж. Молекулярная биология (М.: Мир 1979).

[3]. Инге-Вечтомов С.Г. Генетика с основами селекции (М.: Высшая школа, 1988)

[4]. Тимофеев А.В. Генетическая информация и национальный генотип. – В книге: В поисках парадигмы нации (нацио-логические очерки). Очерк 7. Москва - Нальчик, Изд.-во АМАН, 1997, с. 188-223.

[5]. Колмогоров А.Н. Теория информации и теория алгоритмов (М.: Наука, 1987)

[6]. Goldberg D.E. Genetic Algoritms in Search, Optimization and Mashine Learning (Addision - Wesley, 1989).

[7]. Koza J.R. Genetic Programming (Bradford/MIT Press, 1992).

[8]. Тимофеев А.В.Адаптивные робототехнические комплексы (Л.: Машиностроение, 1988).

[9]. Каляев А.В.,Тимофеев А.В. Методы обучения и минимизации сложности когнитивных нейромодулей супер-макро-нейрокомпьютера с программируемой архитектурой. - Доклады АН, 1994, т.337, №2, с.180-183.

[10]. Тимофеев А.В. Методы синтеза диофантовых нейронных сетей минимальной сложности. - Доклады АН, 1995, т.337, № 1, с.32-35.

[11]. Timofeev A.V. Intelligent Control Applied to Non-Linear Systems and Neural Networks with Adaptive Architecture. - Journal of Intelligent Control, Neurocomputing and Fuzzy Logic, 1996, v.1, № 1, pp.1-18.

[12]. Тимофеев А.В. Оптимальный синтез и минимизация сложности генно-нейронных сетей по генетическим базам данных — Нейрокомпьютеры: разработка и применение, 2002, № 5-6, с. 34-39.

[13]. Timofeev A. V. Parallel Structures and Self-Organization of Heterogeneous Polynomial Neural Networks for Pattern Recognition and Diagnostics of States. – Pattern Recognition and Image Analysis, 2007, Vol. 17, No. 1, pp. 163–169.

## Информация об авторе

***Тимофеев Адиль Васильевич –*** *заведующий лабораторией информационных технологий в управлении и робототехнике Санкт-Петербургского института информатики и автоматизации Российской академии наук, доктор технических наук, профессор, Заслуженный деятель науки РФ, 199178, Россия, Санкт-Петербург, 14-я линия, д. 39, СПИИРАН, tav@iias.spb.su*

# Pattern Recognition and Forecasting

## "AVO-POLYNOM" RECOGNITION ALGORITHM

**Alexander Dokukin**

***Abstract**: Estimates Calculating Algorithms have a long story of application to recognition problems. Furthermore they have formed a basis for algebraic recognition theory. Yet use of ECA polynomials was limited to theoretical reasoning because of complexity of their construction and optimization. The new recognition method "AVO-polynom" based upon ECA polynomial of simple structure is described.*

***Keywords**: pattern recognition, estimates calculating algorithms, algebraic approach, recognition polynomials.*

***ACM Classification Keywords**: I.5.2 [Pattern Recognition]: Design Methodology – Classifier design and evaluation*

***Conference**: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

### Introduction

ECA or Estimates Calculating Algorithms [1] are a parametrical family of methods for pattern recognition developed in Computing Centre about thirty years ago. The idea of method is simple. Training sample is divided into two parts: actual training and check ones. Closeness to each object of training sample as well as remoteness from it is stimulated, i.e. the estimation of object S belonging to class K is increased if S is close to some representative of K or is far from a representative of K's addition. The value of increasing is determined by the representative's weight.

ECA was widely used for solving applied tasks. In addition, a number of theoretical results have been achieved for its algebraic closure. The most important of them proved existence of correct polynomial over ECA [2]. Yet there was a huge distance between theoretical reasoning and application, since former was based on polynomial constructions over ECA family, while latter on optimization of single ECA by its weights [7].

The major step in applying polynomials to the real world problems was made by reducing correct polynomial's complexity both in number of items and power. The approach was based on maximizing ECA's height, i.e. difference between minimal estimation of regular pair (object, class) and maximal estimation of irregular one [4]. A number of algorithms for minimization of ECA height have been suggested and tested, both precise [5] and approximate [8]. Either of them had a major drawbacks: precise ones being too slow for polynomial construction [6] while approximate ones not precise enough.

Nevertheless during the analysis of different combinations of methods a regularity has been noticed. ECA's of maximal height tend to have good recognition quality in some areas close to their so called center. This fact has been assumed as a basis for a novel recognition method named "AVO-polynom" that is Russian for ECA-polynomial.

## Definitions

The following recognition problem is referred to as a standard problem. We consider two samples of vectors from the n-dimensional feature space: a learning sample and a check one. For definiteness, we assume that the former sample contains m objects: $S_1, \ldots, S_m$, while the latter one contains q objects: $S^1, \ldots, S^q$. We also assume that the set of admissible objects is divided into l classes, which may intersect in general case. The classification of each object in the learning sample is known; it is necessary to reconstruct classification of the check sample.

The family of ECAs is defined as follows.

1. Each feature is ascribed a certain weight $p_i$, $i = \overline{1,n}$.

2. Certain subsets of the set of features, which are referred to as supporting subsets, are singled out. The aggregate of these subsets is denoted by $\Omega_A$. Each supporting set $\omega \in \Omega_A$ has a weight.

3. The proximity function $B_\omega(S, S')$ for two objects in the supporting set is introduced. We will use the threshold proximity function unless specially announced; i.e., two objects $S = (a_1, \ldots, a_n)$ and $S' = (b_1, \ldots, b_n)$ will be regarded close if the following inequalities hold for all supporting features:

$$\rho_i(a_i, b_i) < \varepsilon_i, \quad \forall i \in \omega.$$

Here $\varepsilon_i$, $i = \overline{1,n}$ are called the proximity function thresholds.

4. Each $S_j$ of the learning sample is ascribed its own weight $\gamma(S_j)$, $j = \overline{1,m}$.

5. The estimate of an object class is calculated by the formula

$$\Gamma_j(S^t) = x_0 \cdot \Gamma_0^j(S^t) + x_1 \cdot \Gamma_1^j(S^t),$$
$$\Gamma_0^j(S^t) = \sum_{S_i \in C\widetilde{K}_j} \gamma(S_i) \sum_{\omega \in \Omega_A} p(\widetilde{\omega}) \cdot \overline{B_\omega}(S_i, S^t),$$
$$\Gamma_1^j(S^t) = \sum_{S_i \in \widetilde{K}_j} \gamma(S_i) \sum_{\omega \in \Omega_A} p(\widetilde{\omega}) \cdot B_\omega(S_i, S^t).$$

Here, the following variables and notation are used: $x_1, x_0 \in \{0,1\}$, $\widetilde{K}_j = K_j \bigcap \{S_1, \ldots, S_m\}$, $C\widetilde{K}_j = \{S_1, \ldots, S_m\} \setminus \widetilde{K}_j$, $\overline{B_\omega}(S_i, S^t) = 1 - B_\omega(S_i, S^t)$.

The height of the ECA is defined as the difference between the minimal estimate of a regular pair (object, class) (i.e., the pair whose object belongs to the corresponding class) and the maximal estimate of an irregular pair [4].

Some changes have been made to a classical ECA optimization. First of all, optimization by objects' weights was replaced with optimization by similarity functions thresholds for better flexibility. Secondly, the optimization criterion has been changed too. Instead of recognition quality over whole check sample the height on its subset is considered. The optimization problem is reduced to the search for the values $\varepsilon^*$ of the $\varepsilon$-thresholds of the proximity function, which maximize the functional:

$$\varepsilon^* = \arg\max_{\varepsilon \in (0,\infty)^n} \left( \min_{(i,j) \in M_1} \Gamma_j(S^i) - \min_{(u,v) \in M_0} \Gamma_v(S^u) \right).$$

Here $M_1$ denotes set of regular pares and $M_0$ of irregular ones.

## "AVO-polynom"

The method has been designed to be a part of software system RECOGNITION [3] that applies some restrictions on training sequence. First of all, the input sample has to be divided into training and checking parts. By default the division is made randomly in proportion 2 to 1. This parameter is a single one which can be adjusted by user, and its default value covers most part of tested cases.

Second and the most time consuming part is devoted to finding a set of simple ECAs with better recognition quality. The input sample divided into two parts is further divided to q smaller overlapping ones. Each checking object in combination with all training ones forms a set for training simple ECA. The checking object used is referred to as central object of the ECA. The method of fastest descent [8] is then used to find ECA of maximal height. If positive height can't be achieved the central object is considered as outlier and corresponding ECA is dropped out.

The local nature of each recognition operator achieved is taken in account by dividing its contribution by distance to the central object. I.e. final estimations are calculated by formula

$$\Gamma_j(S) = \sum_{i=\overline{1,n}} \frac{\Gamma_j^i(S)}{d(S, S^i)}.$$

The second multiplier can be expressed in terms of ECA with use of specific distance functional. Thus, the whole construction represents second degree polynomial over ECA.

In the next section "AVO-polynom" will be compared to some over recognition methods. They are simple ECA [7], logical regularities and linear machine [3].This choice is not accidental. Simple ECA shows advantages of using polynomial instead of single item. Logical regularities have similar nature since it founds some typical hyper parallelepipeds in feature space. Linear machine shows results of completely different approach.

## Testing results

The testing was performed with the set of seven real world tasks from the UCI Repository of Machine Learning Databases. All samples have been pre-divided into training and testing ones. The latter was used only for quality estimation. Here is the list of used samples: Abalone, Breast-canser, Ionosphere, Echocardiogram, Hepatitis,

Image, Credit. Testing results are described in following table:

| Task | Simple ECA | Logical regularities | Linear Machine | AVO-polynom |
|---|---|---|---|---|
| Abalone | 57.3 | **-** | **65.5** | 62.3 |
| Breast canser | **96.3** | 94.1 | 95.5 | 96.1 |
| Ionosphere | 81.9 | 89.6 | 85.2 | **98.7** |
| Echocardiogram | 76.1 | 59.2 | 70.4 | **77.4** |
| Hepatitis | 79.5 | 83.1 | 78.3 | **88.0** |
| Image | 89.0 | **93.2** | 93.7 | 89.4 |
| Credit | **86.2** | 77.9 | 85.9 | **86.2** |

In general "AVO-polynom" performed on the same level with best methods, but some results deserve to be mentioned specially. For example in Abalone task the best result has been achieved with Linear Machine, but AVO-polynome has far surpassed Simple ECA and Logical regularities. In some other tasks AVO-polynom have shown simply the best results.

## Acknowledgements

The work is presented with financial support of RFBR (Projects 08-01-00636-a, 08-07-00437-a) and grant of the President of Russian Federation "Scientific School – 5294.2008.1".

## Bibliography

[1] Yu.I. Zhuravlev, Well-Posed Algebras over a Set of Ill-Posed (Heuristic) Algorithms I, Kibernetica, No. 4, 14–21 (1977).

[2] Yu.I. Zhuravlev, Well-Posed Algebras over a Set of Ill-Posed (Heuristic) Algorithms II, Kibernetika No. 6, 21–27 (1977).

[3] Yu.I. Zhuravlev, V.V. Ryazanov, O.V. Senko, RECOGNITION. Mathematical methods. Software System. Practical Solutions. (in Russian), Moscow, Phasis, 2006, ISBN 5-7036-0106-8.

[4] Yu.I. Zhuravlev, I.V. Isaev, Construction of Recognition Algorithms Correct for a Given Control Sample, Zh. Vych. Mat. Mat. Fiz. 19 (3), 726–738 (1979).

[5] A.A. Dokukin, Generalization of the Method for Constructing Maximum-Height Estimate-Calculating Algorithms to Recognition Problems, Pattern Recognition and Image Analysis, 2006, Vol. 16, No. 4, pp. 689–694.

[6] A.A. Dokukin, On complexity of searching the optimal ECA (in Russian), Reports to All-Russia Conference MMPO-12, 2006.

[7] V.V. Ryazanov, Optimization of estimates calculating algorithms by representativeness parameters of precedents (in Russian), Zh. Vych. Mat. Mat. Fiz. 16 (6), 1559--1570 (1976).

[8] A.A. Dokukin, On construction of samples for testing approximate methods for optimization of estimates calculating algorithms (in Russian), Zh. Vych. Mat. Mat. Fiz. 46 (5), 978--983 (2006).

## Authors' Information

***Alexander Dokukin*** *– Researcher; Dorodnicyn Computing Centre of Russian Academy of Sciences, 40, Vavilova St., Moscow, Russian Federation; e-mail: dalex@ccas.ru*

# СЛОЖНЫЕ ЗАДАЧИ РАСПОЗНАВАНИЯ ОБРАЗОВ И ВОЗМОЖНОСТИ ИХ РЕШЕНИЯ

**Виктор Краснопрошин, Владимир Образцов**

***Аннотация**: Рассматривается задача распознавания образов с обучением. Вводится понятие локальной разрешимости такой задачи и показано, что при некоторых, достаточно конструктивных условиях, задача распознавания является локально разрешимой. Получены критерий и два достаточных условия локальной разрешимости.*

***Ключевые слова**: Задача распознавания образов с обучением, локальный подход, критерий и достаточны условия локальной разрешимости.*

***ACM Classification Keywords:** I. Computing Methodologies; I.5 Pattern Recognition; I.5.1 Models; Subject descriptor: Models Deterministic*

***Conference**: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

## Введение

Задача распознавания образов с обучением, как и любая другая задача информатики, может оказаться сложной. Понятие сложности может быть определено по-разному. Чтобы не быть связанными конкретными свойствами задачи, мы определим сложность задачи как некоторую совокупность характеристик, следствием которых является структурируемость информации. К числу таких характеристик можно отнести, к примеру, большую размерность задачи или большой объем обучающей и контрольной выборок.

Надо заметить, что в рамках детерминистского подхода [Журавлев, 1978] вопросы сложности почти не рассматривались. Поэтому в принципиальном смысле важен следующий вопрос: *можно ли в рамках указанного подхода развить технику решения сложных задач распознавания образов*?

В данной работе показано, что ответ на сформулированный выше вопрос является положительным. Для этого нами введено понятие локальной разрешимости задачи распознавания образов с обучением и для широкого класса моделей алгоритмов определены критерий и достаточные условия локальной разрешимости. В содержательном смысле предложенный подход близок к широко используемой в математике технике, суть которой звключается в декомпозиции задачи.

Полученные результаты свидетельствуют, что понятие сложности задачи распознавания является вполне конструктивным. А т.к. практические задачи с большими размерностью и/или объемами выборок становятся все более актуальными, то и результаты решения подобных задач приобретают несомненную важность.

## Локально разрешимые задачи распознавания

Рассмотрим произвольную модель распознающих операторов $\mathcal{M}$ и некоторую задачу распознавания $Z = (I_0, \tilde{S}^q)$ из $Z_2^{ql}$ [Журавлев, 1978]. Предположим, что задано $t$ подмножеств $\tilde{S}_1^q, ..., \tilde{S}_t^q$ и $\tilde{S}_m^1, ..., \tilde{S}_m^t$ ($t \in \square$) контрольной $\tilde{S}^q$ и обучающей $\tilde{S}_m$ выборок соответственно, таких что

$$\begin{cases} (\tilde{S}_i^q \neq \varnothing, \tilde{S}_m^i \neq \varnothing)\ \forall i \in \{1,2,...,t\} \\ (\tilde{S}^q = \bigcup_{i=1}^{t} \tilde{S}_i^q, \tilde{S}_i^q \bigcap \tilde{S}_j^q = \varnothing \text{ if } i \neq j)\ 1 \le i,j \le t \\ (\tilde{S}_m^i \subseteq \tilde{S}_m)\ \forall i \in \{1,2,...,t\} \end{cases} \quad (1)$$

Информацию $Z_i = ((\tilde{S}_m^i, I(\tilde{S}_m^i)), \tilde{S}_i^q)$, $i = 1,2,...,t$ назовем **подзадачей** задачи $Z$. Нетрудно заметить, что при $t > 1$ подзадачи $Z_1,...,Z_t$ однозначно определяются подмножествами $\tilde{S}_i^q$ и $\tilde{S}_m^i$ ($i = 1,2,...,t$), удовлетворяющими условию (1). Обратное утверждение, в общем случае, неверно.

Предположим, что модель $\mathcal{M}$ представляется совокупностью параметрических функций $(\xi_\delta^{(m)}, \eta_\lambda^{(l)})$ с областью изменения параметров $\Omega \times \Lambda$. При фиксированных $(\delta, \lambda)$ набор $(\xi_\delta^{(m)}, \eta_\lambda^{(l)})$ определяет распознающий оператор $B(\xi_\delta^{(m)}, \eta_\lambda^{(l)}) \in \mathfrak{M}$, т.е.:

$$\mathfrak{M}(\xi_\delta^{(m)}, \eta_\lambda^{(l)}) = \bigcup_{\delta \in \Omega} \bigcup_{\lambda \in \Lambda} B(\xi_\delta^{(m)}, \eta_\lambda^{(l)}) \quad (2)$$

Используя условие (1) каждой подзадаче $Z_i$ ($i = 1,2,...,t$) и модели $\mathcal{M}$ можно поставить в соответствие **подмодель** $\mathfrak{M}_i$, в которой набор функций (определяющий ее в указанном выше смысле) является сужением [Мальцев, 1970] исходного набора на подмножество $\tilde{S}_m^i$ ($i = 1,2,...,t$). Полученные при этом подмодели будем называть **локальными** [МЭ, 1977] в $\mathcal{M}$.

Таким образом, подзадачи $Z_i$ ($i = 1,2,...,t$) порождают в модели $\mathcal{M}$ некоторую совокупность локальных подмоделей $\mathfrak{M}_1,...,\mathfrak{M}_t$. Нетрудно заметить, что имеет смысл соответствующее сужение распознающего оператора $B(\xi_\delta^{(m)}, \eta_\lambda^{(l)})$, которое мы обозначим $B_i$ $(i = 1,2,...,t)$.

Пусть $A_{\mathfrak{M}}$ - некоторая модель алгоритмов, порожденная распознающими операторами $\mathcal{M}$ и решающим правилом $c \in C(c_0, c_1)$. Задачу $Z = (I_0, \tilde{S}^q)$ назовем **локально-разрешимой** в модели $A_{\mathfrak{M}}$, если

$$\begin{gathered} \exists Z_1,...,Z_t (t > 1)\, \exists B \in \mathfrak{M}\ \forall c \in C(c_0, c_1)\, \forall S^u \in \tilde{S}^q \\ (c(B(I_0, S^u)) = c(B(I_0^i, S^u))) \end{gathered} \quad (3)$$

Непосредственно из определения для таких задач получаем

$$((\mathfrak{M}_i(Z_i) \bigcap R_c(Z_i) \neq \varnothing)_{i=1}^t \Rightarrow (\mathfrak{M}(Z) \bigcap R_c(Z) \neq \varnothing)).$$

Т.е. корректность локальных подмоделей на задачах $Z_1,...,Z_t$ с необходимостью влечет корректность исходной модели на Z. Подход к построению корректных алгоритмов, основанный на таком свойстве информации является локальным [МЭ, 1977].

Основную задачу данного подхода можно сформулировать следующим образом:

> *Необходимо определить условия на $\mathcal{M}$ и $Z \in Z_2^{ql}$, при которых задача распознавания является локально разрешимой в соответствующей модели алгоритмов $A_{\mathfrak{M}}$.*

В дальнейшем ограничимся случаем, когда совокупность подзадач $Z_1,...,Z_t (t > 1)$ удовлетворяет дополнительному условию:

$$(\tilde{S}_m = \bigcup_{i=1}^{t} \tilde{S}_m^i,\ \tilde{S}_m^i \bigcap \tilde{S}_m^j = \varnothing \ \text{ if } \ i \neq j)\ 1 \le i,j \le t \quad (4)$$

Нетрудно видеть, что каждой задаче Z однозначно соответствует некоторый набор подзадач $Z_1,...,Z_t$, для

которых имеют место условия (1), (4). Верно и обратное, т.е. каждому набору подзадач $Z_1,...,Z_t$ можно поставить в соответствие некоторую задачу $Z \in Z_2^{ql}$. Иными словами, в условиях (1), (4), соответствие между Z и $Z_1,...,Z_t (t>1)$ взаимно-однозначно с точностью до перестановок объектов в $\tilde{S}_m$ и $\tilde{S}^q$.

## Критерий локальной разрешимости

Покажем возможность построения алгоритмов распознавания с использованием описанного выше локального подхода и определим условия локальной разрешимости задач из множества $Z_2^{ql}$.

В [Журавлев, 1978] введено понятие распознающего оператора линейно зависящего от параметров. Формальное объединение таких операторов названо линейной моделью. Опишем данную модель в виде (2). Для этого используем некоторую идеализацию процесса построения решений задачи Z в $M(\xi_\delta^{(m)},\eta_\lambda^{(l)})$. Предположим, что он реализуется в два этапа: на первом – строится проекция объектов $\tilde{S}^q$ на обучающую выборку $\tilde{S}_m$, а на втором – полученные оценки проектируются на классы $K_1,...,K_l$. Определим наборы функций и $\xi_\delta^{(m)}=(\xi_\delta^1,...,\xi_\delta^m)$ и $\eta_\lambda^{(l)}=(\eta_\lambda^1,...,\eta_\lambda^l)$ такие, что

$$\forall S \in \{S\}\ (\xi_\delta^i : S \times S_i \to \square)\ \forall S_i \in \tilde{S}_m (i=1,2,...,m)\ \forall \delta \in \Omega \tag{5}$$

$$(\eta_\lambda^j : (\square)^m \to \square)\ \forall \lambda \in \Lambda\ (j=1,2,...,l) \tag{6}$$

Тогда распознающие операторы модели $M(\xi_\delta^{(m)},\eta_\lambda^{(l)})$ представимы в виде суперпозиции

$$B=\eta_\lambda^l \circ \xi_\delta^m,\ \text{где}\ \delta \in \Omega, \lambda \in \Lambda \tag{7}$$

Нетрудно заметить, что такие операторы (по построению функций $\xi_\delta^{(m)}$ и $\eta_\lambda^{(l)}$) реализуют отображение $Z \in Z_2^{ql}$ в пространство вещественных матриц $\square^{ql}$.

Пусть $L(\square^m,\square^l)$ - пространство линейных операторов из $\square^m$ в $\square^l$. Модель (2) с распознающими операторами (7), для которых

$$\eta_\lambda^{(l)} \in L(\square^m,\square^l)$$

назовем **полулинейной** (обозначим ее $M(\xi_\delta^{(m)},L^{ml})$), а соответствующее семейство $A_M$ - полулинейным.

Заметим, что рассмотрение таких моделей не уменьшает общности полученных в дальнейшем результатов. Так в [Журавлев, 1978] показано, что многие известные эвристические модели (в том числе – с разделяющими гиперплоскостями, потенциальных функций, вычисления оценок) являются полулинейными в указанном выше смысле. В тоже время, существуют модели (например, с предварительным преобразованием информации из Z [Krasnoproshin, 2006]) в которых процесс построения решений реализуется другими наборами функций типа (5), (6). Однако и они, в свою очередь, могут быть сделаны полулинейными.

Обозначим $\eta_\lambda^{(l)}(i,j)$ - матрицу пространства $\square^{ml}$ $(1 \le i \le m, 1 \le j \le l)$, соответствующую линейному оператору $\eta_\lambda^{(l)} \in L(\square^m,\square^l)$.

**Теорема 1**. Пусть $M(\xi_\delta^{(m)},L^{ml})$ - полулинейная модель с произвольными функциями $\xi_\delta^{(m)}$ вида (5). Задача $Z \in Z_2^{ql}$ локально разрешима в $A_M$ тогда и только тогда, когда

$$\exists Z_1,...,Z_t (t>1)\ \exists \xi_\delta^{(m)}\ \ \eta_\lambda^{(l)} \in L(\square^m,\square^l)$$
$$\forall i=\{1,2,...,t\}\ \forall S^u \in \tilde{S}_i^q\ \ (u=1,2,...,q)\ \forall j \in \{1,2,...,l\}$$
$$(\sum_{S_v \in \tilde{S}_m^i} \xi_\delta^v (S^u,S_v)\cdot \eta_\lambda^l (v,j)=0)$$

Теорема 1 дает критерий локальной разрешимости произвольной задачи $Z \in Z_2^{ql}$ в полулинейной модели $A_M$. Условия теоремы можно использовать как при исследовании локальной разрешимости задач, так и для построения соответствующих алгоритмов. Однако более конструктивным в этом смысле является следующее условие:

$$\exists Z_1,...,Z_t (t>1)\ \exists \xi_\delta^{(m)}\ \forall i \in \{1,2,...,t\}\ \ \forall S^u \in \tilde{S}_i^q\ \ \forall S_v \notin \tilde{S}_m^i\ \ (\xi_\delta^v (S^u,S_v)=0)$$

Легко показать, что оно является достаточным для локальной разрешимости задач $Z_2^{ql}$ в полулинейных моделях $A_M$.

## Достаточные условия локальной разрешимости

Пусть $A_M$ - произвольная полулинейная модель распознающих алгоритмов. Рассмотрим условия на $Z \in Z_2^{ql}$, при которых эти задачи являются локально разрешимыми в $A_M$.

**Линейная независимость подзадач**

Предположим, что $\Re$ - обычное евклидово конечномерное пространство. Подзадачи $Z_1,...,Z_t (t>1)$ задачи $Z \in Z_2^{ql}$ назовем **линейно-независимыми**, если

$$\forall i,j \in \{1,2,...,t\}\ \ (l(\tilde{S}_i^q \bigcup \tilde{S}_m^i) \bigcap l(\tilde{S}_j^q \bigcup \tilde{S}_m^j)=\varnothing)\ \text{ при } i \neq j, \qquad (8)$$

где $l$ – линейная оболочка в евклидовом пространстве $\Re$.

Обозначим $L(\Re,\Re)$ - пространство линейных операторов, сопряженное к $\Re$, т.е.

$$\forall R \in L(\Re,\Re)\ \ (R:\Re \to \Re).$$

Введем функции $\xi_\delta^{(m)}$ вида (5)

$$\forall S \in \{S\}\ \forall S_u \in \tilde{S}_m\ \ (\xi_{\delta R}^{(m)}(S,S_u)=\delta < R(S),R(S_u) >), \qquad (9)$$

где $< \cdot,\cdot >$ - скалярное произведение в евклидовом пространстве $\Re$, $R(S)$ - линейное преобразование объекта $S \in \{S\}$ оператором $R \in L(\Re,\Re)$ и $\delta$ - некоторый числовой параметр. Обозначим для краткости $M(\xi_{\delta R}^{(m)},L^{ml})$ - полулинейную модель с функциями (9).

**Теорема 2**. Пусть $M(\xi_{\delta R}^{(m)},L^{ml})$ - полулинейную модель распознающих операторов и $Z$ - произвольная задача из $Z_2^{ql}$. Если в $Z$ существуют линейно-независимые подзадачи $Z_1,...,Z_t (t>1)$, то задача $Z$ локально разрешима в соответствующей модели $A_M$.

**Характеристическая независимость подзадач**

В дальнейшем полагаем $\Re=\square^n$. Определим в пространстве $\square$ характеристическую функцию $\gamma_{R_0}$ произвольного подмножества $R_0 \subset \square$ такую, что

$$\forall x \in \square \qquad \gamma_{R_0}(x)=\begin{cases}1, & if\ \ x \in R_0\\ 0, & otherwise.\end{cases}, \qquad (10)$$

Рассмотрим в $\Re^n$ подмножество $(R_1 \times .... \times R_n)$ и введем отображение

$$\forall X = (x_1,...,x_n) \in \Re^n \ (\gamma_{(R_1,...,R_n)}(X) = (\gamma_{R_1}(x_1),...,\gamma_{R_n}(x_n))), \tag{11}$$

с элементами в виде (9). Полученный при таком отображении вектор $(\gamma_{R_1}(x_1),...,\gamma_{R_n}(x_n))$ назовем **характеристическим** для $X \in \Re^n$ по $\gamma_{(R_1,...,R_n)}$. Введем также

$$\overline{\gamma}_{(R_1,...,R_n)}(X) = 1_{B_2^n} - \gamma_{(R_1,...,R_n)}(X),$$

где $1_{B_2^n}$ - единичный вектор пространства $B_2^n$. Для фиксированных $\gamma_0 \in B_2^n$ и $\gamma_{(R_1,...,R_n)}$ определим в $\Re^n$ подмножество

$$R^{\gamma_0}_{\gamma_{(R_1,...R_n)}} = \{X \in \Re^n \,|\, (<\gamma_{(R_1,...,R_n)}(X), \gamma_0>) > 0 \,\&\, (<\gamma_{(R_1,...,R_n)}(X), \overline{\gamma}_0>) = 0\}. \tag{12}$$

Нетрудно видеть, что $R^{\gamma_0}_{\gamma_{(R_1,...R_n)}} \neq \varnothing$ при условии, что $\gamma_0 \neq 0_{B_2^n}$ (где $0_{B_2^n}$ - нулевой вектор пространства $B_2^n$).

Рассмотрим некоторые свойства подмножества (14), порожденные отображениями (11).

**Лемма 1**. Пусть $\gamma^0, \gamma^1 \in B_2^n$ и $\gamma_{(R_1,...,R_n)}$ - некоторое отображение (11). Тогда

$$(R^{\gamma^0}_{\gamma_{(R_1,...R_n)}} \bigcap R^{\gamma^1}_{\gamma_{(R_1,...R_n)}} = \varnothing) \Leftrightarrow (<\gamma^0, \gamma^1> = 0) \tag{13}$$

Непосредственно из определения нетрудно получить способ порождения подмножеств $R^{\gamma^0}_{\gamma_{(R_1,...R_n)}}$, содержащих некоторую заданную совокупность $X^{(m)} = (X_1,...,X_m) \subset \Re^n$. Действительно, зафиксируем произвольное ненулевое для всех $X \in X^{(m)}$ отображение (11) и определим вектор

$$\gamma_{(R_1,...,R_n)}(X^{(m)}) = (\gamma_{R_1}(\bigcup_{i=1}^{m} x_{i1}),...,\gamma_{R_n}(\bigcup_{i=1}^{m} x_{in})). \tag{14}$$

где

$$\gamma_{R_j}(\bigcup_{i=1}^{m} x_{ij}) = \begin{cases} 1, if\,(\exists i \in \{1,2,...,m\}\ (\gamma_{R_j}(x_{ij}) = 1) \\ 0, otherwise. \end{cases}.$$

Тогда, по построению имеем

$$X^{(m)} \subset R^{\gamma_{(R_1,...R_n)}(X^{(m)})}_{\gamma_{(R_1,...R_n)}}.$$

Исходя из проведенных рассуждений, можно ввести следующее определение. Подзадачи $Z_1,...,Z_t (t > 1)$ задачи $Z$ назовем **характеристически-независимыми**, если

$$\exists \gamma_{(R_1,...,R_n)} \begin{cases} \forall S \in (\tilde{S}_m \bigcup \tilde{S}^q)\ (\gamma_{(R_1,...,R_n)}(S) \neq 0_{B_2^n}) \\ \forall i \forall j \neq i\ (<\gamma_{(R_1,...,R_n)}(\tilde{S}_i^q \bigcup \tilde{S}_m^i), \gamma_{(R_1,...,R_n)}(\tilde{S}_j^q \bigcup \tilde{S}_m^j) >= 0) \end{cases}. \tag{15}$$

Покажем, что вопрос построения таких подзадач при фиксированном отображении $\gamma_{(R_1,...,R_n)}$ сводится к вопросу о приводимости специальной квадратной матрицы к блочно-диагональной форме с *t* блоками на главной диагонали.

Пусть $\gamma_{(R_1,...,R_n)}$ - произвольное отображение (15), удовлетворяющее на $Z$ условию

$$\forall S \in (\tilde{S}_m \bigcup \tilde{S}^q)\ (\gamma_{(R_1,...,R_n)}(S) \neq 0_{B_2^n}).$$

Нетрудно заметить, что для таких отображений условие (14) эквивалентно следующему

$$\forall i, j \in \{1,2,...,t\} \quad \forall S' \in (\tilde{S}_i^q \bigcup \tilde{S}_m^i)\ \forall S'' \in (\tilde{S}_j^q \bigcup \tilde{S}_m^j) \quad (<\gamma_{(R_1,...,R_n)}(S'), \gamma_{(R_1,...,R_n)}(S'')> = 0). \tag{16}$$

Предположим теперь, что подзадачи $Z_1,...,Z_t (t>1)$ задачи $Z$ для некоторого $\gamma_{(R_1,...,R_n)}$ удовлетворяют (15). Пусть, кроме того, в $Z$ выборки $\tilde{S}_m$ и $\tilde{S}^q$ упорядочены таким образом, что вначале расположены объекты соответствующие подзадаче $Z_1$ и т.д. Рассмотрим матрицу

$$\chi_{\gamma_{(R_1,...,R_n)}}(\tilde{S}_m \bigcup \tilde{S}^q) = \left\| \begin{array}{c} \chi(S_1,S_1).....\chi(S_1,S_{m+q}) \\ ..... \\ \chi(S_{m+q},S_1).....\chi(S_{m+q},S_{m+q}) \end{array} \right\|, \quad (17)$$

где для всех $S_i, S_j \in (\tilde{S}_m \bigcup \tilde{S}^q)$ $(1 \le i, j \le m+q)$

$$\chi(S_i,S_j) = \begin{cases} 1, if (< \gamma_{(R_1,...,R_n)}(S_i), \gamma_{(R_1,...,R_n)}(S_j) > \neq 0), \\ 0, otherwise. \end{cases}.$$

Очевидно, что при сделанных предположениях построенная матрица будет иметь блочно-диагональную форму с *t* блоками на главной диагонали. Верно и обратное. Если при произвольной нумерации объектов в $(\tilde{S}_m \bigcup \tilde{S}^q)$ матрица $\chi_{\gamma_{(R_1,...,R_n)}}(\tilde{S}_m \bigcup \tilde{S}^q)$ для некоторого $\gamma_{(R_1,...,R_n)}$ приводима к блочно-диагональной форме, и в каждый блок на главной диагонали попадают объекты из $\tilde{S}_m$ и $\tilde{S}^q$, то для $Z$ можно указать характеристически-независимые подзадачи $Z_1,...,Z_t$, где *t* – число блоков в полученной матрице.

Вопрос о приведении матрицы (17) к блочно-диагональной форме элементарно решается с помощью методов, изложенных в [Тьюарсон, 1977]. В частности, если ввести матрицу

$$\chi^2_{\gamma_{(R_1,...,R_n)}}(\tilde{S}_m \bigcup \tilde{S}^q) = \chi(\chi_{\gamma_{(R_1,...,R_n)}}(\tilde{S}_m \bigcup \tilde{S}^q)),$$

с элементами $(1 \le i, j \le m+q)$

$$\chi_{ij} = \begin{cases} 1, if \sum_{u=1}^{m+q} \chi(S_i,S_u) \cdot \chi(S_j,S_u) > 0, \\ 0, otherwise, \end{cases}.$$

то можно воспользоваться теоремой 3.5.1 из [Тьюарсон, 1977]. Обозначим $E^{m+q}_{m+q}$ - единичную относительно коммутативного умножения матрицу пространства $\square^{m+q\,m+q}$. Тогда простой переформулировкой указанной теоремы получаем следующий критерий приводимости матрицы (17) к блочно-диагональной форме

**Лемма 2**. Пусть $\chi^{2^h}_{\gamma_{(R_1,...,R_n)}}(\tilde{S}_m \bigcup \tilde{S}^q) \neq E^{m+q}_{m+q}$ для всех $h \le [\log_2(m+q)]$. Тогда матрица $\chi_{\gamma_{(R_1,...,R_n)}}(\tilde{S}_m \bigcup \tilde{S}^q)$ приводима к блочно-диагональной форме в том и только том случае, если

$$\exists h < [\log_2(m+q)] \ (\chi^{2^{h+1}}_{\gamma_{(R_1,...,R_n)}}(\tilde{S}_m \bigcup \tilde{S}^q) \equiv \chi^{2^h}_{\gamma_{(R_1,...,R_n)}}(\tilde{S}_m \bigcup \tilde{S}^q)).$$

Покажем теперь, что задача, в которой можно указать характеристически-независимые подзадачи $Z_1,...,Z_t (t>1)$, является локально разрешимой в некоторой полулинейной модели $A_M$. Для этого необходимо ввести соответствующие функции $\xi^{(m)}_\delta$ вида (5).

Пусть $Z_1,...,Z_t$ - характеристически-независимые подзадачи произвольной задачи $Z^{ql}_2$. Поставим в соответствие каждому объекту $S$ обучающей выборки $\tilde{S}_m$ характеристический вектор (13) той подзадачи $Z_i$ $(i \in \{1,2,...,t\})$, в которую относится $S$, т.е.

$$(S \in \tilde{S}^q_i \bigcup \tilde{S}^i_m) \Rightarrow (\gamma^*_{(R_1,...,R_n)}(S) = \gamma_{(R_1,...,R_n)}(\tilde{S}^q_i \bigcup \tilde{S}^i_m)).$$

Заметим, что для всех $S_u \in \tilde{S}_m$ $(u = 1,2,...,m)$ такой вектор определен однозначно. Введем в пространстве $\Box^n$ функции $\xi_\delta^{(m)}$ следующим образом

$$\forall S \in \{S\}\ (\xi_{\delta\chi}^{(u)}(S,S_u) = \chi^*(S,S_u)\cdot\xi_\delta^u(S,S_u)), 1 \le u \le m\ ,$$

где

$$\chi^*(S,S_u) = \begin{cases} 1, if\ (< \gamma_{(R_1,...,R_n)}(S), \gamma^*_{(R_1,...,R_n)}(S_u) >\neq 0), \\ 0, otherwise. \end{cases}.$$

Обозначим для краткости $M(\xi_{\delta\chi}^{(m)}, L^{ml})$ - полулинейную модель с такими функциями.

**Теорема 3**. Пусть $M(\xi_{\delta\chi}^{(m)}, L^{ml})$ - полулинейную модель распознающих операторов и $Z$ - задача распознавания из $Z_2^{ql}$. Если в $Z$ можно указать характеристически-независимые подзадачи $Z_1,...,Z_t (t > 1)$ , то задача $Z$ локально разрешима в соответствующей модели $A_M$ .

Отметим, что наиболее сложным при построении характеристически-независимых подзадач $Z_1,...,Z_t (t > 1)$ задачи $Z \in Z_2^{ql}$ является вопрос выбора отображений (15).

## Заключение

В работе описан один из возможных подходов к решению задачи распознавания образов в случаях, когда можно говорить о сложности априорной информации. В принципиальном смысле предлагаемый подход показывает, что со сложностью, которая является следствием большой размерности, можно справляться стандартным для математики способом – через декомпозицию задачи.

В настоящей работе рассматривается случай, когда на информации можно определить отношение эквивалентности. Показано, что для достаточно широкого класса алгоритмов, можно понизить сложность решения задачи распознавания. Сделано это на примере реализации корректных алгоритмов [Журавлев, 1978]. Полученные результаты могут послужить хорошей основой, как для дальнейших теоретических исследований, так и для решения конкретных практических задач.

## Библиография

[Мальцев, 1970] Мальцев А.И. Алгебраические системы. – М.: Наука, 1970. – 392 с.

[МЭ, 1977] Математическая энциклопедия. – М.: Советская энциклопедия. – 1977. – Т.1. – С. 207-209.

[Журавлев, 1978] Журавлев Ю. И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. — 1978. — Т. 33. — С. 5–68.

[Тьюарсон, 1977] Тьюарсон Р. Разреженные матрицы. – М.: Мир, 1977. - 189 с.

[Krasnoproshin, 2006] V.V.Krasnoproshin V.A.Obraztsov Problem of Solvability and Choice of Algorithms for Decision Making by Precedence // Pattern Recognition and Image Analysis. 2006. Vol. 16. no 2.- p.p.155-169.

## Информация об авторах

***Виктор Краснопрошин*** *– заведующий кафедрой МО АСУ, ФПМИ, Белорусский государственный университет, пр-т Независимости, 4, Минск, 220050, Беларусь; e-mail: krasnoproshin@bsu.by*

***Владимир Образцов*** *– доцент кафедры МО АСУ, ФПМИ, Белорусский государственный университет, пр-т Независимости, 4, Минск, 220050, Беларусь; e-mail: obraztsov@bsu.by*

# ЗАДАЧИ ПОМЕХОУСТОЙЧИВОГО АНАЛИЗА И РАСПОЗНАВАНИЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ, ВКЛЮЧАЮЩИХ ПОВТОРЯЮЩИЕСЯ УПОРЯДОЧЕННЫЕ НАБОРЫ ВЕКТОР–ФРАГМЕНТОВ[1]

**Александр Кельманов, Людмила Михайлова, Сергей Хамидуллин**

***Аннотация**: Рассматриваются некоторые задачи помехоустойчивого off-line анализа и распознавания числовых и векторных последовательностей, включающих повторяющиеся наборы квазипериодических фрагментов или векторов. Обоснованы эффективные алгоритмы решения этих задач, гарантирующие оптимальность решения по критерию максимального правдоподобия, в случае, когда помеха аддитивна и является гауссовской последовательностью независимых одинаково распределенных случайных величин.*

***Ключевые слова**: структурированная последовательность, упорядоченный набор вектор-фрагментов, помехоустойчивое обнаружение и распознавание, дискретная экстремальная задача, off-line алгоритм.*

***ACM Classification Keywords**: F.2. Analysis of Algorithms and Problem Complexity, G.1.6. Optimization, G2. Discrete Mathematics, I.5. Pattern Recognition*

***Conference**: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

## Введение

Объектом исследования настоящей работы являются проблемы анализа и распознавания структурированных данных – числовых и векторных последовательностей, в составе которых имеются повторяющиеся, чередующиеся и перемежающиеся информационно значимые фрагменты или векторы. Предмет исследования – некоторые варианты проблемы помехоустойчивого off-line анализа и распознавания последовательностей, включающих повторяющиеся упорядоченные наборы вектор-фрагментов в качестве структурных элементов, в предположении, что скрытые в шуме фрагменты или векторы из искомых наборов совпадают с компонентами упорядоченного эталонного набора векторов, принадлежащего заданному конечному множеству (словарю). Цель работы – обоснование алгоритмов решения этих задач.

Рассмотрим две содержательные задачи. Пусть в первой из них источник сообщений передает информацию об активном состоянии некоторого физического объекта в виде эталонного набора импульсов, имеющих одну и ту же известную длительность, но различную форму. Каждому импульсу соответствует некоторое промежуточное активное состояние объекта. Порядок импульсов фиксирован. Пассивному состоянию соответствует отсутствие каких-либо импульсов. На приемную сторону через канал передачи поступает последовательность квазипериодически чередующихся импульсов, искаженная аддитивным шумом. Термин «квазипериодически» означает, что интервал между двумя последовательными импульсами не одинаков, а лишь ограничен сверху и снизу некоторыми константами. Моменты времени появления импульсов в принятой (наблюдаемой) зашумленной последовательности

[1] Работа поддержана грантами РФФИ 09-01-00032, 07-07-00022 и грантом АВЦП Рособразования 2.1.1/3235.

неизвестны. Требуется обнаружить упорядоченные наборы импульсов в наблюдаемой последовательности, т.е. определить моменты времени, в которые объект находился в активном состоянии.

Во второй содержательной задаче предполагается, что на приемную сторону поступает информация от различных физических объектов, число которых конечно. Каждому объекту однозначно соответствует известный уникальный упорядоченный векторный набор, элемент которого – результат измерения каких-либо характеристик этого объекта в промежуточном активном состоянии. Число промежуточных активных состояний у физических объектов не одинаково. В пассивном состоянии все измеряемые характеристики равны нулю. Упорядоченная совокупность промежуточных активных состояний соответствует активному состоянию этого объекта в целом. На приемную сторону поступает искаженная шумом квазипериодическая последовательность результатов измерения характеристик от неизвестного объекта. Требуется определить (распознать), от какого объекта поступила информация.

Ситуации, в которых возникают сформулированные содержательные задачи, характерны, в частности, для электронной разведки, геофизики, гидроакустики, телекоммуникации и других приложений. В обеих задачах возможны два случая, когда число принятых импульсов или число ненулевых векторных наборов в последовательности известно и неизвестно. Эти случаи для двух сформулированных содержательных задач проанализированы в настоящей работе.

## Формальная постановка задач

Пусть $\mathbf{x}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$, где $\mathcal{N} = \{1,2,\ldots,N\}$, – последовательность векторов евклидова пространства. Допустим, что эта последовательность имеет следующую структуру

$$\mathbf{x}_n = \begin{cases} \mathbf{u}_1, & n \in \mathcal{M}_1, \\ \mathbf{u}_2, & n \in \mathcal{M}_2, \\ \ldots, & \ldots, \\ \mathbf{u}_L, & n \in \mathcal{M}_L, \\ \mathbf{0}, & n \in \mathcal{N} \setminus \bigcup_{j=1}^{L} \mathcal{M}_j, \end{cases} \tag{1}$$

где $\bigcup_{j=1}^{L} \mathcal{M}_j \subseteq \mathcal{N}$, причем $\mathcal{M}_i \cap \mathcal{M}_j = \varnothing$, если $i \neq j$.

Положим $|\mathcal{M}_j| = M_j$, $j = 1,2,\ldots,L$, и $\{n_1,\ldots,n_M\} = \bigcup_{j=1}^{L} \mathcal{M}_j$, где $M = \sum_{j=1}^{L} M_j$. В дополнение к этому допустим, что

$$\mathcal{M}_j = \{n_m \mid m \equiv j (\mathrm{mod}\, L),\ 1 \le m \le M\},\ j = 1,\ldots,L, \tag{2}$$

причем элементы набора $(n_1,\ldots,n_M)$, соответствующие номерам ненулевых векторов в последовательности (1), удовлетворяют ограничениям

$$1 \le T_{\min} \le n_m - n_{m-1} \le T_{\max} \le N-1, \quad m = 2,\ldots,M, \tag{3}$$

где $T_{\min}$ и $T_{\max}$ – натуральные числа.

Ограничения (3) устанавливают допустимый интервал между двумя ближайшими номерами ненулевых векторов в последовательности (1). Эти ограничения можно трактовать как условие квазипериодичности повторов ненулевых векторов в последовательности (1).

Из (1)-(3) видно, что последовательность $\mathbf{x}_n$ включает $\lfloor M/L \rfloor$ полных повторов векторного набора $(\mathbf{u}_1,...,\mathbf{u}_L)$ и, возможно, один неполный набор. Элементы повторяющегося набора $(\mathbf{u}_1,...,\mathbf{u}_L)$ будем интерпретировать как информационно значимые векторы. Доступной для анализа будем считать последовательность

$$\mathbf{y}_n = \mathbf{x}_n + \mathbf{e}_n, \quad n \in \mathcal{N}, \tag{4}$$

где $\mathbf{e}_n$ – вектор помехи (ошибки измерения), независимый от вектора $\mathbf{x}_n$. Заметим, что $\mathbf{x}_n = \mathbf{x}_n(n_1,...,n_M,\mathbf{u}_1,...,\mathbf{u}_L)$. Положим

$$S(n_1,...,n_M,\mathbf{u}_1,...,\mathbf{u}_L) = \sum_{n\in\mathcal{N}} \| \mathbf{y}_n - \mathbf{x}_n \|^2 \tag{5}$$

и рассмотрим следующие задачи среднеквадратического приближения.

**Задача 1a.** *Дано*: последовательность $\mathbf{y}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$, структура которой описывается формулами (1)-(4), набор $(\mathbf{u}_1,...,\mathbf{u}_L)$ ненулевых векторов из $\mathcal{R}^q$ и натуральное число $M$. *Найти*: набор $(n_1,\ldots,n_M)$ номеров такой, что целевая функция (5) минимальна.

**Задача 1b.** *Дано*: последовательность $\mathbf{y}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$, структура которой описывается формулами (1)-(4), набор $(\mathbf{u}_1,...,\mathbf{u}_L)$ ненулевых векторов из $\mathcal{R}^q$. *Найти*: набор $(n_1,\ldots,n_M)$ номеров и его размерность $M$ такие, что целевая функция (5) минимальна.

Задачи 1a и 1b отражают сущность проблемы оптимального обнаружения по критерию минимума суммы квадратов уклонений заданного повторяющегося набора информационно значимых векторов в ненаблюдаемой последовательности, структура которой описывается формулами (1)-(3). Отличие этих задач состоит в том, что в первой из них число ненулевых информационно значимых векторов считается заданным, а во второй – неизвестным, т.е. является оптимизируемой величиной.

Положим $\mathbf{w} = (\mathbf{u}_1,...,\mathbf{u}_L)$. Допустим в дополнение к (1)-(4), что $\mathbf{w} \in \mathcal{W}$, причем $|\mathcal{W}| = K$, где

$$\mathcal{W} \subset \{(\mathbf{u}_1,...\mathbf{u}_L) \mid \mathbf{u}_j \in \mathcal{R}^q, 0 < \| \mathbf{u}_j \|^2 < \infty, j = 1,\ldots,L;\ L \in \{1,...,L_{\max}\}\}. \tag{6}$$

Здесь $\mathcal{W}$ – множество (словарь) векторных наборов (слов) мощности $K$, размерность которых не превосходит $L_{\max}$.

Рассмотрим еще две задачи среднеквадратического приближения.

**Задача 2a.** *Дано*: множество $\mathcal{W}$, $|\mathcal{W}| = K$, наборов векторов из $\mathcal{R}^q$, последовательность $\mathbf{y}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$, структура которой описывается формулами (1)-(4) и (6), а также натуральное число $M$. *Найти*: векторный набор $\mathbf{w} \in \mathcal{W}$ такой, что целевая функция (5) минимальна на множестве допустимых наборов $(n_1,\ldots,n_M)$.

**Задача 2b.** *Дано*: множество $\mathcal{W}$, $|\mathcal{W}| = K$, наборов векторов из $\mathcal{R}^q$, последовательность $\mathbf{y}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$, структура которой описывается формулами (1)-(4) и (6), *Найти*: векторный набор $\mathbf{w} \in \mathcal{W}$ такой, что целевая функция (5) минимальна на множестве допустимых наборов $(n_1,\ldots,n_M)$.

Задачи 2a и 2b соответствуют проблеме распознавания последовательностей, включающих повторяющиеся наборы чередующихся векторов, скрытых в ненаблюдаемой последовательности (1).

В задаче 2a число ненулевых векторов в последовательности считается заданным, а в задаче 2b – неизвестным.

Легко установить, что к минимизации функции (5) и к таким же сформулированным выше четырем задачам приводит статистический подход к проблемам обнаружения и распознавания, если считать, что $\mathbf{e}_n$ в формуле (4) есть выборка из $q$-мерного нормального распределения с параметрами $(\mathbf{0}, \sigma^2\mathbf{I})$, где $\mathbf{I}$ единичная матрица, а в качестве критерия решения задачи использовать максимум функционала правдоподобия.

## Редуцированные оптимизационные задачи

Раскрывая квадрат нормы в формуле (5), получим

$$S = \sum_{n\in\mathcal{N}} \| \mathbf{y}_n \|^2 + \sum_{j=1}^{L} M_j \| \mathbf{u}_j \|^2 - 2\sum_{j=1}^{L} \sum_{n\in\mathcal{M}_j} \left\langle \mathbf{y}_n, \mathbf{u}_j \right\rangle$$

$$= \sum_{n\in\mathcal{N}} \| \mathbf{y}_n \|^2 + \sum_{m=1}^{M} \| \mathbf{u}_{(m-1) \bmod L+1} \|^2 - 2\sum_{m=1}^{M} \left\langle \mathbf{y}_{n_m}, \mathbf{u}_{(m-1) \bmod L+1} \right\rangle,$$

где $\langle \cdot,\cdot \rangle$ – скалярное произведение.

Первое слагаемое в правой части полученного выражения – константа. При фиксированных $M$ и $(\mathbf{u}_1,...,\mathbf{u}_L)$ второе слагаемое также является константой. Поэтому имеем следующие редуцированные оптимизационные задачи, к которым сводятся задачи 1a и 1b.

**Задача SRTVS-F** (Searching for Recurring Tuples of Vectors in a Sequence, when $M$ is Fixed). *Дано*: последовательность $\mathbf{y}_0,\ldots,\mathbf{y}_{N-1}$ векторов из $\mathcal{R}^q$, набор $(\mathbf{u}_1,\ldots,\mathbf{u}_L)$ ненулевых векторов из $\mathcal{R}^q$ и натуральное число $M$. *Найти*: набор $(n_1,...,n_M)$ номеров такой, что

$$\sum_{m=1}^{M} \left\langle \mathbf{y}_{n_m}, \mathbf{u}_{l(m,L)} \right\rangle \to \max,$$

где $l(m \mid L) = (m-1) \bmod L + 1$, при ограничениях (3).

**Задача SRTVS-NF** (Searching for Recurring Tuples of Vectors in a Sequence, when $M$ is Not Fixed). *Дано*: последовательность $\mathbf{y}_0,\ldots,\mathbf{y}_{N-1}$ векторов из $\mathcal{R}^q$ и набор $(\mathbf{u}_1,\ldots,\mathbf{u}_L)$ ненулевых векторов из $\mathcal{R}^q$. *Найти*: набор $(n_1,...,n_M)$ номеров такой, что

$$\sum_{m=1}^{M} \{2\left\langle \mathbf{y}_{n_m}, \mathbf{u}_{l(m,L)} \right\rangle - \| \mathbf{u}_{l(m,L)} \|^2\} \to \max, \qquad (7)$$

где $l(m \mid L) = (m-1) \bmod L + 1$, при ограничениях (3).

Точные полиномиальные алгоритмы решения этих редуцированных оптимизационных задач обоснованы в [1-3]. Трудоемкости алгоритмов решения задач SRTVS-F и SRTVS-NF есть величины $O[M(T_{\max} - T_{\min} + q)N]$ и $O[L(T_{\max} - T_{\min} + q)N]$ соответственно.

Задачи 2a и 2в сводятся к решению следующих экстремальных задач.

**Задача SVTVP-F** (Searching for a Vector Tuple in the Vocabulary of Patterns, when $M$ is Fixed). *Дано*: последовательность $\mathbf{y}_0,\ldots,\mathbf{y}_{N-1}$ векторов из $\mathcal{R}^q$, натуральное число $M$ и словарь $\mathcal{W}$, $|\mathcal{W}|=K$, упорядоченных наборов векторов из $\mathcal{R}^q$. *Найти*: векторный набор $\mathbf{w}\in\mathcal{W}$ такой, что выполняется (7), при ограничениях (3).

Задача **SVTVP-NF** (Searching for a Vector Tuple in the Vocabulary of Patterns, when $M$ is Not Fixed). *Дано*: последовательность $\mathbf{y}_0,\ldots,\mathbf{y}_{N-1}$ векторов из $\mathcal{R}^q$ и множество (словарь) $\mathcal{W}$, $|\mathcal{W}|=K$, упорядоченных наборов (слов) векторов из $\mathcal{R}^q$. *Найти*: векторный набор $\mathbf{w}\in\mathcal{W}$ такой, что выполняется (7), при ограничениях (3).

Точные полиномиальные алгоритмы решения этих экстремальных задач обоснованы в [4-5]. Временные сложности алгоритмов решения задач SVTVP-F и SVTVP-NF есть величины $O[KM(T_{\max}-T_{\min}+q)N]$ и $O[KL_{\max}(T_{\max}-T_{\min}+q)N]$ соответственно.

Алгоритмы решения приведенных редуцированных задач лежат в основе алгоритмов помехоустойчивого анализа и распознавания структурированных последовательностей, включающих повторяющиеся наборы чередующихся вектор-фрагментов. Эти алгоритмы гарантируют оптимальность решения как по критерию максимального правдоподобия в случае, когда помеха аддитивна и является гауссовской последовательностью независимых одинаково распределенных величин, так и по критерию минимума суммы квадратов уклонений.

## Численное моделирование

Результаты численных экспериментов, представленные ниже в качестве примера, носят чисто иллюстративный характер. Они лишь демонстрируют работу алгоритмов и сущность рассмотренных задач для одномерных последовательностей.

На рис. 1 а изображена сгенерированная последовательность $X$, включающая 3 повтора набора фрагментов. На рис. 1 б представлена последовательность $Y$, подлежащая обработке (в этом примере уровень помехи превышает уровень сигнала). На рис. 1 в приведена последовательность $\hat{X}$, полученная с помощью алгоритма обнаружения, в условиях, когда число $M$ задано. Прямоугольными рамками очерчены места расположения обнаруженного набора, найденные алгоритмом в зашумленной последовательности. Числовые данные под графиками соответствуют заданным (рис. 1 а) и найденным (рис. 1 б и 1 в) начальным номерам фрагментов. Рисунок иллюстрирует практически безупречную работу алгоритма в условиях, когда уровень сигнала ниже уровня помехи.

На рис. 2 представлены кривые оценок нормированной среднеквадратической ошибки $e(\sigma)=\mathbf{E}\,\|X-\hat{X}\|^2/e^u$, где $\mathbf{E}$ – символ математического ожидания, $e^u$ – оценка сверху для $\|X-\hat{X}\|^2$. Кривая 1 получена с помощью алгоритма обнаружения при неизвестном числе $M$ фрагментов, а кривая 2 – с помощью алгоритма, ориентированного на ситуацию, когда это число известно. Результаты получены при обработке одних и тех же 25000 сгенерированных последовательностей, в составе которых повторялся набор из трех фрагментов; места расположения фрагментов в последовательностях генерировались с помощью датчика случайных чисел.

Рис. 3 иллюстрирует зависимость от уровня помехи вероятности ошибки распознавания последовательностей, включавших повторы двух различных эталонных наборов, в составе которых имелось по три вектора. Теоретические оценки верхней и нижней границ вероятности ошибки

распознавания $\alpha^u(\sigma)$ и $\alpha^d(\sigma)$ в виде графиков приведены под номерами 1 и 4. Кривые 2 и 3 получены в условиях, когда число $M$ было неизвестно и известно соответственно.

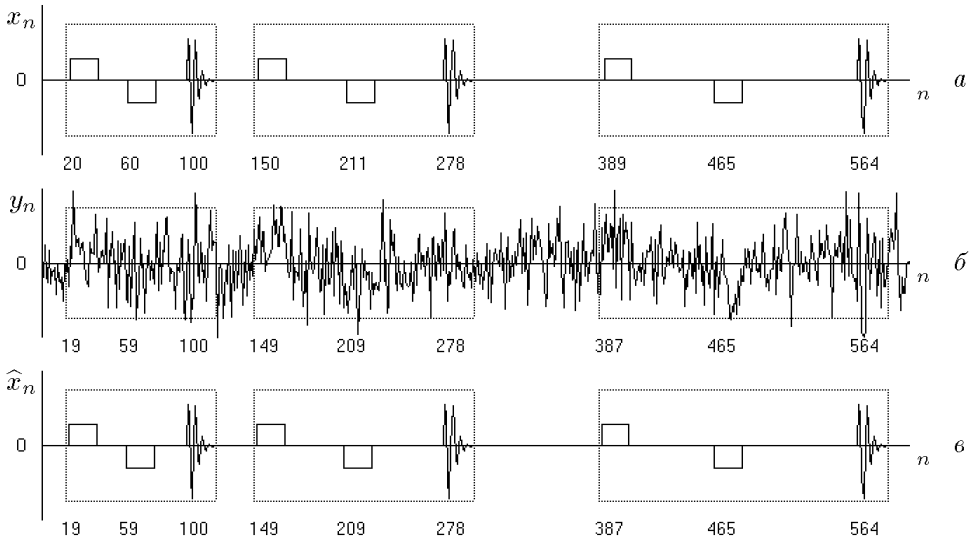



Рис. 1

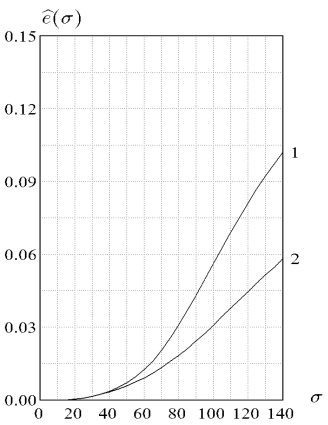



Рис. 2

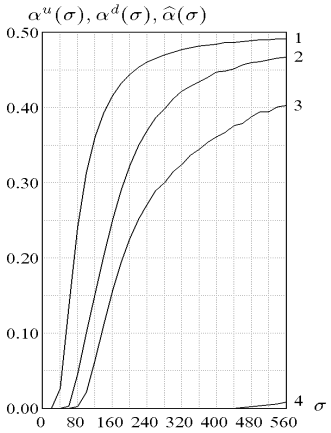



Рис. 3

Оценка вероятности ошибки распознавания при каждом значении $\sigma$ подсчитана по формуле $\hat{\alpha} = (v_1 + v_2)/2$, где $v_1$ и $v_2$ – число неверно опознанных последовательностей, сгенерированных по каждому эталонному набору. Моделировалась байесовская процедура принятия решения с равновероятными гипотезами (наборами). Каждая точка экспериментальной кривой $\hat{\alpha}$ получена в результате усреднения 25000 значений. Рис. 2 и 3 демонстрируют легко доказуемый факт, что ошибка обнаружения и вероятность ошибки распознавания будут меньше в ситуации, когда число ненулевых фрагментов в последовательности известно, чем в ситуации, когда это число неизвестно.

## Заключение

Рассмотренные задачи входят в большое семейство актуальных задач [6], к которым сводятся типовые проблемы помехоустойчивого off-line анализа и распознавания структурированных данных в виде числовых и векторных последовательностей, включающих повторяющиеся, чередующиеся и

перемежающиеся информационно значимые векторы или фрагменты. В настоящей работе представлены эффективные алгоритмические решения четырех ранее не изученных задач из этого семейства.

Открытым остается вопрос о разрешимости обобщения рассмотренных задач обнаружения и распознавания на тот случай, когда вместо набора фрагментов, элементы которого упорядочены в соответствии с фиксированным набором векторов, требуется найти набор фрагментов с точностью до всевозможных перестановок элементов фиксированного векторного набора. Алгоритмы решения этих задач представляют значительный интерес для ряда упомянутых во введении приложений. Обоснование алгоритмов решения этих задач представляется делом ближайшей перспективы.

## Благодарности

Работа поддержана грантами РФФИ 09-01-00032, 07-07-00022 и грантом АВЦП Рособразования 2.1.1/3235.

## Литература

[1] Kel'manov A.V., Mikhailova L.V., Khamidullin S.A. A Posteriori Detection of a Recurring Tuple of Reference Fragments in a Quasi-Peridic Sequence // Computational Mathematics and Mathematical Physics. 2008, Vol. 48, No. 12, pp. 2276-2288.

[2] Кельманов А.В., Михайлова Л.В., Хамидуллин С.А. Об одной задаче поиска упорядоченных наборов фрагментов в числовой последовательности // Дискретный анализ и исследование операций. 2009 (принята в печать).

[3] Kel'manov A.V., Mikhailova L.V., Khamidullin S.A. Optimal Detection of a Recurring Tuple of Reference Fragments in a Quasiperiodic Sequence // Numerical Analysis and Applications. 2008. Vol. 1, No.3, pp. 255-268.

[4] Кельманов А.В., Михайлова Л.В., Хамидуллин С.А. Распознавание квазипериодической последовательности, включающей повторяющийся набор фрагментов // Сибирский журнал индустриальной математики. 2008, Т. 11, №2 (34). С. 74-87.

[5] Кельманов А.В., Михайлова Л.В., Хамидуллин С.А. Алгоритм распознавания квазипериодической последовательности, включающей повторяющийся набор фрагментов // Тез. докл. 15-й международной конф. «Проблемы теоретической кибернетики» (Казань, 2-7 июня 2008). Под ред. Ю.И. Журавлева.- Казань: Отечество, 2008.- С. 45.

[6] http://math.nsc.ru/~serge/qpsl

## Информация об авторах

***Александр Кельманов*** *– д.ф.-м.н., ведущий научный сотрудник, Институт математики им. С.Л. Соболева Сибирского отделения РАН, проспект академика Коптюга, 4, Новосибирск, 630090, Россия; Новосибирский государственный университет, ул. Пирогова, 2, Новосибирск, 630090, Россия; e-mail: kelm@math.nsc.ru*

***Людмила Михайлова*** *– к.ф.-м.н., старший научный сотрудник, Институт математики им. С.Л. Соболева Сибирского отделения РАН, проспект академика Коптюга, 4, Новосибирск, 630090, Россия; e-mail: mikh@math.nsc.ru*

***Сергей Хамидуллин*** *– к.т.н., старший научный сотрудник, Институт математики им. С.Л. Соболева Сибирского отделения РАН, проспект академика Коптюга, 4, Новосибирск, 630090, Россия; e-mail: kham@math.nsc.ru*

# ПОСТРОЕНИЕ ЛОГИКО-ВЕРОЯТНОСТНЫХ МОДЕЛЕЙ ВРЕМЕННЫХ РЯДОВ С ИСПОЛЬЗОВАНИЕМ ЦЕПЕЙ МАРКОВА

**Светлана Неделько**

***Abstract**: The method of logic and probabilistic models constructing for multivariate heterogeneous time series is offered. There are some important properties of these models, e.g. universality. In this paper also discussed the logic and probabilistic models distinctive features in comparison with hidden Markov processes. The early proposed time series forecasting algorithm is tested on applied task.*

***Keywords**: multivariate heterogeneous time series, pattern recognition, classification, deciding functions, logic and probabilistic models.*

***ACM Classification Keywords**: G.3 Probability and statistics: time series analysis, Markov processes, multivariate statistics, nonparametric statistics; G.1.6. Numerical analysis: optimization.*

***Conference**: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

## Введение

Задачи анализа и прогнозирования многомерных разнотипных временных рядов в настоящее время представляют большой интерес для исследования. В зависимости от предположений о прогнозируемой функции, постановки задач делятся на статистические (вероятностные) и детерминированные. В вероятностной постановке задач прогнозирования и идентификации модели математической моделью временного ряда $Z(t)$ выступает случайный процесс.

Среди процессов с дискретным временем большой интерес для задач прогнозирования представляют стационарные процессы с конечной длиной значимой предыстории, т. е. такие процессы, для которых распределение величин в момент $t$ не зависит от $t$ и всецело определяется реализовавшимися значениями на предыдущих $d$ моментах времени. Случайные процессы с такими свойствами называются *марковскими процессами*. Процессы рассмотренного типа удобны для анализа тем, что их реализацию можно свести к обычной таблице данных, после чего воспользоваться методами статистического анализа.

При $d = 1$ временной ряд с дискретным множеством значений (состояний) называется цепью Маркова, которая полностью определяется матрицей переходных вероятностей $p_{ij} = P(z(t) = j / z(t-1) = i)$.

Вероятности можно непосредственно оценить по выборочной реализации как частоты соответствующих переходов. В случае $d > 1$ матрица переходов становится многомерной, и ее непосредственное оценивание по частотам уже при небольших $d$ требует большой длины обучения. В этом случае необходимо использовать специализированные методы прогнозирования. Существенно расширяет область применимости марковских цепей подход, основанный на использовании скрытых марковских процессов.

Скрытые марковские модели [Baum, 1970] строятся обычно в рамках параметрического подхода, при этом вид условных распределений на наблюдаемых переменных для скрытых состояний задается эвристически (экспертно). Для подбора параметров модели используется, в основном, критерий максимума правдоподобия.

Другим подходом к аппроксимации процесса, задающего временной ряд, выступает построение логико-вероятностных моделей [Lbov, Nedel'ko, 2001]. В отличие от скрытых марковских моделей, здесь задается лишь класс разбиений, области которых и рассматриваются в качестве состояний, являющихся наблюдаемыми. При построении логико-вероятностных моделей можно использовать различные варианты [Неделько, 2008] критериев информативности.

## Постановка задачи

Имеется $n$-мерный временной ряд $v = \left\{ z^t \;\middle|\; t = \overline{1,N} \right\},\ z^t = \left( z_1^t,...,z_n^t \right),\ z_j^t \in Z_j$. Здесь $Z_j$ – множество допустимых значений $j$-й переменной ряда, $Z = \prod_{j=1}^{n} Z_j$. В наборе переменных $Z_1,...,Z_n$ могут быть одновременно переменные разных типов.

Пусть ряд является реализацией случайного $n$-мерного процесса $z(t)$ с дискретным временем, который задается переходной (условной) вероятностной мерой $P : \Lambda \times Z^d \to [0,1]$, где $\Lambda$ – σ-алгебра подмножеств из $Z$, а $d$ – длина предыстории, которая определяет распределение в заданный момент.

Требуется на основе имеющихся данных $v$ построить прогноз временного ряда в моменты времени $t > N$ в соответствии с некоторым критерием. Также можно ставить задачу аппроксимации условного распределения, т. е. построения модели описывающего ряд процесса. Найденная модель используется для анализа ряда, а также прогнозирования его значений.

Введем обозначения $X \equiv Z^d$, $Y \equiv Z$, причем $X$ будем использовать для обозначения пространства предысторий, а $Y$ — для пространства значений в момент времени, для которого делается прогноз. Тогда переходную меру можно записать как $P\left[ Z/z(t-1),\ z(t-2),\ ...,\ z(t-d) \right] \equiv P\left[ Y/x \right]$.

В данных обозначениях вероятность события $E_Y \in \Lambda$, $E_Y \subseteq Z$, записывается как

$$P\left( z(t) \in E_Y / z(t-1),\ z(t-2),\ ...,\ z(t-d) \right) \equiv\ P\left( E_Y / x \right) \equiv P\left[ Y/x \right]\left( E_Y \right).$$

Заметим, что круглые скобки используются для указания аргумента функции, а квадратные — как часть обозначения меры.

Во введенных обозначениях прогнозируемые значения есть $y_j(t) = z_j(t),\ j = 1,...,n$, а переменные, используемые для прогнозирования,

$$x_j(t) = z_j(t-1),\ x_{j+n}(t) = z_j(t-2)\ ,\ ...\ ,\ x_{j+n(d-1)}(t) = z_j(t-d),\ j = 1,...,n.$$

Размерность пространства прогнозирующих переменных есть $m = nd$.

## Логико-вероятностная модель

В работе будем прогнозировать многомерный разнотипный временной ряд в классе логических решающих функций, что сводится к задаче построения логико-вероятностной модели ряда. Пусть одновариантная решающая функция имеет вид $f : X \to Y$, многовариантное решающее правило представляется в виде $f : X \to S$, где решение $s \in S$ представляет собой множество пар

$s = \left\{ \left( E_Y^k, \tilde{p}_k \right) \;\middle|\; k = \overline{1,M} \right\},\ \tilde{p}_k$ – оценка условной вероятности $P\left( E_Y^k / x \right),\ E_Y^k \subseteq Y$.

Тогда логико-вероятностной моделью является многовариантное решающее правило $f_L = \left\{\left(E_X^l, s_l\right) \middle| \; l = \overline{1,L}\right\}$, где $E_X^l \subseteq X$ образуют разбиение $\alpha_L = \left\{E_X^l \;\middle|\; l = \overline{1,L}\right\}$ пространства $X$.

Решающую функцию можно строить на основе восстановления условного распределения либо в заданном классе в соответствии с эмпирическим критерием $K$. Используемый в данной работе подход предполагает частичную аппроксимацию распределений. Аппроксимирующей моделью случайного процесса будем называть случайный процесс, заданный на сигма-алгебре, являющейся подмножеством исходной. Таким образом, модель определяет вероятности не всех событий. Более того, практически мы будем рассматривать только модели, определяющие вероятности лишь на множествах конечной алгебры, порожденной множествами некоторого конечного разбиения пространства переменных. Вообще, говоря о сигма-алгебре, мы везде подразумеваем, что в частном случае она может быть конечной.

*Определение*. Носителем модели назовем совокупность таких множеств ее σ-алгебры, которые не включают в себя других элементов σ-алгебры.

*Определение*. Сложностью модели назовем мощность ее носителя.

*Определение*. Класс моделей $\Phi$ называется универсальным, если для любого $P[Z]$ и любого $\varepsilon > 0$ найдется модель $\overline{P}[Z] \in \Phi$, для которой $K(P[Z]) - K(\overline{P}[Z]) < \varepsilon$.

Содержательно, универсальность означает, что моделями класса можно сколь угодно точно аппроксимировать любое распределение (на изначально фиксированной σ-алгебре), при этом под точностью аппроксимации понимается близость значений критерия информативности.

Если пространство переменных непрерывно, то универсальный класс должен содержать, очевидно, бесконечное число моделей сколь угодно большой сложности. Однако, при этом он не обязан содержать моделей бесконечной сложности.

*Определение.* Класс моделей $\Phi$ назовем замкнутым, если для любых двух его моделей существует модель, носитель которой содержит пересечение носителей этих двух моделей.

*Гипотеза*. Если класс моделей $\Phi$ является замкнутым и лебегово замыкание объединения σ-алгебр моделей класса дает исходную σ-алгебру, то такой класс является универсальным.

Содержательно здесь утверждается, что универсальность класса моделей определяется тем, что модели в совокупности порождают ту же σ-алгебру событий, на которой задана вероятностная мера. Из данного утверждения следует, что класс логико-вероятностных моделей является универсальным.

Заметим, что сформулированное предположение тесно связано с универсальностью класса логических решающих функций [Лбов, Старцева, 1999], а также со свойством universal consistency решающих функций.

## Критерий информативности

Одним из критериев информативности, используемых для построения логико-вероятностной модели, является критерий на основе дивергенции, который может быть записан в виде

$$K_C(P[X,Y]) = \int \ln \frac{dP[Y/x]}{dP[Y]}\, dP[X,Y] = \int \ln \frac{dP[X,Y]}{dP[X]dP[Y]}\, dP[X,Y].$$

Известно, что дивергенция является неотрицательной величиной и равна нулю, только если распределения совпадают. Видно, что выражение критерия свелось к дивергенции между совместным распределением и произведением маргинальных распределений, что является величиной,

характеризующей степень зависимости между $X$ и $Y$. Заметим, что марковский случайный процесс полностью определяется вероятностной мерой $P[X,Y]$.

Под $\mu[Z]$ будем понимать меру Лебега, если $Z$ – непрерывное пространство, и считающую меру, если $Z$ – дискретно. Для разнотипного $Z$ мера $\mu[Z]$ будет определяться как сумма лебеговых мер непрерывных компонент множества.

*Свойство*. Значения дивергенции лежат в диапазоне $[0,+\infty]$, в частности, существуют марковские случайные процессы, для которых $K_C(P[X,Y]) = +\infty$.

Рассмотрим следующий пример. Пусть $Z = [0,1]$, $X = Y = Z$. Зададим случайный процесс через плотности

$$\frac{dP[X]}{d\mu[X]} = \begin{cases} 1, & x \in [0,1] \\ 0, & x \notin [0,1] \end{cases}, \qquad \frac{dP[Y/x]}{d\mu[Y]} = \begin{cases} 0{,}5, & y \in [0,1],\ y \neq x \\ 0, & y \notin [0,1] \end{cases}, \quad P(y = x/x) = 0{,}5.$$

При данном процессе значение единственной переменной в следующий момент времени с вероятностью $0{,}5$ остается прежним, а с вероятностью $0{,}5$ изменяется на случайное значение из интервала $[0,\ 1]$. Для построенного примера $K_C(P[X,Y]) = +\infty$.

Таким образом, дивергенция не просто неограничена, но может принимать бесконечные значения.

*Определение*. Энтропией меры $P[Z]$ будем называть величину $H(P[Z]) = -\int \ln \frac{dP[Z]}{d\mu[Z]}\, dP[Z]$.

Дивергенцию можно легко выразить через энтропию

$$K_C(P[X,Y]) = H(P[X]) + H(P[Y]) - H(P[X,Y]).$$

Здесь мы предполагаем конечность всех энтропий.

## Связь дивергенции и правдоподобия

Пусть $\varphi(y) = \frac{dP[Y]}{d\mu[Y]}$ – безусловная плотность вероятности, $\varphi(y/x) = \frac{dP[Y/x]}{d\mu[Y]}$ – условная плотность вероятности, а $\varphi(x,y) = \frac{dP[X,Y]}{d\mu[X,Y]}$ – совместная плотность вероятности, и $v = (z^1,...,z^N)$ – выборка, представляющая собой реализацию случайного процесса. Для простоты рассмотрим случай $d = 1$.

Функция правдоподобия выборки есть

$$\pi(v) = \ln \varphi(z^1) + \sum_{i=2}^{N} \ln \varphi(z^i / z^{i-1}) = \sum_{i=2}^{N} \ln \varphi(z^{i-1}, z^i) - \sum_{i=2}^{N} \ln \varphi(z^i).$$

Пусть $\widetilde{P}[Y]$ – эмпирическая (выборочная) мера (каждой выборочной точке, кроме первой, приписано значение $\frac{1}{N-1}$), а $\widetilde{P}[X,Y]$ – эмпирическая (выборочная) мера на парах, составленных из соседних выборочных значений (каждой паре приписано значение $\frac{1}{N-1}$). Тогда функция правдоподобия может быть записана в виде

$$\pi(v) = \int \ln \varphi(x,y)\, d\widetilde{P}[X,Y] - \int \ln \varphi(y)\, d\widetilde{P}[Y].$$

При достаточно больших объемах выборки эмпирическую меру можно приближенно заменить вероятностной мерой (в соответствии с которой получена выборка). Тогда в правой части получим разность энтропий

$$\pi(\nu) \approx H(P[Y]) - H(P[X,Y]).$$

Нетрудно убедиться, что это выражение получается и в случае $d > 1$.

Заметим, что полученное выражение отличается от $K_C(P[X,Y])$ на величину $H(P[X])$.

Данное отличие очень существенно. Если дивергенция определяется исключительно вероятностными мерами, то значение правдоподобия зависит от выбора меры μ. Так, например, при умножении меры μ на константу $A$ к величине правдоподобия прибавится слагаемое $-\ln A$. Таким образом, абсолютное значение правдоподобия не имеет содержательного смысла, а имеет смысл лишь отношение правдоподобия (или разность логарифмов).

На самом деле, нам как раз и нужно сравнивать модели, поэтому разность логарифмов правдоподобия была бы подходящей, если бы ее можно было вычислить во всех практических ситуациях. Однако для нахождения отношения правдоподобия необходимо, чтобы сравниваемые модели определяли вероятности на одном и том же множестве событий. При этом разные логико-вероятностные модели содержат оценки вероятностей для разных разбиений (совокупностей областей), и вовсе не обязаны включать в себя оценки вероятностей для областей из пересечения разбиений. Поэтому критерий правдоподобия, в отличие от дивергенции, практически не пригоден для оценивания качества логико-вероятностных моделей.

Заметим, что указанное обстоятельство является одним из наиболее существенных отличий рассматриваемого подхода, основанного на построении логико-вероятностных моделей, от подхода, связанного с построением скрытых марковских моделей, в котором требуется, чтобы модели были сравнимы по правдоподобию.

## Метод прогнозирования

Зафиксируем некоторое разбиение $\lambda = \left\{ E^{\omega} \subseteq Z \;\middle|\; \omega = \overline{1,k} \right\}$, $\bigcup_{\omega=1}^{k} E^{\omega} = Z$, $\omega \neq \varpi \Rightarrow E^{\omega} \cap E^{\varpi} = \varnothing$, пространства $Z$. Теперь исходному многомерному ряду $v$ можно сопоставить одномерную символьную последовательность $w = \left\{ \omega^t \;\middle|\; z^t \in E^{\omega^t},\ t = \overline{1,N} \right\}$. Случайному процессу $z(t)$ будет соответствовать процесс $\omega(t)$, переходные вероятности для которого обозначим

$$p_{\omega_0|\omega_1,\omega_2,\ldots,\omega_d} = P\left(\omega(t) = \omega_0 / \omega(t-1) = \omega_1,\ \ldots,\ \omega(t-d) = \omega_d\right).$$

Критерий качества, основанный на дивергенции, вводится следующим образом:

$$K_C(\lambda) = \sum_{\omega_0=1}^{k} \ldots \sum_{\omega_d=1}^{k} p_{\omega_0\omega_1\ldots\omega_d} \ln \frac{p_{\omega_0|\omega_1,\omega_2,\ldots,\omega_d}}{p_{\omega_0}}.$$

Приближенное к оптимальному разбиение $\lambda$ пространства переменных ряда $Z$ ищется алгоритмом направленного поиска LRP [Лбов, Старцева, 1999]. Оценивается матрица переходных вероятностей

между состояниями, представляющими собой области из разбиения. Критерием качества построенного решения является мера информативности матрицы переходных вероятностей между состояниями.

## Задача прогнозирования состояния ионосферы

Ionosphere — массив данных, содержащий результаты электромагнитного зондирования ионосферы, представленный в коллекции задач UCI (University of California, Irvine) Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets.html). Имеется 351 объект, характеризующийся 34 переменными. Каждая из 17 пар переменных есть действительная $Z_1$ и мнимая $Z_2$ части комплексной величины, соответствующей некоторой характеристике объекта. Известно, что для 225 объектов характеристики соответствуют хорошему прохождению определенного сигнала через ионосферу, а для 126 объектов плохому прохождению этого сигнала. Будем обозначать эти случаи как классы 1 и 0. Целью является прогнозирование класса (0 или 1), т.е. ставится задача классификации.

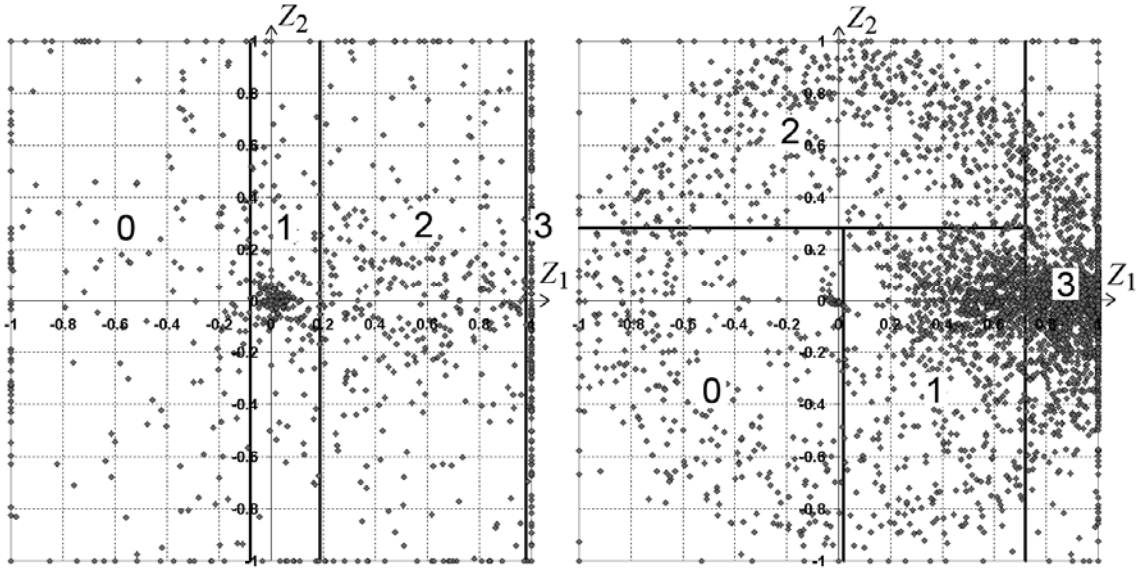



Рис. 1. Области наиболее информативного разбиения в пространстве значений характеристик сигнала для классов 0 – слева, и 1 – справа.

На основе предварительного анализа массива данных было сделано предположение, что указанные переменные соответствуют значениям одной физической величины, измеренной в последовательные моменты времени, т.е. что строки исходной таблицы можно интерпретировать как отрезки двумерного временного ряда длиной 17.

Метод классификации предлагается следующий. Временные ряды каждого класса объединяются в один временной ряд, для которого предложенным выше методом находятся закономерности. На рисунке 1 представлены объединенные временные ряды для обоих классов (отображены только точки, соответствующие значениям ряда, при этом траектории движения не показаны ввиду громоздкости), а также области разбиения пространства характеристик сигнала для моделей по каждому классу отдельно.

В таблице 1 приведены матрицы оценок переходных вероятностей между областями для классов 0 и 1 (левая и правая части таблицы). При этом в первой строке и первом столбце указаны оценки априорных вероятностей нахождения в соответствующих состояниях.

*Табл. 1.* Оценки переходных и априорных вероятностей для классов 0 – слева, и 1 – справа.

| $p_\omega$ | | 0,18 | 0,34 | 0,16 | 0,32 |
|---|---|---|---|---|---|
| | $\omega$ | **0** | **1** | **2** | **3** |
| 0,18 | **0** | 0,33 | 0,27 | 0,12 | 0,27 |
| 0,34 | **1** | 0,14 | 0,57 | 0,07 | 0,22 |
| 0,16 | **2** | 0,15 | 0,16 | 0,50 | 0,19 |
| 0,32 | **3** | 0,14 | 0,23 | 0,11 | 0,52 |

| $p_\omega$ | | 0,08 | 0,24 | 0,15 | 0,52 |
|---|---|---|---|---|---|
| | $\omega$ | **0** | **1** | **2** | **3** |
| 0,08 | **0** | 0,78 | 0,12 | 0,00 | 0,09 |
| 0,24 | **1** | 0,01 | 0,74 | 0,03 | 0,21 |
| 0,15 | **2** | 0,10 | 0,03 | 0,78 | 0,09 |
| 0,52 | **3** | 0,00 | 0,09 | 0,05 | 0,86 |

Для каждой из 351 отдельной реализации вычислялось правдоподобие по отношению к обеим моделям. В пространстве полученных значений правдоподобия строилось решающая функция (правило классификации). При этом ошибка на скользящем экзамене составила 11,4%. В случае априорной классификации (объектам приписывается преобладающий класс 1) ошибка составляет 0,36. Это говорит о том, что построенные логико-вероятностные модели действительно являются информативными и отражают свойства, специфичные для классов.

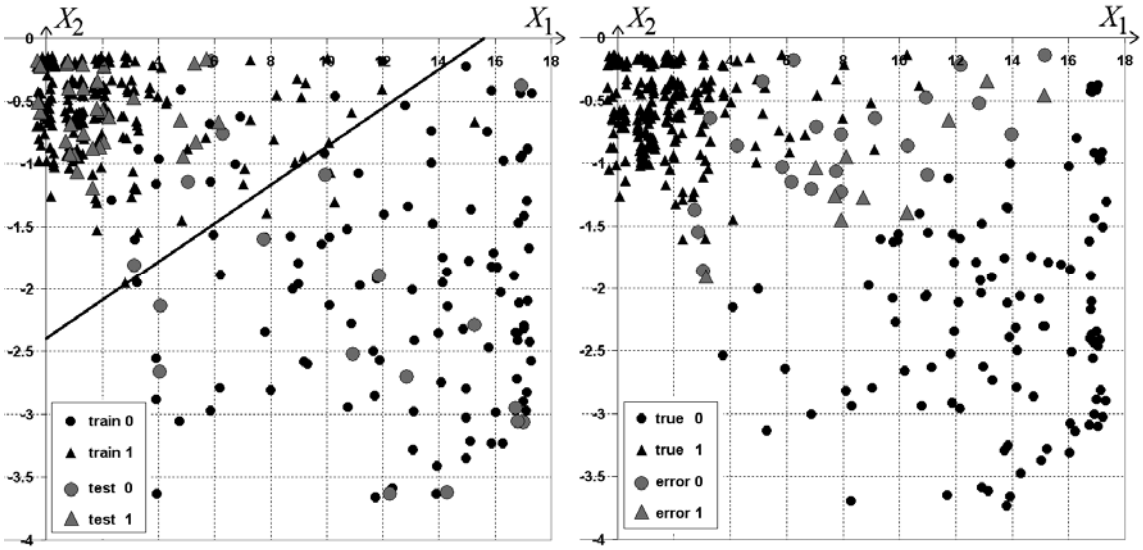



Рис.2. Реализации классов в пространстве значений правдоподобия и голосования по методу ближайшего соседа по подпространствам с выделением контрольной выборки (слева) и скользящего экзамена (справа).

Полученные значения качества классификации близки к значениям, достигаемым другими методами классификации, не интерпретирующими данные как временные ряды, т.е. использующими только «статические» свойства данных. Предложенный метод, наоборот, использует только «динамические» закономерности. Естественно ожидать, что комбинированное использование обоих типов информации может улучшить результат.

В качестве дополнительного метода был взят метод ближайшего соседа. Для этого метода при вычислении расстояния в исходном 34-мерном пространстве ошибка на скользящем экзамене составила 13%; для модификации метода с голосованием по 17 подпространствам ошибка на скользящем экзамене составила 13,7%. Таким образом, эти вариации метода можно считать равноценными по качеству. В случае $k$ ближайших соседей ошибка получается более 13% и растет с ростом $k$.

Вариант метода ближайшего соседа с голосованием по 17 двумерным подпространствам позволяет сформировать признак $X_1$ – количество «голосов» за принадлежность объекта классу 0. Признак $X_2$ – величина правдоподобия объекта к логико-вероятностной модели, построенной по всем реализациям первого класса.

При построении линейного классификатора в пространстве $(X_1, X_1)$ ошибка на скользящем экзамене составила 9,1%. Результаты представлены на рис. 2. На левой диаграмме изображены точки обучающей (train) и контрольной выборки (test) для нулевого и первого классов, а также линейная разделяющая функция. На правой диаграмме изображены объекты, правильно (true) и неправильно (error) классифицированные в процессе скользящего экзамена. При этом значения $(X_1, X_1)$ для каждого объекта соответствуют модели, построенной без использования этого объекта для обучения.

Применение динамической модели, учитывающей упорядоченность данных, дает меньшую ошибку классификации по сравнению с методом ближайшего соседа. При этом одновременный учет статических и динамических свойств данных позволяет еще уменьшить этот показатель.

## Заключение

В работе рассмотрен метод анализа многомерного разнотипного временного ряда, основанный на построении логико-вероятностной модели, представляющей собой марковскую цепь, заданную на состояниях, выбираемых по реализации в соответствии с критерием максимума информативности. На примере решения прикладной задачи продемонстрирована возможность использования метода для формирования признаков, позволяющих классифицировать временные последовательности.

## Благодарности

Статья частично финансированна из проекта **ITHEA XXI** Института Информационных теории и Приложений FOI ITHEA и Консорциума FOI Bulgaria (www.ithea.org, www.foibg.com).

## Литература

[Baum, 1970] L. Baum et. al. A maximization technique occuring in the statistical analysis of probabilistic functions of markov chains. Annals of Mathematical Statistics, 1970. 41. P. 164–171.

[Lbov, Nedel'ko, 2001] G.S. Lbov, V.M. Nedel'ko. A Maximum informativity criterion for the forecasting several variables of different types. // Computer data analysis and modeling. Robustness and computer intensive methods. Minsk, 2001, vol 2. P. 43–48.

[Неделько, 2008] С.В. Неделько. Исследование статистической устойчивости логико-вероятностных моделей временного ряда // Научный вестник НГТУ, 2008, №4(33), с. 43-52.

[Лбов, Старцева, 1999] Г.С. Лбов, Н.Г. Старцева. Логические решающие функции и вопросы статистической устойчивости решений. Институт математики СО РАН, Новосибирск, 1999, 211 с.

## Информация об авторе

***Светлана Неделько*** *– ассистент кафедры Высшей математики НГТУ, 630092, Россия, г. Новосибирск, проспект К. Маркса, 20, e-mail: nedelko@math.nsc.ru*

# ОБ ОДНОЙ ЗАДАЧЕ РАСПОЗНАВАНИЯ ПОСЛЕДОВАТЕЛЬНОСТИ, ВКЛЮЧАЮЩЕЙ ПОВТОРЯЮЩИЙСЯ ВЕКТОР[1]

**Алексей Долгушев, Александр Кельманов**

***Аннотация**: Рассматривается дискретная экстремальная задача, к которой сводится один из вариантов проблемы помехоустойчивого off-line распознавания векторных последовательностей, включающих в качестве элемента квазипериодически повторяющийся вектор евклидова пространства. Обоснован эффективный алгоритм решения задачи, гарантирующий оптимальность решения по критерию максимального правдоподобия в случае, когда помеха аддитивна и является гауссовской последовательностью независимых одинаково распределённых случайных величин.*

***Ключевые слова**: помехоустойчивое распознавание, векторная последовательность, повторяющийся вектор, максимум правдоподобия, дискретная экстремальная задача, off-line алгоритм.*

***ACM Classification Keywords**: F.2. Analysis of Algorithms and Problem Complexity, G.1.6. Optimization, G2. Discrete Mathematics, I.5. Pattern Recognition.*

***Conference**: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

## Введение

Объект исследования работы – проблемы оптимизации в задачах анализа данных и распознавания образов. Предмет исследования – дискретная экстремальная задача, к которой сводится один из вариантов проблемы помехоустойчивого off-line распознавания векторной последовательности, как последовательности, включающей квазипериодически повторяющийся вектор, совпадающий с некоторым вектором из заданного алфавита векторов евклидова пространства. Цель работы – обоснование алгоритма решения этой задачи. Рассматриваемая задача является обобщением задачи, изученной в [1].

Одна из возможных содержательных трактовок задачи состоит в следующем. Источник сообщений через канал связи с помехой передает информацию об активном и пассивном состояниях некоторого физического объекта в виде упорядоченного набора – вектора – измеряемых характеристик. В пассивном состоянии значения каждой компоненты этого вектора равны нулю. Имеется конечная совокупность физических объектов. Каждому объекту соответствует уникальный набор измеряемых информационно важных характеристик. На приёмную сторону поступает зашумлённая последовательность квазипериодически перемежающихся векторов, в которой кроме информационно значимого вектора, соответствующего активному состоянию объекта, имеются посторонние неизвестные ненулевые векторы-вставки. Термин «квазипериодически» означает, что интервал между двумя последовательными ненулевыми векторами не одинаков, а лишь ограничен сверху и снизу некоторыми константами. Число повторов информационно значимого вектора, а также число векторов-вставок известны. Требуется определить (распознать), от какого из объектов была принята последовательность. Ситуации, в которых требуется решение подобной задачи, характерны, в частности, для геофизики, технической диагностики, электронной разведки и других приложений (см., например, [2] и цитированные там работы).

[1] Работа поддержана грантами РФФИ 09-01-00032, 07-07-00022 и грантом АВЦП Рособразования 2.1.1/3235.

## Формальная постановка задачи

Пусть векторная последовательность $\mathbf{x}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$, где $\mathcal{N} = \{1,2,\ldots,N\}$, обладает свойством

$$\mathbf{x}_n = \begin{cases} \mathbf{u}, & n \in \mathcal{M}_1, \\ \mathbf{w}_n, & n \in \mathcal{M}_2, \\ \mathbf{0}, & n \in \mathcal{N} \setminus (\mathcal{M}_1 \cup \mathcal{M}_2), \end{cases} \quad (1)$$

где $\mathcal{M}_1 \cup \mathcal{M}_2 \subseteq \mathcal{N}$, $\mathcal{M}_1 \cap \mathcal{M}_2 = \varnothing$.

Допустим, что $\mathbf{u} \in \mathcal{A}$, $\mathcal{A} \subset \{\mathbf{u} \mid \mathbf{u} \in \mathcal{R}^q, 0 < \| \mathbf{u} \|^2 < \infty\}$, где $\|\cdot\|$ – норма вектора, и $|\mathcal{A}| = K$. Пусть $\mathbf{w}_n \in \{\mathbf{w} \mid \mathbf{w} \in \mathcal{R}^q, 0 < \| \mathbf{w} \|^2 < \infty\}$, $n \in \mathcal{M}_2$.

Положим $|\mathcal{M}_j| = M_j$, $j = 1, 2$, и $M = M_1 + M_2$. Вектор $\mathbf{u}$ будем интерпретировать как информационно значимый вектор, множество $\mathcal{A}$ – как алфавит информационно значимых векторов, вектор $\mathbf{w}_n$, $n \in \mathcal{M}_2$, – как вектор-вставку, а $M_1$ и $M_2$ – соответственно как число повторов информационно значимого вектора и число векторов-вставок в последовательности (1). Допустим, кроме того, что элементы набора $(n_1,\ldots,n_M)$, образующего совокупность $\{n_1,\ldots,n_M\} = \mathcal{M}_1 \cup \mathcal{M}_2$, удовлетворяют ограничениям

$$1 \le T_{\min} \le n_m - n_{m-1} \le T_{\max} \le N - 1, \quad m = 2,\ldots,N. \quad (2)$$

Ограничения (2), в которых $T_{\min}$ и $T_{\max}$ – константы, задают допустимый интервал между ближайшими номерами двух ненулевых векторов последовательности (1).

Доступной для анализа будем считать последовательность

$$\mathbf{y}_n = \mathbf{x}_n + \mathbf{e}_n, \quad n \in \mathcal{N}, \quad (3)$$

где $\mathbf{e}_n$ – вектор помехи (ошибки измерения), независимый от вектора $\mathbf{x}_n$. Заметим, что $\mathbf{x}_n = \mathbf{x}_n(\mathcal{M}_1, \mathcal{M}_2, \mathbf{u}, \{\mathbf{w}_n, n \in \mathcal{M}_2\})$, $n \in \mathcal{N}$. Положим

$$S(\mathcal{M}_1, \mathcal{M}_2, \mathbf{u}, \{\mathbf{w}_n, n \in \mathcal{M}_2\}) = \sum_{n \in \mathcal{N}} \| \mathbf{y}_n - \mathbf{x}_n \|^2 \quad (4)$$

и рассмотрим следующую задачу.

**Задача 1**. *Дано*: последовательность $\mathbf{y}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$, и множество (алфавит) $\mathcal{A}$, $|\mathcal{A}| = K$. *Найти*: вектор $\mathbf{u} \in \mathcal{A}$, непересекающиеся подмножества $\mathcal{M}_1$ и $\mathcal{M}_2$ множества $\mathcal{N}$, а также множество $\{\mathbf{w}_n \mid n \in \mathcal{M}_2, \mathbf{w}_n \in \mathcal{R}^q, 0 < \| \mathbf{w}_n \|^2 < \infty\}$ такие, что целевая функция (4) минимальна при ограничениях (2) на элементы набора $(n_1,\ldots,n_M)$, которые образуют совокупность $\{n_1,\ldots,n_M\} = \mathcal{M}_1 \cup \mathcal{M}_2$.

К этой задаче сводится один из вариантов проблемы помехоустойчивого распознавания последовательности (1), как структуры, включающей повторяющийся ненулевой вектор, совпадающий с некоторым элементом из заданного алфавита векторов, которая кроме этого вектора содержит неизвестные ненулевые векторы-вставки. В [3] показано, что к решению задачи 1 приводит статистическая формулировка проблемы, если считать, что $\mathbf{e}_n$ в формуле (3) есть выборка из $q$-мерного нормального распределения с параметрами $(\mathbf{0}, \sigma^2\mathbf{I})$, где $\mathbf{I}$ - единичная матрица, а в качестве критерия решения задачи использовать максимум функционала правдоподобия.

## Редуцированная экстремальная задача

Нетрудно аналитически убедиться, что для любых допустимых $\mathcal{M}_1$ и $\mathcal{M}_2$ при фиксированном векторе $\mathbf{u} \in \mathcal{A}$ минимум функционала (4) по неизвестным векторам $\mathbf{w}_n$, $n \in \mathcal{M}_2$, доставляется векторами $\hat{\mathbf{w}}_n = \mathbf{y}_n$, $n \in \mathcal{M}_2$, и равен

$$S_{\min}(\mathcal{M}_1, \mathcal{M}_2, \mathbf{u}) = \sum_{n \in \mathcal{N}} \| \mathbf{y}_n \|^2 - \sum_{n \in \mathcal{M}_1} \{2\langle \mathbf{y}_n, \mathbf{u} \rangle - \| \mathbf{u} \|^2\} - \sum_{n \in \mathcal{M}_2} \| \mathbf{y}_n \|^2, \quad (5)$$

где $\langle \cdot, \cdot \rangle$ – скалярное произведение векторов.

Первый член в правой части равенства (5) является константой. Поэтому имеем следующую оптимизационную задачу, к которой сводится минимизация целевой функции (4).

**Задача SVVGA** (Searching for a Vector in a Given Vector Alphabet). *Дано*: последовательность $\mathbf{y}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$, множество (алфавит) $\mathcal{A}$, $|\mathcal{A}| = K$, векторов из $\mathcal{R}^q$ и натуральные числа $M_1$ и $M_2$. *Найти*: вектор $\mathbf{u} \in \mathcal{A}$ и непересекающиеся подмножества $\mathcal{M}_1$ и $\mathcal{M}_2$ множества $\mathcal{N} = \{1,2,\ldots,N\}$, доставляющие максимум целевой функции

$$G(\mathcal{M}_1, \mathcal{M}_2, \mathbf{u}) = \sum_{n \in \mathcal{M}_1} (2\langle \mathbf{y}_n, \mathbf{u} \rangle - \| \mathbf{u} \|^2) + \sum_{n \in \mathcal{M}_2} \| \mathbf{y}_n \|^2 , \quad (6)$$

при условии, что имеют место ограничения $|\mathcal{M}_1| = M_1$, $|\mathcal{M}_2| = M_2$ на мощности искомых подмножеств, а элементы этих подмножеств образуют объединенный набор номеров $(n_1,\ldots,n_M)$ размерности $M = M_1 + M_2$, компоненты которого удовлетворяют ограничениям (2).

## Алгоритм решения задачи

Положим

$$g_1(n, \mathbf{u}) = 2\langle \mathbf{y}_n, \mathbf{u} \rangle - \| \mathbf{u} \|^2, \; g_2(n) = \| \mathbf{y}_n \|^2 \;, \; \mathbf{u} \in \mathcal{A}, \; n \in \mathcal{N} .$$

Тогда целевую функцию (6) можно переписать в виде

$$G(\mathcal{M}_1, \mathcal{M}_2, \mathbf{u}) = \sum_{n \in \mathcal{M}_1} g_1(n, \mathbf{u}) + \sum_{n \in \mathcal{M}_2} g_2(n) \; . \quad (7)$$

В настоящей работе показано, что максимум $G_{\max}$ целевой функции (7) вычисляется по формуле

$$G_{\max} = \max_{\mathbf{u} \in \mathcal{A}} G_{\max}(\mathbf{u}), \quad (8)$$

где

$$G_{\max}(\mathbf{u}) = \max_{\mathcal{M}_1, \mathcal{M}_2} G(\mathcal{M}_1, \mathcal{M}_2 \,|\, \mathbf{u}), \; \mathbf{u} \in \mathcal{A},$$

– условный максимум функции (7), который находится по правилу

$$G_{\max}(\mathbf{u}) = \max_{n \in \omega_M(M)} \max_{t \in \{M_1, M_1+1, \ldots, M\}} G_n(M_1, t, M \,|\, \mathbf{u}), \quad (9)$$

где $M = M_1 + M_2$. Значения функции $G_n(M_1, t, M \,|\, \mathbf{u})$, $n \in \omega_M(M)$, $t \in \{M_1, M_1+1, \ldots, M\}$, при каждом фиксированном $\mathbf{u} \in \mathcal{A}$ вычисляются по рекуррентным формулам

$$G_n(l,t,m\,|\,\mathbf{u})=\begin{cases} g_1(n,\mathbf{u}), & \text{если } l=1,t=1,m=1,\\ g_1(n,\mathbf{u})+\max\limits_{j\in\gamma^-_{m-1}(n)} G_j(0,0,m-1), & \text{если } l=1,t=2,\ldots,M,m=t,\\ g_1(n,\mathbf{u})+\max\limits_{j\in\gamma^-_{m-1}(n)}\max\limits_{s\in\{l-1,\ldots,m-1\}} G_j(l-1,s,m-1\,|\,\mathbf{u}), & \\ \qquad\qquad \text{если } l=2,\ldots,M_1,t=l,\ldots,M,m=t,\\ g_2(n)+\max\limits_{j\in\gamma^-_{m-1}(n)} G_j(l,t,m-1\,|\,\mathbf{u}), & \\ \qquad\qquad \text{если } l=1,\ldots,M_1,\ t=l,\ldots,M,m=t+1,\ldots,M, \end{cases}$$

где

$$G_n(0,0,m)=\begin{cases} g_2(n), & \text{если } m=1,\\ g_2(n)+\max\limits_{j\in\gamma^-_{m-1}(n)} G_j(0,0,m-1), & \text{если } m=2,\ldots,M, \end{cases}$$

в которых $n\in\omega_m(M)$, причём

$$\omega_m(M)=\{n\,|\quad 1+(m-1)T_{\min}\le n\le N-(M-m)T_{\min}\},\ m=1,\ldots,M,$$

$$\gamma^-_{m-1}(n)=\{k\,|\quad \max\{1+(m-2)T_{\min},\,n-T_{\max}\}\le k\le n-T_{\min}\},\ n\in\omega_m(M),\ m=1,\ldots,M.$$

Так как $\{n_1,\ldots,n_M\}=\mathcal{M}_1\cup\mathcal{M}_2$, поиск непересекающихся подмножеств $\mathcal{M}_1=\{n_{\mu_1},\ldots,n_{\mu_{M_1}}\}$ и $\mathcal{M}_2=\{n_{\nu_1},\ldots,n_{\nu_{M_2}}\}$ множества $\mathcal{N}$ эквивалентен поиску объединённого набора $(n_1,\ldots,n_M)$ и одного из подмножеств $\{\mu_1,\ldots,\mu_{M_1}\}$ или $\{\nu_1,\ldots,\nu_{M_2}\}$ множества $\{1,2,\ldots,M\}$; пусть для определённости искомым является подмножество $\{\mu_1,\ldots,\mu_{M_1}\}$. Для каждого фиксированного $\mathbf{u}\in\mathcal{A}$ определим функции:

$$\widetilde{G}_j(l,m\,|\,\mathbf{u})=\max_{s\in\{l-1,\ldots,m-1\}} G_j(l-1,s,m-1\,|\,\mathbf{u}),$$

$$K_j(l,m\,|\,\mathbf{u})=\arg\max_{s\in\{l-1,\ldots,m-1\}} G_j(l-1,s,m-1\,|\,\mathbf{u}),$$

$$l=2,\ldots,M_1,\ \ m=l,\ldots,M,\ \ j\in\gamma^-_{m-1}(n),$$

где $n\in\omega_m(M)$;

$$I_n(l,t,m\,|\,\mathbf{u})=\begin{cases} n, & \text{если } l=1,t=1,m=1,\\ \arg\max\limits_{j\in\gamma^-_{m-1}(n)} G_j(0,0,m-1), & \text{если } l=1,t=2,\ldots,M,m=t,\\ \arg\max\limits_{j\in\gamma^-_{m-1}(n)} \widetilde{G}_j(l,m\,|\,\mathbf{u}), & \text{если } l=2,\ldots,M_1,t=l,\ldots,M,m=t,\\ \arg\max\limits_{j\in\gamma^-_{m-1}(n)} G_j(l,t,m-1\,|\,\mathbf{u}), & \text{если } l=1,\ldots,M_1,\ t=l,\ldots,M,m=t+1,\ldots,M, \end{cases}$$

где $n\in\omega_m(M)$;

$$J_n(l,m\,|\,\mathbf{u})=K_{I_n(l,m,m|\mathbf{u})}(l,m\,|\,\mathbf{u}),\qquad l=2,\ldots,M_1,\ \ m=l,\ldots,M.$$

где $n\in\omega_m(M)$.

Тогда в соответствии с (8) искомый вектор $\hat{\mathbf{u}}$ находится по правилу

$$\hat{\mathbf{u}} = \arg\max_{\mathbf{u}\in\mathcal{A}} G_{\max}(\mathbf{u}).$$

Последние компоненты оптимальных наборов $(\hat{n}_1,\ldots,\hat{n}_M)$ и $(\hat{\mu}_1,\ldots,\hat{\mu}_{M_1})$ согласно (9) определяются по формуле

$$(\hat{n}_M,\hat{\mu}_{M_1}) = \arg\max_{n\in\omega_M(M)}\max_{t\in\{M_1,M_1+1,\ldots,M\}} G_n(M_1,t,M\,|\,\hat{\mathbf{u}}).$$

Остальные компоненты (при $M_1>1$) искомых наборов находятся по следующему правилу:

$$\hat{n}_{m-1} = \begin{cases} I_{\hat{n}_m}(M_1,\hat{\mu}_{M_1},m\,|\,\hat{\mathbf{u}}), & m = M, M-1,\ldots,\hat{\mu}_{M_1}, \\ I_{\hat{n}_m}(l,\hat{\mu}_l,m\,|\,\hat{\mathbf{u}}), & l = L-1,\ldots,2,\ m=\hat{\mu}_{l+1}-1,\ldots,\hat{\mu}_l; \end{cases}$$

$$\hat{\mu}_{l-1} = J_{\hat{n}_{\hat{\mu}_l}}(l,\hat{\mu}_l\,|\,\hat{\mathbf{u}}), \qquad l = L, L-1,\ldots,2.$$

При этом, если $\hat{\mu}_2-\hat{\mu}_1>1$, то

$$\hat{n}_{m-1} = I_{\hat{n}_m}(1,\hat{\mu}_1,m\,|\,\hat{\mathbf{u}}), \qquad m=\hat{\mu}_2-1,\ldots,\hat{\mu}_1+1,$$

а если $\hat{\mu}_1>1$, то

$$\hat{n}_{m-1} = I_{\hat{n}_m}(1,m,m\,|\,\hat{\mathbf{u}}), \qquad m=\hat{\mu}_1,\ldots,2.$$

Временная сложность алгоритма решения задачи SVVGA есть величина $O[\min\{M_1,M_2\}K(M_1+M_2)^2(T_{\max}-T_{\min}+q)N]$.

Алгоритм решения задачи SVVGA лежит в основе процедуры помехоустойчивого распознавания структурированных данных в виде векторных последовательностей, включающих квазипериодически повторяющийся ненулевой информационно значимый вектор, совпадающий с некоторым вектором из заданного алфавита векторов евклидова пространства. Эта алгоритм гарантирует оптимальность решения по критерию максимального правдоподобия в случае, когда помеха аддитивна и является гауссовской последовательностью независимых одинаково распределенных величин.

## Заключение

Рассмотренная задача входит в большое семейство актуальных задач [3-5], к которым сводятся типовые проблемы помехоустойчивого off-line анализа и распознавания структурированных данных в виде числовых и векторных последовательностей, включающих повторяющиеся, чередующиеся и перемежающиеся информационно значимые фрагменты или векторы. В настоящей работе представлено алгоритмическое решение одной из таких ранее неизученных задач: обоснован точный полиномиальный алгоритм, который является ядром помехоустойчивого алгоритма распознавания.

## Благодарности

Работа поддержана грантами РФФИ 09-01-00032, 07-07-00022 и грантом АВЦП Рособразования 2.1.1/3235.

## Литература

[1] Kel'manov A.V., Khamidullin S.A. Recognizing a Quasiperiodic Sequence Composed of a Given Number of Identical Subsequences // Pattern Recognition and Image Analysis, 2000. Vol.10, No.1. P. 127-142.

[2] Kel'manov A.V., Jeon B. A Posteriori Joint Detection and Discrimination of Pulses in a Quasiperiodic Pulse Train // IEEE Transactions on Signal Processing, Vol. 52, No. 3, March 2004, P. 1-12.

[3] Кельманов А.В. Полиномиально разрешимые и NP-трудные варианты задачи оптимального обнаружения в числовой последовательности повторяющегося фрагмента // Материалы Росс. конф. «Дискретная оптимизация и исследование операций» (Владивосток, 7-14 сентября 2007). – Новосибирск: Изд-во Института математики СО РАН, 2007.– http://math.nsc.ru/ conference/ door07/ DOOR_abstracts.pdf. С. 46-50.

[4] Кельманов А.В. О некоторых полиномиально разрешимых и NP-трудных задачах анализа и распознавания последовательностей с квазипериодической структурой // Доклады 13-ой Всеросс. конф. «Математические методы распознавания образов». Ленинградская обл., г. Зеленогорск, 30 сентября – 6 октября 2007г. М.: МАКС Пресс, – 2007. С. 261-264.

[5] http://math.nsc.ru/~serge/qpsl

## Информация об авторах

***Алексей Долгушев*** *– аспирант, Новосибирский государственный университет, ул. Пирогова, 2, Новосибирск, 630090, Россия, e-mail: dolgushev@math.nsc.ru*

***Александр Кельманов*** *– д.ф.-м.н., ведущий научный сотрудник, Институт математики им. С.Л. Соболева Сибирского отделения РАН, проспект академика Коптюга, 4, Новосибирск, 630090, Россия; Новосибирский государственный университет, ул. Пирогова, 2, Новосибирск, 630090, Россия; e-mail: kelm@math.nsc.ru*

# Features Processing and Transformations

## AN APPROACH TO VARIABLE AGGREGATION IN EFFICIENCY ANALYSIS

**Veska Noncheva, Armando Mendes, Emiliana Silva**

***Abstract**: In the nonparametric framework of Data Envelopment Analysis the statistical properties of its estimators have been investigated and only asymptotic results are available. For DEA estimators results of practical use have been proved only for the case of one input and one output. However, in the real world problems the production process is usually well described by many variables. In this paper a machine learning approach to variable aggregation based on Canonical Correlation Analysis is presented. This approach is applied for efficiency estimation of all the farms in Terceira Island of the Azorean archipelago.*

***Keywords**: Canonical Correlation Analysis, Data Envelopment Analysis, Efficiency, Variable Aggregation*

***ACM Classification Keywords**: H.2.8 Data mining, G.3 Multivariate statistics, G.4 Efficiency*

***Conference**: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

### Introduction

Data Envelopment Analysis (DEA) is becoming an increasingly popular management tool. It is a mathematical programming based technique. The task of the DEA is to evaluate the relative performance of units of a system. It has useful applications in many evaluation contexts.

DEA makes it possible to identify efficient and inefficient units in a framework where results are considered in their particular context. The units to be assessed should be relatively homogeneous and are originally called Decision Making Units (DMUs). DMUs can be manufacturing units, departments of a big organization such as universities, schools, bank branches, hospitals, medical practitioners, power plants, police stations, tax offices, hotels, or a set of farms. DEA is an extreme point method and compares each DMU with only the "best" DMUs.

DEA can be a powerful tool when used wisely. A few of the characteristics that make it powerful are:

- DEA can handle multiple input and multiple output models.
- DMUs are directly compared against a peer or combination of peers.
- Inputs and outputs can have very different units. For example, one variable could be in units of lives saved and another could be in units of dollars without requiring an a priori tradeoff between the two.

The same characteristics that make DEA a powerful tool can also create problems. An analyst should keep these limitations in mind when choosing whether or not to use DEA.

- Since DEA is an extreme point technique, noise such as measurement error can cause significant problems.
- When the number of inputs or outputs is increased, the number of observations must increase at an exponential rate.

- For DEA estimators, useful theoretical results have been obtained only for the case of one input and one output variable.

The approach presented in this paper is focused on measuring efficiency when the number of DMUs is few and the number of explanatory variables needed to compute the measure of efficiency is too large. We approach this problem from a statistical standpoint through variable aggregation. The aggregation in our approach is not fixed.

## Variable Aggregation in DEA

DEA estimators are biased by construction. When the number of exploratory variables is large, unless a very large quantity of data are available, the resulting imprecision will manifest itself in the form of large bias, large variance, and very wide confidence intervals (see [Simar and Wilson, 2008]). Because of it, the question of obtaining an appropriate aggregate input and aggregate output from appropriate individual inputs and outputs, respectively, is an important one. A natural way to define an aggregate input (or an aggregate output) is to assume a linear structure of aggregation of the input variables (and outputs, respectively). One of the most important issues here is the choice of weights in the aggregation.

A subtle technique for the aggregation of inputs or outputs is the use of weight restrictions. This way the unimportant variables will still count in the overall model but only up to the specified limit of 'importance'. Weights choice may be done by the researcher according his opinion about the contribution of each variable. In our machine learning approach the weights are not fixed. They are extracted from data describing the production process under investigation. To achieve this aim we apply Canonical Correlations Analysis (CCA) to aggregate automatically both input and output data sets.

Obviously the input and output sets of variables in a production process are related. We are concerned with determining a relationship between the two sets of variables. The aim is the linear combinations that maximize the canonical correlation to be fond. In CCA such a linear combination is called "canonical variate" and in DEA it will be used as an aggregate variable.

In this paper, we propose CCA to aggregate both input and output variables in order to get final input and output, respectively.

## Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a multidimensional exploratory statistical method. A canonical correlation is the correlation of two latent variables, one representing a set of independent variables, the other a set of dependent variables. The canonical correlation is optimized such that the linear correlation between the two latent variables (called canonical variates) is maximized. There may be more canonical variates relating the two sets of variables. The purpose of canonical correlation is to explain the relation of the two sets of original variables. For each canonical variate we can also assess how strongly it is related to measured variables in its own set, or the set for the other canonical variate.

Both methods Principal Components Analysis (PCA) and CCA have the same mathematical background. The main purpose of CCA is the exploration of sample correlations between two sets of quantitative variables, whereas PCA deals with one data set in order to reduce dimensionality through linear combination of initial variables.

Another well known method can deal with quantitative data. It is Partial Least Squares (PLS) regression. However, the object of PLS regression is to explain one or several response variables (outputs) in one set, by variables in the other one (the input). On the other hand, the object of CCA is to explore correlations between two

sets of variables whose roles in the analysis are strictly symmetric. As a consequence, mathematical principles of both PLS and CCA methods are fairly different.

The canonical coefficients of a canonical variate are standardized coefficients and their magnitudes can be compared. However, the canonical coefficients may be subject to multicollinearity, leading to incorrect judgments. Also, because of suppression, a canonical coefficient may even have a different sign compared to the correlation of the original variable with the canonical variable. Therefore, instead, we interpret the relations of the original variables to a canonical variable in terms of the correlations of the original variables with the canonical variables - that is, by structure coefficients.

## Example: Terceira's Farms' Efficiency Measurement

Terceira is the second biggest island in the Azorean archipelago. The Azores islands belong to Portugal with a population of about 250000 inhabitants. The most part (about 75%) of this population is in S. Miguel and Terceira islands. The main economic activity is dairy and meat farming. In S. Miguel, Terceira and S. Jorge islands, about 24% of the farms produce only milk, other 13% of farms produce only meat and 24% produce both and other cultures as well. The remaining farms produce other agricultural productions. Dairy policy depends on Common Agricultural Policy of the European Union and is limited by quotas. In this context, decision makers need knowledge for deciding the best policies in promoting quality and best practices. One of the goals of our work is to provide Azorean Government with a reliable tool for measurement of productive efficiency of the farms.

In Azores there are about 15.107 farmers. Azorean farms are small - about 8 hectares per farm, what is about the half of the average European farm dimension (15.8 in 2003). The production system is primarily based on grazing (about 95% of the area). In the last years, the most representative expenses – based in data of FADN (Farm Account Database Network) are on concentrates, annual depreciation, rents and fertilizers. The subsidies are important for the dairy farms, and in 2004 they were about 61.6% of all profit. Some of these subsidies are compensations for low selling prices received by farmers, and so they are due after the production of meat and milk, others are incentives to investment and compensation for high prices of production factors. There are also subventions to improve ecological production.

Some research work on the dairy sector in Azores has been already done ([Marote and Silva, 2002], [Silva, et al. 2001]). The beef sector in Azores has been investigated by means of Stochastic Frontier Analysis ([Silva, 2004])).

Any resource used by an Azorean dairy farm is treated as an input variable and because of it the list of variables that provide an accurate description of the milk and meat production process is large. The names of all input variables used in the analysis are the following: EquipmentRepair, Oil, Lubricant, EquipmentAmortization, AnimalConcentrate, VeterinaryAndMedicine, OtherAnimalCosts, PlantsSeeds, Fertilizers, Herbicides, LandRent, Insurance, MilkSubsidy, MaizeSubsidy, SubsidyPOSEIMA, AreaDimension, and DairyCows. The names of output variables are Milk and Cattle. The number of all farms in Terceira is 30.

The analysis of the Terceira's farms efficiency is implemented in R statistical software version 2.8.1 using the DEA, FEAR and CCA packages and routines developed by the authors (see [R Development Core Team, 2007]).

Outliers may influence the results. Because of it we start the data analysis with outlier detection. One outlier obtained in Terceira data was the result of a recording error and it was corrected. We used again the statistical methodology presented in [Wilson, 1993] and implemented in FEAR package to look for new atypical observations. Using the graphical analysis presented in **Figure 1** another three observations could also be identified as outliers. However data from Terceira Island are viewed as having come from a probability distribution and it is quite possible to observe few points with low probability. One would not expect to observe many such points, given their low probability. The fact that a particular observation has low probability of occurrence is not

sufficient to warrant the conclusion that this observation is an error. More errors in the available data are not identified.

The application of canonical correlation analysis aims at highlighting correlations between input and output data sets, called X and Y, respectively. Two preliminary steps calculate the sample correlation coefficients and visualise the correlation matrixes. All sample correlation coefficients are presented in **Table 1** and the correlation matrixes are visualised in **Figure 2**.

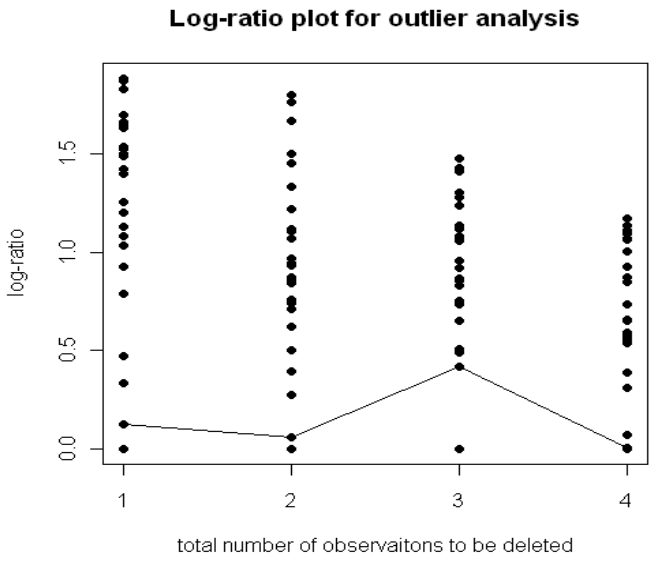



**Figure 1**. Plot produced by the outlier detection procedure.

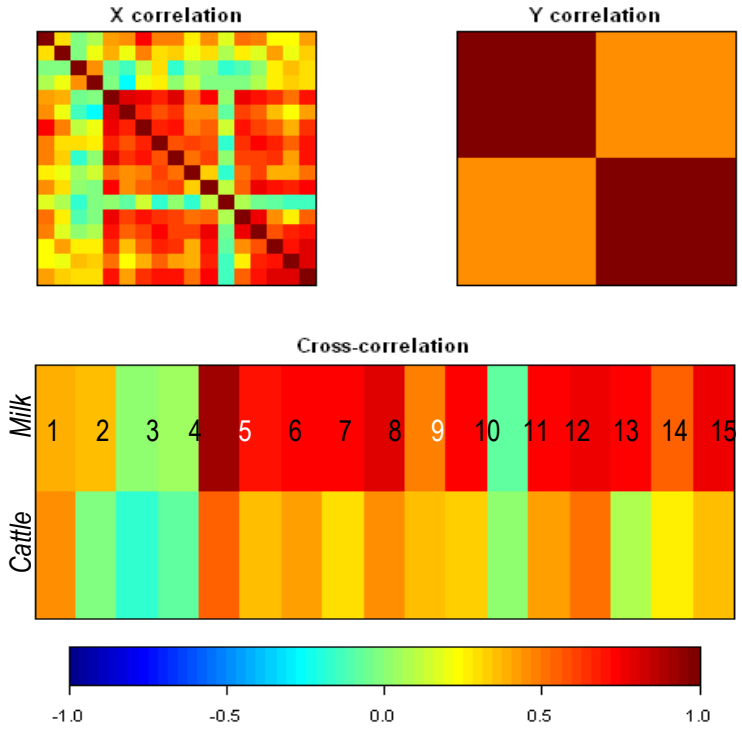



**Figure 2**. Visualisation of sample correlation coefficients.

**Table 1**. Sample correlation coefficients and correlations of the original inputs with both aggregated input and output.

| | Orriginal input variables | Sample correlation coefficient with *Milk* variable | Sample correlation coefficient with *Cattle* variable | Correlation with the aggregated input (structure weight) | Correlation with the aggregated output (structure weight) |
|---|---|---|---|---|---|
| *1* | *EquipmentRepair* | 0.399089550 | 0.449336923 | -0.44487248 | -0.42591381 |
| *2* | *Oil* | 0.349190515 | -0.023206764 | -0.34213524 | -0.32755482 |
| *3* | *Lubricant* | 0.009272362 | -0.171455723 | 0.01024649 | 0.00980983 |
| *4* | *EquipmentAmortization* | 0.051043354 | -0.077088336 | -0.04167289 | -0.03989696 |
| *5* | *AnimalConcentrate* | 0.914685924 | 0.537983929 | -0.96395974 | -0.92287966 |
| *6* | *VeterinaryAndMedicine* | 0.707943660 | 0.370392398 | -0.74087590 | -0.70930276 |
| *7* | *OtherAnimalCosts* | 0.724266952 | 0.407358115 | -0.76117503 | -0.72873682 |
| *8* | *PlantsSeeds* | 0.719946680 | 0.304399253 | -0.74525915 | -0.71349921 |
| *9* | *Fertilizers* | 0.781448807 | 0.452145566 | -0.82269954 | -0.78763940 |
| *10* | *Herbicides* | 0.497643020 | 0.347245965 | -0.53062365 | -0.50801061 |
| *11* | *LandRent* | 0.722516988 | 0.343699321 | -0.75224389 | -0.72018629 |
| *12* | *Insurance* | -0.072519332 | 0.002379461 | 0.07133021 | 0.06829041 |
| *13* | *MilkSubsidy* | 0.746508776 | 0.431464776 | -0.78586254 | -0.75237225 |
| *14* | *MaizeSubsidy* | 0.751413121 | 0.526768325 | -0.80148885 | -0.76733263 |
| *15* | *SubsidyPOSEIMA* | 0.724407535 | 0.083726114 | -0.72469294 | -0.69380945 |
| *16* | *AreaDimension* | 0.536678292 | 0.279164537 | -0.56145996 | -0.53753280 |
| *17* | *DairyCows* | 0.776032879 | 0.348513730 | -0.80562574 | -0.77129323 |

**Figure 2** highlights a significant correlation between Milk and AnimalConcentrate and nearly null correlation between Milk and Lubricant, Milk and EquipmentAmortization, and Milk and Insurance.

The correlation coefficient between the two canonical variates, presenting the production process of Terceira farms, is 0.957.

The canonical weights (canonical coefficients) explain the unique contributions of original variables to the canonical variable. In this example the small canonical coefficients are a result of existing multicollinearity. Some canonical coefficients even have a different sign compared to the correlation of the original variable with the canonical variable. Therefore, we follow the standard approach to interpreting the relations of the original variables to a canonical variable in terms of the correlations of the original variables with the canonical variables - that is, by structure coefficients. The structure weights explain the simple, overall correlation of the original variables with the canonical variable. The structure weights are reported in **Table 1** and **Table 2**. The canonical weights are reported in **Table 3**. From the first two tables we can conclude that both canonical variates are predominantly associated with the following original inputs: *Animal Concentrate, Fertilizers, DairyCows, MaizeSubsidy, MilkSubsidy, OtherAnimalCosts, PlantsSeeds, LandRent, VeterinaryAndMedicine, SubsidyPOSEIMA* and with the original output variable *Milk*.

Computational aspects of the canonical correlation analysis are implemented in CCA package in R (see [González et al., 2008]).

Both, the original inputs and outputs are aggregated into overall measures called aggregate input variate and aggregate output variate, respectively.

**Table 2**. Correlations of the original outputs with both aggregated input and output.

| Original outputs | Correlations with the aggregated input (structure weights) | Correlations with the aggregated output (structure weights) |
|---|---|---|
| ***Milk*** | ***-0.9529591*** | ***-0.9953781*** |
| Cattle | -0.5225409 | -0.5458007 |

**Table 3**. Canonical weights.

| Input variables (X) | Estimated coefficients for the input variate | Output variables (Y) | Estimated coefficients for the output variate |
|---|---|---|---|
| *EquipmentRepair* | 2.839421e-05 | *Milk* | -3.419875e-05 |
| *Oil* | 1.549179e-05 | *Cattle* | 3.778954e-05 |
| *Lubricant* | 1.199566e-03 | | |
| *EquipmentAmortization* | -3.131292e-06 | | |
| *AnimalConcentrate* | -8.497169e-05 | | |
| *VeterinaryAndMedicine* | 1.473172e-05 | | |
| *OtherAnimalCosts* | -5.441544e-06 | | |
| *PlantsSeeds* | -1.021208e-04 | | |
| *Fertilizers* | -1.305625e-06 | | |
| *Herbicides* | 6.589684e-04 | | |
| *LandRent* | 2.583145e-05 | | |
| *Insurance* | 1.655867e-04 | | |
| *MilkSubsidy* | 2.115323e-05 | | |
| *MaizeSubsidy* | -3.555158e-04 | | |
| *SubsidyPOSEIMA* | -6.560970e-05 | | |
| *AreaDimension* | 3.092947e-04 | | |
| *DairyCows* | -2.520118e-02 | | |

Then we use aggregated input and output in the BCC DEA model presented in [Cooper et al., 2007] and described bellow.

An input oriented DEA model aims to minimise inputs while satisfying at least the given output levels. As we mentioned above the dairy policy in Azorean Islands depends on Common Agricultural Policy of the European Union and it is limited by quotas. Because of it we apply an input oriented DEA model.

The input-oriented BCC model evaluates the efficiency of $DMU_o$, $o$=1,…,$n$, by solving the linear program:

$$\min\ \theta_B, \text{ subject to } \theta_B x_o - X\lambda \geq 0,\ Y\lambda \geq y_o,\ e\lambda = 1,\ \lambda \geq 0,$$

where $\theta_B$ is a scalar, $\lambda$ is a column vector with all elements non-negative, $e$ is a row vector with all elements unity, and $n$ is the number of DMUs.

The BCC problem is solved using a two-phase procedure. In the first phase, we minimise $\theta_B$ and, in the second phase, we maximise the sum of the input excesses $s^-$ and output shortfalls $s^+$, keeping $\theta_B=\theta_B^*$. Here $\theta_B^*$ is the

optimal value obtained in the first phase. An optimal BCC solution is represented by ($\theta_B^*$, $\lambda^*$, $s^{-*}$, $s^{+*}$), where $s^{-*}$ and $s^{+*}$ represent the maximal input excesses and output shortfalls, respectively. If an optimal BCC solution ($\theta_B^*$, $\lambda^*$, $s^{-*}$, $s^{+*}$) satisfies $\theta_B^*=1$, $s^{-*}=0$, and $s^{+*}=0$, then the $DMU_o$ is called BCC-efficient. The sum $s^{-*} + s^{+*}$ , called slack, may essentially be viewed as allocative inefficiency.

Computational aspects of the BCC model are implemented in both DEA and FEAR packages in R.

We build the DEA analysis on aggregated measures. **Table 4** contains the DEA estimates of efficiency. All slacks are zeros. The farms 3, 8,14,17 and 20 are BCC-efficient.

For purposes of efficiency measurement, the upper boundary of the production set is of interest. The efficient frontier is the locus of optimal production plans (e.g., minimal achievable input level for a given output) and it is visualised on **Figure 3**.

**Table 4**. Efficiency of Terceira's farms.

| *DMU* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* |
|---|---|---|---|---|---|---|---|---|---|---|
| Efficiency | 0.885 | 0.866 | 1.000 | 0.971 | 0.916 | 0.874 | 0.941 | 1.000 | 0.883 | 0.975 |
| *DMU* | *11* | *12* | *13* | *14* | *15* | *16* | *17* | *18* | *19* | *20* |
| Efficiency | 0.867 | 0.824 | 0.845 | 1.000 | 0.894 | 0.896 | 1.00 | 0.899 | 0.998 | 1.000 |
| *DMU* | *21* | *22* | *23* | *24* | *25* | *26* | *27* | *28* | *29* | *30* |
| Efficiency | 0.960 | 0.861 | 0.861 | 0.890 | 0.870 | 0.882 | 0.962 | 0.882 | 0.858 | 0.782 |

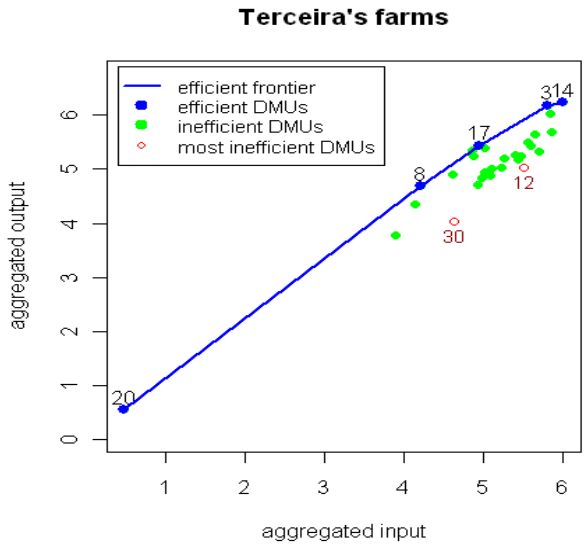



**Figure 3** . The DEA estimator of the production set obtained by the BCC Model.

## Conclusion

In our approach to efficiency measurement CCA provides an aggregation of both input and output units and then DEA provides efficient units. The aggregation can cause additional bias in an DMU's technical efficiency scores. The effects of the input aggregation on efficiency indicators have not been investigated. Estimating the aggregation bias is a question of our future research.

## Acknowledgments

This work has been partially supported by Direcção Regional da Ciência e Tecnologia of Azores Government through the project M.2.1.2/I/009/2008.

## Bibliography

[Cooper et al., 2007] Cooper, W. W., Seiford, L. M. and Tone, K. Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software. Second edition. Springer. New York. ISBN 0-387- 45281-8, 2007.

[González et al., 2008] Ignacio González, Sébastien Déjean, Pascal G. P. Martin, and Alain Baccini. CCA: An R Package to Extend Canonical Correlation Analysis. In: Journal of statistical software Vol. 23, Issue 12, Jan 2008

[Marote and Silva, 2002] Eusébio Marote, Silva, Emiliana. Análise dinâmica da eficiência das explorações leiteiras da ilha Terceira. In: Actas do Congresso de Zootecnia, 12ª ed., 2002

[R, 2008] R Development Core Team. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org, 2008

[Silva, et al. 2001] Emiliana Silva, Julio Berbel, and Amílcar Arzubi. Tipología y análisis no paramétrico de eficiencia de explotaciones lecheras en Azores (Portugal) a partir de datos de RICA-A. In: Economía agraria y recursos naturales: Nuevos enfoques y perspectivas – Actas do Congreso de la Asociación Española de Economía Agraria, 4ª ed., Universidad Pública de Navarra, 2001

[Silva, 2004] Emiliana Silva and Fátima Venâncio. A competitividade das explorações pecuárias no Faial: Recurso a metodologias alternativas. In: Actas do Congresso de Economistas Agrícolas, 4ª ed. , 2004

[Simar and Wilson, 2008] Simar, L., and P.W. Wilson. Statistical Inference in Nonparametric Frontier Models: Recent Developments and Perspectives, in: H. Fried, C.A.K. Lovell, S. Schmidt (eds) The Measurement of Productive Efficiency and Productivity Change, New York, Oxford University Press, 421-521, 2008

[Wilson, 1993] Paul W. Wilson. Detecting outliers in deterministic nonparametric frontier models with multiple outputs. In: Journal of Business and Economic Statistics, 11, pp. 319-323, 1993.

## Authors' Information

***Veska Noncheva*** *– Associate Professor, University of Plovdiv, 24 Tzar Assen Str, Plovdiv 4000, Bulgaria; Researcher, CEEAplA, Azores University, Ponta Delgada 9501-801, Portugal; e-mail: wesnon@uni-plovdiv.bg*

***Armando Mendes*** *– Researcher; CEEAplA, Azores University, Ponta Delgada 9501-801, Portugal; e-mail: amendes@uac.pt*

***Emiliana Silva*** *– Researcher; CEEAplA, Azores University, Angra do Heroísmo 9700-851, Portugal; e-mail: emiliana@uac.pt*

# ON COORDINATION OF EXPERTS' ESTIMATIONS OF QUANTITATIVE VARIABLE*

**Gennadiy Lbov, Maxim Gerasimov**

***Abstract**: In this paper, we consider some problems related to forecasting of quantitative feature. We assume that decision rule is constructed on the base of analysis of empirical information represented in the form of statements from several experts. The criterion of a quality of experts' statements is suggested. The method of forming of united expert decision rule is considered.*

***Keywords**: expert statements, coordination.*

***ACM Classification Keywords**: I.2.6. Artificial Intelligence - knowledge acquisition.*

***Conference**: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

## Introduction

In this work we assume that objects under investigation are described by some set of qualitative and quantitative features, and some independent experts give predictions of estimated quantitative feature. Their statements may be partially or completely identical, supplementary, and/or contradictory. Also, experts' statements may vary from time to time as well as new "knowledge" from new experts may be obtained. Hence, decision rule is constructed on the base of analysis of empirical information, represented in the form of several experts' statements. Obtained decision rule must be free from anomalies as conflict and redundancy.

## Setting of a Problem

Let $\Gamma$ be a population of elements or objects under investigation. By assumption, $L$ experts give estimations of values of unknown quantitative feature $Y$ for objects $a \in \Gamma$, being already aware of their description $X(a)$. We assume that $X(a) = (X_1(a),...,X_j(a),...,X_n(a))$, where the set $X$ may simultaneously contain qualitative and quantitative features $X_j$, $j = \overline{1,n}$. Let $D_j$ be the domain of the feature $X_j$, $j = \overline{1,n}$; $D_Y$ be the domain of the quantitative feature $Y$, $D_Y = [\alpha, \beta] \subset R$. In this paper we assume that the feature space $D$ is a subset of the product set $\prod_{j=1}^{n} D_j$.

Note that $D$ may be not equal to $\prod_{j=1}^{n} D_j$.

Example. $D_1 = \{a,b,c,d\}$, $D_2 = [10,20]$, $D = [a,c]\times[10,15] \bigcup [b,d]\times[12,20]$.

We shall say that a set $E$ is a *rectangular set* in $D$ if $E = \prod_{j=1}^{n} E_j$, $E_j \subseteq D_j$, $E_j = [\alpha_j, \beta_j]$ if $X_j$ is a quantitative feature, $E_j$ is a finite subset of feature values if $X_j$ is a nominal feature.

In this paper, we consider statements $S^i$, $i = \overline{1,M}$; represented as sentences of type "if $X(a) \in E^i$, then $Y(a) = y^i$", where $E^i$ is a rectangular set in $D$. By assumption, each statement $S^i$ has its own weight $w^i$

* The work was supported by the RFBR under Grants N07-01-00331a, 08-07-00136a.

($0 < w^i \le 1$ for individual statements). Such a value is like a measure of "confidence". Each statement $S^i$ corresponds to $\left\langle l^i, E^i, y^i, w^i \right\rangle$, where $l^i$ is a code of expert from whom statement is obtained.

Without loss of generality we may assume that experts themselves have equal "weights".

Denote the initial sets of statements obtained from $l$-th expert by $\Omega^l$, the set of initial statements from all experts by $\Omega$, $\Omega = \bigcup_{l=1}^{L} \Omega^l$.

The problem consists in constructing decision rule that reflects information synthesized from an organized group of expert opinions.

## On Criterion of a Quality of Experts' Statements

Let $y_0(x)$ be the value of the feature $Y$ at the point $x \in D$, i.e. $y_0(x) = Y(a)$ if $X(a) = x$. Let $y_l(x)$ be the estimation of the $y_0(x)$ made by $l$-th expert.

We shall say that the set of the values $y_0(x)$ on $D$ is a *strategy of nature* (denote it by $c$), and the set of the values $y_l(x)$ on $D$ is a *strategy of $l$-th expert* (denote it by $g_l$).

In this paper we assume for simplicity that there exists rectangular sets $V^1, \ldots, V^{T_l} \subseteq D$ such that $D = \bigcup_{t=1}^{T_l} V^t$, $V^{t_i} \bigcap V^{t_j} = \varnothing$ if $i \ne j$, $y_l(x) \equiv \beta^t \ \ \forall x \in V^t$, where $\beta^t$ is a constant.

Thus, we assume that the strategies $g_l$ are piecewise constant in $D$.

Consider value $h$ such that $0 \le h \le 1$. We shall say that $l$-th expert (a strategy $g_l$) has a *competence* $h$ if $\dfrac{|y_0(x) - y_l(x)|}{\beta - \alpha} \le 1 - h \ \ \forall x \in D$.

Define the criterion of a quality of strategy $g_l$ as the integral

$$\eta(g_l) = \frac{\int_D (y_0(x) - y_l(x))^2 dx}{(\beta - \alpha)^2 \mu(D)},$$

where $\mu(D)$ is a measure of the set $D$.

Consider strategies $g_1, \ldots, g_m$. Let $A$ be some algorithm of constructing decision rule on the base of these strategies. Denote the resulted strategy by $g^A$, $g^A = A(g_1, \ldots, g_m)$.

We shall say that an algorithm $A$ is a *linear combination* of strategies $g_1, \ldots, g_m$ if $\exists \lambda_1, \ldots, \lambda_m \ge 0$ such that $\sum_{l=1}^{m} \lambda_l = 1$, $y^A(x) = \sum_{l=1}^{m} \lambda_l y_l(x) \ \ \forall x \in D$.

**Proposition 1.** If strategies $g_1, \ldots, g_m$ have a competence $h$, then their linear combination has a competence at least equal to $h$.

**Proof.** Take any point $x \in D$. Then

$$\left| y_0(x) - y^A(x) \right| = \left| y_0(x) - \sum_{l=1}^{m} \lambda_l y_l(x) \right| = \left| \sum_{l=1}^{m} \lambda_l y_0(x) - \sum_{l=1}^{m} \lambda_l y_l(x) \right| \le \sum_{l=1}^{m} \lambda_l \left| y_0(x) - y_l(x) \right| \le 1 - h.$$

■

**Proposition 2.** There exists an algorithm $A$ such that for any strategies $g_1$ and $g_2$ we have

$$\eta(A(g_1, g_2)) \le \frac{\eta(g_1)+\eta(g_2)}{2}.$$

**Proof.** Consider algorithm $A$ such that $y^A(x) = \frac{y_1(x)+y_2(x)}{2} \quad \forall x \in D$.

Since strategies $g_l$ are piecewise constant in $D$, strategy $g^A$ is piecewise constant in $D$.

Take any point $x \in D$. Then

$$\left(y_0(x) - \frac{y_1(x)+y_2(x)}{2}\right)^2 = \frac{1}{4}\left(y_0(x)-y_1(x)+y_0(x)-y_2(x)\right)^2 =$$

$$= \frac{1}{2}\left(y_0(x)-y_1(x)\right)\left(y_0(x)-y_2(x)\right) + \frac{1}{4}\left(y_0(x)-y_1(x)\right)^2 + \frac{1}{4}\left(y_0(x)-y_2(x)\right)^2 \le$$

$$\le \frac{\left(y_0(x)-y_1(x)\right)^2 + \left(y_0(x)-y_2(x)\right)^2}{2}.$$

■

**Proposition 3.** There exists an algorithm $A$ such that for any strategies $g_1, \ldots, g_m$ we have

$$\eta(A(g_1, \ldots, g_m)) \le \frac{\eta(g_1)+\ldots+\eta(g_m)}{m}.$$

**Proof.** Consider algorithm $A$ such that $y^A(x) = \frac{y_1(x)+\ldots+y_m(x)}{m} \quad \forall x \in D$.

Further proof is similar to the proof of Proposition 2.

■

Note that equality in Proposition 3 is obtained if and only if $y_1(x) \equiv \ldots \equiv y_m(x) \quad \forall x \in D$.

Suppose that strategy of nature $c$ is unknown and there are independent experts with the same competence. From propositions 1 and 3 it follows that the decision rule obtained by the considered algorithm $A$ has at least the same competence and the quality better than average experts' quality.

**Proposition 4.** Let $A$ be the linear combination of independent strategies $g_1, \ldots, g_m$; then the minimum of the value $E\eta(g^A) = E\eta(A(g_1, \ldots, g_m))$ is obtained if $\lambda_1 = \ldots = \lambda_m = \frac{1}{m}$.

**Proof.** Consider values $\varepsilon_l = \lambda_l - \frac{1}{m}$. Note that $\sum_{l=1}^{m} \varepsilon_l = 0$.

Using $E\left(\sum_{l=1}^{m} \varepsilon_l y_l\right)^2 \ge 0$, $E\left(\sum_{l=1}^{m} \varepsilon_l y_l\right) = 0$, $E\left(\sum_{i=1}^{m} y_i \sum_{l=1}^{m} \varepsilon_l y_l\right) = 0$, we get

$$E\left(y_0 - \sum_{l=1}^{m} \lambda_l y_l\right)^2 = E\left(y_0 - \frac{1}{m}\sum_{l=1}^{m} y_l - \sum_{l=1}^{m} \varepsilon_l y_l\right)^2 = E\left(y_0 - \frac{1}{m}\sum_{l=1}^{m} y_l\right)^2 -$$

$$-2E\left(\left(y_0-\frac{1}{m}\sum_{l=1}^{m}y_l\right)\left(\sum_{l=1}^{m}\varepsilon_l y_l\right)\right)+E\left(\sum_{l=1}^{m}\varepsilon_l y_l\right)^2\geq E\left(y_0-\frac{1}{m}\sum_{l=1}^{m}y_l\right)^2.$$

■

## A "Default" Algorithm of Constructing of a Consensus of Several Experts

Further on, we assume that strategy of nature $c$ is unknown.

Let for some point $x\in D$ we have statements from several experts. Consider some "reasonable" algorithm of forming a consensus of experts' statements (denote it by $A$).

Firstly, the algorithm $A$ coordinates each $l$-th expert's statements separately. Suppose that $S^1,\ldots,S^m\in\Omega^l$, $y^i(x)$ be the corresponding estimations of $y_0(x)$ made by $l$-th expert, $i=\overline{1,m}$.

Minimizing value $\sum_{i=1}^{m}w^i(y^i(x)-y)^2$, we get equation $\sum_{i=1}^{m}w^i(y^i(x)-y)=0$. Therefore, put

$$y_l(x)=\frac{\sum_{i=1}^{m}w^i y^i(x)}{\sum_{i=1}^{m}w^i};$$

here $y_l(x)$ is the coordinated opinion of $l$-th expert at the point $x\in D$.

Put $w_l=\max_i\left(1-\frac{2\Delta y^i}{\beta-\alpha}\right)w^i$, where $\Delta y^i=\left|y^i(x)-y_l(x)\right|$.

Secondly, the algorithm $A$ coordinates all experts' statements at the point $x\in D$. Suppose that we have statements from $k$ experts, coordinated as above.

Minimizing value $\sum_{l=1}^{k}w_l(y_l(x)-y)^2$, we get equation $\sum_{l=1}^{k}w_l(y_l(x)-y)=0$. Therefore, put

$$y^A(x)=\frac{\sum_{l=1}^{k}w_l y_l(x)}{\sum_{l=1}^{k}w^l};$$

here $y^A(x)$ is the experts' opinions at the point $x\in D$, coordinated by the algorithm $A$.

After coordination by the algorithm $A$ for all $x\in D$ we have sets in the form of $\widetilde{E}^1$ or $\widetilde{E}^1\setminus(\widetilde{E}^2\cup\widetilde{E}^3\cup\ldots)$ with different predictions, where $\widetilde{E}^i$ are rectangular sets in $D$.

Let us notice that resulted decision rule may suffer from redundancy. Since there are $M$ initial statements, we have up to $2^M$ sets in $D$ with different predictions.

Consider algorithms $B$ of forming a consensus of experts' statements under restrictions on amount of resulted statements. The value

$$F(B) = \frac{\int_D \left(y^A(x) - y^B(x)\right)^2 dx}{\mu(D)}$$

estimates a quality of the algorithm $B$. Here $y^A(x)$ and $y^B(x)$ are the estimations of the $y_0(x)$ prescribed to the point $x \in D$ by the algorithms $A$ and $B$, respectively.

In the general case, the best algorithm $B^* = \arg\min_B F(B)$ is unknown. In the work [1], the heuristic algorithm of forming a consensus of experts' statements for the case of interval prediction is suggested. This algorithm uses distances / similarities between multidimensional sets in heterogeneous feature space [2, 3].

## Conclusion

Suggested method of forming of united decision rule (as the method in [1]) can be used for coordination of several experts statements and different decision rules obtained from learning samples and/or time series. Applications of these methods are relevant to many areas, such as medicine, economics and management.

## Acknowledgements

The work was supported by the RFBR under Grants N07-01-00331a, 08-07-00136a.

## Bibliography

[1] G.Lbov, M.Gerasimov. Interval Prediction Based on Experts' Statements. In: Proc. of XIII Int. Conf. "Knowledge-Dialogue-Solution", 2007, Vol. 2, pp. 474-478.

[2] G.S.Lbov, M.K.Gerasimov. Determining of Distance Between Logical Statements in Forecasting Problems [in Russian]. In: Artificial Intelligence, Ukraine. 2004, Vol.2, pp. 105-108.

[3] A.Vikent'ev. Measure of Refutation and Metrics on Statements of Experts (Logical Formulas) in the Models for Some Theory. In: Int. Journal "Information Theories & Applications", 2007, Vol. 14, No.1, pp. 92-95.

## Authors' Information

***Gennadiy Lbov -*** *Institute of Mathematics, SB RAS, Koptyug St., bl.4, Novosibirsk, Russia;*
*e-mail: lbov@math.nsc.ru*

***Maxim Gerasimov -*** *Institute of Mathematics, SB RAS, Koptyug St., bl.4, Novosibirsk, Russia,*
*e-mail: max_post@ngs.ru*

# ИСПОЛЬЗОВАНИЕ FRIS-ФУНКЦИЙ ДЛЯ РЕШЕНИЯ ЗАДАЧИ SDX

**Ирина Борисова, Николай Загоруйко**

***Аннотация**: Рассматривается задача структуризации избыточного набора информации, выявления основных закономерностей, содержащихся в нем с помощью аппарата FRiS-функций. В результате решения этой задачи (задачи SDX) на основе исходного множества объектов строится его сокращенное описание в терминах классов и существенных признаков. Данное описание снабжено системой правил, позволяющих восстанавливать значения всех признаков на основе существенных и находить место новым объектам в системе построенных классов.*

***Ключевые слова**: Распознавание образов, выбор признаков, натуральная классификация, функция конкурентного сходства.*

***ACM Classification Keywords**: I.5.2. Pattern Recognition*

***Conference**: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

## Введение

Формализация человеческой способности к анализу информации дает возможность частично наделять этой способностью искусственные объекты – компьютеры. Даже самые примитивные модели анализа данных, перенесенные на компьютеры, позволяют достигать значительных результатов, так как использование этих моделей позволяет машинам решать задачи, недоступные человеку из-за своей громоздкости и трудоемкости. Это становится особо актуально в последнее время, когда накопление информации в различных прикладных областях идет с огромной скоростью и ее обработка в принципе невозможна без помощи компьютера.

Одним из наиболее важных этапов обработки информации нам представляется ее систематизация и упрощение – представление в виде, доступном для понимания, более подробного исследования и дальнейшего использования. Человеком для этого используются различные приемы, многие из которых формализованы в рамках предметной области, называемой интеллектуальным анализом данных, и относятся к задачам распознавания образов. Вот основные из них :

1. **Сокращение числа рассматриваемых объектов**. Вместо изучения каждого отдельного представителя выборки рассматриваются классы сходных объектов. Похожесть в рамках класса позволяет заменять множества объектов из этого класса неким эталонным (идеальным) образцом, реализациями которого эти объекты являются.
2. **Упрощение описания классов.** Исходное описание класса в виде прямого перечисления всех объектов, попавших в него, заменяется описанием в виде обобщающего правила (логического решающего правила, линейной разделяющей границы и т.п.). Построение описаний уже существующих классов в виде решающих правил той или иной степени сложности, позволяет более четко представить структуру этих классов, их однородность.
3. **Сокращение числа учитываемых и используемых признаков**. Достигаться оно может как за счет исключения слабых, неинформативных, несущественных, случайных, шумящих признаков,

так и за счет выделения такой подсистемы информативных, существенных признаков, по которой можно восстановить все остальные неслучайные признаки с достаточной степенью точности.

В данной статье предпринимается попытка использовать формальные реализации когнитивных способностей человека для построения алгоритма решения одной из достаточно общих задач распознавания образов, когда перед исследователем оказывается набор данных единой природы (из ограниченной предметной области), представленный в виде таблицы «объект-свойство». При этом относительно представленного набора можно предположить лишь одно – он достаточно полно отражает многообразие объектов этой природы (предметной области) и многообразие признаков, их описывающих. Задачей же исследователя является структуризация этого возможно избыточного набора информации, представление его в виде, удобном для дальнейшего анализа и использования человеком.

Задача такого приведения исходной информации к виду, удобному для восприятия человеком, нами формулируется в терминологии задач комбинированного типа как задача типа **SDX** – задача одновременного формирования образов (задача **S**) с решающими правилами для их распознавания (задача **D**) в наиболее информативном подпространстве признаков (задача **X**).

Задачи основных типов такие как задача построения решающих правил, задача группировки объектов (таксономии), задача выбора системы информативных признаков хорошо известны и давно решаются в области распознавания образов. Однако при решении задач комбинированного типа для решения задач основных типов, из которых они состоят, целесообразно использовать единый подход, опирающийся на одну и ту же базовую гипотезу. В качестве единого базиса для решения различных задач распознавания образов мы используем метод оценки близости между объектами, основанный на функции конкурентного сходства (FRiS-функции).

Использование FRiS-функции позволило нам построить внутренне непротиворечивые и эффективные алгоритмы для решения таких задач комбинированного типа, как **DX** (распознавания с одновременным выбором информативной системы признаков), **SD** (таксономии с одновременным построением решающего правила) , **SX** (таксономия с одновременным выбором информативной системы признаков). Их описание содержатся в ранее опубликованных статьях [1, 2]. Теперь же рассмотрим, как функция конкурентного сходства может быть использована при решении задачи **SDX** (таксономии с одновременным построением решающего правила в пространстве наиболее информативных признаков).

## Функция конкурентного сходства

Кратко напомним, что мы называем функцией конкурентного сходства, и какие предпосылки определяют ее эффективность при решении задач анализа данных.

Человек является самой совершенной из ныне существующих распознающих систем. Если мы хотим, чтобы наши алгоритмы хорошо имитировали человеческие способности решать задачи распознавания, то мы должны использовать ту же меру сходства, которую использует человек. Человек считает сходство категорией не абсолютной, а относительной, и оценивает меру сходства в зависимости от конкурентной ситуации. Для ответа на вопрос «На сколько сильно объект *a* похож на объект *b*?», нужно знать ответ на вопрос «По сравнению с чем?»

Для измерения в шкале отношений меры сходства объекта *z* с конкурирующими объектами *a* и *b* предлагается пользоваться следующими соотношениями:

$F_{a/b}=(r_b-r_a)/(r_a+r_b)$ для сходства Z с объектом *a*

и $F_{b/a}=(r_a-r_b)/(r_a+r_b)$ для сходства Z с объектом *b*.

Здесь $r_a$ и $r_b$ – расстояния от $z$ до $a$ и $b$, соответственно. Функцию $F$ мы и называем функцией конкурентного сходства или FRiS-функцией (от слов **F**unction of **Ri**val **S**imilarity). $F_{a/b}$ принимает значение +1, если $z$ и $a$ неразличимы, -1, если $z$ совпадает с $b$, и 0, если объект $z$ равноудален от объектов $a$ и $b$.

## Формулировка и общая схема решения задачи SDX

Как было сказано выше, человек в силу особенностей своего восприятия предпочитает иметь дело не со всеми $m$ объектами, а с небольшим числом $k$ групп (кластеров) этих объектов, описанных небольшим набором информативных (существенных) признаков Y, выбранных из их исходного множества X. Чтобы быть практически полезной, такое сокращенное описание выборки должно содержать систему решающих правил, в соответствии с которыми каждый новый анализируемый объект может быть отнесен к той или иной группе. Помимо решающих правил сокращенное описание выборки должно содержать систему индуктивных правил, устанавливающих связь между подмножеством существенных признаков и всеми остальными признаками, не вошедшими в базис классификации. По таким правилам для каждого объекта, входящего в образ, по значениям его информативных признаков можно восстанавливать значения остальных признаков.

Это подход согласуется с принципами построения естественных классификаций [3], рассматриваемых рядом авторов, как способ объединения объектов в группы «на основании общих, присущих им свойств, определяющих множество других свойств этих объектов, как известных, так и еще не известных ». При этом «количество свойств объекта, поставленных в функциональную связь с его положением в системе, является максимальным»[4]. Возможность предсказывать значения признаков объектов по их месту в классификации мы будем называть предсказательной способностью классификации.

Рассмотрим вариант этой задачи, когда каждая группа объектов определяется своим типичным представителем (столпом). Новый объект относится к той группе, столп которой оказался ближайшим к этому объекту в пространстве информативных (существенных) характеристик. В качестве прогнозируемых значений признаков, не вошедших в число существенных, для этого объекта берется их значение для соответствующего столпа. Для оценки надежности такого рода прогноз мы используем функцию конкурентного сходства, которая измеряет близость между объектом и эталоном с учетом конкурентной ситуации.

В результате для фиксированного набора столпов $S \subseteq A$, где $A$-исходное множество объектов, и некоторого множества информативных признаков $Y \subseteq X$, где $X$ - исходное множество признаков, определим качество, с которым выбранный набор данных $<S,Y>$ описывает исходный набор $<A,X>$ как:

$$Q_F(S,Y) = \sum_{a \in A} F_X(a, s^* \mid s^* = \arg\min_{s \in S} \rho_Y(a,s))$$

где $F_X$ –функция конкурентного сходства в пространстве $X$, $\rho_Y$ – метрика в пространстве $Y$. Задача же состоит в выборе такой пары $<S,Y>$, которая обеспечит максимум функционалу $Q_F$. Чтобы получить достаточно качественное решение этой сложной задачи мы разобьем ее на две более простые и перейдем к рассмотрению задачи двухуровневой оптимизации:

$$Q_F(Y) = \sum_{a \in A} F_X(a, s^* \mid s^* = \arg\min_{s \in S_Y} \rho_Y(a,s)) \to \max_{Y \subseteq X},$$

где $S_Y = \arg\max_{\substack{S \subseteq A, \\ |S| \le m^*}} \sum_{a \in A} F_Y(a, s^* \mid s^* = \arg\min_{s \in S} \rho_Y(a,s))$.

Набор столпов $S_Y$ для фиксированной подсистемы признаков $Y$ отыскивается с помощью алгоритма таксономии FRiS-Tax[2], который опирается на использование функций конкурентного сходства и в процессе работы строит набор столпов, обеспечивающий максимум среднего значения функции конкурентного сходства по выборке.

При переходе к решению задачи таксономии, мы опираемся на допущение, что в пространстве $Y$ существенных (информативных) характеристик классы, обладающие реальными предсказательными свойствами, должны образовывать компактные сгустки, и, как следствие, отыскиваться с помощью некоторой таксономической процедуры. Выбор же самого пространства $Y$ после определения алгоритма для вычисления $Q_F(Y)$ может осуществляться одной из известных процедур направленного поиска (алгоритмом AdDel, GRAD, СПА), либо локального спуска.

Таким образом, сложная задача SDX сводится к серии более простых, решение которых позволяет представлять исходную выборку объектов в виде, наиболее удобном для анализа пользователем, согласованно выделяя группы похожих объектов, решающее правило для отнесения новых объектов к выделенным группам и информативные (существенные) признаки, наиболее полно, описывающие выборку.

## Проверка на реальных данных

Следующие эксперименты проводились, во-первых, для выяснения того, насколько точно восстанавливает информацию о выборке алгоритм FRiS-SDX, реализующий общую схему решения задачи SDX, описанную в предыдущем параграфе. Во-вторых, ставилась задача оценить, насколько отсутствие информации о выборке (отсутствие априорной информации о разбиении объектов на классы, об информативности описывающих признаков) ухудшает качество решения задач распознавания образов. Насколько оправданным в том или ином случае является переход от основных задач распознавания к комбинированным, и насколько он позволяет восстанавливать эту информацию и тем самым менять качество решения задач в зависимости от того, какова доля информативных признаков в выборке.

За основу была взята таблица, содержащая 64 мерные описания различных вариантов написания 10 арабских цифр. Мы предполагаем, что подобное разбиение является естественным, а практически все признаки – в той или иной степени информативными. Примеры объектов выборки приводятся на Рисунке 1.

Рис.1 Примеры объектов выборки, состоящей из различных вариантов написания арабских цифр.

Обучающая выборка $A$, сформированная на основе этой таблицы, содержала 100 объектов, тестовая выборка $B$ – 655 объектов. Кроме того рассматривались «раздутые» варианты тех же выборок. Так в выборках $A'$ и $B'$ помимо исходных 64 признаков содержалось 64 клона этих признаков с наложенным на них Гауссовым шумом, а также 64 чисто шумовых признака, никак не связанных ни с целевым признаком, ни с исходными описывающими признаками. В итоге, каждый объект в этих выборках описывался уже 192 признаками и соотношение числа в той или иной степени информативных признаков к общему числу признаков было 1:2. По аналогии формировались выборки $A''$ и $B''$. Но в них уже было 1024 шумовых

признака (всего 1152 признака), и доля информативных признаков составляла 1:9. На этих выборках решались следующие типы задачи распознавания :

1. **Задача построения решающего правила** (задача **D)**. Эта задача соответствует случаю, когда известно как разбиение объектов обучающей выборки на классы, так и то, что среди описывающих признаков нет заведомо неинформативных, способных ухудшить качество распознавания. Для ее решения на обучающей выборке запускался алгоритм FRiS-Stolp [1], а эффективность построенного решающего правила оценивалась через качество распознавания обучающей - $Q_{st}$, и тестовой выборки - $Q_{ts}$.
2. **Задача таксономии** (задача **S**). Эта задача соответствует случаю, когда относительно выборки известно, что большая часть признаков информативны, однако информация о принадлежности объектов к классам недоступна. Она решалась с помощью алгоритма FRiS-Tax[2]. Качество таксономии оценивалась следующим образом. Каждому полученному таксону присваивалось имя класса, чьих представителей в нем оказывалось большинство, а затем на полученной выборке строилось решающее правило алгоритмом FRiS-Stolp. Чем выше при этом оказывалось качество распознавания по построенному правилу исходной обучающей выборки $Q_{st}$ и качество распознавания тестовой выборки $Q_{ts}$ тем более похожая на исходную естественную классификацию таксономия у нас получалась .
3. **Задача построения решающего правила с одновременным выбором информативных признаков** (задача **DX**). Эта задача возникает тогда, когда нет уверенности, что все признаки, вошедшие в таблицу-объект-свойство являются информативными, более того высока вероятность появления шумящих, неинформативных признаков, искажающих общую картину. Для ее решения использовался упрощенный вариант алгоритма FRiS-GRAD [1], в котором для направленного поиска системы признаков применялся алгоритм AdDel [5], а информативность каждой тестируемой системы признаков оценивалась через качество описания этой системы признаков системой столпов, построенных алгоритмом FRiS-Stolp. Этот алгоритм запускался на обучающей выборки, а затем полученным решающим правилом в пространстве информативных характеристик распознавалась контрольная выборка.
4. **Задача построения таксономии с одновременным выбором информативных признаков** (задача **SX**). В этом случае недоступной считается как информация об информативности признаков, так и информация о разбиении объектов обучающей выборки на классы. Для решения этой задачи мы использовали тот же алгоритм, что и для решения задачи SDX. Единственным отличием являлось то, что система столпов, которые в последствии могли использоваться как решающее правило, в нем не сохранялись. Качество полученной таксономии, как и в случае задачи **S**, оценивалось через надежность распознавания обучающей и контрольной выборки в выбранном подпространстве признаков в системе классов, сформированной в этой таксономии .
5. **Задача таксономии с одновременным построением решающего правила в пространстве наиболее информативных признаков** (задача **SDX**). Как и в предыдущей задаче, здесь отсутствующей считается информация как об информативности, так и о классовой принадлежности. Эта задача решалась алгоритмом FRiS-SDX, реализующим общую схему, описанную в данной статье. В результате его работы строилась некоторая классификация в пространстве информативных с точки зрения предсказательной способности характеристик. Параллельно строилось решающее правило для распознавания новых объектов. Чтобы оценить качество решения данной задачи, как задачи SDX, мы распознавали исходную обучающую и контрольную выборку относительно построенной классификации по построенному решающему правилу в пространстве информативных характеристик.

По сути две последние задачи в данном случае являются эквивалентными, так как используя алгоритм FRiS-Tax для построения таксономии мы автоматически строим решающее правило, разница лишь в методике оценки качества получаемых решений. В задаче **SX** решающее правило строится отдельно, а в задаче **SDX** для распознавания используется система столпов, полученных в процессе таксономии.

Результаты всех экспериментов, для выборок с различным уровнем шумов приводятся в Таблице 1.

Таблица 1.

| Тип задачи | (*A*,*B*) | | (*A'*,*B'*) | | (*A''*,*B''*) | |
|---|---|---|---|---|---|---|
| | $Q_{st}$ | $Q_{ts}$ | $Q_{st}$ | $Q_{ts}$ | $Q_{st}$ | $Q_{ts}$ |
| D | 0.96 | 0.82 | 0.94 | 0.80 | 0.72 | 0.49 |
| DX | 0.87 | 0.66 | 0.87 | 0.66 | 0.81 | 0.65 |
| S | 0.90 | 0.75 | 0.81 | 0.68 | 0.68 | 0.47 |
| SX | 0.85 | 0.68 | 0.83 | 0.69 | 0.54 | 0.36 |
| SDX | 0.85 | 0.68 | 0.8 | 0.69 | 0.54 | 0.37 |

Как и ожидалось, полученные результаты не дают возможности однозначно ответить на вопрос, следует ли от задач основных типов переходить к задачам комбинированным. Так в случае, когда доля информативных характеристик в выборке велика (пары (*A*,*B*) и (*A'*,*B'*)), выбор информативной подсистемы может ухудшить общее качество решения задачи **DX**. Таким образом, как и предполагалось, отказ от предположения об информативности описывающих признаков и ухудшает качество распознавания, в случае когда эта информации достоверна. Однако, с ростом числа шумящих характеристик в выборке такая процедура становится необходимой и оправданной, что подтверждает эксперимент по решению задачи **DX** на выборках *A''* и *B''*.

При построении таксономии наоборот, добавление процедуры выбора информативной системы признаков оказывается оправданной лишь при относительно небольшом уровне шумов в выборке ( выборка *(A',B')*) и значительно ухудшается с их ростом.

Отказ от информации о классовой принадлежности объектов обучающей выборки также негативно сказывается на качестве получаемых решений, однако, в некоторых случаях это негативное влияние сглаживается на контрольной выборке, которая распознается лучше на более компактной системе классов, построенной в процессе таксономии. Именно поэтому результаты распознавания контрольной выборки в задаче **SX** для выборок (*A,B*) и (*A',B'*), оказываются даже лучше результатов решения задачи **DX** для них же.

Таким образом

Стоит отметить, что подобные результаты также объясняются спецификой конкретной задачи, в которой практически все исходные признаки, описывающие выборку, являются информативными и слабо коррелированными между собой, так как распознавание цифр по их частичному написанию представляется проблематичным. Потому их уменьшение автоматически ведет к потере качества.

## Заключение

1. Показана возможность решения задачи комбинированного типа SDX одновременного выбора классификации S, решающего правила D и информативного подмножества X наблюдаемых объектов.

2. Для оценки предсказательной способности классификации при этом используется среднее значение функции $F_X$ сходства объектов обучающей выборки с эталонами своих образов.

3. Экспериментально показано что информация о разбиении объектов на классы, получаемая в процессе решения задачи SDX, а также SX, достаточно хорошо согласуется с имеющейся естественной классификацией этих объектов. При этом удается сократить число признаков в описании классов.

4. Задачи комбинированного типа целесообразно решать в условиях отсутствия информации об обучающей выборке, при подозрении, что в описании содержатся неинформативные признаки, при отсутствии разбиения на классы.

## Благодарности

Данная работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований, Грант № 08-01-00040.

## Библиография

1. N. G. Zagoruiko, I. A. Borisova, V. V. Dyubanov and O. A. Kutnenko. Methods of recognition based on the function of rival similarity. Pattern Recognition and Image Analysis. 2008. Vol. 18, No.1, pp. 1-6.

2. Борисова И.А. Алгоритм таксономии FRiS-Tax. Научный вестник НГТУ, 2007, №3(28), стр. 3-12.

3. Zagoruiko N., Borisova I. Principles of natural classification. Pattern Recognition and Image Analysis 2005 Vol.15, No1, pp.27-29.

4. Л.А. Субботин Классификация. Москва, 2001.

5. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. Новосибирск: Изд. ИМ СО РАН, 1999.

## Информация об авторах

***Ирина Борисова –*** *Институт Математики СО РАН, пр. Коптюга, дом 4, Новосибирск, 630090, Россия; e-mail: biamia@mail.ru*

***Николай Загоруйко*** *- Институт Математики СО РАН, пр. Коптюга, дом 4, Новосибирск, 630090, Россия; e-mail: zag@math.nsc.ru*

# ВЫЯВЛЕНИЕ ФРАКТАЛОПОДОБНЫХ СТРУКТУР В ДНК-ПОСЛЕДОВАТЕЛЬНОСТЯХ

**Владимир Гусев, Любовь Мирошниченко, Надежда Чужанова**

***Аннотация**: Разработан и реализован алгоритм выявления фракталоподобных структур в ДНК-последовательностях. Фрактальность трактуется как самоподобие, основанное на свойстве симметрии или комплементарной симметрии. Локальные фракталы интересны своей способностью аккумулировать множественные палиндромно-шпилечные структуры с потенциально возможными регуляторными функциями. Выявлены реальные случаи проявления фрактальности в различных геномах: от вирусов до человека. Рассмотрена возможность использования фракталоподобных структур в качестве маркеров, различающих близкие классы последовательностей.*

***Keywords**: DNA sequences, fractal-like structures, repeated fragments, palindrome, complementary palindrome.*

***ACM Classification Keywords**: J. Computer Applications – J.3 Life and medical sciences – Biology and genetics; I. Computing Methodologies- I.5 Pattern recognition – I.5.2. – Design methodology –Feature evaluation and selection.*

***Conference**: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

## Введение

Отдельные фрагменты ДНК характеризуются проявлениями самоподобия, основанного на свойстве симметрии или комплементарной симметрии. По ассоциации с работой [Mandelbrot, 1992], будем называть их локальными фракталами (при отсутствии искажений) или фракталоподобными структурами (при их наличии). Такого рода объекты встречаются в участках аномально низкой сложности, содержащих *повторяющиеся симметричные фрагменты* (палиндромы) или *комплементарные палиндромы* [Gusev, 1999]. Значимость последних в регуляции генетических процессов не вызывает сомнений, тогда как роль обычных палиндромов не столь очевидна. Можно, тем не менее, указать на работу [Bacolla, 2004], в которой приведены примеры нестандартных структур, в образовании которых принимают участие близко расположенные симметричные фрагменты.

Реальные примеры фракталоподобных структур обнаружены нами в зоне начала репликации генома бактериофага λ [Гусев, 1989]. Фракталоподобную (в наших терминах) структуру образует сайт связывания trpR-репрессора E.coli, образованный повторением двух комплементарных палиндромов ACTAGT со вставкой биграммы TA между ними [Karlin, 2005]. В цитируемой работе этот пример приводится в связи с обсуждением аномалий в позиционном распределении комплементарного палиндрома CTAG по длине генома. В недавней работе [Загоскин, 2008] по изучению диминуции хроматина в геноме C.kolensis (пресноводный рачок) было выявлено, что три из четырех исследовавшихся межмикросателлитных локуса длиной от 500 до 750п.н. остаются в геноме соматических клеток после прохождения диминуции, тогда как один из них устраняется в процессе диминуции. Именно в этом фрагменте ДНК обнаружен длинный комплементарный палиндром GGTACGTGCACGTACC, который в двух из пяти вхождений повторяется тандемно. Возможно, возникающие при этом фракталоподобные структуры имеют отношение к объяснению механизма диминуции.

Приведённые примеры свидетельствуют об актуальности изучения фракталоподобных структур и их роли в регуляции основных генетических процессов. Поскольку ни один из известных методов (сложностные профили [Гусев, 1999], сканирующие статистики [Karlin, 1989], алгоритмы отыскания тандемных повторов [Crochemore, 1994]) не гарантируют в общем случае выявления *всех фракталоподобных структур* с

учетом возможного их *наложения* и заданными *ограничениями* на длину базового повтора ("периода"), размер гэпов между соседними его вхождениями и кратность повторений, требуется специальный алгоритм, удовлетворяющий этим ограничениям. *Целью работы* является разработка такого *алгоритма*, его апробация на различных текстах и характеризация выявляемых *фракталоподобных структур.*

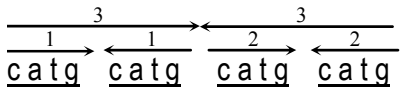
## Основные понятия и предпосылки

Обычный палиндром – это фрагмент, одинаково читаемый в обоих направлениях (например, cattac). Комплементарный палиндром удовлетворяет этому свойству лишь при переименовании элементов алфавита в одном из двух прочтений: a→ t, (a вместо t), t → a, c → g, g → c, что соответствует известному в молекулярной биологии отношению комплементарности. Так, цепочка catatg, является комплементарным палиндромом, поскольку при прочтении ее справа налево с заменой g на c, t на a и т.д. получаем исходную последовательность символов. Элемент самоподобия проявляется в том, что повторение обычного симметричного палиндрома (случай "а") или комплементарного (случай "б") приводит к образованию нового палиндрома (соответственно комплементарного палиндрома) вдвое большей длины:

a) c a t t a c  c a t  t a c

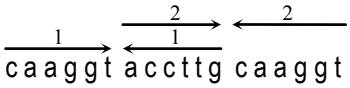
b) c a t  a t g  c a t  a t g

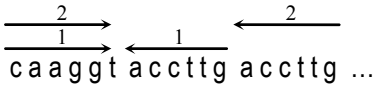
(здесь повторяющиеся фрагменты подчеркнуты; расходящиеся стрелки сверху обозначают палиндром, сходящиеся – комплементарный палиндром). При кратности повторений выше двух возникают множественные структуры (см. 1,2,3 и т.п. в примере с 4-кратным повторением комплементарного палиндрома catg):




Структуры, образуемые точными ("совершенными") тандемными повторами симметричных (в широком смысле) фрагментов, мы называем локальными фракталами. При наличии незначительных искажений внутри повторяющихся фрагментов, равно как и вставок между ними, используем термины "фракталоподобные структуры" или "несовершенные локальные фракталы". Предполагается, что размеры вставок сопоставимы с длинами повторяющихся фрагментов. Функциональная значимость при вставках не теряется: возникают "шпилечные" структуры, участвующие в регуляции многих генетических процессов.

Заметим, что множественные налагающиеся структуры возникают и при повторении левой (случай "а") или правой (случай "б") половинок симметричного в широком смысле фрагмента, даже если для этих половинок не выполняется свойство симметрии:




Однако свойство самоподобия в этом случае теряется.

## Алгоритм обнаружения локальных фракталов

Поиск совершенных структур основан на вывлении всех тандемных повторов в тексте и отбору тех из них, в которых наименьший период (или его циклический сдвиг) является палиндромом или комплементарным палиндромом. Мы используем для выявления тандемных повторов технику сложностного анализа, а именно, ДНК-ориентированный вариант [Gusev, 1999] меры сложности Лемпеля и Зива [Lempel, 1976] реализованный для режима скользящего окна. Размер окна является естественным ограничением на

размеры выделяемых фракталов. Однако его увеличение практически не влияет на трудоемкость, поскольку при сдвиге окна на символ компоненты сложностного разложения не пересчитываются заново, а корректируются, причем далеко не все.

Пусть $\Sigma$ – исходный алфавит; $S$ – текст, составленный из элементов $\Sigma$; $N = |S|$ – длина текста $S$; $S[i]$ – элемент $S$, стоящий в $i$-й позиции ($1 \le i \le N$); $S[i : j]$ – фрагмент $S$, включающий элементы с $i$-го по $j$-й ($1 \le i < j \le N$). Сложностным разложением текста $S$ назовём разбиение его на фрагменты ($S = v_1v_2\ldots v_m$), где $v_1 = S[1]$, а $v_k$ при $k > 1$ определяется следующим образом: если $|v_1v_2\ldots v_{k-1}| = i$, то $v_k$ – наибольший префикс $u$ фрагмента $S[i + 1 : N]$, встречающийся хотя бы дважды в тексте $v_1v_2\ldots v_{k-1}u$. Если такого $u$ не существует, полагаем $v_k = S[i + 1]$. Следует иметь в виду, что $v_k$ и его прототип нерасширяемы вправо, но могут допускать *расширение влево*, трактуемое как удлинение повтора. Число добавляемых слева символов строго меньше длины предыдущего компонента $|v_{k-1}|$. Общее число компонентов в разложении (сложность текста) не меняется, но компоненты лучше согласуются со структурой повторов в тексте. Данное определение сложностного разложения не запрещает наложения (со сдвигом) последнего вхождения $u$ на предшествующее ему, что является индикатором наличия тандемной повторности.

Обозначим через $j(k)$ позицию, с которой начинается предпоследнее вхождение $u$ в $v_1v_2\ldots v_{k-1}u$ (в терминах [Lempel, 1976], $j(k)$ – это указатель ближайшего возможного прототипа для $u$). Если $j(k) + |v_k| \ge i + 1$ (*), т.е. прототип вплотную примыкает к порождаемому компоненту или накладывается на него, то имеет место тандемная повторность с длиной периода $t = i + 1 - j(k)$ и кратностью повторений не меньшей, чем $entier(|v_k| / t) + 1$. Можно показать, что проверка условия (*) значительно упрощается, если работать с *расширенными влево* компонентами разложения. Тогда следует проверять лишь компоненты со значением $j(k) = 1$. Если расположение тандемной структуры синхронизовано с началом окна, для всех значений $j(k) = 1$ выполняется и условие (*). Если тандемов нет, $j(k)$ может равняться 1, но условие (*) не выполняется. Если тандемы встречаются не в начале, а внутри окна, условие (*) может выполняться, но $j(k) \ne 1$ (эти тандемы будут выявлены при последующем движении окна).

Трудоемкость алгоритма в среднем составляет $O(N \log N)$, однако на текстах ограниченной длины он работает быстрее линейного алгоритма, описанного в [Crochemore, 1994]. Такое возможно при существенном различии в значениях мультипликативных констант в обоих алгоритмах (малое значение в нелинейном алгоритме и большое – в линейном).

## Алгоритм обнаружения несовершенных фракталоподобных структур

Рассматриваются структуры, образованные повторяющимися палиндромами (обоих типов) при условии, что расстояние между соседними повторами не превышает заданного порога $r$. Искажение самих палиндромов не допускается. Фрагмент **agagaaga**ct**aga**ttca**aga**tc**aga**, например, при значении $r = 4$ относится к категории интересующих нас фракталоподобных структур с базовым симметричным повтором "aga". Фактически речь идет о выделении кластеров слов определенного типа с числом слов в кластере не меньшим 2 и расстоянием между соседями не большем $r$ Мы вновь используем идеологию *скользящего окна*, размер которого $W$ ограничивает сверху размеры выделяемых структур..

Шаг 1 алгоритма связан с построением $L$-граммного дерева для фрагмента, выделяемого окном. Оно фиксирует полную совокупность $L$-грамм (связных цепочек из $L$ символов), представленных в окне. У $L$-грамм "склеиваются" общие начала ("trie"-структура). Рёбра дерева помечены символами из $L$-граммных цепочек. В узле дерева содержится информация о местах вхождения в текст $L$-граммы, помечающей путь из корня в данный узел. Дерево строится для начального положения окна. Далее, при движении окна по тексту, оно лишь корректируется [Гусев, 2001]. Дерево используется для быстрого поиска палиндромов в пределах окна и проверки условий кластеризуемости. Параметр $L$ примерно соответствует средней длине повторов (6–10 символов).

Шаг 2 алгоритма связан с попыткой обнаружения элементарного (минимального по длине) палиндрома в начале окна, которое рассматривается как центр симметрии. Проще всего воспользоваться "наивным"

алгоритмом, осуществляющим последовательное (до первого несовпадения) сравнение пар символов слева и справа от оси симметрии, которая может проходить через символ (палиндром нечётной длины) или между соседними символами (палиндром или комплементарный палиндром чётной длины). Если палиндром не найден, окно сдвигается на символ, корректируется *L*-граммное дерево (шаг 1) и повторяется шаг 2.

Шаг 3. Если на шаге 2 палиндром найден, с помощью *L*-граммного дерева проверяется наличие других его вхождений в окно анализа. По цепочке символов, образующих палиндром, двигаемся от корня до вершины, содержащей нужную информацию (при длине палиндрома большей *L* продолжение цепочек отслеживается по тексту). Если список вхождений не пуст, проверяются условия кластеризации и принимается решение о наличии (отсутствии) фракталоподобной структуры. Затем окно сдвигается на один символ и осуществляется поиск нового элементарного палиндрома. В случае успеха может быть выявлен новый кластер, наложенный на предыдущий, но с другим повторяющимся ядром (см. фрагмент генома вируса Эпштейна-Барр):

поз. 70490: g g c c g g g g c c g c a g a g g c c g g g g c c

поз. 70491: g c c g g g g c c g c a g a g g c c g g g g c c g

Здесь один и тот же фрагмент текста (сдвиг всего на символ) представлен вверху комбинацией комплементарных палиндромов с ядром ggcc, а внизу - комбинацией обычных палиндромов с ядром gccg.

Трудоемкость алгоритма составляет $O(N \cdot L)$, если считать среднюю (по всем позициям) длину выделяемых палиндромов, не зависящей от *N*, что, как правило, выполняется для ДНК-последовательностей.

## Экспериментальные результаты

Апробация алгоритмов проводилась на вирусных и бактериальных геномах, отдельных генах человека (кодирующие последовательности) и хромосомах генома Arabidopsis thaliana (растение). Результаты можно суммировать следующим образом:

– *невырожденные* по НК-составу фракталоподобные структуры встречаются довольно редко. Характерные длины повторяющихся палиндромов – от 4 до 10, кратность повторения – от 2 до 8 (для совершенных структур – до 5). Исключение составляют участки микросателлитной ДНК с длиной периода 2 или 3 $((ta)^n,(aga)^n$ и т.п., $n \sim 10$ и выше), обладающие специфическими свойствами ввиду вырожденности НК-состава (легкоплавкость и т.п.);

– не симметричные тандемные повторы радикально отличаются от локальных фракталов наличием длинных (десятки символов) периодов (рекордное наблюдавшееся нами значение для несовершенных фракталов – 14, для совершенных – не более 10);

– фракталоподобные структуры со вставками часто возникают в обычных тандемных повторах, содержащих симметричный фрагмент внутри периода (например, $(\overleftarrow{t\,g\,g}\,a\,\overleftarrow{g\,g\,t}\,g\,g\,c\,t\,a)^n$ . Аналогичный пример, но уже с комплементарным палиндромом, демонстрирует фрагмент гена Ecodermal dysplasis 1 у человека (accession number AF060999 в GenBank, поз. 599):

…c a (g g a a t t c c a)(g g g a t t c c t)(***G*** g a a t t c ***C*** a)g g a a…
… (Gly Ile Pro) (Gly Ile Pro) (Gly Ile Pro)…

Здесь комплементарный палиндром (1) длины 8 входит в состав тандемных повторов длины 9 (см. круглые скобки). Повторение (1) "усиливает" конструкцию: возникает структура (2) шпилечного типа. Нам неизвестна её функциональная нагрузка, но интересно отметить, что имевшие место мутации в выделенных заглавными буквами позициях (замена *G* на a и *C* на t), ослабляющие структуру и на нуклеотидном и на аминокислотном уровне, приводили к наследственным заболеваниям;

– наряду с проявлениями фрактальности выделяемые структуры порой демонстрируют другие уникальные свойства. Одна из таких структур обнаружена в четвертой хромосоме генома Arabidopsis th. Её ядро составляет комплементарный палиндром tgtcgatcgaca. Для (a,t)-богатой хромосомы этот идеально сбалансированный по НК-составу фрагмент встречается неожиданно часто – 194 раза. Его инвертированная копия встречается всего 1 раз (аномальная асимметрия!). Все вхождения сосредоточены в диапазоне от 1816917-й до 5664586-й позиции, что при длине хромосомы, равной 18585042нк, следует охарактеризовать как сильную позиционную аномалию. И, наконец, внутри указанного диапазона почти половина всех вхождений являются спаренными, т.е. отстоят друг от друга на 9-10 нк, образуя фракталоподобные структуры с гэпами. Две из них приведены для иллюстрации ниже (поз.1999210 и. 3931735):

**t g t c g a t c g a c a** t c a c c a t g a g **t g t c g a t c g a c a**

**t g t c g a t c g a c a** g a g g t a g t a a **t g t c g a t c g a c a**

Нетрудно видеть, что при строгой консервативности самих палиндромов и расстояний между ними разделяющие их вставки эволюционируют относительно свободно. Большая часть структур связана с ретротранспозонами ("gypsy-like retrotransposon family (Athila)", "non-LTR retrotransposon family (LINE)" и т.п.) Поскольку мобильные элементы рассматриваются как своего рода "батареи" регуляторных элементов, перемещение которых по геному меняет экспрессию генов, можно предполагать, что выделенная структура также имеет отношение к регуляции этого процесса;

– просматривается возможность использования фракталоподобных структур в качестве признаков (маркеров), разделяющих те или иные классы объектов. Рассматривались два класса объектов – геномы вирусов клещевого энцефалита (ВКЭ) и вирусов Повассан (ВП) – представители одного и того же рода флавивирусов. Эти геномы представлены молекулой одноцепочечной РНК длиной около 11 тыс. нуклеотидов, содержащей единственную открытую рамку считывания, в которой последовательно закодированы все структурные и неструктурные белки. Эти (индивидуальные) белки образуются в результате посттрансляционного расщепления единого полипротеина, длина которого у разных штаммов практически не меняется (~ 3414 аминокислотных остатков). Выравнивание кодирующих последовательностей у разных штаммов показывает высокий уровень гомологии (свыше 90%).

Нас интересовали два вопроса: 1) существуют ли на уровне РНК фракталоподобные структуры, разделяющие геномы ВКЭ и ВП? 2) существуют ли фракталоподобные структуры, разделяющие геномы одного класса (ВКЭ) на две группы: инаппарантные штаммы (выделенные от людей с диагнозом «укус клеща», но с бессимптомным течением заболевания) и высоко вирулентные (болезнетворные) штаммы?

Ответ на первый вопрос – положительный. В качестве примера можно указать на совершенный локальный фрактал ccatggccatgg, выявленный в поз. 440 в геномах вируса Повассан: штаммы Nadezdinsk (номер доступа EU670438 в EMBL/Genbank), Spassk-9 (EU770575), Partizansk (EU643649) и LB (LO6436). Другой маркер – tggccatggcca, получающийся циклическим сдвигом из первого, выявлен в поз. 4223. У вируса клещевого энцефалита эти маркеры отсутствуют. Результаты носят предварительный характер ввиду ограниченности исходных данных

Ответ на второй вопрос требует более детального изучения. С одной стороны, можно указать на совершенные фракталы aggaaggaagga и gtggtggtggtg, первый из которых представлен только в инаппарантных штаммах в поз. 5995 (Primorye-212: EU816450, Primorye-253: EU816451 и др.), а второй – в высоковирулентных в поз. 933 (Primorye-94: EU816454, Dalnegorsk: FJ402886). С другой стороны, ввиду близости геномов чаще имеет место ситуация, когда фракталоподобная структура присутствует в варьированной форме как в инаппарантных, так и в высоковирулентных штаммах и требуется детально фиксировать различия между ними. Ниже приведено выравнивание фракталоподобных структур, выявленных в двух высоковирулентных штаммах Primorye-94 (длина структуры *l* = 26, поз. 439, базовый повтор gttg) и Kavalerovo (FJ402885, *l* = 23, поз. 441, базовый повтор tggt). В квадратные скобки заключён более короткий фрагмент, выявленный как фракталоподобная структура в инаппарантных штаммах Primorye-212, Primorye-253 и др. Видно, что зона относительной нестабильности расположена в конце и правее этого фрагмента: здесь имеются разночтения во всех сравниваемых геномах как на нуклеотидном, так и на аминокислотном уровне, прослеживается множественность структур. В штамме Kavalerovo чётко проявлена периодичность $(ctggtt)^3$ – подчёркнута, в штамме Primorye-94 – симметрия (стрелки сверху).

```
                                  V   V   L   L
поз. 439 [g t t g g t t g c t g g t t g t  t g] t c c t g t t g      - Primprye-94
                  | | | | | | | | | | | |  |   | |  | | |  |
поз. 441      t g g t t g c t g g t t c t g g  t t c t g g t        - Kavalerovo
…                 W   L   L   V   L   V   L   V…
```

Наибольший интерес, однако представляет тот факт, что эти фракталоподобные структуры находятся на стыке капсидного белка C и preM – полипептида, что может свидетельствовать об их функциональной или эволюционной значимости*. В известных нам работах по структуре генома ВКЭ роль межпротеиновых интервалов никак не освещена. Добавим также, что на аминокислотном уровне обсуждаемая фракталоподобная структура представлена неслучайным кластером гидрофобных аминокислот, выделяемым с помощью позиционного анализа [Гусев, 2002]

## Заключение

Введено понятие локального фрактала (совершенного и несовершенного) для фрагментов ДНК, характеризующихся проявлениями структурного самоподобия. Локальные фракталы содержат множество палиндромно-шпилечных структур, налагающихся друг на друга. Возможно, они обладают различной функциональной нагрузкой и в этом смысле похожи на многозначные слова в естественном языке. Разработаны и реализованы алгоритмы выявления локальных фракталов в ДНК-последователь-ностях. Проведена их характеризация по результатам обработки ряда геномов. Указаны наиболее существенные отличия от проявлений обычной тандемной повторности. Приведены примеры наиболее интересных фракталоподобных структур.

## Благодарности

Работа выполнена в рамках интеграционного проекта СО РАН № 63

* Наличие высокорегулярных фрагментов относительно небольшой длины на границах между крупными структурными единицами геномов отмечалось нами и другими авторами в связи с анализом алгоритмов сегментации. Такие фрагменты несут на себе следы эврлюционных перестроек и могут являться элементами регуляторных структур. В качестве тестового объекта часто рассматривается геном бактериофага λ с достаточно хорошо изученной модульной структурой (см., например, [Гусев, 1989], [Бородовский, 1990], [Braun, 1998])

## Библиография

[Bacolla, 2004] Bacolla A, Jaworski A, Larson J.E., Jakupciak J.P., Chuzhanova N.A., Abeysinghe S.S., O'Connell C.D., Cooper D.N., Wells R.D. Breakpoints of gross deletions coincide with non-B DNA conformations. In: Proc. Natl. Acad. Sci. USA, 2004, Vol.101, P.14162-14167.

[Crochemore, 1994] Crochemore M., Rytter W. Text Algorithms. In: Text Algorithms, Oxford University Press, New York, Oxford, 1994.

[Gusev, 1999] Gusev V.D., Nemytikova L.A., Chuzhanova N.A. On the complexity measures of genetic sequences. In: Bioinformatics, 1999, Vol.15, No 12, P.994-999.

[Karlin, 1989] Karlin S.Statistical signals in bioinformatics. In: PNAS USA,1989, Vol. 102, No 38, P.13355-13362

[Lempel, 1976] Lempel A., Ziv J. On the complexity of finite sequences. In: IEEE Trans. Inform. Theory, 1976, Vol. IT-22, No.1, P.75-81.

[Mandelbrot, 1992] Mandelbrot B. The Fractal Geometry of Nature. In: The Fractal Geometry of Nature. San Francisco: Freeman, 1992.

[Гусев, 1989] Гусев В.Д., Куличков В.А., Чупахина О.М. Сложностной анализ генетических текстов (на примере фага λ) // Препринт Института математики СО РАН, Новосибирск, 1989.- 49 стр.

[Гусев, 2001] Гусев В.Д., Немытикова Л.А. Учет проявлений повторности, симметрии и изоморфизма в символьных последовательностях // Вычислительные системы, вып. 167. Новосибирск, 2001. С. 11–33

[Гусев, 2002] Гусев В.Д., Немытикова Л.А., Саломатина Н.В. Выявление аномалий в распределении слов или связных цепочек символов по длине текста // Интеллектуальный анализ данных. – Новосибирск, 2002. – вып. 171: Вычислительные системы. С. 51—74.

[Загоскин, 2008] Загоскин М.В., Гришанин А.К., Королёв А.Л., Паленко М.В., Муха Д.В. Характеристика межмикросателлитных последовательностей ДНК до и после диминуции хроматина у Cyclopus kolensis // ДАН, 2008, т. 423, № 4. С. 551-555.

[Бородовский, 1990] Бородовский М.Ю., Певзнер П.А. Зонная структура генома фага лямбда. In: Компьютерный анализ генетических текстов. М., Наука, 1990. С. 62-67.

[Braun, 1998] Jerom V. Braun, Hans-Georg Müller. Statistical methods for DNA sequence segmentation. In: Statistical Science, 1998. Vol. 13, No. 2, P. 142-162.

## Authors' Information

***Владимир Гусев*** *– Старший научный сотрудник, Институт математики им. С.Л. Соболева Сибирского отделения Российской академии наук, Россия, 630090, Новосибирск, пр. ак. Коптюга., 4; e-mail: gusev@math.nsc.ru*

***Любовь Мирошниченко*** *– научный сотрудник, Институт математики им. С.Л. Соболева Сибирского отделения Российской академии наук, Россия, 630090, Новосибирск, пр. ак. Коптюга., 4; e-mail: luba@math.nsc.ru*

***Надежда Чужанова*** *– Reader in Bioinformatics, School of Computing, Engineering and Physical Sciences, University of Central Lancashire, Preston, PR1 2HE Great Britain; e-mail: nchuzhanova@uclan.ac.uk*

# Data Mining and Knowledge Discovery

## STRUCTURING OF RANKED MODELS

**Leon Bobrowski**

***Abstract**: Prognostic procedures can be based on ranked linear models. Ranked regression type models are designed on the basis of feature vectors combined with set of relations defined on selected pairs of these vectors. Feature vectors are composed of numerical results of measurements on particular objects or events. Ranked relations defined on selected pairs of feature vectors represent additional knowledge and can reflect experts' opinion about considered objects. Ranked models have the form of linear transformations of feature vectors on a line which preserve a given set of relations in the best manner possible. Ranked models can be designed through the minimization of a special type of convex and piecewise linear (CPL) criterion functions. Some sets of ranked relations cannot be well represented by one ranked model. Decomposition of global model into a family of local ranked models could improve representation. A procedures of ranked models decomposition is described in this paper.*

***Keywords**: Ranked regression, CPL criterion function, prognostic models, decomposition of ranked models*

***ACM Classification Keywords**: Computing classification systems,*

***Conference**: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

### Introduction

Linear regression models allow to predict the value of dependent variable as the weighted sum of the independent variables [1]. Parameters (weights) of such models can be estimated in a standard way from a set of feature vectors composed of independent variables values and accompanied by values of dependent variable.

Linear ranked models can also be used for the purpose of prognosis [2]. The ranked model is such a linear transformation of feature vector on a line which preserves in the best possible manner a given set of ranked relations defined on pairs of these vectors. Parameters (weights) of models are estimated on the basis of a set of ranked pairs of feature vectors. For this purpose, a special convex and piecewise linear (*CPL*) criterion functions is defined on a given family of ranked pairs of feature vectors. Parameters of the ranked line are found through minimization of a such *CPL* criterion function [3].

Some families of ordering relations between feature vectors can be fully preserved during adequate linear transformation of these vectors on the line. In such case, the ranked line represents all ordering relations between feature vectors. It has been proven that the linear model can reflect all the ranking relations between feature vectors if and only if the sets of positive and negative differences of these vectors, are linearly separable [4]. But there exist such families of order relations which cannot be fully represented by one ranked model. More than one ranked model could be needed for a satisfactory representation of ordering relations between feature vectors. Such problems are discussed in the presented paper.

## Pairs of feature vectors with ranked relations

Let us take into consideration a set C of n-dimensional feature vectors $\mathbf{x}_j[n] = [x_{j1},.....,x_{jn}]^T$:

$$C = \{\mathbf{x}_j[n]\} \quad (j = 1,....., m) \tag{1}$$

Vectors $\mathbf{x}_j[n]$ can be considered as points in the *feature space* $F[n]$ ($\mathbf{x}_j[n] \in F[n]$). The component $x_{ji}$ of the vector $\mathbf{x}_j[n]$ is a numerical result of the $i$-th examination (*feature*) ($i = 1,.....,n$) of a given object or event $O_j$ ($j = 1,.....,m$). The feature vectors $\mathbf{x}_j[n]$ can be of a mixed type, and represent different types of measurements (for example: $x_{ji} \in \{0,1\}$ or ($x_{ji} \in R^1$). The symbol "$\mathbf{x}_j[n] \prec \mathbf{x}_k[n]$" means the ordering relation "*follows*", which is fulfilled for a pair of feature vectors $\{\mathbf{x}_j[n], \mathbf{x}_k[n]\}$ with the indices ($j$, $k$) from the set $J_p$:

$$(\forall (j, k) \in J_p) \ (\mathbf{x}_j[n] \prec \mathbf{x}_k[n]) \Leftrightarrow (\mathbf{x}_k[n] \ \textit{follows} \ \mathbf{x}_j[n]) \tag{2}$$

The relation "$\prec$" between feature vectors $\mathbf{x}_j[n]$ and $\mathbf{x}_k[n]$ ($(j,k) \in J_p$) means that the objects or events $O_j$ and $O_k$ could be in some causal dependence. This relation is determined on the basis of additional knowledge about some (not necessarily all) pairs of objects or events $O_j$ and $O_k$. For example, a medical doctor who compares two patients $O_j$ and $O_k$ with the same disease can declare that the patient $O_j$ is in a more serious condition than the patient $O_k$. A disease model can be designed on such basis and used for the purpose of prognosis. As another example let us consider a *causal sequence* of $k$ events $O_j$:

$$O_{j(1)} \rightarrow O_{j(2)} \rightarrow \ldots\ldots\ldots\ldots \rightarrow O_{j(k)} \tag{3}$$

were the symbol "$O_{j(k)} \rightarrow O_{j(k+1)}$" means that the event $O_{j(k+1)}$ is a consequence of the previous (earlier) event $O_{j(k)}$.

The causal sequence (2) of events $O_j$ results in the below ordering relation among feature vectors $\mathbf{x}_j[n]$:

$$\mathbf{x}_{j(1)}[n] \prec \mathbf{x}_{j(2)}[n] \prec \ldots\ldots\ldots\ldots \prec \mathbf{x}_{j(k)}[n] \tag{4}$$

The ordering relation (4) forms the *sequential pattern* $J_p(\mathbf{x})$ of feature vectors $\mathbf{x}_j[n]$ [2].

Let us consider a linear transformation $y = \mathbf{w}[n]^T\mathbf{x}[n]$ of $n$-dimensional feature vectors $\mathbf{x}_j[n]$ ($\mathbf{x}_j[n] \in R^n$) on the points $y_j$ of the line $R^1$ ($y_j \in R^1$):

$$(\forall j \in \{1,.....,m\}) \quad y_j = \mathbf{w}[n]^T\mathbf{x}_j[n] \tag{5}$$

where $\mathbf{w}[n] = [w_1,.....,w_n]^T$ is the weight vector.

The problem of how to design such a linear transformation $y = w[n]^Tx[n]$ (5) which preserves the relation "$\prec$" for all or almost all pairs of indices (j,k) from some set $J_p$ (2) has been analyzed in the paper [2].

*Definition* 1: Feature vectors $\mathbf{x}_j[n]$ with indices $j$ from the set $J_p$ (2) constitute the *linear ranked pattern* $J_p(\mathbf{x}[n])$ if and only if there exists such $n$-dimensional weight vector $\mathbf{w}_p^*[n]$, that the below implication takes place for all ordering relations (2) defined by the set $J_p$ (2):

$$(\exists \mathbf{w}_p^*[n] \in R^n) \ (\forall (j,k) \in J_p) \ \ \mathbf{x}_j[n] \prec \mathbf{x}_k[n] \Rightarrow \mathbf{w}_p^*[n]^T\mathbf{x}_j[n] < \mathbf{w}_p^*[n]^T\mathbf{x}_k[n] \tag{6}$$

In this case, the ordering relations "$\mathbf{x}_j[n] \prec \mathbf{x}_k[n]$" are fully preserved on the ranked *line* $y = \mathbf{w}_p^*[n]^T\mathbf{x}[n]$.

## Differential sets $R^+$ and $R^-$

The procedure of discovering the ranked linear patterns $J_p(\mathbf{x}[n])$ (6) and the ranked line designing has been based on the concept of the positively and negatively oriented dipoles $\{\mathbf{x}_j[n], \mathbf{x}_{j'}[n]\}$, where $j < j'$ [2], [4].

*Definition* 2: The ranked pair $\{\mathbf{x}_j[n], \mathbf{x}_{j'}[n]\}$ of the feature vectors $\mathbf{x}_j[n]$ and $\mathbf{x}_{j'}[n]$ $((j, j') \in J_p^+$, *where* $j < j')$ constitutes the *positively oriented dipole,* if and only if $\mathbf{x}_j[n] \prec \mathbf{x}_{j'}[n]$.

$$(\forall (j, j') \in J_p^+, \textit{where } j < j') \quad \mathbf{x}_j[n] \prec \mathbf{x}_{j'}[n] \tag{7}$$

*Definition* 3: The ranked pair $\{\mathbf{x}_j[n], \mathbf{x}_{j'}[n]\}$ of the feature vectors $\mathbf{x}_j[n]$ and $\mathbf{x}_{j'}[n]$ $((j, j') \in J_p^-$, *where* $j < j'$ constitutes the *negatively oriented dipole* $((j, j') \in J_p^-)$, if and only if $\mathbf{x}_{j'}[n] \prec \mathbf{x}_j[n]$.

$$(\forall (j, j') \in J_p^-, \textit{where } j < j') \quad \mathbf{x}_{j'}[n] \prec \mathbf{x}_j[n] \tag{8}$$

*Definition* 4: The line $y(\mathbf{w}[n]) = \mathbf{w}[n]^T\mathbf{x}[n]$ (5) is *fully ranked* if and only if

$$\begin{gathered}(\forall (j, j') \in J_p^+, \textit{where } j < j') \quad \mathbf{w}[n]^T\mathbf{x}_j[n] < \mathbf{w}[n]^T\mathbf{x}_{j'}[n], \textit{ and} \\ (\forall (j, j') \in J_p^-, \textit{where } j < j') \quad \mathbf{w}[n]^T\mathbf{x}_{j'}[n] < \mathbf{w}[n]^T\mathbf{x}_j[n]\end{gathered} \tag{9}$$

where $J_p^+ \cup J_p^- = J_p$.

Let us introduce the positive set $R^+$ and the negative set $R^-$ of the differential vectors $\mathbf{r}_{jj'}[n] = \mathbf{x}_{j'}[n] - \mathbf{x}_j[n]$ on the basis of the sets of indices $J_p^+$ (7) and $J_p^-$ (8).

$$\begin{aligned}R^+ &= \{\mathbf{r}_{jj'}[n] = (\mathbf{x}_{j'}[n] - \mathbf{x}_j[n]) : (j,j') \in J_p^+\} \\ R^- &= \{\mathbf{r}_{jj'}[n] = (\mathbf{x}_{j'}[n] - \mathbf{x}_j[n]) : (j,j') \in J_p^-\}\end{aligned} \tag{10}$$

We examine a separation of the sets $R^+$ and $R^-$ (10) by such a hyperplane $H(w[n],\theta)$ in the feature space $F[n]$ that passes through the point 0 ($\theta = 0$), where:

$$H(\mathbf{w}[n], \theta) = \{\mathbf{x}[n] : \mathbf{w}[n]^T\mathbf{x}[n] = \theta\} \tag{11}$$

*Definition* 5: The differential sets $R^+$ and $R^-$ (10) are linearly separable in the feature space $F[n]$ by the hyperplane $H(\mathbf{w}[n], 0)$ with the threshold $\theta$ equal to zero ($\theta = 0$) if and only if the below inequalities hold:

$$\begin{gathered}(\exists \mathbf{w}'[n]) \; (\forall (j, j') \in J_p^+) \quad \mathbf{w}'[n]^T\mathbf{r}_{jj'}[n] > 0, \textit{ and} \\ (\forall (j, j') \in J_p^-) \quad \mathbf{w}'[n]^T\mathbf{r}_{jj'}[n] < 0\end{gathered} \tag{12}$$

The hyperplane $H(\mathbf{w}'[n],0)$ (11) separates the sets $R^+$ and $R^-$ (10) if and only if all the above inequalities (12) with the vector $\mathbf{w}'[n]$ are fulfilled.

*Remark* 1: All the implications (6) are fulfilled on the line $y(\mathbf{w}'[n]) = \mathbf{w}'[n]^T\mathbf{x}[n]$ (5) if and only if the hyperplane $H(\mathbf{w}'[n],0))$ (11) separates (12) the sets $R^+$ and $R^-$ (10).

## Convex and piecewise linear criterion function $\Phi(\mathbf{w}[n])$

Designing the separating hyperplane $H(\mathbf{w}[n], 0)$ (11) could be carried out through the minimisation of the convex and piecewise linear (*CPL*) criterion function $\Phi(\mathbf{w}[n])$ similar to the perceptron criterion function [2]. Let us introduce for this purpose the positive penalty function $\varphi_{j\,j'}{}^+(\mathbf{w}[n])$ and the negative penalty function $\varphi_{j\,j'}{}^-(\mathbf{w}[n])$ :

$$(\forall (j,j') \in J_p^+) \qquad \varphi_{jj'}{}^+(\mathbf{w}[n]) = \begin{cases} 1 - \mathbf{w}[n]^T\mathbf{r}_{jj'}[n] & \textit{if } \mathbf{w}^T[n]\mathbf{r}_{jj'}[n] < 1 \\ 0 & \textit{if } \mathbf{w}[n]^T\mathbf{r}_{jj'}[n] \geq 1 \end{cases} \tag{13}$$

and

$$(\forall (j,j') \in J_p^-) \qquad \varphi_{jj'}{}^-(\mathbf{w}[n]) = \begin{cases} 1 + \mathbf{w}[n]^T\mathbf{r}_{jj'}[n] & \textit{if } \mathbf{w}^T[n]\mathbf{r}_{jj'}[n] > -1 \\ 0 & \textit{if } \mathbf{w}[n]^T\mathbf{r}_{jj'}[n] \leq -1 \end{cases} \tag{14}$$

$$0 \qquad \textit{if } \mathbf{w}[n]^{\mathrm{T}}\mathbf{r}_{jj'}[n] \le -1$$

The criterion function $\Phi(\mathbf{w}[n])$ is the sum of the penalty functions $\varphi_{jj'}^{+}(\mathbf{w}[n])$ and $\varphi_{jj'}^{-}(\mathbf{w}[n])$:

$$\Phi(\mathbf{w}[n]) = \sum_{(j,j')\in J_p^{+}} \gamma_{jj'}\varphi_{jj'}^{+}(\mathbf{w}[n]) + \sum_{(j,j')\in J_p^{-}} \gamma_{jj'}\varphi_{jj'}^{-}(\mathbf{w}[n]) \tag{15}$$

where $\gamma_{jj'}$ ($\gamma_{jj'} > 0$) is a positive parameter (*price*) related to the dipole $\{\mathbf{x}_j[n], \mathbf{x}_{j'}[n]\}$ ($j < j'$).

$\Phi(\mathbf{w}[n])$ (14) is the convex and piecewise linear (*CPL*) criterion function as the sum of such type of penalty functions as $\phi_{jj'}^{+}(\mathrm{w}[n])$ and $\phi_{jj'}^{-}(\mathrm{w}[n])$ The basis exchange algorithms, similarly to linear programming, allow one to find the minimum of such function efficiently, even in the case of large multidimensional data sets $R^{+}$ and $R^{-}$ (9) [3]:

$$\Phi^{*} = \Phi(\mathbf{w}^{*}[n]) = \min_{\mathbf{w}} \Phi(\mathbf{w}[n]) \ge 0 \tag{16}$$

The optimal parameter vector $\mathbf{w}^{*}[n]$ and the minimal value $\Phi^{*}$ of the criterion function $\Phi(\mathbf{w}[n])$ (15) can be applied to solving a variety of data mining tasks. In particular, the ranked line $y = (\mathbf{w}^{*}[n])^{\mathrm{T}}\mathbf{x}[n]$ (5) can be found in this way. The below *Lemma* has been proved [2]:

*Lemma* 1: The minimal value $\Phi(\mathbf{w}^{*}[n])$ (16) of the criterion function $\Phi(\mathbf{w}[n])$ (15) is equal to zero if and only if all the inequalities (9) are fulfilled on the line $y(\mathbf{w}^{*}[n]) = (\mathbf{w}^{*}[n])^{\mathrm{T}}\mathbf{x}[n]$ (5).

By taking into account *Remark* 1, we can prove that the minimal value $\Phi(\mathbf{w}^{*}[n])$ (16) of the nonnegative criterion function $\Phi(\mathbf{w}[n])$ (15) is equal to zero if and only if the differential sets $R^{+}$ and $R^{-}$ (10) are linearly separable (12).

## Linear models based on ranked relations family

Family $F_p$ of ranked relations $''\mathbf{x}_j(k) \lessdot \mathbf{x}_{j'}(k)''$ can be defined by the sets $J_p^{+}$ (7) and $J_p^{-}$ (8) of pairs of indices ($j$, $k$).

$$F_p = \{\mathbf{x}_j[n] \lessdot \mathbf{x}_k[n]: (j, k) \in J_p\}, \textit{ where } J_p = J_p^{+} \cup J_p^{-} \tag{17}$$

*Definition* 6: The family $F_p$ is *transient* if the ranked relations $''\mathbf{x}_j(k) \lessdot \mathbf{x}_{j'}(k)''$ from this family fulfill the following implication:

$$\textit{If } ''\mathbf{x}_j(k) \lessdot \mathbf{x}_k(k)'' \textit{ and } ''\mathbf{x}_k[n] \lessdot \mathbf{x}_l[n]'', \textit{ then } ''\mathbf{x}_j[n] \lessdot \mathbf{x}_l[n]'' \tag{18}$$

*Definition* 7: The family $F_p$ the ranked relations is *complete* for the set C (1) if the ranked relations $''\mathbf{x}_j[n] \lessdot \mathbf{x}_k[n]''$ is defined for each pair $\{\mathbf{x}_j[n],\mathbf{x}_k[n]\}$ of elements of this set.

*Theorem* 1: The complete family $F_p$ (17) of relations $''\mathbf{x}_j[n] \lessdot \mathbf{x}_k[n]''$ defines the linear ranked pattern $J_p(\mathbf{x}[n])$ in the feature space $F[n]$ (*Definition* 1) if and only if this family is transient.

*Proof*: If the family $F_p$ defines the linear ranked pattern $J_p(\mathbf{x}[n])$, then there exists such weight vector $\mathbf{w}_p^{*}[n]$ with the length equal to one ($||\mathbf{w}_p^{*}[n]|| = 1$) that the below implication (6) takes place:

$$(\forall (j,k) \in J_p)\ \mathbf{x}_j[n] \lessdot \mathbf{x}_k[n] \Rightarrow y_j < y_k \tag{19}$$

where $y_j = \mathbf{w}_p^{*}[n]^{\mathrm{T}}\mathbf{x}_j[n]$ (5) is the point on the line $y = \mathbf{w}_p^{*}[n]^{\mathrm{T}}\mathbf{x}[n]$ which is equal to the projection of the feature vector $\mathbf{x}_j[n]$ on this line. The transient relation is fulfilled among all the points $y_j$ on the line. Therefore, the transient relation (18) has to be fulfilled also among feature vectors $\mathbf{x}_j[n]$. On the other hand, if the ranked relations $''\mathbf{x}_j[n] \lessdot \mathbf{x}_k[n]''$ from the transient family $F_p$ are defined for each pair $\{\mathbf{x}_j[n],\mathbf{x}_k[n]\}$ of elements $\mathbf{x}_j[n]$ of the set $C$, then the projection points $y_j$ fulfill the implication (6). □

## Linearly separable learning sets $C_k$

We assume that each learning set $C_k$ is composed of $m_k$ labeled feature vectors $\mathbf{x}_j(k)$ assigned in accordance with additional knowledge to the $k$-th category (class) $\omega_k$ ($k = 1,\ldots,K$):

$$C_k = \{\mathbf{x}_j(k)\} \quad (j \in J_k) \tag{20}$$

where $J_k$ is the set of indices $j$ of the feature vectors $\mathbf{x}_j(k)$ belonging to the class $\omega_k$.

Vectors $\mathbf{x}_j(k)$ can be treated as examples or prototypes for the category $\omega_k$. The learning sets $C_k$ (20) are *separable* in the feature space $F[n]$, if they are disjoined in this space. It means that the following rule is fulfilled:

*if* $k \neq k'$, *then* $C_k \cap C_{k'} = \varnothing$.

*Definition* 8: The learning sets $C_k$ (20) are *linearly separable* in the $n$-dimensional feature space $F[n]$ if each of the sets $C_k$ can be fully separated from the sum of the remaining sets $C_i$ by some hyperplane $H(\mathbf{w}_k,\theta_k)$ (11):

$$\begin{gathered}(\forall k \in \{1,\ldots,K\})\ (\exists\, \mathbf{w}_k,\theta_k)\ (\forall \mathbf{x}_j(k) \in C_k)\quad (\mathbf{w}_k)^T\mathbf{x}_j(k) > \theta_k \\ and\ \ (\forall \mathbf{x}_j(k) \in C_i,\ i \neq k)\ \ (\mathbf{w}_k)^T\mathbf{x}_j(k) < \theta_k\end{gathered} \tag{21}$$

In accordance with the relation (21), all the vectors $\mathbf{x}_j(k)$ belonging to the learning set $C_k$ are situated on the positive side ($(\mathbf{w}_k)^T\mathbf{x}_j(k) > \theta_k$) of the hyperplane $H(\mathbf{w}_k,\theta_k)$ (11) and all feature vectors $\mathbf{x}_j(i)$ from the remaining sets $C_i$ are situated on the negative side ($(\mathbf{w}_k)^T\mathbf{x}_j(k) < \theta_k$) of this hyperplane. The linear separability (21) of the learning sets $C_k$ (20) exists among others in the case of the linearly independent feature vectors $\mathbf{x}_j(k)$ [2].

*Definition* 8: The family $F_{k,k'}$ of ordering relations $''\mathbf{x}_j(k) \prec \mathbf{x}_{j'}(k')''$ ($(j, j') \in J_p$ (2)) among labeled feature vectors (17) from different learning sets $C_k$ and $C_{k'}$ (20) is *consistent* with these sets, if and only if all the pairs $\{\mathbf{x}_j(k),\mathbf{x}_{j'}(k)\}$ are ordered in the same manner. This means that:

$$F_{k,k'} = \{\mathbf{x}_j(k) \prec \mathbf{x}_{j'}(k'):\ \mathbf{x}_j(k) \in C_k \ and\ \mathbf{x}_{j'}(k') \in C_{k'},\ where\ k \neq k'\} \tag{22}$$

Let us remark that the above definition excludes ordering relations $''\mathbf{x}_j(k) \prec \mathbf{x}_{j'}(k)''$ among labeled feature vectors $\mathbf{x}_j(k)$ and $\mathbf{x}_{j'}(k)$ (17) from the same learning sets $C_k$.

*Definition* 8: Two learning sets $C_k$ and $C_{k'}$ are *linearly separable* (18) if there exists such hyperplane $H(\mathbf{w}_k,\theta_k)$ (11) which separates these sets:

$$\begin{gathered}(\exists\, \mathbf{w}_k,\theta_k)\ (\forall \mathbf{x}_j(k) \in C_k)\quad (\mathbf{w}_k)^T\mathbf{x}_j(k) > \theta_k \\ and\ \ (\forall \mathbf{x}_{j'}(k') \in C_{k'})\quad (\mathbf{w}_k)^T\mathbf{x}_{j'}(k') < \theta_k\end{gathered} \tag{21}$$

*Lemma* 2: If the learning sets $C_k$ and $C_{k'}$ (20) are separated (23) by the hyperplane $H(\mathbf{w}[n],\theta)$ (11) in the feature space $F[n]$, then the line $y(\mathbf{w}[n]) = \mathbf{w}[n]^T\mathbf{x}[n]$ is *fully ranked* (9) in respect to an arbitrary consistent family $F_{k,k'}$ (22) of ordering relations $''\mathbf{x}_j(k) \prec \mathbf{x}_{j'}(k)''$ between elements $\mathbf{x}_j(k)$ and $\mathbf{x}_{j'}(k')$ of these sets.

*Lemma* 3: If the line $y(\mathbf{w}[n]) = \mathbf{w}[n]^T\mathbf{x}[n]$ is *fully ranked* (9) in respect to the consistent family $F_{k,k'}$ (22) of ordering relations $''\mathbf{x}_j(k) \prec \mathbf{x}_{j'}(k')''$ (which are constituted by all elements $\mathbf{x}_j(k)$ and $\mathbf{x}_{j'}(k')$ of the learning sets $C_k$ and $C_{k'}$ then these sets are linearly separable (23).

The above Lemmas point out the links between linear ranked models (9) and linear separability (23) of the learning sets $C_k$ and $C_{k'}$ (20).

## Decomposition of linear ranked models

As it results from the *Theorem* 1, the transient property of the complete family $F_p$ (21) of ranked relations $''\mathbf{x}_j[n] \prec \mathbf{x}_k[n]''$ assures that this family can be fully represented (6) on a line (5). The minimal value $\Phi^*$ (16) of the criterion function $\Phi(\mathbf{w}[n])$ (15) is equal to zero in this case.

The minimal value $\Phi^*$ (16) of the criterion function $\Phi(\mathbf{w}[n])$ (15) defined by arbitrary family $F_p$ (17) of ranked relations allows to determine the degree of linearity of this family. The minimal value $\Phi^*$ (16) is greater than zero if the family $F_p$ (21) is not linear (6). It has been proved that the minimal value $\Phi^*$ (16) of the criterion function $\Phi(\mathbf{w}[n])$ (15) is *monotonical* in respect to reducing the relation family $F_p$ (21) [4]. It means that:

$$(F_p \supset F_{p'}) \Rightarrow (\Phi_p^* \geq \Phi_{p'}^*) \tag{24}$$

where $\Phi_p^*$ is the minimal value (16) of the criterion function $\Phi_p(w[n])$ (15) defined by ranked relations from the family $F_p$ (17).

We can infer on the basis of the implication (24) that neglecting sufficient number of ranked relations $''\mathbf{x}_j[n] \prec \mathbf{x}_k[n]''$ in the family $F_p$ (17) allows to reduce to zero the minimal value $\Phi_p^*$ (16) of the criterion function $\Phi_p(w[n])$ (15). The multistage procedure of decomposing a global ranked model based on ranked relations family $F_p$ (21) into a family of local ranked models can be based on the implication (24). During the first stage a possibly large subset $F_1$ ($F_1 \subset F_p$) of ranked relations is discovered. which can be represented in a satisfactory manner on some line (5) Then, the family $F_p$ (17) is reduced to $F_{p'}$ by neglecting relations from the subset $F_1$ ($F_{p'} = F_p - F_1$). The reduced family $F_{p'}$ is then used to enhance the second linear model representing relations from the subset $F_2$. In this way the family $F_p$ (21) can be reduced to zero after finite number stages and global ranked model can be replaced by a family of local ranked models.

Another procedure of decomposing the relations family $F_p$ (21) and a global ranked model can be based on consistent subsets $F_{k,k'}$ (22) of ranked relations (2) between labeled feature vectors $\mathbf{x}_j(k)$ and $\mathbf{x}_{j'}(k')$ from selected learning sets $C_k$ and $C_{k'}$ (20). In accordance with the *Lemma* 2, if the learning sets $C_k$ and $C_{k'}$ are linearly separable, then the subset $F_{k,k'}$ (22) of relations (2) is linear and can be fully represented on the ranked line. Such conditions are shown on the Fig. 1.

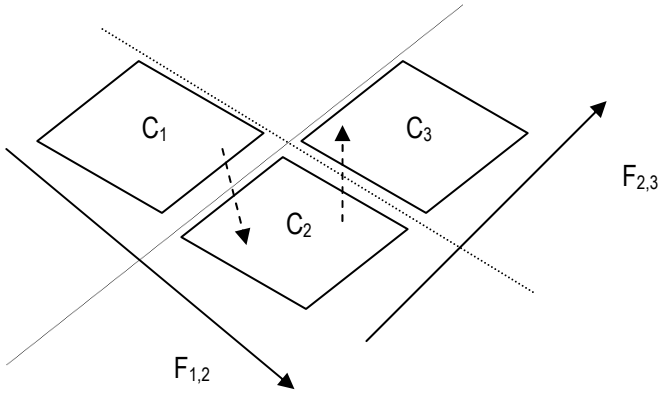



Fig. 1. An example of decomposition of nonlinear family $F_p$ (21) of ranked relations into two linear subsets $F_{1,2}$ and $F_{2,3}$ defined by (25).

Three learning sets $C_1$, $C_2$ and $C_3$ are represented on the above Figure. Each learning set $C_k$ is composed of a large number of two dimensional feature vectors $\mathbf{x}_j(k) = [x_{j1}, x_{j2}]^T$ which can be visualized as points on the plane. We can assume that the vectors $\mathbf{x}_j(k)$ has been generated in accordance with an uniform distribution with a specific *rhombus* shape for each learning set $C_k$.

Let us define the family $F_{k,\,k+1}$ (22) as a set of ranked relations $''\mathbf{x}_j(k) \prec \mathbf{x}_{j'}(k+1)''$ between elements $\mathbf{x}_j(k)$ and $\mathbf{x}_{j'}(k+1)$ of the learning sets $C_k$ and $C_{k+1}$ ($k = 1, 2$):

$$F_{k,\,k+1} = \{\, \mathbf{x}_j(k) \prec \mathbf{x}_{j'}(k+1),\ \textit{where } k = 1 \textit{ or } k = 2 \tag{25}$$

We can remark that the family $F_p$ (25) is not linear, but the subsets $F_{1,2}$ and $F_{2,3}$ (22) of this set $F_p$ are linear. As a result, the global linear model cannot represent all ranked relations from the family $F_p$ (25), but two local models based the subsets $F_{1,2}$ and $F_{,2,3}$ allow to represent all ranked relations.

## Concluding remarks

Linear ranked models can be applied for solving many problems of exploratory data analysis [2]. For example, this approach has been used for designing survival analysis models or in modeling causal sequence of liver diseases.

One of the important problems in ranked modeling is decomposing nonlinear family $F_p$ (17) of ranked relations into linear subsets. The presented paper gives some theoretical insight into these problems where the family has the structure $F_{k,k'}$ (22) based on some learning sets $C_k$ (20).

There are still many unanswered questions concerning decomposition of ranked models. Some of them concern the need for efficient and reliable procedures of local models enhancement when there is no specific assumption about the structure of the relations family $F_p$ (17).

## Acknowledgements

This work was supported by the by the KBN grant 3T11F01130, and partially financed by the grantS/WI/2/2009 from the Białystok University of Technology, and by the grant 16/St/2009 from the Institute of Biocybernetics and Biomedical Engineering PAS.

## Bibliography

1. Duda O. R., Hart P. E., and Stork D. G.: *Pattern Classification,* J. Wiley, New York, 2001
2. L. Bobrowski, "Ranked modelling with feature selection based on the CPL criterion functions", in: *Machine Learning and Data Mining in Pattern Recognition*, Eds. P. Perner et al*., Lecture Notes in Computer Science* vol. 3587, Springer Verlag, Berlin 2005
3. Bobrowski L. and Niemiro W.: "A method of synthesis of linear discriminant function in the case of nonseparabilty". *Pattern Recognition* 17, pp.205-210,1984
4. Bobrowski : "Ranked linear models and sequential patterns recognition", pp. 1-7 in: *Pattern Analysis & Applications*, Volume 12, Issue1 (2009)

## Authors' Information

***Leon Bobrowski*** *– Faculty of Computer Science, Białystok Technical University; Wiejska 45A, 15-351 Białystok, Poland, e-mail:* leon@ibib.waw.pl

# CHAIN SPLIT AND COMPUTATIONS IN PRACTICAL RULE MINING

**Levon Aslanyan, Hasmik Sahakyan**

***Abstract***: *A novel association rule mining algorithm is composed, using the unit cube chain decomposition structures introduced in [HAN, 1966; TON, 1976]. [HAN, 1966] established the chain split theory. [TON, 1976] invented an excellent chain computation framework which brings chain split into the practical domain. We integrate these technologies around the rule mining procedures. Effectiveness is related to the intention of low complexity of rules mined. Complexity of the procedure composed is complementary to the known Apriori algorithm which is defacto standard in rule mining area.*

***Keywords***: *Data mining, unite cube.*

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

***Conference***: *The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

## Introduction

Association rule mining (ARM) is a part of data mining theory. Data Mining is known as a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns or knowledge in data. Existing algorithms are complex computationally, and efficency vs. accuracy issue of algorithms is still open. In association rule mining rules are logical implications of the form $X \to Y$. The mining problem is to generate all implications that have several property estimates greater than the user specified minimum. One of the most used algorithms is *Apriori [KOT, 2006].*

Let we are given a set $I = \{x_1,...,x_n\}$ of $n$ different items. $X \subseteq I$ is itemset and $X$ is $k$-itemset when $|X| = k$. Given a database $D$ with records (transaction, itemset), and we say that $T \in D$ supports $X$, if $X \subseteq T$. We consider the standard concepts of **support** and **confidence**

$$supp(X) = |\{T \in D \mid X \subseteq T\}| / |D|, \text{ and}$$

$$conf(X \to Y) = supp(X \cup Y) / supp(X).$$

As a rule ARM processes the rule mining in 2 tasks; first is to find frequent subsets (that have transaction support above minimum) and second one is to generate association rules themselves. Several ARM construct the frequent subsets by growing. Theoretically, while growing, it constructs not only the maximal elements of hierarchy but may also construct all their subsets. An alternative approach to accelerate the rule mining is considered in this paper, intending to implement the known research results on n-cube geomrtry and algorithmic recognition of Monotone Boolean Functions to the rule mining area.

## Constrained Monotone Boolean Reconstruction

ARM, and its frequent subsets generation (FSG) stage in particular can be described in terms of Monotone Boolean functions. Consider unit cube $B^n$ of dimension $n$ which consists of all binary $n$-vectors. We apply to $n$-cube geometry terms – layer, neighbor vertices, chain, etc. [AS, 1979]. Each cube vertex

$\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)$ can be viewed as transaction, where $\alpha_{i_j} = 1$ indicates that item $i_j$ involved in the transaction, otherwise $\alpha_{i_j} = 0$. Boolean Function, formed in this way, equals 0 if the vertex-subset is frequent, and 1, if not.

Practically all frequent subsets in a typical application problem are placed on very low layers of $B^n$. We may suppose that a value $k$ is known so that all frequent subsets belong to layers lower than $k$. A different question is how precise is the known boundary $k$, but being given $k$, FSG applies for effective solutions of reconstruction of monotone Boolean functions with 0' below the $k$-th layer. Regular ARM starts work from the 0 layer and continues it till some $k$-th layer. The alternative way of solving FSG uses chain split of $B^n$. It is that $B^n$ below the $k$-th layer can be split into the $C_n^k$ disjoint chains providing some special characteristics [HAN, 1966]. Chains are related through the property of conditional complements and as consequence - if values of function are known on $n-2p-1\,(0 \le p \le \lfloor n/2 \rfloor)$ chains then applying monotonity on n-2p+1 chains we receive on them at most 2 new undetermined vertices of function. Second valuable component that enforced our algorithm is that the chain system allows calculations in virtue without archiving and search over the chains [TON, 1976]. Being short we formulate a typical result and then explain the algorithm informally:

**Theorem 1.** Minimal number $\varphi(n)$ of **"example"** type operations required for recognizing arbitrary monotone Boolean functions $f(x_1, x_2, ..., x_n)$ with 0's only below the $k$-th layer, $0 \le k \le \lfloor n/2 \rfloor$, equals

$$C_n^k + C_n^{k-1}.$$

The statement considers a specifically constrained set of Boolean functions achieving in this way more precise and lower estimate for complexity of reconstructing algorithms. Theoretically, the use of this concept requires the set of all $C_n^k$ chains of considered area of $B^n$, which is computationally hard, requiring large memory areas and recursions. Resolving this trade off we engage the [TON, 1976] approach which does not require to keep the chains in memory and calculating, instead, the necessary information having the vertex given by its coordinates.

## Chain Computation Algorithm

This part explains the FSG tasks of ARM by the chain technique. The memory and computational resource reductions as mentioned are the results achieved. If the base algorithm - ***Apriori*** may require $C_n^k + C_n^{k-1} + ... + C_n^0$ steps to restore the Constrained Monotone Boolean Function, then the steps required by alternative algorithm will be not grater than $C_n^k + C_n^{k-1}$.

Large data volumes which appear in data mining applications require low computaional algorithms for composite optimisation problems where data mining is the recurrent task of the total algorithmic solution. ***Apriori*** alternative algorithm by this work uses the specific theoretical know-how which reduces required computations. The system is developed and applied in solving pratical problems – network intrusion detection by LOG records of application software systems is an example of applications.

Now let us stay on description of chain computation framework. Imagine a set of vertical chains connected to each other through the special set of horizontal passes through sets of vertices. These are the chains splitting $B^n$. The procedure working on this set of chains produces a knowledge system which finally becomes the result of algorithm. In our case this will be partial values of function on chains. Vertices in which function is yet unknown might occupy some middle intervals of chains because of monotonity. This structure in its size is smaller than the considered area of $B^n$ with its chain split. We intend to generate the same resulting knowledge by computations

which involve the chain split elements and their coordinates. The rules about chains and passes properties are also simply applied. This work style guarantees the minimal possible memory use of algorithm. Chain computation on considered area of $B^n$ is through the following set of procedures:

(1) computation of the consecutive number of a given vertex on its chain;
(2) computation of consecutive numbers of all neighbor vertices for the given one;
(3) characterization of chain lengths adjoint to the neighbor vertices for the given one;
(4) computation of consecutive number of the next upper vertex to the given one on its chain;
(5) computation of consecutive number of the next vertex below to the given one on its chain;
(6) enumeration of all minimal vertices of all chains of given length;
(7) enumeration of all maximal vertices of all chains of given length;
(8) computation of conditional compliment and its parameters;
(9) computation of all down neighbor vertices to the given one.

It is to mention that the set (1)-(9) is just one example set of chain computation style procedures. These are simple computational tasks. The scenario of FSG we consider is not the unique and several modifications and extensions are possible and useful concerning the application problem conditions. Discuss several characteristic fragments of chain computation rule mining algorithm by procedures (1) - (9).

In our approach it is important that the Boolean function describing itemset frequencies equals 1 above the layer $k$ (the best estimate, given by applied problem). Consider all vertices of layer $k$. For each vertex compute the chain length passing through this vertex and the consecutive number of this vertex on its chain. Working instrument is (1) in this stage and let $R_n$ denotes the chain split of $B^n$. Firstly, procedure computes some values $K_n(\widetilde{\alpha})$ for each vertex $\widetilde{\alpha}$ of $k$-th layer, and then (1) states that vertices $\widetilde{\alpha} \in L, L \in R_n$ are $K_n(\widetilde{\alpha})$-th consecutive vertices on their chains $L$. After this, length of chain $L$ is computed taking into consideration properties of chain split. Described fragment is recursive part of total algorithm.

In a later stage, among the vertices of $k$-th layer we separate all those that are the last vertices of their chains. The chain length of all these vertices equals $n-k-k=n-2k=l$. Ask operator $A_f$ ("example" operator) for the values of considered function on these vertices. After this we apply to the chains of length $l+2$ and extend the results received from $A_f$ to these chains. Determination of all last vertices of the chains of length $l+2$ is by procedure (6), $R(n,l+2,l+3)=\{\widetilde{\alpha} \in B^n \| \|\widetilde{\alpha}\| = (n-l-2/2)$ and $\widetilde{\alpha}$ obeys a property $C\}$, where $C$ is a simple checkable property. Next is to apply $\widetilde{\alpha}_{(+1)}$ (down by the chains) $l+2$ times to each vertex $\widetilde{\alpha} \in R(n,l+2,l+3)$ which constructs the first vertex $\widetilde{\beta}$ of an $l+2$ chain.

On the general step - a chain of length $l+m$ is considered. The first and last vertices of this chain are found and, then the first vertex of a chain of length $l+m-2$ is computed, and the same way the last vertex of this chain, which is the compliment to the pre-final vertex of the chain of length $l+m$. All the values for vertices of chains of length $l+m-2$ are known at this stage, and extension by monotonity to the chains of length $l+m$ and computation on reminder vertices by the operator $A_f$ is to be applied.

## Conclusion

Frequent subset generation is always based on computations on monotone Boolean functions. Monotone function domain is known as complex although the optimal algorithms of recognition are known. Monotone recognition in data mining appears with constrains, which helps to construct less complex tasks and the way to this is through a set of simple computational tasks on the chains mentioned above. The concepts were effectively implemented in intrusion detection analysis by the set of LOG files of applied software systems.

## Acknowledgement

The paper is partially financed by the project **ITHEA XXI** of the Institute of Information Theories and Applications FOI ITHEA and the Consortium FOI Bulgaria. www.ithea.org, www.foibg.com.

## Bibliography

[AS, 1979] L. Aslanyan. Isoperimetry problem and related extremal problems of discrete spaces, Problemy Kibernetiki, 36, pp. 85-126 (1976).

[HAN, 1966] G. Hansel. Sur le nombre des functions booleennes monotones de n variables, C.R. Acad. Sci. Paris, 262, serie A (1966), 1088.

[TON, 1976] G. P. Tonoyan. Chain decomposition of n dimensional unit cube and reconstruction of monotone Boolean functions, JVM&F, v. 19, No. 6 (1976), 1532-1542.

[KOT, 2006] S. Kotsiantis and D. Kanellopoulos. Association Rules Mining: A Recent Overview, GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82.

[AS, 2008] L. Aslanyan and R. Khachatryan. Association rule mining inforced by the chain decomposition of an n-cube, Mathematical Problems of Computer Science, XXX, 2008, ISSN 0131-4645.

## Authors' Information

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

***Hasmik Sahakyan*** *– Leading Researcher, Institute for Informatics and Automation Problems, P.Sevak St. 1, Yerevan 14, Armenia, e-mail: hasmik@ipia.sci.am*

# METHODS OF REGULARITIES SEARCHING BASED ON OPTIMAL PARTITIONING

**Oleg Senko, Anna Kuznetsova**

***Abstract***: *The purpose of discussed optimal valid partitioning (OVP) methods is uncovering of ordinal or continuous explanatory variables effect on outcome variables of different types. 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. Statistical validity associated with revealed regularities is estimated with the help of permutation test repeating search of optimal partition for each permuted dataset. Method for output regularities selection is discussed that is based on validity evaluating with the help of two types of permutation tests.*

***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*

***Conference***: *The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

## Introduction

In present paper the optimal valid partitioning (OVP) approach to data analysis is discussed. The OVP procedures calculate the sets of optimal partitions of one-dimensional admissible intervals of single variables or two-dimensional admissible areas of pairs of variables and evaluate statistical validity of regularities associated with these partitions. It must be noted that applying standard techniques ( F-test, Chi-square and others) for assessing validity by the same datasets which previously has been used for boundaries calculating come across problem of multiple testing (`see [Mazumdar, 2000]). So validity estimates appeared to be too optimistic. One of the ways to calculate adequate estimate is randomized splitting of initial data on two subsets. The first one is used for the boundaries calculating and the second one is used for evaluating of statistical validity. But such approach leads to significant loss of both boundaries exactness and validity levels due to decrease of observations numbers in two datasets. The another way to verify nonrandom character of differences between dependent variable levels in groups of observations formed by partitions is using permutation tests. Discussed below technique that is based on random permutations allows using the same dataset for both purposes: boundaries search and evaluating statistical significance. One more advantage of permutation tests is absence of necessity for any suppositions about variables distribution or any restrictions on groups sizes. Today rather many examples of successful use of permutation technique in different types of tasks [O'Gorman, 2001], [Abdolell, 2002]. Variants of OVP methods using search of optimal partitions inside families of different complexity levels was previously considered by [Senko,1998], [Kuznetsova,2000], [Senko,2003]. Suppose that we study dependence of variable $Y$ on explanatory variables $X_1,\ldots,X_n$ by some empirical dataset $\tilde{S}_0$. Various types of dependent variable are admissible: $Y$ may be continuous variables that are directly observed, vectors of probabilities of several types of events at points in $X$ space, survival curves and so on. The observations from data set $\tilde{S}_0$ must include vectors of independent variables $\mathbf{x}$ and information $\mathbf{y}$ related to dependent variable

$Y$. Existence of some common procedure is supposed for evaluating mean values of *Y* by sets of observations . In case $Y$ is directly observed continuous variable $Y$ is simply value of $Y$ and abovementioned evaluating procedure is reduced to calculating of normal means, evaluating procedure is also reduced to calculating of normal means ( fractions of events types ) when $Y$ is probabilities vector and $Y$ is binary vector indicating type of events, in case $Y$ is survival curve $Y$ is pair including time of last observation and binary indicating if patient is alive. In the last case the Kaplan-Mayer technique is the example of evaluating procedure. The variant of OVP for this type of tasks will be referred to as standard OVP or simply OVP.

But sometimes tasks occur where training set does not contain direct $Y$-descriptions of single objects but includes only mutual distances between $Y$-descriptions. However, OVP methods may be applied in such tasks also with the help of special quality functional.. The variant of OVP using only mutual distances between $Y$-descriptions will be referred to as OVP based on mutual distances or OVPMD.

## 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 $q_1, \ldots, q_r$. The partition $\tilde{Q}$ produces partition of dataset $\tilde{S}_0$ on subsets $\tilde{S}_1, \ldots, \tilde{S}_r$, where $\tilde{S}_j$ ($j = 1, \ldots, r$) 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 $\tilde{S}_j$. Besides integral functional $F_I(\tilde{Q}, \tilde{S}_0)$ local functional $F_L(\tilde{Q}, \tilde{S}_0)$ is possible that is defined as $F_L(\tilde{Q}, \tilde{S}_0) = \max_{j=1,\ldots,r} \{\rho[\hat{y}(\tilde{S}_0), \hat{y}(\tilde{S}_j)] m_j\}$. Unlike integral functional $F_I(\tilde{Q}, \tilde{S}_0)$ local functional $F_L(\tilde{Q}, \tilde{S}_0)$ allows to pick out the most distant from remaining part of $\tilde{S}_0$ subregion of partition. The optimal value of quality

functional in dataset $\tilde{S}$ will be further referred to as $F_I^o(\tilde{S})$ or $F_L^o(\tilde{S})$. In case of OVP-MD The integral quality functional $F_I(\tilde{Q},\tilde{S}_0)$ is defined as the sum:

$$F_I(\tilde{Q},\tilde{S}_0)=\sum_{i=1}^{r}\Big\{\sum_{s_j\in\tilde{S}_i}\sum_{s_{j'}\in\tilde{S}_o\setminus\tilde{S}_i}\rho_y(s_j,s_{j'})-\frac{m_i(m-m_i)}{m(m-1)}\sum_{s_j\in\tilde{S}_i}\sum_{s_{j'}\in\tilde{S}_i}\rho_y(s_j,s_{j'})\Big\},$$

where $m_j$ - is number of observations in subset $\tilde{S}_j$. The local functional $F_L(\tilde{Q},\tilde{S}_0)$ in case of OVP-MD is defined as $F_L(\tilde{Q},\tilde{S}_0)=\max\limits_{i=1,\ldots,r}\{\sum\limits_{s_j\in\tilde{S}_i}\sum\limits_{s_{j'}\in\tilde{S}_o\setminus\tilde{S}_i}\rho_y(s_j,s_{j'})-\frac{m_i(m-m_i)}{m(m-1)}\sum\limits_{s_j\in\tilde{S}_i}\sum\limits_{s_{j'}\in\tilde{S}_i}\rho_y(s_j,s_{j'})\}$

## Regularities validation

For validation of found optimal partitions the permutation test (PT) is used. Advantage of permutation tests is freedom from constraints on probability distribution and size of samples (Senko and Kuznetsova (2006)). The initial variant (PT-1) is based on testing basic null hypothesis that variable $Y$ is fully independent on involved explanatory variables. The optimal value of quality functional $F_*^o$ (it may be $F_I^o$ or $F_L^o$) is used as PT-1 statistics. Let optimal partition of variable $X'$ admissible interval was found inside families I or II or optimal partition of variables $X', X''$ joint admissible area was found inside family III for dataset $\tilde{S}_0=\{(\mathrm{Y}_1,\mathbf{x}_1),\ldots,(\mathrm{Y}_m,\mathbf{x}_m)\}$. Let $F_*^o(\tilde{S}_0)$ is the optimal value of used quality functional. To evaluate statistical validity of discovered regularity set of random permutations $\{\pi_1,\ldots,\pi_N\}$ is calculated with the help of random numbers generator. Initial dataset $\{(\mathrm{Y}_1,\mathbf{x}_1),\ldots,(\mathrm{Y}_m,\mathbf{x}_m)\}$ and permutations $\{\pi_1,\ldots,\pi_N\}$ give rise to permuted datasets $\{\tilde{S}_1^r,\ldots,\tilde{S}_N^r\}$, where $\tilde{S}_j^r=\{(\mathrm{Y}_{\pi_j(1)},\mathbf{x}_1),\ldots,(\mathrm{Y}_{\pi_j(m)},\mathbf{x}_m)\}$. For each dataset $\tilde{S}_{\pi_j}^r$ from $\{\tilde{S}_1^r,\ldots,\tilde{S}_N^r\}$ optimal partition is searched inside the same family for the same variable (variables) and by optimizing the same quality functional that were previously used in case of $\tilde{S}_0$. Let $N_{gt}[F_*^o(\tilde{S}_0)]$ is the number of datasets in $\{\tilde{S}_1^r,\ldots,\tilde{S}_N^r\}$ for which $F_*^o(\tilde{S}_*^r)>F_*^o(\tilde{S}_0)$. The ratio $N_{gt}[F_*^o(\tilde{S}_0)]/N$ is used as estimate of PT-1 p-value for regularity discovered in $\tilde{S}_0$ with the help of optimal partitioning. .

The second variant (PT-2) is based on testing more complicated null hypothesis that variable $Y$ is independent on involved explanatory variables only inside some apriori defined subregions of $X$-space. Let explanatory variables admissible region in $X$-space is partitioned on subregions $q_1^a,\ldots,q_p^a$. This partition produces the partition of dataset $\tilde{S}_0$ on subsets $\tilde{S}_1^a,\ldots,\tilde{S}_p^a$. The following Monte-Carlo procedure of $p$-values estimating was used in second PT variant. Datasets $\{\tilde{S}_1^{ar},\ldots,\tilde{S}_N^{ar}\}$ are generated from $\tilde{S}_0$ with the help of permutations $\{\pi_1^{ar},\ldots,\pi_N^{ar}\}$. As in the first variant only $Y$-components positions are permuted and the order of $X$-components remains fixed. Unlike permutations $\{\pi_1^r,\ldots,\pi_N^r\}$ from the first variant permutations $\{\pi_1^{ar},\ldots,\pi_N^{ar}\}$ do not include transpositions between $Y$-components of observations belonging to different

subsets from $\{\tilde{S}_1^a,\ldots,\tilde{S}_p^a\}$. The procedure of $p$-values calculating by generated datasets $\{\tilde{S}_1^{ar},\ldots,\tilde{S}_{\mathrm{N}}^{ar}\}$ completely coincides with the procedure of $p$-values calculating in the first variant. The $p$-values evaluating the independence of $Y$ inside subregions $q_1^a,\ldots,q_p^a$ and calculated by PT-2 will be referred to as $p_2(q_1^a,\ldots,q_p^a)$-values.

## Forming set of output regularities

The set of output regularities is selected from the set of found optimal partitions using calculated $p$-values. To simplify the discussion we shall not differ further between regularity and describing it optimal partition. The first and simplest way is selecting in output set only regularities with calculated $p$-values less than previously defined threshold $p_{thr}$. The OVP procedures using this way of selecting will be referred to as OVP-CIS (complexity independent selecting). But series of experiments at simulated data [Senko, 2006] demonstrated that OVP-CIS procedure resuts to falling into output set of so called partially false "regularities" with high validity according PT-1. But the cause of this validity actually is dependence of output only on one of variables describing found "regularity". So another variant of OVP procedure (OVP-CDS) will be discussed below. The basic idea underlying this modification of OVP method is selecting to output set only those optimal partitions from more complicated families II , III or IV where variations between induced groups of observations can not be explained from the viewpoint of previously found regularities from simplest family I. In other words selecting of partitions from complicated families in OVP-CDS (complexity dependent selecting) is based on testing if $Y$ is independent on explanatory variable (variables) inside subregions belonging to simple regularities involving these explanatory variable (variables). So OVP-CDS includes different selecting modes for optimal partitions from family I and optimal partitions from more complicated families. Selecting of partitions from family I in OVP-CDS always precede selecting of optimal partitions from families II and III. Then the second variant of permutation test is used to evaluate the validity of the last. Assume that uncovered regularities from family I involving variables $X'$ and $X''$ are contained in the output set. The first from these simple regularities includes subregions $q_1', q_2'$ and second regularity includes subregions $q_1'', q_2''$. Then optimal partition from family II involving variable $X'$ is put to the output set only if $p_2(q_1', q_2')$-values is less than threshold $p_{thr.}$. Optimal partition from families III or IV involving variables $X'$ and $X''$ is placed to the output set only if both inequality $p_2(q_1', q_2') < p_{thr.}$ and $p_2(q_1'', q_2'') < p_{thr.}$ are satisfied. In case output regularities from family I do not involve variables used in optimal partitions from more complicated families II and III the selecting procedure for the last partitions are the same as in OPV-CIS.

## Examples

**Example 1 .**The task of utera mioma relapse predicting from immunological parameters. The group of 6 patients with relapse is compared with 15 patients for which relapse took place before 2 years after operation. Univariate regularity with two boundary point is represented at Fig. 1.

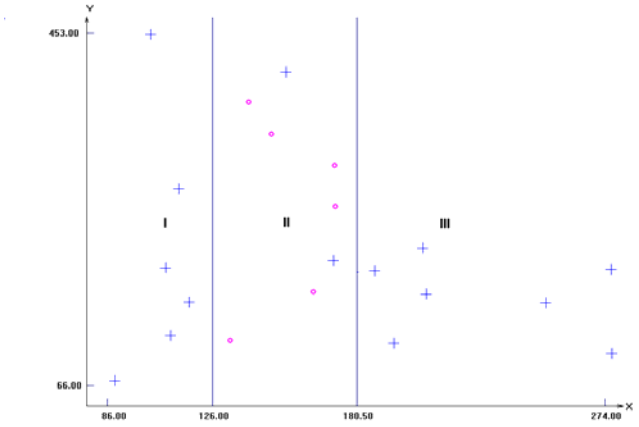



**Fig. 1 –** Optimal 1-dimensional regularity with two boundary points related to dependence of relapse occurrence on variable. Var. 1 correspond to X, var. 2 correspond to Y, .Quadrant I – number of patients without relapse(+) -6, number of patients with relapse (o) – 0;Quadrant I I– without relapse -2, with relapse – 6;Quadrant III – without relapse -7, with relapse 0;It is seen from figure 1 that variable 1 values in patients with relapse are concentrated inside middle interval: 126.0 <var1<180.5.

Figure 1

| | ANOVA | Kolmogorov-Smirnov Test | Mann-Whitney U Test | OVP |
|---|---|---|---|---|
| p-value | 0.672450 | >0.1 | 0.755497 | 0.013 (PF-II,PT-1) |

**Example 2 .** Group of 23 territorial units in Russian Federation with positive migration balance is compared with group of 53 territorial units with negative migration balance. Two-variate regularity with two boundary point related to Task 1.

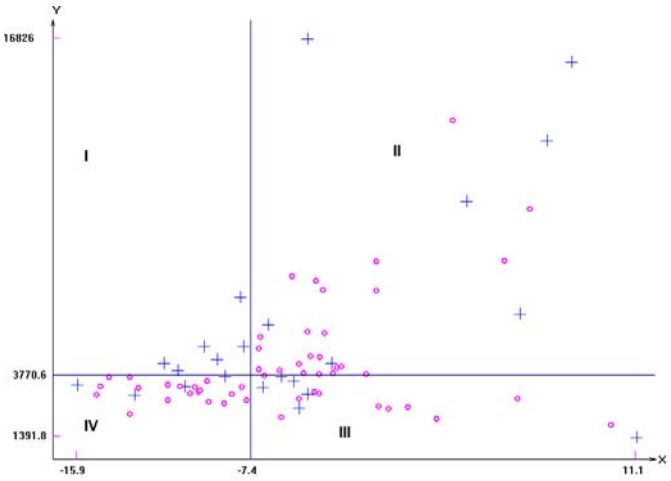



**Fig. 2 –** Optimal 2-dimensional regularity related to dependence of migration balance on variables 8 and 9 Var. 8 correspond to X, var. 9 correspond to Y, .Quadrant I – number of regions with positive balance (+) -6, number of regions with negative balance(o) – 0; Quadrant I I– positive balance -7, negative balance – 24;Quadrant III – positive balance -6, negative balance – 10;Quadrant IV – positive balance -4, negative balance – 19.

Figure 2

It is seen from figure 1 strong dependence of migration balance on variable 3 in case var2<-7.4, but in case var2>-7.4 a distinct dependence of migration balance on variable 3 is not observed. Statistical validity of regularity according PT-1 is p=0.014

Table 1. Validity according standard statistical tests and OVP technique

| | ANOVA | Kolmogorov-Smirnov Test | Mann-Whitney U Test | OVP |
|---|---|---|---|---|
| p-value var 2 | 0.686 | p>0.1 | 0.768 | 0.46 (PF-I, PT-1) |
| p-value var 3 | 0.0398 | P>0.1 | 0.062889 | 0.17(PF-I, PT-1) |
| 2- variate p-value | 0.109 | - | - | 0.014(PF-III, PT-1) |

ANOVA F-test reveals valid (p=0.0398) difference between two groups of regions by variable 3 This difference may be related to group of 4 regions in quadrant II with positive balance and high values of variable 3. All univariate tests did not discover any difference between groups of regions by variable 2. No difference was indicated also by 2-variate ANOVA.

## Conclusion

The new method for uncovering empirical regularities in data was represented. The method allows to find out regularities related to effect of ordinal or continuous explanatory variables on outcome. Method may be used in tasks with different types of dependent variables; binary scalar outcome, scalar or vector continuous variable, survival curve. Besides method may be used when outcome is not described directly but data contains mutual distances between outcome descriptions for different objects. Method is based on validity estimates with the help of permutation tests. These estimates are free from constraints on probability distribution and sample size. Using of permutation test modification (PT-2) allows to select only regularities with statistically founded inclusion of all constituents (features or boundaries).

## Bibliography

[Abdollel, 2002] Abdolell M., LeBlanc M., Stephens D., Harrison R.V. Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. //Statistics in Medicine, 2002, 21:3395-3409.

[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.

[Mazumdar, 2000] Mazumdar, M., Glassman, JR. Tutorial in Biostatistics. Categorizing a prognostic variable: review of methods, coding for easy implementation and applications to decision making about cancer treatment. Statistics in Medicine.2000, 19:113-132.

[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 18th 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., Zabotina 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

## Authors' Information

***Oleg Senko*** *– Leading researcher in Dorodnicyn Computer Center of Russian Academy of Sciences, Russia, 119991, Moscow, Vavilova, 40,* senkoov@mail.ru

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

# ОЦЕНИВАНИЕ РИСКА РЕГРЕССИОННОЙ МОДЕЛИ В СЛУЧАЕ НЕИЗВЕСТНОГО РАСПРЕДЕЛЕНИЯ[1]

**Татьяна Ступина, Виктор Неделько**

***Аннотация**: В данной работе поднимается достаточно актуальная проблема оценивания качества решения в условиях отсутствия информации о распределениях. Для задачи регрессионного анализа рассматривается альтернативная функция риска, построенная ранговым методом. Отражены положительные и отрицательные стороны такого подхода. Статистическим моделированием получены точечные оценки эмпирической функции риска, отражающие обоснованность применения рангового метода в условия «полной неопределённости».*

***Ключевые слова**: функция риска, эмпирическая функция риска, ранговая регрессия, класс линейных решающих функций.*

***ACM Classification Keywords**: G3 Вероятность и Статистика – Корреляционный и Регрессионный анализ.*

***Conference**: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

## Введение

Подход к обработке экспериментальных данных зависит от специфики конкретной области и конечной цели, которая ставится в задаче. В различных областях знаний, целью которых является обнаружение причинно-следственных связей, могут быть использованы одинаковые методы не всегда приводящие к удовлетворительному решению. Чаще всего причина кроется в недостатке априорной информации об изучаемом объекте (явлении) или в некорректной применимости того или иного метода (алгоритма) к обрабатываемым данным. Уточнение же модели, как правило, происходит уже в процессе обработки данных экспертами или в случаях наличия достаточной априорной информации, что не всегда бывает возможным в случае автоматизированной обработки информации и необходимости быстрого принятия решения.

Таким образом, на первом этапе эффективней было бы предложить эксперту модель, полученную наиболее универсальным методом, для её последующего уточнения или вообще принятия решения об её концептуальном изменении. Неотъемлемым этапом в построении модели является её оценка – оценка качества модели. Хорошо известным и широко применяемым способом оценивания качества модели является функция риска [В.Н. Вапник, 1984]. Несмотря на достаточно широкое применение регрессионного анализа во многих прикладных областях знаний задача оценивания риска регрессионной модели и до настоящего времени остаётся актуальной. Это связано с отсутствием универсального метода оценивания качества модели, построенной по выборкам ограниченного объёма в условиях полной неопределённости (отсутствие какой-либо информации о распределениях) [Дж. Себер 1980]. Для задачи распознавания образов предложен подход к эмпирическому оцениванию риска методом численного моделирования, который даёт практически приемлемые оценки [В.М. Неделько, 2008].

[1] Работа выполнена при финансовой поддержке гранта РФФИ 08-01-00944-а

На практике для оценивания риска обычно используют оценки скользящего контроля, как точечные оценки без указания доверительной вероятности. При этом скользящий контроль во многих случаях полагается наилучшим способом оценивания риска, хотя к настоящему времени неизвестны имеющие практически приемлемую точность интервальные оценки риска, основанные на скользящем контроле. В работе [В.М. Неделько, 2008] для задачи распознавания двух образов было показано, что в некоторых случаях на основе эмпирического риска могут быть получены более точные интервальные оценки риска, чем на основе скользящего экзамена. Более того, метод построения эмпирических доверительных интервалов потенциально позволяет использовать не только рассмотренные эмпирические функционалы качества, но и другие характеристики выборки и метода обучения.

В представленной работе получены эмпирические оценки ранговой регрессионной модели из класса линейных функций. Построение решений в данном классе функций не предполагает выполнение классических требований как при восстановлении линейных регрессионных функций. И ещё одним положительным моментом является возможность построения решения в разнотипном пространстве переменных в классе логических решающих функций [Т.А. Ступина, 2006]. Результаты представлены графически и таблично. Проведена сравнительная характеристика эмпирического риска с риском, построенным по контрольной выборке.

## Основные понятия

Пусть $D_X$ – пространство значений переменных, используемых для прогноза, а $D_Y$ – пространство значений прогнозируемых переменных, и пусть *C* – множество всех вероятностных мер на заданной $\sigma$-алгебре подмножеств множества $D = D_X \times D_Y$.

При каждом $c \in C$ имеем вероятностное пространство: $\langle D, B, P_c \rangle$, где *B* – $\sigma$-алгебра, $P_c[D]$ – вероятностная мера (в квадратных скобках мы указываем не аргумент функции, а множество, на котором задана $\sigma$-алгебра). Параметр $c$ будем называть *стратегией природы*. Решающей функцией называется соответствие $f : D_X \to D_Y$ из некоторого класса решающих функций $\Phi$.

Качество принятого решения оценивается заданной функцией потерь $L : Y^2 \to [0, \infty)$. Функция потерь задает цену ошибки как меру несоответствия принятого решения $f(x)$ и истинного значения *y*.

Под риском будем понимать средние потери:

$$R(c, f) = \int_D L(y, f(x)) dP_c[D].$$

Заметим, что значение риска зависит от стратегии природы *c* — распределения, которое в общем случае является неизвестным.

Пусть $v = \{(x^i, y^i) \in D \mid i = \overline{1, N}\}$ — случайная независимая выборка из распределения $P_c[D]$. Эмпирический риск определим как средние потери на выборке:

$$\tilde{R}(v, f) = \frac{1}{N} \sum_{i=1}^{N} L(y^i, f(x^i)).$$

Оценка риска на контрольной выборке определяется как

$$R^*(v^*, f) = \frac{1}{N^*} \sum_{i=1}^{N^*} L(y_i^*, f(x_i^*)),$$

где $v^* = \left\{\left(x_i^*, y_i^*\right) \in D \;\middle|\; i = \overline{1, N^*}\right\}$ – «новая» случайная независимая выборка из распределения $P_c[D]$.

Пусть $Q : \{v\} \to \Phi$ – алгоритм (метод) построения решающих функций, а $f_{Q,v} \in \Phi$ – функция из класса решающих функций $\Phi$, построенная по выборке *v* алгоритмом Q.

Функционал скользящего экзамена определяется как

$$\breve{R}(v, Q) = \frac{1}{N} \sum_{i=1}^{N} L\left(y^i, f_{Q,v_i'}\left(x^i\right)\right),$$

где $v_i' = v \setminus \left\{\left(x^i, y^i\right)\right\}$ – выборка, получаемая из $v$ удалением *i*-го наблюдения.

Задача построения решающей функции (модели) заключается в выборе подходящего алгоритма Q и в оценивании риска принятого решения.

Доверительный интервал для *R* будем задавать в виде $[0, \hat{R}(v)]$.

Здесь мы ограничиваемся односторонними оценками, поскольку на практике для риска важны именно оценки сверху. Таким образом, в данном случае построение доверительного интервала эквивалентно выбору функции $\hat{R}(v)$, которую будем называть оценочной функцией или просто оценкой (риска).

При этом должно выполняться условие:

$$\forall c, \quad P\left(R \le \hat{R}(v)\right) \ge \eta,$$

где $\eta$ – заданная доверительная вероятность.

Известные на данный момент оценки риска строятся не как функции непосредственно выборки, а через композицию $\hat{R}(v) = R_e\left(\overline{R}(v)\right)$, то есть как функции значений некоторого эмпирического функционала $\overline{R}(v)$, в качестве которого обычно выступает эмпирический риск или скользящий экзамен [В.Н. Вапник, 1984].

Эмпирический функционал здесь выступает в роли точечной оценки риска, на основе которой строится интервальная оценка.

## Функция риска построения ранговой регрессии

Пусть $y = f(x)$ — решающая функция, являющаяся некоторой аппроксимацией целевой зависимости, $f \in \Phi$.

Определим риск следующим образом

$$R(c, f) = \max_{A \in \Psi_X} \left| P\left(x \in A, y > f(x)\right) - P\left(x \in A, y < f(x)\right) \right|,$$

где $\Psi_X \subseteq \Lambda_X$ — некоторое подмножество $\Lambda_X$ — $\sigma$-алгебры подмножеств из $D_X$.

Если $\Psi_X = \Lambda_X$, то

$$R(c, f) = \int_{D_X} \left| \beta^+(x) - \beta^-(x) \right| dP(x),$$

где $\beta^+(x) = P\left(y > f(x) \middle| x\right), \; \beta^-(x) = P\left(y < f(x) \middle| x\right)$.

Чтобы риск можно было оценить по выборке, нужно ограничить $\Psi_X$, например, множеством интервалов.

Как вариант, в качестве риска можно использовать расстояние Монжа между $\beta^+(x)$ и $\beta^-(x)$.Так же можно попробовать определить расстояние Монжа без использования дополнительной метрики в $D_X$. Очевидно, что всегда существует $f^*(x)$, для которой риск равен нулю. Это условная медиана, являющаяся оптимальной решающей функцией относительно заданного риска.

Учитывая, что $\beta^+(x) = 1 - \beta^-(x)$ функцию риска представим в следующем виде:

$$R(c, f) = \int_{D_X} |2\beta(x) - 1| dP(x),$$

где $\beta(x)$ - порядок квантили $f(x)$.

Без ограничения общности будем рассматривать $f \in \Phi$ - класс линейных функций. Приоритетной стороной рассматриваемого рангового риска является то, что решения, полученные относительно него являются робастными, т.е. устойчивыми к большим случайным выбросам. Отметим также, что при выполнении классических требований к восстановлению линейных регрессионных зависимостей (ошибки независимы и нормально распределены, регрессоры не случайны) оптимальная решающая функция, представленная условным математическим ожиданием, также является оптимальной относительно рангового критерия.

## Выборочный функционал риска. Алгоритм построения решения

Алгоритмом $Q$ по выборке $\nu$ объёма $N$ строим эмпирическую функцию $f$ из класса линейных функций $\Phi$. Качество построенной функции будем оценивать по эмпирическому риску:

$$\tilde{R}_f = \sum_{i=1}^{M} \sum_{x \in D_X^i} |2\tilde{\beta}(x) - 1| \cdot \tilde{p}(x),$$

где $\tilde{p}(x) = \frac{N_i}{N}$, $N_i = |D_X^i|$, $\tilde{\beta}(x) = \frac{N_i^1}{N_i}$, $N_i^1 = |D_i^1|$, $D_i^1 = \{(x, y) \in V_N \mid y < f(x), x \in D_X^i, y \in D_Y\}$.

Тогда оптимальной решающей функцией в заданном классе относительно рангового критерия будет функция $\tilde{f}(x) = \arg\min_{f \in \Phi} \tilde{R}_f$.

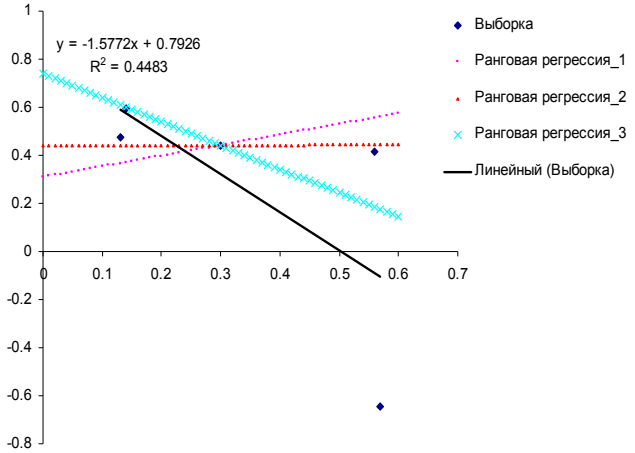



Рис. 1 Линейные регрессии, построенные ранговым методом и по МНК-методу

В целях изучения свойств эмпирического рангового риска будем рассматривать произвольный алгоритм построения линейной зависимости, процедуру и способ разбиения исходного признакового пространства $D_X = \bigcup D_X^i$. Тем самым мы практически охватываем всевозможные способы (алгоритмы) восстановления линейных зависимостей. Оценки эмпирического риска, полученные таким способом, будут являться практически оптимальными. Следовательно, появляется возможность исследования качества решения, построенного некоторым направленным алгоритмом относительно рангового критерия.

Для построения оценки эмпирического риска будем рассматривать оценку риска по контрольной выборке $R^*$ как несмещённой оценки риска [В.Н. Вапник, 1984], представленной в первом параграфе. Риск по контрольной выборке задается аналогично эмпирическому риску, но для элементов контрольной выборки $v^* = \left\{ \left( x_i^*, y_i^* \right) \in D \;\middle|\; i = \overline{1, N^*} \right\}$.

На рисунке 1 мы приведём небольшой показательный пример, демонстрирующий приоритетное свойство линейной регрессии, построенной по ранговому методу в условиях малого объёма выборки, $N = 5$, равномерно распределенной случайной составляющей со среднеквадратическим отклонением равным 0,1 и с 20% выбросами. Истинная линейная функциональная зависимость в примере представляется простым уравнением $f(x) = 0.5$. Несмотря на неоднозначность решения, в примере ранговые регрессионные функции, очевидно, менее отличаются от истинной. По крайней мере, восстановленная по выборке функция достаточно близкая, в метрике $L^2$ или в метрике $C$, к истинной является элементом множества решений, имеющих одинаковые значения эмпирической функции риска. В принципе, при введении дополнительных условий, на основании некоторой априорной информации, можно построить алгоритм, определяющий единственное решение из данного множества. Этот вопрос в данной работе мы пока не рассматриваем.

## Построение эмпирической оценки риска

Под эмпирической оценкой понимается величина, полученная оцениванием минимальной доверительной вероятности по некоторому эвристически выбранному множеству распределений. Если это множество выбрано достаточно «широким», то естественно ожидать, что полученная оценка будет близка к истинной. Возможность доверия таким оценкам может быть аргументирована следующим соображением. Если целенаправленным эвристическим поиском не удалось построить распределения, при котором доверительная вероятность была бы меньше заданной величины, то можно ожидать, что и в реальной задаче распределение окажется таким, что оценка останется справедливой.

| $E\tilde{R}$ | $ER^*$ | $\sigma^2$ |
|---|---|---|
| 0.12 | 0.36 | 0.1 |
| 0.16 | 0.29 | 0.2 |
| 0.17 | 0.25 | 0.3 |
| 0.21 | 0.27 | 0.4 |

Таб. 1 Оценка эмпирического риска в зависимости от уровня шума

В таблице 1 приведены значения точечных оценок эмпирической функции, построенные статистическим моделированием. Результаты подчёркивают достаточно интересный факт. При плохих распределениях оценка «рангового» риска практически равна значению риска, полученного на контроле, как на распределении. Этот результат даёт нам основание применять эмпирическую оценку риска как

достаточно хорошую при построении ранговой регрессии в случае неизвестного распределения. Проведя дополнительное объёмное моделирование по всевозможным распределениям, можно построить эмпирические доверительные интервалы для функции риска, аналогично тому, как в это было сделано для задачи распознавания двух образов [В.М. Неделько 2008].

## Заключение

Несмотря на достаточно хорошо изученные и широко применяемые методы регрессионного анализа, в данной работе поднимается достаточно актуальная проблема оценивания качества решения в условиях отсутствия информации о распределениях. Была рассмотрена и исследована альтернативная функция риска, построенная ранговым методом для задачи восстановления регрессионной зависимости. Отражены положительные и отрицательные стороны такого подхода. Статистическим моделированием получены точечные оценки эмпирической функции риска, отражающие обоснованность применения данного метода в условия «полной неопределённости». Нетривиальной и интересной задачей остаётся создание направленного алгоритма построения эмпирической ранговой регрессии относительно исследуемого риска. Некоторые идеи лежат прямо на поверхности и достаточно скоро будут реализованы авторами работы.

## Благодарности

Работа выполнена при финансовой поддержке гранта РФФИ 08-01-00944-а.

## Библиография

[Дж. Себер 1980] Дж. Себер. Линейный регрессионный анализ. Изд-во М: Мир, 450с.

[В.Н. Вапник, 1984] В.Н. Вапник. Алгоритмы и программы восстановления зависимостей. Изд-во, М: Наука, 805с.

[В.М. Неделько 2008] В.М. Неделько. Об интервальном оценивании риска для решающей функции. Таврический вестник информатики и математики, Изд-во НАН Украины, 2008, с. 97-103.

[T.A. Ступина 2006] T.A. Stupina. Recognition of the Heterogeneous Multivariate Variable. Proceeding of the international conference, 2006 (KDS'2006), Varna (Bulgaria), Vol 1 – pp. 199-202.

## Информация об авторах

***Татьяна Ступина*** *– Институт Нефтегазовой Геологии и Геофизики СО РАН, проспект Коптюга 3, Новосибирск, 630090, Россия, e-mail: stupinata@ipgg.nsc.ru*

***Виктор Неделько*** *– Институт Математики СО РАН, проспект Коптюга 4, Новосибирск, 630090, Россия, e-mail: nedelko@math.nsc.ru*

# МЕТОД ВЫДЕЛЕНИЯ ЗНАЧИМЫХ ДАННЫХ НА ИЗОБРАЖЕНИЯХ ИЗОХРОМНЫХ ЛИНИЙ ДЛЯ СИСТЕМ БЕСКОНТАКТНОГО ИЗМЕРЕНИЯ ВНУТРИГЛАЗНОГО ДАВЛЕНИЯ

**Наталия Белоус, Виктор Борисенко, Виктор Левыкин, Дмитрий Макивский, Анна Зайцева**

***Аннотация***: *Глаукома – это болезнь глаза, причиной которой является повышение внутриглазного давления. Если глазное давление при глаукоме вовремя не снизить до нормы, может погибнуть зрительный нерв, что приведет к необратимой слепоте. На сегодняшний день предложен принципиально новый способ измерения внутриглазного давления, базирующийся на обследовании роговицы глаза человека в поляризованном свете, что позволяет видеть на ней специфическую интерференционную картину. В работе авторами предлагается метод, позволяющий провести распознавание изображения глаза человека, отснятого в поляризованном свете, и выделить на исходном изображении данные, необходимые для разработки системы бесконтактного измерения внутриглазного давления. Проведенный анализ показал, что на сегодняшний день не существует аналогов реализации данного метода. Программная реализация метода позволит разработать программно-аппаратный комплекс, на порядок превосходящий существующие аналоги по стоимости и простоте исполнения, а также бесконтактно, быстро и точно измерять внутриглазное давление.*

***Ключевые слова***: *Внутриглазное давление, глаукома, диагностика, распознавание изображения, обработка изображения, изохрома, изоклина.*

***ACM Classification Keywords***: *I.5 Pattern Recognition, I.5.2 Design Methodology - Feature evaluation and selection.*

***Conference***: *The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

## Введение

Сегодня при неизменных темпах роста науки и техники общество не может обходиться без компьютерной техники. Согласно данным Мировой организации здравоохранения, нагрузка на глаза человека выросла в 100 раз в 2000 году по сравнению с 1900 годом. В будущем эта цифра будет только увеличиваться. Так как почти все время деятельности человека будет связано с ЭВМ, а это означает постоянное напряжение мышц глаз, что в 95 процентах будет приводить к нарушению внутреннего давления глаз. Следовательно, можно констатировать, что болезни, связанные с заболеваниями глаз, становятся критической проблемой современной медицины.

Каждая клетка живого организма имеет определенный тонус, т.е. некоторый уровень внутреннего давления. Являясь следствием биохимических процессов, внутренний тонус обусловливает форму каждого живого элемента и в конечной степени его функцию [http://mv_vizion.ru, 2008].

Внутриглазная жидкость выполняет важные функции по обеспечению глаза питательными веществами и формирует внутриглазное давление. Внутриглазное давление выполняет следующие физиологические функции: расправляет все внутриглазные оболочки, создает в них тургор, придает правильную сферическую форму глазному яблоку, что необходимо для функционирования оптической системы глаза. При нарушении работы механизмов притока и оттока внутриглазной жидкости возникают заболевания глаза, связанные с внутриглазным давлением. Наиболее опасной болезнью, связанной с нарушением оттока внутриглазной жидкости, является глаукома кости, является глаукома [http://www.glaukoma.info/#anathomy,2008].

Глаукома – это болезнь глаза, причиной которой является повышение внутриглазного давления. Если глазное давление при глаукоме вовремя не снизить до нормы, может погибнуть зрительный нерв, что приведет к необратимой слепоте. При глаукоме страдает зрительная функция глаза. В начале человек просто начинает хуже видеть, затем нарушается периферическое зрение, ограничивается зона видимости и в итоге может наступить слепота. Причем изменения эти необратимы, поэтому так важно, вовремя начать лечение глаукомы (рис.1).

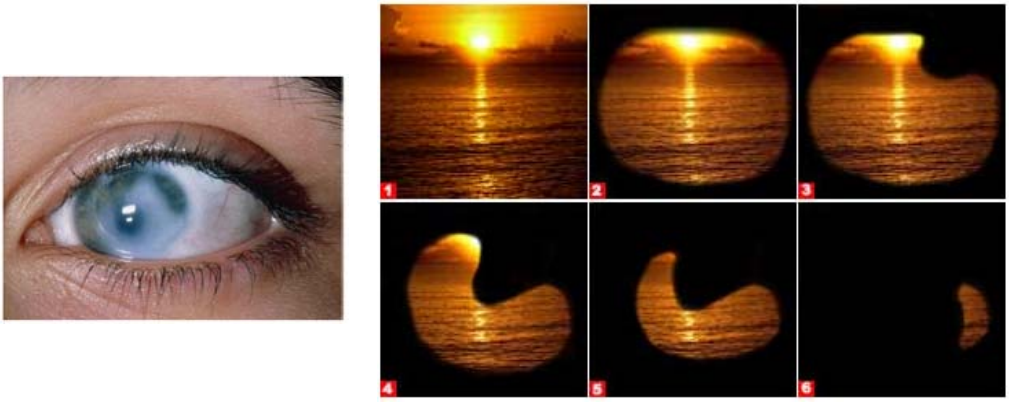



Рисунок 1 – Потеря зоны видимости при глаукоме

На данное время для лечения этой болезни разработано большое количество медикаментозных препаратов, различных физиологических методик снижения внутриглазного давления и комплексов упражнений для глаз, которые помогают полностью остановить прогресс болезни, однако диагностика данного заболевания не имеет оптимального решения. Методы диагностики повышения внутриглазного давления, разработанные и используемые на сегодняшний день, имеют различные недостатки, такие как сложность проведения процедуры, большие затраты времени на проведение процедуры, дороговизна оборудования, неточность измерения, невозможность автоматизации вследствие особенностей процесса измерения и прочее [http://www.glaukoma.info/#anathomy, 2008].

## Новый медицинский подход к измерению внутриглазного давления

В современной медицине разработан и запатентирован [Кочин О.В., 2008] принципиально новый способ измерения внутриглазного давления. Обследование роговицы в поляризованном свете позволяет увидеть на ней специфическую интерференционную картину. Данная картина представляет собой фигуру подобную ромбу. Эта интерференционная картина формируется цветными коллинеарными контурами, называемыми изохромами. Цвет изохромного рисунка зависит от цвета радужки человека, на фоне которой наблюдается интерференционное изображение. Также на рисунке можно выделить элемент в виде мальтийского креста, который положен на вертикальную и горизонтальную диагональ изохромного ромба, он имеет название изоклина. Для диагностики внутреннего давления глаза необходимо измерять параметры изохромной фигуры. Приведенный подход имеет ряд преимуществ, по сравнению с другими подходами, к измерению внутриглазного давления:

- высокая точность измерений;
- высокая скорость процесса получения данных, необходимых для диагностики внутриглазного давления;
- является бесконтактным, а значит и безболезненным для пациента и не требующим применения дополнительных медицинских препаратов;
- применим для людей, измерение внутриглазного давления у которых стандартными методами затруднено, например, у младенцев при диагностике врожденной глаукомы;

- возможность применения ЭВМ для автоматизации процесса;
- более дешевая реализация автоматического оборудования по данному подходу, по сравнению с другими подходами к измерению внутриглазного давления;
- простота использования диагностического оборудования, которое может быть построено на основе данного способа [Кочин О.В., 2008, Кочина М.Л., 2008].

Для автоматизации описанного выше подхода к измерению внутриглазного давления актуально разработать метод для распознавания и выделения диагностических данных на исходных изображениях. Исходными изображениями для данного подхода являются снимки глаза человека, получаемые в поляризованном свете.

Разработка метода и программного обеспечения, предназначенного автоматизировать подход измерения внутриглазного давления по изображениям изохромных линий, производится впервые и не имеет аналогов в мире. Разработка методов распознавания изображений для решения данной проблемы также ранее не производилась.

## Выделение рабочей области на изображении глаза человека

Изображение глаза человека в поляризованном свете несет большое количество помех, преимущественно точечного вида. Контуры элементов изображения, вследствие особенностей съемки, слегка размыты. Информация, которую необходимо снять с изображения для измерения внутриглазного давления, заключена не в частных значениях яркости изображения, а в положении контура изохромы первого порядка. Таким образом, к исходным изображениям целесообразно применить медианный фильтр для очистки изображения от помех. Данный прием сгладит экстремумы яркости, возникшие в результате помех при съемке, и не приведет к потере информативности изображения.

Поскольку на изображении глаза в поляризованном свете может присутствовать до трех изохром, приступать к поиску столь специфического элемента не представляется возможным каким-либо из разработанных, на текущий момент, методом или алгоритмом. Поэтому в первую очередь необходимо ограничить зону поиска. В данной предметной области изохрома первого порядка располагается в пределах радужки глаза человека, а границами выступают контур зрачка и контур роговицы. Поэтому для измерения внутриглазного давления в первую очередь необходимо выделить зону радужки глаза на изображении.

В первую очередь необходимо выделить на изображении контур зрачка, поскольку данный элемент является наиболее легкоузнаваемым и крупным элементом изображения, а область зрачка характеризуется наименьшей яркостью на изображении. Выделив границу зрачка, найдем минимальную границу кольцеобразной области радужки глаза человека, в пределах которой необходимо производить поиск изохромы первого порядка.

Для нахождения зрачка учтем следующие особенности изображения глаза отснятого фотокамерой в монохромном цвете:

- наиболее темная область на изображении глаза – область зрачка;
- зрачок имеет эллипсоидную форму на изображении, приближенную к кругу. Отклонение от круглой формы обусловлено тем, что кривизна роговицы больше кривизны глазного яблока, а это приводит к небольшим погрешностям при съемке, если направление взгляда человека в момент съемки не направлено точно в объектив камеры;
- необходимо учитывать, что ресницы, попавшие на изображение при съемке, приведут к появлению помех, данные помехи по контрастному тону сопоставимы с изображением зрачка.

Для решения задачи поиска зрачка в первую очередь необходимо очистить изображение от помех, созданных ресницами и веками глаз. Ресницы располагаются по контуру изображения глаза. Для

устранения данных помех авторами разработана радиальная фильтрация. Идея радиальной фильтрации заключается в повышении яркости пикселя на относительную величину расстояния от центра изображения до этого пикселя. Преобразование будем проводить в полярной системе координат. Данную операцию можно представить по формулам (1-4):

$$l = \sqrt{\left(\frac{N}{2} - n\right)^2 + \left(\frac{M}{2} - m\right)^2} - r_0 \text{ , n = 1..N, m = 1..M,} \quad (1)$$

$$\alpha = \begin{cases} 0, npu \quad l \le 0 \\ l, npu \quad l > 0, \end{cases} \quad (2)$$

$$\Delta f_{m,n} = f^0_{m,n} + \alpha, \quad (3)$$

$$f^1_{m,n} = \begin{cases} \Delta f_{m,n}, npu \quad 0 < \Delta f_{m,n} \le \max f \\ \max f, npu \quad \Delta f_{m,n} > \max f \end{cases}, \quad (4)$$

где $l$ – расстояние от изображения до пикселя;

n,m – текущие координаты пикселя на изображения;

N – ширина изображения;

M – высота изображения.

$\alpha$ – величина корректировки значения яркости пикселя;

$f^0_{m,n}$ – первоначальное значение яркости пикселя;

$\Delta f_{m,n}$ – значение яркости пикселя после преобразования до введения ограничивающего порога $\max f$;

$\max f$ - максимальное значение яркости для заданной системы;

$f^1_{m,n}$ – выходное значение яркости пикселя, полученное после преобразования;

После проведенной радиальной фильтрации все наиболее темные точки на изображении будут располагаться внутри области зрачка. Для проведения дальнейшего процесса поиска изохромы первого порядка на изображении выбирается несколько точек с минимальной яркостью. Для каждой из этих точек строится прямоугольный треугольник, так, чтобы точка минимальной яркости принадлежала катету треугольника (рис.2а). Гипотенуза треугольника будет являться диаметром круга, в который вписан треугольник, а, следовательно, диаметром зрачка. Для каждой точки минимальной яркости построим четыре прямоугольных треугольника, проведя через точку две перпендикулярные друг другу хорды. Поскольку границы зрачка на исследуемом изображении слегка размыты, то поиск производится по нескольким точкам. Полученные результаты, по координатам центров, разобьем на кластеры методом К-средних, задав радиус кластера размером в одну условную единицу (пиксель). По результатам кластеризации выбирается центр зрачка как центр самого многочисленного кластера. Радиус зрачка $r_{\min}$ определяется как среднее значение радиуса по всем объектам кластера. Таким образом, выделили минимальную границу кольцеобразной зоны радужки, в пределах которой проводится поиск изохромы.

Для поиска максимальной границы кольцеобразной области радужки глаза человека, в пределах которой необходимо проводить поиск изохромы первого порядка, возвратимся к исходному изображению и проведем обработку изображения градиентным методом. Авторами был модифицирован градиентный метод, путем задания направления градации от центра зрачка к границам изображения. Данное направление градирования обусловлено радиальной формой основных элементов изображения роговицы. Результаты градирования приведены на рисунке 2б.

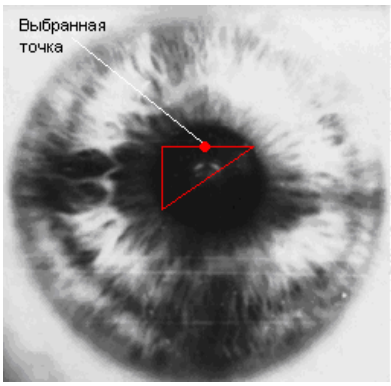



Рисунок 2а) исходное изображение с точкой минимума и треугольником, вписанным в контур зрачка

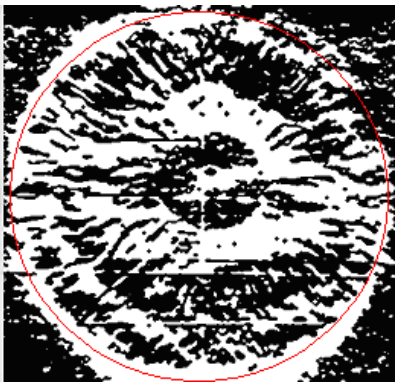


Рисунок 2б) изображение глаза после проведения операции градирования, и нахождения контура с максимальной энергии

Искомый элемент представляет собой кругообразный контур на градированном изображении с центром в точке приближенной к точке центра зрачка и максимальным радиусом, среди подобных кругообразных структур. Поиск производится по формулам (5-6).

$$Q_{x,y,r} = \sum f_{x,y,\Delta r}, \quad \Delta r = r_{\min} \cdots \frac{M}{2}, \tag{5}$$

$$Q_{\max} = \max Q_{x,y,\Delta r}, \tag{6}$$

где $f_{x,y,r}$ – значение яркости пикселя, лежащего на конуре круга с центром в координатах x,y и радиусом r;

$Q_{x,y,r}$ – энергия контура круга с центром в координатах x,y и радиусом r;

Среди всех контуров с энергией $Q_{x,y,r}$, найденных по формуле (6), выбирается контур с максимальной энергией $\max Q_{x,y,r}$ Данный контур является контуром роговицы. Радиус контура с энергией $\max Q_{x,y,r}$ будет являться радиусом ($r_{\max}$) круга, ограничивающего максимальную границу кольцеобразной области радужки глаза человека.

Таким образом, были определены центр и радиусы контуров зрачка и роговицы глаза человека (рис. 3). Эти данные позволяют существенно ограничить область поиска и приступить к распознаванию изоклинной и изохромной линий глаза человека.

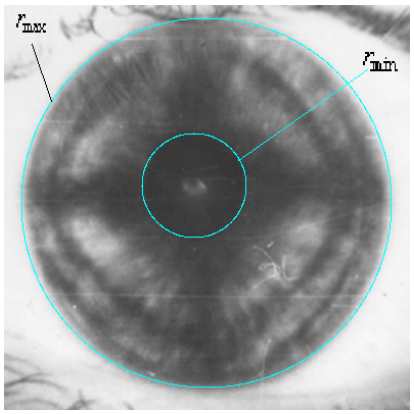



Рисунок 3 - изображение с выделенной областью, в пределах которой производится поиск изохромы первого порядка

## Выделение положения изоклины и контура изохромы

Изохрома первого порядка на изображении глаз человека имеет ромбообразную форму, а изоклина представляет собой пересечение диагоналей ромба изоклины. Выделив на изображении крайние точки, в пределах радужки, ветвей изоклины получим точки, углов изохромного ромба. Для поиска координат ветвей изоклины построим диаграмму распространения области яркости, по уровню энергии сопоставимой с яркостью зоны окружающей зрачок. Диаграмма яркости строится по формулам (7-10).

$$P^{1}_{x+1,y} = P(x+1,\Delta y), \quad при \quad f(x+1,\Delta y) - f_T(x,y) > \theta, \quad \Delta y = y..n, \tag{7}$$

$$P^{0}_{x+1,y} = P(x+1,\Delta y), \quad при \quad f(x+1,\Delta y) - f_T(x,y) > \theta, \quad \Delta y = 1..y, \tag{8}$$

$$P^{T+1}_{x+1,y} = (P^{1}_{x+1,y} - P^{0}_{x+1,y})/2, \tag{9}$$

где $P^{1}_{x+1,y}$ – первая точка повышения яркости на заданную величину $\theta$ при $\Delta y$ изменяющемся от $y$ до $n$;

$P^{0}_{x+1,y}$ – первая точка повышения яркости на заданную величину $\theta$ при $\Delta y$ изменяющемся от 1 до $y$;

x,y – координаты точки P;

$P(x+1,\Delta y)$ - точка перехода яркости;

$f(x+1,\Delta y)$ - величина яркости точки $P(x+1,\Delta y)$;

$f_T(x,y)$ - величина яркости точки $P^{T}_{x,y}$;

$P^{T}_{x,y}$ - точка серединного значения диаграммы, на предыдущем шаге;

$P^{T+1}_{x+1,y}$ - точка серединного значения диаграммы, на текущем шаге.

Таким образом, продвигаясь от зрачка к контуру роговицы по направлению распространения ветви изоклины, будем строить на каждом этапе окно с размерами 1 x n, и выбирать точку n/2. Размер окна n определяется пределами распространения области яркости изоклины на векторе перпендикулярном направлению распространения ветви изоклины. Начальной выбирается точка, лежащая на контуре зрачка.

На заключительном этапе распознавания предлагается установить положение изохромы, по найденным ранее точкам углов изохромного ромба и провести подстройку точек, т.н. методом «активного контура» [Сойфер В.А., 2003]. Для применения метода активного контура необходимо задать приблизительное положение точек изохромы и направление поиска. Зная начальное положение стороны ромба изохромы, зададим направление работы для метода активный контур от границы зрачка к границе роговицы глаза как показано на рисунке 4. Для повышения точности и скорости выделения изохромы зададим яркость искомой изохромы, как величину среднеквадратичного отклонения яркости для всех точек, найденных в процессе выделения изоклины.

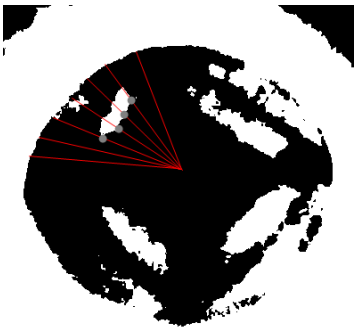


Рисунок 4 – Препарирование изображения и поиск точек изохромы

Найденные таким образом точки будут находиться на границе контура изоклины первого порядка. Поученные точки, а также точки окончания ветвей изоклины, будут представлять собой изохрому первого порядка глаза человека, снятого в поляризованной свете.

## Заключение

Результаты исследования данной работы, а именно разработанный метод, были использованы при разработке экспериментального образца программного обеспечения, позволяющего проводить измерение внутриглазного давления. Исходным материалом для разработанного программного обеспечения являются изображения глаза человека в поляризованном свете. При помощи экспериментального

образца программного обеспечения было проведено тестирование разработанного метода выделения значимых данных на изображении глаза человека в поляризованном свете. Тестирование экспериментального образца программного обеспечения на снимках глаза человека показало высокую эффективность разработанного метода. Данный метод позволяет точно определить положение точек контура изохромы первого порядка на 96.6% исходных изображений. Для оставшихся 3.4% диагностических случаев понадобилось провести повторную съемку глаза человека. По результатам тестирования экспериментального образца программного обеспечения можно сделать выводы, что разработанный метод позволяет быстро и эффективно решать задачу поиска изохромных линий на изображении глаза человека, освещенного поляризованным светом и в будущем построить систему бесконтактной диагностики внутриглазного давления.

## Библиография

[http://mv_vizion.ru, 2008] Офтальмология. Внутриглазное давление [Электронный ресурс] / Информация о болезни глаз. – Режим доступа: www/ URL: http://mv_vizion.ru/bolezni_glaz_vnutriglaznoe.htm/ - 10.09.2008 г. - Загл. с экрана.

[http://www.glaukoma.info/#anathomy,2008] Глаукома. Анатомия и физиология путей оттока внутриглазной жидкости [Электронный ресурс] / Информация для пациентов о глаукоме. – Режим доступа: www/ URL: http://www.glaukoma.info/#anathomy – 15.09.2008г. - Загл. с экрана.

[Кочина М.Л., 2008] Кочина М.Л. Бесконтактные методы диагностики патологии глаза с использованием излучения оптического диапазона //18-th InternationalCrimean Conference “Microwave & Telecommunication Technology”, September 8-12,2008, p. 58-59..

[Кочин О.В., 2008] Патент на корисну модель «Спосіб виміру внутрішньо очного тиску», Кочина М.Л., Кочин О.В., 33640 Україна, МПК (2006), A 61B3/16, A 61B8/10. Заявл.23.10.2007.Опубл.10.07.2008.

[Сойфер В.А., 2003] Методы компьютерной обработки изображения. [Текст]: учеб./ Гашников В.М., Глумов Н.И., Попов С.Б., Чернов В.М., Сойфер В.А., - М.: ФИЗМАТЛИТ, 2003. – 784 с.

## Информация об авторах

***Белоус Наталия***– *заведующий лабораторией «Информационные технологии в системах обучения и машинного зрения», к.т.н., профессор каф. ПО ЭВМ Харьковского национального университета радиоэлектроники. Харьков. Украина. e-mail: belous@kture.kharkov.ua*

***Левыкин Виктор –*** *д.т.н., профессор Харьковского национального университета радиоэлектроники, зав. каф. ИУС. Харьков. Украина. e-mail: levykin@kture.kharkov.ua*

***Борисенко Виктор Петрович –*** *к.т.н. доцент каф. ИУС. Харьковского национального университета радиоэлектроники. Харьков. Украина.*

***Макивский Дмитрий*** *– магистр кафедры ПО ЭВМ Харьковского национального университета радиоэлектроники. Харьков. Украина. e-mail: mak.spectrum@gmail.com*

***Зайцева Анна*** *– магистр кафедры ИУС Харьковского национального университета радиоэлектроники. Харьков. Украина.*

# DEVELOPING OF DISTRIBUTED VIRTUAL LABORATORIES FOR SMART SENSOR SYSTEM DESIGN BASED ON MULTI-DIMENSIONAL ACCESS METHOD

**Oleksandr Palagin, Volodymyr Romanov, Krassimir Markov, Vitalii Velychko, Peter Stanchev, Igor Galelyuka, Krassimira Ivanova, Ilia Mitov**

***Abstract**: In the article it is considered preconditions and main principles of creation of virtual laboratories for computer-aided design, as tools for interdisciplinary researches. Virtual laboratory, what are offered, is worth to be used on the stage of the requirements specification or EFT-stage, because it gives the possibility of fast estimating of the project realization, certain characteristics and, as a result, expected benefit of its applications. Using of these technologies already increase automation level of design stages of new devices for different purposes. Proposed computer technology gives possibility to specialists from such scientific fields, as chemistry, biology, biochemistry, physics etc, to check possibility of device creating on the basis of developed sensors. It lets to reduce terms and costs of designing of computer devices and systems on the early stages of designing, for example on the stage of requirements specification or EFT-stage. An important feature of this project is using the advanced multi-dimensional access method for organizing the information base of the Virtual laboratory.*

***Keywords**: Virtual Laboratory; Computer-Aided Design; Access Methods; Distributed System.*

***ACM Classification Keywords**: J.6 Computer-Aided Engineering –Computer-Aided Design (CAD); D.4.3 File Systems Management – Access Methods; K.4.3 Organizational Impacts – Computer-Supported Collaborative Work.*

***Conference**: The paper is selected from Seventh International Conference on Information Research and Applications – i.Tech 2009, Varna, Bulgaria, June-July 2009*

## Introduction

Fast spreading of market relations and competition between manufacturers of different (including scientific) production and information services makes very actual the acceleration of development of theory and methods of computer-aided design of computer devices and biosensors. Actual design of devices and systems, which is often used, needs a lot of time, material and human resources. If one needs to make a small set of devices by means of actual design, the price of final production becomes very high. Therefore, manufactures of computer devices get very complicated issue, which consists in time and price reduction of new devices design. Only after solving of this issue the new devices of own design will be able to become competitive on domestic and world markets.

To minimize these design expenses to reach high level of competitive recently side by side with actual design it is begun to use a virtual design. These methods realized by means of virtual laboratories of computer-aided design (VLCAD), which are based on advanced access methods and worth to be used on the stage of the requirements specification or EFT-stage, because it gives the possibility of fast estimating of the project realization, certain characteristics and, as a result, expected benefit of its applications.

Market analysis and joint discussion confirm the acute necessity in the developing of new virtual design methods and in the creating on their base open VLCAD, main feature of which is possibility to use such remote laboratory by specialists in different science branches, without education in information technologies and instrumentation.

## Preconditions and Main Principles of Virtual Laboratory Creation

One of problems, which are met by developers of new devices for different fields of science and engineering, is existence of more than 15 thousands of such fields or disciplines to date. Naturally to carry out researches or create a new device developers must have knowledge from disciplines, which refer to developed device. Therefore it is important to orientate new computer technology for interdisciplinary researches, which occur on boundary of several science fields or disciplines.

Urgency of these researches is caused by absence of computer technology of smart devices designing for interdisciplinary researches in Ukraine and Bulgaria. It does not allow to test on computer models the performance of designed devices, which are created on the base of new effects or sensors. To date to develop new device or to check the possibility of its creations and operation it is necessary to invite specialists in information technology, electronics and circuit technology on the commercial base. Getting results in such way is very expensive and, as usual, is not supported with necessary funds. This again confirms acute necessity of design technology development and creating on their base the open virtual laboratories, the main feature of which is possibility to use these virtual laboratories by specialists from different science fields, especially non-specialists in the field of information technology and instrument making.

Good solution of this problem is to create on the base of information technologies the special hardware-software tools [Palagin and Sergiyenko , 2003], which in convenient mode (for example, with help of dialogues) allows sensor developer to check possibility of creating of new devices and the device model. Such tool has to give possibility to create a model set of certain device (e.g. functional, electrical, operational etc.), including prior parameters calculations, project of circuit board and set of design documentations (e.g. cost, performance, validity, size, reliability etc.). Description of sensor or its model should be incoming data for such design system.

Now on the world market there are a lot of software for computer-aided design (CAD), which allow to automotive design of new devices and systems and analyze them in different ways [Gavrilov, 2000]. But for skilled usage of such CAD software it is necessary to have special skills in circuit technology, electronics and instrument engineering, and also know this CAD software perfectly. It is clear, that sensor developers, who are mainly chemists, biologists, biochemists, physicists etc, have no enough possibility and skills to use such complicated CAD software for designing of new devices on basis of developed sensors. In such case they need help of CAD specialists. But it is very expensive service. Therefore in most cases sensor developer leave sensor "in quiet" and switch his attention to another tasks.

It is necessary to note, that only by paying attention to the design process of computer devices it will be possible to reach a high level of competitiveness of scientific developments, what lets in the future to take up notable place on the world market. It is easily to see, that most devices have the same structure, to be exact, they consist of sensor, measuring channel, data processor, interface and additional subsystems. That’s why process of designing could be easily formalized.

To solve this problem within the bounds of international Ukrainian-Bulgarian project it is began developing of virtual laboratory for computer-aided design for computer device designing [Palagin et al, 2007]. The VLCAD is being created on the virtual methods of design [Galelyuka, 2008]. Offered virtual laboratory are created on the base of formalized representation of theoretic knowledge, principles of organization, methods and facilities of computer-aided design and testing information-measuring systems and devices, in particular on the base of subject field ontology. For VLCAD creating it is used the methodology of system integration [Palagin and Kurgaev, 2003] concerning base methods and tools, on which it is created. In the methodology basis it is putted system approach to tasks of analysis and synthesis of both VLCAD component and object of designing, and, first of all, forming knowledge system of interdisciplinary nature and its computer ontology. Proposed VLCAD is open system.

Mentioned VLCAD allows sensor developer to:

- check possibility of creating of devices and computer facilities (including portable devices) on basis of developed sensors without involving specialists in circuit technology and instrument engineering at the stage of EFT-project. It allows reducing terms and costs on this stage;
- avoid expensive actual tests on the stage of device creating by replacing with virtual methods of designing and testing;
- prepare set of design documentations on designed device in the automotive mode without involving corresponding specialists. Next stage is to send design documentations to contract production for creating of test party of devices.

Terms "Virtual laboratory" and "Virtual design" appear lately, so, as usual, they are absent almost in all dictionaries. The word "Virtual" appeared in word literature a long time ago. "Virtuality" has almost all features of empirical reality with the exception of its direct presence. So, it is "reality, which is absence" or "present absence". Also, "virtual" is one, which has no physical embodiment. "Virtual reality" is comprehended as a part of reality, which is modeled by computer device. Since any laboratory is a part of reality, so taking into account above-stated, there can be formulated next term of "virtual laboratory": virtual laboratory is imagined laboratory, which has all features of real laboratory and is modeled by means of software and hardware.

In general, virtual laboratory is some information environment, which lets to conduct researches in the case, when there is no direct access to test subject. Researches can be conducted by means of mathematical models and with using of remote access to test object.

Somebody may work with physical objects in two ways:

- emulation of physical objects with defined level of approximation to reality;
- remote access to physical objects with defined capabilities of interacting.

The first method lets to get completely virtual analog of some environment, what is very practical. Disadvantage of this method is complexity of model creating, which is very approximate to reality.

The second method provides maximal approximating to reality. But it requires creating and supporting of remote access to test objects, but the number of access channels is limited. Server of laboratory setup, besides access to equipments, is able to give background and methodological materials to researcher. Remote experiment in most cases is conducted in such way. Researcher communicates laboratory setup server and send data for experiment. Server software conducts experiment and sends results as tables, graphics to researcher.

For realization of VLCAD it is decided to use the first method. But the second method is not set aside and in future it will be probably used as additional tool.

Virtual laboratories, in which experiments are conducting by means of mathematical models, differ from previous one by using mathematical or other model instead of real test object. These laboratories have no laboratory setup.

## Creating of VLCAD

Before VLCAD creating, first of all, it is necessary to determine features of VLCAD as tool for interdisciplinary researches and what functions it has to have.

In general, VLCAD is a system for computer-aided design, but with certain difference. This difference is that for using any CAD system it is necessary to have deep knowledge in this software, instrument engineering, circuit technology and electronics. It is expected, that for using VLCAD users need only experience in work with computer. Design process by means of VLCAD is much regulated and is going on dialog mode with additional help messages. So, the main feature of VLCAD as tool for interdisciplinary researches is orientation of this system in the side of usual users, which are nospecialists in the field of information technology, instrument

engineering and circuit technology. It make practicable to develop new device or verify possibility of such development by such specialists, as biologists, ecologists, medics, biochemists at el.

For such VLCAD creating, first of all, it is necessary to execute next actions:

- improve design process on the base of using mathematical methods and computer tools [Palagin et al, 1993];
- automate process of searching, processing and issuing of information;
- use methods of optimal and variant designing, effective mathematical models of design object, components and materials;
- create multi-dimensional hierarchical databases with integrated data of reference type, needed for computer-aided design;
- improve quality of designed document execution;
- increase creative part of designer work at the expense of automation of noncreative routine work;
- unify and standardize design methods;
- train specialists, including students, masters etc.;
- implement interaction with automatic systems of different levels and purposes.

To define place of VLCAD in the design process it is necessary to take into account world experience of design engineers of computer and portable devices. Integrated scheme of design process with proper outlet documentation and the place of VLCAD in design process are shown on fig. 1. As one can see VLCAD covers early stages of designing.

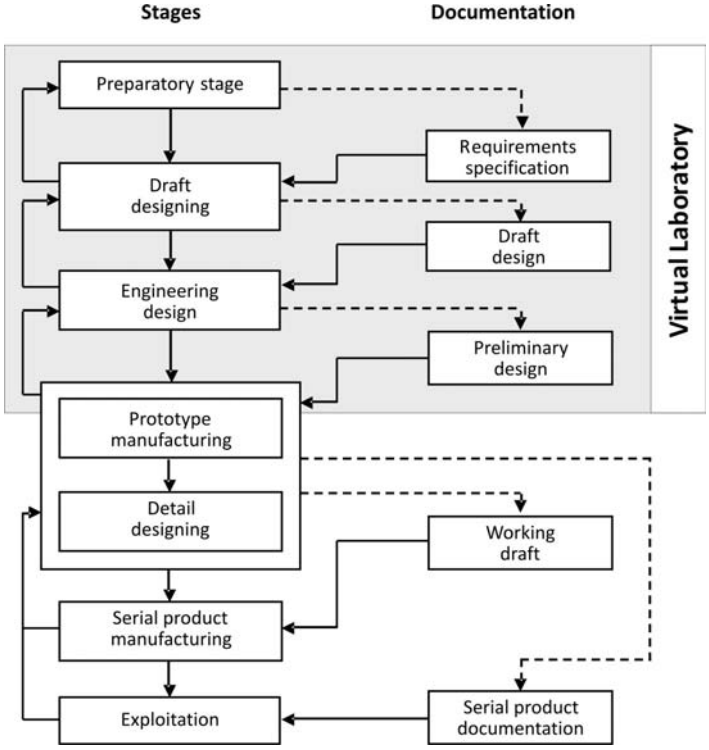



Fig. 1. Integrated scheme of design process with proper outlet documentation

Since VLCAD has many features of CAD system it is rationally to use methodology of CAD system creating during VLCAD developing, but taking into account features of VLCAD. It is necessary to note, that now there are several conceptions of CAD system creating. Full-automatic and man-machine systems are the most widespread. First systems are difficult to build and, in some cases, it is impossible to create such full-automatic system, because design process is heterogeneous, has many internal and external connections and includes a lot of undefined factors. To take into account these undefined factors it is necessary to use creative opinion of designer.

Taking into account described above we can state, that creating of VLCAD for computer device design is very important scientific-technical problem, and implementation of such VLCAD needs certain investment. Received experience and analysis of world literature let us to separate out next main principles of such virtual laboratories creating:

1. Virtual laboratory is man-machine system. All design systems, which had been developed and now are being developed, are computer-aided, and designer is the main part of these systems. Human in such systems has to solve tasks, which cannot be well defined, and problem, which human by using own heuristic abilities may solve better and more effective than computer. Close interaction between human and computer during design process is one of principles of development and exploitation of any CAD systems for computer device designing.

2. Virtual laboratory is hierarchical system, which use comprehensive approach to automation of all design levels. Level hierarchy is presented in system structure as hierarchy of subsystems.

3. Virtual laboratory is set of informational-concerted subsystems. This very important principle refers not only to connections between large subsystems, but to connections between separate parts of subsystems. Informational compliance means, that almost all possible sequences of design tasks are served by informational-concerted programs. Two programs are informational-concerted if all data in these programs are part of numeric arrays and do not need transformations during sending from one program to another and inversely. So, results of one program can be incoming data for another program.

4. Virtual laboratory is open system, which are permanently expanding. Permanent progress of technology, designed objects, computer technology and computational mathematics lead to appearance of new, more perfect mathematical models and programs, which replace old analogs. So, VLCAD has to be open system and be able to use new methods and tools.

5. Virtual laboratory is specialized system with maximum using of unified units. Requirements of high efficiency and universality for any system are, as a rule, conflicting or competitive. It is reasonable to develop VLCAD on the base of unified parts. Necessary condition of unification is searching of common principles in the modeling, analysis and synthesis of technical objects.

Computer technology, what are offered by us, is hardware-software complex, what consist of personal computers or work stations with set of necessary peripheral items, connected in local and worldwide networks, such as Internet, and is supplied with all software. Using of these technologies already increase automation level of design stages of new devices for different purposes, including devices for interdisciplinary researches.

Today such complex systems, as VLCAD and CAD, are developed as knowledge-oriented systems, main feature of which is informational integration. Informational integration is the main application area of ontology using. Ontology, as a rule, contains hierarchy of concepts of knowledge domain and describes important features of every concept by means of mechanism "attribute–value". Connection between concepts may be described by means of additional logical statements. Constants refer to one or several concepts. This and another ontology features let to use ontology in different fields of knowledge, increasing effect from application of different methods and modes of work with information or creating on their base new more effective methods [Palagin, 2005]. Especially efficiency of ontology application can be shown in such science intensive fields, as knowledge engineering and knowledge management, objects and processes modeling, databases designing, informational integration and data mining [Gladun, 1994].

Analysis of literature and certain application domain lets to specify requirements to ontology, on the base of which VLCAD is developing [Palagin et al, 2007], [Galelyuka, 2008]:

- Ontology has to include conceptual knowledge, but not episodic ones.
- Ontology has to be specified and internal concerted with structure, names and content for all defined conceptions.
- Ontology has to be structured and simple for understanding and searching of conceptions.
- Ontology has to be limited by certain application domain for defining of used conceptions. Ontology has not to include all possible information about application domain.

## VLCAD storage space

As a storage space for VLCAD a multi-dimensional access method, called ArM32, property of FOI Creative Ltd. may be used. It is built on the base of the Multi-Domain Information Model (MDIM) [Markov, 2004].

The ArM32 elements are organized in a hierarchy of numbered information spaces with variable ranges. There is no limit for the ranges the spaces. Every element may be accessed by correspond multidimensional space address given via a coordinate array.

The Multi-Domain Information Model (MDIM), presented in [Markov, 2004], is a step in the process of development of tools for data-base organization. Its main idea is to permit practically unlimited access to multi-dimensional information structures. In MDIM there exist two main constructs – numbered information spaces and basic information elements.

The Basic information element is an arbitrary long string of machine codes (bytes). When it is necessary the string may be parceled out by lines. The length of the lines may be variable. In ArM32 the length of the string may vary from zero up to 1GB. There is no limit for the number of strings in an archive but theirs total length plus internal indexes could not exceed the limit for the length of a single file of the operating system.

Basic information elements are united in numbered sets, called numbered information spaces of range 1. The numbered information space of range n is a set, which elements are numerically ordered information spaces of range n-1.

ArM32 allows using of information spaces with different ranges in the same archive (file).

The main ArM32 operations are reading, writing, appending, inserting, removing, replacing and deleting of a basic information element or any it's part.

The ArM32 numbered information spaces are ordered and main operations within spaces take in account this order. So, from given space point (element or subspace) we may search the previous or next empty or non empty point (element or subspace). In is convenient to have operation for deleting the space as well as for count its nonempty elements or subspaces.

ArM32 engine supports multithreaded concurrent access to the information base in real time.

Very important feature of ArM32 is the possibility not to occupy disk space for empty structures (elements or spaces). Really, only non empty structures need to be saved on external memory.

## Conclusion

For increasing of competitiveness of science products it is necessary to develop new hardware-software tools, what is applicable for using in interdisciplinary researches. Virtual laboratory for computer-aided design can serves as example of such tool. In the article it is considered preconditions and main principles of such virtual laboratories creation, main purpose of which is to give possibility for sensor developers to verify ability of creating new devices on the base of their sensors on the early stages of designing, particularly on the stage of requirements specification or EFT-stage.

The features of ArM32 are appropriate for building the information base of VLCAD. The multi-dimensional information spaces make possible the effective creating of complex information structures using small amount of resources which is very important for VLCAD. At the first place the ontology's' representing and knowledge formation processes as well as intelligent recognition and classification are realizable.

## Acknowledgements

This work is partially financed by Bulgarian National Science Fund under the joint Bulgarian-Ukrainian project **D 002-331 / 19.12.2008** "Developing of Distributed Virtual Laboratories Based on Advanced Access Methods for Smart Sensor System Design" as well as Ukrainian Ministry of Education under the joint Ukrainian-Bulgarian project No: **145 / 23.02.2009** with the same name.

## Bibliography

[Gavrilov, 2000] Gavrylov L. Computer-aided design (CAD) systems for analog and analog-digital devices // Electronic component. – 2000. – № 3. – P. 61–66. (In Russian)

[Galelyuka, 2008] Galelyuka I. Elements of theory and tools for virtual designing of computer devices and systems of automation of biological objects experimental researches: Thesis for the candidate's degree of the technical sciences on the specialty 05.13.06 – Information technologies / I. Galelyuka. – Kiev, 2008. – 20 p. (In Ukrainian)

[Gladun, 1994] V. P. Gladun. Processes of New Knowledge Formation. Sofia, SD Pedagog 6, 1994, 192 p, (in Russian).

[Markov, 2004] K. Markov. Multi-Domain Information Model. Int. Journal "Information Theories and Applications", Vol.11, No.4, 2004, pp. 303-308.

[Palagin et al, 1993] Palagin O., Denisenko E., Belycjkyy R., Sigalov V. Microprocessor system for data processing: designing and debugging / editor Beh A. – Kiev: Naukova dumka, 1993. – 352 p. (In Russian)

[Palagin and Sergiyenko, 2003] Palagin O., Sergiyenko I. Virtual scientific-innovative centers: conception of creating and perspectives of development // Control systems and computers. – 2003. – № 3. – P. 3–11. (In Russian)

[Palagin and Kurgaev, 2003] Palagin O., Kurgaev A. Problem orientation in the development computer architecture // Cybernetics and system analysis. – 2003. – № 4. – C. 167–180. (In Russian)

[Palagin, 2005] Palagin O., Yakovlev Yu. System integration of computer facilities. – Vinnitsa: Universum-Vinnitsa. – 2005. – 680 c. (in Russian)

[Palagin et al, 2007] Palagin O., Romanov V., Sachenko A., Galelyuka I., Hrusha V., Kachanovska M., Kochan R. Virtual Laboratory for Computer-Aided Design: Typical Virtual Laboratory Structure and Principles of Its Operation // Proceeding of 4th IEEE Workshop "Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2007)". – Dortmund, Germany. – 6–8 September, 2007. – P. 77–81.

## Authors' Information

***Oleksandr Palagin*** *– Depute-director of V.M. Glushkov's Institute of Cybernetics of National Academy of Sciences of Ukraine, Academician of National Academy of Sciences of Ukraine, Doctor of technical sciences, professor; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine; e-mail: palagin_a@ukr.net*

***Volodymyr Romanov*** *– Head of department of V.M. Glushkov's Institute of Cybernetics of National Academy of Sciences of Ukraine, Doctor of technical sciences, professor; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine; e-mail: dept230@insyg.kiev.ua, VRomanov@i.ua*

***Krassimir Markov*** *– Assoc. Professor; Institute of Mathematics and Informatics, BAS, Acad. G.Bontchev St., bl.8, Sofia-1113, Bulgaria; e-mail: markov@foibg.com*

***Vitalii Velychko*** *– Doctoral Candidate; V.M.Glushkov Institute of Cybernetics of NAS of Ukraine, Prosp. Akad. Glushkov, 40, Kiev-03680, Ukraine; e-mail: glad@aduis.kiev.ua*

***Peter Stanchev*** *– Professor, Kettering University, Flint, MI, 48504, USA Institute of Mathematics and Informatics – BAS; Acad. G.Bontchev St., bl.8, Sofia-1113, Bulgaria; e-mail: pstanche@kettering.edu*

***Igor Galelyuka*** *– Research fellow of V.M. Glushkov's Institute of Cybernetics of National Academy of Sciences of Ukraine; Candidate of technical science; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine; e-mail: galib@gala.net*

***Krassimira Ivanova*** *– Researcher; Institute of Mathematics and Informatics, BAS, Acad. G.Bonthev St., bl.8, Sofia-1113, Bulgaria; e-mail: ivanova@foibg.com*

***Ilia Mitov*** *– PhD Student of the Institute of Mathematics and Informatics, BAS, Acad. G.Bontchev St., bl.8, Sofia-1113, Bulgaria; e-mail: mitov@foibg.com*