

Новые классы метрических пространств измеримых множеств и мультимножеств в кластерном анализе*

А. Б. Петровский¹

Понятия близости, сходства или различия исследуемых объектов играют важную роль во многих теоретических и практических задачах принятия решений, искусственного интеллекта, распознавания образов, обработки разнородной информации и др. Эти понятия обычно формализуются путем определения расстояния между объектами в признаковом пространстве. В работе предложены новые классы метрических пространств измеримых множеств и мультимножеств, рассмотрены свойства метрик. Показаны возможности применения новых понятий при агрегировании и кластерном анализе многопризнаковых объектов.

Введение

В различных теоретических и практических проблемах принятия решений, искусственного интеллекта, распознавания образов, обработки разнородной информации и других предметных областей часто возникает необходимость оценить близость, сходство или различие исследуемых объектов, основываясь на их свойствах, которые выражены признаками объектов. Признаки, характеризующие свойства объектов, могут быть количественными, качественными, смешанными. Понятия близости, сходства или различия объектов задаются с помощью расстояния между объектами в признаковом пространстве.

Вместе с тем имеется достаточно широкий круг задач, где анализируемые объекты характеризуются многими разнородными признаками, и, кроме того, одни и те же объекты могут существовать в нескольких экземплярах с отличающимися значениями признаков, свертка которых или не-

* Работа выполнена при поддержке программ фундаментальных исследований РАН «Математическое моделирование и интеллектуальные системы» и ОИТВС РАН «Фундаментальные основы информационных технологий и систем», РФФИ (проекты 04-01-00290, 05-01-00666), гранта Президента Российской Федерации НШ 1964.2003.1 для поддержки ведущих научных школ.

¹ 117312, Москва, проспект 60-летия Октября, 9, ИСА РАН, rab@isa.ru.

возможна, или математически некорректна. В качестве примеров таких задач укажем классификацию и ранжирование объектов, оцененных несколькими экспертами по многим качественным критериям, распознавание графических символов, обработку текстовых документов. Множественность и повторяемость факторов, описывающих объекты, усложняет и затрудняет решение таких задач.

Удобной математической моделью для описания многопризнаковых объектов является мультимножество или множество с повторяющимися элементами. Кратность элементов — существенная особенность мультимножества, отличающая его от множества и позволяющая считать мультимножество качественно новым математическим понятием. В работе предложены новые классы метрик в пространствах измеримых множеств и мультимножеств, рассмотрены свойства метрик. Указаны примеры применения метрик в задачах классификации, кластерного анализа и упорядочения многопризнаковых объектов.

1. Представление многопризнаковых объектов

Многопризнаковые объекты A_i , $i = 1, \dots, n$ обычно принято представлять как векторы или кортежи $q_i = (q_{i1}^{e_1}, \dots, q_{im}^{e_m})$ в пространстве $Q = Q_1 \times \dots \times Q_m$, где $Q_s = \{q_s^{e_s}\}$ — непрерывная или дискретная шкала s -го признака, $e_s = 1 \div h_s$, $s = 1, \dots, m$. Ситуация существенно осложняется, если одному и тому же объекту A_i может соответствовать не один, а несколько m -мерных векторов с различающимися значениями признаков. Подобные ситуации возникают, например, когда объект A_i оценивается k независимыми экспертами по m критериям, или когда необходимо одновременно учесть m параметров объекта A_i , измеренных k различными способами. В таких случаях объект A_i представляется в m -мерном пространстве Q уже не одним вектором q_i , а группой, состоящей из k векторов $\{q_i^{(1)}, \dots, q_i^{(k)}\}$ вида $q_i^{(j)} = (q_{i1}^{e_1(j)}, \dots, q_{im}^{e_m(j)})$, $j = 1, \dots, k$, которая должна рассматриваться как единое целое. При этом, очевидно, измеренные разными способами значения параметров, как и индивидуальные оценки, данные экспертами, могут быть похожими, различающимися и даже противоречивыми, что в свою очередь может приводить к несравнимости m -мерных векторов $q_i^{(j)}$, характеризующих один и тот же объект A_i .

Совокупность таких «составных» объектов может иметь в пространстве Q сложную структуру, достаточно трудную для анализа. Непросто ввести в этом пространстве и метрику для измерения расстояний между объектами. Указанные трудности можно преодолеть, если воспользоваться

иным способом представления многопризнаковых объектов, основанным на формализме мультимножеств [Петровский, 2003], который позволяет одновременно учесть различные комбинации значений количественных и качественных признаков, а также их многозначность.

Введем вместо прямого произведения m шкал признаков $Q = Q_1 \times \dots \times Q_m$ обобщенную последовательную шкалу признаков — множество $G = \{Q_1, \dots, Q_m\}$, состоящее из m групп признаков, и представим объект A_i в таком символическом виде:

$$A_i = \{k_{A_i}(q_1^1) \cdot q_1^1, \dots, k_{A_i}(q_1^{h_1}) \cdot q_1^{h_1}, \dots, \\ k_{A_i}(q_m^1) \cdot q_m^1, \dots, k_{A_i}(q_m^{h_m}) \cdot q_m^{h_m}\}, \quad (1)$$

где число $k_{A_i}(q_s^{e_s})$ указывает, сколько раз признак $q_s^{e_s} \in Q_s$ встречается в описании объекта A_i , знак $\cdot$ обозначает кратность вхождения признака $q_s^{e_s}$. Например, при многокритериальной оценке объекта A_i несколькими экспертами число $k_{A_i}(q_s^{e_s})$ равно числу экспертов, давших объекту A_i оценку $q_s^{e_s}$ по критерию Q_s . Объект A_i можно записать и более единообразно как $A_i = \{k_{A_i}(x_1) \cdot x_1, \dots, k_{A_i}(x_h) \cdot x_h\}$, определив элементы множества $G = \{x_1, \dots, x_h\}$ следующим образом:

$$x_1 = q_1^1, \quad x_2 = q_1^2, \quad \dots, \quad x_{h_1} = q_1^{h_1}, \quad x_{h_1+1} = q_2^1, \quad \dots, \quad x_{h_1+h_2} = q_2^{h_2}, \quad \dots, \\ x_{h_1+\dots+h_{m-1}+1} = q_m^1, \quad \dots, \quad x_{h_1+\dots+h_m} = q_m^{h_m},$$

где $h = h_1 + \dots + h_m$. Множество G характеризует свойства совокупности объектов $A = \{A_1, \dots, A_n\}$. Такая запись объектов A_i представляет их как множества с повторяющимися элементами $x_j \in G$ или мультимножества.

2. Мультимножества и операции над ними

Дадим краткий обзор теории мультимножеств [Петровский, 2003]. Мультимножеством A , порожденным обычным множеством $G = \{x_1, x_2, \dots\}$, все элементы которого различны, называется совокупность групп элементов вида $A = \{k_A(x) \cdot x | x \in G, k_A(x) \in \mathbb{Z}_+\}$. Здесь $k_A: G \rightarrow \mathbb{Z}_+ = \{0, 1, 2, \dots\}$ называется функцией числа экземпляров мультимножества, определяющей кратность вхождения элемента $x_i \in G$ в мультимножество A , что обозначено символом $\cdot$. Если $k_A(x) = \chi_A(x)$, где $\chi_A(x) = 1$ при $x \in A$ и $\chi_A(x) = 0$ при

$x \notin A$, то мультимножество A становится обычным множеством A . Обычное множество $\text{Supp}A = \{x | x \in G, \chi_{\text{Supp}A}(x) = \min(k_A(x), 1)\}$ называется носителем мультимножества A . Мощность мультимножества $|A| = \sum_x k_A(x)$ определяется как общее число экземпляров всех его элементов; размерность мультимножества $/A/ = \sum_x \chi_A(x) = |\text{Supp}A|$ — как общее число различных элементов. Мультимножество называется пустым $\emptyset$, если $k_{\emptyset}(x) = 0$, постоянным $C_{[t]}$, если $k_{C_{[t]}}(x) = t = \text{const}$, и максимальным Z , если $k_Z(x) = \max_{A \in A} k_A(x), \forall x \in G$. Величина $\text{hgt}A = \max_{x \in G} k_A(x)$ называется высотой

мультимножества. Обычное множество A является, таким образом, постоянным мультимножеством $A_{[1]}$ высоты $\text{hgt}A = 1$.

Введем следующие операции над мультимножествами:

объединение	$A \cup B = \{k_{A \cup B}(x) \cdot x \mid k_{A \cup B}(x) = \max(k_A(x), k_B(x))\};$
пересечение	$A \cap B = \{k_{A \cap B}(x) \cdot x \mid k_{A \cap B}(x) = \min(k_A(x), k_B(x))\};$
сложение	$A + B = \{k_{A+B}(x) \cdot x \mid k_{A+B}(x) = k_A(x) + k_B(x)\};$
вычитание	$A - B = \{k_{A-B}(x) \cdot x \mid k_{A-B}(x) = k_A(x) - k_{A \cap B}(x)\};$
симметрическая разность	$A \Delta B = \{k_{A \Delta B}(x) \cdot x \mid k_{A \Delta B}(x) = k_A(x) - k_B(x) \};$
дополнение	$\overline{A} = Z - A = \{k_{\overline{A}}(x) \cdot x \mid k_{\overline{A}}(x) = k_Z(x) - k_A(x)\};$
умножение на число	$t \cdot A = \{k_{t \cdot A}(x) \cdot x \mid k_{t \cdot A}(x) = t \cdot k_A(x), t \in \mathbb{Z}_+\};$
умножение	$A \bullet B = \{k_{A \bullet B}(x) \cdot x \mid k_{A \bullet B}(x) = k_A(x) \cdot k_B(x)\};$
n -я степень	$A^n = \{k_{A^n}(x) \cdot x \mid k_{A^n}(x) = (k_A(x))^n\};$
прямое произведение	$A \times B = \{k_{A \times B} \cdot \langle x_i, x_j \rangle \mid k_{A \times B} = k_A(x_i) \cdot k_B(x_j), x_i \in A, x_j \in B\};$
прямая n -я степень	$(\times A)^n = \{k_{(\times A)^n} \cdot \langle x_1, \dots, x_n \rangle \mid k_{(\times A)^n} = \prod_{i=1}^n k_A(x_i), x_i \in A\}.$

Носители операций над мультимножествами удовлетворяют следующим соотношениям:

$$\text{Supp}(A \cup B) = \text{Supp}(A + B) = (\text{Supp}A) \cup (\text{Supp}B);$$

$$\text{Supp}(A \cap B) = \text{Supp}(A \bullet B) = (\text{Supp}A) \cap (\text{Supp}B);$$

$$\text{Supp}(A \Delta B) = (\text{Supp}(A - B)) \cup (\text{Supp}(B - A));$$

$$\text{Supp}(t \cdot A) = \text{Supp} A = \text{Supp}(A^n);$$

$$\text{Supp}(A \times B) = (\text{Supp} A) \times (\text{Supp} B); \text{Supp}(\times A)^n = (\times \text{Supp} A)^n.$$

Существует двойственность операций над мультимножествами, аналогичная законам де Моргана для множеств:

$$\overline{A \cup B} = \overline{A} \cap \overline{B}, \quad \overline{A \cap B} = \overline{A} \cup \overline{B};$$

$$\overline{A + B} = \overline{A} - B = \overline{B} - A, \quad \overline{A - B} = \overline{A} + B, \quad \overline{A - B} = B - A.$$

Заметим, что некоторые свойства, которыми обладают операции над множествами, у мультимножеств отсутствуют. В то же время появляются новые свойства, не имеющие аналогов для множеств. Например,

$$A + \overline{A} = Z, \quad Z - \overline{A} = A; \quad A \cap \overline{A} \neq Z, \quad A \cap \overline{A} \neq \emptyset; \quad A - \overline{A} \neq \overline{A} - A \neq \emptyset.$$

При переходе от мультимножеств к множествам и замене $k_A(x)$ на $\chi_A(x)$ многие из выше сформулированных утверждений для мультимножеств останутся справедливыми и для множеств, но некоторые из них могут изменить или потерять смысл. Например, операции арифметического сложения и умножения на число для множеств неопределимы, вычитание множеств определяется иначе, а арифметическое умножение множеств будет совпадать с их пересечением.

3. Метрические пространства мультимножеств

Различные классы метрических пространств мультимножеств (A, d) определяются следующими метриками (псевдометриками):

$$d_{1p}(A, B) = [m(A \Delta B)]^{1/p}, \quad d_{2p}(A, B) = [m(A \Delta B)/m(Z)]^{1/p};$$

$$d_{3p}(A, B) = [m(A \Delta B)/m(A \cup B)]^{1/p}, \quad (2)$$

где p — целое число, m — мера мультимножества, действительная неотрицательная функция, заданная на алгебре мультимножеств $L(Z)$. Алгеброй мультимножеств называется семейство мультимножеств $L(Z)$, замкнутое относительно операций объединения, пересечения, сложения и дополнения. Максимальное мультимножество Z является единицей алгебры, а пустое мультимножество $\emptyset$ — нулем. Мера мультимножества m удовлетворяет условию сильной аддитивности $m(\sum_i A_i) = \sum_i m(A_i)$ и обладает свойствами: $m(\emptyset) = 0$; слабая аддитивность $m(\bigcap_i A_i) = \sum_i m(A_i)$ для $A_i \cap A_j = \emptyset$; монотонность

$m(\mathbf{A}) \leq m(\mathbf{B}) \Leftrightarrow \mathbf{A} \subseteq \mathbf{B}$; симметричность $m(\mathbf{A}) + m(\overline{\mathbf{A}}) = m(\mathbf{Z})$; непрерывность $\lim_{i \rightarrow \infty} m(\mathbf{A}_i) = m(\lim_{i \rightarrow \infty} \mathbf{A}_i)$; эластичность $m(t \bullet \mathbf{A}) = tm(\mathbf{A})$. Заметим, что мощность мультимножества $|\mathbf{A}|$ также будет мерой мультимножества.

Меру мультимножества можно определить различными способами, например, как линейную комбинацию функций кратности: $m(\mathbf{A}) = \sum_i w_i k_A(x_i)$, $w_i > 0$. В этом случае метрики приобретают вид

$$d_{1p}(\mathbf{A}, \mathbf{B}) = \left(\sum_{x_i \in G} w_i |k_A(x_i) - k_B(x_i)| \right)^{1/p};$$

$$d_{2p}(\mathbf{A}, \mathbf{B}) = \left(\sum_{x_i \in G} w'_i |k_A(x_i) - k_B(x_i)| \right)^{1/p}, \quad w'_i = 1 / \sum_{j=1}^h w_j;$$

$$d_{3p}(\mathbf{A}, \mathbf{B}) = \left(\frac{\sum_{x_i \in G} w_i |k_A(x_i) - k_B(x_i)|}{\sum_{x_i \in G} w_i \max[k_A(x_i), k_B(x_i)]} \right)^{1/p}.$$

Назовем соответствующие метрики основной, полностью усредненной и локально усредненной. Основная метрика $d_{1p}(\mathbf{A}, \mathbf{B})$ является метрикой типа Хемминга, традиционно используемой во многих приложениях. Полностью усредненная метрика $d_{2p}(\mathbf{A}, \mathbf{B})$ характеризует различие между двумя мультимножествами $\mathbf{A}$ и $\mathbf{B}$, отнесенное к расстоянию, максимально возможному в исходном пространстве. Локально усредненная метрика $d_{3p}(\mathbf{A}, \mathbf{B})$ задает различие, отнесенное к максимально возможной «общей части» $\mathbf{A} \cap \mathbf{B}$ только этих двух мультимножеств в исходном пространстве. Функции $d_{2p}(\mathbf{A}, \mathbf{B})$ и $d_{3p}(\mathbf{A}, \mathbf{B})$ удовлетворяют условию нормировки $0 \leq d(\mathbf{A}, \mathbf{B}) \leq 1$. Функция $d_{3p}(\mathbf{A}, \mathbf{B})$ не определена для $\mathbf{A} = \mathbf{B} = \emptyset$, поэтому по определению принимается $d_{3p}(\emptyset, \emptyset) = 0$.

При замене k_A на χ_A метрические пространства измеримых мультимножеств превращаются в соответствующие метрические пространства измеримых множеств, которые будут обладать многими аналогичными свойствами. Так, основная метрика вида $d_{11}(\mathbf{A}, \mathbf{B}) = m(\mathbf{A} \Delta \mathbf{B})$ или $d_{11}(\mathbf{A}, \mathbf{B}) = |\mathbf{A} \Delta \mathbf{B}|$ приведена во многих монографиях по теории множеств и называется расстоянием Фреше—Никодима—Ароншьяна или метрикой меры. Локально усредненная метрика $d_{31}(\mathbf{A}, \mathbf{B}) = m(\mathbf{A} \Delta \mathbf{B})/m(\mathbf{A} \cap \mathbf{B})$ называется расстоянием Штейнхауса, а $d_{31}(\mathbf{A}, \mathbf{B}) = |\mathbf{A} \Delta \mathbf{B}|/|\mathbf{A} \cap \mathbf{B}|$ — биотопическим

расстоянием [Деза, Лоран, 2001]. Расстояние между множествами вида $d_{12}(A, B) = [m(A\Delta B)]^{1/2}$ упомянуто в книге [Орлов, 1979]. Семейства метрик общего вида $d_{1p}(A, B) = [m(A\Delta B)]^{1/p}$, $d_{2p}(A, B) = [m(A\Delta B)/m(Z)]^{1/p}$ и $d_{3p}(A, B) = [m(A\Delta B)/m(A \cup B)]^{1/p}$ в пространствах измеримых множеств и измеримых мультимножеств введены впервые автором [Петровский, 2003].

4. Свойства метрик и метрических пространств

Расстояния между мультимножествами обладают рядом геометрических свойств, которые можно соотнести с различными видами преобразований пространств (A, d_{qp}) измеримых мультимножеств.

При любом фиксированном целом числе p основная метрика $d_{1p}(A, B)$ и полностью усредненная метрика $d_{2p}(A, B)$:

- инвариантны относительно «мягкого» сдвига на произвольное мультимножество $d_{1p(2p)}(A, B) = d_{1p(2p)}(A + C, B + C)$;
- инвариантны относительно «жесткого» сдвига на произвольное дизъюнктивное мультимножество $d_{1p(2p)}(A, B) = d_{1p(2p)}(A \cup C, B \cup C)$, если $A \cap C = B \cap C = \emptyset$;
- инвариантны относительно исключения общей части $d_{1p(2p)}(A, B) = d_{1p(2p)}(A - C, B - C)$, если $C \subseteq A \cap B$, и $d_{1p(2p)}(A, B) = d_{1p(2p)}(C - A, C - B)$, если $A \cap B \subseteq C$;
- инвариантны относительно симметричных отображений

$$d_{1p(2p)}(A, B) = d_{1p(2p)}(A \cap B, A \cap B) = d_{1p(2p)}(A - B, B - A) = d_{1p(2p)}(\bar{A}, \bar{B}).$$

При $p = 1$ основная метрика $d_{11}(A, B)$ и полностью усредненная метрика $d_{21}(A, B)$

- d -выпуклы относительно мультимножества C , находящегося между мультимножествами A и B , т. е. $d_{11(21)}(A, B) = d_{11(21)}(A, C) + d_{11(21)}(C, B)$, если $A \cap B \subseteq C \subseteq A \cup B$;
- обладают свойством максиминной аддитивности $d_{11(21)}(A, B) = d_{11(21)}(A \cup C, B \cup C) + d_{11(21)}(A \cap C, B \cap C)$ относительно произвольного мультимножества C .

Локально усредненная метрика $d_{3p}(A, B)$:

- инвариантна относительно отображения $d_{3p}(A, B) = d_{3p}(A \cap B, A \cap B)$;
- d -выпукла при $p = 1$ и $C = A \cap B$ $d_{31}(A, B) = d_{31}(A, A \cap B) + d_{31}(A \cap B, B)$.

Основная метрика $d_{1p}(A, B)$ эластична относительно растяжения мультимножеств $d_{1p}(A, B) = (1/t)^{1/p} d_{1p}(t \bullet A, t \bullet B)$, а полностью и локально усредненные метрики $d_{2p}(A, B)$ и $d_{2p}(A, B)$ инвариантны относительно растяжения $d_{2p(3p)}(A, B) = d_{2p(3p)}(t \bullet A, t \bullet B)$, где t — положительное целое число.

Основная метрика $d_{1p}(A, B) = [m(A \Delta B)]^{1/p}$ и полностью усредненная метрика $d_{2p}(A, B) = [m(A \Delta B)/m(Z)]^{1/p}$ являются непрерывными, равномерно непрерывными и равностепенно непрерывными функциями, а локально усредненная метрика $d_{3p}(A, B) = [m(A \Delta B)/m(A \cup B)]^{1/p}$ — кусочно непрерывной функцией своих переменных почти всюду на соответствующем метрическом пространстве при любом фиксированном числе p .

Справедливы следующие утверждения. Метрические пространства измеримых мультимножеств и множеств (A, d_{1p}) , (A, d_{2p}) гомеоморфны, полны, сепарабельны. Кроме того, метрические пространства измеримых мультимножеств и множеств (A, d_{11}) , (A, d_{21}) изометричны пространству ограниченных числовых последовательностей l_1 и метрически выпуклы.

5. Агрегирование многопризнаковых объектов

Широко применяемым методом исследования естественных группировок большого числа объектов и связей между ними является кластерный анализ, в результате которого исходная совокупность объектов $A = \{A_1, \dots, A_n\}$ агрегируется в небольшое количество групп $A = \{X_1, \dots, X_R\}$. В кластерном анализе объединение объектов в группы производится, исходя из их сходства или различия, которое оценивается степенью близости объектов, рассматриваемых как точки признакового пространства.

Принципиальными моментами кластерного анализа являются: выбор выражения, определяющего расстояние между объектами в признаковом пространстве; выбор алгоритма группирования объектов; разумная интерпретация сформированных групп. Выбор вида пространства и типа метрики зависит от свойств анализируемых объектов. Для рассмотренных выше многопризнаковых объектов наиболее адекватно метрическое пространство измеримых мультимножеств (A, d) с основным, полностью или локально усредненным расстоянием, например, d_{11} , d_{21} , d_{31} .

Кластер традиционно формируется как теоретико-множественное объединение наиболее близких объектов [Anderberg, 1973]. Новые типы операций над мультимножествами открывают новые возможности для группирования многопризнаковых объектов. Например, группа X_i объектов может быть получена как сумма $X_i = \sum_i A_i$, объединение $X_i = \cup_i A_i$ или пересечение $X_i = \cap_i A_i$ мультимножеств A_i , описывающих объекты A_i , либо

как линейная комбинация различных мультимножеств вида $X_i = \sum t_i \bullet A_i$, $X_i = {}_i t_i \bullet A_i$ или $X_i = {}_i t_i \bullet A_i$, $t_i > 0$. Когда группа X_i образуется в результате объединения или пересечения объектов A_i , происходит усиление лучших свойств (максимальных значений признаков x_j) или соответственно худших свойств (минимальных значений признаков x_j), присутствующих у отдельных членов группы. При сложении объектов агрегируются все свойства x_j всех членов группы X_i .

Для генерации группы объектов в алгоритмах кластерного анализа обычно используются следующие подходы: минимизировать различие (максимизировать сходство) между объектами внутри группы; максимизировать различие (минимизировать сходство) между группами объектов. Далее будем считать для простоты, что различие между объектами A_i внутри группы X_i , между некоторым объектом A_i и группой объектов X_i , между группами объектов X_i , представленными как точки или группы точек в признаковом пространстве, определяется одинаковым образом и задается одним из вышеприведенных расстояний d_{11} , d_{21} , d_{31} .

6. Иерархический кластерный анализ

Рассмотрим общую схему процедуры иерархической кластеризации совокупности объектов A , представленных мультимножествами $A_1, \dots, A_n$, когда число R генерируемых кластеров $\{X_1, \dots, X_R\}$ заранее не определено.

Шаг 1. Положить $R = n$, где R — число кластеров, n — число объектов A_i . Тогда каждый кластер X_i совпадает с некоторым объектом A_i , $i = 1, \dots, R$.

Шаг 2. Вычислить расстояния $d(X_a, X_b)$ между всеми возможными парами кластеров X_a и X_b , $1 \leq A, B \leq R$, $a \neq b$, используя одну метрику d_{11} , d_{21} , d_{31} метрического пространства мультимножеств.

Шаг 3. Найти пару наиболее близких кластеров X_u, X_v таких, что

$$d(X_u, X_v) = \min_{a,b} d(X_a, X_b), \quad (3)$$

и сформировать новый кластер путем сложения $X_r = X_u + X_v$, объединения $X_r = X_u \cup X_v$ или пересечения $X_r = X_u \cap X_v$ мультимножеств, или путем линейной комбинации одной из этих операций.

Шаг 4. Уменьшить число кластеров на единицу: $R = n - 1$. Если $R = 1$, то остановиться и вывести результат, например, в виде дендрограммы. Если $R > 1$, перейти к следующему шагу.

Шаг 5. Вычислить расстояния $d(X_a, X_b)$ между новыми парами кластеров для всех $1 \leq A, B \leq R, a \neq b \neq r$. Вернуться к шагу 3. ■

Процесс иерархической кластеризации заканчивается, когда все объекты будут объединены в несколько классов или в один единственный класс. Процедура может быть также прервана на некотором шаге, когда индекс различия превысит некоторый пороговый уровень. Методы оптимизации дендрограмм, основанные на поиске удовлетворительного порогового уровня $d_{\text{opt}}(X_u, X_v)$ для расстояния [Miyamoto, 1987], позволяют дать разумную интерпретацию сформированным группам объектов.

Характер образования кластеров и результаты процесса кластеризации в значительной мере определяется видом используемой метрики d . Основная метрика d_{11} и полностью усредненная метрика d_{21} дают практически одинаковые результаты. При кластеризации сначала происходит слияние «небольших» (имеющих небольшое число признаков) объектов, мало различающихся между собой, а затем начинается агрегирование более «крупных» объектов. Получающиеся в итоге группы более или менее сопоставимы по числу входящих в них объектов, но сильно отличаются друг от друга наборами характеризующих их признаков. Процесс кластеризации с локально усредненной метрикой d_{31} начинается с группирования похожих объектов «среднего» и «большого» размеров, имеющих существенные «общие» наборы признаков, к которым далее присоединяются все более и более отличающиеся «небольшие» объекты. Итоговые группировки объектов, полученные в первом и во втором случаях, могут различаться достаточно сильно.

Отметим, что процедура иерархической кластеризации, начиная с шага 3, может сильно ветвиться из-за большого числа эквивалентных пар кластеров X_u, X_v , находящихся в признаковом пространстве на одинаковом минимальном расстоянии. Поэтому появляется много вариантов для дальнейшего слияния кластеров. И, как следствие, в итоге могут сформироваться несовпадающие конечные группы объектов даже при использовании одной и той же метрики. Наименьшее число таких конечных групп возникает, если агрегирование объектов осуществляется путем сложения мультимножеств, наибольшее — при пересечении мультимножеств. Использование локально усредненной метрики d_{31} порождает меньшее число точек ветвления алгоритма по сравнению с основной метрикой d_{11} и полностью усредненной метрикой d_{21} .

Чтобы уменьшить число возможных комбинаций при выборе пар соединяемых кластеров, предлагается вместо единственного критерия бли-

зости кластеров (3) использовать несколько различных критериев. В этом случае алгоритм иерархической кластеризации модифицируется.

Шаг 3.1. Найти все эквивалентные пары наиболее близких кластеров X_u, X_v , соответствующих условию (3), и с помощью одной из операций (суммирования, объединения или пересечения) сформировать новые кластеры $X_{rl}, l = 1, \dots, g$ для всех g эквивалентных пар кластеров.

Шаг 3.2. Найти кластер $X_r^\#$, минимизирующий некоторый дополнительный критерий, например, компактность кластера, выражаемую функционалом

$$K(X_r^\#) = \min_l \sum_{i,j \in I_{rl}} d(A_i, A_j) / N_{rl},$$

где N_{rl} равно числу объектов A_i в кластере X_{rl} . Перейти к шагу 4.

7. Неиерархический кластерный анализ

В методах неиерархического кластерного анализа число кластеров R считается фиксированным и заданным заранее. В алгоритмах кластеризации часто используется понятие центра A_t° кластера $X_t, t = 1, \dots, R$, который находится как решение задачи минимизации некоторого функционала от расстояния, например, следующего вида:

$$J(A_t^\circ, X_t) = \min_j \sum_{i \in I_t} d(A_i, A_j). \quad (4)$$

Функционал $J(A_t^\circ, X_t)$ по своему содержанию близок критерию $K(X_r^\#)$ компактности кластера и характеризует своего рода «кучность» распределения объектов в признаковом пространстве. Заметим, что в нашем случае центр A_t° кластера X_t может совпадать с одним из реально имеющихся объектов A_i из совокупности A , или быть так называемым «фантомным» объектом, сконструированным из признаков x_j в виде (1), но реально отсутствующим в исходной совокупности объектов A .

Обобщенная схема неиерархической кластеризации совокупности объектов $A = \{A_1, \dots, A_n\}$, представленных мультимножествами, включает следующие основные этапы.

Шаг 1. Выбрать некоторое начальное разбиение совокупности объектов A на R кластеров $A = \{X_1, \dots, X_R\}$.

Шаг 2. Распределить все объекты A_i по кластерам X_t ($t = 1, \dots, R$) в соответствии с некоторым правилом. Например, вычислить расстояния $d(A_i, X_t)$ между каждым объектом A_i и кластерами X_t , и поместить объект A_i в ближайший из кластеров X_r , который определяется условием $d(A_i, X_r) = \min_{1 \leq t \leq R} d(A_i, X_t)$. Или определить центр A_t°

для каждого кластера X_t , решив уравнение (4), и поместить каждый объект A_i в кластер с ближайшим центром A_s° , задаваемый условием $d(A_i, A_s^\circ) = \min_{1 \leq t \leq R} d(A_i, A_t^\circ)$.

Шаг 3. Если после распределения всех объектов A_i по кластерам X_t объекты не изменят своей кластерной принадлежности, заданной первоначальным разбиением, то остановиться и вывести результат. В противном случае вернуться к шагу 2. ■

Результаты классификации объектов можно оценить качеством разбиения, в частности, как решение следующей оптимизационной задачи:

$$H(X_{\text{opt}}) = \min \sum_{t=1}^R J(A_i, X_t), \quad (5)$$

где функционал $J(A_i^\circ, X_t)$ определяется, например, выражением (4). В общем случае решение задачи (5) неоднозначно, поскольку функционал $H(X_{\text{opt}})$ качества разбиения является функцией, имеющей много локальных экстремумов. Конечный результат зависит и от первоначального (близко или далеко от оптимального) распределения объектов по классам.

Часто в процедурах кластеризации вместо минимизации расстояния d между объектами, характеризующего их различия, используется максимизация разнообразных показателей сходства объектов s . Можно задать следующие показатели (индексы) сходства объектов, представленных мультимножествами:

$$s_1(A, B) = 1 - m(A \Delta B) / m(Z); \quad s_2(A, B) = m(A \cap B) / m(Z); \\ s_3(A, B) = m(A \cap B) / m(A \cup B).$$

Функции s_1, s_2, s_3 обобщают на случай мультимножеств известные индексы сходства объектов, такие соответственно как коэффициент простой согласованности, мера сходства Рассела—Рао, коэффициент Жаккара или мера Роджерса—Танимото [Anderberg, 1973]. Коэффициент простой согласованности s_1 и мера сходства Рассела—Рао s_2 связаны соотношением $s_1(A, B) = s_2(A, B) + s_2(\bar{A}, \bar{B})$, которое представляет

собой одно из возможных бинарных разложений максимального мультимножества Z на блоки из покрытий и перекрытий мультимножеств [Петровский, 2000].

Заключение

В работе представлены новые классы пространств измеримых множеств и мультимножеств и новые виды метрик, которые могут найти свое применение при решении разнообразных задач классификации, сортировки, ранжирования многопризнаковых объектов. Представление многопризнаковых объектов с помощью мультимножеств позволяет расширить круг рассматриваемых проблем, а также более простым способом находить решение известных задач. Примеры использования предложенных метрик в задачах многокритериального принятия решений по противоречивым данным приведены в [Петровский, 2003, 2004].

При решении практических задач с помощью кластерного анализа [Petrovsky, 2003] может оказаться полезным следующий прием. Вначале производится иерархическая кластеризация исходной совокупности объектов, и формируются несколько возможных разбиений объектов. Эти разбиения затем анализируются неиерархическими методами, и находится наилучшее или наиболее приемлемое из разбиений.

В ряде случаев возникает необходимость решить обратную задачу кластерного анализа, например, построить классификацию множества $G = \{x_1, \dots, x_h\}$ свойств объектов, используя совокупность объектов $A = \{A_1, \dots, A_n\}$. Или требуется одновременно расклассифицировать и объекты, и их свойства. В последнем случае применяются методы двойной классификации, и строятся две последовательности: группы объектов $\{X_1, \dots, X_R\}$ и группы свойств $\{G_1, \dots, G_L\}$. Содержательную интерпретацию полученных группировок можно дать, например, с помощью многомерного шкалирования [Айвазян и др., 1989]. При этом, однако, могут возникнуть ситуации, особенно если множество G свойств объектов состоит из элементов разного «качества», когда содержательная интерпретация отдельных кластеров свойств G_w окажется затруднительной.

Литература

1. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика. Классификация и снижение размерности. М.: Финансы и статистика, 1989.

2. Деза М. М., Лоран М. Геометрия разрезов и метрик / Пер. с англ. М.: МЦНМО, 2001.
3. Орлов А. И. Устойчивость в социально-экономических моделях. М.: Наука, 1979.
4. Петровский А. Б. Метрические пространства мультимножеств // Доклады Академии наук. 1995. Т. 344. № 2. С. 175–177.
5. Петровский А. Б. Комбинаторика мультимножеств // Доклады Академии наук. 2000. Т. 370. № 6. С. 750–753.
6. Петровский А. Б. Пространства множеств и мультимножеств. М.: УРСС, 2003.
7. Петровский А. Б. Многокритериальное принятие решений по противоречивым данным: подход теории мультимножеств // Информационные технологии и вычислительные системы. 2004. № 2. С. 56–66.
8. Anderberg M. R. Cluster Analysis for Applications. New York: Academic Press, 1973.
9. Miyamoto S. Cluster analysis as a tool of interpretation of complex systems // Working paper WP-87-41. IIASA. Laxenburg, Austria, 1987.
10. Petrovsky A. Cluster Analysis in Multiset Spaces // Information Systems Technology and its Applications. Gesellschaft für Informatik / Ed. M. Goldevsky, S. Liddle, H. Mayr. Bonn, 2003. P. 109–119.