УДК 343.346.8:004.056.53

Е.В. Бодянский, В.М. Струков, Д.Ю. Узлов

ЗАДАЧА ОЦЕНКИ БЛИЗОСТИ МНОГОМЕРНЫХ ОБЪЕКТОВ АНАЛИЗА ДАННЫХ

Введение. Задача оценки близости многомерных объектов достаточно хорошо исследована, предложены различные способы метризации пространства признаков, способы оценки близости многомерных объектов, которые применяются в задачах кластеризации, классификации, ассоциации [1-4]. Вместе с тем, на практике встречаются подобные задачи, обладающие некоторыми особенностями, которые не позволяют непосредственно применять классические способы и методики. Одной из таких практически важных задач является задача оценки близости многомерных объектов, информация о которых накапливается в базах данных подразделений информационного обеспечения полиции Украины. В качестве таких объектов в данном случае являются лица, предметы и события. Характерными особенностями накопленных массивов данных являются:

1) большие объемы данных (до нескольких десятков, а иногда и сотен миллионов записей); причем количество этих данных с каждым днем увеличивается;

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

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

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

Первая и вторая особенности обуславливают очень большую размерность массивов обрабатываемых данных, что позволяет отнести данную задачу к категории Big Data Mining [5].

Третья особенность обусловлена достаточно строгой регламентацией процесса регистрации и ввода данных о происшествиях в интегрированную информационно-поисковую систему (ИИПС) органов внутренних дел, правила которой достаточно подробно изложены в [6,7].

Четвертая особенность является следствием нестрогого соблюдения операторами правил регистрации и ввода, описанных в [6], вследствие чего в базах данных ИИПС возникают пропуски в данных. В связи с этим непосредственно применять известные метрики и, соответственно, использовать основанные на них алгоритмы кластеризации, ассоциации или классификации не представляется возможным.

Исследование характеристик массивов данных, хранящихся в ИИПС, позволяет сделать вывод о том, что, как правило, это данные, измеряются в числовых, категориальных и ранговых шкалах. Работа с данными этих типов достаточно хорошо исследована и описана в литературе [1-5]. Вместе с тем, наличие описанных выше особенностей используемых массивов не дает возможности применять непосредственно известные алгоритмы для их обработки.

Постановка задачи. Для обеспечения возможности применения классических алгоритмов кластеризации, классификации и ассоциации в задачах обработки исследуемых массивов данных необходимо формализовать перечисленные особенности этих массивов и на этой основе разработать метрики для соответствующих типов данных, а также комбинированную метрику в многомерном пространстве признаков.

Изложение основного материала. Введем следующие обозначения:

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

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

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

Полностью заполненная (идеальная) матрица “объект-свойство” имеет вид:

об ъект	x_{i1}	x_{i2}	…	x_{ij}	$x_{i,j+1}$	…	x_{ip}	…	$x_{i,n-1}$	x_{in}
	Наименование шкалы									

	cat	num	…	cat	cat	…	rank	…	rank	num
	Наименование признака (свойства)									
	цвет волос	возраст	…	сем. полож.	пол	…	телосл.	…	доход	рост
x_1	блондин	18	…	женат	м	…	худощ	…	низкий	150
⋮	⋮	⋮		⋮	⋮		⋮		⋮	⋮
x_i	брюнет	22	…	разведен	м	…	толст	…	высокий	180
⋮	⋮	⋮		⋮	⋮		⋮		⋮	⋮
x_m	шатен	40	…	незам.	ж	…	очень толст	…	средний	165

Фрагмент матрицы с пропусками имеет вид:

	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i6}
⋮						
x_i	брюнет	22	none	none	толст	180
x_l	none	40	незам.	none	none	165
⋮						

Здесь none – отсутствующие данные в ячейке.

Обозначим далее:

n_i – число пропусков в объекте x_i, (в примере $n_i = 2$),

n_l – число пропусков в объекте x_l, (в примере $n_l = 3$),

n_{il} – число общих пропусков (в данном примере – в четвертом столбце $n_{il} = 1$);

$x_i \cap x_l \neq \varnothing$.

С учетом введенных обозначений, используя модель “частичного расстояния” [8], расстояние d_{il} между объектами x_i и x_l в общем виде можно записать следующим образом:

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

$$\delta_{il} = \begin{cases} 1, (x_{ij} \neq none) \vartheta (x_{lj} \neq none), \\ 0, (x_{ij} = none) ! (x_{lj} = none). \end{cases} \qquad (2)$$

Очевидно, что для различных шкал измерений расстояние между значениями признаков x_i

и x_l в выражении (1) будут вычисляться по-разному.

1. Количественные метрики.

Способы вычисления расстояний (метрики) в числовых шкалах известны [1,2]. Для использования их в выражении (1) целесообразно выполнить следующую нормализацию:

$$x_{jmin} \le x_{ij} \le x_{jmax},$$

$$x_{ij} = \frac{x_{ij} - x_{j\min}}{x_{j\max} - x_{j\min}}, \qquad (3)$$

$$0 \le x_{ij} \le 1$$

Наиболее часто используемыми количественными метриками являются евклидова и манхэттенская. С учетом введенных выше обозначений и выражений (1)-(3) выражение для евклидовой метрики в нашей задаче может быть записано следующим образом:

$$d_{il}^{numE} = \frac{1}{n - n_i - n_l + n_{il}} \sum_{j=1}^{n} (x_{ij} - x_{lj})^2 \delta_{il}; \qquad (4)$$

$$0 \le d_{il}^{numE} \le 1. \qquad (5)$$

Выражение для манхэттенской метрики можно записать так:

$$d_{il}^{numBC} = \frac{1}{n - n_i - n_l + n_{il}} \sum_{j=1}^{n} \left| x_{ij} - x_{lj} \right| \delta_{il}; \qquad (6)$$

$$0 \le d_{il}^{numBC} \le 1. \qquad (7)$$

2. Категориальная метрика.

Категориальная метрика применяется для множеств значений, которые выражают какие-то неизмеримые качества объектов, например, цвет волос (шатен, блондин, брюнет), к которым применимы только операции отношения «равно» или «не равно». Традиционно категориальная метрика выражается следующей формулой:

$$d_{il}^{cat} = \sum_{j=1}^{n} \delta(x_{ij}, x_{lj}); \qquad (8)$$

$$\delta(x_{ij}, x_{lj}) = \begin{cases} 1, если\ x_{ij} \neq x_{lj}, \\ 0, если\ x_{ij} = x_{lj}. \end{cases} \qquad (9)$$

Так, например, если все соответствующие значения признаков объектов x_i, x_l совпадают – $x_{ij} = x_{lj}$ для всех $j = 1, 2, \ldots, n$, то в этом случае $d_{il}^{cat} = 0,$ если же все соответствующие значения признаков различны, то $d_{il}^{cat} = n,$ Таким образом:

$$0 \le d_{il}^{cat} \le n. \qquad (10).$$

В нашей задаче удобнее применить не метрику (10), а ее нормированный вариант:

$$d_{il}^{cat} = \frac{1}{n} \sum_{j=1}^{n} \delta(x_{ij}, x_{lj}), \qquad (11)$$

$$0 \le d_{il}^{cat} \le 1. \qquad (12)$$

Для совместного использования различных шкал с учетом наличия пропусков данных целесообразно применять модифицированную категориальную метрику:

$$d_{il}^{cat'} = \frac{1}{n - n_i - n_l + n_{il}} \sum_{j=1}^{n} \delta(x_{ij}, x_{lj}) \delta_{il}, \qquad (13)$$

$$0 \le d_{il}^{cat'} \le 1. \qquad (14)$$

3. Ранговая метрика.

Ранговая метрика применяется в тех случаях, когда значения не имеют числового выражения, но между ними существуют отношения порядка – «больше», «меньше», «равно».

Пусть j-тый признак имеет R_j рангов: $r_j = 1, 2, \ldots, R_j$; т.е. вместо x_{ij} обрабатывается $x_{ij}^{r_j}, \forall i = 1, 2, \ldots, N.$ Таким образом, в соответствующей клетке матрицы стоит лингвистическая переменная $x_{ij}^{r_j}$, т.е. $x_i = \left\{x_{ij}^{r_j}\right\}.$

Расстояние в ранговой метрике может быть введено на основе распределения частот:

$$f_j^{r_j} = \frac{N_j^{r_j}}{N_j}, \qquad (15)$$

где N_j – вследствие наличия пропусков данных может быть не равно N;

$N_j^{r_j}$ - число появлений r_j-того ранга в j-том столбце.

Вводя накопительные частоты:

$$F_j^1 = \frac{f_j^1}{2}, \qquad F_j^{r_j} = \frac{f_j^{r_j}}{2} + \sum_{q=1}^{r_j-1} f_j^q; \qquad (16)$$

$$\sum_{q=1}^{R_j} f_j^q = 1;$$

можем ранги заменить их числовыми значениями, основанными на частотах появлений[9]:

$$x_{ij}^1 = \frac{f_j^1}{2}, \qquad x_{ij}^{r_j} = x_{ij}^{r_j-1} + 0{,}5(f_j^{r_j-1} + f_j^{r_j}), \qquad (17)$$

Выполняя далее нормализацию полученных выражений для приведения переменных в нашей задаче к единому основанию – в интервал [0, 1]:

$$x_{ij}^{r_j'} = \frac{x_{ij}^{r_j} - x_{ij}^1}{x_{ij}^{R_j} - x_{ij}^1}, \qquad (18)$$

можно записать расстояние d_{il}^{rank} между x_i и x_l в ранговой метрике в виде:

$$d_{il}^{rank} = \frac{1}{n - n_i - n_l + n_{il}} \sum_{j=1}^{n} \left| x_{ij}^{r_j'} - x_{il}^{r_j'} \right| \delta_{il}, \qquad (19)$$

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

В случае использования всех трех метрик расстояние между x_i и x_l будет иметь следующий вид:

$$d_{il} = \frac{1}{3}(d_{il}^{num'} + d_{il}^{cat'} + d_{il}^{rank'}), \qquad (20)$$

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

Предложенная модель может применяться в случаях, когда все признаки равноценны с точки зрения определения меры близости. В реальных задачах часто возникает необходимость определять степень близости между объектами криминальных учетов не по всем признакам, а

по некоторому подмножеству значимых в данной ситуации признаков, а иногда даже по одному какому-то критическому для конкретных случаев признаку. Кроме того, даже в случае учета всех признаков их значимость для определения степени близости между объектами, как правило, неравнозначна. В связи с этим введем в выражение (20) коэффициенты значимости признаков k_j, которые определяются экспертами-аналитиками в ходе решения конкретных задач:

$$d_{il} = \sum_{j=1}^{n} k_j d_{il}^{j} \delta_{il}; \qquad (21)$$

где $d_{il} \in \{d_{il}^{num'}, d_{il}^{cat'}, d_{il}^{rank'}\}$.

Пронормировав коэффициенты k_j :

$$k_j^{'} = \frac{k_j}{\sum_{j=1}^{n} k_j}, \qquad (22)$$

$$\sum_{j=1}^{n} k_j^{'} = 1,$$

с учетом (22) выражения (4), (6), (13) и (19) для определения степени близости между объектами x_i и x_l в рассмотренных метриках можно записать следующим образом:

$$d_{il}^{numE'} = \frac{1}{n - n_i - n_l + n_{il}} \sum_{j=1}^{n} k_j^{'} (x_{ij} - x_{lj})^2 \delta_{il}; \qquad (23)$$

$$d_{il}^{numBC'} = \frac{1}{n - n_i - n_l + n_{il}} \sum_{j=1}^{n} k_j^{'} \left| x_{ij} - x_{lj} \right| \delta_{il}; \qquad (24)$$

$$d_{il}^{cat''} = \frac{1}{n - n_i - n_l + n_{il}} \sum_{j=1}^{n} k_j^{'} \delta(x_{ij}, x_{lj}) \delta_{il}; \qquad (25)$$

$$d_{il}^{rank'} = \frac{1}{n - n_i - n_l + n_{il}} \sum_{j=1}^{n} k_j^{'} \left| x_{ij}^{r_j^{'}} - x_{il}^{r_j^{'}} \right| \delta_{il}, \qquad (26)$$

Подставив выражения (24)-(26) в (21) получим выражение для определения расстояния между объектами x_i и x_l в общем виде:

$$d_{il} = \frac{1}{n - n_i - n_l + n_{il}} \left(\sum_{j=1}^{n} k_j' \left| x_{ij} - x_{lj} \right| \delta_{il} + \sum_{j=1}^{n} k_j' \delta(x_{ij}, x_{lj}) \delta_{il} + \sum_{j=1}^{n} k_j' \left| x_{ij}^{r_j'} - x_{il}^{r_j'} \right| \delta_{il} \right); \qquad (27)$$

Для корректности вычислений расстояния по обобщенной формуле введем "флажки" для каждой из метрик:

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

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

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

Тогда, подставив их в (27), получим итоговое выражение для определения расстояния между объектами x_i и x_l в общем виде:

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

$$\left. \sum_{j=1}^{n} b_j^{cat} k_j' \delta(x_{ij}, x_{lj}) \delta_{il} + \sum_{j=1}^{n} b_j^{rank} k_j' \left| x_{ij}^{r_j'} - x_{il}^{r_j'} \right| \delta_{il} \right). \qquad (28)$$

Проиллюстрируем применение построенной модели на примере. Пусть имеется следующая матрица “объект-свойство”:

	Пол	*Возраст (на вид)*	*Рост*	*Телосложение*	*Цвет волос*	*Вид преступления*
	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i6}
x_1	М	молодой	средний	упитанный	none	грабеж
x_2	Ж	none	средний	толстый	светлый	кража
x_3	М	пожилой	низкий	худой	седой	кража
x_4	М	none	высокий	худой	none	бандитизм
x_5	Ж	средний	none	крепкий	желтый	кража

С помощью сформулированной метрики (28) определим расстояние между объектами x_2 и x_4.

$n=6$, $n_2=1$, $n_4=2$, $n_{24}=1$. Для простоты вычислений примем $k_1=k_2=k_3=k_4=k_5=k_6=1$.

$$d_{24}=\frac{1}{n-n_2-n_4+n_{24}}\Big(\sum_{j=1}^{6}b_j^{num}k_j^{'}\left|x_{ij}-x_{lj}\right|\delta_{24}+$$

$$\sum_{j=1}^{6}b_j^{cat}k_j^{'}\delta(x_{ij},x_{lj})\delta_{24}+\sum_{j=1}^{6}b_j^{rank}k_j^{'}\left|x_{ij}^{r_j^{'}}-x_{il}^{r_j^{'}}\right|\delta_{24}. \qquad (29)$$

$$\frac{1}{n-n_2-n_4+n_{24}}=\frac{1}{6-1-2+1}=0{,}25;$$

$$k_1^{'}=k_2^{'}=k_3^{'}=k_4^{'}=k_5^{'}=k_6^{'}=\frac{1}{6};$$

Первое слагаемое в (29) будет равно 0, поскольку

$$b_1^{num}=b_2^{num}=b_3^{num}=b_4^{num}=b_5^{num}=b_6^{rank}=0.$$

Три признака в примере измеряются в ранговой метрике:

$$b_1^{rank}=0,\quad b_2^{rank}=1,\quad b_3^{rank}=1,\quad b_4^{rank}=1,\quad b_5^{rank}=0,\quad b_6^{rank}=0.$$

Три признака - в категориальной:

$$b_1^{cat}=1,\quad b_2^{cat}=0,\quad b_3^{cat}=0,\quad b_4^{cat}=0,\quad b_5^{cat}=1,\quad b_6^{cat}=1.$$

Вычислим компоненты третьего слагаемого.

Для второго признака (x_{i2}).

$R_2=3$, $\quad x_{i2}^{rank}=\{x_{i2}^1,x_{i2}^2,x_{i2}^3\}=\{$молодой, средний, пожилой$\}$,

$$f_2^1=\frac{N_2^1}{N_2}=\frac{1}{3},\ f_2^2=\frac{N_2^2}{N_2}=\frac{1}{3},\ f_2^3=\frac{N_2^3}{N_2}=\frac{1}{3},$$

$$F_2^1=\frac{f_2^1}{2}=\frac{1}{6},\ f_2^2=\frac{N_2^2}{N_2}=\frac{1}{3},\ f_2^3=\frac{N_2^3}{N_2}=\frac{1}{3},$$

$$F_2^2=\frac{f_2^2}{2}+f_2^1=\frac{1}{6}+\frac{1}{3}=\frac{1}{2},$$

$$F_2^3=\frac{f_2^3}{2}+f_2^1+f_2^2=\frac{1}{6}+\frac{1}{3}+\frac{1}{3}=\frac{5}{6}.$$

Значения второго признака во второй и четвертой строках не определено: x_{22}=*none;* x_{42}=*none.* Таким образом, перейдя от лингвистических переменных в исходной матрице к их частотным эквивалентам в соответствии с (17), получим:

$$x_{i2}^1 = F_2^1 = \frac{1}{6};\ \ x_{i2}^2 = F_2^2 = \frac{1}{2};\ \ x_{i2}^3 = F_2^3 = \frac{5}{6}.$$

Пронормировав полученные значения в соответствии с (18), получим:

$$x_{i2}^{1'} = 0, x_{i2}^{2'} = \frac{\frac{1}{2} - 0}{\frac{5}{6} - 0} = \frac{6}{10} = \frac{3}{5};\ \ x_{i2}^{3'} = 1.$$

Выполняя аналогичные действия для третьего и четвертого столбцов матрицы, получим:

R_3=3, $x_{i3}^{rank} = \{x_{i3}^1, x_{i3}^2, x_{i3}^3\} = \{\text{низкий, средний, высокий}\}$,

$$x_{i3}^{1'} = 0, x_{i3}^{2'} = \frac{1}{2}; x_{i3}^{3'} = 1;$$

R_4=3,

$$x_{i4}^{rank} = \{x_{i4}^1, x_{i4}^2, x_{i4}^3, x_{i4}^4\} = \{\text{худой, крепкий, упитанный, толстый}\},$$

$$x_{i4}^{1'} = 0, x_{i4}^{2'} = \frac{3}{7}; x_{i4}^{3'} = \frac{5}{7}, x_{i4}^{4'} = 1.$$

Подставляя полученные значения в исходную матрицу, получим:

	Пол	*Возраст (на вид)*	*Рост*	*Телосложение*	*Цвет волос*	*Вид преступления*
	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i6}
x_1	М	0	$\frac{1}{2}$	$\frac{5}{7}$	none	грабеж
x_2	Ж	none	$\frac{1}{2}$	1	светлый	кража
x_3	М	1	0	0	седой	бандитизм
x_4	М	none	1	0	none	кража
x_5	Ж	$\frac{3}{5}$	none	$\frac{3}{7}$	желтый	кража

Далее, в соответствии с (29) вычисляем расстояние в предложенной метрике между

вторым и четвертым объектами:

$$d_{24} = \frac{1}{4}\Big(\sum_{j=1}^{6} b_j^{num} k_j^{'} \left| x_{ij} - x_{lj} \right| \delta_{24} + \sum_{j=1}^{6} b_j^{cat} k_j^{'} \delta(x_{ij}, x_{lj}) \delta_{24} +$$

$$\sum_{j=1}^{6} b_j^{rank} k_j^{'} \left| x_{ij}^{r_j^{'}} - x_{il}^{r_j^{'}} \right| \delta_{24} = \frac{1}{4}\Big(\frac{1}{6} \sum_{j=1}^{6} b_j^{cat} k_j^{'} \delta(x_{ij}, x_{lj}) \delta_{24} + \frac{1}{6} \sum_{j=1}^{6} b_j^{rank} k_j^{'} \left| x_{ij}^{r_j^{'}} - x_{il}^{r_j^{'}} \right| \delta_{24}\Big) =$$

$$\frac{1}{24}\left(\left(1+0+0+0+0+0\right) + \left(0+0+\frac{1}{2}+1+0+0 \right) \right) = \frac{5}{48}.$$

Заключение. Предложенный способ определения близости многомерных объектов, учитывающий особенности формирования баз данных криминальных учетов, 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. В частности, таких задач как поиск преступлений по аналогии, определение круга подозреваемых по определенному преступлению или по группе преступлений и т.п.

Список литературных источников.

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

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

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

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

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

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

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

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

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