ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2022. Т. 16. Вып. 1. С. 40–46

О ПРИМЕНЕНИИ ТОПОЛОГИЧЕСКОГО ПОДХОДА К АНАЛИЗУ ПЛОХО ФОРМАЛИЗУЕМЫХ ЗАДАЧ ДЛЯ ПОСТРОЕНИЯ АЛГОРИТМОВ ВИРТУАЛЬНОГО СКРИНИНГА КВАНТОВО-МЕХАНИЧЕСКИХ СВОЙСТВ ОРГАНИЧЕСКИХ МОЛЕКУЛ. ЧАСТЬ 1: ОСНОВЫ ПРОБЛЕМНО ОРИЕНТИРОВАННОЙ ТЕОРИИ*

И. Ю. Торшин[1]

Аннотация: Топологический подход к анализу плохо формализованных задач и теория хемографов являются расширениями алгебраического подхода к распознаванию, развиваемого в научной школе академика РАН Ю. И. Журавлёва. В первой части статьи предложен проблемно ориентированный формализм для разработки алгоритмов скрининговых оценок квантово-механических (КМ) свойств молекул по их химической структуре. Предложены способы введения метрик на множествах молекул и процедуры порождения «синтетических» признаковых описаний, основанные на согласовании значений «экспертной» метрики на множестве значений свойств молекул со значениями настраиваемой метрики на множестве структур молекул.

Ключевые слова: алгебраический подход; хемоинформатика; размеченные графы; комбинаторный анализ разрешимости

DOI: 10.14357/19922264220106

1 Введение. Проблемная область

Алгебраический подход к анализу проблем распознавания, классификации включает формулировку и исследование критериев разрешимости и регулярности задач [1–3], корректности и полноты моделей алгоритмов [4–6]. Исследуются алгоритмы в виде композиций $A = B \circ C \circ D$, где B — распознающий оператор; C — корректирующая операция (корректор); D — решающее правило. Исходно в рамках алгебраического подхода наибольшее внимание уделяется именно формулировке и исследованию критериев разрешимости и регулярности задач распознавания/классификации, корректности и полноты моделей алгоритмов [1].

Исследование выполнимости этих критериев и «обучение» алгоритмов подразумевает, что задано формальное описание: *матрица информации* (набор признаковых описаний объектов) и *информационная матрица* (отнесение объектов к определенным классам). В то же время во многих прикладных областях (биология, химия, медицина) встречаются задачи, формальные описания которых могут быть получены очень многими способами.

Прикладные задачи, для которых не существует однозначного метода выделения объектов, определения признаковых описаний и классов объектов на основе имеющегося «исходного описания» считаются *плохо формализуемыми* [4]. Адекватная формализация задач позволяет улучшить аккуратность и обобщающую способность соответствующих алгоритмов распознавания. С этой целью в русле алгебраического подхода разрабатываются топологическая и метрическая теории анализа данных [7–9].

К плохо формализуемым относятся и рассматриваемые в настоящей работе задачи типа «структура–свойство», где существенную неоднозначность представляют собой способы порождения признаковых описаний молекул.

В статье представлены результаты совместного применения топологического подхода к плохо формализованным задачам [4] и методов теории анализа хемографов [8, 9] для разработки алгоритмов виртуального скрининга КМ свойств молекул.

*Работа выполнена при поддержке РФФИ (проекты 19-07-00356, 18-07-00944, 20-07-00537).

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

2 Введение. О виртуальных скринингах молекул

Квантовая механика — один из ярких примеров успешного применения теории вероятностей, теории операторов, теории групп и функционального анализа в теоретической физике. Тем не менее точное аналитическое решение уравнения Шредингера имеется только для атома водорода, а высокоточные полуэмпирические схемы КМ-расчетов вычислительно затратны. В то же время задачи поиска перспективных материалов и лекарств подразумевают проведение скринингов пространства всех возможных «малых» органических молекул (более 10^{11} структур [10]).

Для осуществления первоначальных стадий таких виртуальных скринингов необходимы алгоритмы для оценочных расчетов КМ-свойств молекул, позволяющие выделять подмножества молекул с экстремальными значениями КМ-свойств. Практически важны (а) точность, (б) обобщающая способность, (в) производительность вычислений и (г) интерпретируемость таких алгоритмов. В случае оценочных расчетов КМ-свойств молекул необходимо особо акцентировать интерпретируемость таких алгоритмов в терминах теории химической связи на уровне, доступном для химиков-практиков. Одно из возможных решений — алгоритмы, построенные на основании теории хемографов в рамках топологического подхода к анализу данных.

3 Основы топологической теории анализа данных

Алгебраический подход требует определения множества *начальных информаций* (I_i) и множества *конечных информаций* (I_f) [1]. Исследуемые алгоритмы $A(\theta) : \mathrm{I}_i \to \mathrm{I}_f$ (θ — вектор внутренних параметров) «обучаются» на основе конечных множеств прецедентов $\mathrm{Pr} \subset \mathrm{I}_i \times \mathrm{I}_f$. Топологическая теория анализа данных позволяет проводить систематические исследования различных способов определения множеств I_i и I_f и так называемого множества оценок (область значений распознающего оператора B).

Пусть $\mathbf{X} = \{x_1, x_2, \ldots, x_\alpha, \ldots, x_{N_0}\}$ — множество исходных описаний объектов, $\mathbf{X} \subseteq S$, где S — пространство таких описаний. Пусть J_{ob} — пространство допустимых признаковых описаний объектов и определена функция $D : S \to J_{\mathrm{ob}}$. Определим $J_{\mathrm{ob}} \subseteq \mathrm{I}_1 \times \mathrm{I}_2 \times \cdots \times \mathrm{I}_k \times \cdots \times \mathrm{I}_{n+1}$, $\mathrm{I}_i \subseteq \mathrm{I}_1 \times \mathrm{I}_2 \times \cdots \times \mathrm{I}_k \times \cdots \times \mathrm{I}_n$ и $\mathrm{I}_f \subseteq \mathrm{I}_{n+1} \times \mathrm{I}_{n+2} \times \cdots \times \mathrm{I}_{n+l}$ на основе $\mathrm{I}_k = \{\lambda_{k_1}, \lambda_{k_2}, \ldots, \lambda_{k_b}, \ldots, \lambda_{k_{|\mathrm{I}_k|-1}}, \Delta\}$ — множеств значений k-х компонент признакового описания, $k = 1, \ldots, n+l$, где Δ — неопределенность; n — число признаков; l — число таргетных (прогнозируемых) переменных.

Формализация задачи, т. е. переход от множества $\mathbf{X} \subseteq S$ к множеству прецедентов $\mathbf{Q} \subseteq J_{\mathrm{ob}}$, заключается в определении таких функций $\Gamma_k : S \to \mathrm{I}_k$, $k = 1, \ldots, n+l$, что $D(x_\alpha) = (\Gamma_1(x_\alpha) \times \cdots \times \Gamma_k(x_\alpha) \times \cdots \times \Gamma_{n+l}(x_\alpha))_\Delta$ [4, 6], $\mathrm{m}_\alpha = (\Gamma_1(x_\alpha) \times \cdots \times \Gamma_n(x_\alpha))_\Delta$ — *вектор значений признаков*, а $t_\alpha = (\Gamma_{n+1}(x_\alpha) \times \cdots \times \Gamma_{n+l}(x_\alpha))_\Delta$ — *вектор значений l таргетных переменных*, образующих *матрицу информации* $\mathbf{M}_{\mathrm{I}} = (m_\alpha)$, $m_\alpha \in \mathrm{I}_i$, и *информационную матрицу* $\mathbf{M}_{\mathrm{F}} = (t_\alpha)$, $t_\alpha \in \mathrm{I}_f$, соответственно. Функции D соответствует функция $\varphi : 2^S \to 2^{J_{\mathrm{ob}}}$, $\varphi(\mathbf{X}) = \{D(x_\alpha) | x_\alpha \in \mathbf{X}\}$, формирующая множество прецедентов $\mathbf{Q} = \{q_i | q_i = (m_i, t_i)\} \subseteq \mathrm{I}_i \times \mathrm{I}_f$, $q_i[k] \in \mathrm{I}_k$, $i = 1, \ldots, |\mathbf{X}|$, $\mathbf{Q} = \varphi(\mathbf{X})$ [4]. Для Γ_k определена функция полного прообраза значения $\lambda_{k_b} \in \mathrm{I}_k$ в множестве $\mathbf{X}$, $\Gamma_k^{-1}(\lambda_{k_b}) \subseteq \mathbf{X}$, а для функции D — функция полного прообраза объекта, $D^{-1}(q) = \bigcap\limits_{k=1,n} \Gamma_k^{-1}(q[k])$.

Определение 1. *Регулярным будем называть множество прецедентов $\mathbf{Q}$, для которого выполнено* $\forall\limits_{\mathbf{Q}^2}(q_1, q_2) : D^{-1}(q_1) \neq D^{-1}(q_2)$. Если множества $\mathbf{X}$ и $\mathbf{Q}$ изоморфны, так что $\forall\, x \in \mathbf{X} : x = D^{-1}(D(x))$, то $\mathbf{X}$ также регулярно [4]. В дальнейшем рассматриваются только регулярные множества $\mathbf{X}/\mathbf{Q}$.

В подавляющем большинстве реальных задач множества I_k выбираются произвольно. Для получения более информативных «синтетических признаковых описаний» [4] множество $\mathbf{U}(\mathbf{X}) = \{\Gamma_k^{-1}(\lambda_{k_b})\}$ рассматривается как предбаза *топологии* $\mathbf{T}(\mathbf{X}) = \{\emptyset, \mathrm{I}, a \cup b, a \cap b : a, b \in \mathbf{U}(\mathbf{X})\}$, где $\mathrm{I} = \{\mathbf{X}\}$ — единичный элемент. Введение отношения порядка на элементах $\mathbf{T}(\mathbf{X})$ позволяет частично упорядочить элементы топологии $\mathbf{T}(\mathbf{X})$ в *решетку* $\mathbf{L}(\mathbf{T}(\mathbf{X})) = \{a \vee b, a \wedge b : a, b \in \mathbf{T}(\mathbf{X}), (a \geq b) \text{ или } (a \leq b)\}$ так, что $a \leq b \equiv a \subseteq b$ и $a \vee b = \sup(a, b)$ — объединение, а $a \wedge b = \inf(a, b)$ — пересечение множеств a и b. Показано, что при регулярности $\mathbf{X}/\mathbf{Q}$ (определение 1) решетка $\mathbf{L}(\mathbf{T}(\mathbf{X}))$ — булева и элементы $\mathbf{U}(\mathbf{X})$ являются вершинами $\mathbf{L}(\mathbf{T}(\mathbf{X}))$ [4]. При этом возникают три фундаментальные разновидности признаков: *булевы* (вершина решетки $\Gamma_k^{-1}(1)$), *категорные* (антицепи решетки) и *числовые* (цепи решетки) [4].

Таким образом, k-я числовая величина, заданная на $\mathbf{Q}$ посредством I_k, соответствует некоторой цепи $A_k(\mathbf{X})$ в $\mathbf{L}(\mathbf{T}(\mathbf{X}))$, образованной множествами $u(\lambda_{k_b}) = \bigcup_{\beta=1}^{b} \Gamma_k^{-1}(\lambda_{k_\beta})$, $\lambda_{k_{b-1}} \leq \lambda_{k_b} \leq \lambda_{k_{b+1}}$. Каждому значению λ_{k_b} соответствует множество объектов $u(\lambda_{k_b})$ и единственное дополнение $\neg u(\lambda_{k_b})$

(так как решетка булева). Совокупность точек $\mathrm{cdf}\left(A_k(\mathbf{X})\right) = \{(\lambda_{k_b}, |u(\lambda_{k_b})|/N)\}$ представляет *эмпирическую функцию распределения* (э.ф.р.) k-й переменной.

4 Порождение «синтетических» числовых признаков

Булевой решетке $\mathbf{L}(\mathbf{T}(\mathbf{X}))$ сопоставляется *метрическое пространство значений признаков* $\mathbf{M}_{\mathbf{L}}(\mathbf{L}(\mathbf{T}(\mathbf{X})), \rho_{\mathbf{L}})$ с метрикой $\rho_{\mathbf{L}} : \mathbf{L}^2 \to \mathbf{R}^+$. Над $\mathbf{L}(\mathbf{T}(\mathbf{X}))$ также определяется *метрическое пространство объектов* $\mathbf{M}_q[\mathbf{L}(\mathbf{T}(\mathbf{X}))](\mathbf{Q}, \rho_q)$ с метрикой $\rho_q : \mathbf{Q}^2 \to \mathbf{R}^+$. Для определения метрик $\rho_{\mathbf{L}}$ и ρ_q вводятся понятия *изотонной оценки* на решетки и *окрестности* элемента в топологии [4].

Определение 2. *Оценка на решетке* $\mathbf{L}$ *есть такая функция* $v : \mathbf{L} \to \mathbf{R}^+$, *что* $\underset{\mathbf{L}}{\forall} a, b : v[a] + v[b] = v[a \wedge b] + v[a \vee b]$. *Оценка изотонна, если* $\underset{\mathbf{L}}{\forall} a, b : a \subseteq b \Rightarrow v[a] \geq v[b]$ (например, высота элемента $h[x]$). Функция $\rho(x, y) = v[x \vee y] - v[x \wedge y]$, где v изотонна, является метрикой типа $\rho_{\mathbf{L}}$ [4].

Определение 3. *Окрестность* $u(x)$ *точки* $x \in \mathbf{X}$ *в топологии* $\mathbf{T}(\mathbf{X})$ *есть произвольное* $u \in \mathbf{T}(\mathbf{X})$, $x \in \mathbf{u}$; $u(x)$ *разделяет* $x, y \in \mathbf{X}$ *при* $(x \in u) \neq (y \in u)$.

Любой объект $q \in \mathbf{Q}$ представлен в $\mathbf{L}(\mathbf{T}(\mathbf{X}))$ как набор окрестностей $\Gamma_k^{-1}(q[k])$, $k = 1, \ldots, n$, что позволяет определить метрику $\rho_q(q_1, q_2)$ как функцию $f_q : \mathbf{R}^n \to \mathbf{R}^+$, $\rho_q(q_1, q_2) = f_q((\rho_{\mathbf{L}}(\Gamma_k^{-1}(q_1[k]), \Gamma_k^{-1}(q_2[k])))$, $k = 1, \ldots, n$.

Определение 4. *Для аддитивных* f_q *расстояние* $\rho_q(q_1, q_2)$ *между объектами* q_1 *и* q_2 *определяется посредством линейной комбинации* $S\left(\sum_{k=\overline{1,n}} \omega_k s\left(\rho_{\mathbf{L}}\left(\Gamma_k^{-1}(q_1[k]), \Gamma_k^{-1}(q_2[k])\right)\right)\right)$, *где* $S, s : \mathbf{R} \to \mathbf{R}^+$ — *произвольные «сглаживающие» функции, а* ω_k — *вес* k*-го признакового описания.*

Если $S = 1$, $s = 1/|\mathbf{Q}|$, $\rho_{\mathbf{L}}(x, y) = (x \neq y)$, $(\omega_k = 1)$, а Γ_k^{-1} и Γ_k определены для $\mathrm{I}_k = [0, 1]$, то ρ_q — метрика Хэмминга. При $S = \sqrt[p]{}$, $s(x) = x^p$, $\rho_{\mathbf{L}}(x, y) = |x \Delta y|$, $\mathrm{I}_k \subset \mathbf{R}$, ρ_q — взвешенная метрика Минковского и т. д. В рамках топологического подхода порождение синтетических признаков осуществляется посредством метрики $\rho_{\mathbf{L}}$ либо метрики ρ_q (последнее требует определения $\rho_{\mathbf{L}}$).

5 Синтетические признаки на основании метрики $\rho_{\mathbf{L}}$

Определение 5. *Расстояние* $\rho_A : \mathbf{A}(\mathbf{X})^2 \to \mathbf{R}^+$ *между цепями* $a = \langle a_1, \ldots, a_i, \ldots \mathrm{I}\rangle$ *и* $b = \langle b_1, \ldots, b_j, \ldots \mathrm{I}\rangle$ *есть сумма расстояний между соответствующими элементами*:

$$\rho_{\mathbf{A}}(a, b) = \min\left(\sum_{i=1,|a|} \rho_{\mathbf{L}}\left(a_i, \underset{b_j \in b}{\arg\min}\, \rho_{\mathbf{L}}\left(a_i, b_j\right)\right), \sum_{i=1,|b|} \rho_{\mathbf{L}}\left(b_j, \underset{a_i \in a}{\arg\min}\, \rho_{\mathbf{L}}\left(b_j, a_i\right)\right)\right).$$

Пусть $\mathbf{A}(\mathbf{X})$ — множество всех цепей $\mathbf{L}(\mathbf{T}(\mathbf{X}))$. Алгоритмы прогнозирования k-й переменной соответствуют цепям в $\mathbf{A}(\mathbf{X})$, наиболее близким к цепи $\mathbf{A}_k(\mathbf{X})$. Пусть $\mathbf{A}(\mathbf{X})_{1,n} \subset \mathbf{A}(\mathbf{X})$ соответствует цепям над элементами $\mathbf{U}(\mathbf{X})$, $k = 1, \ldots, n$. Тогда искомые алгоритмы соответствуют решению задачи

$$aa = \underset{a \in \mathbf{A}(\mathbf{X})_{i,n}}{\arg\min}\ \rho_{\mathbf{A}}(A_k(\mathbf{X}), a)\,. \tag{1}$$

6 Синтетические признаки на основании метрики ρ_q

В рамках данного подхода согласовываются значения настраиваемой «признаковой» метрики ρ_q с вектором весов (ω_k)

$$\underset{\{\omega_k\}}{\arg\min} \sum_{m=1}^{|\mathbf{X}|} \sum_{j \neq m}^{|\mathbf{X}|} \mathbf{L}_{\mathrm{F}}\left(\rho_q\left((\omega_k), \mathrm{X}_m, \mathrm{X}_j\right) - \rho_e(\mathrm{X}_m, \mathrm{X}_j)\right), \tag{2}$$

где $\mathbf{L}_{\mathrm{F}}$ — та или иная функция потерь. Например, в настоящей работе в качестве $\mathbf{L}_{\mathrm{F}}$ используется функция модуля $\mathbf{L}_{\mathrm{F}}(x) = |x|$. Практически важным случаем экспертной метрики ρ_e является «скалярная» метрика (например, модуль разности значений).

Теорема 1. *При использовании скалярной метрики* ρ_e *условие* (1) *для метрики Хэмминга* ρ_q *соответствует аддитивной схеме учета признаков.*

Доказательство. Для ρ_e в виде модуля разности выполнены все три аксиомы метрики (так как они выполнены для любых трех коллинеарных точек). Пусть нулевой элемент входит во все множества I_k, так что можно определить расстояние от нулевого элемента до любого другого элемента множества I_k посредством экспертной метрики ρ_e. В ρ-конфигурации объектов, образованной метрикой ρ_q, за нулевую точку с номером m_0 можно, вообще говоря, принять любой объект (например, центральный объект, соответствующий минимальной сумме расстояний до всех остальных). Пусть в m-сумме в выражении (2) одна из j-сумм соответствует нулевому элементу $\mathbf{X}_{m_0}$.

Использование скалярной метрики ρ_e в условии (2) соответствует проекции $|\mathbf{X}|$-мерной ρ_q-конфигурации на одномерную числовую ось значений I_k. Любая из j-сумм в (1) включает все объекты из $\mathbf{X}$ и покрывает все множество значений I_k. Разница между любыми двумя значениями из I_k (т. е. $\rho_e(\mathbf{X}_m, \mathbf{X}_j)$) также является отрезком на числовой оси k-й переменной, соответствующим сдвигу значений в множестве I_k на константу λ_{k_β}.

Так как каждая из j-сумм в (2) соответствует одной и той же (с точностью до константы) задаче согласования метрик, то задача (2) с $\mathbf{L}_\mathrm{F}(x) = |x|$ может быть переформулирована через расстояния от нулевого элемента, т. е. произведен переход от оценки попарных расстояний к суммированию по всем объектам:

$$\arg\min_{\{\omega_k\}} \sum_{m \neq m_0}^{|\mathbf{X}|} |\rho_q(\{\omega_k\}, \mathbf{X}_{m_0}, \mathbf{X}_m) - \mathbf{T}_m|\,, \qquad (3)$$

где $\mathbf{T}_m = \rho_e(\mathbf{X}_{m_0}, \mathbf{X}_m)$ — значение прогнозируемой k-й числовой переменной для объекта $\mathbf{X}_m$ (так как, по построению, точка $\mathbf{X}_{m_0}$ соответствует нулевому значению). Если ρ_q определена как метрика Хэмминга (см. комментарий к определению 4), то задача в постановке (3) соответствует представлению ρ_q в виде линейной формы, т. е. аддитивной схеме учета признаков. Теорема доказана.

Таким образом, в случае скалярных ρ_e расстояния ρ_q служат своего рода «синтетическими» числовыми признаками объектов в множествах $\mathbf{X}/\mathbf{Q}$. Для практического применения (2) необходимо определить функции Γ_k, а для определения метрических пространств $\mathbf{M}_\mathbf{L}$ и $\mathbf{M}_q$ — метрики $\rho_\mathbf{L}$ и ρ_q. В случае задач типа «структура–свойство» молекул для этого применяется теория хемографов.

7 О теории анализа размеченных графов

Для порождения признаковых описаний молекулярных структур в рамках теории анализа размеченных графов вводится понятие хемогра́фа [8]. Признаковые описания порождаются над разметками хемографов [9].

Определение 6. *Граф* $G = (\mathbf{V}, \mathbf{E})$ *— совокупность множества вершин* $\mathbf{V} = \mathbf{V}(G)$ *и множества ребер* $\mathbf{E} = \mathbf{E}(G)$, $\mathbf{E} \subset \mathbf{V}^2$. *Хемограф (*$\chi$*-граф) — конечный, связный, неориентированный, размеченный граф без петель, с кликовым числом не более* 3. Множество $\mathbf{\Gamma} = \{(\mathbf{V}, \mathbf{E}) | \mathbf{V} \subset \mathbf{N}, \mathbf{E} \subset \mathbf{N}^2\}$, где $\mathbf{N}$ — натуральный ряд, есть *множество всех графов*.

Множество вершин χ-графа $\mathbf{X}$ соответствует множеству атомов молекулы, а множество ребер — множеству химических связей молекулы. Хемографы строятся на основе «внешних» (декартовых) координат ядер атомов $\mathbf{R}(X) = \{\vec{R}_i(X) | \vec{R}_i \in \mathbf{R}^3\}$, $i = 1, \ldots, |\mathbf{V}(\mathrm{X})|$, или «внутренних» координат (межатомные расстояния), так что χ-графу X сопоставлен ряд матриц $\mathbf{M}(X) = (m_{ij}(X))$, $m_{ij} \in \mathbf{R}$, $i, j = 1, \ldots, |\mathbf{V}(X)|$ (матрица межатомных расстояний, матрица смежности и др.). Химические формулы в различных «машинных» форматах (SMILES, XYZ и др.) суть упрощенные представления матриц $\mathbf{M}(X)$ или координат $\mathbf{R}(X)$.

Признаковые описания χ-графов вводятся на основании анализа специальной разновидности подграфов χ-графов — цепей и узлов, для описания которых в работах [8, 9] вводится комплекс понятий и соответствующих обозначений: *множество всех замкнутых подграфов* χ*-графа* X, $\mathbf{\Pi}(X)$ (включает *множество связных подграфов* $\mathbf{S}(X)$, множество *цепей* $\mathbf{C}(X)$, *множество* χ*-узлов* $\mathbf{K}(X)$), *операции объединения/пересечения множества подграфов* ($\breve{\Pi} = \bigcup_{i=1}^{|\Pi|} \pi_i = (\bigcup_{i=1}^{|\Pi|} \mathbf{v}_i, \breve{\bigcup}_{i=1}^{|\Pi|} \mathbf{e}_i)$; $\widehat{\Pi} = \bigcap_{i=1}^{|\Pi|} \pi_i$), *условие образования* графа $\breve{O} = X$ множеством $\mathbf{O} \subseteq \mathbf{\Pi}(X)$, оператор $\hat{c}^n$ вычисления *цепей длины* n, *множество смежности* (*окружение*) вершины v: $\Gamma(v) = \hat{v}\hat{e}v$, *узел* вершины v, $(\Gamma(v), \hat{e}v)$ *и др.* [8]. Такие подграфы необходимы для исследования изоморфизма χ-графов.

Определение 7. *Графы* G_1 *и* G_2 *изоморфны (*$G_1 \simeq G_2$*) если существует взаимно однозначное соответствие между их вершинами и ребрами.* Изоморфизм $G_1 \simeq G_2$ соответствует существованию функции перенумерации $\mu_\mathbf{I} : \mathbf{N} \to \mathbf{N}$, так что из $G_2 = \mu_\mathrm{I}(G_1)$ следует $G_1 \simeq G_2$. Определяется класс $\mathbf{I}(G) = \{g \in \mathbf{\Gamma} | \exists\, \mu_\mathbf{I} : G = \mu_\mathbf{I}(g)\}$.

Определение 8. *Пусть задан алфавит меток* $\mathbf{Y} = \{\upsilon_1, \upsilon_2, \ldots, \upsilon_{n(Y)}\}$. *Функция разметки* $\mu_\mathbf{V} : \mathbf{V} \to \mathbf{Y}$ *сопоставляет метку каждой вершине* χ*-графа.*

Для комбинаторного анализа свойств изоморфизма графов и для порождения признаковых описаний χ-графов в теории анализа размеченных графов вводятся *множество* χ*-цепей* $\tilde{\mathbf{Y}}$ над алфавитом $\mathbf{Y}$ (*в том числе подмножества всех* χ*-цепей длины* n, $\tilde{\mathbf{Y}}^n \subset \tilde{\mathbf{Y}}$), множество χ*-узлов* $\hat{\mathbf{Y}} = \bigcup_{k=2} \hat{\mathbf{Y}}(k)$, $\hat{\mathbf{Y}}(k) = \{\mathbf{Y} \times 2^{\mathbf{Y}^k}\}$, понятия *инварианта графа* ($\iota : \mathbf{\Gamma} \to \mathbf{R}^n, n \in \mathbf{N} : \forall\, a \in \mathbf{\Gamma} : b \in \mathbf{I}(a) \Rightarrow \iota(b) = \iota(a)$), *кортеж-инварианта* $\iota : \mathbf{\Gamma} \to \mathbf{R}^n$, $n \geq 2$, *полноты инварианта* $\forall\, a \in \mathbf{\Gamma} : b \in \mathbf{I}(G) \Leftrightarrow \iota(a) = \iota(b)$, *изомерных по инварианту* ι *графов*, *функции разметки* χ*-цепей* $\mu_c : \mathbf{C} \to \tilde{\mathbf{Y}}$ и χ*-узлов* $\mu_\kappa : \mathbf{K} \to \hat{\mathbf{Y}}$, χ*-фрагментов* как элементов множеств $\hat{\mu}_c^{-1}\alpha$ и $\hat{\mu}_\kappa^{-1}\kappa$, *оператора вхождения множества подграфов* $\boldsymbol{\pi} \subset \mathbf{\Gamma}$ в χ-граф $\mathbf{X}$ $\hat{\beta}[\mathbf{X}] : 2^\mathbf{\Gamma} \to \{0, 1\}$, *опе-*

ратора числа вхождений множества подграфов в $\mathbf{X}$ $\hat{\eta}[\mathbf{X}] : 2^{\mathbf{\Gamma}} \to \mathbf{N}$ и др. [9].

Эти понятия позволяют проводить комбинаторный анализ изоформизма χ-графов и порождать признаковые описания χ-графов как $\hat{\eta}[\mathbf{X}]\hat{\mu}_c^{-1}\alpha$, $\hat{\eta}[\mathbf{X}]\hat{\mu}_\kappa^{-1}\kappa$, $\hat{\beta}[\mathbf{X}]\hat{\mu}_c^{-1}\alpha$, $\hat{\beta}[\mathbf{X}]\hat{\mu}_\kappa^{-1}\kappa$ (лемма 1 в работе [7]). Результат последовательного применения $\hat{\mu}_c^{-1}$ к $\boldsymbol{\alpha} = \{\alpha \in \tilde{\mathbf{Y}}\}$ обозначим $\hat{\boldsymbol{\mu}}_c^{-1}\boldsymbol{\alpha} = \{\hat{\mu}_c^{-1}\alpha, \alpha \in \boldsymbol{\alpha}\}$, оператора $\hat{\mu}_\kappa^{-1}$ ко множеству χ-узлов $\boldsymbol{\kappa} = \{\kappa \in \hat{\mathbf{Y}}\} - \hat{\boldsymbol{\mu}}_\kappa^{-1}\boldsymbol{\kappa} = \{\hat{\mu}_\kappa^{-1}\kappa, \kappa \in \boldsymbol{\kappa}\}$, оператора $\hat{\beta}$ ко множеству $\tilde{\boldsymbol{\pi}} = \{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_n\}$, $\tilde{\boldsymbol{\pi}} \subset \mathbf{\Gamma}$ обозначим $\hat{\boldsymbol{\beta}}\tilde{\boldsymbol{\pi}} = \{\hat{\beta}\boldsymbol{\pi}_1, \hat{\beta}\boldsymbol{\pi}_2, \ldots, \hat{\beta}\boldsymbol{\pi}_n\}$.

Комбинаторный анализ разрешимости задачи анализа изоморфизма χ-графов сводится к установлению локальной полноты некоторых кортеж-инвариантов [9]. Пусть $\iota : \mathbf{\Gamma} \to \mathbf{R}^n$, $n \geq 2$, — кортеж-инвариант, построенный над некоторым *множеством из* n *элементарных инвариантов* $\boldsymbol{\iota}_e \subset \mathbf{E}$. Пусть для графа G выражение $\boldsymbol{\iota}_i(G) = \{\iota_i(G), i = 1, \ldots, n\}$ будет означать множество значений инвариантов из $\boldsymbol{\iota}_e$. Определим *функцию нумерации элементарных инвариантов* $\lambda : \boldsymbol{\iota}_e \to \mathbf{N}$ и *оператор формирования кортеж-инварианта*. *Оператор формирования кортеж-инварианта* $\hat{\iota} : 2^{\mathbf{E}} \to \mathbf{R}^n$ по заданному $\boldsymbol{\iota}_e$ определим как $\hat{\iota}\boldsymbol{\iota}_e = (\iota_j, \iota_k, \ldots, \iota_l)$, $\iota_j, \iota_k, \ldots, \iota_l \in \boldsymbol{\iota}_e$, $\lambda(\iota_j) = 1 \leq \lambda(\iota_k) < \cdots < \lambda(\iota_l) = n$, а значение i-го элемента кортежа $\hat{\iota}\boldsymbol{\iota}_e$ обозначим $\hat{\iota}[i]\boldsymbol{\iota}_e(G) = \iota(G)|\lambda(\iota) = i$. В теореме 2 показана взаимосвязь между изоморфизмом χ-графов и полнотой кортеж-инвариантов.

Теорема 2. $\forall a, b \in \mathbf{\Gamma} : |\mathbf{I}(a) \cap \mathbf{I}(b)| > 0 \Leftrightarrow \Leftrightarrow \underset{i=\overline{1,n}}{\exists} i : \hat{\iota}[i]\boldsymbol{\iota}_e(a) \neq \hat{\iota}[i]\boldsymbol{\iota}_e(b)$, *так что* $\hat{\iota}\boldsymbol{\iota}_e$ *— полный кортеж-инвариант*.

Доказательство приведено в работе [7]. Следствия теоремы важны для анализа множества прецедентов Pr.

Следствие 1. Полные инварианты могут быть образованы над подмножествами множеств $\tilde{\mathbf{Y}}$ и $\hat{\mathbf{Y}}$, если для каждого χ-графа $\mathbf{X}$ множества прецедентов Pr существуют $\boldsymbol{\kappa}'(\mathbf{X}) = \{\kappa \in \boldsymbol{\kappa}|\hat{\eta}[\mathbf{X}]\hat{\mu}_\kappa^{-1}\kappa = 1\}$ и $\boldsymbol{\alpha}'(\mathbf{X}) = \{\alpha \in \boldsymbol{\alpha}'|\hat{\eta}[\mathbf{X}]\hat{\mu}_c^{-1}\alpha = 1\}$ такие, что при $\boldsymbol{\pi}'(\mathbf{X}) = \hat{\boldsymbol{\mu}}_\kappa^{-1}\boldsymbol{\kappa}'(\mathbf{X}) \cup \hat{\boldsymbol{\mu}}_c^{-1}\boldsymbol{\alpha}'(\mathbf{X})$ $\breve{\boldsymbol{\pi}}(\mathbf{X}) = \mathbf{X}$.

Следствие 2. Пусть в конечном $\mathrm{Pr} \subset \mathbf{\Gamma}$ каждый граф G помечен меткой изоморфности $\mathrm{iso}(G) : \mathbf{I}(\mathbf{\Gamma}) \to \mathbf{N}$. Разрешимость

$$\underset{a,b \in \mathrm{Pr}}{\forall} \mathrm{iso}(a) \neq \mathrm{iso}(b) \Rightarrow \iota(a) \neq \iota(b)$$

эквивалентна полноте

$$\underset{a,b \in \mathrm{Pr}}{\forall} \mathrm{iso}(a) \neq \mathrm{iso}(b) \Rightarrow \underset{i=\overline{1,|\chi|}}{\exists} i : \hat{\iota}[i]\chi(a) \neq \hat{\iota}[i]\chi(b).$$

Следствие 3. Задача распознавания изоморфных графов разрешима тогда и только тогда, когда $\sum_{a \in \mathrm{Pr}} |\mathbf{i}\boldsymbol{\mu}(a, \iota, \mathrm{Pr}) \backslash \mathbf{i}(a, \mathrm{Pr})| = 0$, где $\mathbf{i}\boldsymbol{\mu}(G, \iota, \mathrm{Pr})$ — множество изомерных G графов; $\mathbf{i}(G, \mathrm{Pr})$ — множество изоморфных G графов.

Следствие 4. Инвариант ι полон при $r_\iota(\iota, \mathrm{Pr}) = 1 - (1/|\mathrm{Pr}|^2) \sum_{a \in \mathrm{Pr}} \mathbf{i}\boldsymbol{\mu}(a, \iota, \mathrm{Pr}) \backslash \mathbf{i}(a, \mathrm{Pr})| = 1$.

Следствие 5. Пусть $\phi(\hat{\iota}\boldsymbol{\chi}, i, \mathrm{Pr}) = \{(a, b)|a, b \in \mathrm{Pr}, \hat{\iota}[i]\boldsymbol{\chi}(a) \neq \hat{\iota}[i]\boldsymbol{\chi}(b)\}$ и $\phi(\hat{\iota}\boldsymbol{\chi}, i, \mathrm{Pr}) \cap \phi(\hat{\iota}\boldsymbol{\chi}, j, \mathrm{Pr}) = \emptyset$. Тогда $r_\iota(\iota, \mathrm{Pr}) = \sum_{i=1}^{|\chi|} \varphi_\iota(\hat{\iota}\boldsymbol{\chi}, i, \mathrm{Pr})$, где $\varphi_\iota(\hat{\iota}\boldsymbol{\chi}, i, \mathrm{Pr}) = |\phi(\hat{\iota}\boldsymbol{\chi}, i, \mathrm{Pr})| \backslash |\mathrm{Pr}|^2$.

8 Топологическая теория анализа χ-графов

На основании алфавита $\mathbf{Y}$ строятся различные множества χ-цепей длины m, $\tilde{\mathbf{Y}}^m$, $m = 1, \ldots, m_{\max}$, множество всех χ-узлов $\hat{\mathbf{Y}}$ и затем соответствующие элементарные χ- и кортеж-инварианты. Для χ-графа $X \in \mathbf{X}$ вычисляются множество всех χ-цепей X, $\tilde{\mathbf{Y}}(X)$, и множество всех χ-узлов X, $\hat{\mathbf{Y}}(X)$. Для любых $\alpha \in \tilde{\mathbf{Y}}(X)$, $\kappa \in \hat{\mathbf{Y}}(X)$ пусть $\hat{\beta}[X]\hat{\mu}_c^{-1}\alpha$, $\hat{\beta}[\mathrm{X}]\hat{\nu}_\kappa^{-1}\kappa$ — булевы инварианты. Тогда для выбранных $\boldsymbol{\alpha} \subseteq \tilde{\mathbf{Y}}$ и $\boldsymbol{\kappa} \subseteq \hat{\mathbf{Y}}$ таких, что $|\boldsymbol{\alpha}| + |\boldsymbol{\kappa}| = n$, вычисляются множества $\hat{\boldsymbol{\mu}}_c^{-1}\boldsymbol{\alpha}$ и $\hat{\boldsymbol{\mu}}_\kappa^{-1}\boldsymbol{\kappa}$. Если множества $\mathbf{I}_k$ определены как $[0, 1]$, то множество элементарных инвариантов $\boldsymbol{\iota}_e = \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\mu}}_c^{-1}\boldsymbol{\alpha} \cup \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\mu}}_\kappa^{-1}\boldsymbol{\kappa}$, а если $\mathbf{I}_k \subset \mathbf{N}$, то $\boldsymbol{\iota}_e = \hat{\boldsymbol{\eta}}\hat{\boldsymbol{\mu}}_c^{-1}\boldsymbol{\alpha} \cup \hat{\boldsymbol{\eta}}\hat{\boldsymbol{\mu}}_\kappa^{-1}\boldsymbol{\kappa}$. Определив функцию нумерации инвариантов $\lambda : \boldsymbol{\iota}_e \to \mathbf{N}$, определяем функцию Γ_k как $\Gamma_k(X) = \hat{\iota}[k]\boldsymbol{\iota}_e(X)$, а функцию D — как $D(X) = \hat{\iota}\boldsymbol{\iota}_e(\mathrm{X})$.

Таким образом, предбаза $\mathbf{U}(\mathbf{X})$ топологии $\mathbf{T}(\mathbf{X})$ состоит из подмножеств χ-графов, соответствующих χ-цепям или χ-узлам. Для $D(\mathbf{X}) = \hat{\iota}\boldsymbol{\iota}_e(\mathbf{X})$ задача (2) реализуется как *хемометрический анализ* [7]:

$$\arg\min_{\{\omega_k\}} \sum_{m_0=1}^{|\mathrm{X}|} \sum_{m \neq m_0}^{|\mathrm{X}|} \left| \sum_{k=1}^{n} \omega_k \hat{\iota}[k]\hat{\boldsymbol{\beta}}[X_{m_0}]\boldsymbol{\pi} \oplus \right.$$
$$\left. \oplus\, \hat{\iota}[k]\hat{\boldsymbol{\beta}}[X_m]\boldsymbol{\pi} - |T_m - T_{m_0}| \right|,$$

где $\boldsymbol{\pi} \subseteq \hat{\boldsymbol{\mu}}_c^{-1}\boldsymbol{\alpha} \cup \hat{\boldsymbol{\mu}}_\kappa^{-1}\boldsymbol{\kappa}$ — некоторое «опорное» множество подграфов; T_m — оцениваемое свойство m-й молекулы, а выражение под знаком суммы по k соответствует ρ_q-метрике Хэмминга. Применяя теорему 1 и «сглаживающую» функцию $S : \mathbf{R} \to \mathbf{R}^+$ (определение 4), получаем:

$$\arg\min_{\{\omega_k\}} \sum_{m=1}^{|\mathrm{X}|} \left| S\left(\sum_{k=1}^{n} \omega_k s\left(\hat{\iota}[k]\hat{\boldsymbol{\beta}}[X_m]\boldsymbol{\pi} \right) \right) - T_m \right|.$$

9 Заключение

Для построения проблемно ориентированных метрик и алгоритмов прогнозирования КМ свойств

молекул по их химической структуре разработанные признаковые описания хемографов анализируются посредством теории топологического анализа данных. В результате анализа топологии $\mathbf{T}(\mathbf{X})$ и решетки $\mathbf{L}(\mathbf{T}(\mathbf{X}))$ с использованием условий (1) и (2) могут быть получены различные алгоритмы прогнозирования произвольной числовой переменной на основе структуры хемографов. Эти алгоритмы одновременно являются алгоритмами порождения «синтетических» числовых признаков, информативных относительно этой переменной. Такие «синтетические» признаки могут использоваться в алгоритмических конструкциях алгебраического подхода и в обычных методах машинного обучения (регрессия, нейронные сети, метрические методы и др.). Сопоставление полученных моделей алгоритмов с формализмом квантовой механики, равно как и экспериментальная апробация соответствующих алгоритмов, будут представлены во второй части статьи.

Литература

1. *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики, 1978. № 33. С. 5–68.
2. *Рудаков К. В., Торшин И. Ю.* Вопросы разрешимости задачи распознавания вторичной структуры белка // Информатика и её применения, 2010. Т. 4. Вып. 2. С. 25–35.
3. *Рудаков К. В., Торшин И. Ю.* Анализ информативности мотивов на основе критерия разрешимости в задаче распознавания вторичной структуры белка // Информатика и её применения, 2011. Т. 5. Вып. 4. С. 40–50.
4. *Torshin I. Yu., 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.
5. *Torshin I. Yu., 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.
6. *Torshin I. Yu., 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. 4. P. 654–667. doi: 10.1134/S1054661819040175.
7. *Ruddigkeit L., van Deursen R., Blum L., Reymond J.* Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17 // J. Chem. Inf. Model., 2012. Vol. 52. No. 11. P. 2864–2875. doi: 10.1021/ci300415d.
8. *Torshin I. Yu., Rudakov K. V.* On the application of the combinatorial theory of solvability to the analysis of chemographs. Part 1: Fundamentals of modern chemical bonding theory and the concept of the chemograph // Pattern Recognition Image Analysis, 2014. Vol. 24. No. 1. P. 11–23.
9. *Torshin I. Yu., Rudakov K. V.* On the application of the combinatorial theory of solvability to the analysis of chemographs. Part 2: Local completeness of invariants of chemographs in view of the combinatorial theory of solvability // Pattern Recognition Image Analysis, 2014. Vol. 24. No. 2. P. 196–208.
10. *Torshin I. Yu., Rudakov K. V.* Topological data analysis in materials science: The case of high-temperature cuprate superconductors // Pattern Recognition Image Analysis, 2020. Vol. 30. No. 2. P. 262–274. doi: 10.1134/S1054661820020157.

Поступила в редакцию 30.03.21

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. PART 1: THE BASICS OF THE PROBLEM-ORIENTED THEORY

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: The topological approach to the analysis of poorly formalized problems and the theory of chemographs are extensions of Zhuravlev’s algebraic approach to recognition. In the first part of the article, a problem-oriented

formalism is proposed aimed at development of algorithms for screening assessments of the quantum-mechanical properties of molecules on the basis of their chemical structure. Methods for introducing metrics on sets of molecules and procedures for generating "synthetic" feature descriptions are proposed. The latter are generated by matching the values of some "expert" metric on the set of molecular properties to a tunable metric on the set of molecular structures.

Keywords: algebraic approach; chemoinformatics; labeled graphs; combinatorial solvability analysis

DOI: 10.14357/19922264220106

Acknowledgments

This work was supported in part by RFBR grants 19-07-00356, 18-07-00944, 20-07-00537.

References

1. Zhuravlev, Yu. I. 1978. Ob algebraicheskom podkhode k resheniyu zadach raspoznavaniya ili klassifikatsii [On algebraic approach to recognition and classification problems]. *Problemy kibernetiki* [Cybernetic Problems] 33:5–68.
2. Rudakov, K. V., and I. Yu. Torshin. 2010. Voprosy razreshimosti zadachi raspoznavaniya vtorichnoy struktury belka [Questions of solvability of the problem of recognition of the secondary structure of a protein]. *Informatika i ee Primeneniya — Inform Appl.* 4(2):25–35.
3. Rudakov, K. V., and I. Yu. Torshin. 2011. Analiz informativnosti motivov na osnove kriteriya razreshimosti v zadache raspoznavaniya vtorichnoy struktury belka [Analysis of the informativeness of motives based on the criterion of solvability in the problem of recognizing the secondary structure of a protein]. *Informatika i ee Primeneniya — Inform Appl.* 5(4):40–50.
4. Torshin, I. Yu., 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.
5. Torshin, I. Yu., 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.
6. 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.
7. Ruddigkeit, L., R. van Deursen, L. Blum, and J. Reymond. 2012. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* 52(11):2864–2875. doi: 10.1021/ci300415d.
8. Torshin, I. Yu., and K. V. Rudakov. 2014. On the application of the combinatorial theory of solvability to the analysis of chemographs. Part 1: Fundamentals of modern chemical bonding theory and the concept of the chemograph. *Pattern Recognition Image Analysis* 24(1):11–23.
9. Torshin, I. Yu., and K. V. Rudakov. 2014. On the application of the combinatorial theory of solvability to the analysis of chemographs. Part 2: Local completeness of invariants of chemographs in view of the combinatorial theory of solvability. *Pattern Recognition Image Analysis* 24(2):196–208.
10. Torshin, I. Yu., and K. V. Rudakov. 2020. Topological data analysis in materials science: The case of high-temperature cuprate superconductors. *Pattern Recognition Image Analysis* 30(2):262–274. doi: 10.1134/S1054661820020157.

Received March 30, 2021

Contributor

Torshin Ivan Yu. (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