On the Procedures of Generation of Numerical Features Over Partitions of Sets of Objects in the Problem of Predicting Numerical Target Variables

I. Yu. Torshin[a],* and K. V. Rudakov[a],**

[a] *Dorodnicyn Computing Centre, Informatics and Control Federal Research Center, Russian Academy of Sciences, ul. Vavilova 44, Moscow, 119333 Russia*

**e-mail: tiy1357@yandex.ru*

***e-mail: rudakov@ccas.ru*

Abstract—Analysis of criteria for the solvability/regularity of problems and of the correctness of algorithms is applied here to the problem of prediction of the values of numerical variables. It is shown that partial regularity is a necessary and sufficient condition for the solvability of the corresponding system of the classification problems. Cross-validation experiments conducted on several datasets from the field of biomedicine (non-invasive diagnostics of magnesium concentration in blood plasma), bioinformatics (prediction of the protein secondary structure), and solid-state physics (prediction of the properties of high-temperature superconductors) have demonstrated the effectiveness of the developed methods for generating "synthetic" informative numerical features and for increasing the accuracy of prediction of the numerical target variables.

Keywords: algebraic approach, regularity of problems, topologies and lattices, subquadratic algorithms, big data

DOI: 10.1134/S1054661819040175

1. INTRODUCTION

Poorly formalized problems (i.e., classification/prediction problems for which there are no unequivocal methods for defining the objects, the classes, and the feature descriptions of the objects) are widely studied in biomedicine, chemoinformatics, bioinformatics, solid-state physics, applied linguistics, and other fields of modern science [1]. A systematic approach to the study of formalization methods for such problems is fundamentally important for the improvement of the accuracy and generalizing ability of the appropriate algorithms.

In [1], the present authors demonstrated that the formalization of any poorly formalized problem can be considered as a series of successive transitions from the set of original descriptions to a particular topology, then to a lattice, and then to a certain metric space. The formalization of a problem provides a general structure to describe information on objects: *a set of initial information* (I_i) and *a set of final information* (I_f), which allows one to form *a set of precedents* (a subset of the product $I_i \times I_f$), and then to apply the constructions of the algebraic approach to the solution of recognition/classification/prediction problems [2, 3].

Within the algebraic approach developed in the scientific school of Academician Yu.I. Zhuravlev, for given sets I_i and I_f, one analyzes the properties of the set of precedents and of the algorithms $A(\theta) : I_i \to I_f$ (θ is a vector of internal parameters of an algorithm) that solve the problem. The algorithm A is often an element of an *algorithmic model* $M[\hat{\Theta}]$ ($\hat{\Theta}$ is a method for calculating the vectors θ) and is constructed as a superposition $A(\theta) = B(\theta) \circ C(\theta) \circ D(\theta)$ involving a *recognition operator* B, a *correction operation* C, and a *decision rule* D [2]. The properties of *solvability/regularity* of the problems (existence theorems of a solution) and *correctness/completeness* of the algorithmic models (which characterize the quality of the solutions) are fundamental components of the algebraic approach to recognition.

Earlier, the present authors formulated fundamental principles of the factorization [4] and of the metric [5] approaches to the analysis of poorly formalized problems and obtained metric forms of the solvability, regularity, correctness, and completeness criteria of the algebraic approach. In particular, the analysis of the compactness of metric configurations [6, 7] has allowed the authors to obtain sufficient conditions for the existence of correct algorithms. Cross-validation forms of the criteria allow one to assess not only the quality of formalization but also the extent of "overfit-

Received February 1, 2019; revised February 27, 2019; accepted July 1, 2019

ISSN 1054-6618, Pattern Recognition and Image Analysis, 2019, Vol. 29, No. 4, pp. 654–667. © Pleiades Publishing, Ltd., 2019.

tedness" of the procedures used for the selection or generation of the feature descriptions within the algorithms studied [5].

The criteria obtained in [1–7] were formulated for classification problems. In other words, the target "output" variables of these algorithms represent Boolean variables that predict the membership of an object in relation to a certain class of objects (so-called *qualitative* prediction). At the same time, in many applied problems, especially in the problems arising in natural sciences, one actually needs a *quantitative* prediction, i.e., calculation of numerical values of some "output" variables.

Solutions of the problems of predicting numerical target variables in computer science are often sought as "neural networks," (including "deep learning" networks). Note that the latter word combination is a newfangled term for the paradigm of constructing
2 multilevel perceptrons, which have been known in cybernetics since the early 1960s [8] at the very least. Within this paradigm, the input information presented in a set of precedents (first-level features) serves as a basis for calculating the values of some intermediate, so-called "hidden," variables (features of the second and the higher levels), and by the values of these "synthetic" variables one calculates the values of the target variable(s).

Despite extensive media coverage with the use of such terms as "neural nets," "deep learning," and so on, real research data indicate that the claimed increase in the "recognition accuracy" achieved with the use of these methods is accompanied by an appreciable decrease in the generalizing ability due to the most apparent "overfitting" phenomenon. In the case of classification problems, computational experiments with the algorithms of the "deep learning" type show that the error rate ν_{ob} on a "training" dataset (which is used to adjust the internal parameters θ by method $\hat{\Theta}$) is, as a rule, much lower than the error rate ν_k on an entirely independent test dataset. The difference $\Delta\nu =$
1 $\nu_k - \nu_{ob}$ is called the *overfittedness*, and $\Delta\nu > 0$ indicates that the corresponding classification algorithm is *overfit* with respect to the datasets studied.

1 In other words, positive values of overfittedness imply too strong a "fitting" of some elements of the vector θ to the specific training sample used and that is responsible for the decrease in generalizing ability. The problem of overfitting indicates the utter necessity of cross-validation testing, since a combinatorial calculation of errors within a sliding control setup of data analysis characterizes the generalizing ability of an algorithm considerably better than any "theoretical" probability of overfitting [9].

Even without a rigorous analysis, it is entirely clear that, in the case of deep-learning-type prediction systems, the addition of each new level of "hidden" features increases the number of the algorithm's parameters (i.e., $|\theta|$) thus increasing complexity of the model which only further promotes the overfitting of the algorithms thus developed. To reach a balance between the generalizing ability of algorithms and the complexity of the prediction models (the number, dimension, configuration of the levels of variables, and so on), one should operate directly with the data layers, which are by definition "hidden" in so-called "deep learning." In this case, one successively adds the variables using a cross-validation control of the generalizing ability, thus assessing the efficiency of the algorithms at each step of the "machine learning" procedure (using, in particular, the solvability/regularity and correctness/completeness functionals [4] as guidelines).

Here, we present elements of a theoretical substantiation for the procedures of generation of "synthetic" numerical features over partitions of the sets of objects in the framework of predicting numerical target variables. The formalism has been developed within the scope of topological [1] and metric [5] approaches to the formalization of the problems and the algebraic approach to recognition [2, 3]. We consider the properties of the partitions of sets of objects produced in accordance with the values of a numerical target variable. Then we present the results of appropriate experiments that demonstrate the practical applicability of the developed approaches to data mining.

2. MAIN DEFINITIONS

Let $X = \{x_1, x_2, \ldots, x_\alpha, \ldots, x_{N_0}\}$ be a set of original descriptions of objects; $X \subseteq S$, where S is the space of admissible objects; let J_{ob} be the space of admissible descriptions of objects; and let the function $D : S \to J_{ob}$ assign an object $s \in S$ its admissible description D(s). A *sampling operator* of the set X, $\hat{\zeta}$, forms a *set of samples* $\hat{\zeta}X = \{a_1, a_2, \ldots, a_k, \ldots a_{|\hat{\zeta}X|} | a_k \subset X\}$ similar to some procedures of formation of subsamples of objects in a cross-validation experiment.

Let $I_i \subseteq I_1 \times I_2 \times \ldots \times I_k \times \ldots \times I_n$ and $I_f \subseteq I_{n+1} \times I_{n+2} \times \ldots \times I_{n+l}$, where I_k are the sets of values of the k-th feature descriptions (including information on the output parameters defined by the elements of the set I_f) and $I_k = \{\lambda_{k_1}, \lambda_{k_2}, \ldots, \lambda_{k_b}, \ldots, \lambda_{k_{|I_k|-1}}, \Delta\}$ are the sets of all possible values of the k-th component of the formal description, $k = 1, \ldots, n + l$, where Δ is the uncertainty (the value of the variable is undefined). The formalization of a problem corresponds to the definition of the functions $\Gamma_k : S \to I_k$, $k = 1, \ldots, n + l$, that generate the values of the appropriate components of the corresponding formal description on the basis of the original description $x_\alpha \in X$, so that $q_i[k] = \Gamma_k(x_\alpha)$. In this case, the set $\Gamma_k^{-1}(\lambda_{k_b}) \subseteq X$ is the complete preim- 3

age of $\lambda_{k_b} \in I_k$ in the set X. Recall two well-known definitions:

Definition 1. *Suppose given arbitrary sets* A, B, C, *and* D *and functions* $f : A \to B$ *and* $g : C \to D$. *Then the Cartesian product of the functions* f *and* g *is defined as a function* $f \times g : A \times C \to B \times D$ *such that* $(f \times g)(a,c) = (f(a), g(c))$ *for any* $a \in A$ *and* $c \in C$.

Definition 2. *Suppose given sets* A *and* B *and a function* h *of m arguments*, $h : A^m \to B$. *Then a diagonalization of the function* h *is a function* $h_\Delta : A \to B$ *such that* $h_\Delta(a) = h(a,\ldots,a)$ *for any* $a \in A$.

Define a function $D : S \to J_{ob}$ as *a diagonalization of the Cartesian product of functions* Γ_k. For given $X \subseteq S$, $J_{ob} \subseteq I_1 \times I_2 \times \ldots \times I_k \times \ldots \times I_{n+1}$, and $Q \subseteq J_{ob}$, the function D is defined as $D(x_\alpha) = (\Gamma_1(x_\alpha) \times \ldots \times \Gamma_k(x_\alpha) \times \ldots \times \Gamma_{n+1}(x_\alpha))_\Delta$. The function D naturally corresponds to the function φ: $2^S \to 2^{J_{ob}}$, $\varphi(X) = \{D(x_\alpha) | x_\alpha \in X\}$.

Thus, the *formalization* of a problem corresponds to the definition of a function $\varphi: 2^S \to 2^{J_{ob}}$ for the transition from some set of original descriptions of objects in the problem domain ($X = \{x_\alpha\}$) to the *set of precedents* $Q = \{q_i | q_i = (m_i, \iota_i)\} \subseteq I_i \times I_f$, $q_i[k] \in I_k$), $i = 1 \ldots N$. Each element of the set Q represents a concatenation of the i-th rows of the corresponding *information matrix* $\hat{M}(m_i) = \begin{pmatrix} m_1 \\ \ldots \\ m_N \end{pmatrix}$, $m_i \in I_i$, and the *matrix of information* $\hat{M}(\iota_i) = \begin{pmatrix} \iota_1 \\ \ldots \\ \iota_N \end{pmatrix}$, $\iota_i \in I_f$. The definition of a function φ for a specific problem is a subject of the corresponding problem-oriented theory and allows one to form the corresponding *set of precedents* $\varphi(a)$ and *classification/prediction problem* $Z(\varphi(a))$ for an arbitrary sample $a \in \hat{\zeta} X$.

3. PROBLEMS OF PREDICTING NUMERICAL TARGET VARIABLES AND GENERATING SYNTHETIC NUMERICAL FEATURES WITHIN A LATTICE-THEORETICAL INTERPRETATION OF THE HETEROGENEOUS FEATURE DESCRIPTIONS

The methods of discrete mathematics (lattice theory, graph theory, etc.) provide a natural tool for the analysis of real-world data, which are always represented in the discrete form due to the limited number of precedents, finite number of feature descriptions, finite accuracy of the measurements, and so on. Heterogeneous feature descriptions of objects, including information on the "input" and "output" variables, belong to one of three classes: (1) Boolean features, (2) so-called "categorial" features, and (3) numerical features. This classification is not arbitrary but is based on the fundamental properties of lattices that arise during the analysis of poorly formalized problems [1].

Earlier, we have demonstrated (Theorem 3 in [1]) that, under regularity conditions for a set X, a lattice L(T(X)) formed over an appropriate topology T(X) is Boolean. This implies the uniqueness of the complement of each element of the lattice ($x \wedge \neg x = \varnothing$, $x \vee \neg x = I$, $x, \varnothing, I \in L(T(X))$, where I is the identity element of the lattice, corresponding to the set X).

For an adequate formalization of a problem, an arbitrary element of the lattice corresponds to a single value of a feature, either original or "synthetic." Therefore, for any feature with a finite number of values $I_k = \{\lambda_{k_1}, \lambda_{k_2}, \ldots, \lambda_{k_b}, \ldots, \lambda_{k_{|I_k|-1}}, \Delta\}$, there is a set of disjoint subsets $\Gamma_k^{-1}(\lambda_{k_b})$ of the set X, which map to some vertices of the Boolean lattice L(T(X)). To each such subset, there corresponds a single complement, $\neg\Gamma_k^{-1}(\lambda_{k_b}) = X \backslash \Gamma_k^{-1}(\lambda_{k_b})$.

The analysis of the mutual arrangement of these subsets of the set X in the lattice L(T(X)) points to the existence of three fundamentally different types of features. If a k-th feature is *Boolean* (i.e., $I_k = [0,1]$), all objects with this feature are related to the *lattice vertex* corresponding to the set $\Gamma_k^{-1}(1)$, while all other objects are contained in the complement set $\Gamma_k^{-1}(0) = X \backslash \Gamma_k^{-1}(1)$.

The "categorical" ("enumerative") features do not imply ordering of the values λ_{k_b} of the set I_k; therefore, they are projected onto certain *antichains of the lattice* (by definition, an antichain $a \subseteq L(T(X))$: $\forall_a x \neq y : \neg(x \subseteq y) \wedge \neg(y \subseteq x)$).

In the case of a *numerical* k-th feature, the set I_k is linearly ordered ($\lambda_{k_{b-1}} \leq \lambda_{k_b} \leq \lambda_{k_{b+1}}$). Therefore, numerical features are projected onto specific *chains*: the linearly ordered subsets of the Boolean lattice L(T(X)). By definition, a chain $c \subseteq L(T(X))$: $\forall_c x \neq y : (x \subseteq y) \vee (y \subseteq x)$, so that the chains corresponding to the numerical features consist of the sets $\Gamma_k^{-1}(\lambda_{k_1})$, $\Gamma_k^{-1}(\lambda_{k_1}) \cup \Gamma_k^{-1}(\lambda_{k_2})$, ..., $\bigcup_{\beta=1}^{b} \Gamma_k^{-1}(\lambda_{k_\beta})$, ..., I. The imaginary "motion" from the minimum element $\Gamma_k^{-1}(\lambda_{k_1})$ of the corresponding chain to the maximum element of any chain (the identity, I) corresponds to the enumeration of the values $\lambda_{k_1}, \lambda_{k_2}, \ldots, \lambda_{k_b}, \ldots$ of the numerical feature in increasing order. It is also obvious

that the quantiles of a numerical feature correspond to some subchains of fixed length.

Each value λ_{k_b} of the k-th numerical feature corresponds, on the one hand, to the subset $u(\lambda_{k_b}) = \bigcup_{\beta=1}^{b} \Gamma_k^{-1}(\lambda_{k_\beta})$ of the set of objects (which is also represented by some lattice vertex) and, on the other hand, to the partition of the corresponding chain into two subchains, the lower subchain $\left\langle \Gamma_k^{-1}(\lambda_{k_1}),\ldots,u(\lambda_{k_b})\right\rangle$ and the upper subchain $\left\langle u(\lambda_{k_{b+1}}),\ldots,I\right\rangle$.

Thus, *a numerical target variable is represented by a linearly ordered set of subsets of the set of objects (precedents)*. To each value λ_{k_b} of the k-th target variable, $k \geq n+1$, $k \leq n+l$, there corresponds a set of objects $u(\lambda_{k_b})$ and its complement $\neg u(\lambda_{k_b})$, which, respectively, define two disjoint classes of objects in the set of precedents: $\varphi(u(\lambda_{k_b})) = \{q_i | q_i[k] \leq \lambda_{k_b}\}$ and $\varphi(\neg u(\lambda_{k_b})) = \{q_i | q_i[k] > \lambda_{k_b}\}$. It is obvious that the collection of the points $\{(\lambda_{k_b}, |\{q_i | q_i[k] \leq \lambda_{k_b}\}|/N)\}$ on the plane represents an *empirical distribution function* (*EDF*) of the k-th numerical feature.

Within the theoretical framework developed, the definition of a pair of classes $C_{kb}^{+} = \varphi(u(\lambda_{k_b}))$ and $C_{kb}^{-} = \varphi(\neg u(\lambda_{k_b}))$, $C_{kb}^{+} \cup C_{kb}^{-} = Q$, $C_{kb}^{+} \cap C_{kb}^{-} = \varnothing$, allows one to define the corresponding classification problem. Suppose that some of the columns $q_i[k]$ in the set of precedents $Q = \{q_i | q_i = (m_i, \iota_i)\}$ are uncertain or "undefined" (i.e., contain the value "Δ"), for example, the range of columns $m_i[t]$, $t = t_1,\ldots,t_2$, $1 \leq t_1 \leq t_2 \leq n$, $I_t \subset R \cup \Delta$, and the range of columns $\iota_i[d]$, $d = d_1,\ldots,d_2$, $n+1 \leq d_1 \leq d_2 \leq n+l$, $I_d = [0,1,\Delta]$. Then, calculating the values in the d-th columns $\hat{M}(\iota_i[d])$ as $\iota_i[d] = (q_i \in C_{kb}^{+})$, we obtain a *system of classification problems of objects with respect to the classes of values of the k-th numerical variable* $\mathbf{Z}(Q,k) = \{Z(Q,k,\beta) = Z((m_i, \iota_i[d_0+\beta] = (q_i \in C_{k\beta}^{+})))\}$, where $n+1 \leq k$, $d_0 + \beta \leq n+l$, $\beta = 1,\ldots,|I_k| - 1$. The following theorem is obvious.

Theorem 1. *Suppose that an algorithm* $\hat{A}(k)$ *correctly solves the problem of predicting the k-th numerical variable with the desired accuracy level* ε, *i.e.*, $\left\|\hat{A}(k)\hat{M}(m_i) - \hat{M}(\iota_i[k])\right\| \leq \varepsilon$. *A correct algorithm* $\hat{A}(k)$ *exists if and only if there exists a series of correct algorithms for solving each problem of the system* $\mathbf{Z}(Q,k)$. Suppose that, for each value λ_{k_b}, there exists a correct algorithm $\hat{A}_{kb}$, $\hat{A}_{kb}\hat{M}(m_i) = \hat{M}(\iota_i[d_0+b])$, that reproduces the information matrix $\hat{M}(\iota_i[d_0+b])$ on the basis of the information matrix $\hat{M}(m_i)$. Then, each of the values λ_{k_b} represented in the k-th description of the i-th object $\iota_i[k]$ is uniquely assigned a Boolean vector, equal to the vector $(\iota_i[d+\beta])$, in which all positions corresponding to $\lambda \leq \lambda_{k_b}$ are ones and all positions corresponding to $\lambda > \lambda_{k_b}$ are zeros. In other words, for an arbitrary object with $\beta = 1,\ldots,b$, there holds $\iota_i[d_0+\beta] = 1$, $\beta = 1,\ldots,b$, while, for $\beta = b+1,\ldots,|I_k|-1$, there holds $\iota_i[d_0+\beta] = 0$; i.e., the sequence of numbers $\iota_i[d_0+\beta]$ monotonically decreases with an inflection point at $\beta = b$. Thus, the set of correct algorithms $\{\hat{A}_{kb}\}$ for solving problems from the system $\mathbf{Z}(Q,k)$ allows one to point out an exact value of the k-th numerical variable for an arbitrary object, which corresponds to the inflection point $\beta = b$. The latter is equivalent to the existence of a correct algorithm $\hat{A}$. Conversely, if there exists a correct algorithm $\hat{A}(k)$, then the reproduction of monotone Boolean vectors $(\iota_i[d_0+\beta])$ is trivial. The theorem is proved.

Corollary. *A necessary condition for the correctness of all algorithms from the set* $\{\hat{A}_{kb}\}$ *is the monotonic increase of a function defined by the set of ordered pairs* $\left\{\left(\lambda_{k_b}, \frac{1}{N}\sum_{i=1,N} \hat{A}_{kb}\hat{M}(m_i)\right)\right\}$. In the case of correct algorithms $\hat{A}_{kb}$, this function is identical to the EDF of the k-th numerical variable.

Theorem 1 shows that, on the basis of the results of calculations by classification algorithms for solving problems of the system $\mathbf{Z}(Q,k)$, each of which indicates the membership of an arbitrary object in a certain numerical interval, one can actually predict the numerical variable (or the most likely range of values of the variable). In the case of correct algorithms $\hat{A}_{kb}$, such a prediction is unique.

Note that real classification algorithms are not, as a rule, correct, especially when analyzed in cross-validation setup. Hence, in the actual computational experiments, some results of classification $\hat{A}_{k\beta}\hat{M}(m_i)[d_0+\beta]$ do not correspond to the values $\iota_i[d_0+\beta]$, so that the above-described monotonicity of the number sequence $\iota_i[d_0+\beta]$ would be disturbed. The estimates of the degree of the monotonicity disturbance of $\iota_i[d_0+\beta]$ (by combinatorial functionals, approximation by a sigmoidal function, etc.) can be used to assess the quality of the predictions of the numerical variable.

It is also important to note that the results of classification corresponding to the value λ_{k_b}, $\hat{A}_{kb}\hat{M}(m_i)[d_0+b]$, or the results of prediction of the values of the k-th variable, $\hat{A}(k)\hat{M}(m_i)$, obtained by some real algorithms $\hat{A}(k)$, $\hat{A}_{kb}$, and others, can be considered as some *"synthetic" features*, which have not been ini-

tially represented in the information matrix $\hat{M}(m_i)$. These synthetic features, both numerical and Boolean, can be placed in the t-th positions of the vectors $m_i[t]$ of the information matrix, thus forming a certain *"derived" information matrix*, $\hat{M}(m_i')$, or the *information matrix with second-level features*. Successively continuing the process of generation of the synthetic features according to the multilevel perceptron paradigm, one can obtain information matrices with the features of the third level $\hat{M}(m_i'')$, of the fourth level $\hat{M}(m_i''')$, and so on.

Thus, finding adequate solutions to the system of classification problems $\mathbf{Z}(Q,k)$ is fundamentally important for solving the problem of predicting a numerical variable and for finding some synthetic features that are informative with respect to the target variable under consideration. Within the algebraic approach, the search for such solutions starts from studying the solvability/regularity properties of the classification problems involved.

4. SOLVABILITY/REGULARITY CONDITIONS OF THE CLASSIFICATION PROBLEMS INVOLVED IN PREDICTING A NUMERICAL TARGET VARIABLE

The fundamental criteria of *solvability* and *regularity* of classification problems, *correctness* of algorithms, and *completeness* of algorithmic models, are analyzed in [4, 5] by means of the factorization and of the metric approach to the data mining. In the factorization paradigm, heterogeneous feature descriptions are reduced in one way or other to Boolean features. In the metric paradigm, one specifies methods for measuring distances between objects and features and then carries out an analysis of the metric configurations (ρ-configurations) thus obtained.

The *solvability* of a problem is defined as the consistency of the corresponding set of precedent. Formally, a problem is *solvable* if the set of algorithms (the algorithmic model) $M[\hat{\Theta}]$ is nonempty. The *regularity* of a problem is the requirement of a sort of "collective solvability" of a problem: a problem from a set of problems Z is *regular* if it is solvable and all problems from the equivalence class (neighborhood) are solvable; i.e., the regularity of a problem is sufficient for its solvability. By the *correctness* of an algorithm or an algorithmic model is meant the correspondence of the algorithm or the model to the sets of precedents. The *completeness* of an algorithmic model $M[\hat{\Theta}]$ implies that, for every regular problem, there is at least one correct algorithm in the algorithmic model $M[\hat{\Theta}]$.

Suppose we are given a set of original descriptions of objects X, a formalization method φ, sets I_i and I_f, and a space $J_{ob} \subseteq I_i \times I_f$. For any $a \subseteq X$, a set of precedents $\varphi(a) \subseteq J_{ob}$ is defined that consists of objects $q_i = (m_i, \iota_i)$, where m_i correspond to the values of n given feature descriptions $q_i[1],\ldots, q_i[k],\ldots, q_i[n]$, $q_i[k] \in I_k$ (I_k is the set of values of the k-th feature), and ι_i contains information $q_i[n+1],\ldots, q_i[n+l]$ on the membership of objects in each of the l classes. Then, for given feature selection mask $\chi = (\gamma_1,\ldots,\gamma_j,\ldots\gamma_n)$, $\gamma_j \in [0,1]$, and factorization method $\{\delta_k(v_1,v_2) : I_k^2 \to [0,1],\ k = 1\ldots n\}$ that checks the membership of the values v_1, v_2 of the k-th feature in the same equivalence class of the values of features, we formulate the solvability and regularity criteria of the corresponding problem with Boolean output variable (a classification problem $Z(\varphi(a))$) [4]:

- *solvability criterion of problem* $Z(\varphi(a))$:

$$(1)\ \bigvee_{\varphi(a)} q_1, q_2 : \iota_1 \neq \iota_2 \Rightarrow \exists_{1\ldots n}\, j : \neg\delta_k(q_1[j], q_2[j]) \wedge \gamma_j(\varphi(a));$$

- *regularity criterion of problem* $Z(\varphi(a))$:

$$(2)\ \bigvee_{\varphi(a)} q_1, q_2 : \exists_{1..n}\, j : \neg\delta_j(q_1[j], q_2[j]) \wedge \gamma_j(\varphi(a)),$$

where $\gamma_j(\varphi(a))$ denotes a method for calculating the elements of the mask χ with respect to the set of precedents φ(a).

For given algorithm $A : I_i \to I_f$ and vector of parameters $\theta \in R^{n'}$ of the algorithm, which reflects the "internal settings" of the algorithm, we formulate a *correctness criterion of the algorithm A*:

$$(3)\ A(m_i,\theta,\chi) : \bigvee_{\varphi(a)} (m_i,\iota_i) : A(m_i,\theta,\chi) = \iota_i.$$

For criteria (1)–(3), we have obtained the respective combinatorial functionals $r_1(\varphi(a),\chi)$, $r_2(\varphi(a),\chi)$, and $r_3(\varphi(a),A)$ that characterize the "degree" of satisfiability of the criteria for specific $\varphi(a)$, χ, and A.

Criteria (1)–(3) make it possible to formulate various methods for calculating feature selection masks (for example, on the basis of the dead-end property of masks with respect to the solvability/regularity criteria [4, 5]) according to some set of precedents $\varphi(a)$. The masks $\chi(\varphi(a)) = (\gamma_1(\varphi(a)),\ldots, \gamma_j(\varphi(a)), \ldots\gamma_n(\varphi(a)))$ thus obtained imply the existence of two classes of feature values: *informative* ($\gamma_j = 1$) and noninformative ($\gamma_j = 0$) values. For efficient selection of features, it is expedient to rank the features according to some functional of the "informativity" estimate (a smaller informativity rank corresponds to a greater value of the informativity estimate) [4].

The *metric forms of criteria* (1)–(3) can be obtained if the estimate of the informativity of features with respect to the class C^+ is a metric. Suppose that the

ρ-configuration $\boldsymbol{\rho}(\psi)$ describing the interactions of features from the set ψ contains a point $i(C^+)$, and let a method for selecting a neighborhood of the i-th point be given, $\bar{O}(i,r)$. Then the informative values of the features correspond to the points in the neighborhood of the point $i(C^+)$, and the feature selection mask is calculated as $\chi = (\gamma_j = (\rho[j] \in \bar{O}(i(C^+),r)))$. Substituting these operations into definitions (1)–(3), we obtain the corresponding parametric criteria whose satisfiability depends on the radius of the neighborhood. These criteria imply, in particular, that the condition $\exists_{\rho(\psi)} \rho[j],\ j \neq i(C^+):\rho(i(C^+),j) < \varepsilon,\ \varepsilon = \min \rho_{ij}$, $\varepsilon > 0$, is a constructive criterion for assessing the "quality" of the generated sets of feature descriptions [5].

The results presented in [1, 4, 5] allow us to obtain the solvability and regularity criteria for the prediction problem of the k-th numerical variable. According to the Theorem 1, the existence of a correct solution to a given problem is equivalent to the existence of correct solutions to the system of classification problems $\mathbf{Z}(Q,k)$. Since a correct algorithm is possible only under the condition of solvability of the problem [4], the *solvability of the prediction problem of the k-th numerical variable is equivalent to the solvability of each problem of the system* $\mathbf{Z}(Q,k)$:

$$(1.1)\ \bigvee_{\mathbf{Z}(\varphi(a),k)} Z(Q,k,b) \bigvee_{Q} q_1, q_2 : \iota_1[d+b] \neq \iota_2[d+b]$$
$$\Rightarrow \exists_{j=1}^{n} j : \neg\delta_j(q_1[j], q_2[j]) \wedge \gamma_j(Q),$$
$$n+1 \leq k, \quad d+b \leq n+l, \quad b = 1,\ldots,|I_k|-1.$$

For further exposition, we define a *condition of the partial regularity of the set of precedents Q with respect to the k-th numerical variable*:

$$(2.1)\ \bigvee_{b\in I_k} \bigvee_{\varphi(\Gamma_k^{-1}(\lambda_{k_b}))} q_1 \bigvee_{Q\setminus\varphi(\Gamma_k^{-1}(\lambda_{k_b}))} q_2 : \exists_{j=1}^{n} j : \neg\delta_j(q_1[j], q_2[j]) \wedge \gamma_j(Q).$$

The partial regularity condition (2.1) implies that, for a given factorization method $\delta_j()$, $j = 1\ldots n$, the feature descriptions of all objects included in the set of precedents $\varphi(\Gamma_k^{-1}(\lambda_{k_b}))$ formed for each value λ_{k_b} of the k-th numerical variable are different from the feature descriptions of all the other objects of the precedent set Q.

Theorem 2. *For the solvability condition* (1.1) *to hold, it is necessary and sufficient that the partial regularity conditions of the set of precedents Q with respect to the k-th numerical variable should hold*. In the lattice-theoretical interpretation, to each value λ_{k_b} of the k-th numerical feature, there corresponds a subset $u(\lambda_{k_b}) = \bigcup_{\beta=1}^{b} \Gamma_k^{-1}(\lambda_{k_\beta})$ that splits the corresponding chain of the Boolean lattice L(T(X)) into lower $\langle \Gamma_k^{-1}(\lambda_{k_1}),\ldots, u(\lambda_{k_b})\rangle$ and upper $\langle u(\lambda_{k_{b+1}}),\ldots, I\rangle$ subchains. Consider any two "neighbor" values of the k-th variable, $\lambda_{k_{b-1}}$ and λ_{k_b}, $b = 2,\ldots,|I_k|-1$, which correspond to the two "neighbor" problems $Z(Q, k, b-1)$ and $Z(Q,k,b)$ in the system of problems $\mathbf{Z}(Q,k)$. To satisfy the condition (1.1), each of the problems $Z(Q, k, b-1)$ and $Z(Q,k,b)$ should be solvable. Comparison of the two pairs of the lattice subchains corresponding to the values $\lambda_{k_{b-1}}$ and λ_{k_b} shows that, to the subset of objects $u(\lambda_{k_b})\setminus u(\lambda_{k_{b-1}}) = \Gamma_k^{-1}(\lambda_{k_b})$, there correspond mutually exclusive memberships of the classes for $b-1$ and b, so that $\iota[d_0 + b - 1] = \neg\iota[d_0 + b]$. Therefore, for the simultaneous solvability of any neighbor problems $Z(Q, k, b-1)$ and $Z(Q,k,b)$, the feature descriptions of elements of the subset $\varphi(\Gamma_k^{-1}(\lambda_{k_b}))$ should contain features that allow one to distinguish the elements of the set $\varphi(\Gamma_k^{-1}(\lambda_{k_b}))$ from all the other elements of the subset $\varphi(u(\lambda_{k_{b-1}}))$ and from all elements of the subset $\varphi(X\setminus u(\lambda_{k_b}))$. The fulfillment of this requirement for all $b = 1,\ldots,|I_k|-1$ corresponds to the partial regularity (2.1) in the hypothesis of the theorem. Conversely, the fulfillment of the partial regularity guarantees the solvability of each problem of $Z(Q,k,b)$, since, in each of these problems, all elements of each subset $\varphi(\Gamma_k^{-1}(\lambda_{k_b}))$ always belong to the same class of objects defined by the values $\iota[d_0 + b]$. The theorem is proved.

Corollary. *The "complete regularity" of the set of precedents Q, defined by condition* (2.1) *is sufficient for the existence of a partial regularity in the hypothesis of the theorem*. The regularity condition (2) is stronger than the condition of partial regularity (2.1), since it even guarantees the distinguishability between objects within each of the subsets $\varphi(\Gamma_k^{-1}(\lambda_{k_b}))$ of a given set of precedents.

The satisfiability and combinatorial testing of the conditions (1), (2), (1.1), and (2.1) depend on the choice of the binary functions $\delta_j()$ defining the membership of two values of the j-th feature in the same equivalence class. In general, the combinatorial testing of the conditions (1), (2), (1.1), and (2.1) is performed

in quadratic time, $o(|Q|^2)$. When $\delta_j()$ can be represented as a product of values of some unary function, $\delta_j(v_1, v_2) = \tilde{\delta}_j(v_1) \wedge \tilde{\delta}_j(v_2)$, a transition is possible from quadratic to subquadratic testing time of the conditions (1–2.1).

Theorem 3. *If each of the functions $\delta_j()$ can be represented on the set of precedents Q as a product $\delta_j(v_1, v_2) = \tilde{\delta}_j(v_1) \wedge \tilde{\delta}_j(v_2)$, $v_1, v_2 \in I_j$, then this fact is sufficient for the combinatorial testing of the conditions of solvability and complete and partial regularity in subquadratic time of about $o(|Q|\ln|Q|)$.* The possibility of calculating a value of the function $\delta_j(v_1, v_2)$ as a product of $\tilde{\delta}_j(v_1)$ and $\tilde{\delta}_j(v_2)$ corresponds to some Boolean feature generated on the basis of the tested numerical feature when calculating the conditions (1), (2), (1.1), and (2.1). If each of the functions $\delta_j()$ is factorized in a similar manner, then the information matrix of the set of precedents Q is uniquely transformed into a set of Boolean vectors. Such a set of vectors can be ordered with the use of subquadratic sorting algorithms in logarithmic time $o(|Q|\ln|Q|)$; low-dimensional Boolean vectors admit a so-called "index ordering" in a quasilinear time $o(|Q|)$. The regularity condition (2) corresponds to the absence of the repeated Boolean vectors, which is checked by a single pass of a sorted array in time $o(|Q|)$. When testing the solvability (1), (1.1) and the partial regularity (2.1) conditions, repeated vectors are allowed under the condition that each of them corresponds to a single class corresponding to the value λ_{k_b} of the k-th numerical variable, which is tested in time $o(\sum_{b=1}^{|I_k|}\left|\Gamma_k^{-1}(\lambda_{k_b})\right| \cdot \ln\left|\Gamma_k^{-1}(\lambda_{k_b})\right|) \ll o(|Q|^2)$ (since $|Q| = \sum_{b=1}^{|I_k|}\left|\Gamma_k^{-1}(\lambda_{k_b})\right|$ and, as a rule, $|I_k| \ll |Q|$). The theorem is proved.

The condition $\delta_j(v_1, v_2) = \tilde{\delta}_j(v_1) \wedge \tilde{\delta}_j(v_2)$ can easily be satisfied when the j-th feature is Boolean or categorial. In the case of numerical values of the features, this condition can be satisfied when defining the corresponding quantiles of values (in particular, those based on the data on the measurement accuracy of numerical values, EDF modes, etc.). If one uses certain statistical significance criteria for numerical features as a basis for defining $\delta_j()$, then the function $\delta_j()$ can hardly be represented as the product of some unary $\tilde{\delta}_j()$.

Note that the testing of the conditions (1), (2), (1.1), and (2.1) may involve data on the "synthetic" features, whose values are placed in the t-th positions of vectors of the information matrix $(m_i[t])$, including synthetic numerical features, synthetic Boolean features of the second level (representative sets), features of the third level (representative sets of representative sets), and so on. In any case, one should calculate the "informativity" of the original and of the synthetic features before testing the conditions (1), (2), (1.1), and (2.1).

5. THE "INFORMATIVITY" ESTIMATES OF THE ORIGINAL AND SYNTHETIC FEATURES

Within the confines of the theory of combinatorial solvability [4–7], the term "informativity" is used in the sense of a numerical estimation of the relative distribution of the values of a feature between the classes of objects, rather than in the sense of estimation of the Kolmogorov complexity [10] or in the sense of algorithmic information theory [11]. The more frequently a certain subset of a feature value is encountered in a given class of objects of the set of precedents $Q = \varphi(X)$ and lower – in all the other classes, the higher will be the informativity of this feature value in respect to this class.

The calculation of the feature selection masks $(\gamma_j(Q))$ when testing the solvability and the regularity conditions on the basis of ordering features according to an informativity estimate allows one to perform an efficient selection of the informative feature values. In particular, the iterative addition of features (arranged in decreasing order of the informativity functional) performed until the solvability/regularity criterion is satisfied allows one to find cul-de-sac (dead-end) masks $(\gamma_j(Q))$ [12–14].

Suppose that the d-th column of the matrix of information $\hat{M}(\iota_i)$ of the set of precedents Q, $I_d = [0,1]$ determines the membership of objects in a class $C^+ \subseteq Q$ with respect to which the informativity estimate $\Lambda(d, j): I_j \times I_d \to R$ of the j-th feature is calculated. Let $J_d(h)$ be a renumbering function such that $\Lambda(d,J_d(1)) \geq \Lambda(d,J_d(2)) \geq \ldots \geq \Lambda(d,J_d(h)) \geq \ldots \geq \Lambda(d,J_d(n))$, $h = 1,\ldots,n$, and there exists the inverse function, $J_d^{-1}(j)$, which calculates the *rank* of informativity of the j-th feature with respect to the class C^+. Then the elements $\gamma_j(Q)$ of the feature selection mask are calculated according to the maximum admissible informativity rank $h_{max}(\gamma_j(Q) = (J_d^{-1}(j) \leq h_{max}))$ or using the minimum admissible informativity Λ_{min} $(\gamma_j(Q) = (\Lambda(d, j) \geq \Lambda_{min}))$ or as dead-end masks [13, 14], etc. The choice of a specific method for calculating a mask $(\gamma_j(Q))$ should be made by an expert with regard to the specificities of a particular problem.

To calculate the estimates $\Lambda(d, j)$, we can apply various functionals that estimate the differences between the distribution frequencies of feature values in the classes [12], and so on. Moreover, the method of calculating the weights of the features that is used during "training" of an algorithm can be simultaneously considered as a method for estimating the infor-

mativity of the features (the value of the weight of a feature would be then the estimate of "informativity"). Such *ad hoc* estimates of the informativity of features, although allowing one to solve the corresponding technical problems during the analysis of conditions (1), (2), (1.1), and (2.1), are of purely empirical character. The lattice-theoretical approach, developed in the present series of papers, to the analysis of problem formalization allows one to obtain informativity estimates of a more fundamental character.

In [1], we demonstrated that the "interactions" between heterogeneous feature descriptions (including classes of objects) can be considered in terms of the relations between the corresponding chains, antichains, and vertices of the lattice L(T(X)). In the lattice L(T(X)), the class C^+ corresponds to the vertex $\Gamma_d^{-1}(1)$, and the class $C^- = Q\backslash C^+$, to the vertex $\Gamma_d^{-1}(0)$. Consider the relations of the class C^+ with the j-th feature.

Let the j-th feature be Boolean, $I_j = [\lambda_{j_1} = 0, \lambda_{j_2} = 1]$, so that the lattice points $\Gamma_j^{-1}(1)$ and $\Gamma_j^{-1}(0)$ correspond to this feature. All possible relations between the feature and a class are described in terms of four subsets of the set X: $\mathbf{a}(\lambda_{j_2}) = \Gamma_d^{-1}(1) \wedge \Gamma_j^{-1}(1)$, $\mathbf{b}(\lambda_{j_2}) = \Gamma_d^{-1}(0) \wedge \Gamma_j^{-1}(1)$, $\mathbf{c}(\neg\lambda_{j_2}) = \Gamma_d^{-1}(1) \wedge \Gamma_j^{-1}(0)$, and $\mathbf{d}(\neg\lambda_{j_2}) = \Gamma_d^{-1}(0) \wedge \Gamma_j^{-1}(0)$. The cardinalities of these subsets can apparently be placed in a 2×2 factor table to which one applies the exact Fisher's test. The probability of a random set of values in the factor table is estimated on the basis of the obvious combinatorial formula and corresponds to the hypergeometric distribution: $p(d,j,\lambda_{j_2}) = \binom{|\mathbf{a}|+|\mathbf{b}|}{|\mathbf{a}|}\binom{|\mathbf{c}|+|\mathbf{d}|}{|\mathbf{c}|}\Big/\binom{|\mathbf{a}|+|\mathbf{b}|+|\mathbf{c}|+|\mathbf{d}|}{|\mathbf{a}|+|\mathbf{c}|}$. The stronger the difference between the distributions of the values of the j-th feature over the classes C^+ and C^-, the smaller the value of p, so that the informativity estimate corresponds to the functionals $1 - p$, $1/p$, etc. It is well known that, in the case of sufficiently large values of $|\mathbf{a}|$, $|\mathbf{b}|$, $|\mathbf{c}|$, and $|\mathbf{d}|$, it is appropriate to apply the Pearson criterion ($\chi 2$) to speed up the computations.

When the j-th feature is either categorial or numerical, $I_j = \{\lambda_{j_1},\lambda_{j_2},\ldots,\lambda_{j_b},\ldots,\lambda_{j_{|I_j|-1}},\Delta\}$, each value of the feature corresponds to a lattice vertex $\Gamma_j^{-1}(\lambda_{j_b})$. Then the relations between the vertices $\Gamma_j^{-1}(\lambda_{j_b})$, $\Gamma_d^{-1}(1)$, and $\Gamma_d^{-1}(0)$ are estimated by the set of values $p(d,j,\lambda_{j_b})$, so that the informativity estimate corresponds to the functional of the type $1 - \sum_b p(d,j,\lambda_{j_b})$, etc.

When all features in the information matrix $\hat{M}(m_i)$ are numerical (which is important in the case of the problem of predicting a numerical target variable), the lattice-theoretical approach points to the possibility of applying another fundamental method of nonparametric statistics: exact or asymptotic forms of the Kolmogorov criterion [15, 16]. In the lattice L(T(X)), the j-th numerical feature corresponds to a chain $c_j = \langle \Gamma_j^{-1}(\lambda_{j_1}),\ldots,u(\lambda_{j_b}),\ldots,I\rangle$ in which the cardinality of the sets is described by the EDF of the j-th numerical feature, $\phi_j(\lambda_{j_b}) = |u(\lambda_{j_b})|/N$. The relations between the chain c_j and the classes C^+ and C^- are described by the set of conjunctions $u(\lambda_{j_b}) \wedge C^+$, $u(\lambda_{j_b}) \wedge C^-$, which correspond to the splitting of $\phi_j(\lambda_{j_b})$ into two EDFs $\phi_j^+(\lambda_{j_b}) = |u(\lambda_{j_b}) \wedge C^+|/|C^+|$ and $\phi_j^-(\lambda_{j_b}) = |u(\lambda_{j_b}) \wedge C^-|/|C^-|$. To estimate the differences between the EDFs ϕ_j^+ and ϕ_j^- one can use a number of statistical functionals, including the *maximum deviation* between two EDFs $D(\phi_j^+,\phi_j^-) = \max_x |\phi_j^+(x) - \phi_j^-(x)|$. The values of $D(\phi_j^+,\phi_j^-)$ allow us to assess the satisfiability of the corresponding nonparametrical criteria (for instance, the Kolmogorov–Smirnov or Kolmogorov–Bol'shev criteria). Note that characterizing the differences in the distributions of the values of the j-th feature between the classes C^+ and C^- by the values of the function $D(\phi_j^+,\phi_j^-)$ represents an informativity estimate of the j-th numerical feature with respect to the class C^+.

Thus, depending on the specific features of the information matrix $\hat{M}(m_i)$ of the given set of precedents Q, we can calculate various "informativity" estimates of the j-th feature, $j = 1,\ldots,n$, with respect to the d-th class of objects, $\Lambda_\alpha(d, j)$, $\alpha = 1,\ldots,N_\Lambda$. To each estimate $\Lambda_\alpha(d, j)$, there corresponds a special ranking n of features when calculating the masks $\chi_\alpha(Q)$ used for testing the solvability/regularity conditions. The analysis of the relationships between different estimates $\Lambda_\alpha(d, j)$ presents a separate research problem.

Within the present study, we assume that the chosen estimate $\Lambda(d, j)$ can, in a sense, be used for searching for solutions to the prediction problem of a numerical variable. The applicability criteria of a particular method for calculating the estimate $\Lambda(d, j)$ can be formulated on the basis of the properties of the function $\tilde{\Lambda}_j(\lambda_{k_b}) : I_k \to R$ defined by the set of pairs $\{(\lambda_{k_b}, \Lambda(d_0 + b, j))\}$ and of the function $\tilde{J}_j(\lambda_{k_b}) : I_k \to N$ defined by the set $\{(\lambda_{k_b}, J_{d_0+b}^{-1}(j))\}$.

When splitting the chain c_k into lower and upper subchains corresponding to the value λ_{k_b} for different λ_{k_b} of the same j-th feature, we obtained different estimates and ranks of informativity, different frequencies of feature values in the classes C^+_{kb} and C^-_{kb}, and other numerical characteristics of the features. In other words, depending on the interval of values of the k-th numerical variable, different predictors make different contributions, which can be observed as a variation in the weights of the features obtained when adjusting the corresponding algorithms. Accordingly, the functions $\tilde{\Lambda}(\lambda_{k_b})$ and $\tilde{J}(\lambda_{k_b})$ may exhibit different properties along the axis of values of the predicted numerical variable:

• *Quasi-constant* behavior: the informativity estimate of the j-th feature is independent of the interval of values of the k-th variable. It corresponds to similar distribution frequencies of the values of the j-th feature over all pairs of classes C^+_{kb} and C^-_{kb} and, hence, to low informativity of the feature.

• *Quasimonotonic* behavior: certain values of the j-th feature are encountered considerably more frequently for higher (or lower) values of the k-th numerical variable.

• *Quasi-quadratic* behavior: certain values of the j-th feature are more informative for the endpoints of the interval of values of the k-th variable (or, conversely, for the median of the interval).

• *Multimodality*: a periodic character of the relationship is possible between the predicted variable and the feature; this question can be studied by autocorrelation and other methods of signal analysis.

Each of these properties of the functions $\tilde{\Lambda}_j(\lambda_{k_b})$ and $\tilde{J}_j(\lambda_{k_b})$ can be estimated quantitatively over the system of problems $\mathbf{Z}(Q,k)$. For example, the monotonicity property can be estimated by a combinatorial functional and calculating the linear, sigmoidal, and other approximations. The experimentally established presence of the monotonicity properties in one or other form on one or other interval of the range of values points to the preservation of regularity within a given interval of values. Generally, the greater the difference of the observed behavior of the functions $\tilde{\Lambda}_j(\lambda_{k_b})$ and $\tilde{J}_j(\lambda_{k_b})$ from the quasi-constant behavior, the higher the informativity of the j-th feature with respect to partitions into classes C^+_{kb}/C^-_{kb}. The visualization of the functions $\tilde{\Lambda}_j(\lambda_{k_b})$ and $\tilde{J}_j(\lambda_{k_b})$ is an excellent instrument for the expert analysis of the feature descriptions involved in a particular problem.

Analysis of the informativity of the features is important not only for testing the solvability/regularity conditions, but also for choosing the generation procedures of synthetic features. It is quite obvious that the use of particular "synthetic" features is expedient only when their informativity in some sense is higher than the informativity of the original feature descriptions. Therefore, instead of dividing features according to levels and methods of their generation, one should combine all features, both original and "synthetic," into a single table of precedents (the t-th positions of the vectors of the information matrix, $m_i[t]$) and then analyze the informativity of the entire set of both original and synthetic features.

6. THE PROCEDURES OF GENERATING SYNTHETIC NUMERICAL FEATURES AND THE ALGORITHMS FOR PREDICTION OF NUMERICAL VARIABLES

Within the approach proposed, both the prediction of the k-th numerical variable and the generation of a numerical feature informative with respect to the k-th numerical variable are performed by algorithm $\hat{A}(k,\theta) : J_{ob} \to R$ (θ is a vector of parameters), which calculates the column of the corresponding values on the basis of the information matrix $\hat{M}(m_i)$, $I = 1, ..., N$. In the case of prediction of a numerical variable, the column $\hat{A}(k,\theta)\hat{M}(m_i)$ is the sought answer, while, in the case of generation of a synthetic numerical feature, the results of calculations are placed in the t-th "undefined" column (i.e., the column containing a value "Δ," see Section 2 of the paper) ($m_i[t]$), $t = t_1,...,t_2$, $1 \le t_1 \le t_2 \le n$, with the formation of a certain "derived" matrix $\hat{M}(m'_i)$.

The algorithm $\hat{A}(k,\theta)$ can predict the values of the k-th variable "directly" (for instance, a regression approach). Alternatively, the prediction can be based on the classification algorithms $\hat{A}_{kb}(\theta)$ that solve the problems from the system $\mathbf{Z}(Q,k)$. In the latter case, the EDF of the k-th numerical variable is formed for each point of which or for special intervals of which (percentiles, modes, etc.) the classes C^+_{kb}/C^-_{kb} are formed. On the basis of ranking with respect to the functional $\Lambda(d, j)$ and testing the solvability/regularity criteria, informative features defined by the mask $(\gamma_j(Q))$ are selected; if necessary, an expert analysis of the functions $\tilde{\Lambda}(\lambda_{k_b})$ and $\tilde{J}(\lambda_{k_b})$ is carried out for each selected feature.

When all features of the objects are Boolean (i.e., when the original feature descriptions can be factorized according to the equivalence relation $\delta_j(v_1, v_2) = \tilde{\delta}_j(v_1) \wedge \tilde{\delta}_j(v_2)$), the algorithms $\hat{A}_{kb}(\theta)$ can be represented as logical rules (disjoint normal forms, DNFs). In this case, the vector of parameters θ, $\theta \in 2^{[0,1,\Delta]^n}$ contains information on the corresponding representative sets [2]. An analysis of the informativity of individual features and of representative sets and their

combinations can be performed by means of the B-covering algorithm [17], using the methods for establishing metric condensations [7–9] or the methods of aggregation of Boolean variables within the classification theory of feature values [7]. Either of these methods allows one to perform a subquadratic search for the most efficient systems of logical rules.

Within the algebraic approach, each classification algorithm $\hat{A}_{kb}(\theta)$ is constructed as a superposition $\hat{A}_{kb}(\theta(Q)) = B_{kb}(\theta(Q)) \circ C_{kb}(\theta(Q)) \circ D_{kb}(\theta(Q))$ which includes the parameter vector θ calculated as a result of "training" the algorithm $\hat{A}_{kb}(\theta)$ on the set of precedents Q. As a rule, the correction operation $C(\theta)$ gives a certain numerical estimate of the membership of the i-th object in the corresponding class. Accordingly, both the algorithm $\hat{A}_{kb}(\theta)$ and the partial composition "recognition algorithm + corrector" $B_{kb}(\theta(Q)) \circ C_{kb}(\theta(Q))$ generate synthetic numerical features.

In the simplest case analyzed in the present study, the vector $\theta \in R^n$, $\theta = (\theta_1,\ldots,\theta_j,\ldots,\theta_n)$, contains the weights of features, and the linear operator $B_{kb}(\theta(Q))$ is defined as $(\theta(Q) \circ (\gamma_j(Q)))\ldots^T$, where "$\circ$" is the Hadamard (element-wise) product of vectors and $\ldots^T$ is the matrix transposition operation. Accordingly, the result of application of $B_{kb}(\theta(Q))$ to the i-th object m_i represents a linear form $(\theta(Q) \circ (\gamma_j(Q)))m_i^T$. As the correction operation $C_{kb}(\theta(Q)) : R \to R$, one uses a linear transformation, logarithm, exponential function, power function, and other elementary functions. The prediction algorithm $\hat{A}(k,\theta)$ may represent a regression formula or be calculated using "neural networks" or other known approaches.

The systems of equations $\hat{A}_{kb}\hat{M}(m_i) = \hat{M}(\iota_i[d_0 + b])$ and $\left\|\hat{A}(k)\hat{M}(m_i) - \hat{M}(\iota_i[k])\right\| \le \varepsilon$ appearing in Theorem 1 provide a basis for estimating the values of vectors of feature weights $\theta \in R^n$. As a rule, any of these systems of equations is either underdetermined or overdetermined and therefore does not have a unique solution. Therefore, one can apply various methods to determine the values of (θ_j): methods of computational linear algebra (singular decomposition, etc.), neural networks, stochastic approximation [18], and so on. In particular, stochastic approximation implemented in the form of an iterative multistart procedure allows one to develop subquadratic algorithms for calculating (θ_j) that terminate the calculations upon reaching a given convergence criterion or, conversely, a specified divergence criterion.

The methods of generation of synthetic features mentioned here can be used for reducing problems with heterogeneous numerical features to a problem with numerical features. Under operators $\hat{A}(k,\theta)$ and others, any Boolean feature can be replaced by or complemented with synthetic numerical features.

7. CROSS-VALIDATION ESTIMATES OF SOLVABILITY, REGULARITY, AND CORRECTNESS

The informativity indices of the j-th feature $(\Lambda(d, j), J_d^{-1}(j), \tilde{\Lambda}_j(\lambda_{k_b}), \tilde{J}_j(\lambda_{k_b}))$ and the masks $(\gamma_j(Q))$ are calculated on the basis of a given set of precedents Q. For a given sampling operator $\hat{\zeta}$ of the set X of original descriptions of objects, one obtains different sets of precedents$Q = \varphi(a)$, $a \in \hat{\zeta}X$, each of which corresponds to different values of the above-mentioned informativity indices. As a result, to each j-th feature, original or synthetic, there corresponds a set of estimates $\{\Lambda(d, j)|Q = \varphi(a), a \in \hat{\zeta}X\}$ rather than a single number $\Lambda(d, j)$, a set of functions rather than a single functions $\tilde{\Lambda}_j(\lambda_{k_b})$, and so on. These differences will affect the selection of features when testing the solvability, regularity, and correctness conditions.

In [4, 5], for a given sampling operator $\hat{\zeta}$, we obtained the following cross-validation forms of criteria that include testing over the set of samples $\hat{\zeta}X$:

- *solvability criterion of problems* $Z(\varphi(a))$, $a \in \hat{\zeta}X$:

$$(1.2)\ \underset{\hat{\zeta}X}{\forall} a,b,a \neq b \underset{\varphi(a)}{\forall} q_1,q_2{:}\iota_1 \neq \iota_2 \Rightarrow \underset{1..n}{\exists} j{:}\neg\delta_j(q_1[j],q_2[j]) \wedge \gamma_j(\varphi(b)) = 1;$$

- *regularity criterion of problems* $Z(\varphi(a))$, $a \in \hat{\zeta}X$:

$$(2.2)\ \underset{\hat{\zeta}X}{\forall} a,b,a \neq b \underset{\varphi(a)}{\forall} q_1,q_2{:}\underset{1..n}{\exists} j{:}\neg\delta_j(q_1[j],q_2[j]) \wedge \gamma_j(\varphi(b)) = 1;$$

- *correctness criterion of the algorithm* A_h:

$$(3.2)\ \underset{\hat{\zeta}X}{\forall} a \neq b \underset{\varphi(a)}{\forall} (m_i,\iota_i) : A_h(m_i(\varphi(a)),\hat{\Theta}\varphi(b),\chi(\varphi(b))) = \iota_i.$$

For criteria (1.2), (2.2), and (3.2), we have obtained appropriate combinatorial functionals. For example, the functional $r_{1c}(\hat{\zeta}X) = \frac{1}{Y(Y-1)}\sum_{i=1}^{Y}\sum_{j=1,j\neq i}^{Y} r_1(\varphi(a_i), \chi_1(\varphi(a_j)))$, $Y = |\hat{\zeta}X|$, corresponds to testing (1.1) on all "training" – "testing" pairs of sample datasets, while the functional $r_{1l}(\hat{\zeta}X) = \frac{1}{Y}\sum_{i=1}^{Y} r_1(\varphi(a_i), \chi_1(\varphi(a_i)))$ estimates the results of testing on a single, training, sample over all samples $\hat{\zeta}X$. The difference $r_{1l}(\hat{\zeta}X) - r_{1c}(\hat{\zeta}X)$ describes some "overfittedness" related to the algorithm for calculating a mask $\chi_1()$, i.e., to the feature selection procedure on the basis of the solvability criterion. Functionals for cross-validation estimates of regularity $r_{2c}(\hat{\zeta}X)$ and $r_{2l}(\hat{\zeta}X)$ and correctness $r_{3c}(\hat{\zeta}X)$ and $r_{3l}(\hat{\zeta}X)$ are obtained in a similar manner.

In addition to these functionals, it is practically convenient to use similar cross-validation functionals based on the known functionals of descriptive statistics. When predicting the values of a numerical variable in regression analysis, one can use correlation coefficient and the residual. By analogy with the combinatorial functionals $r_{1l}(\hat{\zeta}X)$ and $r_{1c}(\hat{\zeta}X)$, one can obtain *cross-validation functionals for the correlation coefficients* $r_{c,l}(\hat{\zeta}X)$, $r_{c,c}(\hat{\zeta}X)$, and others.

8. EXPERIMENTAL TESTING OF THE METHODS

The developed approaches to the synthesis of informative numerical features have been practically tested on a unique sample of biomedical data on patients (n = 400) containing information on 140 diagnoses of ICD-10, blood test results, cardiointervalographic (heart rate variability, HRV) examination, completion of clinical questionnaires, etc. (552 indicators of patients state in total) [19]. The problem was raised of quantitative prediction of the magnesium concentration in blood plasma on the basis of noninvasive examination of a patient (HRV data, clinical symptomatics, medical history). From the viewpoint of the problem area, this problem is quite hard, since it includes the widely spaced levels of the biological systems [20]: magnesium ion concentrations in blood (the lowest level of the structural organization of a biological system, the atomic level) and clinical indicators of the patient's state (a high level of a biological system, the organism level).

The efficiency of solving the problem was estimated in cross-validation experiments including ten random partitions of the dataset into "training–test" samples with the size ratio 1 : 1. The original dataset was made regular (Theorem 2 with corollary). As the informativity estimate $\Lambda(d, j)$, we used functionals based on the exact Fisher's test; the masks $(\gamma_j(Q))$ were calculated in accordance with (2.1) by iterative testing of the partial regularity criterion [12–14] with the use of the EDF modes as the factorization method δ_j (this corresponds to the hypothesis in the Theorem 3). To generate synthetic numerical features, we used a composition $B_{kb}(\theta(Q)) \circ C_{kb}(\theta(Q))$. For linear recognition operators $B_{kb}(\theta(Q))$, we used linear transformation, logarithm, the exponential function, and the power function as the correction operator C_{kb}. In the case of the logical rule method, we used Boolean values of the algorithms $\hat{A}_{kb}$ as synthetic features (see Theorem 1). The results of the experiments are summarized in Table 1.

The SVD proved to be the worst alternative for generating the synthetic features, especially when SVD was also used to predict MgPK ($r_{c,l}(\hat{\zeta}X) = 0.92$ and $r_{c,c}(\hat{\zeta}X) = 0.19$ – apparently, a prominent overfitting due to the fact that the system of equations $\hat{A}\hat{M}(m_i) = \hat{M}(\iota_i[k])$ is apparently overdetermined). A similar situation was observed in the case when we used neural networks for generating "synthetic" numerical features. At the same time, iterative procedures of stochastic approximation, although in practice they did not always reach a verifiable convergence, demonstrated much better results ($r_{c,l}(\hat{\zeta}X) = 0.45$ and $r_{c,c}(\hat{\zeta}X) = 0.45$), increasing the accuracy of numerical prediction to a practically acceptable level (the standard deviation of the MgPK variable values in the test dataset was 0.16 mmol/l, while the actual measurement accuracy of the concentrations was about 0.05 mmol/l).

It is interesting that the method of logical rules allows one to generate synthetic Boolean features (i.e., the results of calculation by the algorithms $\hat{A}_{kb}$, as in Theorem 1) that are useful for predicting MgPK. The use of a relatively small number (typically, 5 to 8) of informative Boolean variables for establishing the corresponding representative sets to obtain the matrix $\hat{M}(m_i')$ and a stochastic approximation or SVD to predict MgPK by the matrix $\hat{M}(m_i')$ also allowed us to obtain an acceptable quality of predictions ($r_{c,l}(\hat{\zeta}X) = 0.52$ and $r_{c,c}(\hat{\zeta}X) = 0.40$).

On the whole, the best result of prediction was obtained when using logical rules and stochastic approximation for generating features ($r_{c,l}(\hat{\zeta}X) = 0.53$ and $r_{c,c}(\hat{\zeta}X) = 0.49$, Fig. 1; standard deviation of MgPK concentrations in the test of 0.10 mmol/l). The addition of further levels (i.e., the generation of the "derived" matrices $\hat{M}(m_i'')$ and $\hat{M}(m_i''')$, etc.) did not increase the prediction accuracy.

Table 1. Cross-validation estimates of the effectiveness of various approaches to the generation of synthetic numerical features and predicting a numerical variable in the problem of quantitative prediction of magnesium concentration in blood plasma (the variable MgPK). The values of the functionals $r_{c,l}(\hat{\zeta}X)$ and $r_{c,c}(\hat{\zeta}X)$ are given only for the best model (b.m.) of generation of synthetic features in the corresponding series of experiments. C_{kb} is the correcting operation in the model, m1 is linear transformation, m4 is logarithm, m5 is exponential function, and m6 is square root. To calculate the vector θ of weights, we used singular vector decomposition (SVD), a three-level artificial neural network (ANN), or a multistart stochastic approximation (MSA) procedure. For generating synthetic Boolean features, we used a method of logical rules (LR). "None" indicates that only the original matrix $\hat{M}(m_i)$ was used.

Generation of synthetic features (calculation of $\hat{M}(m_i')$)	Prediction of the variable MgPK	$r_{c,l}(\hat{\zeta}X)$	$r_{c,c}(\hat{\zeta}X)$	b.m.
None	SVD	0.92	0.18	m6
None	ANN	0.21	0.12	m1
None	MSA	**0.42**	**0.40**	m1
SVD	SVD	0.92	0.19	m5
SVD	ANN	0.25	0.07	m1
SVD	MSA	0.51	0.27	m1
ANN	SVD	0.82	0.16	m1
ANN	ANN	0.29	0.18	m1
ANN	MSA	0.44	0.30	m1
MSA	SVD	0.82	0.27	m1
MSA	ANN	0.22	0.25	m1
MSA	MSA	**0.45**	**0.45**	m5
LR	SVD	0.54	0.39	m5
LR	ANN	0.27	0.27	m1
LR	MSA	**0.52**	**0.40**	m1
MSA, LR	MSA	**0.53**	**0.49**	m6

Analysis of the synthetic features obtained by the methods of logical rules and stochastic approximation, in addition to increasing the quality of predictions, has allowed us to obtain results important for a deeper understanding of the problem area. The analysis of the profiles of $\Lambda(d, j)$ and $J_d^{-1}(j)$ for informative features has shown that, for predicting very low values of MgPK (<0.4 mmol/l, extremely strong deficiency of magnesium), the most informative indicators were HRV spectral indicators %HF, LF, and others; the scoring assessment of the patient's health state; and the dynamometry data. To predict a pronounced deficiency of magnesium (MgPK < 0.7 mmol/l), the most important features were the periodicity and the variability parameters of HRV (RRNN, SDNN, etc.), the scoring assessment of the patient's health state, and the indicators from the diet questionnaire. In the case of a moderate magnesium deficiency (MgPK < 0.82 mmol/l), the most informative indicators were the dynamometry and a few specific HRV indices. Finally, in the case of magnesium sufficiency (MgPK > 0.85 mmol/l), the most informative indicators were the periodicity/variability of HRV and the diet questionnaire. The relationships established are in agreement with the complex physiological effects of magnesium ions known from the literature [21].

In conclusion, we would like to emphasize that the optimal combination of specific methods of generation of synthetic features and of prediction of a target numerical variable essentially depends on the particular problem under study and the datasets involved. For example, if we restrict the number of input variables in the above-described problem of predicting MgPK to the HRV data (24 indicators in total), then the best solutions (albeit of poor quality, $r_{c,l}(\hat{\zeta}X) = 0.31$ and $r_{c,c}(\hat{\zeta}X) = 0.16$) will be obtained using SVD for generating features and stochastic approximation for predicting the MgPK variable.

The application of the described approaches to the problem of predicting secondary protein structure [14] has shown that the best solutions ($r_{c,c}(\hat{\zeta}X) = 0.87$) are obtained when using logical rules and stochastic approximation for generating features and neural networks for predicting the values of variables.

Analysis of a sample of crystal structures of high-temperature cuprate superconductors (HgBa2CuO4+d [22], etc.) has shown that the opti-

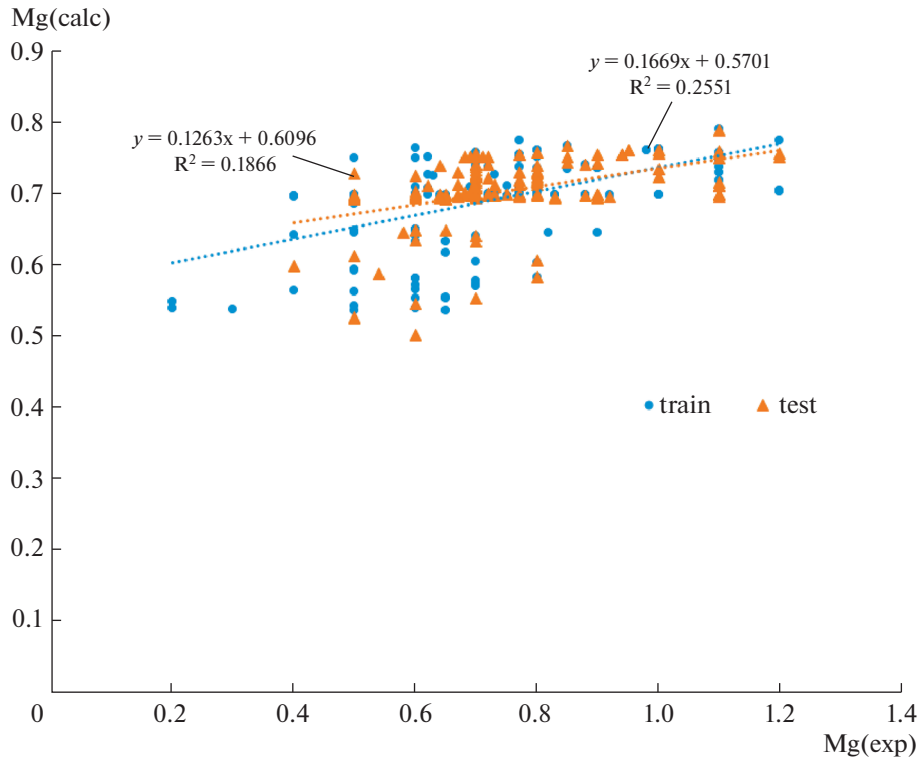

Fig. 1. (Color online) Example of the correlation between experimentally determined and predicted levels of magnesium in blood plasma (mmol/l) when using logical rules and stochastic approximation for generating numerical features and stochastic approximation for predicting.

mal solution of the problem is obtained when one uses logical rules for generating the "synthetic" features and stochastic approximation for predicting the values of the numerical variable (the critical temperature of a superconductor T_c, $r_{c,c}(\hat{\zeta}X) = 0.77$).

9. CONCLUSIONS

In the case of poorly formalized problems, there exist (infinitely) many methods for generating features and, accordingly, for the feature descriptions of one and the same problem. The solvability and regularity criteria of the classification problems involved (especially the partial regularity criterion) allowed the selection of the most informative features for the procedures of generation of synthetic features, which were subsequently used for constructing recognition/classification or the numerical prediction algorithms. The augmentation of the set of original features with synthetic features has allowed us to increase the quality of predictions of numerical variables in the datasets tested. The subquadratic algorithms, proposed in this paper for data mining and for predicting the output numerical variables, are easily scaled when implemented on multiprocessor systems and thus are quite useful for the analysis of so-called "big data" in different areas.

ACKNOWLEDGMENTS

We are grateful to Prof. O.A. Gromova for useful discussions on the expert analysis of biomedical data.

FUNDING

This work was supported by the Russian Foundation for Basic Research, project nos. 19-07-00356, 18-07-01022, 17-07-01419, 16-07-01129, and 18-07-00944.

CONFLICT OF INTEREST

We declare that we have no conflict of interests related to the preparation and publication of this article.

REFERENCES

1. I. Yu. Torshin and K. V. Rudakov, "On the theoretical basis of metric analysis of poorly formalized problems of recognition and classification," Pattern Recogn. Image Anal. **25** (4), 577–587 (2015).
2. Yu. I. Zhuravlev, "Correct algebras over sets of incorrect (heuristic) algorithms. I," Cybern. **13** (4), 489–497 (1977).
3. K. V. Rudakov, "On some universal constraints for classification algorithms", USSR Comput. Math. Math. Phys. **26** (6), 75–81 (1986).

4. I. Yu. Torshin and K. V. Rudakov, "Combinatorial analysis of the solvability properties of the problems of recognition and completeness of algorithmic models. Part 1: Factorization approach," Pattern Recogn. Image Anal. **27** (1), 16–28 (2017).
5. I. Yu. Torshin and K. V. Rudakov, "Combinatorial analysis of the solvability properties of the problems of recognition and completeness of algorithmic models. Part 2: Metric approach within the framework of the theory of classification of feature values," Pattern Recogn. Image Anal. **27** (2), 184–199 (2017).
6. I. Yu. Torshin and K. V. Rudakov, "On metric spaces arising during formalization of recognition and classification problems. Part 1: Properties of compactness," Pattern Recogn. Image Anal. **26** (2), 274–284 (2016).
7. I. Yu. Torshin and K. V. Rudakov, "On metric spaces arising during formalization of problems of recognition and classification. Part 2: Density properties," Pattern Recogn. Image Anal. **26** (3), 483–496 (2016).
8. A. G. Ivakhnenko and V. G. Lapa, *Cybernetic Predictive Devices* (Naukova Dumka, Kiev, 1965) [in Russian]. ISBN 978-5-458-61159-6
9. K. V. Vorontsov, *Combinatorial Theory of Reliability of Learning by Precedents*, Doctoral Dissertation in Mathematics and Physics (Dorodnicyn Computing Centre, Russian Academy of Sciences, Moscow, 2010).
10. A. N. Kolmogorov, "Combinatorial foundations of information theory and the calculus of probabilities," Russ. Math. Surv. **38** (4), 29–40 (1983).
11. R. J. Solomonoff, "A formal theory of inductive inference. Part I," Inf. Control **7** (1), 1–22 (1964). https://doi.org/10.1016/S0019-9958(64)90223-2
12. I. Yu. Torshin, "On solvability, regularity, and locality of the problem of genome annotation," Pattern Recogn. Image Anal. **20** (3), 386–395 (2010).
13. I. Yu. Torshin, "The study of the solvability of the genome annotation problem on sets of elementary motifs," Pattern Recogn. Image Anal. **21** (4), 652–662 (2011).
14. K. V. Rudakov and I. Yu. Torshin, "The motif information analysis based on the solvability criterion for the protein secondary structure recognition," Inform. Primen. (Inf. Appl.) **6** (1), 79–90 (2012) [in Russian].
15. N. L. Bol'shev and N. V. Smirnov, *Mathematical Statistics Tables* (Nauka, Moscow, 1983) [in Russian].
16. A. N. Kolmogoroff, "Sulla determinazione empirica di una legge di distribuzione," Giorn. Ist. Ital. Attuari **4** (1), 83–91 (1933).
17. I. Yu. Torshin, "Optimal dictionaries of the final information on the basis of the solvability criterion and their applications in bioinformatics," Pattern Recogn. Image Anal. **23** (2), 319–327 (2013).
18. M. B. Nevel'son and R. Z. Has'minskii, *Stochastic Approximation and Recursive Estimation*, Translations of Math. Monographs, Vol. 47 (American Mathematical Society, Providence, RI, 1973; Nauka, Moscow, 1972).
19. E. Yu. Egorova, I. Yu. Torshin, O. A. Gromova, A. I. Martynov, "The use of cardiointervalography for diagnostic screening and evaluation of the efficiency of correction of magnesium deficiency and comorbid conditions," Terapevticheskiy Arkhiv (Ther. Arch.) **87** (8), 16–28 (2015) [in Russian].
20. I. Yu. Torshin, *Sensing The Change*: *From Molecular Genetics To Personalized Medicine*, in *Bioinformatics in the Post-Genomic Era Series* (Nova Science Publ., New York, 2009). ISBN 1-60692-217-0
21. O. A. Gromova and I. Yu. Torshin, *Magnesium and the "diseases of civilization"* (GEOTAR-Media, Moscow, 2018) [in Russian]. ISBN 978-5-9704-4527-3
22. I. Yu. Torshin, V. A. Aleshin, and E. V. Antipov, "Synthesis and properties of the high-temperature superconductor HgBa2CuO4+d," Sverkhprovodimost': Fizika, Khimiya, Tekhnika (Supercond.: Phys., Chem., Technol.) **7** (10–12), 1579–1587 (1994) [in Russian].

Translated by I. Nikitin

Ivan Yur'evich Torshin. Born 1972. Graduated from the Department of Chemistry, Moscow State University, in 1995. Received candidates degrees in chemistry in 1997 and in physics and mathematics in 2011. Currently is a senior researcher at Dorodnicyn Computing Centre, an associate professor at Moscow Institute of Physics and Technology, lecturer at the Faculty of Computational Mathematics and Cybernetics, Moscow State University, leading scientist at the Russian Branch of the Trace Elements Institute for UNESCO, and a member of the Center of Forecasting and Recognition. Author of 450 publications in peer-reviewed journals in biology, chemistry, medicine, and informatics and of 9 monographs: 5 in Russian and 4 in English (the series "Bioinformatics in Post-genomic Era", Nova Biomedical Publishers, NY, 2006-2009).

Konstantin Vladimirovich Rudakov. Born 1954. Russian mathematician, corresponding member of the Russian Academy of Sciences, Head of the Department of Computational Methods of Forecasting at the Dorodnicyn Computing Centre, Informatics and Control Federal Research Center, Russian Academy of Sciences, and Head of the Intelligent Systems Chair at the Moscow Institute of Physics and Technology.

SPELL: 1. overfittedness, 2. perceptrons, 3. preimage, 4. cardiointervalographic