УДК 343.346.8:004.056.53

Е.В. Бодянский[1], В.М. Струков[2], Д.Ю. Узлов[3]

[1] *Харьковский национальный университет радиоэлектроники, Харьков*
[2] *Харьковский национальный университет внутренних дел, Харьков*
[3] *Главное управление национальной полиции в Харьковской области, Харьков*

## ОБОБЩЕННАЯ МЕТРИКА В ЗАДАЧЕ АНАЛИЗА МНОГОМЕРНЫХ ДАННЫХ С РАЗНОТИПНЫМИ ПРИЗНАКАМИ

*Работа посвящена задаче оценки близости многомерных объектов, признаки которых измеряются в разных шкалах, а обрабатываемые данные имеют большую размерность и содержат специфические текстовые поля и пропуски. К данным с такими специфическими особенностями не применимы непосредственно классические алгоритмы кластеризации и классификации. Предложена обобщенная метрика в многомерном пространстве таких объектов, которая позволяет строить алгоритмы кластеризации, классификации и ассоциации, основанные на ней, с использованием классических методов.*

***Ключевые слова:*** *многомерные объекты, кластеризация, классификация, шкала измерений, количественная метрика, категориальная метрика, ранговая метрика, текстовая метрика.*

### Введение

В последние годы актуальность задачи аналитической обработки больших объемов данных возрастает. Особенно это касается неструктурированных и многомерных данных, описываемых нечисловыми признаками.

На сегодня достаточно хорошо исследованы задачи обработки однотипных данных – числовых, текстовых. Классические алгоритмы аналитической обработки больших массивов данных (Data ClassifiCation, Data Mining, Text Mining, Knowledge Discovery) работают именно с такими типами данных [1–3; 5–7] и непосредственное применение их для обработки разнотипных данных невозможно.

Вместе с тем, на практике встречаются задачи, требующие обрабатывать именно большие массивы разнотипных данных. Иногда эти данные, к тому же, обладают некоторыми специфическими особенностями, которые даже в случае их однотипности не позволяют непосредственно применять классические методы и алгоритмы.

Одной из таких практически важных задач является задача аналитической обработки многомерных объектов, описываемых признаками различных типов, информация о которых накапливается в базах данных подразделений информационного обеспечения полиции Украины. В качестве обрабатываемых объектов в данном случае выступают лица, предметы и события криминального характера. Информация об этих объектах регистрируется в соответствующих криминальных учетах [8] путем заполнения некоторых регламентированных стандартных форм [9]. Специфическими особенностями этих массивов данных являются [9–11]:

1) большие объемы данных (сотни тысяч, а иногда и миллионы записей); причем количество этих данных с каждым днем увеличивается;

2) большое количество признаков, характеризующих объекты (до сотни и больше признаков);

3) различная природа признаков (как правило, нечисловая);

4) в текстовых полях вследствие особенностей динамического формирования реальных массивов часто встречаются повторяющиеся куски текстов;

возможность наличия пропусков (отсутствие значений там, где они должны находиться) в массивах в силу ряда субъективных и объективных причин.

В исследуемой задаче выполняется обработка данных следующих типов: 1) числовые; 2) текстовые; 3) категориальные; 4) ранговые (порядковые).

В работе [4] предложена метрика, учитывающая часть перечисленных выше особенностей. Данная работа является развитием и обобщением предложенного в [4] подхода.

**Целью данной статьи** является разработка обобщенной оценки близости объектов с перечисленными выше особенностями.

### Обобщенная метрика в пространстве многомерных объектов с разнотипными признаками

Введем следующие обозначения:

$X=\{x_{ij}\}$, – матрица “объект-свойство”, в которой $x_{ij}$ – значение j-го свойства (признака) i-го объекта, i=1, 2, …, m; j = 1, 2, …, n;

шкалы измерений: категориальная (номинальная, бинарная) – cat, ранговая(порядковая) – rank, числовая(интервальная, относительная) – num; текстовая – txt;

$x_i=(x_{i1}, \ldots, x_{ij}, \ldots, x_{in})^T \in R^n$;

none – отсутствующие данные в ячейке;

$n_i$ – число пропусков в объекте $x_i$;

$n_l$ – число пропусков в объекте $x_l$;

$n_{il}$ – число общих пропусков;

$x_i \cap x_l \neq \varnothing$.

Для учета пропусков в данных введем следующую вспомогательную переменную:

$$\delta_{il} = \begin{cases} 1, (x_{ij} \neq \text{none}) \wedge (x_{lj} \neq \text{none}); \\ 0, (x_{ij} = \text{none}) \vee (x_{lj} = \text{none}). \end{cases} \quad (1)$$

С учетом введенных обозначений и выражения (1) расстояние $d_{il}$ между объектами $x_i$ и $x_l$ в обобщенном виде можно записать следующим образом:

$$d_{ij} = \frac{1}{n - n_i - n_l + n_{il}} \sum_{j=1}^{n} Op_j(x_{ij}, x_{lj})\delta_{il}. \quad (2)$$

Здесь $Op_j(x_{ij},x_{lj})$ – оператор вычисления расстояния между объектами $x_i$ и $x_l$ для j-того признака.

Для учета значимости признаков введем их весовые коэффициенты $k_j$, которые определяются или задаются экспертами-аналитиками в ходе решения конкретных задач:

$$d_{il} = \sum_{j=1}^{n} k_j d_{il}^j \delta_{il}. \quad (3)$$

Пронормируем коэффициенты $k_j$:

$$k_j' = \frac{k_j}{\sum_{j=1}^{n} k_j}; \qquad \sum_{j=1}^{n} k_j' = 1.$$

Для учета шкалы измерения каждого признака введем вспомогательные переменные:

$$b_i^p = \begin{cases} 1, \text{если } i-\text{тый признак измеряется в} \\ \quad p-\text{той метрике}; \\ 0, \text{в противном случае}, \end{cases}$$

где p – p-тая метрика, p=1, 2, …, $N_s$, $N_s$ – количество щкал измерения, используемых в задаче.

Тогда итоговое обобщенное выражение для вычисления расстояния между объектами $x_i$ и $x_l$ будет иметь следующий вид:

$$d_{il} = \frac{1}{n - n_i - n_l + n_{il}} \sum_{p=1}^{Ns} \sum_{j=1}^{n} k_j' b_j^p \bullet \left(Op_j^p(x_{ij}, x_{lj})\right)\delta_{il}. \quad (4)$$

Рассмотрим применение полученного выражения к массивам данных, описанным выше, т.е. в случае использования числовых, категориальных, ранговых и текстовых значений.

Для числовых значений признаков будем использовать следующие выражения [4]:

– для евклидовой метрики:

$$d_{il}^{numE'} = \frac{1}{n - n_i - n_l + n_{il}} \sum_{j=1}^{n} k_j'(x_{ij} - x_{lj})^2 \delta_{il};$$

$$0 \le d_{il}^{numE} \le 1.$$

– для манхэттенской метрики:

$$d_{il}^{numBC'} = \frac{1}{n - n_i - n_l + n_{il}} \sum_{j=1}^{n} k_j' \left|x_{ij} - x_{lj}\right| \delta_{il};$$

$$0 \le d_{il}^{numBC} \le 1.$$

С вычислительной точки зрения более предпочтительной является манхэттенская метрика и для применения евклидовой метрики на практике требуется достаточно веское обоснование. Поэтому, в дальнейшем, говоря о числовой метрике, будем иметь в виду именно манхэттенскую метрику.

Для категориальных значений воспользуемся следующим выражением[4]:

$$d_{il}^{cat''} = \frac{1}{n - n_i - n_l + n_{il}} \sum_{j=1}^{n} k_j' \delta(x_{ij}, x_{lj})\delta_{il}; 0 \le d_{il}^{cat'} \le 1.$$

$$0 \le d_{il}^{cat'} \le 1.$$

В ранговой метрике расстояние $d_{il}^{rank}$ между $x_i$ и $x_l$ можно записать в виде[4]:

$$d_{il}^{rank'} = \frac{1}{n - n_i - n_l + n_{il}} \sum_{j=1}^{n} k_j' \left|x_{ij}^{r_j'} - x_{il}^{r_j'}\right| \delta_{il}:$$

$$0 \le d_{il}^{rank} \le 1,$$

где $x_{ij}^{r_j'}$, $x_{il}^{r_j'}$ – нормированные ранговые значения (приведенные к интервалу [0,1]) признаков $x_i$ и $x_l$, соответственно:

$$x_{ij}^{r_j'} = \frac{x_{ij}^{r_j} - x_{ij}^{1}}{x_{ij}^{R_j} - x_{ij}^{1}},$$

где $r_j$ – ранг j-того признака; $r_j$=1, 2, …, $R_j$.

Текстовые поля. Наличие в векторах $x_i$ признаков, представляющих собой текстовые поля различной длины, вносит в исследуемую задачу дополнительную сложность.

Сущность ее состоит в том, что, во-первых, обработка текстовых корпусов представляет собой отдельную достаточно сложную проблему, исследование которой к настоящему времени вылилось в самостоятельное направление – Text Mining [1–3].

Во-вторых, текстовые поля в исследуемой задаче в силу сложившихся процессуальных особен-

ностей их формирования обладают определенной спецификой – наличием многократно повторяемых избыточных кусков текста.

В этом случае применение для оценки близости таких текстовых объектов традиционно используемых метрик, основанных на частотах встречаемости отдельных термов или коллокаций, в частности Dtf-idf [3], будет искажать реальную ситуацию.

Кроме того, такая метрика должна учитывать описанные выше особенности обрабатываемых массивов данных. Одной из известных текстовых метрик, которая в наибольшей степени позволяет учесть все перечисленные особенности, является коэффициент Жаккарда, в основу которого положено использование множественных представлений текстовых объектов – тезаурусов (словарей) [3]:

$$d_{ij}^{t} = \frac{|D_i \cap D_j|}{|D_i \cup D_j|},$$

где $D_i$ и $D_j$ – множественные представления текстовых объектов $p_i$ и $p_j$, соответственно: $D_i = \{ t_1^i, t_2^i, \ldots, t_{k_i}^i \}$, $t_j^i$ - j-й терм множества $D_i$, $k_i$ – количество слов в словаре множества $D_i$.

С учетом вышеперечисленных особенностей исследуемых массивов данных и их текстовых полей, а также введенных обозначений искомая оценка близости значений текстовых признаков может быть сформулирована следующим образом:

$$d_{il}^{t} = \frac{1}{n - n_i - n_l + n_{il}} \sum_{j=1}^{n} k_j' \frac{|D_{ij} \cap D_{lj}|}{|D_{ij} \cup D_{lj}|} \delta_{il};$$

$$0 \le d_{il}^{t} \le 1.$$

Далее, используя введенные обозначения и полученные выражения, сформулируем конкретный вид выражения для определения расстояния между объектами $x_i$ и $x_l$.

С этой целью для корректного учета разнотипности признаков введем следующие служебные переменные:

$$b_i^{num} = \begin{cases} 1, \text{если } j-\text{тый признак измеряется в числовой метрике;} \\ 0, \text{в противном случае;} \end{cases}$$

$$b_i^{cat} = \begin{cases} 1, \text{если } j-\text{тый признак измеряется в категориальной метрике;} \\ 0, \text{в противном случае;} \end{cases}$$

$$b_i^{rank} = \begin{cases} 1, \text{если } j-\text{тый признак измеряется в ранговой метрике;} \\ 0, \text{в противном случае;} \end{cases}$$

$$b_i^{txt} = \begin{cases} 1, \text{если } j-\text{тый признак измеряется в текстовой метрике;} \\ 0, \text{в противном случае.} \end{cases}$$

Тогда, подставив эти переменные и выражение (5) в выражение (4), получим итоговую формулу для определения расстояния между объектами $x_i$ и $x_l$ в следующем виде:

$$d_{il} = \frac{1}{n - n_i - n_l + n_{il}} \Big( \sum_{j=1}^{n} b_j^{num} k_j' \left| x_{ij} - x_{lj} \right| \delta_{il} + \sum_{j=1}^{n} b_j^{cat} k_j' \delta(x_{ij}, x_{lj}) \delta_{il} + \sum_{j=1}^{n} b_j^{rank} k_j' \left| x_{ij}^{r_j'} - x_{il}^{r_j'} \right| \delta_{il} + \sum_{j=1}^{n} b_j^{txt} k_j' \frac{|D_{ij} \cap D_{lj}|}{|D_{ij} \cup D_{lj}|} \delta_{il} \Big).$$

## Выводы

В данной работе исследована задача оценки близости многомерных объектов с разнотипными признаками, содержащих пропуски в данных. Проанализированы и сформулированы особенности реальных массивов данных. Обоснована невозможность применения классических алгоритмов анализа данных (Data Mining, Text Mining) для обработки исследуемых массивов данных. Предложена обобщенная метрика для оценки близости объектов с заданными особенностями, позволяющая свести ее к скалярным числовым значениям. Это позволило свести задачу к классическому численному виду и обеспечило принципиальную возможность применить для ее решения известные методы и алгоритмы. Кроме того, использование коэффициентов приоритетности признаков дает возможность в интерактивном режиме оператору-аналитику регулировать количество анализируемых признаков в соответствии с решаемой конкретной задачей, что существенно в условиях большого их количества. Предложенный подход позволяет применять классические алгоритмы кластеризации, классификации и ассоциации для решения практических задач выявления неявных и скрытых связей между объектами криминальных учетов в базах данных информационных систем органов внутренних дел, а также в других предметных областях, где обрабатываются массивы данных с такими особенностями.

## Список литературы

*1. Han L., Kamber M. Data Mining: Concepts and Techniques. – Amsterdam: Morgan Kaufman Publ., 2006. – 754 p.*

*2. Aggarwal C.C. Data Mining. – Cham: Springer Ltd. Publ. Switzerland, 2015. – 734 p.*

*3. Aggarwal C.C., Reddy C.K. Data Clustering. Algorithms and Applications. – New York:CRC Press, Tailor & Francik Group, 2014. – 648 p.*

*4. Бодянский Е.В. Задача оценки близости многомерных объектов анализа данных / Е.В. Бодянский, В.М. Струков, Д.Ю. Узлов// УСиМ. – 20016. – № 6. – С. 67-72.*

*5. Hathaway R.J., Bezdek J.C. Fuzzy c-means clustering of incomplete data // IEEE Trans. On Systems, Man and Cybernetics. – 2001. – 31. – N. 5. – Pp. 735-744.*

*6. Brouwer R.K. Fuzzy set covering of a set of ordinal attributes without parameter sharing / R.K. Brouwer // Fuzzy Sets and Systems. – 2006. – 157. – N 13. – Pp. 1775-1786.*

*7. Pedricz W., Chen Sh.-M. Information Granularity, Big Data and Computational Intelligence. – Cham: Springer, 2015. – 444 p.*

*8. Інструкція про єдиний облік злочинів. [Електронний ресурс]. – Режим доступу: http://zakon4.rada.gov.ua/laws/show/v0020900-02/page.*

*9. Методичні рекомендації щодо алгоритму дій користувачів з організації формування Інтегрованої інформаційно-пошукової системи органів внутрішніх справ України: від 16.01.2014 № 727/3р.*

*10. Westphal C. Data Mining for Intelligence, Fraud and Criminal Detection. Advanced Analytic & Information Sharing Technologies / C. Westphal. – Boca Raton: CRC Press, 2009. – 426 p.*

*11. Mena J. Investigative Data Mining for Security and Criminal Detection. – Amsterdam: Elsevier Science, 2003. – 452 p.*

*Поступила в редколлегию 26.05.2017*

**Рецензент:** д-р. техн. наук проф. В.О. Филатов, Харьковский национальный университет радиоэлектроники, Харьков.

## УЗАГАЛЬНЕНА МЕТРИКА В ЗАДАЧІ АНАЛІЗУ БАГАТОВИМІРНИХ ДАНИХ З РІЗНОТИПНИМИ ОЗНАКАМИ

Є.В. Бодянський, В.М.Струков, Д.Ю. Узлов

*Робота присвячена задачі оцінки близькості багатовимірних об'єктів, ознаки яких вимірюються в різних шкалах, а оброблювані дані мають велику розмірність і містять специфічні текстові поля і прогалини. До даних з такими специфічними особливостями неможливо безпосередньо застосовувати класичні алгоритми кластеризації та класифікації. Запропоновано узагальнену метрику в багатовимірному просторі таких об'єктів, яка дозволяє будувати алгоритми кластеризації, класифікації та асоціації, засновані на ній, з використанням класичних методів.*

***Ключові слова:*** *багатовимірні об'єкти, кластеризація, класифікація, шкала вимірів, кількісна метрика, категоріальна метрика, рангова метрика, текстова метрика.*

## GENERALIZED METRICS IN THE PROBLEM OF ANALYSIS OF MULTIDIMENSIONAL DATA WITH DIFFERENT SCALES

E. Bodyanskiy, V. Strukov, D. Uzlov

*The work is devoted to the problem of evaluating the proximity of multidimensional objects, the characteristics of which are measured in different scales, and the data being processed are of large dimension and contain specific text fields and omissions. Data with such specific features can not be directly processed with the classical algorithms of clustering and classification. A generalized metric is proposed in the multidimensional space of such objects, which makes it possible to build algorithms for clustering, classifications and associations based on it, using classical methods.*

***Keywords:*** *multidimensional objects, clusterization, classification, measurement scale, quantitative metric, categorical metric, rank metric, text metric.*