PROCESSING SETS OF CLASSES' LOGICAL REGULARITIES

Anatoliy Gupal, Maxim Novikov, Vladimir Ryazanov

Abstract: *The paper considers two methods for processing sets of logical regularities of classes (LRC) found by training samples analysis. The first approach is based on the minimization of logical descriptions of classes. As a result of solving the problem of linear discrete optimization, the shortest logic description of each class is found. Each training object satisfies at least to one LRC of found irreducible subset of logical regularities. The second approach is based on the clustering of the set of LRC and selecting standards of derived clusters. The clustering problem is reduced to the clustering of representations of LRC set. Here each LRC is represented in the form of binary vector with different informative weight. A modification of the known method of "variance criterion minimization" for the case where the objects have different information weights is proposed. We present the results of illustrative experiments.*

Keywords: *classification, logical regularity of class, feature, clustering*

ACM Classification Keywords: *I.2.4 Artificial Intelligence Knowledge Representation Formalisms and Methods – Predicate logic; I.5.1 Pattern Recognition Models – Deterministic, H.2.8 Database Applications, Data mining*

Introduction

Methods of classification by precedents got now widely fame and spread in various fields of science, production and social life. There are many practical problems where the initial data (training samples) have the form $\{w_i, \mathbf{x}_i, i = 1, 2, ..., m\}$, $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{in})$ is a feature description of object, $w_i \in \{1, 2, ..., l\}$ is a number of "class" of the object. We will further assume that the features of the object are real values, the object itself and its feature descriptions will be identified, training sample comprises representatives of each class, the training sample is noncontradictory (there are no equal objects in different classes). The main problem is to construct algorithm A that for any new feature description $\mathbf{x}$ will calculate its class: $w = A(\mathbf{x})$. Currently, there are a lot of different approaches and specific algorithms for solving classification problems based on different principles and ideas. Statistical algorithms are based on probabilistic hypotheses about the structure of classes and suggest the existence of a priori probabilities of classes and conditional probability distributions for each class of objects [Duda et al, 2000]. Methods based on the construction of separating surfaces, find the optimal

separating surfaces for training data (linear, polynomial, piecewise-linear, linear combinations of potential functions, etc.) and use them for further classification [Duda et al, 2000; Vapnik, 1995]. Neural network approaches are the methods to find the optimal superposition of threshold functions and are attempt to simulate the human thinking [Wasserman, 1992]. Logical approaches (models of partial precedent [Zhuravlev, 1978; Zhuravlev et al, 2006] and decision trees [Quinlan, 1993]) are natural generalization of the binary representations of functions in disjunctive normal form. Many approaches are in fact hybrid and combine several different principles. The present work is related to models of a partially precedent approach. Along with the creation of the classifier A on the training set, practical user is interested in illustrative communication by analytical formulas between baseline characteristics and classes, in descriptions of the classes themselves. In this regard, the logical approaches have certain advantages. They can be used to construct as recognition algorithms and also analytical description of classes. Article is devoted to the second part. The LRCs may be similar and even equal. LRC may be some degenerate solution, the number of calculated LRC can be quite large. Practical user is interested in the same small number of LRC, but different and informative. This situation undoubtedly requires the development and use of various means of processing LRC.

Recognition algorithms based on the weighted voting over system of logical regularities

Let us consider the problem of recognition by precedents (supervised classification) in the following standard statement [Zhuravlev, 1978]. We believe that given a set M of objects $\mathbf{x}$. Each object is given in terms of feature values $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}, x_i \in R$, and $M = \{\mathbf{x}\} = \bigcup_{i=1}^{I} K_i, K_i \cap K_j = \varnothing, i, j = 1,2,\ldots,I, i \neq j$. Subsets K_i are called classes. The initial information about M and its partition to classes given in the form of training sample $X = \{\mathbf{x}\} \subseteq M$, containing representatives of all classes. Training sample is consistent, i.e. not containing equal objects of different classes. Required to set up an algorithm A that classifies any object $\mathbf{x}$ to one of the classes, or refuse to classify the object.

We will write and consider $(a \le x)$ the expression equal to one when it is running, and zero otherwise. Next, we use the following definition LRC [Ryazanov, 2007].

Definition 1. The predicate

$$P^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x}) = \&_{i\in\Omega_1}(a_i \le x_i) \&_{i\in\Omega_2}(x_i \le b_i)$$

is a logical regularity of class (LRC) $K_t, t = 1,2,\ldots,I$, if the following conditions are satisfied

1. $\exists \mathbf{x}_j \in X \cap K_t : P^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x}_j) = 1.$

2. $\forall \mathbf{x}_j \notin X \cap K_t : P^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x}_j) = 0.$

3. $F\left(P^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x})\right) = \text{extr}\, F\left(P^{\mathbf{a}^*,\mathbf{b}^*,\Omega_1^*,\Omega_2^*}(\mathbf{x})\right)$, where $\mathbf{a}^*,\mathbf{b}^* \in R^n, \mathbf{a}^* \le \mathbf{b}^*, \Omega_1^*, \Omega_2^* \subseteq \{1,2,\ldots,n\}$, F - some criterion of predicate quality.

In the last condition we consider the problem of local maximizing of the criterion $F\left(P^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x})\right) = \left|\left\{\mathbf{x}_j \in X \cap K_t : P^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x}_j) = 1\right\}\right|$. Two LRC of class K_t will be called equivalent if their values on the objects of training sample are the same. The set $N_\lambda = \left\{\mathbf{x} : P_\lambda^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x}) = 1\right\}$ is called as interval for LRC $P_\lambda^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x})$. Since (for simplicity), we believe that all the features are real, there are a continuum of LRC equivalent for any LRC. For any set of equivalent LRC there is the only minimal LRC. In [Kovshov, 2008] are shown the various approximate and exact methods of finding minimal LRC. One can show that they have the form $P^{\mathbf{a},\mathbf{b}}(\mathbf{x}) = \mathop{\&}_{i\in 1,2,\ldots,n} (a_i \le x_i \le b_i)$. Thus, we assume that using the training set X for each class K_t is calculated a lot of some LRC: $\mathbf{P}_t = \left\{P_i^{\mathbf{a}_i,\mathbf{b}_i}(\mathbf{x})\right\}$. We put in correspondence the set of intervals $\mathbf{N}_t$ for each set $\mathbf{P}_t$. In [Ryazanov, 2007] presented a method for calculating the coefficients γ_λ in the expression for estimates of the classes in the form $\Gamma_t(\mathbf{x}) = \sum_{P_\lambda \in \mathbf{P}_t} \gamma_\lambda P_\lambda^{\mathbf{a}_\lambda,\mathbf{b}_\lambda}(\mathbf{x})$. Further, two approaches are proposed to processing sets $\mathbf{P}_t$.

Building shortest logical descriptions of classes

In [Zhuravlev et al, 2006] proposed the based on the minimization of logical descriptions of classes method for processing a set of LRC. This approach is similar to the minimization of partially defined Boolean functions. Suppose that for a certain class K_t some system $\mathbf{P}_t$ of LZC was found. From their definition it follows that as an approximation of the characteristic function of the class can take the disjunction of all LRC. This function is equal to 1 on all objects of training of K_t and 0 on all objects of training from other classes.

Definition 2 [Zhuravlev et al, 2006]. Logical description of a class K_t is a function $D_t(\mathbf{x}) = \bigvee_{P_i \in \mathbf{P}_t} P_i^{\mathbf{a}_i,\mathbf{b}_i}(\mathbf{x})$.

LRC can be calculated using the recognition system [Zhuravlev et al., 2006]. By analogy with the minimization of partially defined Boolean functions definitions of the shortest and the minimal logical descriptions of classes are introduced.

Definition 3 [Zhuravlev et al, 2006]. Shortest logical description of the class K_t is a function $D_t^s(\mathbf{x}) = \bigvee_{P_i \in \tilde{\mathbf{P}}_t \subseteq \mathbf{P}_t} P_i^{\mathbf{a}_i,\mathbf{b}_i}(\mathbf{x})$ where

1. $D_t^s(\mathbf{x}_i) = D_t(\mathbf{x}_i), i = 1,2,...,m$
2. $\left|\tilde{\mathbf{P}}_t\right| \to \min\limits_{\tilde{\mathbf{P}}_t \subseteq \mathbf{P}_t}$.

Figures 1 and 2 are examples of the covering of sets of training objects from the class K_t by intervals N_i corresponding to all LRC of $\mathbf{P}_t$ and LRC of $\tilde{\mathbf{P}}_t$. In the example, two marked by circles with a black and a white center classes of two-dimensional objects have been considered.

Here, as an approximation of the characteristic function is sufficient to use 3 logical regularities. The shortest logical description of the class is a result of the solution of integer linear programming problem (here N is the number of all LRC):

$$\sum_{i=1}^{N} y_i \to \min,$$

$$\sum_{i=1}^{N} P_i(\mathbf{x}_j) y_i \geq 1, \forall \mathbf{x}_j \in K_t, y_i \in \{0,1\}.$$

The set of unit values of y_i defines a subset of predicates $\tilde{\mathbf{P}}_t$.

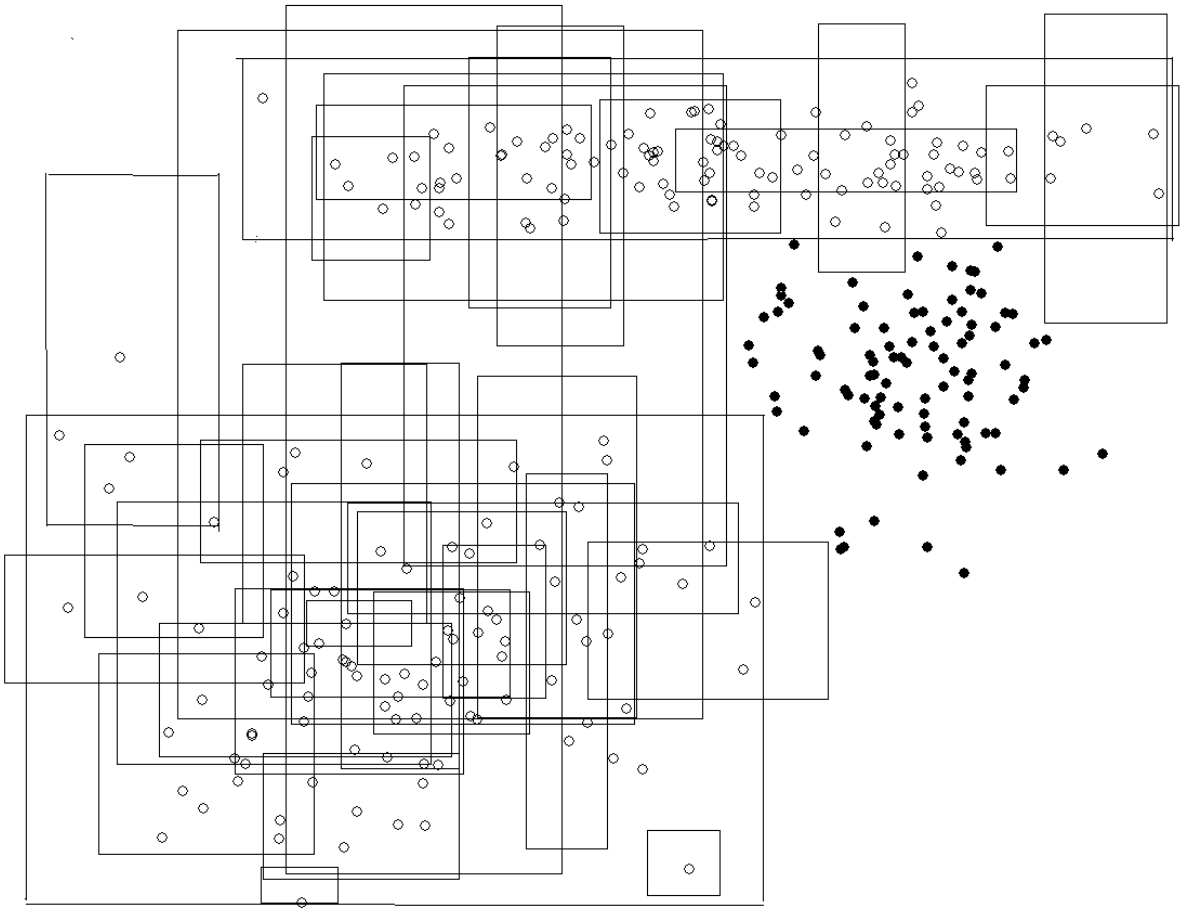


Figure 1. Covering points of the first class of training sample by found intervals from $\mathbf{N}_t$

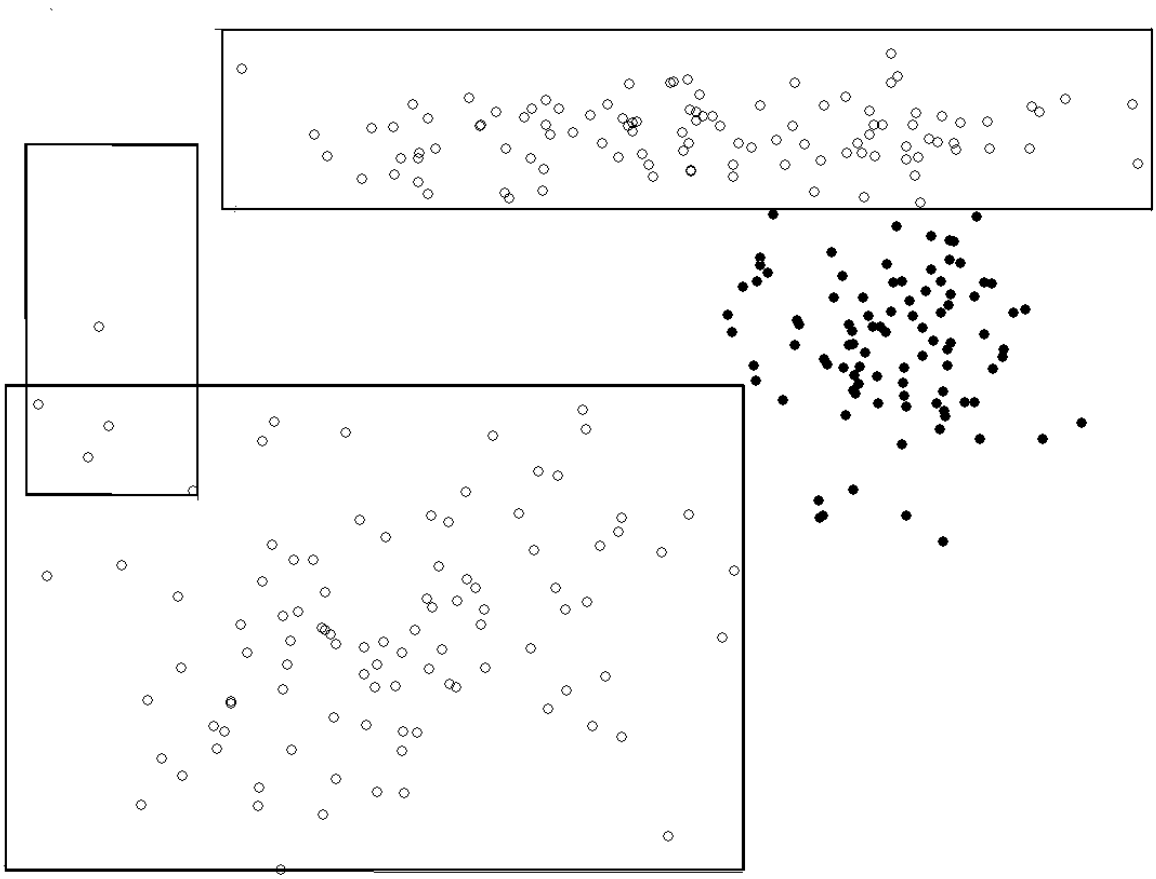


Figure 2. Covering points of the first class of training sample by found intervals from $\tilde{\mathbf{N}}_t$

Processing of sets of LRC using cluster analysis

The second approach to processing sets of LRC is based on cluster analysis. The general idea is as follows. When processing a set of vectors we can solve the problem of clustering for 2, 3, ... clusters depending on the prior knowledge or desire. Its "standard" is calculated (for example, the sample mean vector) for each cluster. The resulting system "of standards" is taken as the result of processing of the original set of vectors of precedents. In our case, the objects are clustering functions (LRC). As a result of clustering of the set LRC for l classes and computation for each cluster its standards we get new predicates, which in general can not be in the original set. These predicates are generally "partial" LRC, i.e. they can take the value 1 on a few of objects of other classes. These predicates are "sufficiently" different, they are measured. Thus, by the initial set of LRC we can calculate and evaluate a given number of "sufficiently" different partial logical regularities. Figures 3 and 4 are examples for illustration. On Figure 3 points of a class (the class "circles") are covered by the system of LRC. Figure 4 shows the same example, but is shown only two intervals. Intervals are significantly different and mainly cover a significant number of points of the class under consideration.

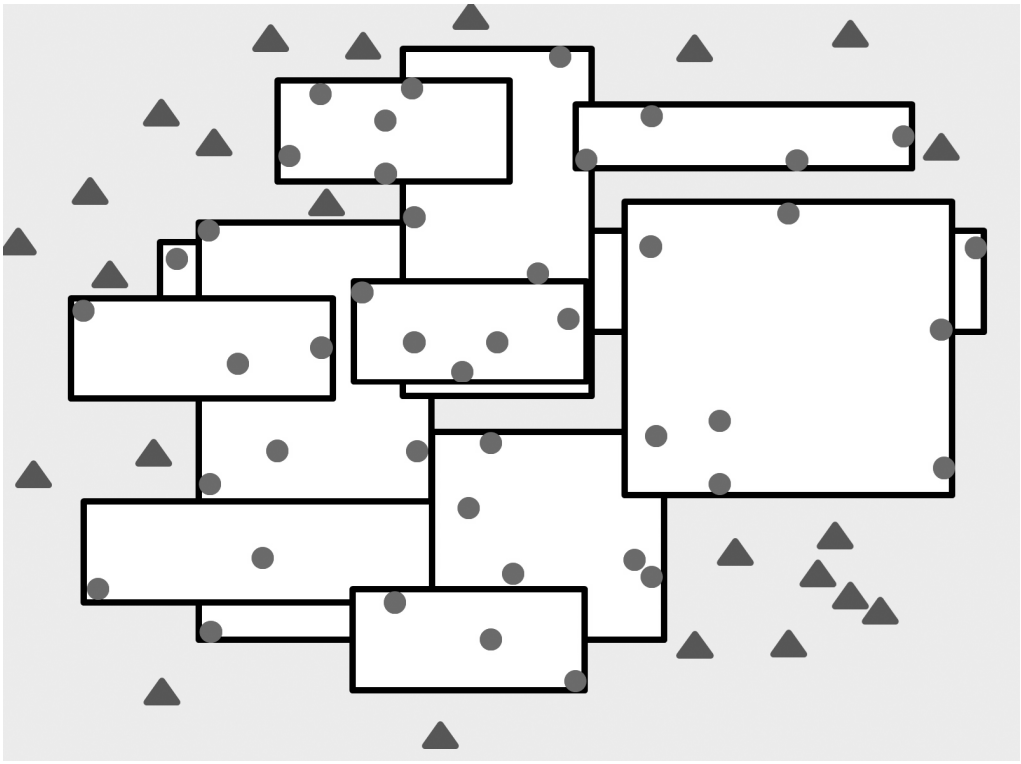


Figure 3. The points covered by a large number of class intervals

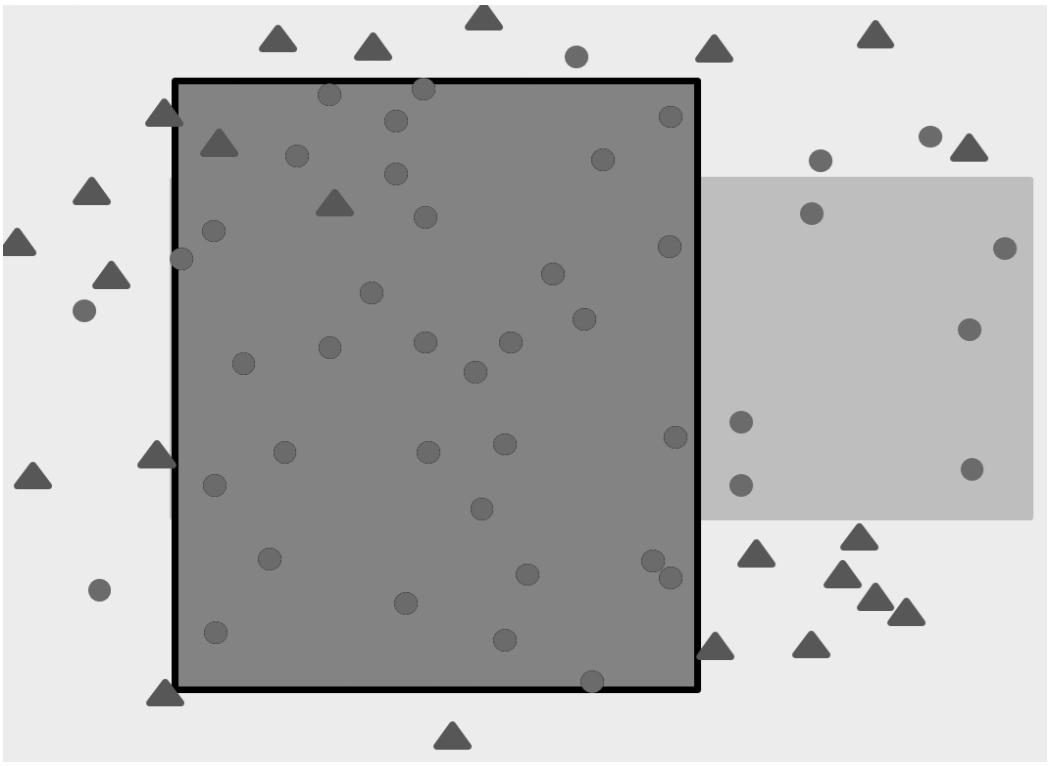


Figure 4. "Substantial" number of class points is covered by two intervals containing perhaps a small number of elements of other classes

To implement this idea we must to create a method of clustering of a set of functions and to calculate the "standard" for a variety of functions that form a cluster. The functions must be weighted. There are two possible ways. The first way is connected with the generalization of existing approaches to clustering the set of vectors to the set of functions. The second way is connected with a one-to-one representation LRC in the form of vectors. In this case, the application of existing clustering methods is possible. The present work is devoted to the second approach.

Put in one-to-one correspondence of each $P_i^{\mathbf{a}_i,\mathbf{b}_i}(\mathbf{x})$ of $\mathbf{P}_t$ the binary vectors as follows: $P_i^{\mathbf{a}_i,\mathbf{b}_i}(\mathbf{x}) \Leftrightarrow \mathbf{z}_i = (\mathbf{z}_{i1}, \mathbf{z}_{i2}, ..., \mathbf{z}_{ih}), \mathbf{z}_{ij} \in \{0,1\}, j = 1,2,...,h$. Here $h = |K_t|$, the vector $\mathbf{z}_i$ marks the original training objects from the class K_t in which the predicate $P_i^{\mathbf{a}_i,\mathbf{b}_i}(\mathbf{x})$ is unity. The weight of each vector $\mathbf{z}_i$ (and corresponding LRC $P_i^{\mathbf{a}_i,\mathbf{b}_i}(\mathbf{x})$) equals to a fraction of class objects K_t for which LRC is equal to 1. Thus, the original problem is reduced to the clustering of the set of binary vectors $\mathbf{z}_i = (\mathbf{z}_{i1}, \mathbf{z}_{i2}, ..., \mathbf{z}_{ih})$ with known weights $y(\mathbf{z}_i)$ and calculation of the mean of each cluster. Further, each of the sample mean is associated with some in general partial LRC. As a basic clustering method we take the method based on the minimization of variance criterion. Suppose we have fixed the number of clusters l.

Clustering on l clusters by minimizing the variance criterion is formulated in the following way:

$$J(\mathrm{K}) = \sum_{i=1}^{l} \sum_{\mathbf{z}_t \in K_i} y_t \left\| \mathbf{z}_t - \mathbf{m}_i \right\|^2 \to \min_{\mathrm{K}} \tag{1}$$

where $y_t = y(\mathbf{z}_t)$, $\mathrm{K} = \bigcup_{i=1}^{l} K_i, K_i \cap K_j = \varnothing, i \neq j, i, j = 1,2,...,l$, $\mathbf{m}_i = \dfrac{\sum_{\mathbf{z}_t \in K_i} y_t \mathbf{z}_t}{\sum_{\mathbf{z}_t \in K_i} y_t}$.

In a standard clustering algorithm we have $y_i = 1, i = 1,2,...,m$. It is known that in this case, while minimizing the criterion $J(\mathrm{K}) = \sum_{j=1}^{l} \sum_{\mathbf{x}_i \in K_j} \left\| \mathbf{z}_i - \mathbf{m}_j \right\|^2 = \sum_{j=1}^{l} J_j(K_j)$, the local optimality condition for $\mathrm{K} = \{K_1, K_2, ..., K_l\}$ is the implementation of inequalities

$$\frac{n_i}{(n_i - 1)} \left\| \hat{\mathbf{z}} - \mathbf{m}_i \right\|^2 - \frac{n_j}{(n_j + 1)} \left\| \hat{\mathbf{z}} - \mathbf{m}_j \right\|^2 \leq 0 \tag{2}$$

for any pair K_i, K_j, and any $\hat{\mathbf{z}} \in K_i$. This means that any movement of the object of the cluster to which it belongs to any other cluster does not lead to a reduction of the dispersion criterion. We can

prove that in the general case there is an analogue (3) of condition (2) for the task (1), when there are not unitary weights. For simplicity, we write $\sum y$ instead of $\sum_{\mathbf{z}_t \in K_i} y_t$, and $\sum y\mathbf{z}$ instead of $\sum_{\mathbf{z}_t \in K_i} y_t \mathbf{z}_t$.

We have the expressions $\mathbf{m}_i^* = \mathbf{m}_i - \frac{\hat{y}(\hat{\mathbf{z}} - \mathbf{m}_i)}{(\sum y - \hat{y})}$ and $\mathbf{m}_j^* = \mathbf{m}_j + \frac{\hat{y}(\hat{\mathbf{z}} - \mathbf{m}_j)}{(\sum y + \hat{y})}$ when transferring $\hat{\mathbf{z}}$ from K_i to K_j.

Then $K_i \to K_i^* = K_i \setminus \{\hat{\mathbf{z}}\}, K_j \to K_j^* = K_j \cup \{\hat{\mathbf{z}}\}$ and

$$\frac{\sum\limits_{K_i} y\hat{y}}{(\sum\limits_{K_i} y - \hat{y})} \left\| \hat{\mathbf{z}} - \mathbf{m}_i \right\|^2 - \frac{\sum\limits_{K_j} y\hat{y}}{(\sum\limits_{K_j} y + \hat{y})} \left\| \hat{\mathbf{z}} - \mathbf{m}_j \right\|^2 \le 0 \tag{3}$$

As a result, the vector $\mathbf{m}_i = (m_{i1}, m_{i2}, \ldots m_{ih})$ is calculated for each cluster. By construction, $m_{ij} \in \{0, \alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iu}\}, 0 \le \alpha_{i\sigma} < \alpha_{i,\sigma+1} \le 1$ and values $\alpha_{i\sigma}$ are calculated by the obtained clustering. The standard of each cluster will be assumed the Boolean vector $\mathbf{b}_i = (b_{i1}, b_{i2}, \ldots, b_{ih}), b_{ij} \in \{0,1\}$, where $b_{ij} = \begin{cases} 1, & m_{ij} \ge \theta_i, \\ 0, & \text{otherwise.} \end{cases}$ Here, θ_i is chosen from a finite set $D_i = \{0, \alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iu}\}$. The vector $\mathbf{b}_i$ is in correspondence to each selection of θ_i, the selection of the optimal value of the vector is carried out by solving the problem of one-dimensional optimization $\Phi(P(\mathbf{x}, \theta_i)) \to \max\limits_{\theta_i \in D_i}$. Note that, in general, partial logical regularities of classes are correspondent to the $\mathbf{b}_i$.

There are different criteria Φ of quality of partial logical regularities of class, which correspond to a choice of θ_i. An example of a criterion $\Phi(P(\mathbf{x}, \theta))$ may be the criterion $\Phi(P(\mathbf{x}, \theta)) = \sqrt{p_i} - \sqrt{n_i}$ [Cohen et al, 1999], where p_i is a number of training objects of the class K_i in which the corresponding to the selected θ_i predicate $P(\mathbf{x})$ is performed, n_i is the number of training objects of other classes, for which the predicate is executed. There are various other evaluation criteria for $P(\mathbf{x})$, that take into account the number of sampling objects in the class and the number of other training objects.

The experimental results on model and practical problems

Practical experiment was performed on problem „breast" [Mangasarian et al, 1990]. This problem of recognition of breast cancer by precedents contained 218 first-class objects (benign tumor) and 126 objects of the second class (malignant neoplasm). Each object is described by 9 discrete features which are denoted as $X1, X2, ..., X9$. For the second class with an approximate algorithm [Zhuravlev et al, 2006] was found 17 logical regularities and the shortest logical description of the class from 7 conjunctions. Below in Table 1, the shortest logical description is represented on the left side. For each conjunction, its serial number from the overall list LRC and share objects of the second training class that have been met are given.

Table 1. Comparison of clustering standards on data „breast"

(3.5<=X5) (4.5<=X6) (X8<=6.5) {60, 0.28} V (6.5<=X4) {61, 0.40} V (6.5<=X3) {77, 0.48} V (X6<=4.1)(4.0<=X8) {78, 0.14} V (6.5<=X1)(3.3<=X6) {79, 0.54} V (2.0<=X3)(4.5<=X7) {81, 0.67} V (2.0<=X1)(5.8<=X9) {83, 0.10}	$\Phi(P(\mathbf{x},\theta))/K_2$	n_i	Logical regularities
	0.84	8	$(4 \le x_2 \le 10)(3 \le x_3 \le 10)(2 \le x_5 \le 10)(2 \le x_7 \le 10)$
	0.78	5	$(3 \le x_1 \le 10)(2 \le x_3 \le 10)(2 \le x_5 \le 10)(4 \le x_6 \le 10)$
	0.67	4	$(2 \le x_3 \le 10)(2 \le x_5 \le 10)(5 \le x_7 \le 10)(1 \le x_9 \le 8)$

In the right part of Table 1, there are shown the standards of resulting clustering of 17 logical regularities of second class on 3 clusters by the use of criterion [Cohen et al, 1999] for construction of standards. There are shown the values of quality and power of obtained clusters. The problem of minimizing the number of features in the standards was not considered. The arrows indicate the LRC of the shortest logical description that are closest to the calculated standards (here the Hamming distance between the 126-dimensional binary vectors was used).

Conclusion

In this paper we have proposed two approaches to processing a set of logical regularities of class based on finding of the shortest logical descriptions and methods of cluster analysis. An illustrative comparison of these approaches on the same medical problem was conducted. Both approaches are useful tools for analyzing sets of LRCs. The first approach gives us a method to select practically a small number of LRC of their complete list, using the principle of minimum complexity class descriptions. In the second

approach, new 1, 2, 3, ... standard LRCs (generally partial LRCs) representing the original set are calculated based on the existing set of LRCs. The number of standards is equal to the number of clusters, which is set by the user. Of interest is the creation and use of other models of clustering LRC sets, as well as the use of different criteria and their quality evaluation.

Acknowledgements

This work was supported by the Program of the Presidium of RAS №15 „Information, control and intelligent technologies and systems", Program №2 of Mathematical Sciences Department of RAS, RFBR № 14-01-90413 Ukr_a, 14-01-90019 Bel_a, 14-01- 00824 a, 13-01-12033 ofi m, 13-01-90616 Arm_a, 12-01-00912-a, 15-01-05776-a.

Bibliography

[Cohen et al, 1999] William W. Cohen and Yoram Singer Simple, Fast, and Effective Rule Learner, AAAI/IAAI 1999: 335-342.

[Duda et al, 2000] Duda, R. O., Hart, P. E., and Stork, D. G. Pattern Classification. John Wiley and Sons, 2nd edition, 2000

[Kovshov et al., 2008] Kovshov N.V., Moiseev V.L., and Ryazanov V.V. "Algorithms for finding logical regularities in pattern recognition problems", Computational Mathematics and Mathematical Physics, M .: Nauka. T.48, 2008, N 2, pp. 329-344.

[Mangasarian et al, 1990] Mangasarian O. L., Wolberg W.H.: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 - 18.

[Quinlan, 1993] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993

[Ryazanov, 2007] Ryazanov V.V. "Logical regularities in pattern recognition (parametric approach)", Computational Mathematics and Mathematical Physics, T.47, №10, 2007, p.1793-1808

[Vapnik, 1995] Vapnik, V.:The Nature of Statistical Learning Theory. Springer-Verlag, 1995.

[Wasserman, 1992] Wasserman, F. Neural Computing. Theory and Practice. - M.: Mir, 1992. - 240 p.

[Zhuravlev et al, 2006] Yu.I.Zhuravlev, V.V.Ryazanov, O.V.Senko, Recognition. Mathematical methods. Software system. Practical applications. Izd.vo "Fazis", Moscow, 2006, 178 pp.

[Zhuravlev, 1978] Yu.I.Zhuravlev, On the algebraic approach to solving the problems of recognition and classification. Problems of Cybernetics. M .: Nauka, 1978. Issure.33, pp. 5-68.

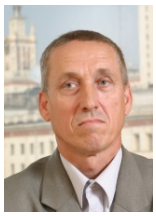
Authors' Information

Anatoliy Gupal *– Corresponding Member of the NAS of Ukraine, Heard of Department; V. M. Glushkov Institute of Cybernetics, Ukraine, 40 Acad. Glushkova Ave., 03680, Kyiv; e-mail: gupal_anatol@mail.ru*

Major Fields of Scientific Research: Bioinformatics, Bayesian network, Markov model

Maxim Novikov *– Junior Engineer at Samsung RnD Institute Russia, Russia, 127018 Moscow, Dvintsev street,12/1; e-mail: maxim.s.novikov@gmail.com*

Major Fields of Scientific Research: Pattern recognition, Data mining, Signal processing



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

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