О ЗАДАЧАХ ОПТИМИЗАЦИИ, ВОЗНИКАЮЩИХ ПРИ ПРИМЕНЕНИИ ТОПОЛОГИЧЕСКОГО АНАЛИЗА ДАННЫХ К ПОИСКУ АЛГОРИТМОВ ПРОГНОЗИРОВАНИЯ С ФИКСИРОВАННЫМИ КОРРЕКТОРАМИ*

И. Ю. Торшин[1]

Аннотация: Корректирующие операции (корректоры) в мультиалгоритмических конструкциях алгебраического подхода могут строиться на основе известных физических моделей и/или многоуровневых описаний физических объектов. В рамках топологического подхода к анализу плохо формализованных задач поиск включаемых в корректор алгоритмов может рассматриваться как задача комбинаторной оптимизации либо как задача минимизации некой функции потерь. Исследование окрестностей цепей в решетке подмножеств объектов позволило получить ряд критериев ранговой оптимизации, перспективных для решения задач прогнозирования числовых таргетных переменных. Формализм апробирован на задаче взаимодействия лиганд–рецептор в рамках хемокиномного анализа лекарств (данные ProteomicsDB). Наилучшие результаты прогнозирования констант EC_{50} наблюдались именно при использовании полученных ранговых критериев: при усреднении по 300 биологическим активностям коэффициент корреляции на контроле составил $0{,}86 \pm 0{,}20$.

Ключевые слова: топологический анализ данных; теория решеток; задачи оптимизации; регрессия; хемоинформатика

DOI: 10.14357/19922264230201 **EDN:** IGSPEW

1 Введение

В рамках алгебраического подхода исследуются конструкции вида $\hat{A}_{(\theta_A)} = \hat{D}_{(\theta_D)} \circ \hat{C}_{(\theta_C)} \circ \hat{B}_{(\theta_B)}$, где $\hat{B}$ — распознающий оператор; $\hat{C}$ — корректирующая операция (корректор); $\hat{D}$ — решающее правило; θ_D, θ_C, θ_B и $\theta_A = (\theta_D, \theta_C, \theta_B)$ — векторы параметров [1]. Алгоритмы $\hat{A}_{(\theta_A)}$ применяются к входной *матрице информации* $M_{\text{вх}}$ для получения выходной *информационной матрицы* $M_{\text{вых}}$, причем в случае корректного алгоритма $M_{\text{вых}} = \hat{A}_{(\theta_A)} M_{\text{вх}}$. Обучение алгоритма по множеству прецедентов $Q = (M_{\text{вх}}, M_{\text{вых}})$ рассматривается как способ вычисления вектора $\theta_A(Q)$. Алгоритм, обученный по Q, *ε-корректен относительно контрольного* $Q' = (M'_{\text{вх}}, M'_{\text{вых}})$, если $L(M'_{\text{вых}}, \hat{A}_{(\theta_A(Q))} M'_{\text{вх}}) \le \varepsilon$, где L — та или иная функция потерь. Для оценки обобщающей способности используются разнообразные комбинаторные функционалы [2, 3].

Важным направлением исследований в научной школе Ю. И. Журавлёва – К. В. Рудакова стало изучение разрешимости и регулярности задач, где множества прецедентов $Q \subset I_i \times I_f$ определены над множествами *начальных* (I_i) и *конечных информаций* (I_f) [4]. При поиске ε-корректных алгоритмов $\hat{A}_{(\theta_A)}$ для решения разрешимых задач утверждается возможность настройки векторов θ_C и θ_B с получением ε-корректного алгоритма при произвольном $\hat{D}_{(\theta_D)}$ [1].

В некоторых задачах возможна фиксация корректора $\hat{C}$ в соответствии с проблемной областью. Например, в физической химии и в биохимии для описания взаимодействия лиганд–рецептор используется уравнение Хилла–Ленгмюра, в котором фигурирует концентрация лиганда C:

$$\frac{E_{\max}}{E(C)} = 1 + \left(\frac{\text{EC}_{50}}{C}\right)^s, \tag{1}$$

где E — измеряемое в эксперименте значение отклика, оценивающего связывание (например, интенсивность свечения флуоресцентной метки); $E_{\max}$ — максимальное значение отклика, EC_{50} — константа процесса (концентрация вещества, вызывающая 50% от максимального отклика, т. е. точка перегиба S-образной кривой $E(C)$). Выражение (1) — термодинамическое описание равновес-

*Работа выполнена при поддержке гранта РНФ (проект № 23-21-00154) с использованием инфраструктуры Центра коллективного пользования «Высокопроизводительные вычисления и большие данные» (ЦКП «Информатика») ФИЦ ИУ РАН (г. Москва).

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, tiy135@yahoo.com

ного связывания лиганда L рецептором R в соответствии с уравнением квазихимической реакции

$$R + nL \leftrightarrow RL_n.$$

Приведение (1) к виду

$$\ln\left(\frac{E_{\max}}{E(C)} - 1\right) = -s\ln(C) + s\ln(\mathrm{EC}_{50}),$$

где s — коэффициент Хилла (угол наклона касательной в точке EC_{50}), указывает на возможность линейной аппроксимации вида

$$y_i(x_i) = bx_i + a\,,\;\; b = -s\,,\;\; a = s\ln(\mathrm{EC}_{50}),$$

что позволяет вычислять значения констант EC_{50} и s для n_e экспериментальных точек (например, методом наименьших квадратов):

$$b = \frac{n_e\sum_{i=1}^{n_e} y_i x_i - \sum_{i=1}^{n_e} y_i \sum_{i=1}^{n_e} x_i}{n_e\sum_{i=1}^{n_e} x_i^2 - \left(\sum_{i=1}^{n_e} x_i\right)^2};$$
$$a = \frac{1}{n}\left(\sum_{i=1}^{n_e} y_i - b\sum_{i=1}^{n_e} x_i\right).$$

Предположим, что молекулы лигандов заданы в виде множества хемогра́фов $\{G_j\}$, $j = 1,\ldots,N$, а для набора концентраций $\{C_i\}$, $i = 1,\ldots,n_e$, из физико-химических экспериментов получены значения $E_j(C_i)$ и вычислены значения констант $\mathrm{EC}_{50}(j)$. Если для такого набора данных можно построить алгоритмы хемометрического анализа [5, 6], которые на основании G_j позволяют прогнозировать $E_j(C_i)$, то выражение (1) может быть использовано как фиксированный «физический» корректор $\hat{C}$ алгоритма $\hat{A}$ при $D \equiv 1$. Для реализации этого «имитационного» алгоритма прогнозирования необходимо построить ε-корректные распознающие операторы $\hat{B}_{(\theta_{B_i}),i}$, $|E_j(C_i) - \hat{B}_{(\theta_{B_i}),i}G_j| \le \varepsilon$ [5].

Схема порождения алгоритма $\hat{A} = \hat{D}\circ\hat{C}\circ\hat{B}$ может использоваться не только для решения «финальных» задач $Z(M_{\text{вх}}, M_{\text{вых}})$, но и для важных промежуточных задач, таких как порождение более информативных «синтетических» признаковых описаний объектов (например, вида $\hat{B}_{(\theta_{B_i}),i}G_j$). В результате получаются вложенные алгоритмические структуры, которые могут описываться алгоритмами α-го уровня:

$$\hat{A}^{(\alpha)}_{(\theta^\alpha_A)} = \hat{D}^{(\alpha)}_{(\theta^\alpha_D)}\circ\hat{C}^{(\alpha)}_{(\theta^\alpha_C)}\circ\hat{B}^{(\alpha)}_{(\theta^\alpha_B)}.$$

Построение операторов $\hat{B}_{(\theta_{B_i}),i}$, $\hat{B}^{(\alpha}_{(\theta^\alpha_B)}$, $\hat{C}^{(\alpha)}_{(\theta^\alpha_C)}$ и др. целесообразно осуществлять в рамках топологического подхода к анализу данных [4, 7].

2 Исследование окрестностей цепей в решетке подмножеств объектов

В рамках топологического подхода алгоритм прогнозирования k-й таргетной переменной (например, представленный распознающим оператором вида $\hat{B}_{(\theta_{B_k}),k}$) находится в результате перебора цепей решетки в окрестности цепи, заданной k-й переменной [5], который может рассматриваться (1) как задача комбинаторной оптимизации (поиск множеств, формирующих соответствующую цепь решетки) или (2) задача минимизации того или иного функционала или «функции потерь» (невязка и др.). При задании метрики на элементах решетки (ρ_L) и определении расстояния между цепями (ρ_A) рассмотрим возможность сведения задачи комбинаторной оптимизации к задаче минимизации особой формы функционала (так называемая ранговая оптимизация).

Основы разрабатываемого топологического формализма изложены в [5, 6]. Вкратце: заданы множество исходных описаний объектов $\mathbf{X} = \{x_1,\ldots,x_{N_0}\}$, $\mathbf{X}\subseteq S$, множества значений признаков $I_k = \{\lambda_{k_1},\lambda_{k_2},\ldots,\lambda_{k_b},\ldots,\lambda_{k_{|I_k|}},\Delta\}$, $b = 1,\ldots,|I_k|$, функции $\Gamma_k : S \to I_k$, $k = 1,\ldots,n+l$, где n — число признаков; l — число таргетных (прогнозируемых) переменных; Δ — неопределенность. Тогда определены пространство допустимых признаковых описаний объектов $J_{\text{ob}} \subseteq I_i \times I_f$ ($I_i \subseteq I_1\times\cdots\times I_n$, $I_f \subseteq I_{n+1}\times\cdots\times I_{n+1}$), функции $D : S \to J_{\text{ob}}$, $D(x_\alpha) = (\Gamma_1(x_\alpha)\times\cdots\times\Gamma_k(x_\alpha)\times\cdots\times\Gamma_{n+1}(x_\alpha))_\Delta$ и $\varphi(\mathbf{X}) = \{D(x_\alpha)|x_\alpha\in\mathbf{X}\}$. Принимается, что множество $\mathbf{X}$ и *множество объектов* $Q = \varphi(\mathbf{X})$, $|Q| = N$, *регулярны* ($\exists\,\varphi^{-1} : \mathbf{X} = \varphi^{-1}(Q)$) [5].

Множество $\mathrm{U}(\mathbf{X}) = \{\Gamma_k^{-1}(\lambda_{k_b})\}$ рассматривается как предбаза *топологии*

$$T(\mathbf{X}) = \{\varnothing, I, a\bigcup b, a\cap b : a,b\in\mathrm{U}(\mathbf{X})\},$$

где $I = \{\mathbf{X}\}$ — единичный элемент. Топологии $T(\mathbf{X})$ соответствует *решетка*

$$L(T(\mathbf{X})) = \{a\vee b, a\wedge b : a,b\in T(\mathbf{X}),\\ (a\ge b)\text{ или }(a\le b)\}.$$

При регулярности $\mathbf{X}/Q$ решетка $L(T(\mathbf{X}))$ — булева, так что *булевы* признаки — вершины, *категориальные* признаки — антицепи, а *числовые* — цепи в $L(T(\mathbf{X}))$ [7]. Тогда k-му числовому признаку с множеством значений I_k, $\lambda_{k_{b-1}} \le \lambda_{k_b} \le \lambda_{k_{b+1}}$, соответствует цепь $A_k(\mathbf{X}) = A(I_k,\mathbf{X}) =$

$= \langle u(\lambda_{k_1}), \ldots, u(\lambda_{k_b}), \ldots \rangle$ в $L(T(\mathbf{X}))$, образованная множествами

$$u(\lambda_{k_b}) = \bigcup_{\beta=1}^{b} \Gamma_k^{-1}(\lambda_{k_\beta});\ \ A_k(\mathbf{X}) \in \mathbf{A}(\mathbf{X}),$$

где $\mathbf{A}(\mathbf{X})$ — множество всех цепей решетки $L(T(\mathbf{X}))$. *Эмпирическая функция распределения* (э. ф. р.) k-го признака определяется через совокупность точек $\mathbf{cdf}(A_k(\mathbf{X})) = \{(\lambda_{k_b} \in I_k, |u(\lambda_{k_b})|/N)\}$, а значение э. ф. р. в точке λ вычисляется как $\mathrm{cdf}(\lambda, A_k(\mathbf{X})) = |u(\lambda_{k_b})|/N$, $\lambda_{k_{b-1}} \le \lambda \le \lambda_{k_b}$ методами кусочно-линейной аппроксимации и т. п.

Булевой решетке $L(T(\mathbf{X}))$ сопоставлено *метрическое пространство значений признаков* $M_L(L(T(\mathbf{X})), \rho_L)$ с метрикой $\rho_L : L^2 \to R^+$. Если $v : L \to R^+$ — *изотонная оценка* ($\forall_L a, b : v[a]+v[b] = v[a \wedge b] + v[a \vee b], \underset{L}{\forall} a, b : a \supseteq b \Rightarrow v[a] \ge v[b]$), то функция $\rho(x, y) = v[x \vee y] - v[x \wedge y]$ — ρ_L-метрика [7]. При заданной ρ_L *расстояние* $\rho_A : \mathbf{A}(\mathbf{X})^2 \to R^+$ *между цепями* $a = \langle a_1, \ldots, a_i, \ldots \rangle$ и $b = \langle b_1, \ldots, b_j, \ldots \rangle$ может быть определено метрикой Хаусдорфа или как функционал от совокупности расстояний между ближайшими элементами цепей:

$$\rho_A(a, b) = \min\left(\sum_{i=1,|a|} \rho_L(a_i, \arg\min_{b_j \in b} \rho_L(a_i, b_j)), \right.$$
$$\left. \sum_{i=1,|b|} \rho_L(b_j, \arg\min_{a_i \in a} \rho_L(b_j, a_i)) \right).$$

Важно, что при любом определении ρ_A подразумевается однозначное соответствие элементов цепей a и b. При заданных $A_k(\mathbf{X})$ и множестве допустимых цепей $A(\mathbf{X})_{1,n}$ искомый ε-корректный алгоритм соответствует решению задачи оптимизации [5]:

$$aa = \underset{a \in \mathrm{A}(\mathbf{X})_{1,n}}{\arg\min}\ \rho_A\left(A_k(\mathbf{X}), a\right) | A(\mathbf{X})_{1,n} \subset \mathbf{A}(\mathbf{X}). \quad (2)$$

Выражение (2) описывает задачу комбинаторной оптимизации, при решении которой перебираются цепи из множества цепей $A(\mathbf{X})_{1,n}$. Для решения (2) необходимо определить метрики ρ_L и ρ_A, множество $A(\mathbf{X})_{1,n}$ и способ перебора цепей, как это делается при прогнозировании классов значений числовых переменных [4]. В настоящей работе рассмотрен подход, в котором $\hat{B}_{(\theta_{B_k}),k} = f_{\theta_k} : I_i \to I_k$, а θ_k вычисляются в результате минимизации некоторой «функции потерь» на выборке обучения.

В рамках такого «регрессионного» подхода будем считать, что f_{θ_k} вычисляет значения из некоторого $I_{\theta_k} \subseteq I_k$, что соответствует цепи $A(I_{\theta_k}, \mathbf{X}) = \langle A_1^{\theta_k}, \ldots, A_\beta^{\theta_k}, \ldots \rangle$, $\beta = 1, \ldots, |I_{\theta_k}|$, т. е. подцепи $A(I_k, \mathbf{X}) = \langle A_1^k, \ldots, A_b^k, \ldots \rangle$, $b = 1, \ldots, |I_k|$. При $I_{\theta_k} \subseteq I_k$ существует функция перенумерации $\delta : \mathbf{N} \to \mathbf{N}$, так что $b = \delta(\beta)$ и э. ф. р. для $y = f_{\theta_k}(x)$ вычисляется как $\mathrm{cdf}(y, A_k(\mathbf{X}))$. Множество цепей $A(\mathbf{X})_{1,n}$ в (2) зададим как $\{A(I_{\theta_k}, \mathbf{X})\}$ для всех $\theta_k \in \Theta_k$. Введем дополнительное понятие *мощностно́й функции расстояния* и рассмотрим задачу (2) в контексте анализа э. ф. р. k-й таргетной переменной.

Определение 1. Пусть $A, B \in L(T(\mathbf{X}))$ — два произвольных множества. Выражение $r(A, B) = ||A| - |B||$ назовем мощностно́й функцией расстояния между множествами A и B.

Теорема 1. *$r(A, B)$ является метрикой.*

Доказательство. Для метрической функции расстояния должны быть выполнены аксиомы тождества, симметричности и треугольника. Выполнение аксиомы тождества $r(A, A) = 0$ очевидно вследствие тождественности множества A самому себе, а аксиомы симметричности $r(A, B) = r(B, A)$ — вследствие операции модуля в определении $r(\,)$. Докажем выполнимость аксиомы треугольника. Рассмотрим случай $|A| \le |B| \le |C|$, в котором длины трех сторон треугольника равны $|B| - |A|$, $|C| - |B|$ и $|C| - |A|$. Тестируя выполнение аксиомы для треугольника, вершины которого соответствуют множествам A, B и C, запишем три неравенства: $|B| - |A| + |C| - |B| \ge |C| - |A|$ (т. е. $|C| \ge |C|$), $|C| - |B| + |C| - |A| \ge |B| - |A|$ (т. е. $|C| \ge |B|$) и $|C| - |A| + |B| - |A| \ge |C| - |B|$ (т. е. $|B| \ge |A|$). Первое из неравенств выполнено всегда, а второе и третье соответствуют рассматриваемому случаю ($|A| \le |B| \le |C|$). Остальные варианты соотношений $|A|$, $|B|$ и $|C|$ сводятся к рассмотренному посредством подстановки переменных, так что аксиома треугольника выполнена и $r(\,)$ — метрика. Теорема доказана.

Следствие 1. *Метрика $r(A, B)$ отражает расстояния между слоями решетки, в которую входят рассматриваемые множества A и B.* По определению в слое решетки расположены элементы одной высоты в L.

Следствие 2. *Пусть в решетке L задана метрика $\rho_L(A, B) = v(A \cup B - v(A \cap B)$ на основе изотонной оценки $v(\,)$, равной высоте элемента в L, $v(A) = h(A) = |A|$. Тогда $r(A, B) = |\rho_L(A, \varnothing) - \rho_L(B, \varnothing)|$, где $\varnothing$ — нулевой элемент решетки L.* Соответствует следствию 1.

Следствие 3. *В отличие от метрик на основе изотонных оценок $r(A, B)$ не зависит от совместного вхождения объектов в множества A и B.* В определении 1 отсутствуют теоретико-множественные операции «$\cup$», «$\cap$» и др.

Следствие 4. *Значения метрик $\rho_L(A, B)$ и $r(A, B)$ равны для двух произвольных элементов одной цепи*

решетки L. Очевидно из следствия 2 и того, что каждый элемент цепи соответствует определенному слою решетки.

Следствие 5. *Пусть метрика* $\rho_L(A,B)$ *определена как в следствии 2,* $\rho_L(A,B) \le \varepsilon$*, а множества* A *и* B *принадлежат разным цепям. Тогда* $r(A,B) \le \le \rho_L(A,B)$ *и* $r(A,B) \le \varepsilon$. При заданных условиях $\rho_L(A,B) = |A\Delta B| = |A\backslash B| + |B\backslash A| \le \varepsilon$. Очевидно, что $|A| = |A\cap B| + |A\backslash B|$ и $|B| = |A\cap B| + |B\backslash A|$, так что $r(A,B) = ||A\backslash B| - |B\backslash A||$. Так как сумма двух неотрицательных чисел $|A\backslash B|$ и $|B\backslash A|$ в $\rho_L(A,B) \le \le \varepsilon$ не может быть больше их разности в $r(A,B)$, то и $r(A,B) \le \varepsilon$.

Следствие 6. $r(A,B) \le \varepsilon$ *— необходимое условие* $\rho_L(A,B) \le \varepsilon$.

Следствие 7. *Пусть* $A, B \subset \mathrm{X}$. *Если* $r(A,B) \le \le \varepsilon$*, то и* $r(X\backslash A, X\backslash B) \le \varepsilon$. Из дистрибутивности L и принципа двойственности очевидно, что $\rho_L(A,B) = \rho_L(X\backslash A, X\backslash B)$, поэтому при $\rho_L(A,B) \le \le \varepsilon$ и $r(A,B) \le \varepsilon$, и $r(X\backslash A, X\backslash B) \le \varepsilon$.

Следствие 8. *Верхняя оценка* $\rho_L(A,B)$ *равна* $|A|+|B|$*, а среднее верхней и нижней оценок —* $\max(|A|,|B|)$.

Теорема 2. *Задача комбинаторной оптимизации* (2) *сводима к задаче минимизации различий между э. ф. р. таргетной переменной и э. ф. р. искомого алгоритма* f_{θ_k} *при использовании нижней оценки* ρ_L *на основе высоты элемента и задании функции потерь посредством неравенств* $\rho_L(A,B) \le \varepsilon$.

Доказательство. В (2) фигурируют расстояния ρ_L между множествами, входящими в цепи $A(I_k, \mathbf{X})$ и $A(I_{\theta_k}, \mathbf{X})$. Для прогнозирования k-й переменной заданы множество исходных описаний $\mathbf{X}$, множество прецедентов $Q = \varphi(\mathbf{X}) = \{(x_i, y_i^k),\ x_i \in I_i,\ y_i^k \in I_k \subset R,\ i = 1,\ldots,N\}$, f_{θ_k} и набор векторов $\theta_k \in \Theta_k$, причем при любом θ_k из Θ_k область значений f_{θ_k} остается равной I_{θ_k}. Объекты из Q формируют цепь $A(I_k, \mathbf{X})$, а для f_{θ_k} и θ_k определена цепь $\mathrm{A}(I_{\theta_k}, \mathbf{X})$.

При любых ρ_A и ρ_L множества в составе цепей однозначно сопоставимы. Для произвольного $\beta = 1,\ldots,|I_{\theta_k}|$ рассмотрим два множества $A^k_{\delta(\beta)} \in \in A(I_k, \mathbf{X})$ и $A^{\theta_k}_{\beta} \in A(I_{\theta_k}, \mathbf{X})$. ε-корректному алгоритму соответствует условие

$$\forall\, \beta : \rho_L\left(A^k_{\delta(\beta)}, A^{\theta_k}_{\beta}\right) \le \varepsilon,$$

т. е.

$$\sum_{\beta=1}^{|I_{\theta_k}|} \rho_L\left(A^k_{\delta(\beta)}, A^{\theta_k}_{\beta}\right) \le \varepsilon |I_{\theta_k}|.$$

В соответствии с теоремой 1 и ее следствиями при использовании ρ_L на основе изотонной оценки, равной высоте элемента, $r(\,)$ — нижняя оценка ρ_L. Заменим ρ_L на $r(\,)$ и получим

$$\sum_{\beta=1}^{|I_{\theta_k}|} \left|\left|A^k_{\delta(\beta)}\right| - \left|A^{\theta_k}_{\beta}\right|\right| \le \varepsilon\, |I_{\theta_k}|\,.$$

Поделив обе части неравенства на $|Q| = |\mathbf{X}| = N$, получим

$$\sum_{\beta=1}^{|I_{\theta_k}|} \left|\mathrm{cdf}\left(\lambda_{k_{\delta(\beta)}}, A_k(\mathbf{X})\right) - \mathrm{cdf}\left(\lambda_{k_\beta}, A(I_{\theta_k}, \mathbf{X})\right)\right| \le \le \frac{\varepsilon |I_{\theta_k}|}{N}\,.$$

При фиксированных I_{θ_k} и N это условие ε-корректного алгоритма можно рассматривать как регрессионную задачу, подразумевающую минимизацию значения ε по θ_k:

$$\arg\min_{\theta_k \in \Theta_k} \sum_{\beta=1}^{|I_{\theta_k}|} \left|\mathrm{cdf}\left(\lambda_{k_{\delta(\beta)}}, A_k(\mathbf{X})\right) - \right. \\ \left. - \mathrm{cdf}\left(\lambda_{k_\beta}, A(I_{\theta_k}, \mathbf{X})\right)\right|. \quad (3)$$

Теорема доказана.

Заметим, что выражение (3) описывает параметрический критерий задачи оптимизации, где исследователь фиксирует значение параметра I_{θ_k}. В простейшем случае, когда $I_{\theta_k} = [0, \lambda]$, задача (3) редуцируется к задаче прогнозирования классов значений k-й числовой переменной [4].

По определению функции $\mathrm{cdf}(\,)$ ее значения соответствуют некоторой доле объектов множества $\mathbf{X}$, так что в (3) разность значений $\mathrm{cdf}(\,)$ для заданного λ_{k_β} равна доле объектов из $\mathbf{X}$, ошибочно классифицированных относительно класса значений $u(\lambda_{k_\beta})$. Поэтому можно перейти от задачи (3) к аналогичной задаче, в которой ошибка алгоритма f_{θ_k} оценивается не суммированием по значениям в I_{θ_k}, а суммированием ошибок в $\mathrm{cdf}(\,)$ по индивидуальным объектам:

$$\arg\min_{\theta_k \in \Theta_k} \sum_{i=1}^{N} \left|\mathrm{cdf}\left(y_i^k, A_k(\mathbf{X})\right) - \right. \\ \left. - \mathrm{cdf}\left(f_{\theta_k}(x_i), A(I_{\theta_k}, \mathbf{X})\right)\right|. \quad (4)$$

Поскольку $I_{\theta_k} \subseteq I_k$, удобней оценивать значения э. ф. р. от f_{θ_k} используя э. ф. р. таргетной переменной:

$$\arg\min_{\theta_k \in \Theta_k} \sum_{i=1}^{N} \left|\mathrm{cdf}\left(y_i^k, A_k(\mathbf{X})\right) - \right. \\ \left. - \mathrm{cdf}\left(f_{\theta_k}(x_i), A_k(\mathbf{X})\right)\right|. \quad (5)$$

Очевидно, что в задачах (3)–(5) подразумевается минимизация отклонений рангов/процентилей значений функции f_{θ_k} с функцией потерь в виде

модуля. При выборе в качестве ρ_L, например, квадрата $r(\,)$ в задаче типа (5) будет минимизироваться сумма квадратов невязок этих рангов/процентилей. Назовем (3), (4) и др. задачами *ранговой оптимизации*. Рассмотрим описанный сценарий, полученный в рамках топологического подхода к анализу данных, в контексте регрессии/аппроксимации функций.

3 Ранговая оптимизация и задачи регрессии/аппроксимации

Рассмотрим случай одномерной вещественной функции, заданной набором точек $\mathrm{Pr} = \{(x_i, y_i)\}$, $i = 1, \ldots, N$, $x_i \in [a, b]$, $y_i \in [c, d]$. Пусть Pr соответствует некоторой «истинной» функции f_t, $y_i = f_t(x_i)$. Будем аппроксимировать набор Pr функцией f_θ, $y_i = f_\theta(x_i) + o(x_i, \theta)$. Пусть функции f_t и f_θ непрерывны, дифференцируемы и интегрируемы на интервале $[a, b]$. Необходимо найти вектор параметров θ, минимизирующий абсолютные значения $o(x_i, \theta)$ в соответствии с заданным критерием оптимизации.

Критерии оптимизации можно формулировать исходя из идеи минимизации площади $S(f_t, f_\theta)$, соответствующей суммарному отличию f_t и f_θ на $[a, b]$. В идеальном случае $S = 0$ ($o(x) \equiv 0$), иначе $f_t(x) = f_\theta(x) + o(x)$, $o(x) > 0$, и площадь S может быть оценена корнем из интеграла $\int_a^b (f_\theta\, dx - f_t\, dx)^2$, соответствующего критерию $\min_\theta \sum_{i=1}^N (y_i - f_\theta(x_i))^2$. Аналогично можно получить критерии минимизации взвешенной суммы невязок, критерий метода наименьших модулей и др. Также возможна оценка площади S взятием интеграла по Лебегу, когда суммируются ошибки в x при заданном y, и т. п.

Пусть при заданном Pr площадь S оценивается как сумма некоторых вкладов «w» отдельных точек, $S = \sum_{i=1}^N w(x_i, y_i) + o(\mathrm{Pr})$. Примем, что каждой точке сопоставлен одинаковый вклад μ в площадь S (с точностью до $o(\mathrm{Pr})/N$), так что $S \sim \mu N$. Такое допущение может быть оправдано при достаточно большом N и при достаточно равномерном распределении точек вдоль соответствующих осей (ситуация, характерная для «больших данных», производимых современными физико-химическими технологиями сбора данных). Тогда можно рассмотреть приближенные оценки площади S (и соответствующие критерии минимизации), основанные не на *разностях* в значениях y_i и $f_\theta(x_i)$, а на *подсчете числа точек* (*объектов*) *в* Pr *и в его подмножествах*.

Вернемся к случаю произвольного множества $\mathbf{X}$, $Q = \varphi(\mathbf{X}) = \{(x_i, y_i^k), x_i \in I_i, y_i^k \in I_k\}$, $I_k = (\lambda_{k_b}), k = 1, \ldots, n$. Пусть задано подмножество $\zeta = (\lambda_{k_\alpha}) \subseteq I_k$, $\alpha = 1, \ldots, m$, $m = |I_k|$. В одномерном случае каждому значению λ_{k_α} соответствует горизонтальная линия, параллельная оси абсцисс, а в случае произвольного $x_i \in I_i$ — гиперплоскость. Для заданного λ_{k_α} вычислим число объектов со значениями y_i^k ниже λ_{k_α}, $n_\alpha^{t\leq} = |u(\lambda_{k_\alpha})|$, и y_i^k выше λ_{k_α}, $n_\alpha^{t>} = |\mathbf{X} \backslash u(\lambda_{k_\alpha})|$. Определим аналогичные числа для λ_{k_α} в цепи $A(I_{\theta_k}, \mathbf{X})$, производимой алгоритмом f_{θ_k}, $n_\alpha^{\theta_k \leq} = |\{(x, y) \in Q | f_{\theta_k}(x) \leq \lambda_{k_\alpha}\}|$ и $n_\alpha^{\theta_k >} = |\{(x, y) \in Q | f_{\theta_k}(x) > \lambda_{k_\alpha}\}|$. При вкладе каждой из точек, равном μ, оценим значение площади S_α для выбранного λ_{k_α} как разность в числе точек (объектов), у которых значение y_i^k ниже порогового значения λ_{k_α}, и объектов, у которых значение y_i^k выше λ_{k_α}:

$$S_\alpha = \left(\left|n_\alpha^{t\leq} - n_\alpha^{\theta_k \leq}\right| + \left|n_\alpha^{t>} - n_\alpha^{\theta_k >}\right|\right) \mu\,.$$

По построению

$$n_\alpha^{t\leq} + n_\alpha^{t>} = n_\alpha^{\theta_k \leq} + n_\alpha^{\theta_k >} = |Q|,$$

так что

$$S_\alpha = 2\left|n_\alpha^{t\leq} - n_\alpha^{\theta_k \leq}\right| \mu\,,$$

что эквивалентно

$$S_\alpha = \\ = 2\mu N \left|\mathrm{cdf}\left(\lambda_{k_{\delta(\alpha)}}, A_k(\mathbf{X})\right) - \mathrm{cdf}\left(\lambda_{k_\alpha}, A(I_{\theta_k}, \mathbf{X})\right)\right|.$$

Оценим площадь $S(f_t, f_\theta)$ как математическое ожидание S_α по всем элементам множества ζ:

$$S = \frac{1}{m} \sum_{\alpha=1}^{m} S_\alpha.$$

Считая μ, N и m константами и минимизируя S по θ_k, приходим к задаче (3).

Таким образом, задачи ранговой оптимизации в рамках топологического подхода (минимизация расстояния между цепями решетки) также могут быть получены исходя из специфического способа оценки различий $S(f_t, f_\theta)$ между «истинной» функцией f_t и ее аппроксимацией f_θ. Перспективно использовать комбинации критериев (3)–(5), что позволит одновременно минимизировать и отличия индивидуальных объектов, и отличия значений э. ф. р. в паре «переменная–алгоритм».

Получаемые критерии оптимизации относятся к параметрическим — в качестве параметра выступает подмножество множества I_k. При определенном выборе подмножества ζ получаются конструкции, идеологически близкие к задачам квантильной регрессии [8]. При назначении весов значениям λ_{k_α} и исследовании э. ф. р. значений S_α можно получить более сложные критерии.

4 Результаты экспериментального тестирования формализма

Формализм апробирован на задаче взаимодействия лиганд–рецептор, в которой значения $\mathrm{EC}_{50}(j)$ прогнозируются исходя из химической структуры молекул. Решения задач в постановках (3), (5) позволяют прогнозировать не только сами значения $\mathrm{EC}_{50}(j)$, но и значения откликов $E_j(C_i)$, для которых затем используется корректор в виде (1). При прогнозировании $\mathrm{EC}_{50}(j)$ и $E_j(C_i)$ на основе хемогра́фа G_j в качестве множества начальных информаций I_i использовалось множество хемоинвариантов над алфавитом специальных меток (см. ниже). Алгоритмы $f_{\theta_k} : I_i \to R$ строились в виде композиций вложенных корректирующих функций нижнего уровня (порождение синтетических признаков) для фиксированного числа моделей $n_{\mathrm{mod}} : f_{\theta_k} = g(f_1(\sum \omega_k^j x_k), \ldots, f_l(\sum \omega_k^j x_k), \ldots)$, $l = 1, \ldots, n_{\mathrm{mod}}$, где g — внешняя корректирующая функция; f_l — внутренние корректирующие функции (модели порождения синтетических числовых признаков); n_{mod} — их число. Суммирование $\sum \omega_j^k x_j$ проводится по компонентам вектора $\mathbf{x} \in I_i$, $k = 1, \ldots, |\mathbf{x}|$. Использовались линейные, нелинейные, монотонные и немонотонные функции-корректоры g и f_l (более 20 монотонных и немонотонных преобразований, в том числе описанных в работе [6]). Векторы параметров настраивались мультистартовой стохастической оптимизацией в рамках кросс-валидационного дизайна [6].

При прогнозировании EC_{50} исходя из $E_j(C_i)$ использовались вложенные алгоритмические структуры, описываемые алгоритмами 2-го уровня:

$$\hat{A}^{(2)}_{(\theta^2_A)} = \hat{C}^{(2)}_{(\theta^2_C)} \circ \hat{B}^{(2)}_{(\theta^2_B)}.$$

В качестве корректирующей операции $\hat{C}^{(2)}_{(\theta^2_C)}$ использовалось уравнение Хилла (1), а в качестве распознающих операторов $\hat{B}^{(2)}_{(\theta^2_B)}$ — функции $g(f_1(\sum \omega_k^j x_k), \ldots)$, настраиваемые на множествах откликов $E_j(C_i)$.

Для хемографа $X \in \mathbf{\Gamma}$ ($\mathbf{\Gamma}$ — множество всех хемографов, основанное на алфавите меток Y) хемоинварианты порождались на основании множеств χ-цепей длины $\tilde{Y}^m(X)$ и χ-узлов $\hat{Y}(X)$ [6]. Вкратце: пусть задано множество подграфов (χ-цепей и χ-узлов) $\boldsymbol{\pi} = \{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_n\} \subset \mathbf{\Gamma}$. Определим оператор вхождения подграфа π в хемограф X как

$$\hat{\beta}[X]\pi = (|\pi \cap \mathbf{\Pi}(X)| > 0)\,, \; \mathbf{\Pi}(X) = \tilde{Y}^m(X) \cup \hat{Y}(X),$$

а последовательное применение $\hat{\beta}$ к $\boldsymbol{\pi}$ — булев вектор

$$\hat{\boldsymbol{\beta}}[X]\boldsymbol{\pi} = \left(\hat{\beta}[X]\boldsymbol{\pi}_1, \hat{\beta}[X]\boldsymbol{\pi}_2, \ldots, \hat{\beta}[X]\boldsymbol{\pi}_n\right).$$

Для множества хемографов $\mathbf{X}$ множество начальных информаций

$$I_i = \bigcup_{k=1}^{|\mathbf{X}|} \hat{\boldsymbol{\beta}}[X_k] \left(\tilde{Y}^m(X_k) \cup \hat{Y}(X_k)\right),\; m = 5$$

(соответствует оптимальным результатам тестирования регулярности по Журавлёву [6]).

Тестирование алгоритмов f_{θ_k} и $\hat{A}^{(2)}_{(\theta^2_A)}$ проводилось на выборке данных из ProteomicsDB (https://www.proteomicsdb.org), содержащей данные по C_i (C_i = 1, 3, 10, 100, 1000, 3000, 30 000 нмоль/л), $E_j(C_i)$ и $\mathrm{EC}_{50}(j)$ для 300 ферментов-киназ (так называемый кино́м человека, часть протеома) и ряда лекарств. Киназы представляют собой таргетные белки известных и перспективных лекарственных средств. Наилучшие результаты прогнозирования $\mathrm{EC}_{50}(j)$ получались при (1) пренебрежении эффектами атомов водорода (т. е. при использовании более простых Y-алфавитов), (2) использовании линейного распознающего оператора в сочетании с немонотонными корректорами (нейронные сети, решающие деревья, полиномиальные функции и др.), 3) совместном использовании критериев (3) и (5). Результаты экспериментов суммированы в таблице.

Как при использовании линейной, так и нейросетевой $g(\,)$, применение критериев ранговой регрессии (3) и (5) способствовало снижению различий между значениями коэффициента корреляции на обучении и контроле. Наиболее выраженный эффект наблюдался для (5), в то время как минимизация отклонений э. ф. р. по (3) имела вспомогательное значение. Наилучший результат получен при использовании нейросетевой g, настраиваемой в соответствии с обоими ранговыми критериями ($r_c = 0{,}86 \pm 0{,}20$).

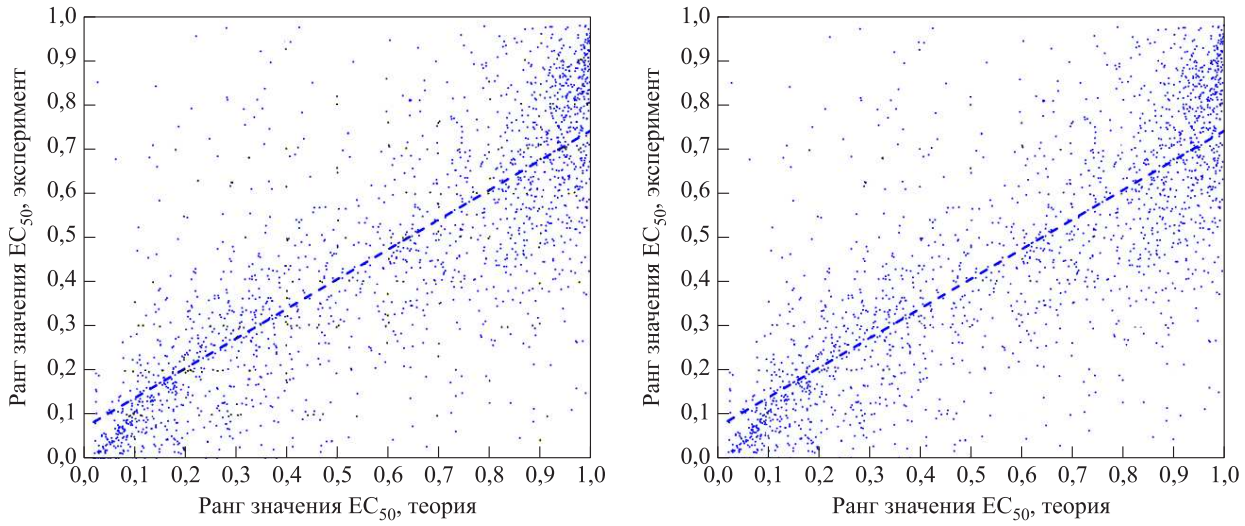
В отличие от описанного выше «прямого» прогнозирования EC_{50} прогнозирование $E_j(C_i)$ с использованием корректора (1) оказалось менее успешным (см. рисунок). Точность прогнозирования отдельных значений $E_j(C_i)$ была сопоставима с приведенной в таблице ($r_c \sim 0{,}85 \pm 0{,}21$), но данная схема прогнозирования отличалась существенно более низким качеством на контроле ($r_c \sim 0{,}63 \pm 0{,}24$).

В целом критерии (5) и (3) могут успешно использоваться не только для хемокиномного анализа, но и для хемотранскриптомного анализа лигандов, поиска эффективных и безопасных средств для фармакотерапии COVID-19, в хемомикробиомном анализе и др. (см. ресурс www.chemoinformatics.ru).

Тестирование расчетов $EC_{50}(j)$ для 300 киназ человека

Эксперимент	r	r_c	CO	CO_c
Обычная регрессия, g — линейная	$0{,}67 \pm 0{,}25$	$0{,}45 \pm 0{,}26$	$0{,}24 \pm 0{,}15$	$0{,}25 \pm 0{,}14$
Ранговая регрессия по (5), g — линейная	$0{,}68 \pm 0{,}23$	$0{,}48 \pm 0{,}25$	$0{,}20 \pm 0{,}19$	$0{,}22 \pm 0{,}20$
Ранговая регрессия по (3), g — линейная	$0{,}67 \pm 0{,}25$	$0{,}47 \pm 0{,}25$	$0{,}23 \pm 0{,}21$	$0{,}22 \pm 0{,}22$
Ранговая регрессия по (3) и (5), g — линейная	$0{,}68 \pm 0{,}23$	$0{,}47 \pm 0{,}25$	$0{,}20 \pm 0{,}20$	$0{,}21 \pm 0{,}20$
Обычная регрессия, g — нейросеть	$0{,}89 \pm 0{,}13$	$0{,}79 \pm 0{,}13$	$0{,}18 \pm 0{,}12$	$0{,}13 \pm 0{,}11$
Ранговая регрессия по (5), g — нейросеть	$\mathbf{0{,}88 \pm 0{,}15}$	$\mathbf{0{,}83 \pm 0{,}28}$	$\mathbf{0{,}05 \pm 0{,}03}$	$\mathbf{0{,}05 \pm 0{,}03}$
Ранговая регрессия по (3), g — нейросеть	$0{,}89 \pm 0{,}13$	$0{,}81 \pm 0{,}16$	$0{,}18 \pm 0{,}17$	$0{,}17 \pm 0{,}17$
Ранговая регрессия по (5) и (3), g — нейросеть	$\mathbf{0{,}88 \pm 0{,}15}$	$\mathbf{0{,}86 \pm 0{,}20}$	$\mathbf{0{,}03 \pm 0{,}02}$	$\mathbf{0{,}04 \pm 0{,}03}$

Примечания: r — коэффициент ранговой корреляции на обучении, r_c — на контроле; CO — стандартное отклонение на обучении, CO_c — на контроле; кросс-валидационный дизайн (10 разбиений, «случай–контроль» 6 : 1). В качестве нейросети использовалась 2-слойная сеть с функцией активации softmax.



(*а*) (*б*)

Прогнозирование EC_{50} с использованием схемы «прямого» прогнозирования ($y = 0{,}67x + 0{,}0706$, $R^2 = 0{,}622$) (*а*) и схемы прогнозирования $E_j(C_i)$ с корректором в виде уравнения Хилла–Ленгмюра ($y = 0{,}5959x + 0{,}018$, $R^2 = 0{,}4027$) (*б*). Приведены результаты для контрольной выборки

5 Заключение

Перспективным направлением повышения точности работы алгоритмов машинного обучения стала разработка математического инструментария, позволяющего порождать проблемно-ориентированные теории для решения конкретных прикладных задач. В частности, топологический подход к распознаванию позволяет систематически анализировать различные способы порождения признаковых описаний плохо формализованных задач распознавания/классификации. Как показали результаты настоящей работы, выбор определенных метрик ρ_L и ρ_A в рамках топологического подхода соответствует порождению специфических критериев рангового характера, оптимизация которых позволяет улучшить показатели обучения соответствующих алгоритмов. С разработанными критериями обучение моделей и оценка качества прогнозирования проводятся на основе сохранения отношений порядка значений, а не самих значений таргетной переменной (что крайне важно, например, с точки зрения теоретической химии). В зависимости от конкретных задач э. ф. р. могут быть представлены в дискретной форме (как в настоящей работе) либо в виде аппроксимаций посредством непрерывных функций. Возможна разработка более сложных ранговых критериев на основе статистических функционалов.

Литература

1. *Журавлёв Ю. И.* Избранные научные труды. — М.: Магистр, 1998. 420 с.
2. *Torshin I. Y., Rudakov K. V.* Combinatorial analysis of the solvability of the problems of recognition, completeness of algorithmic models. Part 1: Factorization approach //

Pattern Recognition Image Analysis, 2017. Vol. 27. No. 1. P. 16–28. doi: 10.1134/S1054661817010151.

3. *Torshin I. Y., Rudakov K. V.* 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 Recognition Image Analysis, 2017. Vol. 27. No. 2. P. 184–199. doi: 10.1134/S1054661817020110.
4. *Torshin I. Y., Rudakov K. V.* On the procedures of generation of numerical features over partitions of sets of objects in the problem of predicting numerical target variables // Pattern Recognition Image Analysis, 2019. Vol. 29. No. 3. P. 654–667. doi: 10.1134/S1054661819040175.
5. *Торшин И. Ю.* О применении топологического подхода к анализу плохо формализуемых задач для построения алгоритмов виртуального скрининга квантово-механических свойств органических молекул I: Основы проблемно ориентированной теории // Информатика и её применения, 2022. Т. 16. Вып. 1. С. 39–45. doi: 10.14357/19922264220106.
6. *Торшин И. Ю.* О применении топологического подхода к анализу плохо формализуемых задач для построения алгоритмов виртуального скрининга квантово-механических свойств органических молекул II: Сопоставление формализма с конструктами квантовой механики и экспериментальная апробация предложенных алгоритмов // Информатика и её применения, 2022. Т. 16. Вып. 2. С. 35–43. doi: 10.14357/19922264220205.
7. *Torshin I. Y., Rudakov K. V.* On the theoretical basis of metric analysis of poorly formalized problems of recognition and classification // Pattern Recognition Image Analysis, 2015. Vol. 25. No. 4. P. 577–587. doi: 10.1134/S1054661815040252.
8. *Koenker R., Bassett G.* Regression quantiles // Econometrica, 1978. Vol. 46. No. 1. P. 33–50. doi: 10.2307/1913643.

Поступила в редакцию 05.10.22

ON OPTIMIZATION PROBLEMS ARISING FROM THE APPLICATION OF TOPOLOGICAL DATA ANALYSIS TO THE SEARCH FOR FORECASTING ALGORITHMS WITH FIXED CORRECTORS

I. Yu. Torshin

Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: Corrective operations (correctors) in multialgorithmic constructions of the algebraic approach can be based on known physical models and/or multilevel descriptions of physical objects. At the same time, within the framework of the topological approach to the analysis of poorly formalized problems, the search for algorithms included in the corrector can be considered as a combinatorial optimization problem or as a problem of minimizing a certain loss function. The study of the neighborhoods of chains in the lattice of subsets of objects made it possible to obtain a number of rank optimization criteria that are promising for solving the problems of predicting numerical target variables. The formalism was tested on the problem of ligand–receptor interaction within the framework of the chemokine analysis of drug molecules (data from ProteomicsDB). The best results of predicting constants were observed when using the obtained rank criteria (correlation coefficient on a sliding control 0.86 ± 0.20 averaging over 300 biological activities).

Keywords: topological data analysis; lattice theory; optimization problems; regression; chemoinformatics

DOI: 10.14357/19922264230201 **EDN:** IGSPEW

Acknowledgments

The research was funded by the Russian Science Foundation, project 23-21-00154. The research was carried out using the infrastructure of the Shared Research Facilities "High Performance Computing and Big Data" (CKP "Informatics") of FRC CSC RAS (Moscow).

References

1. Zhuravlev, Yu. I. 1998. *Izbrannye nauchnye trudy* [Selected scientific works]. Moscow: Magistr. 420 p.
2. Torshin, I. Y., and K. V. Rudakov. 2017. Combinatorial analysis of the solvability of the problems of recognition, completeness of algorithmic models. Part 1: Factorization approach. *Pattern Recognition Image Analysis* 27(1):16–28. doi: 10.1134/S1054661817010151.
3. Torshin, I. Y., and K. V. Rudakov. 2017. 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 Recognition Image Analysis* 27(2):184–199. doi: 10.1134/ S1054661817020110.

4. Torshin, I. Yu., and K. V. Rudakov. 2019. On the procedures of generation of numerical features over partitions of sets of objects in the problem of predicting numerical target variables. *Pattern Recognition Image Analysis* 29(4):654–667. doi: 10.1134/S1054661819040175.
5. Torshin, I. Yu. 2022. O primenenii topologicheskogo podkhoda k analizu plokho formalizuemykh zadach dlya postroeniya algoritmov virtual'nogo skrininga kvantovomekhanicheskikh svoystv organicheskikh molekul I: Osnovy problemno orientirovannoy teorii [On the application of a topological approach to analysis of poorly formalized problems for constructing algorithms for virtual screening of quantum-mechanical properties of organic molecules I: The basics of the problem-oriented theory]. *Informatika i ee Primeneniya — Inform Appl.* 16(1):39–45. doi: 10.14357/19922264220106.
6. Torshin, I. Yu. 2022. O primenenii topologicheskogo podkhoda k analizu plokho formalizuemykh zadach dlya postroeniya algoritmov virtual'nogo skrininga kvantovomekhanicheskikh svoystv organicheskikh molekul II: Sopostavlenie formalizma s konstruktami kvantovoy mekhaniki i eksperimental'naya aprobatsiya predlozhennykh algoritmov [On the application of a topological approach to analysis of poorly formalized problems for constructing algorithms for virtual screening of quantum-mechanical properties of organic molecules II: Comparison of formalism with constructions of quantum mechanics and experimental approbation of the proposed algorithms]. *Informatika i ee Primeneniya — Inform Appl.* 16(2):35–43. doi: 10.14357/19922264220205.
7. Torshin, I. Y., and K. V. Rudakov. 2015. On the theoretical basis of metric analysis of poorly formalized problems of recognition and classification. *Pattern Recognition Image Analysis* 25(4):577–587. doi: 10.1134/ S1054661815040252.
8. Koenker, R., and G. Bassett. 1978. Regression quantiles. *Econometrica* 46(1):33–50. doi: 10.2307/1913643.

Received October 5, 2022

Contributor

Torshin Ivan Y. (b. 1972) — Candidate of Science (PhD) in physics and mathematics, Candidate of Science (PhD) in chemistry, senior scientist, A. A. Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation; tiy135@yahoo.com