Combinatorial Analysis of the Solvability Properties of the Problems of Recognition and Completeness of Algorithmic Models. Part 1: Factorization Approach

I. Y. Torshin[a,*], K.V. Rudakov[a,b,]**

[a] *Moscow Institute of Physics and Technology, Institutskii per. 9, Dolgoprudny, Moscow oblast, 141700 Russia.*
[b] *Dorodnicyn Computing Centre, Federal Research Center "Informatics and Control," Russian Academy of Sciences, ul. Vavilova 40, Moscow, 119333 Russia*
*e-mails: *tiy1357@yandex.ru, **rudakov@ccas.ru*

Abstract—In poorly formalized problems of recognition and classification, there are plenty of methods to generate feature descriptions of objects, which are subsequently studied by the methods of computer science. Using the factorization approach (reduction of a recognition/classification problem to a binary form), here we obtain combinatorial criteria for the solvability and regularity properties of the problems and criteria of correctness of algorithms and completeness of algorithmic models (the analysis of which is an important component of the algebraic approach to the synthesis of correct algorithms). The study presents a hierarchy of the criteria and cross-validation methods for assessing their feasibility. Based on the hierarchy of criteria, we formulate a general approach to the analysis of poorly formalized problems.

Keywords: algebraic approach, metric analysis, the theory of classification of feature values, compact metric spaces, clustering, combinatorial theory of solvability

DOI: 10.1134/S1054661817010151

1. INRODUCTION

Within the algebraic approach to the solution of recognition and/or classification problems, a *set of initial information* (I_i) and a *set of final information* (I_f) are defined, and an *algorithm* solving a problem is sought as a mapping $A(\theta) : I_i \to I_f$ [1–4], where θ stands for a vector of "internal parameters" or "settings" of the algorithm. The range of values of θ depends on the specific method of construction of the algorithm. The algorithm A is a function and is constructed as a superposition $A(\theta) = B(\theta) \circ C(\theta) \circ D(\theta)$, which includes a recognition operator B, a correcting operation C, and a decision rule D.

By a *parametric algorithmic model* is meant a set of parametric mappings for which a method $\hat{\Theta}$ of calculating the vectors of parameters θ_h is defined that is used in the model, so that $M^*[\hat{\Theta}] = \{(A_h(\theta_h), \theta_h), A_h(\theta_h) = B_h(\theta_h) \circ C_h(\theta_h) \circ D_h(\theta_h) | A_h(\theta_h) : I_i \to I_f\}$. The elements of $M^*[\hat{\Theta}]$ can be used for solving a whole class of problems, rather than a single problem. The problems of this class are defined by universal constraints I_s ("*structural information*", using terminology of Zhuravlev's school) which distinguish a family of admissible algorithms $M[\hat{\Theta}, I_s]$ from the general model $M^*[\hat{\Theta}]$) and **local constraints** (*sets of precedents*, which are the subsets of $I_i \times I_f$) [5–9].

In various fields of modern natural science, there are *poorly formalized* problems, for which it is impossible to uniquely define the sets I_i and I_f [10–12]. This peculiarity of the poorly formalized problems substantially complicates the application of the constructions of the algebraic approach to recognition. Accordingly, the necessity arises to develop an appropriate formalism, which, first, would allow one to find "optimal", in a sense, definitions of I_i and I_f, and, second, would be consistent with the methods of analysis of the fundamental properties of problems and of algorithmic models [3, 4]: the *solvability* and the *regularity* of recognition problems, the *correctness* of the (models of) algorithms, and the *completeness* of the algorithmic models.

The property of *solvability* of problems is defined as the consistency of universal and local constraints. A problem is *solvable* if the set $M[\hat{\Theta}, I_s]$ is nonempty, i.e., if at least some of the mappings $A : I_i \to I_f$ satisfy both universal and local constraints simultaneously. In the general case, algorithms from $M[\hat{\Theta}, I_s]$ may impose certain constraints on I_i (the choice of one or another subset of "informative" features, assignment of weights to features, and so on). Under some constraints on I_i, the solvability of this problem cannot be guaranteed when analyzing the correspond-

Received September 10, 2016

ing set of precedents. The exact formal definition of the solvability criterion of a problem depends on the specific method of representation of initial and final information on the objects and the corresponding universal constraints I_s.

The *regularity* of a problem is the requirement of a kind of "collective solvability" of a subset of related problems. Suppose given a partition of a set Z of problems under test (for which the universal constraints I_s are satisfied) into equivalence classes (or a "neighborhood of a problem" is defined [1–4]). A problem from the set Z is *regular* if it is solvable and all the problems from the equivalence class (neighborhood) are solvable as well. The regularity of a problem is sufficient for its solvability. An exact definition of the regularity criterion depends on the specific method of definition of the neighborhood of a problem.

The *correctness* of an algorithm or an algorithmic model implies the compliance of the algorithm or the algorithmic model with a local constraint (i.e., a set of precedents). An algorithmic model $M[\hat{\Theta}, I_s]$ is *complete* if, for any regular problem in $M[\hat{\Theta}, I_s]$, there exists a correct algorithm. In other words, a model possessing the property of completeness provides the solution of all regular problems (for a given system of universal constraints I_s) and thus is essentially nonimprovable in the class of algorithmic models under test [13]. Exact definitions of the correctness and completeness criteria are obtained for specific systems of universal constraints.

In the scientific school of Yu.I. Zhuravlev and K.V. Rudakov, academicians of the Russian Academy of Sciences, an in-depth theoretical analysis has been carried out of the properties of solvability/regularity, correctness, and completeness, which demonstrated the universality of the constructions of the algebraic approach to solving recognition/classification problems. A combinatorial analysis of the solvability and regularity criteria is of significant practical importance and allows one to increase the efficiency of the search for efficient algorithms for solving a number of problems [14–16].

For the practical analysis of specific recognition/classification problems, it seems important to draw parallels between the above-described concepts, which were developed within algebraic approach, and some empirical terminology widely used in the field of computer science. Naturally, the concepts from computer science are not strictly equivalent to the above-considered concepts of the algebraic approach. Nevertheless, it can be said that the (Zhuravlev's) correctness of an algorithm corresponds to the 100% *accuracy of the algorithm* on a learning sample. The (Zhuravlev's) completeness of an algorithmic model is comparable to the maximum *generalizing ability* of the algorithms of this model under testing on various samples of precedents.

The concepts of "accuracy", "generalizing ability", and the closely related concept of "overfitting" and "overfittedness" are important for the practical analysis of specific statements of problems and algorithms used for their solution. When carrying out such an analysis within the framework of, say, statistical theory of machine leaning [17], the main problem consists in obtaining estimates for the error probability of the algorithm on test objects. Computational experiments show that, as a rule, the error frequency υ_{ob} on a learning sample is much lower than the error frequency υ_k on the test sample. The difference $\Delta\upsilon = \upsilon_k - \upsilon_{ob}$ is called *overfittedness*, and $\Delta\upsilon > 0$ points to the existence of *overfitting* of the parameters of the algorithm under study. In other words, overfitting corresponds to too stringent "fitting" of some "internal parameters" of the algorithm to a specific leaning sample, which reduces the generalizing ability.

The problem of estimating the probability of overfitting and the generalizing ability has not yet been fully solved. In practice numerous functionals, each of which formalizes some definition of the generalizing ability, are applied. The adequacy of the theoretical estimates of the generalizing ability essentially depends on the original axiomatic, in particular, on the method of formalization of the concepts of generalizing ability and overfitting [18]. An important result of [18] is an illustration of the fact that a combinatorial calculation of errors in a cross-validation setup (such as "sliding control", for instance) characterizes the generalizing ability of the algorithm much better than any of the known "theoretical" probabilities of overfitting.

Thus, on one hand, in the algebraic theory of recognition, there is a theoretical concept of completeness of an algorithmic model that characterizes some fundamental "versatility" of this model (for an arbitrary set of precedents, there is a correct algorithm in the model). On the other hand, there are practically well-established empirical methods for estimating the "generalizing ability," which are based on intuitively clear concept of "cross-validation" and which allow one to estimate the overfitting of algorithms with the use of some estimation functionals.

At the same time, the analysis of poorly formalized problems faces an obvious difficulty that makes the direct application of both the constructions of algebraic approach to the synthesis of correct algorithms and of the combinatorial methods for estimating overfitting/generalizing ability of these constructions almost impossible. In the case of poorly formalized problems, there are (infinitely) many methods for generating features, feature values and, accordingly, there are many feature descriptions of the same problem (defined as a certain "initial" description of a sample of objects with membership in classes). In this case, one needs criteria that would allow one to distinguish "adequate" feature descriptions, to construct correct

algorithms for such descriptions, and then to estimate their generalizing ability. The widely known theoretical constructions do not allow one to evaluate the contribution of "latent" sources of overfitting such as the procedures of generation and selection of features and the like.

Poorly formalized problems can be analyzed by two complementary approaches – by *factorization of feature descriptions* (reduction of a problem to the binary form) and by *metrization* (introduction of a metric on the sets of features and on the sets of feature descriptions of objects, and analysis of the compactness and density of the corresponding subspaces of metric spaces). In this paper, within the factorization approach we obtain combinatorial criteria for the above-considered fundamental properties of problems, algorithms, and families (models) of algorithms, and propose cross-validation methods for estimating the satisfiability of these criteria.

2. FACTORIZATION APPROACH TO THE ANALYSIS OF POORLY FORMALIZED PROBLEMS

In the case of poorly formalized problems, the feature descriptions can be generated by various methods. Generally, these feature descriptions of objects are heterogeneous and include binary (i.e., boolean), numerical, and so-called "categorial" features. In contrast to "categorial" features, binary and other numerical ones imply mandatory linear ordering of the values of features.

By a *binary factorization* (binarization), we mean a transformation of some "initial" descriptions of objects as a result of which the feature description of any object includes only binary features.

First, we define the solvability and regularity criteria for the sets of binary features and then generalize them to the case of heterogeneous feature descriptions. The criteria obtained should admit experimental verification on various samples of objects, according to the cross-validation ideology.

Let $X = \{x_1, x_2, \ldots, x_i, \ldots, x_{N_0}\}$ be the set of initial descriptions of objects. Suppose that the *sampling operator* of the set X, $\hat{\zeta}$, forms a collection of subsets X, $\hat{\zeta}X = \{a_1, a_2, \ldots, a_k, \ldots a_{|\hat{\zeta}X|} | a_k \subset X\}$, which we will also call as the *collection of samples*.

A specific method of realization of $\hat{\zeta}$ represents, apparently, one of conditions of the actual computational experiment: this may be a partition of X into p equal parts, $\hat{\zeta}_p X = \{a_1, \ldots, a_p \subset X | \forall k_1 \neq k_2 : a_{k_1} \cap a_{k_2} = \varnothing, |a_{k_1}| = |a_{k_2}| = \lfloor N_0/p \rfloor\}$; formation of p randomized samples with returns, each of which contains a fixed fraction s of objects X, $\hat{\zeta}_p(s)X = \{a_1, \ldots, a_p \subset X | \forall k : |a_k| = \lfloor s \cdot N_0 \rfloor\}$, etc. In short, the operator $\hat{\zeta}$ forms a collection of samples $\hat{\zeta}X$ of the set X similar to the sampling procedures applied during cross-validation.

Formalization of a problem corresponds to the definition of a function $\phi : X \to Q$ for the transition from some set of initial descriptions of objects in the problem domain ($X = \{x_i\}$) to a *set of precedents* ($Q = \{q_i | q_i = (m_i, \iota_i)\}$), each of whose elements represents a pair of ith rows of some matrix of information $\{m_i\}$ and information matrix $\{\iota_i\}$. The construction of the function ϕ for a specific problem is the subject of the appropriate problem-oriented theory [13].

For the statements of problems with n binary features and l classes of objects, the elements of the sets $\{m_i\}$ and $\{\iota_i\}$ belong to the corresponding subspaces of the Boolean cube B^{n+l} ($m_i \in [0,1]^n$, $\iota_i \in [0,1]^l$). Then, a *formalized statement of problem* Z(Pr) with a set of precedents *Pr* implies finding a method of calculation of $\{\iota_i(\mathrm{Pr})\}$ from the values of $\{m_i(\mathrm{Pr})\}$. The solutions of such a problem are given by *algorithms* $A_h : [0,1]^n \to [0,1]^l$ from some parametric *algorithmic model* $M_A[\hat{\Theta}, I_s] = \{A_h\}$. Local constraints of the algorithms of the model are the sets of precedents $\{\phi(a), \ \forall a \in \hat{\zeta}X\}$. The universal constraint I_s is *symmetric* [13], since the order of elements in the set X and the sets $\phi(a)$ is arbitrary.

3. DEFINITION OF SOLVABILITY AND REGULARITY CRITERIA WITHIN THE FACTORIZATION APPROACH

Now, it becomes possible to formulate combinatorial definitions of the criteria of solvability, regularity, correctness, and completeness. Under a symmetric universal constraint, by the solvability of problem $Z(\phi(a))$, $a \in \hat{\zeta}X$, is meant the consistency of the set of precedents $\phi(a)$. The algorithms $A \in M_A$ are functions; hence, the *solvability criterion* of problem $Z(\phi(a))$ is defined as the corresponding existence condition of a function:

$$\underset{\phi(a)}{\forall} (m_1, \iota_1), (m_2, \iota_2) : m_1 = m_2 \Rightarrow \iota_1 = \iota_2. \qquad (1)$$

The algorithms of the model M_A may impose constraints on the subset of "informative" features that are used by the algorithms for the calculations. Let $\chi \in B^n$ be a *mask* that distinguishes the features used in an arbitrary object $q_i = (m_i, \iota_i)$. Then, the *solvability criterion* of $Z(\phi(a))$ on the *subset of features* χ is obtained by a simple transformation of (1):

$$\underset{\phi(a)}{\forall} q_1, q_2 : m_1 \wedge \chi = m_2 \wedge \chi \Rightarrow \iota_1 = \iota_2. \qquad (1.1)$$

Suppose given a mask $\chi = (\gamma_1,\ldots,\gamma_k,\ldots\gamma_n)$ and objects $q_1 = (m_1, \iota_1)$ and $q_2 = (m_2, \iota_2)$, such that $m_1 = (\alpha_1^1,\ldots,\alpha_k^1,\ldots,\alpha_n^1)$, $m_2 = (\alpha_1^2,\ldots,\alpha_k^2,\ldots,\alpha_n^2)$, and $\gamma_k, \alpha_k^1, \alpha_k^2 \in [0,1]$. The sum of digits of the binary number corresponding to the mask χ is called the *trace of the mask*, $tr(\chi) = \sum_{k=1}^{n} \gamma_k$ and represents the number of selected features. The k-th binary feature is said to be *distinguishing* for the objects q_1 and q_2 if these objects belong to different classes and $\alpha_k^1 \neq \alpha_k^2$, while $\gamma_k = 1$.

Theorem 1. *A problem is solvable over the set of precedents if and only if, for every pair of objects belonging to different classes, there is at least one distinguishing feature.*

Proof. Let us write assertion (1.1) in the inverse form, applying the above substitutions for m_1, m_2, and χ, and obtain (1.2):

$$\begin{aligned} &\forall_{\phi(a)}\, q_1, q_2 : \iota_1 \neq \iota_2 \\ &\Rightarrow (\alpha_1^1 \wedge \gamma_1,\ldots,\alpha_k^1 \wedge \gamma_k,\ldots,\alpha_n^1 \wedge \gamma_n) \\ &\neq (\alpha_1^2 \wedge \gamma_1,\ldots,\alpha_k^2 \wedge \gamma_k,\ldots,\alpha_n^2 \wedge \gamma_n) \end{aligned} \tag{1.2}$$

Two binary numbers are unequal if and only if they differ in the value of one or several bits; therefore, we rewrite (1.2) as

$$\forall_{\phi(a)}\, q_1, q_2 : \iota_1 \neq \iota_2 \Rightarrow \exists_{1..n}\, k : \alpha_k^1 \wedge \gamma_k \neq \alpha_k^2 \wedge \gamma_k \tag{1.3}$$

The right-hand side of condition (1.3) can be satisfied only for $\gamma_k = 1$. Thus, the solvability implies the existence of a distinguishing feature, which proves the sufficiency of the assertion of the theorem. The necessity follows from the inversion of the above sequence of transformations, so that (1.3) implies assertion (1.1) and, for $\chi = 2^n - 1$, assertion (1). The theorem is proved.

Corollary 1. *Let* $r_1(\phi(a),\chi) = \dfrac{2}{N(N-1)} \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} (\iota_i \neq \iota_j \Rightarrow \exists_{1..n}\, k : \alpha_k^i \wedge \gamma_k \neq \alpha_k^j \wedge \gamma_k)$. Then the solvability of $Z(\phi(a))$ is equivalent to $r_1(\phi(a),\chi) = 1$.

Problem $Z(\phi(a))$ is *regular* if it is (1) solvable and (2) all the problems in the neighborhood of problem $Z(\phi(a))$ are solvable. Since the neighborhood of a problem can be defined in various ways, we consider the extreme case presented in Theorem 2.

Theorem 2. *Suppose that the neighborhood of problem $Z(\phi(a))$ is defined as all the problems whose information matrices are identical or are subsets of the information matrix of the problem $Z(\phi(a))$, while the matrices of information may take an arbitrary value. Then the problem is regular over the set of precedents $\phi(a)$ if, for any pair of objects, there is at least one distinguishing feature.*

Proof. The regularity criterion of problem $Z(\phi(a))$ can be obtained from condition (1.3). The information matrix of an arbitrary problem from the neighborhood considered consists of the rows $\{m_i, i = 1..|\phi(a)|\}$ of the information matrix $Z(\phi(a))$, so that, for arbitrary matrices of information, any two m_1 and m_2 in the same problem of the neighborhood may correspond to $\iota_1 \neq \iota_2$, while in another, to $\iota_1 = \iota_2$. Therefore, to guarantee the satisfiability of the solvability condition for an arbitrary problem from the neighborhood of the problem $Z(\phi(a))$, the right-hand side of condition (1.3) should be satisfied irrespective of the fulfillment of the left-hand side; i.e., condition (2) should be satisfied for any pairs of objects from $\phi(a)$:

$$\forall_{\phi(a)}\, q_1, q_2 : \exists_{1..n}\, k : \alpha_k^1 \wedge \gamma_k \neq \alpha_k^2 \wedge \gamma_k. \tag{2}$$

Condition (2) guarantees the solvability of an arbitrary problem from the neighborhood considered and, hence, is a *criterion of regularity* of $Z(\phi(a))$ *on the subset of features* χ. The theorem is proved.

Corollary 1. Let $r_2(\phi(a),\chi) = \dfrac{2}{N(N-1)} \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} (\exists_{1..n}\, k : \alpha_k^i \wedge \gamma_k \neq \alpha_k^j \wedge \gamma_k)$. Then the regularity of $Z(\phi(a))$ is equivalent to $r_2(\phi(a),\chi) = 1$.

Corollary 2. Condition (2) can be tested in subquadratic time. "Direct" testing of the satisfiability of (2) for a given mask $\chi = (\gamma_1,\ldots,\gamma_k,\ldots\gamma_n)$ is performed in $O(N^2)$, since it requires the analysis of $\frac{1}{2} N \cdot (N-1)$ pairs of objects, $N = |\phi(a)|$. At the same time, the N binary numbers $\{m_i \wedge \chi\}$ can be ordered (for example, in increasing order) by an effective sorting algorithm in $O(N \cdot \ln N)$. When condition (2) is satisfied, all ordered numbers are pairwise different—the fact that can be checked in time $O(N)$.

The selection procedures of features have a significant impact on the accuracy and specificity of recognition/classification algorithms. Therefore, the methods of calculating the mask $\chi \in B^n$ on the basis of the solvability and regularity criteria are of significant interest.

These calculations become practically feasible when (1) one finds dead-end forms of masks and (2) defines a certain "sensible" linear order of features within the rows of the information matrix and, accordingly, within the sought mask. The analysis of the dead-end property within the combinatorial theory of solvability [14–16] uses the ideology similar to that of the theory of dead-end tests [19]. By *dead-end* masks (for example, with respect to the solvability criterion (1.3)) are meant those masks in which the "zeroing" of

any nonzero γ_k leads to the loss of satisfiability of this criterion.

We will assume that the features in the rows of the information matrix $\{m_i\}$ and, accordingly, in the mask $\chi = (\gamma_1,\ldots,\gamma_k,\ldots\gamma_n)$ are already ordered in accordance with some method of determination of "informativity" of the features so that "more informative" features correspond to smaller values of the *rank of informativity, i.e., the values of* k [14–16]. Then the choice of the features appearing in the dead-end masks consists in finding a distinguishing feature with maximum informativity for each pair of objects from $\phi(a)$, $a \in \hat{\xi}X$.

Theorem 3. *A mask $\chi_1 = (\gamma_1^1,\ldots,\gamma_k^1,\ldots\gamma_n^1)$ is dead-end with respect to the solvability criterion if and only if, for every nonzero k-th position of the mask there is at least one pair of objects in $\phi(a)$, $a \in \hat{\xi}X$, for which the k-th feature is the only distinguishing feature among the features with nonzero values of γ_k^1* in χ_1.

Proof. Suppose that the features in the mask χ_1 are ordered according to some method of calculation of the rank of "informativity" of the features. Define a function $K(i,j)$ that finds a feature with the minimum position number k (i.e., with the maximum "informativity") that allows one to distinguish between the ith and jth objects, i.e., $K(i,j) = \min k : \alpha_k^i \neq \alpha_k^j$. Let $\phi(a)$ be consistent, i.e., criterion (1.3) be satisfied for $\chi = 2^n - 1$. Then, one can calculate a mask $\chi_1 \leq 2^n - 1$ that also guarantees the solvability: $\gamma_k^1 = (\underset{\phi(a)}{\exists}(i,j) : K(i,j) = k)$. This procedure finds a single distinguishing feature for every pair of objects, and if, for some k', $\gamma_{k'}^1 = 1$, the objects are distinguishable, then new features with values $k > k'$ are not added to the mask χ_1.

Suppose that the k'-th feature, $\gamma_{k'}^1 = 1$, found in $N_{k'}$ objects by zeroing $\gamma_{k'}^1$ is removed from χ_1. By construction, an arbitrary k'th feature is added to the mask only if there is no feature with smaller value of k. Conversely, features with values of k greater than k' are added only if the features with $k \leq k'$ are not distinguishing for a certain pair of objects. One or another distinguishing feature is unique for a pair of objects (i,j) only when $\sum_{k=1,n}(\alpha_k^i \oplus \alpha_k^j) = 1$. If $\sum_{k=1,n}(\alpha_k^i \oplus \alpha_k^j) > 1$ for any $N_{k'}$ objects for any $j \neq i$, then the zeroing of $\gamma_{k'}^1$ does not lead to the loss of solvability. When $\sum_{k=1,n}(\alpha_k^i \oplus \alpha_k^j) = 1$ at least for one, ith, object from among $N_{k'}$ objects for a certain value of $j \neq i$, then the k'-th feature is unique for this pair of objects and the elimination of this feature inevitably leads to the loss of solvability. In this case, χ_1 will be dead-end under the hypotheses of the theorem. The necessity is proved by contradiction. The theorem is proved.

Theorem 4. *A mask χ_2 is dead-end with respect to the regularity criterion only if, for every nonzero position of χ_2, there is a pair of objects in the set of precedents for which the k-th feature is the only distinguishing one.*

The proof is analogous to the proof of Theorem 3.

Corollary 1. $tr(\chi_1) \leq tr(\chi_2)$. The left-hand side of condition (1.3), $\iota_1 \neq \iota_2$, excludes the analysis of some pairs of objects that are analyzed during the analysis of condition (2). Accordingly, when calculating a dead-end χ_2, all the positions that are nonzero in the dead-end χ_1 and, possibly, additional positions necessary for distinguishing additional pairs of objects will be nonzero in this mask. In other words, the distinguishing of arbitrary pairs of objects in accordance with the condition (2) requires, in general, a greater number of features compared with the distinguishing of objects of different classes by (1.3).

It follows from Theorems 3 and 4 that the calculation of masks χ_1 and χ_2 requires the enumeration of all pairs of objects, i.e., is performed in time $O(N^2)$, where N is the number of objects. Actually, all poorly formalized problems considered in the present series of papers belong to the field of "BigData" (since they include from tens of millions to billions of objects each of which is described by the values of millions of features) and this requires the involvement of supercomputer technologies for calculating χ_1 and χ_2. When the number of features n is not much greater than the number of objects N and the occurrence frequencies of the features are low (a few percents of the value of N), it becomes possible to significantly reduce the computation time.

Theorem 5. *Suppose that the matrix of information in a consistent set of precedents $\phi(a)$ consists of a single column, which corresponds to a problem with two classes, C^+ and C^-, where $N^+ = |C^+|$, $N^- = |C^-|$, $N = N^+ + N^-$, and $N^+ \approx N^-$. Suppose that the occurrence frequency of an arbitrary kth feature in both classes is low ($n_k^+ \ll N^+$ and $n_k^- \ll N^-$) and the number of features n is not greater than the number of objects, $n \approx N$. Then a dead-end mask χ_1 can be found in time much less than $O(N^+ \cdot N^-)$.*

Proof. Consider a complete bipartite graph $G = \mathrm{K}_{N^+,N^-}$ in which the left part corresponds to the objects of class C^+ and the right part, to the objects of class C^-. To every pair of objects (i,j) tested when calculating χ_1 and χ_0, there corresponds an (i,j)-edge in the graph G. In the "original" graph K_{N^+,N^-}, the k-th

feature corresponds to a subset $\vartheta_k^+ = \{q_{k,1}^+, q_{k,2}^+, \ldots\}$ of n_k^+ vertices of the left part, a subset $\vartheta_k^- = \{q_{k,1}^-, q_{k,2}^-, \ldots\}$ of n_k^- vertices of the right part, and the corresponding subset of edges $\vartheta_k^+ \times \vartheta_k^-$. For a consistent set of precedents $\phi(a)$, the family of alls sets ϑ_k^+ and ϑ_k^- covers all $N = N^+ + N^-$ objects, and the family of all sets $\vartheta_k^+ \times \vartheta_k^-$ covers all the edges of K_{N^+,N^-}.

After carrying out a test for every pair of objects (i, j), we will remove the (i, j)-edge from the graph$G$ so that, after testing all $(N^+ \cdot N^-)$ pairs of objects, G becomes a null graph. Let us carry out a test of the solvability criterion on the set of precedents $\vartheta_k^+ \cup \vartheta_k^-$. When carrying out a test on the set of precedents $\vartheta_k^+ \cup \vartheta_k^-$, it is necessary to calculate K(i,j) for $n_k^+ \cdot n_k^-$ pairs of objects.

Note that the kth feature is not necessarily distinguishing on the set of precedents $\vartheta_k^+ \cup \vartheta_k^-$; the role of distinguishing features is played by other ones. At the same time, if the kth feature has already been chosen as distinguishing (i.e., $\gamma_k^1 = 1$), it certainly distinguishes all objects in $\vartheta_k^+ \cup \vartheta_k^-$ from all the other objects of the set of precedents $\phi(a)$, which implies the analysis of $N_k^{+-} = n_k^+ \cdot (N^- - n_k^-) + n_k^- \cdot (N^+ - n_k^+)$ pairs of objects. In other words, when some other, m-th, feature is chosen as distinguishing when testing solvability on $\vartheta_k^+ \cup \vartheta_k^-$, then this subsequently eliminates the need to analyze N_k^{+-} pairs of objects. Therefore, upon carrying out the analysis of solvability for all sets $\vartheta_k^+ \cup \vartheta_k^-$ by $\sum_{k=1}^{n} n_k^+ \cdot n_k^-$, G becomes a null graph (i.e., a regular graph of degree zero, $r(G) = 0$).

Thus, in the case of a "direct" testing of objects in $\vartheta_k^+ \cup \vartheta_k^-$ with all the other objects of the set of precedents $\phi(a)$, one should compare $N_k^{+-} + n_k^+ \cdot n_k^-$ pairs of objects, while, when testing the set of precedents $\vartheta_k^+ \cup \vartheta_k^-$ of the distinguishing kth feature – only $n_k^+ \cdot n_k^-$ pairs of objects. If, after testing $n_d \le n$ features, the equality $r(G) = 0$ is satisfied at that moment then the testing procedure stops. Since $n_k^+ \ll N^+$, $n_k^- \ll N^-$ and $n \approx N$ by the hypothesis of the theorem, here we obtain that $\sum_{k=1}^{n_d} n_k^+ \cdot n_k^- \ll N^+ \cdot N^-$. The theorem is proved.

Indeed, suppose that the hypothesis of Theorem 5 is satisfied and $n_k^+ < \nu_{\max} \cdot N^+$, $n_k^- < \nu_{\max} \cdot N^-$, and $\nu_{\max} < 0.05$. This is a typical case, for example, for problems of bioinformatics [14, 16], where the occurrence frequencies of binary features ("motives in amino acid sequences") do not exceed a few percents. Then, $\sum_{k=1}^{n_d} n_k^+ \cdot n_k^- < n_d \cdot \nu_{\max}^2 \cdot N^+ \cdot N^- \ll N^+ \cdot N^-$. Therefore, under the condition $n_d \ll 1/\nu_{\max}^2$, the procedure described will be certainly computationally more effective than the "direct" testing of the conditions (1.3, 2) over the set of precedents $\phi(a)$.

Thus, Theorems 3-5 allow one to calculate dead-end masks that include features with maximum informativity and guarantee the criteria of solvability (1.3) and regularity (2) of the corresponding problems.

Note that the conditions (1.3) and (2) were obtained for binary features. It is important to note that the solvability and regularity criteria obtained within the factorization approach can be generalized to the case of an arbitrary collection of features (not only binary, but also numerical and "categorial"). We will distinguish between the above-mentioned binary factorization and *partial factorization*; in the latter procedure, for each feature, equivalence classes of the values of features are introduced: intervals of values of numerical features, subsets of values of "categorial" features, and so on.

Let I_k be a set of values of the kth feature. A partial factorization consists in introducing binary functions $\delta_k(v_1, v_2) : I_k^2 \to [0,1]$ for the kth feature, which test the membership of two values v_1, v_2 of the kth feature in the same equivalence class of features ($\delta_k(v_1, v_2) = 1$ if v_1, v_2 belong to the same equivalence class). Specific methods of determining $\delta_k(v_1, v_2)$ are restricted by the framework of the relevant task-oriented theory.

Under the condition of partial factorization, it becomes possible to generalize the definitions of the criteria of solvability (1.3) and regularity (2), thus obtaining the *solvability* (1.4) *and regularity* (2.1) *criteria for heterogeneous feature descriptions* $m_i = (\varphi_1^i, \ldots, \varphi_k^i, \ldots, \varphi_n^i)$, $\varphi_k^i \in I_k$:

$$\forall_{\phi(a)} q_1, q_2 : \iota_1 \ne \iota_2 \Rightarrow \exists_{1..n} k : \neg\delta_k(\varphi_k^1, \varphi_k^2) \wedge \gamma_k = 1, \quad (1.4)$$

$$\forall_{\phi(a)} q_1, q_2 : \exists_{1..n} k : \neg\delta_k(\varphi_k^1, \varphi_k^2) \wedge \gamma_k = 1. \quad (2.1)$$

A method for calculating the values of the corresponding functionals $r_1(\phi(a), \chi)$ and $r_2(\phi(a), \chi)$ for problems with heterogeneous feature descriptions follows from the conditions (1.4), (2.1) and from the above-mentioned definitions of the functionals.

4. CORRECTNESS AND COMPLETENESS CRITERIA FOR ALGORITHMIC MODELS WITHIN THE FACTORIZATION APPROACH

In itself, the criteria of correctness of algorithms and completeness of an algorithmic model formulated below do not require binary or partial factorization. However, the criterion of the completeness of an algorithmic model necessarily implies that the sets of precedents considered are regular. Therefore, it can be said that the criteria formulated further are obtained within the "factorization approach."

Let $I_i \subseteq I_1 \times I_2 \times \ldots \times I_k \times \ldots \times I_n$, where I_k is a set of values of the k-th features. The correctness of the algorithm $A_h(\theta): I_i \to I_f$ or the algorithmic model $M_A[\hat{\Theta}, I_s] = \{A_h(\theta)\}$ implies that the algorithm and/or the model strictly corresponds to the local constraints (i.e., to some set of precedents $\phi(a)$).

Suppose that the algorithm A_h is a function of several arguments, namely, the initial information $m_i \in I_i$, the *parameters of the algorithm* $\theta \in R^{n_p}$, which represent the "internal settings" of the algorithm, and a *mask* $\chi = (\gamma_1, \ldots, \gamma_k, \ldots \gamma_n)$ which describes the features selected. Note that, in a number of cases, the parameters χ and θ can be interrelated. For example, if the vector θ reflects the "weights of features," then $\theta[k] = 0$ corresponds, effectively, to $\gamma_k = 0$.

Within the framework of the formalism developed, the universal constraint I_s is symmetric, so that the order of elements of the set $\phi(a)$ is arbitrary [13]. Then the *criteria of the correctness of the algorithm* $A_h(m_i, \theta, \chi)$ *and the correctness of the algorithmic model* $M_A[\hat{\Theta}]$ are defined as follows:

$$\forall_{\phi(a)} (m_i, \iota_i) : A_h(m_i, \theta, \chi) = \iota_i, \quad (3)$$

$$\forall_{M_A[\hat{\Theta}]} A_h(\theta_h, \chi_h) : (\forall_{\phi(a)} (m_i, \iota_i) : A_h(m_i, \theta_h, \chi_h) = \iota_i). \quad (3.1)$$

Based on the criterion (3), we can define a combinatorial functional $r_3(\phi(a), A_h)$, which characterizes the "degree" of satisfiability of the criterion for specific $\phi(a)$, A_h, χ, and θ: $r_3(\phi(a), A_h(\chi, \theta)) = \frac{1}{N} \sum_{i=1}^{N} (A_h(m_i, \theta, \chi) = \iota_i)$, so that the correctness of the algorithm $A_h(\chi, \theta)$ over the set of precedents $\phi(a)$ corresponds to $r_3(\phi(a), A_h(\chi, \theta)) = 1$. Similarly, criterion (3.1) corresponds to $r_{31}(\phi(a), M_A[\hat{\Theta}]) = \frac{1}{H} \sum_{h=1}^{H} r_3(\phi(a), A_h(m_i, \theta_h, \chi_h)) = H = |M_A[\hat{\Theta}]|$, and the correctness of the algorithmic model, to $r_{31}(\phi(a), M_A[\hat{\Theta}]) = 1$.

Definitions (3) and (3.1) obviously imply

Theorem 6. *The solvability of problem* $Z(\phi(a))$ *is a necessary condition for the correctness of any algorithm and any algorithmic model. It is clear that all* A_h *in* $M_A = \{A_h\}$ *are functions.*

Corollary 1. It follows from $r_3(\phi(a), \chi) = 1$ that $r_1(\phi(a), \chi) = 1$.

Instead of the correctness of the algorithmic model (3.1) a more general concept of *completeness of an algorithmic model* was developed in the algebraic recognition theory [1–6, 10–13]. A model $M_A[\hat{\Theta}]$ is complete if, for any regular problem, the model $M_A[\hat{\Theta}]$ contains a correct algorithm [13]. In other words, a model possessing the property of completeness provides at least one solution to the regular problems analyzed.

Within the formalism developed, the regularity of problems is a natural property for the analysis of poorly formalized problems. For instance, the first step in the formalization of a problem is the choice of the form of the *axiom of correspondence,* i.e., the choice of fundamental constraints imposed on the "formalizing" mapping ϕ. The strong form of the axiom of correspondence is closely related to the regularity of the sets of precedents analyzed [10].

Let us take a set of initial description of objects X, a sampling operator $\hat{\zeta}_r$ and a formalization method ϕ, such that all samples in $\hat{\zeta}_r X$ correspond to regular problems for a unit mask, i.e., $\forall_{\hat{\zeta}_r X} a : r_2(\phi(a), 2^n - 1) = 1$. Then the criterion of *completeness of the algorithmic model* $M_A[\hat{\Theta}]$ on the set of regular samples $\hat{\zeta}_r X$ is obtained from (3.1) by the substitutions of samples from $\hat{\zeta}_r X$:

$$\forall_{\hat{\zeta}_r X} a \exists_{M_A[\hat{\Theta}]} A_h(\theta_h, \chi_h) : (\forall_{\phi(a)} (m_i, \iota_i) : A_h(m_i, \theta_h, \chi_h) = \iota_i). \quad (4)$$

The combinatorial functional $r_4(\hat{\zeta}_r X, M_A[\hat{\Theta}])$ that corresponds to (4) and whose value is unity for complete $M_A[\hat{\Theta}]$ is calculated as follows: $r_4(\hat{\zeta}_r X, M_A[\hat{\Theta}]) = \frac{1}{Y} \sum_{y=1}^{Y} (\max_{A_h \in M_A[\hat{\Theta}]} r_3(\phi(a_y), A_h))$, $Y = |\hat{\zeta}_r X|$.

A mask $\chi_3(\phi(a), \theta_h, A_h)$ is said to be dead-end for the correctness criterion if the zeroing of an arbitrary position of the mask χ_3 leads to the violation of condition (3). A set of masks $\chi_4 = \{\chi_3(\phi(a), \theta_h, A_h), h = 1 \ldots H, H = |M_A[\hat{\Theta}]|\}$ is said to be dead-end for the satisfiability of the completeness criterion if the removal of an arbitrary mask from χ_4 violates the satisfiability of (4). For fixed θ_h, A_h, $\chi_3(\phi(a), \theta_h, A_h)$ and χ_4 can be calculated by iterative zeroing of the positions of the initial unit mask.

$$\begin{array}{ccc} \textit{Completeness of model } M_A[\hat{\Theta}] & \Rightarrow & \textit{correctness } A_h \in M_A[\hat{\Theta}] \\ \Downarrow & & \Downarrow \\ \textit{regularity of problems} & \Rightarrow & \textit{solvability of problems} \Rightarrow \tau - \textit{sufficiency of feature descriptions} \end{array}$$

Fig. 1. Hierarchy of the combinatorial criteria.

Note that, in contrast to the solvability/regularity criteria of problems, the satisfiability of the criteria of the correctness of algorithms (3) and the completeness of the model $M_A[\hat{\Theta}]$ (4) depends not only on the set of precedents $\phi(a)$ and the mask χ, but also on the definition of A_h and on the values of the vectors of parameters θ_h (i.e., on the method of calculation of the vectors of parameters, $\hat{\Theta}$). Therefore, a combinatorial analysis of the dead-end property of the mask χ with respect to criterion (3) and, the more so, criterion (4) cannot be carried out without regard to the values of θ_h and the specific definition of A_h.

Thus, there is a certain hierarchy of criteria within an algorithmic model (Fig. 1). According to the completeness criterion of an algorithmic model (4), the completeness of the model implies the correctness of the algorithms of the model, and the correctness of an algorithm implies the solvability of the corresponding problems (Theorem 6). The completeness criteria (4) also imply the regularity of the problems.

In this case, the regularity occupies a special place in this hierarchy. First, regularity is the property of problems that guarantees the satisfiability of the solvability – the necessary condition for both correctness and completeness. Second, it follows from the regularity of problem $Z(\phi(a))$ that there exists at least one correct algorithm (the tabular value of the function A_h). At the same time, the regularity of $Z(\phi(a))$ does not guarantee the correctness of the algorithm A_h.

The weakest condition in this hierarchy is the necessary and sufficient condition, obtained in [10], for the existence of a topology over the set of feature descriptions of objects, which can be called a "τ-*sufficiency* of the set of feature descriptions", a "τ-criterion", etc. Without satisfying the τ-criterion, there can be no question of the satisfiability of stronger criteria in the hierarchy.

A given hierarchy of criteria should correspond to the hierarchy of appropriate subsets of features. In particular, to satisfy the regularity criterion, one needs more complex masks than those for satisfying the solvability criterion (Theorem 4). It is intuitively clear that, to satisfy stronger criteria of correctness, and, the more so, of completeness, one may need still more complex masks (i.e., masks with the greater values of the traces). A constructive analysis of this question becomes possible when carrying out a cross-validation analysis of the criteria obtained on real samples of descriptions of objects.

5. CROSS-VALIDATION ANALYSIS OF THE SOLVABILITY, REGULARITY, CORRECTNESS, AND COMPLETENESS CRITERIA WITHIN THE FACTORIZATION APPROACH

According to the cross-validation paradigm, the criteria of solvability, regularity, correctness, and completeness should be applicable not only to one specific problem (defined by a fixed set of precedents $\phi(a),\ a \in \hat{\zeta}X$), but over the whole collection of samples $\hat{\zeta}X$. Among the criteria (1) – (4) obtained above, a collection of samples appears only in criterion (4).

Empirical procedures of cross-validation and "sliding control" are used in computer science for controlling the overfitting of some "end-point" specific algorithms (literally – for specific programs for computers with pre-fixed values of parameters). In this case, one often does not consider the nature of overfitting of algorithms and puts the main focus precisely on the solution of the problem of computational efficiency, the correspondence of the procedures to one or another "accuracy" criterion, and so on. These technical questions are indeed very important, because the practical calculation of functionals of complete sliding control seems to be hardly possible [18].

At the same time, the analysis of the general structure of parametric algorithms $A_h(\theta, \chi)$ and of the corresponding models $M_A[\hat{\Theta}]$ shows that overfitting may arise for quite diverse reasons: (1) certain peculiarities of the construction of the algorithm A_h; (2) inadequate choice of the values of the vector of parameters θ of the algorithm (among other things, due to some peculiarities of the method $\hat{\Theta}$ used for calculating the vector θ); (3) inadequate procedure of selection of features (i.e. procedure of calculation of the mask χ); (4) incompleteness of the model $M_A[\hat{\Theta}]$; (5) some peculiarities of the actually analyzed sets of precedents, even if they correspond to solvable/regular problems, and so on.

The separation and quantification of the sources of overfitting presents most complicated problem, both

theoretically and computationally. The cross-validation forms of the criteria of solvability, regularity, correctness, and completeness formulated below are important tools and allow one to get at least the most general overview of the problem from the viewpoint of algebraic theory of recognition.

For convenience, we will consider a *homogeneous parametric algorithmic model* $M_A[\hat{\Theta}] = \{A_h(\theta_h, \chi_h)|\hat{\Theta}\}$, $h = 1,\ldots,|M_A[\hat{\Theta}]|$ in which the vectors of parameters θ_h of the algorithms have identical range of values (and, accordingly, dimension), the values of θ_h for all algorithms being calculated by a unique method defined by the operator $\hat{\Theta}$. The masks of selection of features χ_h in the homogeneous parametric model are also calculated uniquely for all algorithms of the model, so that each algorithm corresponds to a single vector θ_h and a single mask χ_h. A generalization to the case of inhomogeneous parametric models is made by defining $\hat{\Theta}$ as a set of several methods for calculating θ_h and is out of the scope of the present study.

In practice, the parameters of the algorithm can be calculated using a set of precedents (for example, "weights" of features, parameters of a hyperplane, regression coefficients, etc.) or are specified directly by the experimenter (for example, the maximum complexity of conjunctions and disjunctions in the method of logical rules, the dimension of the hyperspace, the complexity of the regression formula, etc.). We will assume that the definition of the operator $\hat{\Theta}$ takes into account both variants of parameters, so that, as a result, both variants of the parameters are the components of the vector θ_h.

Cross-validation scenario implies the "learning" or "setting" of an algorithm by one, "learning", sample and testing of the set algorithm by another, "control", sample. In the case of a homogeneous parametric model $M_A[\hat{\Theta}]$, the "learning" consists in calculating the vector of parameters θ_h and the mask of the selection of features χ_h. The calculation of θ_h by the set of precedents $\phi(a)$ by $\hat{\Theta}$ will be expressed as $\theta(\phi(a)) = \hat{\Theta}\phi(a)$, $a \in \hat{\zeta}X$.

On the basis of the set of precedents, a mask can be calculated as, for example, the dead-end mask χ_1 (Theorem 3), which guarantees solvability, as the dead-end mask χ_2, which guarantees regularity (Theorem 4), or by other methods. We will write the result of calculation of a mask χ by $\phi(a)$ as $\chi(\phi(a))$, and the result of calculation of the k-th element of the mask, as $\gamma_k(\phi(a))$.

6. CROSS-VALIDATION FORMS OF THE CRITYERIA OF SOLVABILITY AND REGULARITY OF PROBLEMS

Theorems 3 and 4 allow one to calculate dead-end masks, which include features with the maximum possible informativity (i.e., minimal value of rank of informativity, see above) and guarantee the fulfillment of the solvability and the regularity criteria of the corresponding problems. The statements of Theorems 3 and 4 imply, first, the application of heuristic estimation functionals of informativity of features (which allow to calculate the ranks of informativity) and, second, the arbitrariness of the assignment of the value of the rank in case when functional of estimation of informativity produces different ranks for the features with similar informativities (this problem is considered in greater detail in [14–16]).

The values of the heuristic informativity estimation functionals may significantly differ for different samples of objects; thus leading to a change in the linear ordering of features "by informativity" when using various samples from the set of samples $\hat{\zeta}X$ under test. Accordingly, the values of the function K(i,j) for identical pairs of objects will also differ for different samples. For equal values of the informativity estimation functional, features with identical informativity are assigned different ranks of informativity, which implies the violation of the linearity of the ordering.

For these and other reasons it is quite possible that for different samples $a \in \hat{\zeta}X$ one will obtain the different masks $\chi_1(\phi(a))$ and $\chi_2(\phi(a))$. The *cross-validation forms of the solvability criterion* reflects the necessity of "cross" checking on pairs of different sets of precedents:

$$\begin{aligned} &\forall\limits_{\hat{\zeta}X} a,b,a \neq b \ \forall\limits_{\phi(a)} q_1,q_2 : \iota_1 \neq \iota_2 \\ &\Rightarrow \exists\limits_{1..n} k : \neg\delta_k(\varphi_k^1,\varphi_k^2) \wedge \gamma_k(\phi(b)) = 1. \end{aligned} \tag{1.5}$$

The combinatorial functional corresponding to (1.5) can be expressed with the use of the functional $r_1(\phi(a),\chi)$: $r_{1c}(\hat{\zeta}X) = \frac{1}{Y(Y-1)}$ $\sum_{i=1}^{Y}\sum_{j=1,j\neq i}^{Y} r_1(\phi(a_i),\chi_1(\phi(a_j)))$, $Y = |\hat{\zeta}X|$.

The functional $r_{1c}(\hat{\zeta}X)$ corresponds to testing (1.4) on pairs of samples "learning"–"control," the functional $r_{1l}(\hat{\zeta}X) = \frac{1}{Y}\sum_{i=1}^{Y} r_1(\phi(a_i),\chi_1(\phi(a_i))), Y = |\hat{\zeta}X|$, estimates the results of testing on a single, "learning," sample, over all samples in the set $\hat{\zeta}X$. Within the cross-validation paradigm, the difference $r_{1l}(\hat{\zeta}X) - r_{1c}(\hat{\zeta}X)$ describes some "overfittedness" related to the algorithm for calculating $\chi_1()$, i.e., to the selection procedure of features on the basis of the solvability criterion.

Note that in practice it is more expedient to use not so much the functionals $r_{1c}(\hat{\zeta}X)$ and $r_{1l}(\hat{\zeta}X)$ themselves, but the corresponding empirical distribution functions (EDFs) of the values of the functionals r_1. Define sets $A_{1c} = \{r_1(\phi(a_i),\chi_1(\phi(a_j)))|a_i, a_j \in \hat{\zeta}X, a_i \neq a_j\}$ and $A_{1l} = \{r_1(\phi(a_i),\chi_1(\phi(a_i)))|a_i \in \hat{\zeta}X\}$. Define $\hat{\varphi}(x)$ – *the operator of formation of EDFs* by the set $A \subset R$ as $\hat{\varphi}(x)A = \sup|\{B \subseteq A|\forall a \in B : a \leq x\}|/|A|, x \in R$ (for short, we express "$\hat{\varphi}(x)A$" as "$\hat{\varphi}A$"). Then the functionals $r_{1c}(\hat{\zeta}X)$ and $r_{1l}(\hat{\zeta}X)$ are the mathematical expectations of the EDFs $\hat{\varphi}A_{1c}$ and $\hat{\varphi}A_{1l}$, and the difference between these EDFs can be more accurately characterized by the methods of nonparametric statistics developed by Smirnov and Kolmogorov [20–22], rather than simply by the value of the difference of the mathematical expectations.

By analogy with (1.5), we obtain a *cross-validation form of the regularity criterion:*

$$\underset{\hat{\zeta}X}{\forall} a,b,a \neq b \underset{\phi(a)}{\forall} q_1,q_2 : \underset{1..n}{\exists} k : \neg\delta_k(\varphi^1_k,\varphi^2_k) \wedge \gamma_k = 1, \quad (2.2)$$

corresponding to the EDFs $\hat{\varphi}A_{2c}$ and $\hat{\varphi}A_{2l}$ and the functionals $r_{2c}(\hat{\zeta}X)$ and $r_{2l}(\hat{\zeta}X)$, which are calculated by the values of the functional $r_2(\phi(a),\chi)$. The difference $r_{2l}(\hat{\zeta}X) - r_{2c}(\hat{\zeta}X)$ characterizes "overfittedness" when selecting features by calculating $\chi_2()$.

7. ON THE REPRODUCIBILITY OF THE SELECTION OF FEATURES UNDER CROSS-VALIDATION TESTING OF THE SOLVABILITY/REGULARITY CRITERIA

The differences $r_{1l}(\hat{\zeta}X) - r_{1c}(\hat{\zeta}X)$ and $r_{2l}(\hat{\zeta}X) - r_{2c}(\hat{\zeta}X)$ characterize some general "overfittedness" related to the procedures of selection of features. When carrying out a combinatorial analysis of solvability/regularity, of special practical interest is a more particular factor – reproducibility of the selection of the specific separate features in the framework of cross-validation testing.

Let χ be a mask calculated as a result of application of some procedure of selection of features on a set of precedents $\phi(a)$ (for example, it can be the $\chi_1()$ or the $\chi_2()$ mask), so that $\chi(\phi(a)) = (\gamma_1(\phi(a)),\ldots,\gamma_k(\phi(a)),\ldots\gamma_n(\phi(a)))$. When calculating χ by all $Y = |\hat{\zeta}X|$ samples of the set $\hat{\zeta}X$, one obtains a set of masks $\chi(\hat{\zeta}X) = \{\chi(\phi(a))|a \in \hat{\zeta}X\}$. Let $\breve{\chi}(\chi)$ be a combined mask of the set of masks χ, $|\chi| = |\hat{\zeta}X| = Y$,

$$\breve{\chi}(\chi) = (\overset{Y}{\underset{i=1}{\vee}} \gamma^i_1,\ldots,\overset{Y}{\underset{i=1}{\vee}} \gamma^i_k,\ldots,\overset{Y}{\underset{i=1}{\vee}} \gamma^i_n|(\gamma^i_1,\ldots,\gamma^i_k,\ldots\gamma^i_n) \in \chi).$$

Theorem 7. *The combined mask $\breve{\chi}(\chi)$ guarantees the satisfiability of cross-validation forms of the solvability (regularity) criterion on the set of samples $\hat{\zeta}X$ if and only if, for a given formalization method ϕ, criterion (1.4) (criterion (2.1)) is satisfied for every sample in the set $\hat{\zeta}X$.*

Proof. Consider the case of the solvability criterion (1.4); the proof for the regularity criterion (2.1) is analogous. The combined mask $\breve{\chi}(\chi)$ includes all the features marked by nonzero elements of masks from χ. If (1.4) is satisfied for every $\chi(\phi(a)) \in \chi(\hat{\zeta}X), a \in \hat{\zeta}X$, then it is certainly satisfied for the combined mask. Hence, in the case of an arbitrary $\phi(a)$, the combined mask contains all the positions $\chi(\phi(a))$ that guarantee the solvability of $Z(\phi(a))$. Thus, for an arbitrary pair of sets of precedents $\phi(a)$, $\phi(b)$, $a,b \in \hat{\zeta}X$, $\breve{\chi}(\chi)$ contains the positions $\chi(\phi(a))$ and $\chi(\phi(b))$, which guarantees the satisfiability of the cross-validation criterion (1.5) provided that criterion (1.4) is satisfied for all $\phi(a)$, $a \in \hat{\zeta}X$. The theorem is proved.

By Theorem 7, the mask $\breve{\chi}(\chi)$ guarantees the cross-validation satisfiability of the solvability/regularity criteria if the feature descriptions and the above-described procedures of calculating $\chi_1()$ or $\chi_2()$ are sufficient for the satisfiability of (1.4) and/or (2.1) for the sets of precedents derived from the individual samples in $\hat{\zeta}X$. The reproducibility of the sets of features obtained under cross-validation testing can be estimated by comparing various characteristics of the mask $\breve{\chi}(\chi)$ and of the elements of the set χ.

The most general characteristic of an arbitrary mask is its trace. Therefore, the *set of traces* $tr\chi(\hat{\zeta}X) = \{tr(\chi)|\chi \in \chi(\hat{\zeta}X)\}$, the corresponding EDF $\hat{\varphi}tr\chi(\hat{\zeta}X)$, and the numerical functionals of the EDF $\hat{\varphi}tr\chi(\hat{\zeta}X)$ characterize the degree of differences of the elements of χ from each other. It is clear that the closeness of the values of traces in $tr\chi(\hat{\zeta}X)$ to the value of trace of the mask $\breve{\chi}(\chi)$ points to the high reproducibility of the selection of features.

Let us evaluate the contribution of individual features to the implementation of solvability/regularity property in cross-validation. A *occupancy* z_k of the k-th feature is a fraction of the number of samples from $\hat{\zeta}X$ in which this feature figures as a distinguishing one, i.e., $z_k = \frac{1}{Y}\sum_{i=1}^{Y} \gamma_k(\phi(a_i))$, $a_i \in \hat{\zeta}X$. The value $z_k = 1$ indicates that the k-th feature is distinguishing on an arbitrary $\phi(a)$, $a \in \hat{\zeta}X$, and $z_k=0$ indicates that the

k-th feature is never a distinguishing one over all $\phi(a)$, $a \in \hat{\zeta}X$. Apparently, features with the *maximum occupancy* ($z_k = 1$) are of particular interest for the construction of the correct recognition algorithms.

When calculating $\chi_1()$ and $\chi_2()$, to increase the values of occupancies one should redefine the function K(i,j) (see Theorem 3) so that, for an arbitrary pair of objects, K(i,j) calculates the number of the equivalence class of informative features rather than the value of k (i.e., the informativity rank of the feature). Accordingly, in the calculated mask (be it $\chi_1()$ or $\chi_2()$), all the features corresponding to this equivalence class has to be marked. Then, the occupancy of features in the corresponding positions in the mask $\breve{\chi}(\chi)$ will be significantly higher (although in this case there can be no question of the dead-end character of the masks calculated in this manner). Nevertheless, the masks thus obtained will be, in a sense, "minimal": they will correspond to the selection of more informative features.

8. CROSS-VALIDATION FORMS OF THE CRITERIA OF THE CORRECTNESS OF ALGORITHMS AND OF THE COMPLETENESS OF MODELS

The satisfiability of the criteria (3) and (4) depends on the set of precedents $\phi(a)$, the mask χ, the method of definition of A_h, and the vector of parameters θ. In a homogeneous parametric model $M_A[\hat{\Theta}]$, the value of an arbitrary vector of parameters θ is calculated on the basis of some $\phi(a)$ as $\theta = \hat{\Theta}\phi(a)$. The masks used for selecting features are also constructed over a certain set of precedents so that $\chi = \chi(\phi(a))$. In this case, for one algorithm A_h, it seems expedient to calculate χ and θ by one set $\phi(a)$, otherwise the accuracy of the algorithm will be certainly reduced on both "learning" sample and in cross-validation. Accordingly, for a given method $\hat{\Theta}$ of calculating the vector of parameters, the *cross-validation criterion of correctness of an algorithm* is expressed by the application of the corresponding substitutions to condition (3):

$$\forall_{a,b\in\hat{\zeta}X}\ a \neq b\ \forall_{\phi(a)}(m_i,\iota_i): A_h(m_i(\phi(a)), \hat{\Theta}\phi(b), \chi(\phi(b))) = \iota_i. \tag{3.2}$$

Accordingly, one also obtains combinatorial functionals that characterize the "degree" of satisfiability of criterion (3.2) on "learning" $(r_{3l}(A_h,\hat{\zeta}X) = \frac{1}{Y}\sum_{i=1}^{N} r_3(\phi(a_i), A_h(\chi(a_i),\hat{\Theta}a_i))$ and on "control" $(r_{3c}(A_h,\hat{\zeta}X) = \frac{1}{Y(Y-1)}\sum_{i=1}^{Y}\sum_{j=1,j\neq i}^{Y} r_3(\phi(a_i), A_h(\chi(a_j),\hat{\Theta}a_j))$, $Y = |\hat{\zeta}X|$. Just as in the case of functionals $r_{1c}(\hat{\zeta}X)$, $r_{2c}(\hat{\zeta}X)$, etc., instead of $r_{3l}(A_h,\hat{\zeta}X)$ and $r_{3c}(A_h,\hat{\zeta}X)$, one can use EDFs over the sets $\{r_3(\phi(a_i), A_h(\chi(a_i),\hat{\Theta}a_i))\}$ and $\{r_3(\phi(a_i), A_h(\chi(a_j),\hat{\Theta}a_j))|i \neq j\}$, $i,j = 1,\ldots,Y$.

The commonly used criteria for the cross-validation estimation of the "accuracy" of algorithms are based on various combinations of *sensitivity* ("true positive rate", "recall", and in a problem with two classes, the fraction of objects of class C^+) and *specificity* ("true negative rate" and the fraction of objects of class C^-) of the algorithms. The criterion (3.2), thus, is a full-fledged functional of sliding control that reflects the fraction of correct answers of the algorithm in both classes (i.e., the values of $r_3()$).

The completeness criterion (4) already possesses a cross-validation structure because it involves the testing of samples of a regular $\hat{\zeta}_r X$. However, in this case one evaluates the application of the algorithm only to "learning" samples. Accordingly, the *cross-validation form of the completeness criterion of a homogeneous parametric algorithmic model* $M_A[\hat{\Theta}]$ involves testing of algorithms on "control" sets of precedents after setting the parameters using the "learning" sets of precedents:

$$\forall_{\hat{\zeta}_r X}\ a\ \exists_{\hat{\zeta}_r X}\ b \neq a \Big|\ \exists_{M_A[\hat{\Theta}]} A_h : \forall_{\phi(a)}(m_i,\iota_i): A_h(m_i,\hat{\Theta}\phi(b),\chi(\phi(b))) = \iota_i). \tag{4.1}$$

The cross-validation completeness of a homogeneous parametric algorithmic model $M_A[\hat{\Theta}]$ implies the presence of samples in a set of regular samples, $\hat{\zeta}_r X$, using which one can calculate the mask χ and the vector of parameters θ in such a way that at least one of the algorithms $A_h(\chi,\theta)$, $h = 1,\ldots,|M_A[\hat{\Theta}]|$, will be correct. The functional $r_4(\hat{\zeta}_r X, M_A[\hat{\Theta}])$ earlier obtained already characterizes the "accuracy" of the algorithm "on learning", and the combinatorial functional corresponding to criterion (4.1) differs from the functional $r_4()$ only in that it prohibits the testing of "learning" samples: $r_{4c}(\hat{\zeta}_r X, M_A[\hat{\Theta}]) = \frac{1}{Y}\sum_{y=1}^{Y}(\max_{A_h\in M_A[\hat{\Theta}], b\in\hat{\zeta}_r X} r_3(\phi(a_y), A_h(\hat{\Theta}\phi(b),\chi(\phi(b))), a_y \neq b)$, $Y = |\hat{\zeta}_r X|$.

Theorem 8. *The cross-validation correctness of an algorithm A_h is sufficient for the cross-validation completeness of all algorithmic models that include this algorithm if and only if the set of samples considered is regular and the strong form of the axiom of correspondence holds for the formalization method ϕ.*

Proof. Indeed, under criterion (3.2) over some set of samples $\hat{\zeta}X$ for an arbitrary $\phi(b)$, $b \in \hat{\zeta}X$, one

obtains $\hat{\Theta}\phi(b)$ and $\chi(\phi(b))$ that guarantee the correctness of the algorithm A_h. Hence, when A_h with parameters in $\hat{\Theta}\phi(b)$ and $\chi(\phi(b))$ is included in an arbitrary algorithmic model, in this model there is an algorithm, correct over the sets of precedents over the set of samples $\hat{\zeta}X$. The condition of the completeness of the model is satisfied only when all samples in the set $\hat{\zeta}X$ are regular (that is, $\hat{\zeta}X = \hat{\zeta}_r X$), and, when the strong form of the axiom of correspondence holds for each of the samples in $\hat{\zeta}X$, this set contains only regular samples. Thus, all $\phi(b)$, $b \in \hat{\zeta}X$, are regular, and the fulfillment of the condition (3.2) over the $\phi(b)$, $b \in \hat{\zeta}_r X$ implies that criterion (4.1) is true. The theorem is proved.

According to Theorem 8, over regular sets of precedents, the cross-validation criterion of correctness of an algorithm is a stronger condition than the completeness (4.1) of some model including this algorithm. In practice, "ideal" algorithms of this kind are quite rare, since the strict fulfillment of (3.2) implies the zero overfittedness of the algorithm (i.e., for learning on an arbitrary sample, 100% accuracy is achieved on control).

CONCLUSIONS

In this study, we have obtained constructive combinatorial criteria of the solvability/regularity of problems and the correctness and completeness criteria for algorithmic models that admit cross-validation testing on the samples of descriptions of objects. The satisfiability of the solvability/regularity criteria depends on the set of samples $\hat{\zeta}X$, the method of formalization of problems ϕ, and the method for calculating the masks of selection of features $\chi()$. The satisfiability of the correctness and completeness criteria depends on all these parameters and, in addition, on the algorithms A_h of the homogeneous parametric model $M_A[\hat{\Theta}]$ and the method $\hat{\Theta}$ for calculating the vectors of parameters of algorithms.

The presence of a hierarchy of the criteria obtained implies an obvious general approach to the analysis of poorly formalized problems. First, one finds all methods of formalization ϕ over the set of samples $\hat{\zeta}X$ that lead to the fulfillment of the condition of τ-sufficiency of the set of feature descriptions. From among these methods, one selects those that guarantee the fulfillment of the solvability/regularity conditions. Then, one carries out an analysis of the completeness of homogeneous parametric algorithmic models under test, which includes the analysis of the correctness of individual algorithms. The results of the analysis, including empirical estimates of overfitting, are represented by the values of combinatorial functionals obtained in the present study.

For the experimental cross-validation testing of the criteria obtained, one should introduce certain factorization methods, methods of estimating the informativity of features, methods of determining the equivalence classes of features by informativity, etc. Within the formalism developed, it seems adequate to select these methods on the basis of the metric approach to the analysis of given poorly formalized problems – the approach considered in the second part of the present article.

REFERENCES

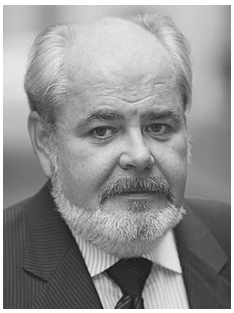
1. Yu. I. Zhuravlev, "Correct algebras for sets of incorrect (heuristic) algorithms. Part I," Kibernetika, No. 4, 5–17 (1977).
2. Yu. I. Zhuravlev, "Correct algebras for sets of incorrect (heuristic) algorithms. Part II," Kibernetika, No. 6, 21–27 (1977).
3. Yu. I. Zhuravlev, "Correct algebras for sets of incorrect (heuristic) algorithms. Part III," Kibernetika, No. 2, 35–43. (1978).
4. Yu. I. Zhuravlev, "On algebraic approach for solving recognition and classification problems," in *Problems of Cybernetics* (Nauka, Moscow, 1978), Issue 33, pp. 5–68 [in Russian].
5. K. V. Rudakov, "On some universal limitations for classification algorithms," Zh. Vychisl. Mat. Mat. Fiz. **26** (11), 1719–1729 (1986).
6. K. V. Rudakov, "Universal and local limitations in the problem on heuristic algorithms correction," Kibernetika, No. 2, 30–35 (1987).
7. K. V. Rudakov, "Completeness and universal limitations in the problem on heuristic classification algorithms correction," Kibernetika, No. **3**, 106–109 (1987).
8. K. V. Rudakov, "The way to apply universal limitations for researching classification algorithms," Kibernetika, No. **1**, 1–5 (1988).
9. K. V. Rudakov, *On Algebraic Theory of Universal and Local Limitations for Classification Problems. Recognition, Classification, Prediction* (Nauka, Moscow, 1989), pp. 176–201 [in Russian].
10. 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), 577–579 (2015).
11. I. Y. 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).
12. I. Y. 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**, 483 (2016).
13. K. V. Rudakov, "The theory of universal and local limitations for recognition algorithms," Doctoral Dissertation in Mathematical Physics (Moscow, Dorodnicyn Computing Centre RAS, 1992).
14. K. V. Rudakov and I. Yu. Torshin, "The way to analyze motifs' information capability according to solvability

criteria in the problem on protein secondary structure recognition," Inf. Ee Prim. **6** (1), 79–90 (2012).

15. Yu. I. Zhuravlev, K. V. Rudakov, and I. Yu. Torshin, "Algebraic criteria of local solvability and regularity as a tool for researching morphology of amino acid regularities," Trudy Mosk. Fiz.-Tekhn. Inst. **3** (4), 45–54 (2011).
16. 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).
17. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer-Verlag, 2009).
18. K. V. Vorontsov, "Combinatory theory of precedent learning reliability," Doctoral Dissertation in Mathematical Physics (Moscow, Dorodnicyn Computing Centre RAS, 2010).
19. S. V. Yablonskii, *Introduction into Discrete Mathematics* (Nauka, Moscow, 1986) [in Russian].
20. A. N. Kolmogorov, *Selected Works. Probability Theory and Mathematical Statistics* (Moscow, 1986) [in Russian].
21. N. V. Smirnov, "The way to approximate random varieties distribution laws according to empirical data," Usp. Mat. Nauk, No. **10**, 179–206 (1944).
22. L. N. Bol'shev and N. V. Smirnov, *Tables of Mathematical Statistics* (Nauka, Moscow, 1983) [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 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 225 publications in peer-reviewed journals in biology, chemistry, medicine, and informatics and of 3 monographs in 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, Federal Research Center "Informatics and Control," Russian Academy of Sciences, and Head of the Chair "Intelligent Systems" at the Moscow Institute of Physics and Technology.