PRIA JOURNAL SPECIAL ISSUE
XV INTERNATIONAL CONFERENCE "INTELLIGENT INFORMATION PROCESSING"(IIP-2024)
Section Papers—Theoretical Advances

Methods of Topological Analysis for Generating More Informative Synthetic Features Based On Support Chains and Arbitrary Metric Distance Functions

I. Yu. Torshin[a],*,**

[a] *Federal Research Center "Computer Science and Control", Russian Academy of Sciences, Moscow, 119333 Russian Federation*
**e-mail: tiy1357@yandex.ru*
***e-mail: tiy135@yahoo.com*

Abstract—The paper presents an analysis of the formalism of topological approach to analyzing poorly formalized problems based on the fundamental concepts of functional analysis. The results of the analysis made it possible to formulate a plethora of new approaches to the definition of the lattice estimates and of the ways of introducing metrics on lattices which arise over the topologies of the feature values. In particular, the use of so-called "support chains" to analyze Boolean lattices formed over Zhuravlev-regular sets of precedents allowed here to formulate a cutting-edge research area that consists in replacing the estimates of the lattice elements with certain types of the functions and/or of the vectors.The analysis also allowed to propose several new approaches to a systematic study of semiempirical (tunable) distance functionals known from the literature. These functional are applied as means to generate feature descriptions in the course of solving various applied problems. The analysis of the precedents' relations between the feature values and the target variable as sets of interactions of Boolean lattice elements indicated the possibility of generating synthetic features using metric distance functions. The paper formulates a few of perspective approaches for (1) estimating the relevance (or "informativeness") of the metrics in respect to the problems to be solved, and for (2) generation/selection of synthetic features, more informative than the initial feature descriptions (that generated the topology and the corresponding lattice). The paper also presents the results of experimental testing the algorithmic approaches based on the formalism developed. The computational experiments were performed with 2400 independent datasets from ProteomicsDB dealing with "molecule-numerical property" type of data. The experiments allowed to produce quite efficient algorithms for predicting numerical properties of the molecules (rank correlation in cross-validation was found to be 0.90 ± 0.23 when averaged over the 2400 datasets). The analysis of the results of experimentation indicated the metrics that most often generate the most informative synthetic features and the forms of corrective operations characterized by the best generalization ability.

Keywords: algebraic approach, feature classification theory, metric analysis, metric condensations, topological neighborhoods

DOI: 10.1134/S105466182560173X

1.INTRODUCTION

The research area called in brief "topological theory of pattern recognition" is an extension of the algebraic approach to pattern recognition developed in the scientific school of Yu. I. Zhuravlev and K. V. Rudakov for poorly formalized problems of pattern recognition, classification, and numerical prediction [13, 14]. The algebraic approach to finding solutions of the pattern recognition/classification problems studies algorithmic constructions of the form $\hat{A}_{(\theta_A)} = \hat{D}_{(\theta_D)} \circ \hat{C}_{(\theta_C)} \circ \hat{B}_{(\theta_B)}$ where $\hat{B}$ is the pattern-recognizing (or simply "recognizing") operator, $\hat{C}$ is the corrective operation (corrector), $\hat{D}$ is the decision rule, and $\theta_A = (\theta_D, \theta_C, \theta_B)$ are the corresponding vectors of the algorithm's parameters [13].

Algorithms of the form $\hat{A}_{(\theta_A)}$ are applied to input data about objects (information matrix M_{inp}) receiving answers from the algorithm $\hat{A}_{(\theta_A)}M_{inp}$. In the case of a *correct* algorithm$M_{out} = \hat{A}_{(\theta_A)}M_{inp}$, where the information matrix M_{out} describes the output data about objects in the corresponding set of precedents Q = (M_{inp}, M_{out}). Algorithm "training" by the set of precedents Q is considered to be a way to compute the vector $\theta_A(Q)$. The algorithm trained by Q is ε-*correct with respect to the test* Q' = (M'_{inp}, M'_{out}) if

Received April 7, 2025; revised April 16, 2025;
accepted April 29, 2025

ISSN 1054-6618, Pattern Recognition and Image Analysis, 2025, Vol. 35, No. 3, pp. 347–358. © Pleiades Publishing, Ltd., 2025.

$L(\mathrm{M}'_{\mathrm{out}}, \hat{\mathrm{A}}_{(\theta_A(\mathrm{Q}))}\mathrm{M}'_{\mathrm{inp}}) \le \varepsilon$, where$L$ is particular loss function. Various combinatorial functionals computed as part of the cross-validation design of computational experiments are introduced to evaluate the generalization ability of the algorithms [8–10].

An important direction of research in searching for the ε-correct algorithms in the scientific school of Yu. I. Zhuravlev–K. V. Rudakov is the study of the solvability and of the regularity of the problems Z(Q), $\mathrm{Q} = (\mathrm{M}_{\mathrm{inp}}, \mathrm{M}_{\mathrm{out}})$ to which the algorithms of the form $\hat{\mathrm{A}}_{(\theta_A)}$ are applied. Here the sets of precedents Q are considered as subsets of the Cartesian product of the set of *initial informations* (I_{i}) and of the set of *finite informations* (I_{f}), $\mathrm{Q} \subset \mathrm{I}_{\mathrm{i}} \times \mathrm{I}_{\mathrm{f}}$. Algorithms of the form $\hat{\mathrm{C}}_{(\theta_C)} \circ \hat{\mathrm{B}}_{(\theta_B)}$ (i.e., without a decision rule) can be applied for predicting the numerical target variables [6].

The results of the previous theoretical studies and a considerable experience in searching for practical applications of the constructs described here indicate that it is reasonable to construct the operators $\hat{\mathrm{B}}_{(\theta_{B_i}),i}$, $\hat{\mathrm{B}}^{(\alpha)}_{(\theta^{\alpha}_{B})}$, $\hat{\mathrm{C}}^{(\alpha)}_{(\theta^{\alpha}_{C})}$ etc. within the framework of the topological approach to the data analysis [6]. The target variables (i.e., the numerical values of any quantity or numerical labels of classes) are represented within the formalism as chains in a lattice constructed over corresponding topology. One of the main goals of this theory of the "topological pattern recognition" is to develop methods of systematic generation and of selection of the synthetic feature descriptions of objects, which would be characterized by greater informativeness than the original features [3].

This paper proposes a development of the formalism in the direction of a more detailed analysis of the structure of the lattice chains arising over arbitrary but Zhuravlev-regular sets of precedents. The practical applicability of the proposed formalism is illustrated by applying it to the problems from the area of pharmacoinformatics.

2. THE NOTATION AND THE DEFINITIONS

In the formalism developed, each object x from the set of initial descriptions of N_0 objects, $\mathbf{X} = \{x_1, \ldots, x_{N_0}\}$, $\mathbf{X} \subseteq \mathrm{S}$, is described by n features using functions $\Gamma_k : \mathrm{S} \to \mathrm{I}_k$ (where$\mathrm{I}_k = \{\lambda_{k_1}, \lambda_{k_2}, \ldots\}$ is the set of values of feature descriptions) and is thus represented by the sets $\{\Gamma_k^{-1}(\Gamma_k(x))\}$, $k = 1, \ldots, n + l$, where l is the number of the target (to be predicted) variables. The value of the t-th target variable of the object x, $t = n + 1, \ldots, n + l$, is computed by the functions $\Gamma_t(x)$.

We explain the meaning of the notation $\Gamma_k^{-1}(\Gamma_k(x))$ as follows. Object descriptions in $\mathbf{X} \subseteq \mathrm{S}$ are represented in some form corresponding to the problem domain (e.g., structures of molecules in chemoinformatics, crystal structures in solid state physics, symbolic sequences in bioinformatics, tables with patients' data in biomedicine etc.). Each function $\Gamma_k(x)$ computes the value of the k-th feature description for the original description of the object, $x \in \mathbf{X}$, used in the problem domain. The functions $\Gamma_k(x)$ introduce thereby a formal description of the object by the k-th feature. The function $\Gamma_k^{-1}(\lambda)$, being a function of the full prototype of the value $\lambda \in \mathrm{I}_k$ of the k-th feature, maps this feature value into a subset of objects characterized by this same feature value. Thus, "$\Gamma_k^{-1}(\Gamma_k(x))$" denotes a specific subset of the set $\mathbf{X}$ associated with a particular value of the k-th feature.

The set of the precedents over the space of admissible feature descriptions of objects $\mathrm{J}_{\mathrm{ob}}$is defined as $\mathrm{Q} = \varphi(\mathbf{X}) = \{\mathrm{D}(\mathrm{x}_\alpha) | \mathrm{x}_\alpha \in \mathbf{X}\}$ by the functions $\mathrm{D} : \mathrm{S} \to \mathrm{J}_{\mathrm{ob}}$ and $\varphi(\mathbf{X}) = \{\mathrm{D}(\mathrm{x}_\alpha) | \mathrm{x}_\alpha \in \mathbf{X}\}$, $\mathrm{D}(\mathrm{x}_\alpha) = (\Gamma_1(\mathrm{x}_\alpha) \times \ldots \times \Gamma_k(\mathrm{x}_\alpha) \times \ldots \times \Gamma_{n+l}(\mathrm{x}_\alpha))_\Delta$. If the Zhuravlev-regularity of the set of the precedents is assumed (which formulated as $\forall \mathrm{x} \in \mathbf{X}$, $\mathrm{x} = \mathrm{D}^{-1}(\mathrm{D}(\mathrm{x}))$), the-n$\mathbf{X}$is isomorphic to Q and both of the sets uniquely correspond to the topology $\mathrm{T}(\mathbf{X}) = \{\varnothing, \{\mathbf{X}\}, \mathrm{a} \cup \mathrm{b}, \mathrm{a} \cap \mathrm{b} : \mathrm{a}, \mathrm{b} \in \mathrm{U}(\mathbf{X}) = \{\Gamma_k^{-1}(\lambda_{k_b})\}\}$ and to the *Boolean lattice* $\mathrm{L}(\mathrm{T}(\mathbf{X})) = \{\mathrm{a} \vee \mathrm{b}, \mathrm{a} \wedge \mathrm{b} : \mathrm{a}, \mathrm{b} \in \mathrm{T}(\mathbf{X})\}$. These descriptions of the set of objects $\mathbf{X}$ remain the same regardless of the type of feature descriptions of objects (Boolean, categorical, or numerical) [7–9]. The concepts of the "lattice estimate", of "homogeneous functions", and of the "operator of the formation of empirical distribution functions" are applied to introduce the metrics along with some auxiliary operations (Definitions 1–4).

Definition 1. A lattice term or an *isotonic estimate* $\mathrm{v} : \mathrm{L} \to \mathrm{R}^+$ over $\mathrm{L}(\mathrm{T}(\mathbf{X}))$ is the function for which the *estimate condition* (**cE**: $\underset{\mathrm{L}}{\forall} a, b : \mathrm{v}[a] + \mathrm{v}[b] = \mathrm{v}[a \wedge b] + \mathrm{v}[a \vee b]$) and *isotonicity condition* (**cI**: $\underset{\mathrm{L}}{\forall} a, b : a \supseteq b \Rightarrow \mathrm{v}[a] \ge \mathrm{v}[b]$) are satisfied. As noted above, the existence of the metric $\rho_0(a, b)$ is guaranteed for v[].

Definition 2. We call arbitrary functions *homogeneous when they* have the same domain of definition (i.e., constructed over the same set of the argument's values) and the same domain of values.

Definition 3. Let there be a finite set of numbers, $A = \{a_1, a_2, \ldots a_i, \ldots, a_n\}$, $a_i \in R$. Define the operator $\hat{\varphi}(x)$ to form an *empirical distribution function* (EDF) of numbers over the set as $a\hat{\varphi}(x)A = \sup\left|\{B \subseteq A | \forall a \in B : a \le x\}\right| / |A|$, $x \in R$ such that $\hat{\phi}(-\infty)A = 0$, $\hat{\phi}(+\infty)A = 1$. We also write $\hat{\phi}(x)A$ as $\hat{\phi}A$ for brevity.

Definition 4. $\hat{\mu}$ is an *operator for computing the mathematical expectation* of the value$x \in A$ by EDF $\hat{\phi}A$ as $\hat{\mu}\hat{\phi}A = \frac{1}{m}\sum_{j=1}^{m} x_j(\hat{\phi}(x_j)A\hat{\phi}(x_{j1})A)$, where $m = |\hat{z}A|$, $x_j = \hat{\iota}^+(j)\hat{z}A$, and the arbitrary $x_0 < \inf(A)$, $x_0 \in R$, where $\hat{z}$ is an operator of forming the set of values in A, $\hat{z}A = B \subseteq A | \forall a \in A : a \in B$, $\forall a, b \in B : a \ne b$, $\hat{\iota}^+$ is a *set ascending ordering operator*; and $\hat{\iota}^+(j)A$ is the j-th element of the ordered set $\hat{\iota}^+A$. The other EDF momentum functionals are defined in similar manner.

3 THE ANALYSIS OF THE TOPOLOGICAL APPROACH CONSTRUCTS IN TERMS OF THE REFLEXIVE AND THE TRANSITIVE RELATIONS

Consider the topological approach to the pattern recognition including the construction of the topology $\mathrm{T}(\mathbf{X})$, of the lattice $\mathrm{L}(\mathrm{T}(\mathbf{X}))$ and the introduction of the appropriate metrics in terms of the fundamental concepts of function theory—the reflexive and the transitive binary relations between arbitrary sets. Only two such relations are used in mathematics: the symmetric equivalence relation of sets and the antisymmetric relation of (partial) order [2].

The precedent-based relation between the feature values $\Gamma_k(x)$ and the t-th target variable given by the set Q projected into the lattice $\mathrm{L}(\mathrm{T}(\mathbf{X}))$ corresponds to the set of pairs $\{(\{\Gamma_k^{-1}(\Gamma_k(x_i)), k = 1, ..., n\}, \Gamma_t^{-1}(\Gamma_t(x_i))), i = 1, ..., N_0\}$. Based on the precedents represented in a regular Q, any pattern recognition algorithm (or "machine learning" in general) builds a model of the relation described by the values of $\Gamma_k(x_i)$ and $\Gamma_t(x_i)$, i.e., between the collection of sets $\{\Gamma_k^{-1}(\Gamma_k(x_i))\}$ and the set $\Gamma_t^{-1}(\Gamma_t(x_i))$.

Focus on two arbitrary sets, $a = \Gamma_k^{-1}(\Gamma_k(x_i))$ and $b = \Gamma_t^{-1}(\Gamma_t(x_i))$. Clearly, the satisfiability of the equivalence or order relation between sets a and b in the general case is out of the question (because the equivalence corresponds to the identity of the k-th feature of the t-th target variable and partial order corresponds to kernel equivalence – i.e., the equivalence of a to the subset b or vice versa). These cases are trivial and correspond as a rule to an "easily solvable" recognition problem.

At the same time, the order relations existing in the lattice $\mathrm{L}(\mathrm{T}(\mathbf{X}))$ generate the supremum $a \vee b$ and the infimum $a \wedge b$ of sets a and b. Therefore, more complex 1 functionals over the sets a, b, $a \vee b$, $a \wedge b$ can be introduced within the topological theory of pattern recognition. The primary interest represent the functionals which describe the relations between the arbitrary a and b in terms of *distances*. If the four metric axioms are satisfied, these distance functionals, $\rho_{\mathrm{L}} : \mathrm{L}^2 \to \mathrm{R}^+$ form the *metric space of feature values*, $\mathrm{M_L}(\mathrm{L}(\mathrm{T}(\mathbf{X})), \rho_{\mathrm{L}})$.

Note that the simplest metric is the functional $\rho_0(a,b) = (\mathrm{v}[a \vee b] - \mathrm{v}[a \wedge b])/N_0$, where $\mathrm{v} : \mathrm{L} \to \mathrm{R}^+$ is an *isotonic estimate* on $\mathrm{L}(\mathrm{T}(\mathbf{X}))$ (see Definition 1). More complex definitions of ρ_{L} are possible by introducing parametric estimates [7] or using metrics known in the literature so that in general there is a number of metrics ρ_m, $m = 1, ..., m_0$. The metrics ρ_m are typically normalized to the interval of values [0...1].

Returning to the consideration of fundamental relations of the theory of functions, we can conclude that the metric $\rho_{\mathrm{L}}(a, b)$ is a functional that numerically evaluates the satisfiability of the equivalence relation between a and b based on order relations (in the form of $a \vee b$, $a \wedge b$). Indeed, $\rho_{\mathrm{L}}(a, b) = 0$ corresponds to the strict satisfaction of the equivalence relation of a and b, and $\rho_{\mathrm{L}}(a, b) = 1$ corresponds to the maximum possible distance between a and b (for example, (a, b) is 1 only for the sets $\varnothing$ and $\{\mathbf{X}\}$, which corresponds to the ends of the maximum chain of the lattice $\mathrm{L}(\mathrm{T}(\mathbf{X}))$.

Thus, the relation between the sets $\{\Gamma_k^{-1}(\Gamma_k(x_i))\}$ and $\{\Gamma_t^{-1}(\Gamma_t(x_i))\}$ can be modeled as the corresponding distance arrays generated by a particular metric ρ_m. In this paper, we investigate the ways of defining such distances and the methods of generating the synthetic feature descriptions over these arrays of the values of the metric distances.

4. ON THE DIFFERENT APPROACHES TO THE DEFINITION OF METRICS ON THE LATTICE L(T(X))

At least three fundamentally different ways of defining metric distances are known in the literature: (1) the metrics based on operations over arbitrary sets, (2) the metrics over the spaces of vectors, and (3) the metrics over the space of functions. Consider these three approaches as applied to the above-described constructs of the topological theory of pattern recognition.

Metrics based on the operations over sets. We mentioned in the Introduction metrics of distance estimate between $a, b \in \mathrm{L}(\mathrm{T}(\mathbf{X}))$ introduced as functionals over

$a \vee b$ (corresponding to $a \cup b$), $a \wedge b$ ($a \cap b$), element height estimates in L(T(**X**)), and other set-theoretic operations over sets *a*, *b*. It was proposed in [6] to use weighted isotonic lattice estimates ($\text{v}_{\boldsymbol{\alpha}} = \sum_{i=0,|\boldsymbol{\alpha}|} \omega_i \text{v}_{\alpha_i}$) based on parametric estimates v_{α_i} (v_{α}^{+}, v_{α}^{-}, d_{α}, etc.) to generate metrics tunable via rank-based optimization of $\boldsymbol{\alpha} \subset \text{L(T}(\mathbf{X}))$, $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \ldots, \alpha_i \ldots\}$. The sets $\boldsymbol{\alpha}$ can be formed on the basis of the "informativeness" of the relevant $\alpha_i \in \boldsymbol{\alpha}$ by the methods of the metric data analysis [11] or on the basis of different subchains of the target numerical variables.

At the same time, there are numerous known functionals of empirical nature which directly operate with sets: distances of Tanimoto, Rand, Russell-Rao, Simpson, Brown-Blanquet, Roger-Tanimoto, Faith, of dispersion, of images, Q_0, of Pearson and various versions of Tversky, Sokal-Sneath, Gower-Legendre, and Yulee distances, among others. [1]. The metric properties of these functionals can be demonstrated by analytical derivations or by a combinatorial analysis over a set of precedents.

Vector metrics. Alternatively to the weighted lattice method and based on a collection of sets $\boldsymbol{\alpha}$ and a number of estimates v_{α_i} for an arbitrary set $a \in \text{L(T}(\mathbf{X}))$, the vector $\overline{\vec{\text{v}}}_{\boldsymbol{\alpha}}[a] = (\text{v}_{\alpha_1}[a], \text{v}_{\alpha_2}[a], \ldots, \text{v}_{\alpha_i}[a] \ldots)$ can be computed and the metrics can be introduced already on the vector space $\vec{\text{v}}_{\boldsymbol{\alpha}}$ by means of the well-known approaches including (weighted/normalized) l_1-metrics, Minkowski l_{p}-metrics, distances of Penrose, Manhattan, Lorentz, Clark, Hellinger, Whittetaker, symmetric $\chi 2$, Mahalanobis (including adjustable weights), intersection distances, of Ruzecki, Roberts, Ellenberg, Gleason, Motyka, Bray-Curtis, Canberra, Kulczynski, and correlation distances (covariance, correlation, cosine, angular, chordal, similarity, Morisita-Horn, Spearman, and Kendall).

Metrics over the space of functions. The functional analysis and the probability theory provide a wide range of tools for defining the distances between functions with the same domain of definition including functionals of Kolmogorov (the maximum deviation), of von Mises, and of Renyi, various metrics (integral L_1, engineering, separations, and similarity of the harmonic mean), distances of Chebyshev, Stepanov, and Kuiper, distance versions of Zolotarev, Kruglov, Burbi-Rao, Bhattacharya, Chizar (including Kullback-Leibler divergence, $\chi 2$, and Hellinger distance) among others [1].

Clearly, some of these distance functions are admittedly not metrics. For example, the symmetry axiom is obviously violated in the Kullback-Leibler divergence; the satisfiability estimate of the triangle axiom for each of these functions requires a separate study. Metrification of these distance functions can be carried out by introducing additional constructs in the definition of the function. In particular, if there is no symmetry, the constructs like min(d(x, y), d(y, x)), max(d(x, y), d(y, x)) and others can be introduced for the function d(x, y). The main problem is "to tie" these approaches to the lattice formalism being developed. Some important concepts defined above in Definitions 1–4 can be used for the purpose. Next, we consider approaches to the analysis of the lattice L(T(**X**)) by way of a certain kind of EDFs.

5.ANALYSIS OF THE LATTICE L(T(X)) USING FUNCTIONSC $\hat{\phi}A$ BASED ON SUPPORT CHAINS

Definitions 2–4 significantly extend the formalism of lattice estimates (Definition 1) by allowing (1) to project L(T(**X**))into the corresponding EDF lattice by means of a certain pre-selected ("support") chain, (2) to measure distances between these EDFs, and (3) to introduce a new approach to the lattice estimates (Definition 1). In particular, a chain corresponding to a *t*th numerical target variable can be selected as a support chain.

Theorem 1. *Let us choose an arbitrary maximal chain* A_{t} *as a "support" for further constructions. Under the condition of regularity of sets in* **X**/Q, *each element of* L(T(**X**)) *corresponds to an empirical distribution function from a set of homogeneous EDFs*. **Proof**. If the regularity condition for **X**/Q is satisfied, the lattice L(T(**X**))is Boolean (Theorem 3 in [7]). The chains in L(T(**X**)) correspond to particular numerical feature descriptions such that the arbitrary (maximal) chain A_{t} in L(T(**X**)) can be represented in the form $\text{A}_{\text{t}} = \langle \text{u}(\lambda_{\text{t}_1}), \ldots, \text{u}(\lambda_{\text{t}_i}), \ldots \text{u}(\lambda_{\text{t}_m}) \rangle$, $\lambda_{\text{t}_i} \in \text{I}_{\text{t}}$, $\text{u}(\lambda_{\text{t}_i}) = \bigcup_{\beta=1}^{i} \Gamma_{\text{t}}^{-1}(\lambda_{\text{t}_\beta})$ where $\text{I}_{\text{t}} = (\lambda_{\text{t}_1}, \ldots, \lambda_m)$ is a strictly monotone sequence of numbers. The value of function Γ_t (including uncertainty) computable for any object in **X** is $\Gamma_t(q)$ for every lattice *atom* $\{q\} \in \text{L(T}(\mathbf{X}))$, the height of the atom is by definition 1 ($\text{h}[\{q\}] \equiv |\{q\}| \equiv 1$)). Since the lattice is Boolean, each of its elements is unique and can be represented as a combination of atoms. Accordingly, any element of the lattice $u \in \text{L(T}(\mathbf{X}))$ is uniquely associated to the set of values of the *t*-th feature $\boldsymbol{\Gamma}_t(u) = \{\Gamma_t(q), q \in u\}$ computed for all lattice atoms included in the element *u*. By applying the operator $\hat{\phi}(x)$ to the set $\boldsymbol{\Gamma}_t(u)$, we obtain the EDF $\hat{\phi}\boldsymbol{\Gamma}_t(u)$ for arbitrary *u*. If the regularity condition for **X**/Q is satisfied, the lattice L(T(**X**)) is uniquely associated to the lattice formed by the numerical sets $\boldsymbol{\Gamma}_t(u)$ for each element of which the function $\hat{\phi}\boldsymbol{\Gamma}_t(u)$ is computable. All these EDFs are homogeneous in construction since they are formed over the same set I_{t}. The theorem is proved.

So, given the support chain $\mathrm{A_t}$ (i.e., a support numerical feature with the domain of values $\mathrm{I_t}$), the set of numbers $\mathbf{\Gamma}_t(u)$, the numerical function $\hat{\phi}(x)\mathbf{\Gamma}_t(u)$, one variable $x \in R\,[\lambda_{\mathrm{t}_1} \ldots \lambda_{\mathrm{t}_{|\mathrm{I_t}|-1}}]$ nontrivially defined for each set $u \in \mathrm{L(T(\mathbf{X}))}$, and a number of functionals including $\hat{\mu}\hat{\varphi}(x)\mathbf{\Gamma}_t(u)$, are thus associated with the element $u \in \mathrm{L(T(\mathbf{X}))}$. Accordingly, the set $\mathbf{\Gamma}_t(u)$, EDF $\hat{\phi}\mathbf{\Gamma}_t(u)$, and the functionals like $\hat{\mu}\hat{\phi}\mathbf{\Gamma}_t(u)$ can be used to determine estimates in the lattice $\mathrm{L(T(\mathbf{X}))}$ based of the selected support chain. In addition, it becomes possible to introduce metric functions of the distance between functions $\hat{\phi}\mathbf{\Gamma}_t(u)$ by methods of functional analysis.

6. ESTIMATES IN THE LATTICE L(T(X)) BASED ON THE SET $\mathbf{\Gamma}_t$ USING THE CONCEPT OF MEASURE

Using Theorem 1, isotonic estimates based on the sets $\mathbf{\Gamma}_t$ are generated by functionals of the form $\mathrm{g} : 2^{\mathrm{I_t}} \to R^+$ such that yO is satisfied with arbitrary $u, v \in \mathrm{L(T(\mathbf{X}))}$ for $\mathrm{g}(\mathbf{\Gamma}_t(u))$ and $\mathrm{g}(\mathbf{\Gamma}_t(v))$ and cI, $\mathrm{g}(\mathbf{\Gamma}_t(u)) \geqslant \mathrm{g}(\mathbf{\Gamma}_t(v))$is satisfied for $u \supseteq v$. Some of the most obvious functionals~~g~~are various definitions of the *measure* of a set, which can be used, in fact, as lattice estimates [6, 7].

We point out the essential similarity between the concept of the estimate in lattice theory and the concept of measure in the functional analysis. Just like the lattice estimate, the measure is positively defined, the measure of an empty set is zero, and the measure of intersection of non-overlapping sets is equal to the sum of the measures of these sets. The estimate condition imposes an additional requirement that if sets intersect, the estimate of their union is equal to the sum of the measures of the sets minus the estimate of their intersection. Thus, any lattice estimate is a measure, but not every measure could be an estimate. In functional analysis, measures can be introduced in various ways, in particular, using discrete weights [2] which can be used to produce new types of the lattice estimates.

Definition 5. Let weights $\mathrm{p}_1, \mathrm{p}_2, \ldots, \mathrm{p}_{|\mathrm{I_t}|-1}$ be associated to the real axis points $\mathrm{I_t} = \{\lambda_{\mathrm{t}_1}, \lambda_{\mathrm{t}_2}, \ldots, \lambda_{\mathrm{t_b}}, \ldots, \lambda_{k_{|\mathrm{I_t}|-1}}, \Delta\}$, $\lambda_{\mathrm{t_b}} \in R$. Then a *measure with discrete weight* computable for an arbitrary set $\mathbf{\Gamma}_t(u)$ is defined as $\mu(\mathbf{\Gamma}_t(u)) = \sum_{\lambda_{\mathrm{t_b}} \in \mathbf{\Gamma}_t(u)} \mathrm{p_b}$.

There can be selected as weights in Definition 5: (1) ranges of values from $\mathrm{I_t}$ ($\lambda_{\mathrm{t_b}} - \lambda_{\mathrm{t_{b-1}}}$, $\lambda_{\mathrm{t_{b+1}}} - \lambda_{\mathrm{t_b}}$, etc.), (2) differences of EDF values ($\mathrm{cdf}(\lambda_{\mathrm{t_{b+1}}}, \mathrm{A_t}(\mathbf{X}))$ – $\mathrm{cdf}(\lambda_{\mathrm{t_b}}, \mathrm{A_t}(\mathbf{X}))$ etc.), (3) weights tuned according to a particular algorithm for solving the problem of correlation of metrics, etc. From this clearly follows

Theorem 2. *The measure* μ *with discrete weight is a lattice estimate*. The statement follows from the consideration of overlapping and non-overlapping sets $\mathbf{\Gamma}_t(u)$ and $\mathbf{\Gamma}_t(v)$ and from the satisfiability of cE from Definition 1.

Corollary 1. *The integral of a summable function using a measure with discrete weight is computed as* $\int_{-\infty}^{+\infty} f(\lambda) d\mu = \sum_{b=1,|\mathrm{I_t}|-1} \mathrm{p_b} f(\lambda_{\mathrm{t_b}})$.

Corollary 2. *The scalar product of summable functions $f(\lambda)$ and $g(\lambda)$ based on a measure with discrete weight is computed as* $(f, g) = \sum_{b=1,|\mathrm{I_t}|-1} \mathrm{p_b} f(\lambda_{\mathrm{t_b}}) g(\lambda_{\mathrm{t_b}})$.

Corollary 3. *The Kolmogorov "charge" functional defined on the set of numbers A using summable $f(\lambda)$ as* $\Phi(\mathrm{A}) = \int_{\mathrm{A}} f(\lambda) d\mu$ *is also a measure*. When using a measure with discrete weights, $\Phi(\mathrm{A}) = \sum_{\lambda_{\mathrm{t_b}} \in \mathrm{A}} \mathrm{p_b}(\lambda_{\mathrm{t_b}}) f(\lambda_{\mathrm{t_b}})$ (Corollary 1).

Corollary 4.*The Kolmogorov charge $\Phi(\mathbf{\Gamma}_t(u))$ is an isotonic estimate on the lattice* $\mathrm{L(T(\mathbf{X}))}$ *at positive definiteness $f(\lambda)$*.The overlap of the area under an arbitrary one-dimensional f in the case of sets $\mathbf{\Gamma}_t(u)$ and $\mathbf{\Gamma}_t(v)$ is equal to $\Phi(\mathbf{\Gamma}_t(u) \cap \mathbf{\Gamma}_t(v))$, which is equal to $\Phi(\mathbf{\Gamma}_t(u \cap v))$ and is equal to the sum of the areas of $\Phi(\mathbf{\Gamma}_t(u))$, $\Phi(\mathbf{\Gamma}_t(v))$ minus the area of the union of sets (which corresponds to the satisfiability of cE in Definition 1). The estimate of Φ is *isotonic* at $f(\lambda) \geqslant 0$.

Corollary 5. *If the frequencies of occurrence of values from* $\mathrm{I_t}$ *are selected as a measure with discrete weight* (*for example*, *the difference form of EDF*) *and $f(\lambda) = \lambda$, then the Kolmogorov charge $\Phi(\mathbf{\Gamma}_t(u))$ corresponds to the mathematical expectation of the t-th value on the set $\mathbf{\Gamma}_t(u)$*.

Corollary 6. *In the case of selecting the measure as in Corollary 5, the selection of any other function f except for $f(\lambda) = \lambda$ will correspond to preferential selection of some values $\lambda \in \mathrm{I_t}$ in the calculation of the charge Φ, i.e., the selection of a class of objects $x \in \mathrm{X}$ according to the values of $\mathbf{\Gamma}_t(x)$*.

It is obvious from Theorem 2 and its corollaries that the introduction of the functional Φ allows us to estimate in a more flexible way the contribution of each value of the target variable $\lambda_{\mathrm{t_b}}$ to the lattice estimate value. After all, not only discrete weights $\mathrm{p_b}$ of values $\lambda_{\mathrm{t_b}}$ are used in $\Phi(\mathrm{A})$, but also the weight function $f(\lambda)$ common for all values.

7. PROSPECTS OF USING METRIC ANALYSIS OF LATTICES WITHOUT USING THE CONCEPT OF THE LATTICE ESTIMATE

Lattice terms $\mathrm{v} : \mathrm{L} \to R^+$ give a scalar estimate of each element of the corresponding lattice L allowing to compare the elements of L among themselves (if the conditions cE and cI of the Definition 1 are satisfiable). Clearly, the association to an arbitrary element u of the lattice L(T($\mathbf{X}$)) of a function $\hat{\phi}\mathbf{\Gamma}_t(u)$ (represented as a vector $\mathbf{\Gamma}_t(u)$) is a much more complex "estimating" description of the set u than any particular scalar estimate v[u]. This chain of reasoning points out two directions for further development of formalism: (1) the introduction of certain functionals that allow to reduce a more complex description in the form $\hat{\phi}\mathbf{\Gamma}_t(u)$ to scalar estimates and (2) the development of new mathematical tools for lattice analysis, where functions of the form $\hat{\phi}\mathbf{\Gamma}_t(u)$ are applied instead of the lattice estimates.

The first direction is partly covered by the results of Theorem 2 with corollaries. The second direction is of interest in that it allows a researcher to introduce metric distance functions without using the construction $\mathrm{v}[x \vee y] - \mathrm{v}[x \wedge y]$ and simply based on the above-described functional analysis approaches. For example, it was shown earlier [6] that the maximum Kolmogorov deviation [2] is a metric on the space of homogeneous EDFs $\hat{\phi}(x)\mathbf{\Gamma}_k(u)$, $u \in$ L(T($\mathbf{X}$)). The metric properties of other distances on the space of homogeneous functions $\hat{\phi}(x)\mathbf{\Gamma}_k(u)$ can be illustrated in a similar manner.

Thus, within the scope of the proposed approach, the precedent relations between the sets $\{\Gamma_k^{-1}(\Gamma_k(x_i))\}$ and $\Gamma_t^{-1}(\Gamma_t(x_i))$ is modeled as corresponding distance arrays generated by a particular metric $\rho_m : \mathrm{L}(\mathrm{T}(\mathbf{X}))^2 \to [0\ldots1]$, $m = 1, \ldots, m_0$. For practical application of the formalism, it is necessary to formulate approaches to the study of the ρ_m properties, ways of estimating the relevance of functions with respect to the problems to be solved, and methods for generating and selecting synthetic features based on ρ_m. The paper presents the results of experimental tests of topological data analysis algorithms on pharmacoinformatics
2 problems (numerous chemokinomic datasets).

8. ON THE STUDY OF PROPERTIES OF DISTANCE FUNCTIONS ρ_m

The working hypothesis of this study is that semiempirical distance functionals on sets, vectors, and functions can be used to generate synthetic features $\{\Gamma_{k'}(x_i)\}$ that are more "informative" than the original features $\{\Gamma_k(x_i)\}$ [1]. The metric properties of the distance functions ρ_m can be investigated analytically or combinatorially using metric axioms [5, 6, 10]. We introduce the concept of a generalized estimated distance function in topological recognition theory to analyze the properties of these functionals,

Definition 6. *We call the "generalized distance estimation function" (or "generalized distance estimator") a construction of the form* $\rho(a,b) = f(g(\mathrm{v}[a \vee b]) - g(\mathrm{v}[a \wedge b]))$ *where f, g are functions monotonic in appropriate parts of the real axis and* $\mathrm{v} : \mathrm{L} \to \mathrm{R}^+$ *is an isotonic estimate for which the estimation condition* (**cE**: $\underset{\mathrm{L}}{\forall} a,b : \mathrm{v}[a] + \mathrm{v}[b] = \mathrm{v}[a \wedge b] + \mathrm{v}[a \vee b]$) and *isotonicity condition* (**cI**: $\underset{\mathrm{L}}{\forall} a,b : a \supseteq b \Rightarrow \mathrm{v}[a] \geq \mathrm{v}[b]$) *are satisfied*. Recall that $\mathrm{v}[a] = |a|$ *i.e., the height of the lattice element*, is the simplest functional for which yO and yI are satisfied.

Theorem 3. *The distance function* ρ *is a generalized distance estimator if and only if* $\rho(a,b) = \rho(a \vee b, a \wedge b)$ *and the terms from* a,b *in the formula for* $\rho(a,b)$ *are a composition of the monotone function and of the isotonic estimate.* The need follows from $a \vee b = (a \vee b) \vee (a \wedge b)$ and $a \wedge b = (a \vee b) \wedge (a \wedge b)$ when substituting $a \vee b$ and $a \wedge b$ for a and b in Definition 6. The equivalence of $\rho(a,b)$ and $\rho(a \vee b, a \wedge b)$ indicates that the expression to compute ρincludes term functionals containing the expressions $a \vee b$ and $a \wedge b$ interchangeable with a and b, i.e., terms of the form $g'(a \vee b)$ and $g'(a \wedge b)$. By the condition of the theorem, these terms include a monotone function from the isotonic estimate, i.e., g' is monotone. Since ρ is a distance function, the g' terms cannot be part of the expression for ρ as a product, sum, ratio, power or sum, but only as a difference, i.e., $\rho(a,b) = f(g'(a \vee b) - g'(a \wedge b))$, which implies sufficiency. The theorem is proved.

Corollary 1. *For the generalized estimator* ρ $\forall \ell \subseteq \mathrm{L}(\mathrm{T}(\mathbf{X}))$: $\Delta_{\vee\wedge}(\ell) \equiv 0$, $\Delta_{\vee\wedge}(\ell) = \sum_{a,b \in \ell} |\rho(a,b) - \rho(a \vee b, a \wedge b)| \cdot 2/|\ell|/(|\ell| - 1)$.

Corollary 2. *Select the "support" set* $a \in \mathrm{L}(\mathrm{T}(\mathbf{X}))$ *and the generalized estimator* ρ. *At* $f(x) = g(x) = x$ $\mathrm{v}_{\mathrm{a},\rho}[\mathrm{b}] = \rho(a,b) = \rho(a \vee b, a \wedge b)$ is an *isotonic estimate*. It follows from that any linear combination of isotonic estimates is an isotonic estimate provided it is positively definite (Theorem 2). It is also checked by direct substitution $\mathrm{v}_{\mathrm{a},\rho}[\mathrm{b}]$ in cE and cI.

Corollary 3. *Distances of Frechet—Nikodym*, *Aman*, *Rand/Schekanovsky*, *Sokal-Sneath* (*variants* 1, 2 *and* 3), *Russell-Rao*, *Roger-Tanimoto*, *Feith*, *Tversky and Yulee are generalized distance estimators*. The statement is verified by analytical checking of the theorem condition from the formulas of distance data (see definitions on p. 261 of the referenceguide [1]).

Corollary 4. *The distances of Simpson*, *Brown-Blanquet*, *Underberg* (*Sokal-Snice variant* 4), *and Gower* (*variant* 2) *are not generalized distance estimators.*

The theorem 3 with corollaries provides analytical and combinatorial tools for examining properties of semiempirical distance functions. If the analytic expression for a given semiempirical$\rho$is simple enough, it is easy to verify the fulfillment of the theorem's condition. If the given semiempirical$\rho$is a generalized estimation distance function, corresponding analytical expressions for functions f, g can be obtained. For example, it is easy to show using the statement of Theorem 3 that the Sokal-Sneath-2 distance $\rho(a,b) = 1 - |a \cap b| / (|a \cup b| + |a \Delta b|)$ is a generalized estimation distance with $f(x) = (e^x - 1)/(0.5e^x - 1)$ and $g(x) = \ln(x)$ (Corollary 3).

But if analytical conclusions are impossible for some reason (high complexity of studied ρ or the absence of the distance function in analytical form), the properties of ρ as a generalized estimator can be studied on subsets ℓ of the lattice $\mathrm{L}(\mathrm{T}(\mathbf{X}))$by computing the values of the functional $\Delta_{\vee\wedge}(\ell)$ (Corollary 1). The methods of selecting the set ℓ may vary involving subsets of chains, representative coverages of the lattice, etc. Corollary 2 makes it possible to generate new, previously unexplored functions of the estimates $\mathrm{v}_{a,\rho}[]$ on the basis of the support set $a \in \mathrm{L}(\mathrm{T}(\mathbf{X}))$ and the semiempiricalρ for which the condition of the theorem is satisfied.

9. ON WAYS TO ASSESS THE RELEVANCE OF THE METRICS ρ_m WITH RESPECT TO THE CLASSIFICATION/PREDICTION TASK

The bijection between the set of precedents Q and the set of initial descriptions of objects $\mathbf{X}$, which exists under Zhuravlev's regularity condition ($\forall \mathrm{x} \in \mathbf{X}$, $\mathrm{x} = \mathrm{D}^{-1}(\mathrm{D}(\mathrm{x}))$, guarantees the unambiguous correspondence of the descriptions of x_i and q_i. Thus, it becomes possible to consider precedent relations defined on Q in terms of the sets $\{\Gamma_k^{-1}(\Gamma_k(x_i))\}$ and $\Gamma_t^{-1}(\Gamma_t(x_i))$ using distances ρ_m on subsets of the set $\mathbf{X}$.

Let a target class of objects be given by means of the α-th value of the t-th variable, $\lambda_{t\alpha} \in \mathrm{I}_t$, $t = n + 1, \ldots, n + l$ as $\mathbf{c}_\alpha = \Gamma_t^{-1}(\lambda_{t\alpha})$. In the case of a numerical variable, each of the elements $\mathrm{u}(\lambda_{t\alpha})$ of the chain At can be taken as $\mathbf{c}_\alpha$. Since $\mathrm{L}(\mathrm{T}(\mathbf{X}))$ is Boolean, the complement of the set $\mathbf{c}_\alpha$, $\overline{\mathbf{c}}_\alpha = \mathbf{X} \backslash \Gamma_t^{-1}(\lambda_{t\alpha})$ is uniquely defined. Thus, the selection the class $\mathbf{c}_\alpha$ gives rise to the classification problem $\mathbf{c}_\alpha/\overline{\mathbf{c}}_\alpha$. Recall that any numerical variable prediction problem can be reduced to a sequence of correctly solvable problems, $\mathbf{c}_\alpha/\overline{\mathbf{c}}_\alpha$ [10].

Let a subset of features, $\mathrm{p} \subseteq [1\ldots n]$, and an element of the lattice $c \in \mathrm{L}(\mathrm{T}(\mathbf{X}))$ be given. Define the function $\boldsymbol{\rho}_{\mathbf{mc}}(x_i, c, \mathrm{p}) = \{\rho_m(c, \Gamma_k^{-1}(\Gamma_k(x_i)), k \in \mathrm{p})\}$. Compute sets of distances, $\boldsymbol{\rho}_{\mathbf{mc}}(x_i, \mathbf{c}_\alpha, \mathrm{p})$ and $\boldsymbol{\rho}_{\mathbf{mc}}(x_i, \overline{\mathbf{c}}_\alpha, \mathrm{p})$ for given ρ_m, p, $\mathbf{c}_\alpha$ and $\overline{\mathbf{c}}_\alpha$ for x_i. We introduce for simplicity the notation $\boldsymbol{\rho}_{\mathbf{m\alpha}}(x_i) = \boldsymbol{\rho}_{\mathbf{mc}}(x_i, \mathbf{c}_\alpha, [1\ldots n])$ and $\boldsymbol{\rho}_{\mathbf{m}\overline{\boldsymbol{\alpha}}}(x_i) = \boldsymbol{\rho}_{\mathbf{mc}}(x_i, \overline{\mathbf{c}}_\alpha, [1\ldots n])$. In addition, a set of distances, $\boldsymbol{\rho}_{\mathbf{m}}(x_i, \mathrm{p}) = \{\rho_{mk_1k_2}(x_i, \mathrm{p}) = \rho_m(\Gamma_{k_1}^{-1}(\Gamma_{k_1}(x_i)), \Gamma_{k_2}^{-1}(\Gamma_{k_2}(x_i)), k_1, k_2 \in \mathrm{p}, k_1 \neq k_2\}$, considered also as a metric configuration $(\rho_{mk_1k_2}(x_i, \mathrm{p}))$, $\boldsymbol{\rho}_{\mathbf{m}}(x_i) = \boldsymbol{\rho}_{\mathbf{m}}(x_i, [1\ldots n])$, is defined for each $x_i \in \mathbf{X}$.

Based on the sets of distances, $\boldsymbol{\rho}_{\mathbf{m\alpha}}(x_i)$ and $\boldsymbol{\rho}_{\mathbf{m}\overline{\boldsymbol{\alpha}}}(x_i)$, relevance estimates of each ρ_m can be introduced as a means of generating synthetic features that are more informative in a sense than the initial Γ_k. The metric ρ_m such that minimizes distances in the list of $\boldsymbol{\rho}_{\mathbf{m\alpha}}(x)$ and maximizes distances (i.e., "approximates" objects to their classes) in the list of $\boldsymbol{\rho}_{\mathbf{m}\overline{\boldsymbol{\alpha}}}(x)$ for all $x \in \mathbf{c}_\alpha$ is more relevant or "informative" with respect to the task $\mathbf{c}_\alpha/\overline{\mathbf{c}}_\alpha$. Two interrelated areas for further research stand out there: (1) to find subsets p of "more informative" features for a fixed ρ_m and (2) to tune/select ρ_m when the subset of feature pis fixed.

For $c' \in \mathrm{L}(\mathrm{T}(\mathbf{X}))$, we define the operation of merging lists of distances, $\boldsymbol{\rho}_{\mathbf{mc}}$, $\vartheta_{\mathbf{mc}}(c', c, \mathrm{p}) = \bigcup_{y \in c'} \boldsymbol{\rho}_{\mathbf{mc}}(y, c, \mathrm{p})$; denote $\vartheta_{\mathbf{m\alpha}}(\mathbf{c}, \mathrm{p}) = \vartheta_{\mathbf{mc}}(\mathbf{c}, \mathbf{c}_\alpha, \mathrm{p})$, $\vartheta_{\mathbf{m}\overline{\boldsymbol{\alpha}}}(\mathbf{c}, \mathrm{p}) = \vartheta_{\mathbf{mc}}(\mathbf{c}, \overline{\mathbf{c}}_\alpha, \mathrm{p})$; compute the setsof distances, 3
$\vartheta_{\mathbf{m\alpha}}(\mathbf{c}_\alpha, \mathrm{p})$, $\vartheta_{\mathbf{m\alpha}}(\overline{\mathbf{c}}_\alpha, \mathrm{p})$; and form the corresponding EDFs, $\hat{\phi}(x)\vartheta_{\mathbf{m\alpha}}(\mathbf{c}_\alpha, \mathrm{p}))$ and $\hat{\phi}(x)\vartheta_{\mathbf{m\alpha}}(\overline{\mathbf{c}}_\alpha, \mathrm{p})$, defined in construction on the segment [0...1].

We introduce the distance functional $d_f : \mathbf{M}_{0..1}^+ \to [0\ldots1]$ (for example, maximum Kolmogorov deviation $\mathrm{D}(f(x), g(x)) = \sup_x |f(x) - g(x)|$; a signed deviation D; von Mises, Renyi metrics; engineering metrics, metrics of Chebyshev, Stepanov etc.) on the space of homogeneous monotonically increasing functions $\mathbf{M}_{0..1}^+ = \{f : [0\ldots1] \to [0\ldots1], x \geq y \Rightarrow f(x) \geq f(y)\}$. The selection of d_f makes possible setting and solution of a number of problems of topological data analysis including:

1) quantitative relevance estimates of ρ_m by computing distances $d_f(\hat{\phi}(x)\vartheta_{\mathbf{m\alpha}}(\mathbf{c}_\alpha, \mathrm{p}), \hat{\phi}(x)\vartheta_{\mathbf{m\alpha}}(\overline{\mathbf{c}}_\alpha, \mathrm{p}))$ for different $\mathbf{c}_\alpha$ (in the case of the numerical t-th variable, estimates are computed for each $\lambda_{t\alpha} \in \mathrm{I}_t$, $\alpha = 1, \ldots, |\mathrm{I}_t|$);

2) formal formulation of optimization problems to increase the separation of classes, $\mathbf{c}_\alpha/\overline{\mathbf{c}}_\alpha$ by tuning ρ_m and/or selecting $\mathrm{p} \subseteq [1\ldots n]$ (for example, $\underset{\rho_m,\mathrm{p}}{\operatorname{argmax}}\, d_f(\hat{\phi}\vartheta_{\mathbf{m\alpha}}(\overline{\mathbf{c}}_\alpha, \mathrm{p}), \hat{\phi}\vartheta_{\mathbf{m\alpha}}(\mathbf{c}_\alpha, \mathrm{p}))$, adjoint problem $\underset{\rho_m,\mathrm{p}}{\operatorname{argmax}}\, d_f(\hat{\phi}\vartheta_{\mathbf{m}\overline{\boldsymbol{\alpha}}}(\overline{\mathbf{c}}_\alpha, \mathrm{p}), \hat{\phi}\vartheta_{\mathbf{m}\overline{\boldsymbol{\alpha}}}(\mathbf{c}_\alpha, \mathrm{p}))$, or problem $\hat{\mu}\hat{\phi}\vartheta_{\mathbf{m\alpha}}(\mathbf{c}_\alpha, \mathrm{p}) \to \min$ & $\hat{\mu}\hat{\phi}\vartheta_{\mathbf{m\alpha}}(\overline{\mathbf{c}}_\alpha, \mathrm{p}) \to \max$);

3) defining ρ_q metrics on the object space [5, 6, 11] (for example, in the form of $d_f(\hat{\phi}\boldsymbol{\rho}_{\mathbf{m}\alpha}(x, \mathrm{p}), \hat{\phi}\boldsymbol{\rho}_{\mathbf{m}\alpha}(y, \mathrm{p})))$, $d_f(\hat{\phi}\boldsymbol{\rho}_m(x, \mathrm{p}), \hat{\phi}\boldsymbol{\rho}_{\mathbf{m}}(y, \mathrm{p})))$ etc.;

4) estimation of metric closeness to the metric of the section by classes $\mathbf{c}_\alpha/\overline{\mathbf{c}}_\alpha$;

5) formulation of the criteria of solvability/regularity of the problem $\mathbf{c}_\alpha/\overline{\mathbf{c}}_\alpha$ and of correctness/completeness of the respective algorithm models [6];

6) estimates of compactness of classes $\mathbf{c}_\alpha$, $\overline{\mathbf{c}}_\alpha$ [7–9].

10. ON THE WAYS TO GENERATE AND SELECT SYNTHETIC FEATURES BASED ON DISTANCE FUNCTIONS

The sets of distances, $\boldsymbol{\rho}_{\mathbf{m}\alpha}(x_i, \mathrm{p})$, $\boldsymbol{\rho}_{\mathbf{m}\bar{\alpha}}(x_i, \mathrm{p}))$ and $\boldsymbol{\rho}_{\mathbf{m}}(x_i)$, as well as individual distances $\rho_m(\mathbf{c}_\alpha, \Gamma_k^{-1}(\Gamma_k(x_i))$, are used not only to assess the relevance of the metric distance ρ_m with respect to $\mathbf{c}_\alpha/\overline{\mathbf{c}}_\alpha$, but also for the formation of synthetic numerical features $\Gamma_{k'}(x_i)$ of the object x_i, $k' = n + l + 1, ..., n + l + n_s$.

The value of the synthetic feature $\Gamma_{k'}(x_i)$ depends on the selection of ρ_m, of classes $\mathbf{c}_\alpha$ and $\overline{\mathbf{c}}_\alpha$ and on the way it is computed including: (1) $\rho_m(\mathbf{c}_\alpha, \Gamma_k^{-1}(\Gamma_k(x_i))$, (2) $\rho_m(\overline{\mathbf{c}}_\alpha, \Gamma_k^{-1}(\Gamma_k(x_i))$, (3) $\rho_m(\mathbf{c}_\alpha, \ldots) - \rho_m(\overline{\mathbf{c}}_\alpha, \ldots)$, (4) 1 – $\rho_m(\mathbf{c}_\alpha, \ldots)$, (5) values of EDF $\hat{\phi}(x)\boldsymbol{\rho}_{\mathbf{m}\alpha}(x_i, \mathrm{p})$ for different x (e.g., corresponding to percentiles $\hat{\phi}\boldsymbol{\rho}_{\mathbf{m}\alpha}(x_i, \mathrm{p}))$, (6) values of $\hat{\phi}(x)\boldsymbol{\rho}_{\mathbf{m}\bar{\alpha}}(x_i, \mathrm{p})$ for different x, (7), and frequency forms of these EDFs $\hat{\phi}(x + \Delta x)\boldsymbol{\rho}_{\mathbf{m}\alpha}(x_i, \mathrm{p}) - \hat{\phi}(x)\boldsymbol{\rho}_{\mathbf{m}\alpha}(x_i, \mathrm{p})$ and $\hat{\phi}(x + \Delta x)\boldsymbol{\rho}_{\mathbf{m}\bar{\alpha}}(x_i, \mathrm{p}) - \hat{\phi}(x)\boldsymbol{\rho}_{\mathbf{m}\bar{\alpha}}(x_i, \mathrm{p})$, where Δx is the selected step value.

In addition, $\mathbf{c}_\alpha$ can be defined for the t-th numerical variable as $\Gamma_t^{-1}(\lambda_{t\alpha})$ or as $\mathrm{u}(\lambda_{t\alpha})$; if $\mathbf{c}_\alpha = \Gamma_t^{-1}(\lambda_{t\alpha})$, then class $\Gamma_t^{-1}(\lambda_{t\alpha+1})$ can be used instead $\Gamma_t^{-1}(\lambda_{t\alpha+1})$; the classes $\mathbf{c}_\alpha/\overline{\mathbf{c}}_\alpha$of the t-th variable can be defined using partitions into different percentiles (which are defined as a subsample of values of $\lambda_{t\alpha} \in \mathrm{I}_t$), etc.

Thus, the proposed schemes generate a great number of synthetic features $\Gamma_{k'}(x_i)$ (10·n or more with n being initial features), which makes it necessary to introduce feature selection procedures. The target variable $\Gamma_t(x_i)$ is numerical and the generated features $\Gamma_{k'}(x_i)$ are also numerical. There are several different approaches in applied mathematics for this case to estimate the relationship $\Gamma_t(x_i)$ and $\Gamma_{k'}(x_i)$.

Correlation estimates (correlation coefficient, correlation metrics) are used for *linear* regularities. The disadvantage of this approach is rare occurrence of linearity in real recognition tasks. One-dimensional approximation can be applied to nonlinear regularities with quality assessment (spline and more complex polynomials, formulas based on Fourier decomposition, etc.). This approach is limited by the number of approximation formulas that can be used.

The apparatus of probability theory/mathematical statistics provides a much more universal toolkit for establishing relationships between tested numerical variables x and y (tests for dependence/independence of variables), including those based on "mutual information" and other concepts of the Kolmogorov school [4].

The most fundamental and practical appear to be testing the relationship between two variables on the basis of the "null hypothesis" of their independence. Suppose there are pairs of tested values of x and y, (x_i, y_i), i = 1, ..., $n_{(x,y)}$, EDF. $\mathrm{F}_{\mathrm{xy}}(x, y)$ characterizes the joint distribution of x and y while EDF $\mathrm{F}_{\mathrm{x}}(x)$ and $\mathrm{F}_{\mathrm{y}}(y)$ are individual distributions of variables. The EDF corresponding to the null hypothesis (independence of x and y) is defined in an obvious way as $\mathrm{F}_{\mathrm{x}}(x)\mathrm{F}_{\mathrm{y}}(y)$.

In order to evaluate differences between $\mathrm{F}_{\mathrm{xy}}(x, y)$ and $\mathrm{F}_{\mathrm{x}}(x)\mathrm{F}_{\mathrm{y}}(y)$, it is necessary to introduce the distance between these functions (so-called "statistics") and to evaluate the significance of differences by means of some statistical test. Functions d_f adapted for the 2D case, for example, the maximum deviation $\mathrm{D}(\mathrm{F}_{\mathrm{xy}}(x, y), \mathrm{F}_{\mathrm{x}}(x)\mathrm{F}_{\mathrm{y}}(y)) = \max(|\mathrm{F}_{\mathrm{xy}}(x_i, y_i) - \mathrm{F}_{\mathrm{x}}(x_i)\mathrm{F}_{\mathrm{y}}(y_i)|)$ can be used as the distance and a Kolmogorov-Smirnov functional, P_{KS} $(\mathrm{D}(\mathrm{F}_{\mathrm{x}}(x)\mathrm{F}_{\mathrm{y}}(y)), n_{(\mathrm{x,y})})$ as the statistical test. Then 1 – P_{KS} characterizes the "informativeness" of x relative to y.

A more universal approach to evaluating the significance of differences between $\mathrm{F}_{\mathrm{xy}}(x, y)$ and $\mathrm{F}_{\mathrm{x}}(x)\mathrm{F}_{\mathrm{y}}(y)$ is to compute directly the selected statistic d_f on sets of pairs of values (x_i, y_i) yielded by a random number generator within a cross-validation design similar to bootstrap testing.

Suppose the *operator* $\hat{\zeta}$ *sampling* a set $\mathbf{X}$, forms a collection of subsets (samples) $\hat{\zeta}\mathbf{X} = \{a_1, a_2, \ldots, a_k, \ldots a_{|\hat{\zeta}X|} \mid a_k \subset \mathbf{X}\}$ and the random procedure is the used random number generator. It is assumed for each sample a_k that $n_{(\mathrm{x,y})} = |a_k|$ and the random values of x and y are used to compute the set of values of the selected statistic, d_f forming the set of numbers $\mathrm{rnd}(\hat{\zeta}\mathbf{X}, d_f) = \{d_f(\mathrm{F}_{\mathrm{xy}}(x_{ij}, y_{ij}), \mathrm{F}_{\mathrm{x}}(x_{ij})\mathrm{F}_{\mathrm{y}}(y_{ij})$, 4
x_{ij}, y_{ij} = random, j = 1, ..., $|a_i|$), i = 1, ..., $|\hat{\zeta}X|\}$.

Then for all $a \in \hat{\zeta}\mathbf{X}$, the value of $\mathrm{P}(d_f, \hat{\zeta}\mathbf{X}, a, k', t) = 1$ – $\hat{\phi}(d_f(\mathrm{F}_{k't}(\Gamma_{k'}(z), \Gamma_t(z)),$
$\mathrm{F}_{k'}(\Gamma_{k'}(z))\mathrm{F}_t(\Gamma_t(z)))|z \in a)\mathrm{rnd}$ $(\hat{\zeta}\mathbf{X}, d_f)$ is the statisti- 4

cal significance of associations between the variables $\Gamma_t(z)$ and according to the statistics on the sample, and the value $1 - \mathrm{P}(d_f, \hat{\xi}\mathbf{X}, a, k', t)$ evaluates the "extent" of the given association by numbers over the range of [0...1].

With this method of evaluating the association $1 - \mathrm{P}(d_f, \hat{\xi}\mathbf{X}, a, k', t)$, the problem of selecting informative features is solved by means of the so-called B-algorithm originally developed for building optimal vocabularies of the final information (as denoted by the letter "B") [5]. This algorithm based on the Zhuravlev's solvability criterion and can select sets of final informations from maximum partial coverage at a minimum of coverage elements. If we replace the computation of the size of intersection of sets by estimates of the association $1 - \mathrm{P}(d_f, \hat{\xi}\mathbf{X}, a, k', t)$, then the B-algorithm will choose the minimum of features with maximum "informativeness", which ensure the solvability of the problem (the most informative features, see Theorems 1, 7, 8 in [5]).

Thus, more informative synthetic features $\Gamma_{k'}(x_i)$ of objects from $x_i \in \mathbf{X}$ are synthesized within the formalism being developed in 5 steps as follows: (1) define a set of initial (generally "low-informative") features $\Gamma_k(x_i)$ and a target variable $\Gamma_t(x_i)$, (2) introduce a collection of metrics ρ_m, estimate their relevance $d_f(\hat{\phi}(x)\vartheta_{\mathbf{m\alpha}}(\mathbf{c}_\alpha, \mathrm{p}), \hat{\phi}(x)\vartheta_{\mathbf{m\alpha}}(\overline{\mathbf{c}}_\alpha, \mathrm{p}))$ for each class $\mathbf{c}_\alpha$ of values of the t-th variable and select the most relevant ρ_m, (3) generate synthetic features $\Gamma_{k'}(x_i)$ via each of the selected ρ_m, (4) select by calculating $1 - \mathrm{P}(d_f, \hat{\xi}\mathbf{X}, a, k', t)$ and applying the B-algorithm the minimum number of features of maximum "informativeness", (5) apply the algorithm for predicting the target variable (which in fact is the Zhuravlev-Rudakov corrector).

11. EXPERIMENTAL TESTING

This algorithm of topological data analysis was tried on a set of pharmacoinformatics datasets related 2 to chemokinomic assays aimed at obtaining quantitative estimates of the inhibition of human proteome kinases with advanced drugs. This set of tasks is related to modeling ligand-receptor interactions in which interaction constant values of the j-th substance, $\mathrm{EC}_{50}(j)$, are predicted based on the chemical structure of molecules (chemographs $G(j)$).The formalism allows to predict "directly" not only the values of $\mathrm{EC}_{50}(j)$, but also individual response values, $E_j(\mathrm{C}_i)$, such that the target variable $\Gamma_t(x_j)$ was defined as a numerical value of the predicted quantity (e.g., the constant $\mathrm{EC}_{50}(x_j)$ for the inhibition of this type of kinase by a molecule corresponding to the chemograph x_j) or as $E_j(\mathrm{C}_i)$.

The developed algorithms were tested on a data sample from ProteomicsDB (https://www.proteomicsdb.org) containing data on C_i (C_i = 1, 3, 10, 100, 1000, 3000, 30000 nmol/L), $\mathrm{E}_j(\mathrm{C}_i)$ and $\mathrm{EC}_{50}(j)$ for 300 enzyme kinases (the so-called "human kinome", a part of proteome) and 250 drug molecules from the ATX list (in total, more than 2400 independent data samples on kinase activity measurements). The practical importance of kinome data analysis is due to the fact that many kinases are target proteins of well-known and advanced drugs.

A set of chemoinvariants over the alphabet of element labels was used in predicting $\mathrm{EC}_{50}(j)$ and $\mathrm{E}_j(\mathrm{C}_i)$ as the set of initial informations I_i at a given concentration C_i based on the chemograph x_j. The initial feature descriptions for the chemograph $X \in \mathbf{\Gamma}$ ($\mathbf{\Gamma}$being the set of all chemographs based on the alphabet of labels Y) were generated as chemoinvariants based on the set of χ chains of length $\tilde{\mathrm{Y}}^{\mathrm{m}}(\mathrm{X})$ and the set of χ nodes, $\hat{\mathrm{Y}}(X)$ [11, 12]. Briefly, suppose there is a set of subgraphs (χ-chains and χ-nodes) $\boldsymbol{\pi} = \{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_n\} \subset \mathbf{\Gamma}$. Define the operator of the presence of a subgraphπ in the chemograph X as a Boolean variable, $\hat{\beta}[X]\pi = (|\pi \cap \mathbf{\Pi}(X)| > 0)$, $\mathbf{\Pi}(X) = \tilde{\mathrm{Y}}^{\mathrm{m}}(X) \cup \hat{\mathrm{Y}}(X)$. Then the result of successive application of the operator $\hat{\beta}$ to an arbitrary set $\boldsymbol{\pi}$ is a Boolean vector, $\hat{\boldsymbol{\beta}}[X]\boldsymbol{\pi} = (\hat{\beta}[X]\boldsymbol{\pi}_1, \hat{\beta}[X]\boldsymbol{\pi}_2, \ldots, \hat{\beta}[X]\boldsymbol{\pi}_n)$. For the given set of chemographs, the set of initial informations is $\mathrm{I}_\mathrm{i} = \bigcup_{k=1}^{|\mathbf{X}|} \hat{\boldsymbol{\beta}}[X_k](\tilde{\mathrm{Y}}^{\mathrm{m}}(X_k) \cup \hat{\mathrm{Y}}(X_k))$. The value m = 5 corresponding to optimal results of combinatorial regularity testing according to Zhuravlev [8, 9] was used.

Two approaches to synthetic feature generation were used including (1) the previously tested method of support functions (the basics of the method are set forth in [10]) and the metric approach proposed in this paper.

When applying the **method of support functions**, the algorithms $f_{\theta_k} : \mathrm{I}_\mathrm{i} \to \mathrm{R}$ were constructed in the form of compositions of nested corrective functions of the lower level (i.e., of the generation of synthetic features) for a fixed number of models $\mathrm{n}_{\mathrm{mod}}$: $f_{\theta_k} = \mathrm{g}(\mathrm{f}_1(\sum \omega_k^j x_k), \ldots, \mathrm{f}_l(\sum \omega_k^j x_k), \ldots)$, $l = 1, \ldots, \mathrm{n}_{\mathrm{mod}}$, where g is the external corrective function, f_l is the internal corrective functions ("models" of generation of synthetic numerical features), and $\mathrm{n}_{\mathrm{mod}}$ is their number.$\sum \omega_j^k x_j$ is summed over the components of the vector $\mathbf{x} \in \mathrm{I}_\mathrm{i}$, $\mathrm{k} = 1, \ldots, |\mathbf{x}|$. Linear, nonlinear, monotone and nonmonotone corrector functions of g and f_l (more than 20 versions of monotonic and non-

2 **Table 1.** Rank correlations between experimental and calculated values of EC_{50} and of other chemokinomic assay values. r,
2 rank correlation coefficient in training; r_c is in control. r and r_c were averaged on 2400 samples of chemokinomic data.

Experiment	r	r_c	SD	SD_c
Ordinary regression, g is linear	0.67 ± 0.25	0.45 ± 0.26	0.24 ± 0.15	0.25 ± 0.14
Rank regression, g is linear	0.68 ± 0.23	0.48 ± 0.25	0.20 ± 0.19	0.22 ± 0.20
Ordinary regression, g is neural network	0.89 ± 0.13	0.79 ± 0.13	0.18 ± 0.12	0.13 ± 0.11
Support functions, opt.1, g is neural network	**0.88 ± 0.15**	**0.83 ± 0.28**	**0.05 ± 0.03**	**0.05 ± 0.03**
Support functions, opt.2, g is neural network	**0.89 ± 0.13**	**0.81 ± 0.16**	**0.18 ± 0.17**	**0.17 ± 0.17**
Support functions, opt.3, g is neural network	**0.88 ± 0.15**	**0.86 ± 0.20**	**0.03 ± 0.02**	**0.04 ± 0.03**
Synthetic $\Gamma_{k'}(x_i)$, corrector is neural network (2 layers)	0.45 ± 0.22	0.22 ± 0.21	0.22 ± 0.2	0.25 ± 0.24
Synthetic $\Gamma_{k'}(x_i)$, corrector is neural network (10 layers)	0.52 ± 0.25	0.21 ± 0.20	0.27 ± 0.22	0.33 ± 0.29
Synthetic $\Gamma_{k'}(x_i)$, corrector is random forest, opt. 1	0.98 ± 0.15	0.67 ± 0.31	0.14 ± 0.11	0.25 ± 0.19
Synthetic $\Gamma_{k'}(x_i)$, corrector is random forest, opt. 2	0.99 ± 0.14	0.71 ± 0.35	0.04 ± 0.02	0.15 ± 0.15
Synthetic $\Gamma_{k'}(x_i)$, polynomial correctors, opt. 1	**0.93 ± 0.11**	**0.90 ± 0.23**	**0.08 ± 0.06**	**0.08 ± 0.08**
Synthetic $\Gamma_{k'}(x_i)$, polynomial correctors, opt. 2	**0.95 ± 0.08**	**0.86 ± 0.27**	**0.06 ± 0.05**	**0.10 ± 0.09**

monotonic transformations including those described in [6]). The parameter vectors were tuned by multistart stochastic optimization within the cross-validation design of the computational experiment [10].

When using the "**metric approach**" proposed in this paper, distance functions on sets, vectors and EDF (65 functions in total from the reference guide [1]) were applied as ρ_m. The vectors for the elements of $L(T(\mathbf{X}))$ were formed from the estimates of v_α^+, v_α^-, d_α [12] for each value of the t-th variable in the class $\mathbf{c}_\alpha$ (quartiles of Γ_t values were used as $\mathbf{c}_\alpha$). The relevance of ρ_m was estimated by the formula $d_f(\hat{\phi}(x)\vartheta_{\mathbf{m\alpha}}(\mathbf{c}_\alpha, \mathrm{p}), \hat{\phi}(x)\vartheta_{\mathbf{m\alpha}}(\overline{\mathbf{c}}_\alpha, \mathrm{p})$ for each $\mathbf{c}_\alpha$ using maximum deviation as d_f. Synthetic features of $\Gamma_{k'}(x_i)$ were generated by all the above methods ($\rho_m(\mathbf{c}_\alpha, \Gamma_k^{-1}(\Gamma_k(x_i))$, EDF features, etc.) and the selection was performed by the B-algorithm using $1 - P(d_f, \hat{\xi}\mathbf{X}, a, k', t)$ values. Just like with the support function method, neural networks with several layers (from 2 to 10) with a softmax activation function, polynomials of various designs (more than 20 formulas, including quasi-polynomial models with elementary functions), and additionally "random forests" of decision trees were used as prediction algorithms.

Consider correlation coefficient values in training and in control as in [6, 12] for a tentative estimate of generalization ability. The sampling operator $\hat{\xi}$ was implemented as a tenfold cross-validation with each object sample divided into 80% (training) and 20% (control). Preliminary experiments showed that the best results of predicting $EC_{50}(j)$ were obtained when 1) the effects of hydrogen atoms were neglected (i.e., simpler Y alphabets were used), and 2) a linear recognition operator was used in combination with a nonmonotonic corrector (neural networks, decision trees, polynomial functions, etc.). The experimental results are summarized in Table 1.

Table 1. Rank correlation between experimental and calculated values of $EC_{50}(j)$ for 300 human kinases under different computational experimental conditions. r, rank correlation coefficient in training; r_c is in control; SD, standard deviation in training, SD_c is in control. Experiments were conducted in a cross-validation design (10 partitions in a case-control ratio of 6:1). A 2-layer network with a softmax activation function was used as a "neural network". The search for optimal parameter values was carried out with multistart stochastic optimization. The best methods are highlighted in bold.

In the case of the support function method, the best result was produced using the neural network g tuned according to both rank criteria ($r_c = 0.86 \pm 0.20$ with a small difference from r in training). The best result of the new "metric" approach with a polynomial corrector ($r_c = 0.90 \pm 0.23$) slightly outperformed the best result of the support function method. The polynomial formulas that most often produced the best results were 1st or 2nd degree polynomials with the products of first-degree variables, 5th degree polynomials, quasi-polynomials of 5th degree with sigmoids, and Fourier polynomials of 3rd degree.

It is interesting to note that neural network correctors regardless of their configurations performed extremely poorly in the case of the "metric" synthetic features $\Gamma_{k'}(x_i)$ ($r = 0.45 \pm 0.22$, $r_c = 0.22 \pm 0.21$) while "random forest" resulted in a significant overtraining (Table 1).

The analysis of the synthetic features $\Gamma_{k'}(x_i)$ included in the best polynomial models showed that features generated using EDF based on support chains (Theorem 1) were the most frequent among more informative features according to the estimate of $1 - \mathrm{P}(d_f, \hat{\xi}\mathbf{X}, a, k', t)$ and the initial features $\Gamma_k(x_i)$ and features based on individual distances $\rho_m(\mathbf{c}_\alpha, \Gamma_k^{-1}(\Gamma_k(x_i)))$ were among the least informative ones. The functions ρ_m most frequently generating informative $\Gamma_{k'}(x_i)$ on the EDF space were maximum Kolmogorov deviation, "oblique" distance, and Lp, Renyi, $\chi 2$, von Mises, and the engineering metrics [1]. These 7 varieties of ρ_m generated on average across all data samples more than 50% of the most informative features $\Gamma_{k'}(x_i)$ selected by the B-algorithm.

12. CONCLUSIONS

Various functionals that estimate distances between sets, vectors, or functions are ubiquitous in applied mathematics. This paper shows that in establishing metric properties of these functionals the toolbox of the formalism of topological recognition theory can be significantly enriched with non-trivial metrics based on empirical and semiempirical distance functions. The approach to the generation of informative synthetic features proposed in the paper implies successive transformations of object descriptions including (1) an initial set of feature values $\Gamma_k(x_i)$, (2) a set of corresponding lattice elements $\Gamma_k^{-1}(\Gamma_k(x_i))$, (3) a set of distances (measured by means of ρ_m) between lattice elements corresponding to classes and features, (4) a set of EDF distances measured with ρ_m, and, finally, (5) a set of synthetic features $\Gamma_{k'}(x_i)$ of the object. The use of multiple metrics at the stage of feature generation allows us to consider the developed formalism as a variant of developing the ideology of ECA (estimates calculating algorithms) of the scientific school of Y. I. Zhuravlev and K.V. Rudakov. Experimental testing of the proposed approach on 2400 homogeneous problems of pharmacoinformatics made it possible to enhance the accuracy and generalization ability of the algorithms compared to the best available solutions.

FUNDING

This work was supported by ongoing institutional funding. No additional grants to carry out or direct this particular research were obtained.

CONFLICT OF INTEREST

The author of this work declares that he has no conflicts of interest.

REFERENCES

1. M. M. Deza and E. Deza, *Encyclopedia of Distances*, 4th ed. (Springer, Berlin, 2016). https://doi.org/10.1007/978-3-662-52844-0
2. A. N. Kolmogorov and S. V. Fomin, *Elements of the Theory of Functions and Functional Analysis* (Nauka, Moscow, 1989; Corier Corporation, 1957).
3. K. V. Rudakov and I. Yu. Torshin, "Selection of informative feature values on the basis of solvability criteria in the problem of protein secondary structure recognition," Dokl. Math. **84**, 871–874 (2011). https://doi.org/10.1134/s1064562411070064
4. G. Sosa-Cabrera, S. Gómez-Guerrero, M. García-Torres, and Ch. E. Schaerer, "Feature selection: A perspective on inter-attribute cooperation," Int. J. Data Sci. Analytics **17**, 139–151 (2023). https://doi.org/10.1007/s41060-023-00439-z
5. I. Yu. Torshin, "Optimal dictionaries of the final information on the basis of the solvability criterion and their applications in bioinformatics," Pattern Recognit. Image Anal. **23**, 319–327 (2013). https://doi.org/10.1134/s1054661813020156
6. I. Yu. Torshin, "On the formation of sets of precedents basedon tables of heterogeneous feature descriptions by methods of topological theory of data analysis," InformatikaiEePrimeneniya **17** (3), 2–7 (2023). https://doi.org/10.14357/19922264230301
7. I. Yu. Torshin and K. V. Rudakov, "On the theoretical basis of metric analysis of poorly formalized problems of recognition and classification," Pattern Recognit. Image Anal. **25**, 577–587 (2015). https://doi.org/10.1134/s1054661815040252
8. 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 Recognit. Image Anal. **27**, 16–28 (2017). https://doi.org/10.1134/s1054661817010151
9. 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 Recognit. Image Anal. **27**, 184–199 (2017). https://doi.org/10.1134/s1054661817020110
10. I. Yu. Torshin and K. V. Rudakov, "On the procedures of generation of numerical features over partitions of sets of objects in the problem of predicting numerical target variables," Pattern Recognit. Image Anal. **29**, 654–667 (2019). https://doi.org/10.1134/s1054661819040175
11. I. Yu. Torshin and K. V. Rudakov, "Topological chemograph analysis theory as a promising approach to the simulation modeling of the quantum-mechanical properties of molecules: Part I. On the generation of feature descriptions of molecules," Pattern Recognit. Image Anal. **31**, 800–810 (2021). https://doi.org/10.1134/s1054661821040246
12. I. Yu. Torshin and K. V. Rudakov, "Topological chemograph analysis theory as a promising approach to simulation modeling of quantum-mechanical properties of molecules. Part II: Quantum-chemical interpre-

tations of chemograph theory," Pattern Recognit. Image Anal. **32**, 205–217 (2022). https://doi.org/10.1134/s1054661821040258

13. Yu. I. Zhuravlev, *Selected Scientific Works* (Magistr, Moscow, 1998).

14. Yu. I. Zhuravlev, K. V. Rudakov, and I. Yu. Torshin, "Algebraic criteria for local solvability and regularity as a tool to investigate the morphology of amino acid sequences," Trudy MoskovskogoFiziko-TekhnicheskogoInstituta (Natsional'nogoIssledovatel'skogoUniversiteta) **3** (4), 67–76 (2011).

Translated by D. Sventsitsky

Publisher's Note. Pleiades Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

AI tools may have been used in the translation or editing of this article.

Ivan Yurevitch Torshin, born 1972, Candidate of Physical and Mathematical Sciences, Candidate of Chemical Sciences, Lead Researcher at the Federal Research Center "Computer Science and Control", Russian Academy of Sciences, Associate Professor at Moscow Institute of Physics and Technology, Lecturer at the Faculty of Computational Mathematics and Cybernetics, Moscow State University. 720 publications in scholarly journals in computer science, medicine, chemistry, and biology; 15 monographs, 11 of them in Russian and 4 in English (in the series "BioinformaticsinPost-GenomicEra", NovaBiomedicalPublishers, NY, 2006-2009).

SPELL: 1. infimum, 2. chemokinomic, 3. setsof, 4. rnd