ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2024. Т. 18. Вып. 2. С. 48–54

О ПОРОЖДЕНИИ СИНТЕТИЧЕСКИХ ПРИЗНАКОВ НА ОСНОВЕ ОПОРНЫХ ЦЕПЕЙ И ПРОИЗВОЛЬНЫХ МЕТРИК В РАМКАХ ТОПОЛОГИЧЕСКОГО ПОДХОДА К АНАЛИЗУ ДАННЫХ. ЧАСТЬ 2. ЭКСПЕРИМЕНТАЛЬНАЯ АПРОБАЦИЯ НА ЗАДАЧАХ ФАРМАКОИНФОРМАТИКИ*

И. Ю. Торшин[1]

Аннотация: Рассмотрение прецедентных отношений между признаками и таргетной переменной в виде наборов элементов булевой решетки указывает на возможность порождения синтетических признаков с использованием метрических функций расстояния. Сформулированы подходы к (1) оценке релевантности («информативности») метрик по отношению к решаемым задачам, (2) порождению и (3) отбору синтетических признаков, более информативных, чем исходные признаковые описания. Представленные результаты топологического анализа 2400 выборок данных «молекула–свойство» из ProteomicsDB позволили получить достаточно эффективные алгоритмы прогнозирования свойств молекул (ранговая корреляция в кросс-валидации — 0,90 ± 0,23). На данной выборке задач установлены метрики, которые наиболее часто порождают информативные синтетические признаки: максимальное уклонение Колмогорова, «косое» расстояние, метрики Lp, Реньи, фон Мизеса. Для решения изученного комплекса задач показано преимущество полиномных корректоров по сравнению с нейросетевыми и с корректорами типа «случайный лес».

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

DOI: 10.14357/19922264240207 **EDN:** OTXCUD

1 Введение

В первой части работы [1] принимается, что задано регулярное множество прецедентов

$$\mathbf{Q} = \{\mathrm{D}(x_i)|x_i \in \mathbf{X}\}$$

на решетке $L(T(\mathbf{X}))$, порожденное на основе множества исходных описаний объектов $\mathbf{X} = \{x_1, \ldots, x_{N_0}\}$. Для индивидуального объекта $x_i \in \mathbf{X}$ прецедентному соотношению между значениями признаками $\Gamma_k(x_i)$ и t-й таргетной переменной соответствует множество пар $\{(\{\Gamma_k^{-1}(\Gamma_k(x_i)), k = \overline{1,,n}\}, \Gamma_t^{-1}(\Gamma_t(x_i))), i = \overline{1,N_0}, \ k = \overline{1,n}, t = \overline{n+1,n+l}\}$, где l — число таргетных переменных. В рамках топологической теории распознавания прецедентное соотношение между множествами $\{\Gamma_k^{-1}(\Gamma_k(x_i))\}$ и $\Gamma_t^{-1}(\Gamma_t(x_i))$ моделируется как соответствующие массивы расстояний, порождаемые той или иной метрикой ρ_m: $L(T(\mathbf{X}))^2 \to [0\ldots1]$, $m = \overline{1,m_0}$. В [1] предложены способы «встраивания» в формализм полуэмирических расстояний на множествах $a \in L(T(\mathbf{X}))$, векторах $\vec{v}_\alpha[a] = (v_{\alpha_1}[a], v_{\alpha_2}[a], \ldots, v_{\alpha_i}[a], \ldots)$ и функциях $\hat{\phi}(x)\mathbf{\Gamma}_t(u)$.

Здесь для практического приложения формализма сформулированы подходы к исследованию свойств ρ_m, способы оценки релевантности функций ρ_m по отношению к решаемым задачам, способы порождения и отбора синтетических признаков, основанных на ρ_m. Представлены результаты экспериментальной апробации на задачах фармакоинформатики.

2 Об исследовании свойств функций расстояния ρ_m

Рабочая гипотеза настоящего исследования состоит в том, что для порождения более «информативных» признаков могут использоваться полуэмпирические функционалы расстояния на множествах, векторах, функциях [2]. Метрические свойства используемых функций расстояния ρ_m могут исследоваться аналитически или комбина-

*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

торно с использованием аксиом метрики [3]. Для анализа свойств этих функционалов в топологической теории распознавания вводится следующее понятие.

Определение 1. Обобщенной оценочной функцией расстояния будем называть конструкцию вида

$$\rho(a,b) = f(g(v[a \vee b]) - g(v[a \wedge b])),$$

в которой f и g — функции, монотонные на соответствующих участках действительной оси; $v : L \to \to R^+$ — изотонная оценка, для которой выполнено условие оценки (**уО**: $\forall_L a,b : v[a] + v[b] = v[a \wedge b] + + v[a \vee b]$) и изотонности (**уИ**: $\forall_L a,b : a \supseteq b \Rightarrow v[a] \geq \geq v[b]$).

Теорема 1. *Функция расстояния ρ считается обобщенной оценочной функцией расстояния тогда и только тогда, когда $\rho(a,b) = \rho(a \vee b, a \wedge b)$, а термы от a и b в формуле для $\rho(a,b)$ представляют собой композицию монотонной функции и изотонной оценки.*

Необходимость следует из $a \vee b = (a \vee b) \vee (a \wedge b)$ и $a \wedge b = (a \vee b) \wedge (a \wedge b)$ при подстановке $a \vee b$ и $a \wedge b$ вместо a и b в определение 1. Эквивалентность $\rho(a,b)$ и $\rho(a \vee b, a \wedge b)$ указывает на то, что в выражение для вычисления ρ входят термы-функционалы, содержащие выражения $a \vee b$ и $a \wedge b$, взаимозаменяемые с a и b, т. е. термы вида $g'(a \vee b)$ и $g'(a \wedge b)$. По условию теоремы эти термы включают монотонную функцию от изотонной оценки, т. е. g' монотонна. Так как ρ — функция расстояния, то g'-термы не могут входить в выражение для ρ в виде произведения, суммы, отношения, степени или суммы, а только в виде разности, т. е.

$$\rho(a,b) = f\left(g'(a \vee b) - g'(a \wedge b)\right),$$

из чего следует достаточность. Теорема доказана.

Следствие 1. Для обобщенной оценочной ρ

$$\forall \ell \subseteq L(T(\mathbf{X})) : \Delta_{\vee\wedge}(\ell) \equiv 0,$$

$$\Delta_{\vee\wedge}(\ell) = \sum_{a,b \in \ell} |\rho(a,b) - \rho(a \vee b, a \wedge b)| \frac{2}{|\ell|/(|\ell|-1)}.$$

Следствие 2. Выберем «опорное» множество $a \in \in L(T(\mathbf{X}))$ и обобщенную оценочную ρ. При $f(x) = = g(x) = x$ $v_{a,\rho}[b] = \rho(a,b) = \rho(a \vee b, a \wedge b)$ — изотонная оценка.

Следует из того, что любая линейная комбинация изотонных оценок — изотонная оценка при условии положительной определенности (теорема 2 в [4]). Также проверяется прямой подстановкой $v_{a,\rho}[b]$ в уО и уИ.

Следствие 3. Расстояния Фреше–Никодима, Амана, Рэнда/Щекановского, Сокала–Сниса (варианты 1, 2 и 3), Рассела–Рао, Роджера–Танимото, Фейта, Тверского и Юле могут служить обобщенными оценочными функциями расстояния.

Следствие 4. Расстояния Симпсона, Брауна–Бланке, Андерберга и Говера-2 не входят в число обобщенных оценочных функций расстояния.

Теорема 1 со следствиями предоставляет аналитический и комбинаторный инструментарий для исследования свойств полуэмпирических функций расстояния. Если заданная ρ служит обобщенной оценочной функцией расстояния, то могут быть получены соответствующие аналитические выражения для функций f и g. Например, расстояние Сокала–Сниса-2

$$\rho(a,b) = 1 - \frac{|a \cap b|}{|a \cup b| + |a \Delta b|}$$

выступает обобщенным оценочным расстоянием с $f(x) = (e^x - 1)/(0{,}5e^x - 1)$ и $g(x) = \ln(x)$. При невозможности аналитической проверки свойства ρ как обобщенной оценочной могут быть изучены на подмножествах ℓ решетки $L(T(\mathbf{X}))$ посредством вычисления значений функционала $\Delta_{\vee\wedge}(\ell)$ (следствие 1).

3 О способах оценки релевантности метрик ρ_m по отношению к задаче классификации/прогнозирования

Биекция между множеством прецедентов $\mathbf{Q}$ и множеством исходных описаний объектов $\mathbf{X}$, существующая при выполнении условия регулярности по Журавлёву ($\forall \mathrm{x} \in \mathbf{X}, \mathrm{x} = D^{-1}(D(\mathrm{x}))$, гарантирует однозначность соответствия описаний x_i и q_i. Это делает возможным рассматривать прецедентные соотношения, заданные на $\mathbf{Q}$, в терминах множеств $\{\Gamma_k^{-1}(\Gamma_k(x_i))\}$ и $\Gamma_t^{-1}(\Gamma_t(x_i))$ с использованием расстояний ρ_m на подмножествах множества $\mathbf{X}$ [1].

Пусть таргетный класс объектов $\mathbf{c}_{\boldsymbol{\alpha}}$ задан посредством α-го значения t-й переменной $\lambda_{t\alpha} \in \mathrm{I}_t$, $t = \overline{n+1, n+l}$, как $\mathbf{c}_{\boldsymbol{\alpha}} = \Gamma_t^{-1}(\lambda_{t\alpha})$. В случае числовой переменной за $\mathbf{c}_{\boldsymbol{\alpha}}$ может приниматься каждый из элементов $u(\lambda_{t\alpha})$ цепи A_t. Так как $L(T(\mathbf{X}))$ булева, то дополнение множества $\mathbf{c}_{\boldsymbol{\alpha}}$, $\overline{\mathbf{c}}_{\boldsymbol{\alpha}} = \mathbf{X} \backslash \Gamma_t^{-1}(\lambda_{t\alpha})$, определено однозначно. Таким образом, выделение класса $\mathbf{c}_{\boldsymbol{\alpha}}$ порождает задачу классификации $\mathbf{c}_{\boldsymbol{\alpha}}/\overline{\mathbf{c}}_{\boldsymbol{\alpha}}$. Любая задача числового прогнозирования может быть сведена к последовательности корректно решаемых задач $\mathbf{c}_{\boldsymbol{\alpha}}/\overline{\mathbf{c}}_{\boldsymbol{\alpha}}$ [5].

Пусть задано подмножество признаков $p \subseteq [1\ldots n]$ и элемент решетки $c \in L(T(\mathbf{X}))$. Определим функцию

$$\boldsymbol{\rho}_{\mathbf{mc}}(x_i, c, p) = \{\rho_m(c, \Gamma_k^{-1}(\Gamma_k(x_i)),\ k \in p)\}.$$

При заданных ρ_m, p, $\mathbf{c}_{\boldsymbol{\alpha}}$ и $\overline{\mathbf{c}}_{\boldsymbol{\alpha}}$ для x_i вычислимы множества расстояний $\boldsymbol{\rho}_{\mathbf{mc}}(x_i, \mathbf{c}_{\boldsymbol{\alpha}}, p)$ и $\boldsymbol{\rho}_{\mathbf{mc}}(x_i, \overline{\mathbf{c}}_{\boldsymbol{\alpha}}, p)$. Обозначим

$$\boldsymbol{\rho}_{\mathbf{m}\boldsymbol{\alpha}}(x_i) = \boldsymbol{\rho}_{\mathbf{mc}}(x_i, \mathbf{c}_{\boldsymbol{\alpha}}, [1\ldots n]);$$
$$\boldsymbol{\rho}_{\mathbf{m}\overline{\boldsymbol{\alpha}}}(x_i) = \boldsymbol{\rho}_{\mathbf{mc}}(x_i, \overline{\mathbf{c}}_{\boldsymbol{\alpha}}, [1\ldots n]).$$

Для $x_i \in \mathbf{X}$ определено множество

$$\boldsymbol{\rho}_{\mathbf{m}}(x_i, p) = \{\rho_{mk_1k_2}(x_i, p) = \rho_m\left(\Gamma_{k_1}^{-1}\left(\Gamma_{k_1}(x_i), \Gamma_{k_2}^{-1}\left(\Gamma_{k_2}(x_i)\right)\right)\right),$$
$$k_1, k_2 \in p,\ k_1 \neq k_2\},\ \boldsymbol{\rho}_{\mathbf{m}}(x_i) = \boldsymbol{\rho}_{\mathbf{m}}(x_i, [1\ldots n]).$$

На основе $\boldsymbol{\rho}_{\mathbf{m}\boldsymbol{\alpha}}(x_i)$ и $\boldsymbol{\rho}_{\mathbf{m}\overline{\boldsymbol{\alpha}}}(x_i)$ вводятся оценки релевантности ρ_m. По отношению к задаче $\mathbf{c}_{\boldsymbol{\alpha}}/\overline{\mathbf{c}}_{\boldsymbol{\alpha}}$ более релевантна или «информативна» такая метрика ρ_m, которая для всех $x \in \mathbf{c}_{\boldsymbol{\alpha}}$ минимизирует расстояния в списке $\boldsymbol{\rho}_{\mathbf{m}\boldsymbol{\alpha}}(x)$ и максимизирует расстояния в списке $\boldsymbol{\rho}_{\mathbf{m}\overline{\boldsymbol{\alpha}}}(x)$ (т. е. «приближает» объекты к их классам). Выделены два взаимосвязанных направления дальнейших исследований:

(1) нахождение подмножеств p признаков, «более информативных» для ρ_m;

(2) настройка/выбор ρ_m при фиксированном p.

Для $c' \in L(T(\mathbf{X}))$ определим $\vartheta_{\mathbf{mc}}$, операцию слияния списков $\boldsymbol{\rho}_{\mathbf{mc}}$:

$$\vartheta_{\mathbf{mc}}(c', c, p) = \bigcup_{y \in c'} \boldsymbol{\rho}_{\mathbf{mc}}(y, c, p).$$

Обозначим

$$\vartheta_{\mathbf{m}\boldsymbol{\alpha}}(\mathbf{c}, p) = \vartheta_{\mathbf{mc}}(\mathbf{c}, \mathbf{c}_{\boldsymbol{\alpha}}, p);\ \vartheta_{\mathbf{m}\boldsymbol{\alpha}}(\mathbf{c}, p) = \vartheta_{\mathrm{mc}}(\mathbf{c}, \overline{\mathbf{c}}_{\boldsymbol{\alpha}}, p),$$

вычислим множества $\vartheta_{\mathbf{m}\boldsymbol{\alpha}}(\mathbf{c}_{\boldsymbol{\alpha}}, p)$ и $\vartheta_{\mathbf{m}\boldsymbol{\alpha}}(\overline{\mathbf{c}}_{\boldsymbol{\alpha}}, p)$ и сформируем эмпирические функции распределения (э.ф.р.) $\hat{\phi}(x)\vartheta_{\mathbf{m}\boldsymbol{\alpha}}(\mathbf{c}_{\boldsymbol{\alpha}}, p)$ и $\hat{\phi}(x)\vartheta_{\mathbf{m}\boldsymbol{\alpha}}(\overline{\mathbf{c}}_{\boldsymbol{\alpha}}, p)$. На пространстве однородных монотонно возрастающих функций

$$\mathbf{M}_{0\ldots1}^{+} = \{f : [0\ldots1] \to [0\ldots1],\ x \geq y \Rightarrow f(x) \geq f(y)\}$$

введем функционал расстояния d_f: $\mathbf{M}_{0..1}^{+} \to [0\ldots1]$ (максимальное уклонение Колмогорова $D(f(x), g(x)) = \sup_x |f(x) - g(x)|$, метрики фон Мизеса, Реньи и др.). Выбор d_f делает возможной постановку ряда задач топологического анализа данных:

(1) количественные оценки релевантности ρ_m как $d_f(\hat{\phi}(x)\vartheta_{\mathbf{m}\boldsymbol{\alpha}}(\mathbf{c}_{\boldsymbol{\alpha}}, p), \hat{\phi}(x)\vartheta_{\mathbf{m}\boldsymbol{\alpha}}(\overline{\mathbf{c}}_{\boldsymbol{\alpha}}, p))$ для разных $\mathbf{c}_{\boldsymbol{\alpha}}$, $\lambda_{t\alpha} \in \mathrm{I}_t$, $\alpha = \overline{1, |\mathrm{I}_t|}$;

(2) задачи оптимизации для увеличения разделения классов $\mathbf{c}_{\boldsymbol{\alpha}}/\overline{\mathbf{c}}_{\boldsymbol{\alpha}}$ ($\arg\max_{\rho_m, p} d_f(\hat{\phi}\vartheta_{\mathbf{m}\boldsymbol{\alpha}}(\overline{\mathbf{c}}_{\boldsymbol{\alpha}}, p), \hat{\phi}\vartheta_{\mathbf{m}\boldsymbol{\alpha}}(\mathbf{c}_{\boldsymbol{\alpha}}, p))$, $\arg\max_{\rho_m, p} d_f(\hat{\phi}\vartheta_{\mathbf{m}\overline{\boldsymbol{\alpha}}}, (\overline{\mathbf{c}}_{\boldsymbol{\alpha}}, p), \hat{\phi}\vartheta_{\mathbf{m}\overline{\boldsymbol{\alpha}}}(\mathbf{c}_{\boldsymbol{\alpha}}, p))$ и др.);

(3) определение ρ_q-метрик на пространстве объектов [2, с. 184–199] (например, в виде $d_f(\hat{\phi}\boldsymbol{\rho}_{\mathbf{m}\boldsymbol{\alpha}}(x, p), \hat{\phi}\boldsymbol{\rho}_{\mathbf{m}\boldsymbol{\alpha}}(y, p)), d_f(\hat{\phi}\boldsymbol{\rho}_{\mathbf{m}}(x, p), \hat{\phi}\boldsymbol{\rho}_{\mathbf{m}}(y, p))$);

(4) оценка близости метрик ρ_q к метрике разреза по классам $\mathbf{c}_{\boldsymbol{\alpha}}/\overline{\mathbf{c}}_{\boldsymbol{\alpha}}$;

(5) формулировка критериев разрешимости/регулярности задачи $\mathbf{c}_{\boldsymbol{\alpha}}/\overline{\mathbf{c}}_{\boldsymbol{\alpha}}$ [6];

(6) оценки компактности классов $\mathbf{c}_{\boldsymbol{\alpha}}$ и $\overline{\mathbf{c}}_{\boldsymbol{\alpha}}$ [3].

4 О способах порождения и отбора синтетических признаков на основании функций расстояния

Множества $\boldsymbol{\rho}_{\mathbf{m}\boldsymbol{\alpha}}(x_i, p)$, $\boldsymbol{\rho}_{\mathbf{m}\overline{\boldsymbol{\alpha}}}(x_i, p)$ и $\boldsymbol{\rho}_{\mathbf{m}}(x_i)$ и отдельные $\rho_m(\mathbf{c}_{\boldsymbol{\alpha}}, \Gamma_k^{-1}(\Gamma_k(x_i)))$ используются для формирования синтетических числовых признаков $\Gamma_{k'}(x_i)$ объекта x_i, $k' = \overline{n + l + 1, n + l + n_S}$. Значение синтетического признака $\Gamma_{k'}(x_i)$ зависит от выбора ρ_m, классов $\mathbf{c}_{\boldsymbol{\alpha}}$ и $\overline{\mathbf{c}}_{\boldsymbol{\alpha}}$ и от способа его вычисления:

(1) $\rho_m(\mathbf{c}_{\boldsymbol{\alpha}}, \Gamma_k^{-1}(\Gamma_k(x_i)))$;

(2) $\rho_m(\overline{\mathbf{c}}_{\boldsymbol{\alpha}}, \Gamma_k^{-1}(\Gamma_k(x_i)))$;

(3) $\rho_m(\mathbf{c}_{\boldsymbol{\alpha}}, \ldots) - \rho_m(\overline{\mathbf{c}}_{\boldsymbol{\alpha}}, \ldots)$;

(4) $1 - \rho_m(\mathbf{c}_{\boldsymbol{\alpha}}, \ldots)$;

(5) значения э.ф.р. $\hat{\phi}(x)\boldsymbol{\rho}_{\mathbf{m}\boldsymbol{\alpha}}(x_i, p)$ при разных x (например, соответствующих процентилям $\hat{\phi}\boldsymbol{\rho}_{\mathbf{m}\boldsymbol{\alpha}}(x_i, p)$);

(6) значения $\hat{\phi}(x)\boldsymbol{\rho}_{\mathbf{m}\overline{\boldsymbol{\alpha}}}(x_i, p)$ при разных x;

(7) $\hat{\phi}(x + \Delta x)\boldsymbol{\rho}_{\mathbf{m}\boldsymbol{\alpha}}(x_i, p) - \hat{\phi}(x)\boldsymbol{\rho}_{\mathbf{m}\boldsymbol{\alpha}}(x_i, p)$ и $\hat{\phi}(x + + \Delta x)\boldsymbol{\rho}_{\mathbf{m}\overline{\boldsymbol{\alpha}}}(x_i, p) - \hat{\phi}(x)\boldsymbol{\rho}_{\mathbf{m}\overline{\boldsymbol{\alpha}}}(x_i, p)$, где Δx — шаг.

Кроме того, $\mathbf{c}_{\boldsymbol{\alpha}}$ может определяться как $\Gamma_t^{-1}(\lambda_{t\alpha})$ или как $u(\lambda_{t\alpha})$; если $\mathbf{c}_{\boldsymbol{\alpha}} = \Gamma_t^{-1}(\lambda_{t\alpha})$, то $\overline{\mathbf{c}}_{\boldsymbol{\alpha}}$ может быть равно $\Gamma_t^{-1}(\lambda_{t\alpha+1})$; классы $\mathbf{c}_{\boldsymbol{\alpha}}/\overline{\mathbf{c}}_{\boldsymbol{\alpha}}$ t-й переменной могут определяться с использованием разбиений на различные процентили (которые определяются как подвыборка значений $\lambda_{t\alpha} \in \mathrm{I}_t$) и т. д.

Таким образом, предлагаемые схемы порождают значительное число синтетических признаков $\Gamma_{k'}(x_i)$ ($10n$ и более при n исходных признаках Γ_k), что делает необходимым введение процедур отбора признаков. Таргетная переменная $\Gamma_t(x_i)$ — числовая, и порождаемые признаки $\Gamma_{k'}(x_i)$ — также числовые. Для данного случая в прикладной математике имеется несколько различных подходов к оценке взаимосвязи $\Gamma_t(x_i)$ и $\Gamma_{k'}(x_i)$: корреляционные оценки (для линейных закономерностей), полиномная аппроксимация с оценкой качества (для нелинейных закономерностей) и методы теории вероятностей / математической статистики, не зависящие от вида закономерности (в том числе на основе «взаимной информации» [7]).

Наиболее фундаментальным представляется тестирование взаимосвязи двух переменных на основе «нулевой гипотезы» об их независимости. Пусть заданы пары тестируемых значений, (x_i, y_i), $i = \overline{1, \mathbf{n}_{(\mathrm{x,y})}}$, э.ф.р. $F_{xy}(x, y)$ характеризует совместное распределение x и y, а э.ф.р. $F_x(x)$ и $F_y(y)$ — индивидуальные распределения переменных. Эмпирическая функция распределения нулевой гипотезы (независимость x и y) определяется как $F_x(x)F_y(y)$.

Для оценки отличий между $F_{xy}(x, y)$ и $F_x(x)F_y(y)$ необходимо ввести расстояние между такими функциями (так называемую «статистику») и оценить достоверность различий посредством того или иного статистического теста. В качестве расстояния можно использовать функции d_f, адаптированные для 2-мерного случая (например, максимальное уклонение $D(\mathrm{F}_{xy}(x, y), \mathrm{F}_x(x)\mathrm{F}_y(y)) = \max(|\mathrm{F}_{xy}(x_i, y_i) - \mathrm{F}_x(x_i)\mathrm{F}_y(y_i)|)$) и статистический тест Колмогорова–Смирнова $P_{\mathrm{KC}}(D(\mathrm{F}_{xy}(x, y), \mathrm{F}_x(x)\mathrm{F}_y(y)), n_{(x,y)})$. Тогда $1 - P_{\mathrm{KC}}$ характеризует «информативность» x относительно y.

Более универсальным подходом к оценке достоверности различий между $\mathrm{F}_{xy}(x, y)$ и $\mathrm{F}_x(x)\mathrm{F}_y(y)$ считается прямое вычисление выбранной статистики d_f на множествах пар значений (x_i, y_i), полученных датчиком случайных чисел.

Пусть *оператор* $\hat{\zeta}$, *семплирующий* множество $\mathbf{X}$, формирует набор семплов

$$\hat{\zeta}\mathbf{X} = \{a_1, a_2, \ldots, a_k, \ldots, a_{|\hat{\zeta}X|} | a_k \subset \mathbf{X}\},$$

а процедура random — датчик случайных чисел (в диапазоне $[0 \ldots 1]$). Для каждого семпла a_k принимается, что $n_{(x,y)} = |a_k|$, и вычисляется множество значений d_f для случайных выборок,

$$\mathrm{rnd}\,(\hat{\zeta}\mathbf{X}, d_f) = \left\{ d_f \left(\mathrm{F}_{xy}(x_{ij}, y_{ij}), \mathrm{F}_x(x_{ij})\mathrm{F}_y(y_{ij}), \right.\right.$$
$$\left.\left. x_{ij}, y_{ij} = \mathrm{random},\ j = \overline{1, |a_i|} \right),\ i = \overline{1, \left|\hat{\zeta}X\right|} \right\}.$$

Для $a \in \hat{\zeta}\mathbf{X}$ значение $P(d_f, \hat{\zeta}\mathbf{X}, a, k', t) = 1 - \hat{\phi}(d_f(\mathrm{F}_{k't}(\Gamma_{k'}(z), \Gamma_t(z)), F_{k'}(\Gamma_{k'}(z))\mathrm{F}_t(\Gamma_t(z)))|z \in a)\mathrm{rnd}\,(\hat{\zeta}\mathbf{X}, d_f)$ — статистическая достоверность «зависимости» $\Gamma_t(z)$ и $\Gamma_{k'}(z)$ по статистике d_f на семпле a, а $1 - P(d_f, \hat{\zeta}\mathbf{X}, a, k', t)$ количественно оценивает зависимость.

При заданном способе оценки зависимости $1 - P(d_f, \hat{\zeta}\mathbf{X}, a, k', t)$ задача отбора информативных признаков решается посредством так называемого B-алгоритма, исходно разработанного для построения оптимальных словарей финальных информаций (чему и соответствует литера «B») [8]. Данный алгоритм, основанный на критерии разрешимости по Журавлёву, позволяет выбирать множества финальных информаций на основе максимального частичного покрытия при минимуме элементов покрытия. Замена мощности пересечения множеств на $1 - P(d_f, \hat{\zeta}\mathbf{X}, a, k', t)$ приведет к тому, что B-алгоритм будет выбирать минимум признаков с максимальной «информативностью» (наиболее информативные признаки, см. теоремы 1, 7 и 8 работы [8]).

Таким образом, в рамках развиваемого формализма синтез более информативных синтетических $\Gamma_{k'}(x_i)$ осуществляется в 5 стадий:

(1) определяется набор исходных (как правило, «низкоинформативных») признаков $\Gamma_k(x_i)$ и таргетная переменная $\Gamma_t(x_i)$;

(2) вводится набор метрик ρ_m, оценивается их релевантность $d_f(\hat{\phi}(x)\vartheta_{\mathbf{m}\boldsymbol{\alpha}}(\mathbf{c}_{\boldsymbol{\alpha}}, p), \hat{\phi}(x)\vartheta_{\mathbf{m}\boldsymbol{\alpha}}(\overline{\mathbf{c}}_{\boldsymbol{\alpha}}, p))$ для каждого класса $\mathbf{c}_{\boldsymbol{\alpha}}$ значений t-й переменной и отбираются наиболее релевантные ρ_m;

(3) посредством каждой из отобранных ρ_m порождаются синтетические признаки $\Gamma_{k'}(x_i)$;

(4) посредством вычислений $1 - P(d_f, \hat{\zeta}\mathbf{X}, a, k', t)$ и B-алгоритма отбирается минимальное число признаков максимальной «информативности»;

(5) применяется алгоритм прогнозирования таргетной переменной (корректор по Журавлёву–Рудакову).

5 Экспериментальная апробация

Формализм апробирован на комплексе задач фармакоинформатики: получение количествен-

Ранговые корреляции между экспериментальными и расчетными значениями EC_{50} и других величин хемокиномного анализа: r — коэффициент ранговой корреляции на обучении; r_c — на контроле. Усреднение r и r_c проводилось по 2400 выборкам хемокиномных данных

Эксперимент	r	r_c
f_{θ_k}-алгоритмы, корректор — нейросеть	$\mathbf{0{,}88 \pm 0{,}15}$	$\mathbf{0{,}86 \pm 0{,}20}$
Синтетические $\Gamma_{k'}(x_i)$, корректор — нейросеть (2 слоя)	$0{,}45 \pm 0{,}22$	$0{,}22 \pm 0{,}21$
Синтетические $\Gamma_{k'}(x_i)$, корректор — нейросеть (10 слоев)	$0{,}52 \pm 0{,}25$	$0{,}21 \pm 0{,}20$
Синтетические $\Gamma_{k'}(x_i)$, корректор — «случайный лес», вариант 1	$0{,}98 \pm 0{,}15$	$0{,}67 \pm 0{,}31$
Синтетические $\Gamma_{k'}(x_i)$, корректор — «случайный лес», вариант 2	$0{,}99 \pm 0{,}14$	$0{,}71 \pm 0{,}35$
Синтетические $\Gamma_{k'}(x_i)$, полиномные корректоры, вариант 1	$\mathbf{0{,}93 \pm 0{,}11}$	$\mathbf{0{,}90 \pm 0{,}23}$
Синтетические $\Gamma_{k'}(x_i)$, полиномные корректоры, вариант 2	$\mathbf{0{,}95 \pm 0{,}08}$	$\mathbf{0{,}86 \pm 0{,}27}$

ных оценок ингибирования киназ протеома перспективными лекарствами (хемокиномный анализ) [9]. Использованы 2400 выборок данных «молекула–свойство» из ProteomicsDB; свойства молекул включили константы EC_{50} и активности для концентраций $(E_j(C_i))$.

Исходные признаки $\Gamma_k(x_i)$ определялись как булевы инварианты над множествами χ-цепей и χ-узлов хемографов x_i, как и в [9]. Таргетная $\Gamma_t(x_i)$ определялась как числовое значение прогнозируемого свойства. В качестве ρ_m использовались функции расстояния на множествах, векторах и э.ф.р. (всего 65 функций из справочника [2]). Классы $\mathbf{c}_{\boldsymbol{\alpha}}$ определялись как квартили значений Γ_t. Векторы элементов $L(T(\mathbf{X}))$ формировались из оценок v_α^+, v_α^- и d_α [4] для каждого $\mathbf{c}_{\boldsymbol{\alpha}}$. Релевантность ρ_m по $d_f(\hat{\phi}(x), \vartheta_{\mathbf{m}\boldsymbol{\alpha}}(\mathbf{c}_{\boldsymbol{\alpha}}, p), \hat{\phi}(x)\vartheta_{\mathbf{m}\boldsymbol{\alpha}}(\overline{\mathbf{c}}_{\boldsymbol{\alpha}}, p))$ оценивалась для каждого $\mathbf{c}_{\boldsymbol{\alpha}}$, d_f — максимальное уклонение. Синтетические признаки $\Gamma_{k'}(x_i)$ порождались всеми перечисленными выше способами; их отбор проводился B-алгоритмом с использованием $1 - P(d_f, \hat{\zeta}\mathbf{X}, a, k', t)$.

В качестве корректоров использовались нейронные сети с несколькими слоями (от 2 до 10) с функцией активации softmax, полиномы различных конструкций (более 20 формул, в том числе квазиполиномные модели с элементарными функциями) и «случайные леса» решающих деревьев. Оператор семплирования $\hat{\zeta}$ был реализован как десятикратная кросс-валидация с делением каждой выборки объектов на 80% (обучение) и 20% (контроль). Результаты экспериментов суммированы в таблице.

Наилучший результат применения нового «топологического» формализма с полиномным корректором ($r_c = 0{,}90 \pm 0{,}23$) немного превзошел наилучший результат применения метода опорных функций (композиций вида $f_{\theta_k} = g(f_1(\sum \omega_k^j x_k), \ldots$ $\ldots, f_l(\sum \omega_k^j x_k))$, см. [9]), для которого $r_c = 0{,}86 \pm$ $\pm\, 0{,}20$. Полиномными формулами, наиболее часто показывавшими наилучший результат, оказались полиномы 1-й или 2-й степеней с произведениями переменных первой степени, полиномы 5-й степени, квазиполиномы 5-й степени с сигмоидами и Фурье-полиномы 3-й степени.

Нейросетевые корректоры всех использованных конфигураций отличались крайне низкими показателями ($r = 0{,}45 \pm 0{,}22$, $r_c = 0{,}22 \pm 0{,}21$), а «случайный лес» приводил к существенному переобучению (см. таблицу). При этом в 290 из 2400 выборок данных (12%) «случайный лес» приводил к улучшению результатов по сравнению с наилучшими полиномными корректорами, а в 1670 из 2400 выборок данных (70%) — к ухудшению.

Анализ синтетических признаков $\Gamma_{k'}(x_i)$, вошедших в наилучшие полиномные модели, показал, что среди более информативных (по оценке $1 - P(d_f, \hat{\zeta}\mathbf{X}, a, k', t)$) признаков чаще всего встречались признаки, порождаемые с использованием э.ф.р. на основе опорных цепей (теорема 1 в 1-й части работы [1]), среди наименее информативных — исходные признаки $\Gamma_k(x_i)$ и признаки на основе отдельных расстояний $\rho_m(\mathbf{c}_{\boldsymbol{\alpha}}, \Gamma_k^{-1}(\Gamma_k(x_i))$. Функциями ρ_m, наиболее часто порождающими информативные $\Gamma_{k'}(x_i)$ на пространстве э.ф.р., оказались максимальное уклонение Колмогорова, «косое» расстояние, метрики Lp, Реньи, $\chi 2$, фон Мизеса, инженерная [2]. В среднем по всем выборкам данных эти 7 разновидностей ρ_m порождали более 50% самых информативных признаков $\Gamma_{k'}(x_i)$, отобранных B-алгоритмом.

6 Заключение

Предлагаемый подход к порождению информативных синтетических признаков подразумевает последовательные трансформации описаний объекта:

(1) исходное множество значений признаков;
(2) множество соответствующих элементов решетки;
(3) множество расстояний (измеряемых посредством ρ_m) между элементами решетки, соответствующими классам и признакам;

(4) множество э.ф.р. расстояний, измеренных заданными ρ_m;

(5) множество синтетических признаков объекта.

Использование многочисленных метрик на стадии порождения признаков позволяет рассматривать развиваемый формализм как вариант развития идеологии ABO (алгоритмы вычисления оценок) научной школы Ю. И. Журавлёва. Экспериментальная апробация предлагаемого подхода на 2400 однородных задачах фармакоинформатики позволила повысить аккуратность и обобщающую способность алгоритмов.

Литература

1. *Торшин И. Ю.* 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 // Информатика и её применения, 2024. Т. 18. Вып. 1. С. 71–77. doi: 10.14357/19922264240110. EDN: RIVOXR.
2. *Деза Е. И., Деза М. М.* Энциклопедический словарь расстояний / Пер. с англ. — М.: Наука, 2008. 444 с. (*Deza E. I., Deza M. M.* Dictionary of distances. — North-Holland: Elsevier, 2006. 412 p. doi: 10.1016/B978-0-444-52087-6.X5000-8.)
3. *Torshin I. Y., Rudakov K. V.* Combinatorial analysis of the solvability properties of the problems of recognition and completeness of algorithmic models. Part 2: Metric approach within the framework of the theory of classification of feature values // Pattern Recognition Image Analysis, 2017. Vol. 27. No. 2. P. 184–199. doi: 10.1134/S1054661817020110.
4. *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. 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.
6. *Torshin I. Y., Rudakov K. V.* Combinatorial analysis of the solvability properties of the problems of recognition and completeness of algorithmic models. Part 1: Factorization approach // Pattern Recognition Image Analysis, 2017. Vol. 27. No. 1. P. 16–28. doi: 10.1134/S1054661817010151.
7. *Sosa-Cabrera G., Gómez-Guerrero S., García-Torres M., Schaerer C. E.* Feature selection: A perspective on inter-attribute cooperation // Int. J. Data Science Analytics, 2024. Vol. 17. P. 139–151. doi: 10.1007/s41060-023-00439-z.
8. *Torshin I. Y.* Optimal dictionaries of the final information on the basis of the solvability criterion and their applications in bioinformatics // Pattern Recognition Image Analysis, 2013. Vol. 23. No. 2. P. 319–327. doi: 10.1134/S1054661813020156.
9. *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.

Поступила в редакцию 09.04.24

ON THE GENERATION OF SYNTHETIC FEATURES BASED ON SUPPORT CHAINS AND ARBITRARY METRICS WITHIN THE FRAMEWORK OF A TOPOLOGICAL APPROACH TO DATA ANALYSIS. PART 2. EXPERIMENTAL TESTING ON PHARMACOINFORMATICS PROBLEMS

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: Consideration of precedent relationships between features and a target variable in the form of sets of Boolean lattice elements indicates the possibility of generating synthetic features using metric distance functions. Approaches to (i) assessing the relevance ("informativeness") of metrics in relation to the problems being solved; (ii) generating; and (iii) selecting synthetic features that are more informative than the original feature descriptions are formulated. The results of topological analysis of 2400 samples of "molecule–property" data from ProteomicsDB made it possible to obtain fairly effective algorithms for predicting the properties of molecules (rank correlation in cross-validation is 0.90 ± 0.23). Using this sample of problems, metrics have been established

that most often generate informative synthetic features: maximum Kolmogorov deviation, “oblique” distance, and Lp, Renyi, and von Mises metrics. To solve the studied set of problems, the advantage of polynomial correctors compared to neural network and random forest correctors is shown.

Keywords: topological data analysis; lattice theory; algebraic approach of Yu. I. Zhuravlev; pharmacoinformatics

DOI: 10.14357/19922264240207 **EDN:** OTXCUD

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. Torshin, I. Yu. 2024. O porozhdenii sinteticheskikh priznakov na osnove opornykh tsepey i proizvol’nykh metrik v ramkakh topologicheskogo podkhoda k analizu dannykh. Chast’ 1. Vklyuchenie v formalizm empiricheskikh funktsiy rasstoyaniya [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]. *Informatika i ee Primeneniya — Inform Appl.* 18(1):71–77. doi: 10.14357/19922264240110. EDN: RIVOXR.
2. Deza, E. I., and M. M. Deza. 2006. *Dictionary of distances*. North-Holland: Elsevier. 412 p. doi: 10.1016/B978-0-444-52087-6.X5000-8.
3. 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. doi: 10.1134/S1054661817020110.
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. Torshin, I. Y., and K. V. Rudakov. 2017. Combinatorial analysis of the solvability of the problems of recognition, completeness of algorithmic models. Part 1: Factorization approach. *Pattern Recognition Image Analysis* 27(1):16–28. doi: 10.1134/S1054661817010151.
7. Sosa-Cabrera, G., S. Gymez-Guerrero, M. García-Torres, and C. E. Schaerer. 2024. Feature selection: A perspective on inter-attribute cooperation. *Int. J. Data Science Analytics* 17:139–151. doi: 10.1007/s41060-023-00439-z.
8. Torshin, I. Y. 2013. Optimal dictionaries of the final information on the basis of the solvability criterion and their applications in bioinformatics. *Pattern Recognition Image Analysis* 23(2):319–327. doi: 10.1134/S1054661813020156.
9. 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.

Received April 9, 2024

Contributor

Torshin Ivan Y. (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