ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2025. Т. 19. Вып. 2. С. 18–25

МЕТРИЗАЦИЯ ДИСКРЕТНЫХ ТОПОЛОГИЧЕСКИХ ПРОСТРАНСТВ В КОНТЕКСТЕ ТЕОРИИ РЕШЕТОК. ЧАСТЬ 2. ПРАКТИЧЕСКИЙ АНАЛИЗ СЛЕДСТВИЙ ТЕОРЕМЫ О РЕГУЛЯРНОСТИ И НОРМАЛЬНОСТИ

И. Ю. Торшин[1]

Аннотация: В первой части доказана теорема, связывающая понятие нормальности топологических пространств с регулярностью по Ю. И. Журавлёву в контексте проблемы метризации признаковых пространств. Следствия теоремы позволяют систематизировать поиск наиболее приемлемых проблемно-ориентированных метрик с учетом индивидуальных весов атомов решетки, порядка атомов, способа формирования метрики (на множества, на функциях, на векторах) и способа настройки параметров метрик. Предложены новые способы порождения синтетических признаков с использованием опорных классов значений или же с использованием целых опорных цепей (отбор наиболее информативных элементарных признаков, максимизация информативности элементарных признаков). Систематизированы перспективные направления дальнейших исследований, в том числе переход к решетке значений признаков и перспективные функционалы порождения синтетических признаков. Приведены результаты соответствующих вычислительных экспериментов, указывающие на снижение значений показателей переобучения и на повышение качества работы алгоритмов числового прогнозирования с использованием предлагаемых способов порождения метрик и синтетических признаков.

Ключевые слова: топологический анализ данных; алгебраический подход; вычислительный эксперимент; синтетические признаки

DOI: 10.14357/1 **EDN:** T

1 Введение

В топологической теории распознавания изучаются методы систематизации порождения метрик и синтетических признаковых описаний, наиболее адекватных прикладным задачам [1]. Порождение синтетических признаков осуществляется посредством метода опорных функций [2] или на основе ρ_L-метрик над булевой решеткой $L(T(\mathbf{X}))$ [3], где $\mathbf{X}$ — множество исходных описаний объектов, изоморфное *множеству прецедентов*

$$Q = \varphi(\mathbf{X}) = \{D(x_i)|x_i \in \mathbf{X}\}\,,$$
$$D(x_i) = (\Gamma_1(x_i) \times \cdots \times \Gamma_k(x_i) \times \cdots \times \Gamma_{n+l}(x_i))_{\Delta}\,;$$

функции $\Gamma_k(x_i)$ вычисляют значения *элементарных признаков* для объекта x_i.

В первой части работы доказана теорема о регулярности и нормальности топологических пространств, показавшая, что регулярность $\mathbf{X}$ по Ю. И. Журавлёву (условие $\varphi^{-1}(\varphi(\mathbf{X}) = \mathbf{X})$ гарантирует нормальность и, следовательно, метризуемость топологического пространства $(\mathbf{X}, T(\mathbf{X}))$ [1]. Из теоремы о регулярности и нормальности были выведены 8 следствий, причем следствия 3–5 указывают на определенные способы порождения метрик над дискретным топологическим пространством $(\mathbf{X}, T(\mathbf{X}))$, однозначно соответствующим булевой решетке $L(T(\mathbf{X}))$.

2 Расстояния на основе функций от множеств (следствие 3)

В данном следствии принимается одинаковый вклад каждого атома решетки в числовую оценку пути между множествами a и b, так что метрики над $L(T(\mathbf{X}))$ представлены в виде (квазилинейных) композиций функций $f : L(T(\mathbf{X})) \to \mathrm{R}$ от множеств a, b, $a/\backslash b$. Если ρ_L-метрика по следствию 3 настраиваема (что обозначим как $\rho_{L\vec{w}}$, где $\vec{w}$ — вектор весов), то настройка $\vec{w}$ может проводиться итеративно (стохастическая оптимизация) или посредством процедур, аналогичных методу наименьших квадратов (МНК).

Рассмотрим расстояния:

(1) на множествах;

(2) на функциях;

(3) на векторах.

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

3 Расстояния по следствию 3 на множествах

В теории множеств важнейшей функцией от множества считается *мощность*, $f(a) = |a|$. При $f(a) = |a|$ метрикам по следствию 3 соответствуют все известные полуэмпирические расстояния на множествах [4], в том числе метрики на основе решеточных оценок [5–7]. Соответствующие определения ρ_L представимы в виде $g_1(|a|) \circ g_2(|b|) \circ g_3(|a/\backslash b|) \circ \ldots$, $g_\alpha : \mathrm{R}^+ \to \mathrm{R}$.

Конкретный вид функций g_α и их композиций определим по результатам анализа известных метрик на множествах [4]:

(1) все такие метрики — дроби;

(2) все числители дробей — линейные комбинации $|a|$, $|b|$, $|a/\backslash b|$, $|\overline{a}|$, $|\overline{b}|$ и др., парных произведений мощностей, корней квадратных от парных произведений;

(3) знаменатели этих дробей по структуре подобны числителям и могут включать минимум/максимум от $|a|$, $|b|$, $|\mathbf{X}|^2$, произведение $|a||\overline{a}||b||\overline{b}|$ или его квадратный корень.

Для $a, b \in L(T(\mathbf{X}))$ определим алфавит мощностей

$$A(a,b) = \{1, |a|, |b|, |a/\backslash b|, |a\backslash/b|, |a\Delta b|, |a/b|, \\ |b/a|, |\overline{a}|, |\overline{a}||a|, \ldots\},$$

так что $A^2(a,b)$ включает все элементы $A(a,b)$ (с единицей) и все пары элементов $A(a,b)$. Пусть $\tilde{A}^2(a,b)$ — вектор, содержащий произведения каждого из элементов $A^2(a,b)$, а $\vec{w}_{A^2}, \vec{w}'_{A^2} \in \mathrm{R}^{|A^2(a,b)|}$ — веса элементов квадрата алфавита. Тогда *обобщенное расстояние между* a, b определено как $\vec{w}_{A^2}\tilde{A}^2(a,b)/(\vec{w}'_{A^2}\tilde{A}^2(a,b))$.

4 Расстояния по следствию 3 на функциях

При задании опорной цепи A_t [7], соответствующей t-й переменной с областью определения I_t, каждому $a \in L(T(\mathbf{X}))$ соответствует множество чисел $\mathbf{\Gamma}_t(a) = \{\Gamma_t(x), x \in a\}$ и определена эмпирическая функция распределения (ЭФР)

$$\hat{\Phi}(x)\mathbf{\Gamma}_t(a) = \sup|\,\{|B \subseteq \mathbf{\Gamma}_t(a)|\forall\,\beta \in B : \\ \beta \le x\}\,|/|\mathbf{\Gamma}_t(a)|, \;\; x \in \mathbb{R}$$

(кратко — $\hat{\Phi}\mathbf{\Gamma}_t(a)$, а функцию плотности обозначим $\hat{\phi}\mathbf{\Gamma}_t(a)$). Использование ЭФР позволяет представить функцию от множества a, фигурирующую в следствии 3, в виде $\hat{\Phi}\mathbf{\Gamma}_t(a)$ и оценивать расстояние $d_{\vec{w}}$ между множествами a и b как расстояние между $\hat{\Phi}\mathbf{\Gamma}_t(a)$ и $\hat{\Phi}\mathbf{\Gamma}_t(b)$, т. е.

$$\rho_{L\vec{w}} = d_{\vec{w}}(\hat{\Phi}\mathbf{\Gamma}_t(a), \hat{\Phi}\mathbf{\Gamma}_t(b)).$$

Расстояния $d_{\vec{w}}$ формулируются в виде функционалов над массивом $\{d\Phi(x,t,a,b) = \hat{\Phi}(x)\mathbf{\Gamma}_t(a) - \hat{\Phi}(x)\mathbf{\Gamma}_t(b), x \in \mathbf{I}_t\}$: расстояние Колмогорова $\max\{|d\Phi(x,t,a,b)|\}$, Крамера – фон Мизеса — $\sum_x d\Phi(x,t,a,b)^2$, Круглова — $\sum_x w(d\Phi(x,t,a,b))$, где $w(x)$ — весовая функция. Настраиваемое расстояние $d_{\vec{w}}$ может быть сформулировано в теоретико-вероятностном формате с использованием функций плотности $p_1 = \hat{\phi}\mathbf{\Gamma}_t(a)$ и $p_2 = \hat{\phi}\mathbf{\Gamma}_t(b)$: расстояния Золотарёва $\sum_x w(x)(p_1(x) - p_2(x))$, «ясности» $1 - \sum_x(p_1 \ln p_2 + p_2 \ln p_1 - (p_1 + p_2)\ln w(x))$, Чизара $\sum_x p_2 w(p_1/p_2)$, Бурби–Рао $\sum_x(0{,}5w(p_1) + 0{,}5w(p_2) - w(0{,}5p_1 + 0{,}5p_2))$ и др. [4]. Заметим, что расстояния Золотарёва и «ясности» включают веса индивидуальных значений $x \in \mathbf{I}_t$, что является одним из случаев расстояний по следствию 4 (индивидуальные веса объектов); расстояния Чизара, Бурби–Рао и Круглова относятся к «промежуточному» случаю расстояний по следствию 4.

В целом все известные $d_{\vec{w}}$ (1) включают агрегацию по $\mathbf{I}_t$ (суммирование или максимум), (2) агрегируются $d\Phi(x,t,a,b)$ или функции плотности, (3) веса при агрегации могут определяться для значений $x \in \mathbf{I}_t$ или для функционалов от значений $d\Phi(x,t,a,b)$, $\hat{\phi}\mathbf{\Gamma}_t(a)$ и $\hat{\phi}\mathbf{\Gamma}_t(b)$. При использовании суммирования по $x \in \mathbf{I}_t$ как единственного способа агрегации «универсальные» функция расстояния между ЭФР строятся как линейные комбинации над алфавитами наподобие $\{w_1(x)d\Phi(x,t,a,b)$, $w_2(d\Phi(x,t,a,b))$, $w_3(x)\hat{\phi}(x)\mathbf{\Gamma}_t(a)$, $w_4(x)\hat{\phi}(x)\mathbf{\Gamma}_t(b)$, $w_5(\hat{\phi}(x)\mathbf{\Gamma}_t(a))$, $w_6(\hat{\phi}(x)\mathbf{\Gamma}_t(b))$, $w_7(\hat{\phi}(x)\mathbf{\Gamma}_t(a) + \hat{\phi}(x)\mathbf{\Gamma}_t(b))\}$.

5 Расстояния по следствию 3 на векторах

Каждому $a \in L(T(\mathrm{X}))$ может быть сопоставлен вектор и введено расстояние на векторном пространстве. Очевидна многочисленность способов построения векторов и выбора способа вычисления компонент векторов. В [3, 7] предложен способ определения компонент векторов элементов как списков оценок над решеткой. В целом методы порождения векторных описаний нуждаются в более систематичном рассмотрении на основании решеточного формализма топологической теории распознавания.

6 Расстояния на основе индивидуальных весов атомов (следствие 4)

В соответствии со следствием 4 вклад каждого атома (одноатомного отрезка пути) в оценку расстояния между множествами a и b индивидуален, а порядок удаления/добавления атомов в множествах $a\backslash b$ и $b\backslash a$ не имеет значения. Функционалы, вычисляющие расстояние, представимы в виде коммутативных композиций функций от атомов, формирующих $a\Delta b$.

Целесообразно рассмотреть данное следствие в контексте двух экстремальных случаев: длины всех одноатомных отрезков равны (следствие 3) и длины всех одноатомных отрезков различны. В промежуточном случае длины одноатомных отрезков одинаковы для определенных групп атомов в наборе $S = \{s_1, s_2, \ldots, s_{|S|}\}$, $S \subset L(T(\mathbf{X}))$, который *становится* фактор-группой над $\mathbf{X}$.

При определении *расстояний по следствию 4 на множествах* целесообразно заменить мощности множеств в формулировках этих расстояний (Фреше–Никодима, Сокала–Сниса, Тверского и др.) на суммы весов атомов, составляющих эти множества. Таким образом, любая непараметрическая функция расстояния на множествах становится параметрической $\rho_{L\vec{w}}$, в которой вектор весов $\vec{w}$ определяется весами объектов.

В *расстояниях на функциях* $\hat{\Phi}(x)\mathbf{\Gamma}_t(a)$ и $\hat{\Phi}(x)\mathbf{\Gamma}_t(b)$, порождаемых t-й опорной цепью, ось абсцисс этих функций соответствует значениям $\mathbf{\Gamma}_t(x_i)$, так что вдоль этой оси объекты x_i могут быть упорядочены по значениям $\mathbf{\Gamma}_t(x_i)$. Например, расстояние Золотарёва, при условии регулярности $\mathbf{X}$ по $\mathbf{\Gamma}_t$, уже оказывается параметризованной метрикой по следствию 4, в которой вектор весов $\vec{w}$ представлен в виде весовой функции $w(x)$.

Определения *расстояния на векторах* $\vec{v}_\alpha[a]$ и $\vec{v}_\alpha[b]$ содержат оценки $v[a]$, включающие мощности множеств. Поэтому, заменяя мощности множеств при вычислении оценок на суммы весов объектов, можно получить параметризованные векторные метрики по следствию 4.

7 Расстояния с учетом порядка атомов (следствие 5)

В соответствии со следствием 5 определен порядок удаления/добавления атомов в множествах $a\backslash b$ и $b\backslash a$, и метрики по следствию 5 ищутся в виде комбинаций функций от последовательностей атомов из множества $a\Delta b$. Принимая во внимание чрезвычайную комбинаторную сложность булевых решеток (число полных цепей равно $N!$), целесообразно рассматривать максимально короткие подпоследовательности атомов из $a\Delta b$. Трактабельным вариантом оказываются функционалы, вводимые по аналогии с *концепцией марковских цепей* и соответствующие последовательностям из двух атомов (т. е. каждый атом соответствует «состоянию конечного автомата»).

8 О критериях выбора значений параметров настраиваемых метрик

Основная проблема при настройке метрик $\rho_{L\vec{w}}$ по следствиям 3–5 — выбор адекватного изучаемой задаче критерия качества, в соответствии с которым настраиваются весовые параметры такой метрики. В силу невозможности применения парадигмы согласования значений метрик [2,3,7] в $L(T(\mathbf{X}))$ остаются только две возможности дальнейших исследований:

(1) оставаясь в парадигме согласования метрик, найти пространство, в котором есть экспертные метрики (переход к расстояниям на объектах);

(2) оставаясь в решетке $L(T(\mathbf{X}))$, найти функционалы, по которым настраивать веса $\vec{w}$ («информативность» метрик, особые синтетические признаки).

9 О переходе от решетки $L(T(\mathbf{X}))$ к расстояниям на объектах

Формальное описание объекта $x_i \in \mathbf{X}$, $D(x_i)$, соответствует представлению любого объекта набором множеств $\{\Gamma_k^{-1}(\Gamma_k(x))\}$. Информация для задания некоторых «экспертных» расстояний между такими множествами, как правило, отсутствует. Применение ρ_L к парам множеств $(\Gamma_k^{-1}(\Gamma_k(x_i)), \Gamma_k^{-1}(\Gamma_k(x_j)))$ подразумевает оценку расстояний $\rho_q : \mathrm{Q}^2 \to \mathrm{R}^+$ между объектами x_i и x_j, $\rho_q(D(x_i), D(x_j))$, что дает возможность использовать четко формализуемые способы введения экспертных метрик: метрику разреза; оценки компактности классов; метрики алгоритмов вычислений оценок Ю. И. Журавлёва; экспертные и прочие оценки «схожести» объектов и др. Задачи числового прогнозирования сводимы к задачам классификации [6].

Громоздкость конструкций $\rho_q(D(x_i), D(x_j))$ диктует необходимость перехода от решетки

$L(T(\mathbf{X}))$ над булеаном множества объектов к решетке значений признаков. Данный переход представляет собой отдельное направление дальнейших теоретических и экспериментальных исследований, включающих среди прочего поиск функционалов ρ_q, наиболее релевантных для исследуемой задачи. Описанный ниже подход с синтетическими k'-ми признаками также может рассматриваться как подход к вычислению ρ_q.

10 Порождение признаков с использованием классов значений

Критерий настройки весов метрики $\rho_{L\vec{w}}$ может быть основан на специальных формах синтетических признаков, позволяющих осуществлять настройку $\rho_{L\vec{w}}$ по значениям t-й таргетной переменной. Пусть опорная цепь A_t разбита на n_z процентилей, заданных $\boldsymbol{\iota} = \{\lambda_0, \lambda_1, \lambda_2, \ldots$ $\ldots, \lambda_z, \ldots, \lambda_{n_z}\} \subset \mathbf{I}_t$, так что для z-го процентиля заданы множества $\Gamma_{tz}^{-1} = \{x|\lambda_{z-1} < (\mathbf{\Gamma}_t(x) < \lambda_z)\}$ и $\mathbf{\Gamma}_t(\Gamma_{tz}^{-1})$.

Определим множество

$$G_1(x_i, t, z) = \left\{\rho_{L\vec{w}}(\Gamma_k^{-1}(\Gamma_k(x_i)), \Gamma_{tz}^{-1}),\ \ k = \overline{1,n}\right\},$$

ЭФР $\hat{\Phi}\mathbf{\Gamma}_t(\Gamma_{tz}^{-1})$ и $\hat{\Phi}G_1(x_i, t, z)$, их математические ожидания $\hat{\mu}\hat{\Phi}\mathbf{\Gamma}_t(\Gamma_{tz}^{-1})$, $\hat{\mu}\hat{\Phi}G_1(x_i, t, z)$ и стандартные отклонения. Эти определения позволяют порождать синтетические признаки $\Gamma_{k'}(x_i)$, $k' > n + l$, на основе линейных и других функционалов, оценивающих $\mathbf{\Gamma}_t(x_i)$. При упорядочении $\hat{\Phi}G_1(x_i, t, z)$ по значениям $\hat{\mu}\hat{\Phi}G_1(x_i, t, z)$ выражение $1-\hat{\mu}\hat{\Phi}G_1(x_i, t, z)$ оценивает вклад расстояний значений признаков $\Gamma_k(x_i)$ до z-го процентиля в оценку значения t-й переменной, так что $\Gamma_{k'}(x_i)$ оценивается суммированием по z:

$$\Gamma_{k'}(x_i) = \sum_{z=1,n_z} \left(1 - \hat{\mu}\hat{\Phi}G_1(x_i, t, z)\right) \hat{\mu}\hat{\Phi}\mathbf{\Gamma}_t(\Gamma_{tz}^{-1}),$$
$$z = \overline{1, n_z}. \quad (1)$$

Значения $\hat{\mu}\hat{\Phi}G_1(x_i, t, z)$ могут нормироваться на максимальное значение:

$$\Gamma_{k'}(x_i) = \sum_{z=1,n_z} \frac{1 - \hat{\mu}\hat{\Phi}G_1(x_i, t, z)}{\max_{\mu z}} \hat{\mu}\hat{\Phi}\mathbf{\Gamma}_t\left(\Gamma_{tz}^{-1}\right),$$
$$\max_{\mu z} = \max(\hat{\mu}\hat{\Phi}G_1(x_i, t, z)). \quad (2)$$

Очевидно, что функционалы (1) позволяют осуществлять прямую настройку $\rho_{L\vec{w}}$ по значениям прогнозируемой t-й переменной $\mathbf{\Gamma}_t(x_i)$, $x_i \in \mathbf{X}$, в том числе с использованием аналитических процедур МНК-типа (см. ниже).

11 Порождение признаков с использованием целых опорных цепей

Эксперименты, проведенные в работе [7], показали, что «метрические» признаки на основе отдельных значений признаков $\Gamma_k(x_i)$ малоинформативны. Признаки на основе ЭФР $\hat{\phi}\boldsymbol{\rho}_{\boldsymbol{m}\overline{\boldsymbol{\alpha}}}(x_i, p)$ [7] показали несколько лучшую информативность, но ограничены заданием «опорных» классов объектов, т. е. множества $\boldsymbol{\iota}$. Возможно ли определение синтетических признаков на основе *всей* опорной цепи A_t, а не z-процентилей? При такой постановке вопроса $\rho_{L\vec{w}}$ применяется к каждому из элементов опорной цепи: (1) при отборе (ранжировании) наиболее информативных $\Gamma_k(x_i)$ либо (2) при повышении информативности каждого $\Gamma_k(x_i)$ путем особой параметризации.

12 Отбор наиболее информативных элементарных признаков

Данный подход подразумевает необходимость отбора признаков (значений признаков), самых информативных для поставленной задачи. Для задачи численного прогнозирования t-й переменной наиболее перспективно оценивать информативность посредством «расщепления» цепи A_t на две подцепи, соответствующие $\Gamma_k(x_i)$ и всем остальным значениям Γ_k [3, 5, 7], так что каждой из двух подцепей соответствует своя ЭФР: $\mathrm{cdf}_{tk}^{+}(x_i)$ и $\mathrm{cdf}_{tk}^{-}(x_i)$. Тогда расстояние d_f между $\mathrm{cdf}_{tk}^{+}(x_i)$ и $\mathrm{cdf}_{tk}^{-}(x_i)$ (по Колмогорову, Крамеру – фон Мизесу и др.) оценивает вклад значения признака $\Gamma_k(x_i)$ в значение t-й переменной, а разность $\hat{\mu}\,\mathrm{cdf}_{tk}^{+}(x_i) - \hat{\mu}\,\mathrm{cdf}_{tk}^{-}(x_i)$ — вклад признака в значение t-й переменной. Соответственно, $\Gamma_{k'}(x_i)$ определяется суммированием:

$$\Gamma_{k'}(x_i) = \sum_{k=1,n} d_f\left(\mathrm{cdf}_{tk}^{+}(x_i), \mathrm{cdf}_{tk}^{-}(x_i)\right) \times$$
$$\times \left(\hat{\mu}\,\mathrm{cdf}_{tk}^{+}(x_i) - \hat{\mu}\,\mathrm{cdf}_{tk}^{-}(x_i)\right). \quad (3)$$

13 Максимизация информативности элементарных признаков

Рассмотрение всей цепи A_t позволяет порождать описания каждого значения k-го признака

в виде одномерной функции. Для $\rho_{L\vec{w}}$ для каждой Γ_k определим множество пар

$$P_k(\lambda_{k\beta}) = \Big\{(\lambda_{tb}, \rho_{L\vec{w}}(\Gamma_k^{-1}(\lambda_{k\beta}), \mathbf{\Gamma}_t^{-1}(\lambda_{tb}))), \\ b = \overline{1, |\mathbf{I}_t|}\Big\}, \ \lambda_{k\beta} \in I_k\,.$$

Рассмотрим $P_k(\lambda_{k\beta})$ как подстановку некоторой действительно-значной и непрерывной функции $f(P_k(\lambda_{k\beta}), x)$, аппроксимирующей расстояния до элементов цепи. Аргумент x функции $f(P_k(\lambda_{k\beta}), x)$ — значения t-й переменной, погруженные в отрезок числовой прямой $[\lambda_1 \ldots \lambda_{|\mathbf{I}_t|}]$. На практике $f(P_k(\lambda_{k\beta}), x)$ удобно определить на основе $P_k(\lambda_{k\beta})$ с использованием разложения над базисом Фурье или над иным базисом в гильбертовом пространстве действительных функций. Описание объекта $q = D(x_i)$ набором («пучком») однородных функций $\mathbf{P}(x_i) = \{f(P_k(\Gamma_k(x_i)), x), k = 1, n\}$ видится максимально полным представлением прецедентных данных, отражающим все расстояния от всех значений признаков $\Gamma_k(x_i)$. Тогда рассматривается задача

$$(w_s) = \\ = \arg\min_{w_s} \sum_{\mathbf{X}} \sum_{k=1,n} \left(\arg\min_x f\left(P_k(\Gamma_k(x_i)), x\right) - \right. \\ \left. - \mathbf{\Gamma}_t(x_i)\right)^2. \quad (4)$$

Одно из возможных решений задачи (4) — введение индивидуальных наборов весов для каждого из значений k (и z, если используется разбиение на n_z процентилей), т. е. $w_s[k, z]$-настраиваемые метрики (wskz-метрики).

14 Об оценках информативности

В [7] предложены оценки информативности k'-й переменной относительно t-й, сводящиеся к измерению расстояний между ЭФР $F_{k't}(x, y)$ и $F_{k'}(x)F_t(y)$ посредством функций d_f. При рассмотрении k'-й и t-й переменных как двух вероятностных схем (в терминологии А. Н. Колмогорова и А. Я. Хинчина) становится возможным рассчитать буквальную «информативность» как разность между количеством информации (средней энтропией по формуле Шеннона) для совместного распределения вероятности $H_{k't}$ и для каждой из переменных по отдельности ($H_{k'}$ и H_t). При независимости k'-й и t-й переменных $H_{k't} = H_{k'} + H_t$, а при взаимозависимости — $H_{k't} < H_t$, так что функционал $H_t - H_{k't}$ оценивает информативность k'-й относительно t-й переменной.

15 МНК-подобные процедуры для настройки $\Gamma_{k'}$

Пусть для данной $\rho_{L\vec{w}}$ $\Gamma_{k'}$ можно представить в виде $\vec{Y}_{k'} = \mathbf{Z}\vec{w}$, где $\vec{Y}_{k'}$ — вектор значений $\Gamma_{k'}$ для всех $x_i \in \mathbf{X}$, $|\vec{Y}_{k'}| = N$, а матрица $\mathbf{Z}$ (размерность $N|\vec{w}|$) отражает способ вычисления $\Gamma_{k'}$ на основе прецедентной информации. Тогда глобальный оптимум $\vec{w}$ для $\Gamma_{k'}(x_i)$ можно вычислить с использованием МНК. Пусть $\vec{Y}$ — столбец матрицы информации, так что $\vec{Y}_{k'} = \vec{Y}$ — условие корректности $\Gamma_{k'}$ и МНК-решение уравнения $\vec{Y} = \mathbf{Z}\vec{w}$ равно

$$\vec{w} = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\vec{Y}.$$

Для $\Gamma_{k'}(x_i)$ по (1) и $\rho_{L\vec{w}}$ в виде расстояния Золотарёва или «ясности» элементы $\mathbf{Z}$ представимы как суммы функционалов от $\hat{\Phi}(x)\mathbf{\Gamma}_t(\,)$или $\hat{\phi}(x)\mathbf{\Gamma}_t(a)$ по k (признакам) и z (процентилям). В случае расстояний Круглова, Чизара, Бурби–Рао выражения для $\rho_{L\vec{w}}$ содержат «вложенные» термы $(w(\hat{\Phi}(x)\mathbf{\Gamma}_t(a) - \hat{\Phi}(x)\mathbf{\Gamma}_t(b))$ в расстоянии Круглова), так что применение МНК требует введения конкретного вида функции $w(\,)$ (кусочно-линейной, полиномиальной, Фурье-аппроксимации и др.).

16 Экспериментальное тестирование предлагаемых методов

Эксперименты были проведены на модельных данных для числового прогнозирования. Признаковые описания объектов представляли собой вектор $\vec{x} = (x_i)$ булевых признаков ($n = 50$), а числовое значение таргетной переменной вычислялось булевым полиномом 3-й степени

$$B(\vec{x}) = \sum_{i=0,n} \sum_{j=0,n} \sum_{l=0,n} w_{ijl} x_i x_j x_l, \quad x_0 = 1\,,$$

со случайными w_{ijl}, корректирующим преобразованием $\mathrm{corr}\,(B(\vec{x}))$ и со стохастическим компонентом $0{,}2\mathrm{random}\,(\,) \cdot |\mathrm{corr}\,(B(\vec{x}))|$. Были сгенерированы 500 модельных выборок. Предварительные эксперименты показали, что использование этой разновидности модельных данных предоставляет намного более строгое тестирование предлагаемых алгоритмических конструкций, чем используемые в предыдущих работах выборки данных по фармако-, хемо- и биоинформатике [2–4, 7].

Синтетические признаки $\Gamma_{k'}(x_i)$ порождались всеми перечисленными в работе [7] способами ($\rho_m(\mathbf{c}_\alpha, \Gamma_k^{-1}(\Gamma_k(x_i))$, ЭФР-признаки $\hat{\phi}(x)\boldsymbol{\rho_{m\alpha}}(x_i, p)$ при разных x и др.), а отбор проводился B-алгоритмом [7]. При $n = 50$ общее число синтетических признаков было равно 7562. В качестве *расстояний ρ_m по следствию* 3 теоремы о регулярности и нормальности были использованы метрики на основе одно- и двухэлементных алфавитов $A(a, b)$ и $B(a, b)$ совместно с 65 известными функциями расстояния на множествах, векторах и ЭФР [4]. В качестве ρ_m *по следствию* 4 — расстояния Золотарёва и «ясности», Круглова, Чизара, Бурби–Рао и wskz-метрики с настройкой $\Gamma_{k'}(x_i)$ посредством МНК.

Для оценки «**информативности**» получаемых синтетических признаков $\Gamma_{k'}(x_i)$ использовались:

(1) функционалы на основании расстояний между $\Gamma_{k'}^{-1}(\Gamma_{k'}(x_i))$ и $\mathbf{\Gamma}_t^{-1}(\mathbf{\Gamma}_t(x_i))$: коэффициент корреляции, нормализованное стандартное отклонение и др.;

(2) различные варианты функций d_f для ЭФР $F_{k't}(x, y)$ и $F_{k'}(x)F_t(y)$;

(3) собственно информативность $H_t - H_{k't}$.

В качестве **алгоритмов прогнозирования** числовых таргетных переменных использовались нейронные сети с несколькими слоями (от 2 до 10) с функцией активации softmax и полиномы различных конструкций (более 20 формул, в том числе квазиполиномные модели с элементарными функциями) в контексте 10-кратной кросс-валидации 50%/50%.

Экспериментальные исследования были проведены в два этапа: уточнение способов оценки «информативности» и отбор наиболее информативных признаков для прогнозирования таргетных переменных. Вычисления по **различным способам оценки «информативности»** порождаемых синтетических признаков проводились на ЭФР-признаках $\hat{\phi}(x)\boldsymbol{\rho_{m\alpha}}(x_i, p)$, вычисляемых на основе непараметрических ρ_m. Эксперименты показали, что большинство способов оценки информативности входили в два расположенных рядом кластера: малый кластер (в центре — метрика Колмогорова и информативность $H_t - H_{k't}$) и большой кластер (12 точек, в центре d_f, соответствующая расстоянию Реньи). Поскольку кластеры 1 и 2 расположены рядом и кластер 1 содержит информативность в виде $H_t - H_{k't}$, то все метрики (кроме расстояний Андерсона, инженерной метрики и коэффициента корреляции $r(\Gamma_{k'}(x_i), \mathbf{\Gamma}_t(x_i))$) могут в действительности рассматриваться как оценки информативности, без кавычек. При этом расстояние Андерсона и инженерная метрика лежат на осях соответствующей метрической диаграммы, а точка, соответствующая коэффициенту корреляции, находится в правом верхнем углу диаграммы.

На втором этапе проводились **эксперименты по прогнозированию модельных таргетных переменных**, включавшие настройку метрик по следствию 4 с использованием синтетических признаков по формулам (1), (2) и (3) и отбор наиболее информативных признаков посредством B-алгоритма на основе метрики Колмогорова (максимальное уклонение) [7]. Наиболее часто в наборы информативных $\Gamma_{k'}$-признаков входили ЭФР-признаки, сконструированные на основании предложенных в предыдущих работах оценочных метрик (43% от всех информативных), причем самыми информативными ρ_m были *непараметрические* метрики по следствию 3 на множествах (оценочные, расстояния Фейта, Симпсона, Фреше–Никодима).

В то же время если брать самые информативные синтетические $\Gamma_{k'}$, то среди них наиболее часто (38% выборок) встречался признак на основании расстояния Золотарёва (по следствию 4), вычисляемого по формуле (1) и настраиваемого процедурой МНК. Интересно отметить, что среди расстояний по следствию 4, настраиваемых по (1), расстояние Золотарёва доминировало: $\Gamma_{k'}$ на основе настраиваемых расстояний Круглова, Чизара, Бурби–Рао не отличались существенно более высоким качеством.

В целом по исследованной выборке, использование синтетических признаков по (1) с метриками по следствию 4, настраиваемыми посредством МНК-процедур, позволило снизить показатели переобучения и улучшить качество работы алгоритмов числового прогнозирования. Так, с использованием только ЭФР-признаков значение коэффициента ранговой корреляции составило $r = 0{,}88 \pm 0{,}09$ на обучении и $r_c = 0{,}56 \pm 0{,}23$ на контроле. В то же время добавление признаков по (1)–(3) способствовало повышению $r_c = 0{,}58 \pm 0{,}18$ и снижению $r = 0{,}75 \pm 0{,}15$, что соответствует улучшению обобщающей способности получаемых алгоритмов.

17 Заключение

В результате анализа условий необходимости и достаточности метризации дискретных пространств $(\mathbf{X}, T(\mathbf{X}))$ была сформулирована и доказано теорема, устанавливающая соответствие между понятиями регулярности множества прецедентов $\varphi(\mathbf{X})$ и нормальности решетки $L(T(\mathbf{X}))$. Следствия этой теоремы позволили сформулировать ряд не исследованных ранее подходов к метризации, включая метрики над алфавитами мощностей

по следствию 3, настраиваемые метрики по следствию 4 и конструкции, связанные с учетом порядка атомов (следствие 5 теоремы). Сформулированы перспективные направления дальнейших исследований:

(1) переход от решетки $L(T(\mathbf{X}))$ к решетке значений признаков (что позволяет систематически исследовать способы введения метрик между объектами);
(2) перспективные функционалы для порождения синтетических признаков;
(3) методы порождения векторных описаний для определения расстояний между элементами $L(T(\mathbf{X}))$;
(4) представление значений признаков произвольных типов (в том числе булевых) в виде непрерывных функций.

Вычислительные эксперименты показали перспективность предлагаемых способов метризации топологического пространства $(\mathbf{X}, T(\mathbf{X}))$ и способов порождения синтетических признаков, способствующих повышению обобщающей способности алгоритмов числового прогнозирования.

Литература

1. *Торшин И. Ю.* Метризация дискретных топологических пространств в контексте теории решеток. Часть 1. О нормальности пространств // Информатика и её применения, 2025. Т. 19. Вып. 1. С. 82–88. doi: 10.14357/19922264250111. EDN: CAWKMO.
2. *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.
3. *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.
4. *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).
5. *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.
6. *Torshin I. Y., Rudakov K. V.* On the procedures of generation of numerical features over partitions of sets of objects in the predicting numerical target variables // Pattern Recognition Image Analysis, 2019. Vol. 29. No. 3. P. 654–667. doi: 10.1134/S1054661819040175.
7. *Торшин И. Ю.* 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 // Информатика и её применения, 2024. Т. 18. Вып. 2. С. 47–53. doi: 10.14357/19922264240207. EDN: OTXCUD.

Поступила в редакцию ??.10.2024
Принята к публикации ??.01.2025

METRIZATION OF DISCRETE TOPOLOGICAL SPACES IN THE CONTEXT OF LATTICE THEORY. PART 2. PRACTICAL ANALYSIS OF THE CONSEQUENCES OF THE THEOREM ON REGULARITY AND NORMALITY

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: In the first part, a theorem has been proved that connects the concept of normality of topological spaces and regularity according to Yu. I. Zhuravlev, in the context of the problem of metrization of feature spaces. The consequences of the theorem allow one to systematize the search for the most acceptable problem-oriented metrics. Promising directions for further research have been systematized, including the transition to a lattice of feature values and promising functionals for generating synthetic features. The results of the corresponding computational experiments are presented.

Keywords: topological data analysis; algebraic approach of Yu. I. Zhuravlev and K. V. Rudakov; computational experiment; synthetic features

DOI: 10.14357/1 **EDN:** T

References

1. Torshin, I. Yu. 2025. Metrizatsiya diskretnykh topologicheskikh prostranstv v kontekste teorii reshetok. Chast' 1. O normal'nosti prostranstv [Metrization of discrete topological spaces in the context of lattice theory. Part 1. On the normality of spaces]. *Informatika i ee primeneniya — Inform Appl.* 19(1):82–88. doi: 10.14357/19922264250111. EDN: CAWKMO.
2. 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.
3. 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.
4. 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.
5. 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. doi: 10.1134/S1054661815040252.
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. Torshin, I. Yu. 2024. O porozhdenii sinteticheskikh priznakov na osnove opornykh tsepey i proizvol'nykh metrik v ramkakh topologicheskogo podkhoda k analizu dannykh. Chast' 2. Eksperimental'naya aprobatsiya na zadachakh farmakoinformatiki [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]. *Informatika i ee Primeneniya — Inform. Appl.* 18(2):47–53. doi: 10.14357/19922264240207. EDN: OTXCUD.

Received ??tober 10, 2024
Accepted ??nuary 15, 2025

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