ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2020, том 492, с. 101–103

ИНФОРМАТИКА

УДК 517

О НЕКОТОРЫХ ФАКТОРИЗАЦИЯХ ПОЛУМЕТРИЧЕСКИХ КОНУСОВ И ОЦЕНКАХ КАЧЕСТВА ЭВРИСТИЧЕСКИХ МЕТРИК В ЗАДАЧАХ АНАЛИЗА ДАННЫХ

© 2020 г. Академик РАН К. В. Рудаков[1,2,*]

Поступило 09.04.2020 г.
После доработки 09.04.2020 г.
Принято к публикации 09.04.2020 г.

Предлагается подход к рассмотрению эвристических метрик, вводимых и применяемых в задачах анализа данных, при котором вся выражаемая числовыми значениями информация о попарных расстояниях сводится к информации о принадлежности метрики как точки полуметрического конуса к соответствующим подконусам – элементам фактор-множеств по предлагаемым отношениям ядерных эквивалентностей для отображений в формальные индексные семейства.

Ключевые слова: интеллектуальный анализ данных, искусственный интеллект, большие данные, полуметрический конус, эвристические метрики, оценки качества метрик

DOI: 10.31857/S2686954320030236

Значительная часть методов интеллектуального анализа данных основана на применении эвристических (не обосновываемых теоретически) метрик, вводимых на множествах анализируемых объектов или ситуаций [1, 2]. Эвристичность метрик как инструментов анализа позволяет ставить и решать вопросы оптимизации их "потребительских качеств", для чего можно применять соответствующие функционалы качества, в том числе и предлагаемые в данном сообщении. Существенно, что обычно числовые величины расстояний оказываются малоинтересными: реально используются только некоторые соотношения расстояний между различными парами объектов. Введенные на пространстве объектов различные метрики, порождающие совпадающие соотношения расстояний, оказываются эквивалентными с точки зрения их использования при решении модельных и прикладных задач. Метрики на конечных множествах считаются точками соответствующего полуметрического конуса, и рассматриваются классы ядерных эквивалентностей для отображений полуметрического конуса в множества некоторых вводимых ниже отношений порядка [3–5]. Отметим, что эти классы оказываются подконусами, определяемыми соответствующими системами равенств и однородных линейных неравенств.

Пусть $\mathbf{S} = \{S_1, \ldots, S_q\}$ и задана полуметрика d: $\mathbf{S}^2 \to \mathbf{R}_+$ (допускается, что $d(S_i, S_j) = 0$ при различных S_i и S_j). Будем также считать, что d – точка полуметрического конуса $\mathbf{Con}^q$ в $\mathbf{R}^p$ при $p = C_q^2$. Множество упорядоченных пар индексов $\{(i, j) | 1 \le i < j \le q)\}$ будем обозначать символом Q и называть пары (i, j) ребрами. Метрика d естественным образом индуцирует линейный порядок $\le_2$ на Q: $((i, j) \le_2 (k, l)) \equiv (d(S_i, S_j) \le d(S_k, S_l))$. Отметим, что, скажем, в алгоритмах вычисления оценок и методах типа "ближайших соседей" фактически используется только порядок $\le_2$. Очевидно, что этот порядок инвариантен по отношению к любым монотонно возрастающим преобразованиям значений расстояний, так что его использование "автоматически" снимает проблемы "единиц измерения" величин расстояния.

Пусть, наоборот, на Q задан некоторый произвольный линейный порядок $\le_0$. Ему легко сопоставить метрику d_0 на $\mathbf{S}$ такую, что соответствующий порядок $\le_2$ будет совпадать с порядком $\le_0$. Действительно, для построения такой метрики d_0 достаточно положить $d_0(S_i, S_j) = 1 + \varepsilon_{ij}$, где $0 < \varepsilon_{ij} < 1$ и ε_{ij} расположены на интервале (0, 1) в нужном порядке. Указанная метрика, содержательно близ-

[1] *Вычислительный центр им. А.А. Дородницына Федерального исследовательского центра "Информатика и управление" Российской академии наук, Москва, Россия*
[2] *Центр хранения и анализа больших данных, Московский государственный университет имени М.В. Ломоносова, Москва, Россия*
*****E-mail: rudakov@frccsc.ru*

кая к “абсолютно неинформативной” метрике пространства изолированных точек, может рассматриваться как в некотором смысле контрпример.

Итак, во многих случаях используется только факт принадлежности метрики к классу эквивалентности по отношению “$\sim$”, где $d_1 \sim d_2$ означает, что для всех i, j, k и l из $\{1, ..., q\}$ выполнено $(d_1(S_i, S_j) \le d_1(S_k, S_l)) \equiv (d_2(S_i, S_j) \le d_2(S_k, S_l))$, или, что то же, используется только факт принадлежности метрики к классу ядерной эквивалентности для отображения π_2 полуметрического конуса $\mathbf{Con}^q$ в подмножество булеана Q^2 – множество линейных порядков на Q: $\pi_2(d) = \{((i,j),(k,l))| d(S_i, S_j) \le d(S_k, S_l)\}$. Отметим, что при этом в каждом классе эквивалентности указанного вида имеются сколь угодно близкие к центральной оси $\mathbf{Con}^q$ метрики.

Отметим, что в метриках типа “$d(S_i, S_j) = 1 + \varepsilon_{ij}$” выполнены не только все необходимые по определению метрики неравенства треугольника, но и вообще все неравенства вида

$$d(S_i, S_j) < d(S_k, S_l) + d(S_l, S_m) \quad (1)$$

и даже вида

$$d(S_i, S_j) < d(S_k, S_l) + d(S_m, S_n) \quad (2)$$

при произвольных различных i, j, k, l, m и n из $\{1, ..., q\}$. Такие метрики можно назвать метриками пространств квазиизолированных точек. Как точки полуметрического конуса, они расположены “около его центральной оси”. Если принять тезис о том, что близость к метрике пространства изолированных точек свидетельствует о “малой информативности” метрики, то в качестве оценки “качества” метрики можно использовать, например, мощности множеств наборов индексов i, j, k, l, m и n, для которых не выполнены неравенства (1) или (2). Очевидно, что для метрик пространств квазиизолированных точек указанные множества пусты.

Пусть $\mathbf{T}_3 = \{(k, l, m)| 1 \le k < m \le q, 1 \le l \le q, l \ne k, l \ne m\}$ и $\mathbf{U}_3 = \boldsymbol{T}_3 \cup \mathbf{Q}$, а $\mathbf{T}_4 = \{(k, l, m, n)| 1 \le k < l \le q, 1 \le m < n \le q\}$ и $\mathbf{U}_4 = \boldsymbol{T}_4 \cup \boldsymbol{Q}$. Тройки (k, l, m) из $\mathbf{T}_3$ будем называть углами. Для всех элементов α и β из $\mathbf{U}_3$ или из $\mathbf{U}_4$ естественным образом определяется вес $W(\alpha) = d(S_i, S_j)$ при $\alpha = (i, j) \in \mathbf{Q}$, $W(\alpha) = d(S_k, S_l) + d(S_l, S_m)$ при $\alpha = (k, l, m) \in \mathbf{T}_3$ и $W(\alpha) = d(S_k, S_l) + d(S_m, S_n)$ при $\alpha = (k, l, m, n) \in \mathbf{T}_4$. Рассмотрим отображения π_3 и π_4 полуметрического конуса $\mathbf{Con}^q$ в множества линейных порядков на $\mathbf{U}_3$ и $\mathbf{U}_4$ соответственно:

$$\pi_3(d) = \{(\alpha,\beta) \mid W(\alpha) \le W(\beta), \alpha \in \mathbf{U}_3, \beta \in \mathbf{U}_3\},$$

$$\pi_4(d) = \{(\alpha,\beta) \mid W(\alpha) \le W(\beta), \alpha \in \mathbf{U}_4, \beta \in \mathbf{U}_4\}.$$

Порядки π_3 и π_4, вообще говоря, более точно описывают метрику d в том смысле, что по ним, естественно, однозначно восстанавливается порядок π_2, но не наоборот. Иначе говоря, содержащие метрику d классы ядерных эквивалентностей для отображений π_3 и π_4 оказываются подклассами ядерных эквивалентностей для π_2.

Кроме того, в отличие от случая π_2, теперь уже неверно, что в каждом классе эквивалентности имеются сколь угодно близкие к оси $\mathbf{Con}^q$ метрики.

Ясно, что классы эквивалентных метрик определяются соответствующими системами линейных неравенств и образуют подконусы конуса $\mathbf{Con}^q$. Для нормированных метрик d в качестве представителей классов эквивалентности можно использовать “ближайшие” к оси $\mathbf{Con}^q$ или, наоборот, “наиболее отдаленные” от оси метрики, которые будут определяться обращением некоторых неравенств треугольника в равенства. Если для оценки удаленности от оси использовать метрику l_1, то такие метрики можно находить как решения задач типа

$$d^* = \arg\min \Sigma w_{ij} d(S_i, S_j)$$
$$\text{или } d^* = \arg\max \Sigma w_{ij} d(S_i, S_j),$$

где

$$w_{ij} = |\{(k, l)| d(S_k, S_l) < d(S_i, S_j)\}| - |\{(k, l)| d(Sk, S_l) > d(S_i, S_j)\}|,$$

при нормировке $\Sigma d(S_i, S_j) = \text{const}$, ограничениях $0 < d(S_i, S_j)$ и ограничениях (линейных неравенствах), задаваемых порядком $\pi_3(d)$ или $\pi_4(d)$.

Описываемый подход позволяет ввести новые теоретико-множественные оценки “качества” конечных метрических конфигураций (в такой роли в задачах кластеризации или классификации обычно используются характеристики типа “отношение средних внутриклассовых расстояний к межклассовым”). Отметим, что указанные оценки можно использовать при решении задач коррекции и комплексирования метрик.

ИСТОЧНИКИ ФИНАНСИРОВАНИЯ

Работа выполнена в ВЦ РАН ФИЦ ИУ РАН в рамках общего плана НИР, в ходе реализации проекта РФФИ № 18–07–00741 и НИР ЦХАБД МГУ имени М.В. Ломоносова “Математические основы интеллектуального анализа больших данных”.

СПИСОК ЛИТЕРАТУРЫ

1. *Журавлев Ю.И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. 1978. Т. 33. С. 5–68.

2. *Воронцов К.В.* Математические методы обучения по прецедентам. Курс лекций МФТИ, 2006.
3. *Деза М., Лоран М.* Геометрия разрезов и метрик. М.: МЦНМО, 2001.
4. *Мальцев А.И.* Алгебраические системы. М.: Наука, 1970. 392 с.
5. *Арутюнов А.В.* Лекции по выпуклому и многозначному анализу. М.: Физматлит, 2014. 188 с.

ON SOME FACTORIZATIONS OF SEMI-METRIC CONES AND QUALITY ESTIMATES OF HEURISTIC METRICS IN DATA ANALYSIS PROBLEMS

Academician of the RAS V. K. Rudakov[a,b]

[a] *Federal Research Center Computer Science and Control of the Russian Academy of Sciences, Moscow, Russian Federation*
[b] *Center of Big Data Storage and Analysis Technologie, Lomonosov Moscow State University, Moscow, Russian Federation*

An approach is proposed to consider heuristic metrics introduced and used in data analysis problems. The approach reduces all information on pairwise distances expressed by numerical values to the information on a metric's belonging as a point of a semi-metric cone to the corresponding subcones which are elements of factor sets for the proposed relations of kernel equivalences for mappings into formal index families.

Keywords: data mining, artificial intelligence, big data, semi-metric cone, heuristic metrics, quality estimates of metrics