ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2024. Т. 18. Вып. 1. С. 85–91

О ПОРОЖДЕНИИ СИНТЕТИЧЕСКИХ ПРИЗНАКОВ НА ОСНОВЕ ОПОРНЫХ ЦЕПЕЙ И ПРОИЗВОЛЬНЫХ МЕТРИК В РАМКАХ ТОПОЛОГИЧЕСКОГО ПОДХОДА К АНАЛИЗУ ДАННЫХ. ЧАСТЬ 1. ВКЛЮЧЕНИЕ В ФОРМАЛИЗМ ЭМПИРИЧЕСКИХ ФУНКЦИЙ РАССТОЯНИЯ*

И. Ю. Торшин[1]

Аннотация: Анализ формализма топологической теории распознавания на основе фундаментальных понятий функционального анализа позволил предложить ранее не исследованные подходы к определению решеточных оценок. В частности, использование опорных цепей для анализа булевых решеток, формируемых над регулярными по Журавлёву множествами прецедентов, указало на новое направление исследований, заключающееся в замене оценок элементов решеток на определенного рода функции и/или векторы. Данное расширение формализма также позволяет проводить систематическое исследование известных в литературе полуэмпирических функционалов расстояния для решения прикладных задач. Обоснованы перспективные направления дальнейшего развития формализма, включающие введение функционалов, редуцирующих описания множеств булевой решетки к скалярным оценкам, и развитие математического аппарата для анализа решеток, в котором вместо оценок фигурируют операции над соответствующими функциями. Последнее направление интересно тем, что позволяет вводить расстояния на решетке без использования оценок.

Ключевые слова: топологический анализ данных; теория решеток; алгебраический подход Ю. И. Журавлёва; функциональный анализ

DOI: 10.14357/19922264240110 **EDN:** RIVOXR

1 Введение

Топологическая теория распознавания стала развитием алгебраического подхода к распознаванию, предложенного научной школой Ю. И. Журавлёва, и предназначена для решения плохо формализованных задач распознавания, классификации и прогнозирования [1]. Одна из основных целей данной теории — разработка методов систематического порождения и отбора синтетических признаковых описаний объектов, которые бы характеризовались большей информативностью, чем исходные признаки [2]. В работе [3] было показано, что веса таких признаков можно эффективно настраивать посредством ранговой оптимизации; разработан формализм для порождения признаковых описаний на основе параметризируемых решеточных оценок [4].

В рамках развиваемого формализма каждый объект x из множества исходных описаний N_0 объектов $\mathbf{X} = \{x_1, \ldots, x_{N_0}\}$, $\mathbf{X} \subseteq \mathrm{S}$, описываемый n признаками посредством функций $\Gamma_k : \mathrm{S} \to \mathrm{I}_k$ (где $\mathrm{I}_k = \{\lambda_{k_1}, \lambda_{k_2}, \ldots\}$ — множества значений признаковых описаний), представлен множествами $\{\Gamma_k^{-1}(\Gamma_k(x))\}$, $k = \overline{1, n+l}$, где l — число таргетных (прогнозируемых) переменных. Значение t-й целевой (таргетной) переменной объекта x, $t = \overline{n+1, n+l}$, вычисляется как $\Gamma_t(x)$. Множество прецедентов над пространством допустимых признаковых описаний объектов J_{ob} определяется как

$$Q = \varphi(\mathbf{X}) = \{D(\mathrm{x}_\alpha) | \mathrm{x}_\alpha \in \mathbf{X}\}$$

посредством $D : \mathrm{S} \to J_{\mathrm{ob}}$ и $\varphi(\mathbf{X}) = \{D(\mathrm{x}_\alpha) | \mathrm{x}_\alpha \in \mathbf{X}\}$, $D(\mathrm{x}_\alpha) = (\Gamma_1(\mathrm{x}_\alpha \times \cdots \times \Gamma_k(\mathrm{x}_\alpha) \times \cdots \times \Gamma_{n+1}(\mathrm{x}_\alpha))_\Delta$. При регулярности множеств $\mathbf{X}/Q$ ($\forall x \in \mathbf{X}, x = D^{-1}(D(\mathrm{x})))\mathbf{X}$ изоморфно Q и обоим множествам сопоставлены *топология* $T(\mathbf{X}) = \{\emptyset, \{\mathbf{X}\}, a \cup b, a \cap b : a, b \in U(\mathbf{X}) = \{\Gamma_k^{-1}(\lambda_{k_b})\}\}$ и *булева решетка* $L(T(\mathbf{X})) = \{a \vee b, a \wedge b : a, b \in T(\mathbf{X})\}$ [5].

Рассмотрим топологический подход к распознаванию с точки зрения фундаментальных понятий теории функций — рефлексивных и транзитивных

*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

бинарных отношений между множествами. В математике применяются два таких отношения: симметричное отношение *эквивалентности* множеств и антисимметричное отношение (частичного) *порядка* [6].

Прецедентному соотношению между значениями признака $\Gamma_k(x)$ и t-й таргетной переменной, заданному множеством Q на решетке $L(T(\mathbf{X}))$, соответствует множество пар $\{(\{\Gamma_k^{-1}(\Gamma_k(x_i)), k = \overline{1,n}\}, \Gamma_t^{-1}(\Gamma_t(x_i))), i = \overline{1,N_0}\}$. На основании представленных в регулярном множестве Q прецедентов любой алгоритм распознавания («машинного обучения») строит некоторую модель соотношения между множествами $\{\Gamma_k^{-1}(\Gamma_k(x_i))\}$ и $\Gamma_t^{-1}(\Gamma_t(x_i))$.

Сфокусируемся на двух произвольных множествах $a = \Gamma_k^{-1}(\Gamma_k(x_i))$ и $b = \Gamma_t^{-1}(\Gamma_t(x_i))$. Очевидно, что говорить о выполнимости отношения эквивалентности или порядка между множествами a и b в общем случае не приходится: ведь эквивалентность соответствует идентичности значения k-го признака значению t-й таргетной переменной, а частичный порядок — ядерной эквивалентности (т. е. эквивалентности a подмножеству b или наоборот). Такие случаи тривиальны и соответствуют «легко решаемым» задачам распознавания.

В то же время существующие в решетке $L(T(\mathbf{X}))$ отношения порядка порождают супремум $a \vee b$ и инфимум $a \wedge b$ множеств a и b. Поэтому в топологической теории распознавания вводятся более сложные функционалы над a, b, $a \vee b$ и $a \wedge b$, расширенно описывающие соотношения между произвольными множествами a и b в терминах *расстояний*. При выполнимости четырех аксиом метрики такие функционалы $\rho_L : L^2 \to R^+$ формируют *метрическое пространство значений признаков* $M_L(L(T(\mathbf{X})), \rho_L)$.

Простейшей метрикой служит функционал

$$\rho_0(a,b) = \frac{v[a \vee b] - v[a \wedge b]}{N_0},$$

где $v : L \to R^+$ — *изотонная оценка* на $L(T(\mathbf{X}))$. Возможны и более сложные варианты определения ρ_L при введении параметрических оценок [4] или использовании известных в литературе метрик, так что в общем случае имеется ряд метрик ρ_m, $m = \overline{1,m_0}$. Как правило, метрики ρ_m нормируются на интервал значений $[0 \ldots 1]$.

Возвращаясь к рассмотрению фундаментальных отношений теории функций, можно сделать вывод о том, что метрика $\rho_L(a,b)$ служит функционалом, численно оценивающим выполнимость отношения эквивалентности между a и b на основе отношений порядка (заданных в форме $a \vee b$, $a \wedge b$). Действительно, $\rho_L(a,b) = 0$ соответствует строгому выполнению отношения эквивалентности a и b, а $\rho_L(a,b) = 1$ — максимально возможному расстоянию между a и b (например, $\rho_0(a,b) = 1$ только для множеств $\emptyset$ и $\{\mathbf{X}\}$, находящихся на концах максимальных цепей решетки $L(T(\mathbf{X}))$).

Таким образом, в рамках топологической теории распознавания соотношение между множествами $\{\Gamma_k^{-1}(\Gamma_k(x_i))\}$ и $\Gamma_t^{-1}(\Gamma_t(x_i))$ моделируется соответствующими массивами расстояний, порождаемыми той или иной метрикой ρ_m. Рассмотрим способы вычисления таких расстояний.

2 О различных способах вычисления метрик на решетке $L(T(\mathbf{X}))$

В литературе известны три принципиально различных способа определения метрических расстояний:

(1) метрики на основе операций над множествами;

(2) метрики над пространством векторов;

(3) метрики над пространством функций.

Рассмотрим три этих подхода в применении к описанным выше конструктам топологической теории распознавания.

Метрики на основе операций над множествами. Во введении были упомянуты метрики оценки расстояния между $a, b \in L(T(\mathbf{X}))$, вводимые как функционалы над $a \vee b$ (соответствует $a \cup b$) и $a \wedge b$ ($a \cap b$), оценками высот элементов в $L(T(\mathbf{X}))$ и другими теоретико-множественными операциями над множествами a и b. В работе [4] для порождения метрик используются взвешенные решеточные оценки

$$v_\alpha = \sum_{i=0,|\boldsymbol{\alpha}|} \omega_i v_{\alpha_i}$$

на основе изотонных оценок v_{α_i}. Формирование наборов $\boldsymbol{\alpha}$ может проводиться на основе «информативности» $\alpha_i \in \boldsymbol{\alpha}$ методами метрического анализа данных [7] или на основе различных подцепей таргетных числовых переменных.

В то же время известны многочисленные функционалы, носящие эмпирический характер и непосредственно оперирующие множествами: расстояния Танимото, Рэнда, Рассела–Рао, Симпсона, Брауна–Бланке, Роджера–Танимото, Фэйта, дисперсии, образов, Q_0, Пирсона; различные варианты расстояний Тверского, Сокала–Сниса, Гоуэра–Лежандра, Юле и др. [8]. Метрические свойства

этих функционалов могут быть продемонстрированы посредством аналитических выводов или комбинаторного анализа на множествах прецедентов.

Векторные метрики. Альтернативно методу взвешенных решеточных оценок на основании набора $\boldsymbol{\alpha}$ и оценок v_{α_i} для произвольного множества $a \in L(T(\mathbf{X}))$ может быть вычислен вектор

$$\vec{v}_{\boldsymbol{\alpha}}[a] = (v_{\alpha_1}[a], v_{\alpha_2}[a], \ldots, v_{\alpha_i}[a], \ldots), \;\; v_{\alpha_i}[a] \in R^+,$$

и введены метрики на пространстве векторов $\vec{v}_{\boldsymbol{\alpha}}$ посредством известных подходов: l_1-метрика, l_p-метрики Минковского, расстояния Пенроуза, Манхэттена, Лоренца, Кларка, Хеллинджера, Уайттеккера, симметрическое χ^2, Махаланобиса (в том числе с настраиваемыми весами), расстояния пересечения, Ружечки, Робертса, Элленберга, Глисона, Мотыки, Брея–Куртиса, Канберры, Кульчинского и корреляционные расстояния (ковариационное, корреляционное, косинус, угловое, хордовое, подобности, Мориситы–Хорна, Спирмана, Кендалла) [8].

Метрики над пространством функций. Функциональный анализ и теория вероятностей предоставляют широкий инструментарий для определения расстояний между функциями с одинаковой областью определения: функционалы Колмогорова (в том числе максимальное уклонение D), фон Мизеса, Реньи, метрики (интегральная L_1, инженерная, разделения, подобности среднего гармонического), расстояния Чебышёва, Степанова, варианты расстояний Золотарёва, Круглова, Бурби–Рао, Бхаттачарья, Чизара (включая вариации расстояний Кульбака– Лейблера, χ^2, Хеллинджера) [8].

Очевидно, что некоторые из этих функций расстояний заведомо не относятся к метрикам. Например, в расстоянии Кульбака–Лейблера нарушена аксиома симметричности; оценка выполнимости аксиомы треугольника для каждой из этих функций требует отдельного исследования. Метризация рассматриваемых функций расстояния может осуществляться посредством введения дополнительных конструкций в определение функции. В частности, при отсутствии симметричности для функции $d(x, y)$ могут быть введены конструкты $\min(d(x, y), d(y, x))$, $\max(d(x, y), d(y, x))$ и др. Основной проблемой все же остается «привязка» этих подходов к разрабатываемому решеточному формализму. Для этого напомним ряд важных понятий.

Определение 1. Решеточный терм, или *изотонная оценка*, $v : L \to R^+$ над $L(T(\mathbf{X}))$ — функция, для которой выполнено *условие оценки* (**уО**: $\forall_L a, b : v[a] + v[b] = v[a \wedge b] + v[a \vee b]$) и *условие изотонности* (**уИ**: $\forall_L a, b : a \supseteq b \Rightarrow v[a] \geq v[b]$). Для изотонной $v[\,]$ гарантировано существование метрики $\rho_0(a, b)$.

Определение 2. *Однородными функциями* будем называть произвольные функции с одинаковыми областями определения и значений.

Определение 3. Пусть задано конечное множество чисел $A = \{a_1, a_2, \ldots, a_i, \ldots, a_n\}$, $a_i \in R$. Определим *оператор $\hat{\phi}(x)$ для формирования эмпирической функции распределения* (э. ф. р.) чисел в множестве A как $\hat{\phi}(x)A = \sup|\{B \subseteq A|\forall\, a \in B : a \leq x\}|/|A|$, $x \in R$, так что $\hat{\phi}(-\infty)A = 0$, $\hat{\phi}(+\infty)A = 1$. Для краткости $\hat{\phi}(x)A$ будем также записывать как $\hat{\phi}A$.

Определение 4. $\hat{\mu}$ — *оператор вычисления математического ожидания* значения $x \in A$ по э. ф. р. $\hat{\phi}A$ как

$$\hat{\mu}\hat{\phi}A = \frac{1}{m}\sum_{j=1}^{m} x_j \left(\hat{\phi}(x_j)A\hat{\phi}(x_{j\,1})A\right),$$

где $m = |\hat{z}A|$; $x_j = \hat{\imath}^+(j)\hat{z}A$, а произвольное $x_0 < \inf(A)$, $x_0 \in R$; $\hat{z}$ — *оператор формирования множества значений* множества A, $\hat{z}A = B \subseteq A|\forall\, a \in A : a \in B, \forall\, a, b \in B : a \neq b$; $\hat{\imath}^+$ — *оператор упорядочения множества по возрастанию*; $\hat{\imath}^+(j)A$ — j-й элемент множества $\hat{\imath}^+A$.

3 Анализ решетки $L(T(\mathbf{X}))$ с использованием функций $\hat{\phi}A$ на основе опорных цепей

Рассмотрим подходы к анализу решетки посредством э. ф. р. Определения 2–4 существенно расширяют формализм решеточных оценок, позволяя (1) проектировать $L(T(\mathbf{X}))$ в соответствующую решетку э. ф. р. посредством некоторой заранее выбранной (опорной) цепи; (2) измерять расстояния между этими э. ф. р.; (3) вводить новые разновидности оценок множеств (см. определение 1). В качестве опорной цепи может быть выбрана, в частности, цепь, соответствующая числовой таргетной переменной.

Теорема 1. *Выберем произвольную максимальную цепь A_t в качестве «опорной» для дальнейших построений. При условии регулярности множеств в $\mathbf{X}/Q$ каждому элементу $L(T(\mathbf{X}))$ сопоставлена э. ф. р. из множества однородных э. ф. р.*

Доказательство. При выполнении условия регулярности для $\mathbf{X}/Q$ решетка $L(T(\mathbf{X}))$ — булева (теорема 3 в [9]). Цепи в $L(T(\mathbf{X}))$ соответствуют тем или иным числовым признаковым описаниям (теорема 1 в [4]), так что произвольная (максимальная) цепь A_t в $L(T(\mathbf{X}))$ представима в виде

$$A_t = \langle u(\lambda_{t_1}), \ldots, u(\lambda_{t_i}), \ldots u(\lambda_{t_m})\rangle,$$
$$\lambda_{t_i} \in I_t,\ \ u(\lambda_{t_i}) = \bigcup\nolimits_{\beta=1}^{i} \Gamma_t^{-1}(\lambda_{t_\beta}),$$

где $I_t = (\lambda_{t_1}, \ldots, \lambda_m)$ — строго монотонная последовательность чисел. Значение функции Γ_t, вычислимое для любого объекта в $\mathbf{X}$, равно $\Gamma_t(q)$ для каждого решеточного *атома* $\{q\} \in L(T(\mathbf{X}))$, высота атома равна 1 ($h[\{q\}] \equiv |\{q\}| \equiv 1$). Поскольку решетка булева, то каждый ее элемент представим в виде комбинации атомов, так что любому элементу решетки $u \in L(T(\mathbf{X}))$ сопоставлено множество значений t-го признака $\mathbf{\Gamma}_t(u) = \{\Gamma_t(q), q \in u\}$ по всем атомам из u. Применяя оператор $\hat\phi(x)$ к множеству $\mathbf{\Gamma}_t(u)$, получаем э. ф. р. $\hat\phi\mathbf{\Gamma}_t(u)$. При выполнении условия регулярности для $\mathbf{X}/Q$ решетка $L(T(\mathbf{X}))$ однозначно сопоставлена решетке, образованной числовыми множествами $\mathbf{\Gamma}_t(u)$, и вычислима $\hat\phi\mathbf{\Gamma}_t(u)$. Все эти э. ф. р. однородны по построению — ведь они сформированы над одним и тем же множеством I_t и принимают значения в диапазоне $[0\ldots1]$, а будучи э. ф. р., характеризуются одинаковой областью значений. Теорема доказана.

Итак, при задании опорной цепи A_t любому элементу решетки $u \in L(T(\mathbf{X}))$ сопоставлено множество чисел $\mathbf{\Gamma}_t(u)$, числовая функция $\hat\phi\mathbf{\Gamma}_t(u)$ и ряд функционалов вида $\hat\mu\hat\phi(x)\mathbf{\Gamma}_t(u)$, которые могут быть использованы для определения оценок в решетке $L(T(\mathbf{X}))$ и/или вычисления расстояний.

4 Оценки в решетке $L(T(\mathbf{X}))$ на основе множеств Γ_t с использованием понятия меры

С использованием теоремы 1 изотонные оценки на основе множеств порождаются функционалами вида $g : 2^{I_t} \to R^+$ так, что при произвольных $u, v \in L(T(\mathbf{X}))$ для $g(\mathbf{\Gamma}_t(u))$ и $g(\mathbf{\Gamma}_t(v))$ выполнено уО, а при $u \supseteq v$ выполнено уИ, т. е. $g(\mathbf{\Gamma}_t(u)) \geq g(\mathbf{\Gamma}_t(v))$. Одни из наиболее очевидных функционалов g — различные *меры* множеств, используемые как решеточные оценки [9].

Понятие оценки в теории решеток и понятие меры в функциональном анализе во многом схожи. Как и $v[\,]$, мера положительно определена, мера пустого множества равна нулю, а мера пересечения непересекающихся множеств равна сумме мер этих множеств. Однако уО выдвигает дополнительное требование: если множества пересекаются, то оценка их объединения равна сумме оценок множеств минус оценка их пересечения, так что любая $v[\,]$ — мера. Меры могут вводиться различными способами [6].

Определение 5. Пусть точкам действительной оси $I_t = \{\lambda_{t_1}, \lambda_{t_2}, \ldots, \lambda_{t_b}, \ldots, \lambda_{k_{|I_t|-1}}, \Delta\}$, $\lambda_{t_b} \in R$, сопоставлены веса $p_1, p_2, \ldots, p_{|I_t|-1}$. Тогда определена *мера с дискретным весом*

$$\mu_{\text{дв}}(\mathbf{\Gamma}_t(u)) = \sum_{\lambda_{t_b} \in \mathbf{\Gamma}_t(u)} p_b.$$

В качестве весов в определении 5 могут быть выбраны: (1) длины интервалов значений из I_t ($\lambda_{t_b} - \lambda_{t_{b-1}}, \lambda_{t_{b+1}} - \lambda_{t_b}$ и др.); (2) разности значений э. ф. р. ($\text{cdf}\,(\lambda_{t_{b+1}}, A_t(\mathbf{X})) - \text{cdf}\,(\lambda_{t_b}, A_t(\mathbf{X}))$ и др.); (3) веса, настраиваемые в соответствии с принципом согласования метрик, и т. д. Очевидна

Теорема 2. *Мера с дискретным весом — оценка.*

Утверждение следует из рассмотрения перекрывающихся и неперекрывающихся множеств $\mathbf{\Gamma}_t(u)$ и $\mathbf{\Gamma}_t(v)$ и выполнимости уО из определения 1.

Следствие 1. *Интеграл от суммируемой функции f с использованием $\mu_{\text{дв}}$ вычисляется как*

$$\int_{-\infty}^{+\infty} f(\lambda)\, d\mu_{\text{дв}} = \sum_{b=1,|I_t|-1} p_b f(\lambda_{t_b}).$$

Следствие 2. *Скалярное произведение суммируемых функций $f(\lambda)$ и $g(\lambda)$ на основе $\mu_{\text{дв}}$ равно*

$$(f, g) = \sum_{b=1,|I_t|-1} p_b f(\lambda_{t_b}) g(\lambda_{t_b}).$$

Следствие 3. *Колмогоровский функционал «заряда» $\Phi(A)$ на множестве чисел A с использованием суммируемой функции $f(\lambda)$, $\Phi(A) = \int_A f(\lambda)\, d\mu$, служит мерой.* При использовании меры с дискретными весами

$$\Phi(A) = \sum_{\lambda_{t_b} \in A} p_b f(\lambda_{t_b})$$

(следствие 1).

Следствие 4. *Колмогоровский заряд $\Phi(\mathbf{\Gamma}_t(u))$ — изотонная оценка на решетке $L(T(\mathbf{X}))$ при положительной определенности $f(\lambda)$.* Перекрывание площади под произвольной одномерной f в случае множеств $\mathbf{\Gamma}_t(u)$ и $\mathbf{\Gamma}_t(v)$ равно $\Phi(\mathbf{\Gamma}_t(u) \cap \mathbf{\Gamma}_t(v))$, что равно $\Phi(\mathbf{\Gamma}_t(u \cap v)$ и равно сумме площадей $\Phi(\mathbf{\Gamma}_t(u))$ и $\Phi(\mathbf{\Gamma}_t(v))$ минус площадь объединения множеств (что соответствует выполнимости уО в определении 1). Оценка Φ *изотонна* при $f(\lambda) \geq 0$.

Из теоремы 2 со следствиями очевидно, что введение «заряда» Φ позволяет более гибко оценивать

вклад каждого значения таргетной переменной λ_{t_b} в значение меры: ведь в $\Phi(A)$ используются не только дискретные веса p_b значений λ_{t_b}, но и весовая функция $f(\lambda)$, общая для всех значений.

5 Перспективы анализа решеток без использования понятия оценки

Решеточные термы $v : L \to R^+$ дают скалярную оценку каждого элемента соответствующей решетки L, позволяющее сравнивать элементы L между собой (при выполнимости уО и уИ). Очевидно, что сопоставление произвольному элементу u решетки $L(T(\mathbf{X}))$ функции $\hat{\phi}\mathbf{\Gamma}_t(u)$ представляется гораздо более сложным «оценочным» описанием множества u, чем скалярная $v[u]$. Эта цепь рассуждений указывает на два направления дальнейшего развития формализма:

(1) введение функционалов, позволяющих редуцировать более сложное описание в виде $\hat{\phi}\mathbf{\Gamma}_t(u)$ к скалярным оценкам;

(2) разработку нового математического аппарата для анализа решеток, в котором вместо оценок $v : L \to R^+$ фигурируют операции над функциями $\hat{\phi}\mathbf{\Gamma}_t(u)$.

Первое направление отчасти покрывается результатами теоремы 2 со следствиями. Второе направление интересно тем, что позволяет вводить метрические функции расстояния без использования конструкции $v[x \vee y] - v[x \wedge y]$, на основании упоминаемых выше подходов функционального анализа (как это было сделано в работе [4] для колмогоровского «максимального уклонения»).

6 Заключение

В прикладной математике повсеместно используются функционалы, оценивающие расстояния между множествами, векторами или функциями. При установлении метрических свойств этих функционалов инструментарий формализма топологической теории распознавания может быть существенно обогащен нетривиальными метриками на основе эмпирических и полуэмпирических функционалов расстояния. Во второй части статьи будут представлены результаты приложения разрабатываемого формализма к комплексу прикладных задач из области фармакоинформатики.

Литература

1. *Журавлёв Ю. И., Рудаков К. В., Торшин И. Ю.* Алгебраические критерии локальной разрешимости и регулярности как инструмент исследования морфологии аминокислотных последовательностей // Труды МФТИ, 2011. Т. 3. № 4. С. 45–54. EDN: OJYMVJ.
2. *Рудаков К. В., Торшин И. Ю.* Анализ информативности мотивов на основе критерия разрешимости в задаче распознавания вторичной структуры белка // Информатика и её применения, 2012. Т. 6. Вып. 1. С. 79–90. EDN: OZHDTV.
3. *Торшин И. Ю.* О задачах оптимизации, возникающих при применении топологического анализа данных к поиску алгоритмов прогнозирования с фиксированными корректорами // Информатика и её применения, 2023. Т. 17. Вып. 2. С. 2–10. doi: 10.14357/19922264230201. EDN: IGSPEW.
4. *Торшин И. Ю.* О формировании множеств прецедентов на основе таблиц разнородных признаковых описаний методами топологической теории анализа данных // Информатика и её применения, 2023. Т. 17. Вып. 3. С. 2–7. doi: 10.14357/19922264230301. EDN: AQEUYO.
5. *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.
6. *Колмогоров А. Н., Фомин С. В.* Элементы теории функций и функционального анализа. — М.: Наука, 1989. 624 с.
7. *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.
8. *Деза Е. И., Деза М. М.* Энциклопедический словарь расстояний / Пер. с англ. — М.: Наука, 2008. 444 с. (*Deza E., Deza M. M.* Dictionary of distances. — North-Holland: Elsevier, 2006. 412 p. doi: 10.1016/B978-0-444-52087-6.X5000-8.)
9. *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.

Поступила в редакцию 15.01.23

ON THE GENERATION OF SYNTHETIC FEATURES BASED ON SUPPORT CHAINS AND ARBITRARY METRICS WITHIN A TOPOLOGICAL APPROACH TO DATA ANALYSIS. PART 1. INCLUSION OF EMPIRICAL DISTANCE FUNCTIONS INTO THE FORMALISM

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 analysis of the formalism of topological recognition theory based on the fundamental concepts of functional analysis made it possible to propose previously unexplored approaches to determining lattice estimates. In particular, the use of support chains for the analysis of Boolean lattices formed over Zhuravlev-regular sets of precedents has pointed to a new direction of research which consists in replacing estimates of lattice elements with certain types of functions and/or vectors. This extension of the formalism also allows for a systematic study of semiempirical distance functionals known in the literature to solve applied problems. Promising directions for further development of the formalism are substantiated including the functionals reducing descriptions of sets of a Boolean lattice to scalars and the development of a mathematical apparatus for the analysis of lattices where operations on the corresponding functions are involved. The latter direction is interesting as it allows defining lattice metrics without using lattice estimates.

Keywords: topological data analysis; lattice theory; algebraic approach by Yu. I. Zhuravlev; functional analysis

DOI: 10.14357/19922264240110 **EDN:** RIVOXR

Acknowledgments

The research was funded by the Russian Science Foundation, project No. 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., K. V. Rudakov, and I. Yu. Torshin. 2011. Algebraicheskie kriterii lokal'noy razreshimosti i regulyarnosti kak instrument issledovaniya morfologii aminokislotnykh posledovatel'nostey [Algebraic criteria of local solvability and regularity as a tool for studying the morphology of amino acid sequences]. *Trudy MFTI* [Proceedings of Moscow Institute of Physics and Technology] 3(4):45–54. EDN: OJYMVJ.
2. Rudakov, K. V., and I. Yu. Torshin. 2012. 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.* 6(1):79–90. EDN: OZHDTV.
3. Torshin, I. Yu. 2023. O zadachakh optimizatsii, voznikayushchikh pri primenenii topologicheskogo analiza dannykh k poisku algoritmov prognozirovaniya s fiksirovannymi korrektorami [On optimization problems arising from the application of topological data analysis to the search for forecasting algorithms with fixed correctors]. *Informatika i ee Primeneniya — Inform Appl.* 17(2):2–10. doi: 10.14357/19922264230201. EDN: IGSPEW.
4. Torshin, I. Yu. 2023. O formirovanii mnozhestv pretsedentov na osnove tablits raznorodnykh priznakovykh opisaniy metodami topologicheskoy teorii analiza dannykh [On the formation of sets of precedents based on tables of heterogeneous feature descriptions by methods of topological theory of data analysis]. *Informatika i ee Primeneniya — Inform Appl.* 17(3):2–7. doi: 10.14357/19922264230301. EDN: AQEUYO.
5. 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.
6. Kolmogorov, A. N., and S. V. Fomin. 1989. *Elementy teorii funktsiy i funktsional'nogo analiza* [Elements of theory of functions and functional analysis]. Moscow: Nauka. 624 p.
7. 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 clas-

sification of feature values. *Pattern Recognition Image Analysis* 27(2):184–199. doi: 10.1134/S1054661817020110.

8. Deza, E., and M. M. Deza. 2006. *Dictionary of distances*. North-Holland: Elsevier. 412 p. doi: 10.1016/B978-0-444-52087-6.X5000-8.
9. 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.

Received January 15, 2023

Contributor

Torshin Ivan Yu. (b. 1972) — Candidate of Science (PhD) in physics and mathematics, Candidate of Science (PhD) in chemistry, leading scientist, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str, Moscow 119333, Russian Federation; tiy135@yahoo.com