ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2023. Т. 17. Вып. 3. С. 2–7

О ФОРМИРОВАНИИ МНОЖЕСТВ ПРЕЦЕДЕНТОВ НА ОСНОВЕ ТАБЛИЦ РАЗНОРОДНЫХ ПРИЗНАКОВЫХ ОПИСАНИЙ МЕТОДАМИ ТОПОЛОГИЧЕСКОЙ ТЕОРИИ АНАЛИЗА ДАННЫХ*

И. Ю. Торшин[1]

Аннотация: Факторизация вкладов различных переменных при анализе разнородных признаковых описаний — насущная задача интеллектуального анализа сложных данных. В работе предложено развитие решеточного формализма топологической теории анализа данных, в рамках которого получены новые способы порождения параметрических оценок и метрик на решетках, образованных над топологиями множеств объектов. Формализм был апробирован на задаче формирования множеств прецедентов для проведения хемомикробиомного анализа. Тогда как порождение множества исходных информаций на основе регрессионных коэффициентов и разности значений материала обучения соответствовало крайне низкой обобщающей способности настраиваемых алгоритмов (коэффициент корреляции на контроле — $0{,}32 \pm 0{,}20$), использование предлагаемых оценок для порождения множеств прецедентов в задачах хемомикробиомики позволило существенно повысить обобщающую способность соответствующих алгоритмов (коэффициент корреляции на контроле — $0{,}79 \pm 0{,}21$).

Ключевые слова: топологический анализ данных; теория решеток; параметризация решеточных термов; микробиом человека; фармакоинформатика, алгебраический подход Ю. И. Журавлёва

DOI: 10.14357/19922264230301 **EDN:** AQEUYO

1 Введение

В биомедицинских исследованиях объектом служит формализованное описание состояния пациента, включающее булевы (диагнозы, прием лекарств и др.), числовые (лабораторные анализы) и категорные (показатели демографии и др.) переменные, графы (формулы лекарств), временны́е ряды и изображения (результаты обследования пациента аппаратными методами). При формализации таких задач важно выделить независимые вклады переменных-признаков в таргетную переменную (отклик) таким образом, чтобы получить высокое качество работы алгоритмов распознавания/классификации. Топологический анализ данных, развиваемый в рамках алгебраического подхода к распознаванию научной школы Ю. И. Журавлёва и К. В. Рудакова [1, 2], позволяет систематически исследовать возможные решения этой задачи.

В настоящей работе данный подход апробирован на задаче формирования множества прецедентов для проведения хемомикробиомного анализа лекарств [3]. В этой прикладной задаче (оценка влияния лекарств на микробиоту) факт использования лекарства, как правило, описан булевым признаком, а откликом служит уровень той или иной бактерии микробиома (числовая переменная). Для выделения независимых вкладов булевых переменных в числовые таргетные переменные предложен системный подход к порождению параметризованных оценок на решетке (решеточных термов) и соответствующих метрик. Такой подход представляет собой теоретическое обобщение «расщепления» эмпирической функции распределения булевым признаком [2] и тесно связан с порождением метрических функций расстояния и с проблематикой так называемых «оценок информативности» (точнее, весовых функций признаков), используемых в комбинаторной теории разрешимости [4].

2 Основные понятия

Основы формализма изложены в [2, 5]. Задано множество исходных описаний объектов $\mathbf{X} = \{x_1, \ldots, x_{N_0}\}$, $\mathbf{X} \subseteq S$, множества значений признаков $I_k = \{\lambda_{k_1}, \lambda_{k_2}, \ldots, \lambda_{k_b}, \ldots, \lambda_{k_{|I_k|-1}}, \Delta\}$, $b = 1, \ldots, |I_k|$, функции $\Gamma_k : S \to I_k$, $k = 1, \ldots, n+l$ (n — число признаков; l — число таргетных (прогнозируемых) переменных; Δ — неопределенность). Определено пространство допустимых признаковых описаний объектов $J_{\text{ob}} \subseteq I_i \times I_f$ ($I_i \subseteq I_1 \times \cdots \times I_n$, $I_f \subseteq I_{n+1} \times \cdots \times I_{n+1}$),

*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

функции $D : S \to J_{\text{ob}}$, $D(x_\alpha) = (\Gamma_1(x_\alpha) \times \cdots \cdots \times \Gamma_k(x_\alpha) \times \cdots \times \Gamma_{n+1}(x_\alpha))_\Delta$ и $\varphi(\mathbf{X}) = \{D(x_\alpha) | x_\alpha \in \in \mathbf{X}\}$. Принимается, что множества $\mathbf{X}$ и $Q = \varphi(\mathbf{X})$, $|Q| = N$, *регулярны*, т. е. $N_0 = N$ и $\exists D^{-1} : \forall x \in \mathbf{X}$, $x = D^{-1}(D(x))$ [5]. Множество $U(\mathbf{X}) = \{\Gamma_k^{-1}(\lambda_{k_b})\}$, образованное функциями полных прообразов значений признаков λ_{k_b}, рассматривается как предбаза *топологии* $T(\mathbf{X}) = \{\varnothing, I, a \cup b, a \cap b : a, b \in U(\mathbf{X})\}$, где $I = \{\mathbf{X}\}$. Топологии $T(\mathbf{X})$ соответствует *решетка* $L(T(\mathbf{X})) = \{a \vee b, a \wedge b : a, b \in T(\mathbf{X})\}$. При регулярности $\mathbf{X}$ и Q решетка $L(T(\mathbf{X}))$ — булева [2].

Теорема 1. *Произвольной цепи* $\langle u_1, \ldots, u_m \rangle$ *решетки* $L(T(\mathbf{X}))$ *можно сопоставить числовой признак с множеством значений* $I_t = (\lambda_{t_1}, \ldots, \lambda_{t_b}, \ldots, \lambda_m)$, $\lambda_{t_{i-1}} \leq \lambda_{t_1} \leq \lambda_{t_{i+1}}$.

Справедливость теоремы следует из того, что линейный порядок любой конечной цепи $\langle u_1 \supseteq \supseteq u_2 \cdots \supseteq u_i \cdots \supseteq u_m \rangle$ изоморфен линейному порядку на конечном множестве чисел $\{\lambda_{t_i}\}$, $\lambda_{t_{i-1}} \leq \leq \lambda_{t_i}$, $i = 2, \ldots, m$.

Следствие 1. *Любая цепь* A_t *в* $L(T(\mathbf{X}))$ *представима в виде* $A_t = \langle u(\lambda_{t_1}), \ldots, u(\lambda_{t_i}), \ldots, u(\lambda_{t_m}) \rangle$, $u(\lambda_{t_i}) = = \bigcup_{\beta=1}^{i} \Gamma_t^{-1}(\lambda_{t_\beta})$, *где* $I_t = (\lambda_{t_1}, \ldots, \lambda_m)$ *монотонна*.

Следствие 2. *При дополнении наблюдаемой области* I_t *значений* t*-го числового признака неопределенностью* Δ *и принятии соглашения* $\Gamma_t(\varnothing) = \Delta$ *соответствующая цепь* A_t *— максимальная цепь решетки* $L(T(\mathbf{X}))$.

Следствие 3. *Для произвольной цепи* t*-го числового признака определена эмпирическая функция распределения* $\text{cdf}\,(\lambda, A_k(\mathbf{X})) = |u(\lambda_{k_b})|/N | \lambda_{k_{b-1}} \leq \lambda \leq \lambda_{k_b}$.

3 Решеточные термы и операции над ними

В топологическом анализе данных булевой решетке сопоставляется *метрическое пространство значений признаков* $M_L(L(T(\mathbf{X})), \rho_L)$ с метрикой $\rho_L : L^2 \to R^+$ [2]. Метрики ρ_L могут определяться на основании формализма оценок (см. ниже) или же на основе известных подходов (расстояния Танимото, Амана, Тверского, Сокала–Сниса, Гоуэра–Лежандра и др.) [6]. 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 $b = \langle b_1, \ldots, b_j, \ldots \rangle$ определено так, что существует однозначное соответствие элементов цепей a и b (например, $\rho_A(a, b) = \min(\sum_{i=1,|a|} \rho_L(a_i, \arg\min_{b_j \in b} \rho_L(a_i, b_j))$, $\sum_{i=1,|b|} \rho_L(b_j, \arg\min_{a_i \in a} \rho_L(b_j, a_i))$. При заданных ρ_L, ρ_A, $A_k(\mathbf{X})$ и множества допустимых цепей $A(\mathbf{X})_{1,n}$ (содержит цепи с длинами от 1 до n) искомый (ε-корректный) алгоритм соответствует решению задачи комбинаторной оптимизации [5]:

$$\arg\min_{a \in \mathbf{A}(\mathbf{X})_{1,n}} \rho_A\,(A_k(\mathbf{X}), a)\,|A(\mathbf{X})_{1,n} \subset \mathbf{A}(\mathbf{X}).$$

При порождении метрических пространств над произвольной решеткой вводится понятие решеточного терма или *оценки* $v : L \to R^+$, 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**: $\forall_L a, b : a \supseteq b \Rightarrow v[a] \geq v[b]$). Изотонность оценки $v[\,]$ важна потому, что позволяет определить метрику $\rho(x, y) = v[x \vee y] - v[x \wedge y]$ [2].

Теорема 2. *Линейная комбинация* $v[\,] = \sum \omega_i v_i[\,]$, $i = = 1, \ldots, m$, *произвольного числа изотонных оценок* $v_i[\,]$ *— изотонная оценка при* $v[\,] \geq 0$.

Теорема доказывается через линейные преобразований условий уО$_i$ и уИ$_i$ для m оценок $v_i[\,]$.

Следствие 1. *Сумма произвольного числа (изотонных) оценок — оценка*.

Следствие 2. *Разность (изотонных) оценок становится оценкой только при условии положительной определенности*.

Следствие 3. *Пусть индекс* i *в выражении для* $v[\,]$ *изменяется от* $i = 0$, $v_0 \equiv 1$. *Тогда можно подобрать такое значение* ω_0, *что* $v[\,] \geq 0$.

При произвольном нелинейном преобразовании $f : R^+ \to R^+$, применяемом к (изотонной) оценке $v[\,]$, выполнимость уО неочевидна. Если $v[\,]$ изотонна, то $f(v[\,])$ может быть изотонна только при условии монотонности f (сигмоида, степенная функция и др.). Для оценки выполнимости уО/уИ $f(v[\,])$ на подмножестве решетки L возможно применение аналитических или комбинаторных подходов. Например, легко показать, что для $f(x) = = x^\alpha$ при $\alpha = 2$ или 0,5 выражение $f(v[\,])$ служит оценкой только на элементах произвольной цепи решетки $L(T(\mathbf{X}))$.

4 Параметрические оценки на основе «опорного элемента»

Выберем подмножество объектов $\mathbf{X}$, $\alpha \in \in L(T(\mathbf{X}))$. Подмножество α может соответствовать k'-му булеву признаку в исходном признаковом описании (тогда $\alpha = \Gamma_{k'}^{-1}(1)$) или синтетическому булеву признаку, что разбивает $\mathbf{X}$ на α и $\overline{\alpha} = \mathbf{X} \backslash \alpha$ и определяет $\nu_\alpha = |\alpha|/|\mathbf{X}| = |\alpha|/N$ и $\nu_{\overline{\alpha}} = 1 - \nu_\alpha$.

«Опираясь» на множество α как параметр, можно породить несколько изотонных решеточных оценок на основе базовой оценки, равной высоте элемента u_i в $L(T(\mathbf{X}))$, $h[u_i] = |u_i|$, частот ν_α

и $\nu_{\overline{\alpha}}$. Оценка $v^+_\alpha[u_i] = |u_i \cap \alpha|/|\alpha|$ соответствует частоте встречаемости элементов из $u_i \subseteq \mathbf{X}$ в $\alpha \subset \mathbf{X}$; оценка $v^-_\alpha[u_i] = |u_i \cap \overline{\alpha}|/|\overline{\alpha}| = |u_i \backslash \alpha|/(N - |\alpha|)$ — встречаемость элементов u_i в $\overline{\alpha}$. В соответствии с теоремой 2 линейные комбинации оценок v^+_α и v^-_α также изотонны при условии положительной определенности.

Теорема 3. *Изотонная оценка $d_\alpha[\,] = v^+_\alpha[\,] - v^-_\alpha[\,]$ существует, когда объекты из множества α встречаются в оцениваемых множествах u_i не реже, чем в среднем по множеству всех объектов $\mathbf{X}$.*

Рассмотрим функционал $\mathrm{d}_\alpha[u_i] = v^+_\alpha[u_i] + \omega v^-_\alpha[u_i]$, $\omega \in R$, и покажем, что $d_\alpha[\,]$ — линейная комбинация метрики $\rho(u_i, \alpha)$ и оценки $h[u_i] = |u_i|$. Приведем v^+_α и v^-_α к общему знаменателю, равному $|\alpha|(N - |\alpha|)$. Подставляя

$$|u_i \cap \alpha| = 0{,}5(|u_i| + |\alpha| - \rho(u_i, \alpha)),$$

после упрощения получим (для произвольного u_i), что

$$d_\alpha[u_i] = k_\alpha|\alpha| + b_\alpha|u_i| - k_\alpha\rho(u_i, \alpha)\,,$$

где

$$b_\alpha = \frac{1/v_\alpha + \omega - 1}{2(N - |\alpha|)}\,;$$
$$k_\alpha = \frac{1/v_\alpha - \omega - 1}{2(N - |\alpha|)}\,.$$

Рассмотрим функционал $d_\alpha[\,]$ (который служит изотонной оценкой для всех $\omega > 0$) в случае $\omega < 0$ и запишем его в виде $d_\alpha[\,] = v^+_\alpha[\,] - |\omega| v^-_\alpha[\,]$. Условию $d_\alpha[u_i] \geq 0$ соответствует

$$\frac{1/v_\alpha + |\omega| - 1}{N - |\alpha|}\,|u_i \bigcap \alpha| \geq \frac{|\omega||u_i|}{N - |\alpha|}\,,$$

т. е.

$$\frac{|i_i \cap \alpha|}{|u_i|} \geq \frac{|\omega|}{1/v_\alpha + |\omega| - 1}\,.$$

Это условие выделяет подмножество $\{u_i\} \subset \subset L(T(\mathbf{X}))$, где $d_\alpha[\,]$ изотонна. При $|\omega| \sim 1$ получаем $|u_i \cap \alpha|/|u_i| \geq \nu_\alpha$, соответствующее условию теоремы. Теорема доказана.

Следствие 1.

$$d'_\alpha[u_i] = \begin{cases} v'_\alpha[u_i] - v'_\alpha[u_i] & \textit{при } \dfrac{|u_i \cap \alpha|}{|u_i|} \geq \nu_\alpha\,; \\ 0 & \textit{в противном случае} \end{cases}$$

есть изотонная оценка.

Функционал $d_\alpha[u_i]$ интересен тем, что может быть увязан с конструктами непараметрической статистики, разработанными в научной школе А. Н. Колмогорова [7]. Как известно, для непараметрических тестов используется так называемое максимальное уклонение D, равное супремуму модуля разности значений двух функций распределения (теоретической и эмпирической или двух эмпирических функций распределения) по всей их области определения.

Теорема 4. *Для «опорного» множества α и произвольной цепи $U = \langle u_1, \ldots, u_i, \ldots\rangle$*

$$D = \sup|d_\alpha[u_i]|$$

при $\omega = -1$ — максимальное уклонение Колмогорова.

Доказательство проводится исходя из теоремы 1 для любой цепи $A = \langle u_i\rangle$, $u_i = u(\lambda_{k_i})$ и анализа числового признака с множеством значений I_k в двух цепях, образованных пересечением каждого элемента u_i цепи A с множествами α и $\overline{\alpha}$ соответственно.

Следствие 1. *Статистическая достоверность значения D оценивается по критерию Колмогорова–Смирнова.*

Оценки v^+_α, v^-_α, d_α и др. могут строиться для набора «опорных» множеств. Пусть задан набор $\boldsymbol{\alpha} \subset L(T(\mathbf{X}))$ подмножеств объектов (т. е. элементов решетки): $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \ldots, \alpha_i \ldots\}$. Каждое из множеств в наборе $\boldsymbol{\alpha}$ порождает (изотонную) параметрическую оценку v_{α_i} (например, $v^+_{\alpha_i}$, $v^-_{\alpha_i}$, их линейные комбинации). Тогда в соответствии с теоремой 2 может быть порождена и настраиваемая изотонная оценка

$$v_\alpha = \sum_{i=0,|\boldsymbol{\alpha}|} \omega_i v_{\alpha_i}, \quad v_{\alpha_0} \equiv 1\,.$$

Альтернативно на основании набора $\boldsymbol{\alpha}$ для каждого $a \in L(T(\mathbf{X}))$ может быть вычислен вектор оценок $(v_{\alpha_1}[a], v_{\alpha_2}[a], \ldots, v_{\alpha_i}[a], \ldots)$, $v_{\alpha_i}[a] \in R^+$, и введены метрики уже на пространстве таких векторов (l_p-метрики, расстояния Пенроуза, Мотыки, Брея–Куртиса, кореляционные расстояния) [6]. Формирование наборов $\boldsymbol{\alpha}$ может проводиться на основе «информативности» соответствующих $\alpha_i \in \boldsymbol{\alpha}$ методами метрического анализа данных и др. [4].

5 О параметрических оценках на основе «опорной цепи»

Крайне перспективным представляется направление порождения оценок, в котором заданная максимальная цепь A_t используется в качестве «опорной», т. е. для порождения параметрических

оценок. Данное направление будет подробно рассмотрено в отдельной работе. Вкратце: произвольная цепь $A = \langle u_i \rangle$ представима в виде $\langle u(\lambda_{t_i}) \rangle$ для некоторого упорядоченного множества чисел I_t (теорема 1). Значение функции Γ_t (включая неопределенность), вычислимое для любого объекта в $\mathbf{X}$, равно $\Gamma_t(q)$ для каждого решеточного атома $\{q\} \in L(T(\mathbf{X}))$, ($h[\{q\}] \equiv |\{q\}| \equiv 1$), так что любому $u \in L(T(\mathbf{X}))$ соответствует множество $\mathbf{\Gamma}_t(u) = \{\Gamma_t(q), q \in u\}$. Определим *оператор* $\hat{\phi}(x)$ для *формирования эмпирической функции распределения* конечного множества $A = \{a_1, a_2, \ldots, a_i, \ldots, a_n\}$, $a_i \in R$, как

$$\hat{\phi}(x)A = \sup \frac{|\{B \subseteq A | \forall\, a \in B : a \leq x\}|}{|A|}, \quad x \in R,$$

и *оператор вычисления математического ожидания* $\hat{\mu}$.

Итак, при задании опорной цепи A_t (т. е. опорного числового признака с областью значений I_t) любому элементу решетки $u \in L(T(\mathbf{X}))$ сопоставлено множество чисел $\mathbf{\Gamma}_t(u)$, числовая функция $\hat{\phi}(x)\mathbf{\Gamma}_t(u)$ одной переменной $x \in R$, нетривиально определенная в каждой точке отрезка $[\lambda_{t_1}, \lambda_{t_{|I_t|-1}}]$, и ряд функционалов (в частности, $\hat{\mu}\hat{\phi}(x)\mathbf{\Gamma}_t(u)$). При выполнении условия регулярности для $\mathbf{X}/Q$ решетка $L(T(\mathbf{X}))$ однозначно сопоставлена решетке, образованной числовыми множествами $\mathbf{\Gamma}_t(u)$, для каждого элемента которой вычислима функция $\hat{\phi}\mathbf{\Gamma}_t(u)$. Множества $\mathbf{\Gamma}_t(u)$, функции $\hat{\phi}\mathbf{\Gamma}_t(u)$ и функционалы наподобие $\hat{\mu}\hat{\phi}\mathbf{\Gamma}_t(u)$ могут быть использованы для определения оценок в решетке $L(T(\mathbf{X}))$, порождаемых на основе выбранного опорного признака.

6 Экспериментальная апробация формализма

Формализм был апробирован на задаче формирования множеств прецедентов для проведения хемомикробиомного анализа [3, 8]. В исходной таблице признаковых описаний представлены булевы переменные, соответствующие 122 лекарствам, которые влияют на 1533 числовые переменные, соответствующие уровням отдельных бактерий или их групп.

Для квантификации вклада лекарства α в изменение уровней t-й бактерии среди разработанных оценок использовалось уклонение D со знаком (теоремы 3 и 4). Для учета направления изменения уровней бактерий уклонение D необходимо домножить на знак $\mathrm{d}_\alpha[\,]$. По тесту Колмогорова–Смирнова вычислялось значение статистической достоверности $p(D)$ для отсечения наименее достоверных ассоциаций «лекарство–эффект». Также использовались более очевидные подходы: коэффициенты линейной многопараметрической регрессии (с отбором признаков по модели «лассо»), разность значений $\hat{\mu}\hat{\phi}\mathbf{\Gamma}_t(\alpha) - \hat{\mu}\hat{\phi}\mathbf{\Gamma}_t(\overline{\alpha})$, доля разности $1 - \hat{\mu}\hat{\phi}\mathbf{\Gamma}_t(\overline{\alpha})/\hat{\mu}\hat{\phi}\mathbf{\Gamma}_t(\alpha)$.

Лекарства с известной химической структурой описывались на основе хемогра́фов G_j, а в качестве множества начальных информаций I_i использовалось множество хемоинвариантов над алфавитом элементных меток. Алгоритмы

Результаты вычислительных экспериментов на выборке 156 327 измерений хемомикробиомной биологической активности (2173 пациентов). Оценены эффекты 122 лекарств. Приведены значения коэффициентов ранговой корреляции для различных моделей формирования множеств прецедентов. Эксперименты проводились в рамках кросс-валидационного дизайна (10 разбиений в соотношении «случай–контроль» 6 : 1). Поиск оптимальных значений параметров проводился мультистартовой стохастической оптимизацией

Вычислительный эксперимент	$n_{\text{акт}}$	r	r_c
Коэффициент многопараметрической регрессии, $p(D) = 0{,}20$	1250	$0{,}72 \pm 0{,}23$	$0{,}32 \pm 0{,}20$
Коэффициент многопараметрической регрессии, $p(D) = 0{,}05$	730	$0{,}73 \pm 0{,}39$	$0{,}38 \pm 0{,}18$
Разность значений, $p(D) = 0{,}20$	1192	$0{,}74 \pm 0{,}19$	$0{,}34 \pm 0{,}22$
Разность значений, $p(D) = 0{,}05$	787	$0{,}78 \pm 0{,}36$	$0{,}39 \pm 0{,}25$
Доля разности значений, $p(D) = 0{,}20$	1292	$0{,}80 \pm 0{,}22$	$0{,}46 \pm 0{,}14$
Доля разности значений, $p(D) = 0{,}10$	1291	$0{,}80 \pm 0{,}27$	$0{,}62 \pm 0{,}18$
Доля разности значений, $p(D) = 0{,}05$	**868**	$\mathbf{0{,}76 \pm 0{,}37}$	$\mathbf{0{,}77 \pm 0{,}22}$
Уклонение D со знаком, $p(D) = 0{,}20$	1286	$0{,}81 \pm 0{,}21$	$0{,}48 \pm 0{,}15$
Уклонение D со знаком, $p(D) = 0{,}10$	1306	$0{,}80 \pm 0{,}26$	$0{,}63 \pm 0{,}18$
Уклонение D со знаком, $p(D) = 0{,}05$	**836**	$\mathbf{0{,}80 \pm 0{,}38}$	$\mathbf{0{,}79 \pm 0{,}21}$

Обозначения: r и r_c — коэффициенты ранговой корреляции на обучении и на контроле соответственно; $n_{\text{акт}}$ — число различных типов хемомикробиомной активности, по которым проводилось усреднение r и r_c; $p(D)$ — верхний порог статистической значимости по тесту Колмогорова–Смирнова.

$f_{\theta_k} : I_i \to R$ строились в виде композиций вложенных корректирующих функций нижнего уровня (т. е. порождения синтетических признаков) для фиксированного числа моделей n_{mod} : $f_{\theta_k} = = = g(f_1(\sum \omega_k^j x_k), \ldots, f_l(\sum \omega_k^j x_k), \ldots)$, $l = 1, \ldots \ldots, n_{\text{mod}}$, как в работе [5]. Тестирование алгоритмов f_{θ_k} проводилось на выборке данных о пациентах (см. таблицу).

Результаты экспериментов показывают, что порождение множества исходных информаций на основе регрессионных коэффициентов и разности значений материала обучения соответствует крайне низкой обобщающей способности настраиваемых алгоритмов. При этом ужесточение порога на значение $p(D)$, давая снижение числа различных типов хемомикробиомной активности, не приводило к увеличению качества алгоритмов. Наилучший результат был получен при использовании в качестве исходной информации значения уклонения D со знаком ($r = 0{,}80 \pm 0{,}38$, $r_c = 0{,}79 \pm 0{,}21$), а более низкие значения порога $p(D)$ приводили к повышению качества распознавания. Отметим неплохой результат и для такого простого функционала, как доля разности значений (см. таблицу).

7 Заключение

В работе впервые проведено систематическое рассмотрение способов введения оценок на решетке (высота элемента, оценки на основании булевых и числовых переменных, линейные комбинации вышеперечисленных оценок и др.). Введение оценок на $L(X)$, по аналогии с понятием меры в функциональном анализе, позволяет порождать параметрические решеточные оценки и затем вводить проблемно-ориентированные метрики, оценивающие расстояние между вершинами решетки. Разработанный формализм был применен для достижения практической цели настоящей статьи — нахождения оптимального способа оценки вкладов переменных при анализе сложных данных хемомикробиомных исследований. Разработанный формализм предоставляет инструментарий для поиска адекватной формализации задач классификации и прогнозирования.

Литература

1. *Журавлёв Ю. И.* Избранные научные труды. — М.: Магистр, 1998. 420 с.
2. *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. doi: 10.1134/S1054661815040252.
3. *Торшин И. Ю., Громова О. А., Захарова И. Н., Максимов В. А.* Хемомикробиомный анализ Лактитола // Экспериментальная и клиническая гастроэнтерология, 2019. Т. 164. № 4. С. 111–121. doi: 10.31146/1682-8658-ecg-164-4-111-121.
4. *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.
5. *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.
6. *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. — Elsevier B.V., 2006. 412 p.)
7. *Колмогоров А. Н., Фомин С. В.* Элементы теории функций и функционального анализа. — М.: Наука, 1989. 624 с.
8. *Forslund S. K., Chakaroun R., Stumvoll M., Bork P.* Combinatorial, additive and dose-dependent drug-microbiome associations // Nature, 2021. Vol. 600. No. 7889. P. 500–505. doi: 10.1038/s41586-021-04177-9.

Поступила в редакцию 02.02.23

ON THE FORMATION OF SETS OF PRECEDENTS BASED ON TABLES OF HETEROGENEOUS FEATURE DESCRIPTIONS BY METHODS OF TOPOLOGICAL THEORY OF DATA ANALYSIS

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: Factorization of the contributions of various variables in the analysis of heterogeneous feature descriptions is an urgent task of complex data mining. The paper proposes the development of the lattice formalism of the

topological theory of data analysis, within which new methods for generating parametric estimates and metrics on lattices formed over the topologies of sets of objects are obtained. The formalism was tested on the problem of forming sets of precedents for conducting chemomicrobiome analysis. Whereas the generation of a set of initial information based on regression coefficients and the difference in the values of the learning material corresponded to an extremely low generalizing ability of custom algorithms (correlation coefficient in the control 0.32 ± 0.20), the use of the proposed estimates for generating sets of precedents in chemomicrobiomics problems made it possible to significantly increase the generalizing ability of the corresponding algorithms (correlation coefficient in control 0.79 ± 0.21).

Keywords: topological data analysis; lattice theory; parametrization of lattice terms; human microbiome; pharmacoinformatics, algebraic approach of Yu. I. Zhuravlev.

DOI: 10.14357/19922264230301 **EDN:** AQEUYO

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. 1998. *Izbrannye nauchnye trudy* [Selected scientific works]. Moscow: Magistr. 420 p.
2. 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.
3. Torshin, I. Yu., O. A. Gromova, I. N. Zakharova, and V. A. Maksimov. 2019. Khemomikrobiomnyy analiz Laktitola [Hemomikrobiomny lactitol analysis]. *Eksperimental'naya i klinicheskaya gastroenterologiya* [Experimental and Clinical Gastroenterology]. 164(4):111–121. doi: 10.31146/1682-8658-ecg-164-4-111-121.
4. 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.
5. 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.
6. Deza, E., and M.-M. Deza. 2006. *Dictionary of distances*. Elsevier B.V. 412 p.
7. Kolmogorov, A. N., and S. V. Fomin. 1989. *Elementy teorii funktsiy i funktsional'nogo analiza* [Elements of the theory of functions and functional analysis]. Moscow: Nauka. 624 p.
8. Forslund, S. K., R. Chakaroun, M. Stumvoll, and P. Bork. 2021. Combinatorial, additive and dose-dependent drug-microbiome associations. *Nature* 600(7889):500–505. doi: 10.1038/s41586-021-04177-9.

Received February 2, 2023

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