ASSEMBLING DECISION RULE ON THE BASE OF GENERALIZED PRECEDENTS

Vladimir Ryazanov, Alexander Vinogradov, Yuryi Laptin

Abstract: *A new approach to analysis of structure of the training sample based on identification and parameterization clusters of local regularities that are considered as generalized precedents of manifestation of partial interrelations in data is investigated. Substantive treatment of non-uniformity in images of empirical distributions in parametric spaces is proposed, and possibilities of use of secondary cluster structure for reduction of complexity of decisions and increase of processing speed, identification and verification of regularities, are studied.*

Keywords: *local dependency, generalized precedent, parametric space, cluster, hyper-parallelepiped, logical regularity, hypercube bitmap, derivative distribution, decision rule*

Introduction

Higher dimensions are the main obstacle at creation analogs of the Hough Transform for abstract feature spaces which would be as effective as it takes place in case of Image Processing and Scene Analysis [de Berg, 2000], [Shapiro, 2001]. We consider here the Hough Transform in some expanded sense as the instrument of coding of statistically reliable local correlations of parameters. In this paper, some broad analogs of the Hough Transform are proposed to be used in problem of choice and optimization of the Decision Rule (DR) in feature spaces of a general kind, on the basis of detection local partial regularities in small-sized subspaces.

It is known that complexity of a problem of detection statistically significant dependences between parameters quickly increases with growth of number of the last. So, development of efficient methods of search and verification of partial dependences binding together limited subsets of parameters becomes now an important problem in the field of Data Analysis. One of the promising areas here consists in search of typical partial dependencies within a pre-selected family. Typicality here means that within the training sample a representative accumulation of objective statistics for dependencies of selected kind can be fulfilled, and estimates created of their statistical significance and their potential contribution to the DR.

The paper investigates the problem of detecting local dependencies in data that appear as features of the training sample geometry. The advantages achieved due representation of geometric forms of clusters, its spatial arrangement and filling in parametric spaces of various kinds, are shown further.

Efficient methods of coding some basic forms of clusters in the feature space R^n, in particular, in the form of a hypercube bitmap, hyper-parallelepiped of elementary logical regularity (ELR), component of Gaussian mixture, and others, are proposed in the framework of this approach [Ryazanov, 2007], [Kovshov, 2008], [Vinogradov, 2010]. Each combined object is considered as a single precedent implementing some local regularity in data - Generalized Precedent (GP) - and is represented by a point of appropriate parametric space. Important information about the presence and severity in the training sample regularities of selected kind becomes represented in the derivative distribution of GPs. Along with selection a code for regularities, a priori and indirect information of various origins can be involved to making the search of GPs more targeted. This approach based on GPs is pioneering, and results of preliminary studies showed significant benefits which arise from its use in problems of finding regularities in data, recognition and prediction [Vinogradov, 2015], [Ryazanov, 2015].

Generalized Precedents

In case of large amounts the problem of analyzing numerous data highlights the priority of processing speed. Under the new conditions simple and well-researched approaches, in particular linear, get rebirth [Berman, 2003]. The most quick are methods in which all calculations can be reduced to comparisons on special linear scales of a particular type. In this series, one of the highly successful approaches turned out to be the one based on the use of Logical Regularities (LR) [Zhuravlev, 2006], [Ryazanov, 2007]. This approach uses data clusters in the form of hyper-parallelepipeds in R^N, each cluster is described by the conjunction of the form $L = \&R_n$, $R_n=(A_n<x_n<B_n)$, and substantially interpreted as recurring joint manifestation of the feature quantities $x=(x_1,x_2,\ldots,x_N)$ on intervals $(A_n<x_n<B_n)$. The principle of proximity precedents of the same phenomenon to each other here is embodied in the requirement of filling the interior of the cluster by objects of the same class. Same time, the geometric shape of the cluster represented by the parameters A_n, B_n, becomes of particular importance. Multiple joint appearances of feature values inside this shape are regarded as substantive independent phenomenon that is called Elementary Logical Regularity (ELR). Thus we get an example of GP in the form of a parametric code $\{A_n<x_n<B_n\}$, $n=1,2,..,N$, for each hyper-parallelepiped of ELR.

A special case of ELR is hypercube of Positional Representation of Data (PRD) [Aleksandrov, 1983]. In this representation, data conversion is executed on bit layers. To perform comparisons $A_n<x_n<B_n$, linearly ordered bit scale of increased depth is used, and, instead of comparison numbers as whole, serial comparisons of descending bits are performed. The PRD in R^N is defined by a bit grid $D^N \subset R^N$ where $|D| = 2^d$ for some integer d. Each grid point $x=(x_1, x_2, ..., x_N)$ corresponds to effectively performed transformation on bit slices in D^N, when the m-th bit in binary representation $x_n \in D$ of n-th coordinate of x becomes $p(n)$-bit of binary representation of the m-th digit of 2^N-ary number that represents vector x as whole. It's supposed here $0<m\leq d$, and function $p(n)$ defines a permutation on $\{1,2, .., N\}$, $p\in S_N$. The

result is a linearly ordered scale *S* of length 2^{dN}, representing one-to-one all the points of the grid in the form of a curve that fills the space D^N densely. For chosen grid D^N an exact solution of the problem of recognition with *K* classes results in *K*-valued function *f*, defined on the scale *S*. As known, *m*-digit in 2^N-ary positional representation corresponds to *n*-dimensional cube of volume $2^{N(m-1)}$. It's called *m*-point. For each *m* the entire set of *m*-points is called *m*-slice. There are just one *d*-point, 2^N ones of *(d-1)*-points, and 2^{dN} ones of *1*-points on the scale *S*. Each of *m*-points, $0<m<d$, can be regarded as separate cluster in D^N. If it's non-empty and filled with data of certain class only, we have got GP [Vinogradov, 2015].

Recently, intensive studies develop related to the generalization of the local parameterization ideas to a wider range of partial dependencies in data. An important case is, for example, normal mixture with constant covariance matrix $\boldsymbol{\sigma}$

$$\sum_{i}^{I} \mu_i \exp\left(-\frac{1}{2}(x_i - x)^T \boldsymbol{\sigma}^{-1}(x_i - x)\right)$$

where the geometry of components as individual clusters is also very specific and described with small number of parameters. As we have seen above, in case of PRD each cluster can be coded with one integer and one real parameter (q^i, μ^i), for homogeneous normal mixture pairs of the form (x^i, μ^i) are sufficient, in the case of ELR - sets of *2N* border marks on the main axes $A^i{}_n, B^i{}_n$, as well as the weight μ^i of the ELR L^i. In many other approaches, to define the basic clusters, their spatial layout and filling, a limited number of parameters also could be enough. Below we present variants of using such parameterizations.

Examples of use GPs in practical tasks

In the first example we mean as GPs certain representations of the training data in the form of samples of new ordinary precedents that fit both the original precedents and classes, as well as results of analysis of the initial training sample. As additional information for each class $K_\lambda, \lambda = 1, 2, ..., l$, sets of LRs of classes $\mathrm{P}_\lambda = \{P_t(\mathbf{x})\}$ are used [Ryazanov, 2007], [Kovshov, 2008], ie, the predicates of the form $P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x}) = \underset{j\in\Omega_1}{\&}(c_j^1 \le x_j)\underset{j\in\Omega_2}{\&}(x_j \le c_j^2)$, $\Omega_1, \Omega_2 \subseteq \{1, 2, ..., n\}, \mathbf{c}^1, \mathbf{c}^2 \in R^n$, where

1) $\exists \mathbf{x}_t \in \tilde{K}_\lambda : P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x}_t) = 1,$

2) $\forall \mathbf{x}_t \notin \tilde{K}_\lambda : P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x}_t) = 0,$

3) $P^{\Omega_1,\mathbf{c}^1,\Omega_2,\mathbf{c}^2}(\mathbf{x})$ is locally optimal for the standard predicate quality criterion.

Here, training sample objects from the class K_λ are designated $\tilde{K}_\lambda$. We considered two schemes of definition of GPs. In the first scheme, sets of objects that satisfy the predicates of P_λ are put in correspondence to the set of precedents $\tilde{K}_\lambda$. We considered an analog of the "nearest neighbor" algorithm.

The object $\mathbf{x}$ belonged in the class, LR of which was considered the nearest, the "distance" $\mathbf{x}$ to this LR was calculated according to the formula

$$d_\alpha(\mathbf{x}) = \frac{\sum\limits_{\mathbf{x}_t : P_\alpha^{\Omega_1, \mathbf{c}^1, \Omega_2, \mathbf{c}^2}(\mathbf{x}_t)=1} \rho(\mathbf{x}, \mathbf{x}_t)}{\left|\{\mathbf{x}_t : P_\alpha^{\Omega_1, \mathbf{c}^1, \Omega_2, \mathbf{c}^2}(\mathbf{x}_t) = 1\}\right|}$$, where ρ is Euclidean metrics in R^n.

Comparison was held on data of credit scoring (2 classes, 15 features, 348 test objects) [Bache, 2013]. The accuracy of standard and modified method of "nearest neighbor" on the test data was, respectively, 75.6% and 77.5% of correct answers.

In the second scheme, we regard as GP the values of all ELR of the object and disjunction of their negations, values of all ELRs of another class and disjunction of their negations (in the classification with 3 classes and more we used the scheme "one against all"). Thus, each object corresponds to a vector of digits $\{0,1\}$. Here, GP is simply a description of the single source object in the new feature space.

Figures 1 and 2 show examples of imaging the original training sample and the test sample in the breast cancer detection task [Mangasarian, 1990]. Objects of different classes are shown in gray and black. Generalized precedents for the test training sample turned out to be linearly separable.

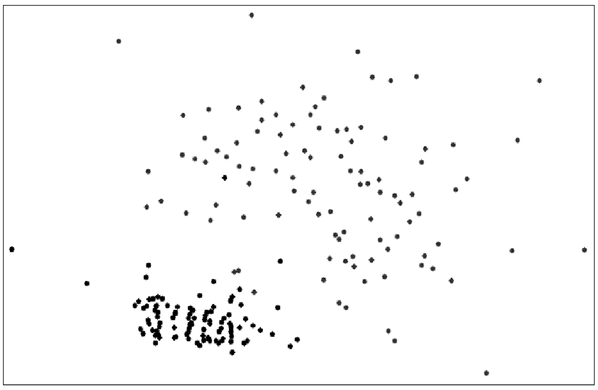

Fig.1. Visualization of the training sample

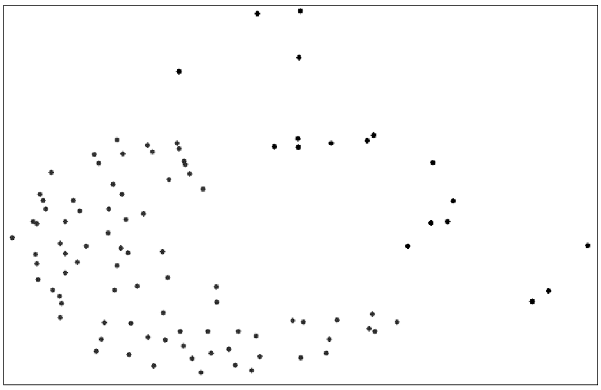

Fig.2. Visualization of the sample of GPs

Results of the comparison of this two test data pattern recognition methods in various problems are presented in the table below. Version of SVM implemented in [Zhuravlev, 2006] was used as classification method.

Task	K	N	Number of training objects	Number of control objects	Exactness on test sample	Exactness on GPs
«breast»	2	9	344	355	94.6 (0.8)	96.1
«credit»	2	15	342	348	80.5 (4.3)	84.5
«image»	7	16	210	2100	68.8 (27.7)	92.0 (0.6)

GPs and typical local dependencies

We will explore those variants of local parameterization, in which typicality of selected geometric specialty results in representativeness of corresponding cluster in the parametric space. In this approach, information about the presence of such clusters is used to optimize the DR starting from analysis of derivative distributions to realization DR in original feature space R^n.

Further we use the coding techniques for lineaments to illustrate applications of the approach in general. In Figure 1-2 it is easy to notice that in two imaged dimensions many objects are grouped in expressed elongated formations - lineaments. Partial dependence between parameters of objects corresponding to features of this type may reflect, for example, the lability of strongly correlated parameters, repeating bias in data logging, etc. In any case, objective specialties of this type must be taken into account at development of improved solutions. We present below the calculation scheme in which local geometrical features of this kind are used for reduction of data volume and simplifying the DR. The goal is to find among these local features the most typical:

a) at the first stage, we construct the set L by finding all ELRs of 2nd kind [Ryazanov, 2007] with rectangular hyper-parallelepipedal boundaries;

b) it is chosen a limited number of parameters characterizing, in the description of each obtained ELR-2, its length and position relatively to the axes. These may be size of the maximum R_l and minimum r_l edges of ELR-2 $l \in L$, or maximum length of edges R_l and guide angles α_n, etc .;

c) one-to-one mapping $f: L \rightarrow \boldsymbol{C}$ of the set L into selected parametric space $\boldsymbol{C}$ is constructed, where clustering is performed including the search for the set of expressed compact clusters $\boldsymbol{C}^T=\{\boldsymbol{c}^t\}$, $t \in T$. Each cluster $\boldsymbol{c}^t$, $t \in T$, represents separate kind of typical lineament.

Thereafter, the reverse assembly of the DR can be done on the base of detected GPs, when typical ELRs-2 are reconstructed in R^n from points c^t of each cluster $\boldsymbol{c}^t$, $t \in T$, by inversing $f^{-1}: \boldsymbol{C} \rightarrow L$. When this, ranked priorities may be given to different elements of $\boldsymbol{C}^T=\{\boldsymbol{c}^t\}$, $t \in T$, depending on the nature of the data, requirements for the solution, etc. For instance, in the first place those ELRs may be used which correspond to lineaments with the highest internal density of objects. Thus, in Fig. 3 a model example of sample with two classes is represented, where systematic error of kind "smear" appears in the data, and the observed distortions are different for two classes. In such a problem it is important to reveal exact boundaries of parameters' spread, and the preferred solution may include, in particular, the replacement of every smear lineament by point (i.e., by an average estimate of the true feature vector).

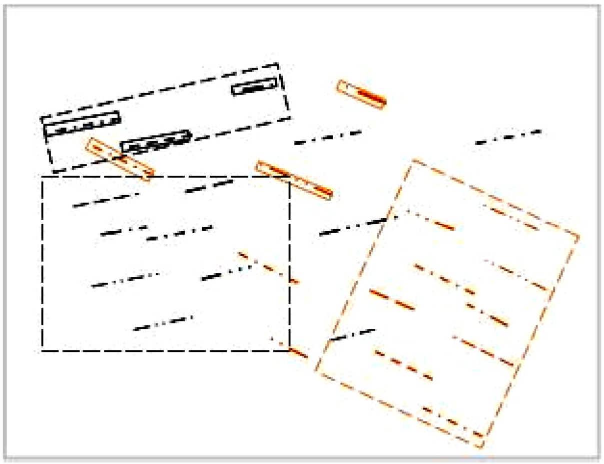

Fig.3. Modeled training sample, K=2, dotted borders show some of found ELRs-2

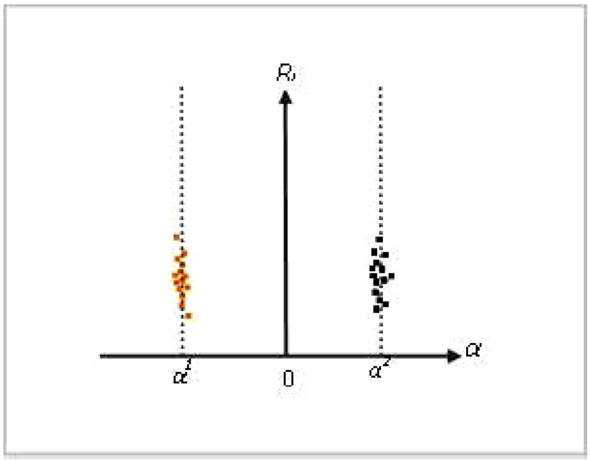

Fig.4. Two essential clusters of ELRs-2 revealed in GP space $\mathbf{C}$ for parameters α, R_l

Fig.4 shows two major compact clusters $\mathbf{c}^1,\mathbf{c}^2 \subset \mathbf{C}$ in parametric space $\mathbf{C}$ for parameters α, R_l, which represent some of constructed in step a) ELRs-2 (small rectangles densely filled by objects of the own class k only, k = 1, 2). Some of the built at step a) ELRs contain objects of different classes k, k = 1, 2, which could result in their exclusion from consideration (large rectangles in Fig.3 are not mapped into the space $\mathbf{C}$ in Fig.4). The screening is similar to that of used in the case of conventional two-dimensional Hough Transform for lines: only significant clusters in parametric space are considered, each of which correspond to elements of the same line, and all the other details of the image are ignored. Likewise, in discussed case the essential clusters $\mathbf{c}^1,\mathbf{c}^2 \subset \mathbf{C}$ unite images $c=f(l) \in \mathbf{C}$, $l \in L$, of the small densely filled ELRs-2 of Fig.3 that are selected not on the basis of intended revealing these properties, but only because of their typicality and multiple occurrence in parametric space $\mathbf{C}$. Similarly, it is the case for any version of Hough Transform, including ones for complicated shapes in higher dimensions [Shapiro, 2001]. Of course, the presence and dominance of lineaments in the empirical distribution - this is just one of many kinds of local geometric specialties which could manifest itself in the training sample. We chose the example with lineaments because of its visibility.

Reversed assembly of the Decision Rule in R^n

Notice in Fig.4 that images of classes in parametric space ***C*** become also linearly separable as that of in Fig.2 due to selection the most typical GPs only. In initial feature space R^n the same sample has more complicated geometry, and on the linear separability of classes there can be no question. We want, if possible, to keep and use these and other potential benefits when returning to the original feature space R^n and constructing DR in it. In simple cases, the reverse transition from the space of GPs ***C*** in the feature space R^n is not a problem. Point $c \in \mathbf{c}^t$ of some cluster $\mathbf{c}^t$, $t \in T$, selected as essential and typical, returns ELR-2 $l = f^{-1}(c)$ that correspond to cluster in R^n of known form (in other tasks among the latter could be as hyper-parallelepipeds of ELR-2, as hypercubes of *m*-digits in 2^N-ary PRD, Gaussian hats, etc.), which can directly participate in the formation of improved DR. For instance, in Fig.3 elements of only clusters $\mathbf{c}^1, \mathbf{c}^2$ when returning in R^n describe in detail the two classes of training sample, and other ELRs may be also used in DR to avoid, perhaps, data overfitting, superfluous computational spending, or another shortcoming.

However, it’s more difficult to maintain and to build in the DR information on those significant clusters in the parametric space, the presence of which provides additional opportunities for global optimization of the DR. We will show how this can be done in case of involvement in DR the property of linear separability of classes found in the space of GPs. We use the same example with lineaments.

Let *x* be the new object. We assume that the plane in Fig.3 represents the maximum R_l and minimum r_l edge sizes of a lineament, and the condition $R_l >> r_l$ holds. The next criterion allows to use in DR the information on the values of parameters R_l, r_l, α^1, α^2 used in the description of clusters $\mathbf{c}^1, \mathbf{c}^2 \subset \mathbf{C}$.

For triple $(R_l, , r_l, \alpha^s)$, $s=1,2$, let's construct two sets $H^s=\{h^s\}$, $s=1,2$, containing all sorts of covering the object *x* hyper-parallelepipeds of ELR-2 with parameters R_l, r_l, α^s. If the object *x* belongs to the class *s*, $s = 1,2$, then it could be found inside a virtual hyper-parallelepiped h^s the more likely, the more typical parameters R_l, r_l, α^s are, and the more representative cluster $\mathbf{c}^s$ is. Let c_i^1, c_j^2 $(i=1,2,\ldots,I,\ j=1,2,\ldots,J)$ be lists of hyper-parallelepipeds of clusters $\mathbf{c}^1, \mathbf{c}^2$, having non-empty intersection with h^s, $h^s \cap c_i^1 \neq \emptyset$. We assume that both lists *I, J* are non-empty, and denote $\rho_i^{s,1}$ the distance between all centers of the hyper-parallelepipeds h^s and c_i^1 along the edge r^s, and similarly, $\rho_j^{s,2}$ – for non-empty intersections $h^s \cap c_j^2$. Then the index of smaller averaged distance among

$$\rho^1 = \frac{1}{I}\sum_{i=1}^{I}\rho_i^{s,1}\ ,\ \rho^2 = \frac{1}{J}\sum_{j=1}^{J}\rho_j^{s,2}$$

points to the more probable class for object *x*, because when $R_l >> r_l$ holds, the intersection of rectangular lineament c_i^1 or c_j^2 $(i=1,2,\ldots,I,\ j=1,2,\ldots,J)$ with hyper-parallelepiped h^s, $s = 1$ or 2, having the same direction of maximal edge, is possible only for smaller values of distances $\rho_i^{s,1}$, $\rho_j^{s,2}$.

Of course, creation and usage of such estimates and criteria imply additional extraordinary efforts which are justified only in case of significant prospects for improvement DR. In other cases, for instance, when sets $H^s=\{h^s\}$, $s=1, 2$, are empty and estimates ρ^1, ρ^2 don’t function, conventional rules of organization of DR are used that refer to usual principles of proximity of the object x to own class in R^n.

Conclusion

Paper presents a new approach to data analysis tasks, including recognition, classification and prognosis, based on the use of concept Generalized Precedent. Approach focuses on application in cases where geometric structure of the sample has typical local details, different for different classes. It can be said that classes differ not only by its location in R^n and the density of filling, but also by detectable "texture" of this filling. As a result of construction of the space of GPs, new "textural features" that may contain significant auxiliary information become added to n basic features of R^n. Potential usefulness and targeted searches of one or another type of local geometric features of classes can be influenced by formulation of the problem, nature of data, direct or indirect demands for quality of solutions, computational cost, etc. Despite this diversity, unifying factor of the approach is the selection of suitable partial parameterization, covering sufficient variety of local dependencies. Essential compact clusters in secondary distribution on the parametric space highlight the most important forms of local dependency between basic features. It is shown that the derivative distribution may be of a complicated structure, but, adequate analysis of this structure allows revealing intrinsic regularities in data and thus achieving essential improvements of the quality of solutions in hard tasks of data analysis. Development of specialized research techniques for analysis of distribution on parametric space and methods of the use of additional information provided is doubtless the prospect of further development of the approach.

Bibliography

[Aleksandrov, 1983] Aleksandrov V.V., Gorskiy N.D. Algoritmy i programmy strukturnogo metoda obrabotki dannykh. Leningrad: Izdatelstvo ”Nauka” 1983, 208 p. (Russian).

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

[de Berg, 2000] de Berg M., van Kreveld M., Overmars M.O. Schwarzkopf O. Computational Geometry (2nd revised ed.), 2000, pp. 291–306.

[Shapiro, 2001] Shapiro L., Stockman G. Computer Vision, Prentice-Hall, Inc. 2001, 608 p.

[Berman, 2003] Berman J. Principles of Big Data, 2003, pp. 1–14.

[Zhuravlev, 2006] Zhuravlev Yu. I., Ryazanov V.V., Senko O.V. RASPOZNAVANIE. Matematicheskie metody. Programmnaya sistema. Prakticheskie primeneniya. Moscow: Izdatelstvo "FAZIS" 2006, 168 p. (Russian).

[Ryazanov, 2007] Ryazanov V.V.: Logicheskie zakonomernosti v zadachakh raspoznavaniya (parametricheskiy podkhod) // Zhurnal vychislitelnoy matematiki i matematicheskoy fiziki, T. 47/10 (2007) str.1793 -1809 (Russian).

[Kovshov, 2008] Kovshov N.V., Moiseev V.L., Ryazanov V.V.: Algoritmy poiska logicheskikh zakonomernostei v zadachakh raspoznavaniya // Zhurnal vychislitelnoy matematiki i matematicheskoy fiziki, T. 48/2 (2008) str. 329 - 344 (Russian).

[Vinogradov, 2010] Vinogradov A., LaptinYu. Usage of Positional Representation in Tasks of Revealing Logical Regularities // Rroceedings of VISIGRAPP-2010, Workshop IMTA-3, Angers, 2010, pp. 100–104.

[Bache, 2013] Bache, K., Lichman, M. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science (2013).

[Vinogradov, 2015] Vinogradov A., Laptin Yu. Using Bit Representation for Generalized Precedents // Int. Workshop OGRW-9, Koblenz, Germany, December 2014, pp.281-283.

[Ryazanov, 2015] Ryazanov V.V., Vinogradov A.P., Laptin Yu.P. Using generalized precedents for big data sample compression at learning // Machine Learning and Data Analysis, 2015, Vol. 1 (1) pp. 1910-1918.

Authors' Information

Vladimir Ryazanov – *Chief of Department, Dorodnicyn Computing Centre, Federal Research Center "Computer Science and Control" of Russian Academy of Sciences; Vavilova 40, 119333 Moscow, Russian Federation; e-mail: rvvccas@mail.ru*

Alexander Vinogradov – *Senior Researcher, Dorodnicyn Computing Centre, Federal Research Center "Computer Science and Control" of Russian Academy of Sciences; Vavilova 40, 119333 Moscow, Russian Federation; e-mail: vngrccas@mail.ru*

Yuryi 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*