

Методы групповой классификации многопризнаковых объектов (часть 1)¹

Аннотация. Рассматривается новый подход к групповой классификации объектов, которые описываются многими качественными и/или количественными признаками и могут присутствовать в нескольких версиях. Обсуждаются новые способы агрегирования многопризнаковых объектов и свойства построенных классов. Методы кластеризации и сортировки объектов основаны на теории метрических пространств мультимножеств.

Ключевые слова: групповая классификация, кластеризация, сортировка, многопризнаковые объекты, метрические пространства мультимножеств.

Введение

Объединение объектов (альтернатив, вариантов, действий, ситуаций, персон и т.п.) в классы (кластеров, категорий, групп) является одним из наиболее популярных средств для обнаружения, извлечения, формализации и фиксации знаний. Свойства объектов обычно задаются множеством характеристик или признаков, значения которых могут быть количественными (числовыми) и/или качественными (символьными и вербальными). Класс включает объекты, обладающие общими особенностями. Классы могут быть определены заранее или появляться и конструироваться в процессе классификации.

Проблемы классификации рассматриваются в принятии решений, анализе данных, распознавании образов и анализе изображений, искусственном интеллекте, биологии, социологии, психологии и других областях. Построенная классификация объектов позволяет выявлять различные взаимные отношения и исследовать возможные связи между объектами, основываясь на их свойствах. Результатами классификации служат новые понятия и правила их создания.

Для решения задач классификации разработаны и используются различные методы, которые можно условно разделить на такие группы: классификация без учителя или кластеризация, классификация с учителем, номинальная и порядковая классификации. В методах кластеризации объекты объединяются в группы (кластеры), исходя из степени их близости, которая формально устанавливается некоторым расстоянием между объектами в пространстве признаков. Число формируемых кластеров может быть произвольным или фиксированным. В методах классификации с учителем ищется общее правило для отнесения объекта к одному из заданных классов, которое строится на основании предварительно полученной информации о принадлежности некоторой части объектов к определенным классам. Методы порядковой и номинальной классификации отличаются наличием или отсутствием упорядоченности классов по какому-то свойству или качеству. В этих методах требуется найти значения признаков объектов или их сочетания, наиболее характерные для каждого класса.

¹ Работа поддержана программами фундаментальных исследований Президиума РАН «Интеллектуальные информационные технологии, математическое моделирование, системный анализ и автоматизация» и ОНИТ РАН «Информационные технологии и методы анализа сложных систем», Российским фондом фундаментальных исследований (проекты 08-01-00247, 08-07-13532, 09-07-00009, 09-07-12111).

Большинство из известных методов классификации оперирует с объектами, представленными многими количественными характеристиками [10-12, 15, 17, 19, 23-25]. В этих случаях каждому объекту, как правило, ставится в соответствие вектор, составленный из числовых значений признаков, или строка таблицы данных «Объекты-Признаки». Если объект описывается качественными признаками, то обычно такие символьные или вербальные переменные преобразуются тем или иным образом в числовые, например, используя лексико-графические шкалы или функции нечеткой принадлежности [16, 26, 29]. При этом, к сожалению, не всегда уделяется должное внимание допустимости и обоснованности подобных преобразований качественных данных в количественные. Существенно меньше методов классификации объектов, описываемых качественными признаками, где эти признаки не трансформируют в числовые [5, 13, 14, 18, 21, 22]. В таких случаях каждый объект представляют кортежем, состоящим из символьных или вербальных значений признаков.

Ситуация усложняется, когда один и тот же объект может существовать в нескольких версиях (экземплярах). Например, характеристики объекта были измерены в разных условиях или с помощью разных приборов, либо объект был независимо оценен несколькими экспертами по многим критериям. Тогда каждому объекту будет соответствовать не один вектор или кортеж, а группа векторов или кортежей, которая должна рассматриваться и обрабатываться как единое целое. При этом, очевидно, значения одноименных компонент разных векторов/кортежей могут различаться и даже быть противоречивыми. Ясно, что совокупность подобных многопризнаковых объектов может иметь весьма сложную структуру, которую трудно анализировать.

Обычно группа объектов, представленных несколькими векторами, заменяется одним вектором, который, например, имеет компоненты, полученные путем усреднения или взвешивания значений признаков всех членов группы, является центром группы или вектором, наиболее близким ко всем векторам в группе. Заметим, однако, что свойства всех объектов, входящих в группу, могут оказаться потерянными в результате такой замены. Для качественных переменных операции усреднения, взвешивания, смешивания и тому подобные преобразо-

вания данных являются математически некорректными и недопустимыми. Тем самым группу объектов, представленных несколькими кортежами, нельзя соединить в единственный кортеж. А значит, нужны новые способы агрегирования подобных объектов и работы с ними.

В данной работе, объекты, описываемые количественными и/или качественными признаками, представлены как мультимножества или множества с повторяющимися элементами. Обсуждаются новые методы работы с группами такого рода объектов. Предложены новые методы кластеризации и сортировки многопризнаковых объектов в метрических пространствах мультимножеств.

1. Мультимножества и метрические пространства мультимножеств

Мультимножество – достаточно известное понятие, которое используется в комбинаторной математике, компьютерных науках и других областях [4, 7, 11, 28]. Мультимножество A , порожденное обычным множеством $X=\{x_1, x_2, \dots\}$, все элементы которого различны, определяются как следующая совокупность групп элементов

$$A = \{k_A(x_1) \circ x_1, \dots, k_A(x_j) \circ x_j, \dots\} = \{k_A(x) \circ x | x \in X, k_A \in \mathbb{Z}_+\}. \quad (1)$$

Здесь $k_A: X \rightarrow \mathbb{Z}_+ = \{0, 1, 2, \dots\}$ называется функцией кратности мультимножества, определяющей число вхождения элемента $x_i \in X$ в мультимножество A , что обозначено символом $\circ$. Мультимножество обобщает понятие обычного множества.

Теоретическая модель мультимножества наиболее подходит для структуризации и анализа совокупности объектов, которые описываются многими качественными и/или количественными признаками, а также могут присутствовать в нескольких версиях. Приведем примеры подобных объектов, представляемых мультимножествами.

Пусть $A = \{A_1, \dots, A_n\}$ представляет собой коллекцию распознаваемых графических объектов: печатных или рукописных символов, линий, изображений, строк или страниц [1]. Множество $X = \{x_1, \dots, x_h\}$ является набором стандартных образцов, состоящим из целых символов или отдельных структурных элементов (фрагментов). В ходе распознавания каждый распознаваемый объект A_i сравнивается с базой образ-

цов и сопоставляется одному или нескольким образцам из стандартного набора с какой-то определенной точностью. Результат распознавания графического объекта A_i можно представить в виде $A_i = \{k_{A_i}(x_1) \circ x_1, \dots, k_{A_i}(x_h) \circ x_h\}$, где число $k_{A_i}(x_j)$ будет равно вычисленной оценке распознавания объекта A_i по отношению к стандартному образцу x_j .

Пусть теперь $A = \{A_1, \dots, A_n\}$ является массивом текстовых документов по некоторой проблематике. Эти документы могут быть, к примеру, статьями, книгами, проектами, патентами, отчетами, обзорами и тому подобным [13]. Множество лексических единиц (дескрипторов, ключевых слов, терминов и прочее) $X = \{x_1, \dots, x_h\}$ образует проблемно-ориентированный терминологический словарь или тезаурус, характеризующий особенности проблематики. Содержание текстового документа A_i можно описать совокупностью повторяющихся лексических единиц x_j и представить его в виде $A_i = \{k_{A_i}(x_1) \circ x_1, \dots, k_{A_i}(x_h) \circ x_h\}$, где $k_{A_i}(x_j)$ будет равно числу лексических единиц x_j , присутствующих в документе A_i . Заметим, что такое представление документа может соответствовать и полному тексту документа, и некоторому его краткому содержанию, индексированному лексическими единицами.

В рассмотренных выше случаях многопризнаковый объект (символ, документ) A_i был представлен как множество повторяющихся элементов (стандартных образцов, лексических единиц) x_j или как мультимножество A_i . Очевидно, что каждый графический объект A_i может многократно появляться внутри распознаваемого текста или изображения, и каждый документ A_i может существовать в нескольких версиях внутри массива текстов. Иначе говоря, могут иметься группы, которые объединяют разные версии (копии) одних и тех же объектов. И каждая такая группа версий многопризнакового объекта A_i , и совокупность объектов $A = \{A_1, \dots, A_n\}$ в целом также может быть представлена как мультимножество над тем же самым множеством X .

Мультимножество A считается конечным, если все $k_A(x)$ конечны. Мультимножество A превращается в обычное множество A при $k_A(x) = \chi_A(x)$, где $\chi_A(x) = 1$, если $x \in A$, и $\chi_A(x) = 0$, если $x \notin A$. Мощность конечного мультимножества $|A| = \sum_x k_A(x)$ определяется как общее число экземпляров всех его элементов; размерность

мультимножества $|A| = \sum_x \chi_A(x)$ – как общее число различных элементов. Говорят, что мультимножества A и B равны ($A=B$), если $k_A(x) = k_B(x)$, и что мультимножество B включено в мультимножество A или является его подмультимножеством ($B \subseteq A$), если $k_B(x) \leq k_A(x)$, $\forall x \in X$.

Определяются следующие операции над мультимножествами [7]: объединение $A \cup B$, $k_{A \cup B}(x) = \max[k_A(x), k_B(x)]$; пересечение $A \cap B$, $k_{A \cap B}(x) = \min[k_A(x), k_B(x)]$; сложение $A+B$, $k_{A+B}(x) = k_A(x) + k_B(x)$; вычитание $A-B$, $k_{A-B}(x) = k_A(x) - k_{A \cap B}(x)$; симметрическая разность $A \Delta B$, $k_{A \Delta B}(x) = |k_A(x) - k_B(x)|$; умножение на скаляр (репродукция) $b \bullet A$, $k_{b \bullet A}(x) = b \cdot k_A(x)$, $b \in \mathbb{N}$; умножение $A \bullet B$, $k_{A \bullet B}(x) = k_A(x) \cdot k_B(x)$; возведение в арифметическую степень A^n ; прямое произведение $A \times B$, $k_{A \times B}(x_i, y_j) = k_A(x_i) \cdot k_B(y_j)$, $x_i \in A$, $y_j \in B$; прямая степень $(\times A)^n$.

Многие из характеристик операций над мультимножествами аналогичны характеристикам операций над обычными множествами. В их числе идемпотентность, инволюция (двойное отрицание), идентичность, коммутативность, ассоциативность и дистрибутивность ряда операций. Как и в случае множеств, не все операции над мультимножествами взаимно коммутативны, ассоциативны и дистрибутивны. Кроме того, в теории множеств не определяются операции сложения, умножения на скаляр, умножения, возведения в арифметическую степень. При переходе от мультимножества к множеству операции умножения и возведения в арифметическую степень превращаются в пересечение, а операции сложения и умножения на скаляр становятся неосуществимыми.

Совокупность $A = \{A_1, \dots, A_n\}$ многопризнаковых объектов можно рассматривать как точки в метрическом пространстве мультимножеств (A, d) . Различные типы метрических пространств задаются следующими псевдометриками:

$$\begin{aligned} d_{1c}(A, B) &= [m(A \Delta B)]^{1/c}; \\ d_{2c}(A, B) &= [m(A \Delta B)/m(Z)]^{1/c}; \\ d_{3c}(A, B) &= [m(A \Delta B)/m(A \cup B)]^{1/c}, \end{aligned} \quad (2)$$

где c – целое число, m – мера мультимножества, а мультимножество Z называется максимальным мультимножеством с функцией кратности $k_Z(x) = \max_{A \in A} k_A(x)$. Мера мультимножества m есть действительная неотрицательная функция,

заданная на алгебре мультимножеств $L(\mathbf{Z})$. Мету мультимножества можно ввести различными способами, например, как мощность мультимножества $m(A)=|A|=\sum_j k_A(x_j)$ или как линейную комбинацию функций кратности $m(A)=\sum_j w_j k_A(x_j)$, $w_j>0$ – важность элемента x_j , $w_i>0$. Функции $d_{2c}(A,B)$ и $d_{3c}(A,B)$ удовлетворяют условию нормировки $0\leq d(A,B)\leq 1$. Функция $d_{3c}(A,B)$ не определена для $A=B=\emptyset$, поэтому по определению принимается $d_{3c}(\emptyset,\emptyset)=0$. Для некоторого фиксированного c метрики d_{1c} и d_{2c} являются непрерывными и однородно непрерывными функциями, а метрика d_{3c} – частично непрерывной функцией почти всюду на метрическом пространстве.

Введенные пространства представляют собой новые виды метрических пространств измеримых мультимножеств, которые отличаются от известных метрических пространств измеримых множеств [2]. Расстояние $d_{1c}(A,B)$ характеризует различие между свойствами двух объектов и является аналогом расстояния Хемминга, традиционного для многих приложений. Расстояние $d_{2c}(A,B)$ задает различие между свойствами двух объектов, отнесенное к свойствам «максимального» объекта, и может быть названо полностью усредненным расстоянием. Расстояние $d_{3c}(A,B)$ отражает различие между свойствами двух объектов, отнесенное к общим свойствам только этих объектов, и может быть названо локально усредненным расстоянием. В случае обычных множеств при $c=1$, $d_{11}(A,B)=m(A\Delta B)$ называется расстоянием Фреше-Никодима-Ароншайна, а $d_{31}(A,B)=m(A\Delta B)/m(A\cup B)$ – расстоянием Штейнхауса [2]. Различные свойства мультимножеств и метрических пространств мультимножеств рассмотрены в [7].

2. Представление и агрегирование многопризнаковых объектов

Задача групповой классификации многопризнаковых объектов в общем случае формулируется следующим образом. Имеется совокупность объектов $A=\{A_1,\dots,A_n\}$, которые описываются m признаками $Q_1,\dots,Q_m$. Каждый признак Q_s имеет свою собственную шкалу $X_s=\{x_s^1,\dots,x_s^{h_s}\}$, $s=1,\dots,m$, градации которой суть качественные или количественные оценки. Каждый из объектов A_i присутствует в k экземплярах, которые, как

правило, отличаются значениями признаков. Например, характеристики объекта были измерены в разных условиях или разными способами, либо объект был независимо оценен несколькими экспертами по многим критериям. Требуется разделить объекты на несколько групп (классов) $C_1,\dots,C_g$, различающихся по своим свойствам, охарактеризовать и интерпретировать свойства этих групп объектов. Число g групп объектов может быть произвольным или заранее заданным, а сами классы упорядоченными или неупорядоченными.

Рассмотрим сначала несколько иллюстративных примеров, показывающих, как можно представить многопризнаковые объекты.

Пусть совокупность A состоит из 10 объектов $A_1,\dots,A_{10}$, имеющих 8 признаков $Q_1,\dots,Q_8$, каждый из которых может принимать одну из оценок по пятибалльной шкале $X=\{1, 2, 3, 4, 5\}$. Допустим, что объектами $A_1,\dots,A_{10}$ являются школьники, а признаками объектов служат их годовые оценки по 8 предметам: Q_1 . Математика, Q_2 . Физика, Q_3 . Химия, Q_4 . Биология, Q_5 . География, Q_6 . История, Q_7 . Литература, Q_8 . Иностранный язык. Градации шкалы оценок X означают следующее: 1 – очень плохо, 2 – плохо, 3 – удовлетворительно, 4 – хорошо, 5 – отлично. Или же в качестве объектов $A_1,\dots,A_{10}$ выступают ответы на вопросы некоторого опросника, с помощью которого изучается мнение группы людей по некоторой проблеме. Признаками объектов будут оценки, данные 8 респондентами $Q_1,\dots,Q_8$, которые кодируются следующим образом: 1 – полностью не согласен, 2 – не согласен, 3 – нейтрален, 4 – согласен, 5 – полностью согласен.

Сопоставим каждому многопризнаковому объекту A_i вектор или кортеж $q_i=(q_{i1}^{e_1},\dots,q_{im}^{e_m})$. Совокупность всех объектов $A=\{A_1,\dots,A_n\}$ и их признаков можно представить таблицей «Объекты-Признаки» $Q=||q_{is}||_{n\times m}$. Строки Q матрицы соответствуют объектам, колонки – признакам, а элементами q_{is} являются значения компонент $q_{is}^{e_s}$ соответствующих векторов/кортежей. Для рассмотренных выше примеров таблица данных Q имеет вид, приведенный в Табл. 1. Таблица данных частично заимствована из работы [15] и представляет собой фрагмент опросника для оценки студентами курса лекций по анализу данных. Так, например, объекту A_1 соответствует вектор/кортеж оценок $q_1=(4, 5, 4, 5, 4, 5, 4, 5)$.

Табл. 1. Таблица данных Q

$A \backslash Q$	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8
A_1	4	5	4	5	4	5	4	5
A_2	4	1	2	1	3	2	2	2
A_3	1	1	3	1	4	1	1	4
A_4	5	3	2	4	4	5	4	5
A_5	4	4	4	4	4	5	4	4
A_6	5	5	4	4	4	5	5	4
A_7	4	1	2	3	3	3	1	2
A_8	4	5	4	2	3	4	5	3
A_9	3	2	3	1	3	3	2	2
A_{10}	5	5	4	5	3	5	5	4

Табл. 2. Таблица данных L

$A \backslash X$	x_1	x_2	x_3	x_4	x_5
A_1	0	0	0	4	4
A_2	2	4	1	1	0
A_3	5	0	1	2	0
A_4	0	1	1	3	3
A_5	0	0	0	7	1
A_6	0	0	0	4	4
A_7	2	2	3	1	0
A_8	0	1	2	3	2
A_9	1	3	4	0	0
A_{10}	0	0	1	2	5

Укажем другой возможный способ задания многопризнаковых объектов с помощью мультимножеств. Выберем множество оценок $X = \{x_1, \dots, x_h\}$ в качестве порождающего множества и поставим в соответствие каждому многопризнаковому объекту $A_i, i=1, \dots, n$ мультимножество $A_i = \{k_{A_i}(x_1) \circ x_1, \dots, k_{A_i}(x_h) \circ x_h\}$ над множеством X . Здесь значение функции кратности $k_{A_i}(x_j)$ показывает, сколько раз оценка x_j присутствует в описании объекта A_i . Совокупность всех объектов $A = \{A_1, \dots, A_n\}$ и их признаков можно представить в таком случае другой таблицей «Объекты-Признаки» $L = \|k_{ij}\|_{n \times h}$, элементами k_{ij} которой являются значения кратности $k_{A_i}(x_j)$ элементов мультимножеств, соответствующих объектам A_i . Таблица данных L для рассмотренных выше объектов $A_1, \dots, A_{10}$ приведена в Табл. 2. Так, например, объекту A_1 соответствует мультимножество оценок $A_1 = \{0 \circ x_1, 0 \circ x_2, 0 \circ x_3, 4 \circ x_4, 4 \circ x_5\}$. Такая форма записи говорит, что объект A_1 имеет 4 оценки x_4 , означающей «хорошо» или «согласен», 4 оценки x_5 , означающей «отлично» или «полностью согласен», а других оценок у объекта A_1 нет. В ряде случаев элементы мультимножества удобнее расположить в обратном порядке – от лучших оценок к худшим – и записать мультимножество как $A_1 = \{4 \circ x_5, 4 \circ x_4, 0 \circ x_3, 0 \circ x_2, 0 \circ x_1\}$. Заметим, однако, что, строго говоря, элементы мультимножества принято считать неупорядоченными.

Пусть теперь каждый объект присутствует в нескольких версиях. Например, для каждого из школьников $A_1, \dots, A_{10}$ заданы два набора семестровых (полугодовых) оценок по тем же самым 8 дисциплинам $Q_1, \dots, Q_8$. Или 8 респондентов $Q_1, \dots, Q_8$ опрашивались дважды, отвечая на одни и те же вопросы $A_1, \dots, A_{10}$. Это означает, что каж-

дый объект будет описываться уже не одним, а двумя векторами/кортежами признаков, или что имеется две версии одного и того же объекта. Например, объект A_1 характеризуется группой векторов/кортежей, состоящей из $q_1^{(1)} = (4, 5, 4, 5, 4, 5, 4, 5)$ и $q_1^{(2)} = (5, 5, 5, 5, 4, 4, 4, 5)$.

В общем случае один и тот же многопризнаковый объект $A_i, i=1, \dots, n$ будет представлен группой векторов/кортежей $\{q_i^{(1)}, \dots, q_i^{(k)}\}$ в декартовом m -мерном пространстве признаков $Q = Q_1 \times \dots \times Q_m$. Здесь $q_i^{(f)} = (q_{i1}^{e_1(f)}, \dots, q_{im}^{e_m(f)})$, $f=1, \dots, k$ задает одну из версий $A_i^{(f)}$ объекта A_i , а компонента $q_{is}^{e_s(f)}$, $s=1, \dots, m$ вектора/кортежа $q_i^{(f)}$ – одна из градаций шкалы $x_j \in X, j=1, \dots, h$.

Заметим, что группа векторов/кортежей, соответствующих объекту A_i , должна рассматриваться как единое целое, несмотря на возможную несравнимость отдельных векторов $q_i^{(f)}$ и/или несогласованность их компонент $q_{is}^{e_s(f)}$. Компоненты векторов являются числовыми переменными. Поэтому группа векторов часто заменяется одним вектором, представляющим всю группу, компоненты которого определяются некими дополнительными соображениями. Например, это может быть вектор, имеющий в качестве компонент усредненные значения соответствующих компонент всех членов группы. Тогда объект A_1 будет представлен вектором «средних» оценок: $q_1 = (4.5, 5.0, 4.5, 5.0, 4.0, 4.5, 4.0, 5.0)$. Однако такой вектор не соответствует никакой конкретной точке m -мерного признакового пространства $Q = Q_1 \times \dots \times Q_m$, образованного дискретной шкалой числовых оценок $X = \{1, 2, 3, 4, 5\}$, в которой отсутствуют нецелые числа. Чтобы можно было оперировать с такими векторами, нужно либо расширить шкалу оценок X , введя промежуточные число-

вые градации, например, $X=\{1.00, 1.25, 1.50, 1.75, 2.00, \dots, 4.00, 4.25, 4.50, 4.75, 5.00\}$, либо считать шкалу оценок X непрерывной. И то, и другое, строго говоря, изменяет первоначальную формулировку задачи.

Когда объекты представлены кортежами с символьными или вербальными компонентами, то есть используется шкала $X=\{1, 2, 3, 4, 5\}$, где, например, 1 – очень плохо, 2 – плохо, 3 – удовлетворительно, 4 – хорошо, 5 – отлично, группу подобных кортежей уже нельзя заменить одним кортежем с «усредненными» оценками, поскольку такая операция математически некорректна. Если группа состоит всего из двух векторов/кортежей, то в таком случае также нельзя ввести единственный объект в качестве типичного представителя группы, например, объект, наиболее близкий ко всем объектам группы. Это справедливо как для числовой, так и символьной шкалы оценок $X=\{1, 2, 3, 4, 5\}$.

Эти трудности можно легко преодолеть, если воспользоваться формализмом теории мультимножеств, который позволяет одновременно учесть все комбинации значений признаков, а также наличие различных версий объектов. Поставим в соответствие каждой версии $A_i^{(f)}$, $i=1, \dots, n$, $f=1, \dots, k$ многопризнакового объекта A_i мультимножество $A_i^{(f)}=\{k_{Ai}^{(f)}(x_1) \circ x_1, \dots, k_{Ai}^{(f)}(x_h) \circ x_h\}$ над множеством оценок $X=\{x_1, \dots, x_h\}$, а каждому объекту A_i – мультимножество $A_i=\{k_{Ai}(x_1) \circ x_1, \dots, k_{Ai}(x_h) \circ x_h\}$ над тем же самым множеством X . Значения функции кратности мультимножества A_i вычисляются, например, по правилу: $k_{Ai}(x_j)=\sum_f k_{Ai}^{(f)}(x_j)$. Так, две версии объекта A_1 представляются двумя мультимножествами оценок $A_1^{(1)}=\{0 \circ x_1, 0 \circ x_2, 0 \circ x_3, 4 \circ x_4, 4 \circ x_5\}$ и $A_1^{(2)}=\{0 \circ x_1, 0 \circ x_2, 0 \circ x_3, 3 \circ x_4, 5 \circ x_5\}$, а сам объект A_1 – мультимножеством $A_1=\{0 \circ x_1, 0 \circ x_2, 0 \circ x_3, 7 \circ x_4, 9 \circ x_5\}$.

Если в рассматриваемой задаче по каким-то причинам значение имеет не общее число разных оценок по всем признакам Q_1-Q_8 , а число разных оценок по каждому из признаков Q_s , например, отметок по каждой дисциплине или ответов каждого респондента, и эти оценки нужно различать, то мультимножество оценок можно определить иначе. Введем гипершкалу признаков, которая объединяет все градации отдельных шкал признаков: $X=X_1 \cup \dots \cup X_m = \{x_1^1, \dots, x_1^{h_1}; \dots; x_m^1, \dots, x_m^{h_m}\}$. Тогда каждому объекту A_i или его версии $A_i^{(f)}$ можно сопоставить мультимножество

$$A_i=\{k_{Ai}(x_1^1) \circ x_1^1, \dots, k_{Ai}(x_1^{h_1}) \circ x_1^{h_1}; \dots; k_{Ai}(x_m^1) \circ x_m^1, \dots, k_{Ai}(x_m^{h_m}) \circ x_m^{h_m}\} \quad (3)$$

над множеством оценок $X=\{x_1^1, \dots, x_1^{h_1}; \dots; x_m^1, \dots, x_m^{h_m}\}$. Здесь функция кратности $k_{Ai}(x_s^{e_s})$ показывает число значений $x_s^{e_s} \in X_s$, $e_s=1, \dots, h_s$, $s=1, \dots, m$, каждого из признаков Q_s , присутствующих в описании объекта A_i . К примеру, объекту A_1 будет теперь соответствовать следующее мультимножество отдельных оценок по всем признакам:

$$A_1=\{0 \circ x_1^1, 0 \circ x_1^2, 0 \circ x_1^3, 1 \circ x_1^4, 1 \circ x_1^5; \\ 0 \circ x_2^1, 0 \circ x_2^2, 0 \circ x_2^3, 0 \circ x_2^4, 2 \circ x_2^5; \\ 0 \circ x_3^1, 0 \circ x_3^2, 0 \circ x_3^3, 1 \circ x_3^4, 1 \circ x_3^5; \\ 0 \circ x_4^1, 0 \circ x_4^2, 0 \circ x_4^3, 0 \circ x_4^4, 2 \circ x_4^5; \\ 0 \circ x_5^1, 0 \circ x_5^2, 0 \circ x_5^3, 2 \circ x_5^4, 0 \circ x_5^5; \\ 0 \circ x_6^1, 0 \circ x_6^2, 0 \circ x_6^3, 1 \circ x_6^4, 1 \circ x_6^5; \\ 0 \circ x_7^1, 0 \circ x_7^2, 0 \circ x_7^3, 2 \circ x_7^4, 0 \circ x_7^5; \\ 0 \circ x_8^1, 0 \circ x_8^2, 0 \circ x_8^3, 0 \circ x_8^4, 2 \circ x_8^5\}.$$

Несмотря на кажущуюся громоздкость записи, такое представление оказывается крайне удобным при сравнении мультимножеств и оперирования с ними с вычислительной точки зрения, поскольку можно выполнять операции одновременно по всем элементам мультимножеств. Заметим, что выражение (3) для любого мультимножества легко записать и в обычном виде (1) $A_i=\{k_{Ai}(x_j) \circ x_j | x_j \in X=\{x_1, \dots, x_h\}\}$, если во множестве $X=\{x_1^1, \dots, x_1^{h_1}; \dots; x_m^1, \dots, x_m^{h_m}\}$ сделать замену переменных: $x_1^1=x_1, \dots, x_1^{h_1}=x_{h_1}$, $x_2^1=x_{h_1+1}, \dots, x_2^{h_2}=x_{h_1+h_2}, \dots, x_m^{h_m}=x_h$, $h=h_1+\dots+h_m$.

Совокупность всех объектов $A=\{A_1, \dots, A_n\}$, каждый из которых имеется в нескольких версиях, и их признаков снова можно представить таблицей «Объекты-Признаки», но имеющей большую размерность. Таблица «Объекты-Признаки» $Q^{\#}=\|q_{is}\|_{kn \times m}$ для объектов $A_1, \dots, A_{10}$ из рассмотренных ранее примеров, которые представлены векторами/кортежами, приведена в Табл.3. Элементами q_{is} таблицы $Q^{\#}$ будут компоненты $q_{is}^{e_s(f)}$ векторов/кортежей, соответствующих версиям объектов A_i . Таблица «Объекты-Признаки» $L^{\#}=\|k_{ij}\|_{kn \times h}$ для объектов $A_1, \dots, A_{10}$, представленных мультимножествами вида (1), приведена в Табл. 4. Элементами k_{ij} таблицы $L^{\#}$ будут значения кратности $k_{Ai}^{(f)}(x_j)$ элементов мультимножеств, соответствующих версиям объектов A_i . Таблица «Объекты-Признаки» $K=\|k_{ij}\|_{n \times h}$, $h=h_1+\dots+h_m$ для объектов, представленных мультимножествами вида (3), дана в Табл. 5. Элементами k_{ij} таблицы K будут значе-

Таблица 3. Таблица данных $Q^{\#}$

$A \backslash Q$	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8
A_1	4	5	4	5	4	5	4	5
	5	5	5	5	4	4	4	5
A_2	4	1	2	1	3	2	2	2
	3	2	1	1	4	3	3	2
A_3	1	1	3	1	4	1	1	4
	1	2	3	1	5	2	1	3
A_4	5	3	2	4	4	5	4	5
	4	4	3	5	4	5	3	4
A_5	4	4	4	4	4	5	4	4
	5	5	3	4	4	4	5	4
A_6	5	5	4	4	4	5	5	4
	4	5	4	4	4	4	5	5
A_7	4	1	2	3	3	3	1	2
	3	2	1	4	2	4	2	3
A_8	4	5	4	2	3	4	5	3
	5	4	5	3	4	5	4	4
A_9	3	2	3	1	3	3	2	2
	4	3	2	2	2	3	3	2
A_{10}	5	5	4	5	3	5	5	4
	3	4	3	4	2	4	2	4

Таблица 4. Таблица данных $L^{\#}$

$A \backslash X$	x_1	x_2	x_3	x_4	x_5
A_1	0	0	0	4	4
	0	0	0	3	5
A_2	2	4	1	1	0
	2	2	3	1	0
A_3	5	0	1	2	0
	3	2	2	0	1
A_4	0	1	1	3	3
	0	0	2	4	2
A_5	0	0	0	7	1
	0	0	1	4	3
A_6	0	0	0	4	4
	0	0	0	5	3
A_7	2	2	3	1	0
	1	3	2	2	0
A_8	0	1	2	3	2
	0	0	1	4	3
A_9	1	3	4	0	0
	0	4	3	1	0
A_{10}	0	0	1	2	5
	0	2	2	4	0

Таблица 5. Таблица данных K

$A \backslash X$	$x_1^1 x_1^2 x_1^3 x_1^4 x_1^5$	$x_2^1 x_2^2 x_2^3 x_2^4 x_2^5$	$x_3^1 x_3^2 x_3^3 x_3^4 x_3^5$	$x_4^1 x_4^2 x_4^3 x_4^4 x_4^5$	$x_5^1 x_5^2 x_5^3 x_5^4 x_5^5$	$x_6^1 x_6^2 x_6^3 x_6^4 x_6^5$	$x_7^1 x_7^2 x_7^3 x_7^4 x_7^5$	$x_8^1 x_8^2 x_8^3 x_8^4 x_8^5$
A_1	0 0 0 1 1	0 0 0 0 2	0 0 0 1 1	0 0 0 0 2	0 0 0 2 0	0 0 0 1 1	0 0 0 2 0	0 0 0 0 2
A_2	0 0 1 1 0	1 1 0 0 0	1 1 0 0 0	2 0 0 0 0	0 0 1 1 0	0 1 1 0 0	0 1 1 0 0	0 2 0 0 0
A_3	2 0 0 0 0	1 1 0 0 0	0 0 2 0 0	2 0 0 0 0	0 0 0 1 1	1 1 0 0 0	2 0 0 0 0	0 0 1 1 0
A_4	0 0 0 1 1	0 0 1 1 0	0 1 1 0 0	0 0 0 1 1	0 0 0 2 0	0 0 0 0 2	0 0 1 1 0	0 0 0 1 1
A_5	0 0 0 1 1	0 0 0 1 1	0 0 1 1 0	0 0 0 2 0	0 0 0 2 0	0 0 0 1 1	0 0 0 1 1	0 0 0 2 0
A_6	0 0 0 1 1	0 0 0 0 2	0 0 0 2 0	0 0 0 2 0	0 0 0 2 0	0 0 0 1 1	0 0 0 0 2	0 0 0 1 1
A_7	0 0 1 1 0	1 1 0 0 0	1 1 0 0 0	0 0 1 1 0	0 1 1 0 0	0 0 1 1 0	1 1 0 0 0	0 1 1 0 0
A_8	0 0 0 1 1	0 0 0 1 1	0 0 0 1 1	0 1 1 0 0	0 0 1 1 0	0 0 0 1 1	0 0 0 1 1	0 0 1 1 0
A_9	0 0 1 1 0	0 1 1 0 0	0 1 1 0 0	1 1 0 0 0	0 1 1 0 0	0 0 2 0 0	0 1 1 0 0	0 0 2 0 0
A_{10}	0 0 1 0 1	0 0 0 1 1	0 0 1 1 0	0 0 0 1 1	0 1 1 0 0	0 0 0 1 1	0 1 0 0 1	0 0 0 2 0

ния кратности $k_{A_i}(x_j)$ элементов мультимножеств, соответствующих объектам A_i .

Группировка объектов является одним из полезных приемов для структуризации совокупности объектов. Классифицируя объекты, мы относим каждый объект к некоторому классу и используем информацию о принадлежности объектов к классам для выработки и коррекции правил классификации. В самом общем смысле правила классификации представляют собой совокупность требований, которые состоят из логических утверждений следующего вида:

$$\text{IF } \langle \text{условия} \rangle, \text{ THEN } \langle \text{решение} \rangle. \quad (4)$$

Здесь терм $\langle \text{условия} \rangle$ определяет требования, которым должны удовлетворять выбираемые объекты. К примеру, это могут быть имена объектов; значения или комбинации значений

признаков, описывающих объекты; ограничения на значения признаков; отношения между объектами; правила сравнения объектов друг с другом или с некоторыми выделенными элементами классов, и тому подобное. Объекты сравниваются по сходству или различию их свойств, которые обычно формализуются с помощью специально задаваемых мер близости. Для выделения определенных элементов классов, например, типичного представителя или центра класса вводят некие требования. Терм $\langle \text{решение} \rangle$ обозначает имя формируемого класса и/или принадлежность объекта к заранее определенному классу при выполнении требуемых условий.

Разнообразие операций над мультимножествами позволяет применять различные способы агрегирования многопризнаковых объектов

в классы. Например, класс C_i объектов A_i можно образовать как сумму $Y_i = \sum_i A_i$, $k_{Y_i}(x_j) = \sum_i k_{A_i}(x_j)$, объединение $Y_i = \bigcup_i A_i$, $k_{Y_i}(x_j) = \max_i k_{A_i}(x_j)$ или пересечение $Y_i = \bigcap_i A_i$, $k_{Y_i}(x_j) = \min_i k_{A_i}(x_j)$ мультимножеств, представляющих рассматриваемые объекты. Класс объектов C_i можно также сформировать как линейную комбинацию соответствующих мультимножеств $Y_i = \sum_i b_i \cdot A_i$, $Y_i = \bigcup_i b_i \cdot A_i$ или $Y_i = \bigcap_i b_i \cdot A_i$. Когда класс C_i строится с помощью сложения мультимножеств, агрегируются все свойства (все значения всех признаков) всех членов группы мультимножеств Y_i . В случае объединения или пересечения мультимножеств происходит усиление самых лучших свойств (максимальных значений всех признаков) или соответственно самых худших свойств (минимальных значений всех признаков), имеющих у индивидуальных членов группы мультимножеств Y_i . К примеру, мультимножество A_i , представляющее объект A_i , формировалось путем сложения мультимножеств $A_i^{(j)}$, представляющих версии этого объекта.

3. Методы групповой кластеризации многопризнаковых объектов

В кластерном анализе совокупность объектов $A = \{A_1, \dots, A_n\}$ распределяется по нескольким группам (кластерам), исходя из понятия близости объектов [10, 15, 19]. Для того чтобы сгенерировать группы объектов, в методах кластерного анализа обычно используются следующие подходы: минимизировать различие (максимизировать сходство) между объектами внутри группы; максимизировать различие (минимизировать сходство) между группами объектов.

Представим объект, описываемый многими качественными (символьными или вербальными) признаками как мультимножество вида (1) или (3). Принимая во внимание формулу (2) при $c=1$ и $m(A) = \sum_j w_j k_A(x_j)$, где $w_j > 0$ – важность признака Q_j , различие между группами объектов можно задать как одно из следующих расстояний в пространстве мультимножеств (A, d) :

$$\begin{aligned} d_{11}(Y_p, Y_q) &= D_{pq}; \quad d_{21}(Y_p, Y_q) = D_{pq}/W; \\ d_{31}(Y_p, Y_q) &= D_{pq}/M_{pq} \end{aligned} \quad (5)$$

Сходство между группами объектов можно выразить одним из следующих индексов:

$$\begin{aligned} s_1(Y_p, Y_q) &= 1 - (D_{pq}/W); \quad s_2(Y_p, Y_q) = L_{pq}/W; \\ s_3(Y_p, Y_q) &= L_{pq}/M_{pq} \end{aligned} \quad (6)$$

Здесь $D_{pq} = \sum_j w_j |k_{Y_p}(x_j) - k_{Y_q}(x_j)|$; $L_{pq} = \sum_j w_j \min[k_{Y_p}(x_j), k_{Y_q}(x_j)]$; $M_{pq} = \sum_j w_j \max[k_{Y_p}(x_j), k_{Y_q}(x_j)]$; $W = \sum_j w_j \sup k_{Y_i}(x_j)$. Значения функций кратности $k_{Y_p}(x_j)$, $k_{Y_q}(x_j)$ зависят от способа объединения объектов в группы. Выражения s_1 , s_2 , s_3 (6) обобщают на случай мультимножеств широко известные неметрические индексы сходства объектов, такие как коэффициент простой согласованности, мера сходства Рассела-Рао, коэффициент Жаккара или мера Роджерса-Танимото соответственно [10, 19].

Рассмотрим основные идеи кластерного анализа для объектов, представленных мультимножествами [8]. Предположим, для простоты, что выражения (5) и (6) определяют одновременно различие и сходство в пространстве признаков между объектами внутри каждой группы, и между некоторым объектом A_i и группой объектов C_i , и между группами объектов.

Общая схема процедуры иерархической кластеризации, в которой число генерируемых кластеров заранее неизвестно, состоит из следующих шагов.

Шаг 1. Положить $g=n$, g – число кластеров C_i , n – число объектов A_i . Тогда каждый кластер C_i состоит из единственного объекта A_i , и $Y_i = A_i$ для всех $i=1, \dots, g$.

Шаг 2. Вычислить расстояния $d(Y_p, Y_q)$ между парами мультимножеств, представляющих кластеры C_p и C_q для всех $1 \leq p, q \leq g$, $p \neq q$, используя одну из метрик (5).

Шаг 3. Найти пару наиболее близких кластеров C_u и C_v , удовлетворяющих условию

$$d(Y_u, Y_v) = \min_{p,q} d(Y_p, Y_q), \quad (7)$$

и сформировать новый кластер C_r , который представляется суммой $Y_r = Y_u + Y_v$, объединением $Y_r = Y_u \cup Y_v$, пересечением $Y_r = Y_u \cap Y_v$, соответствующих мультимножеств или линейной комбинацией одной из этих операций.

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

Шаг 5. Вычислить расстояния $d(Y_p, Y_r)$ между парами новых мультимножеств, представляющих кластеры C_p и C_r для всех $1 \leq p, r \leq g$, $p \neq r$. Вернуться к шагу 3.

Иерархический алгоритм строит дерево или дендрограмму путем последовательного агрегирования объектов в группы. Новые объекты или кластеры C_p и C_q появляются при движении от корня дерева по его ветвям, переходя на каждом шаге в один из ближайших кластеров C_r . Процесс иерархической кластеризации объектов заканчивается, когда все объекты будут объединены в несколько классов или в один единственный класс в зависимости от стоящей задачи. Процедура может быть также прервана на некотором шаге, когда будет выполнено некоторое правило, например, когда величина индекса различия между объектами превысит некоторый пороговый уровень.

Заметим, что на 3 шаге алгоритма может появляться много пар наиболее близких кластеров C_u, C_v , эквивалентных по условию минимальности расстояния $d(Y_p, Y_q)$ в признаковом пространстве. Тем самым существуют разные точки ветвления алгоритма (способы агрегирования мультимножеств) и, значит, можно построить различные итоговые деревья объектов. Специальные исследования показали, что наименьшее число финальных кластеров появляется в результате сложения, а наименьшее число финальных кластеров – в результате пересечения мультимножеств. Использование расстояния 3 ведет к меньшему числу точек ветвления алгоритма по сравнению с расстояниями 1 и 2, применения которых дают схожие результаты.

Чтобы уменьшить число возможных вариантов соединяемых кластеров, можно использовать дополнительные критерии близости объектов, например, критерий компактности кластера, вместо единственного критерия (5). В этом случае модифицированный алгоритм иерархической кластеризации объектов выглядит следующим образом.

Шаг 3.1. Найти все эквивалентные пары наиболее близких кластеров C_u, C_v , представленных мультимножествами Y_u, Y_v , которые соответствуют условию (7), и сформировать новые кластеры C_{rl} ($l=1, \dots, t_r$) для всех эквивалентных пар кластеров с помощью одной из операций – суммирования, объединения или пересечения мультимножеств. t_r – число эквивалентных пар кластеров, находящихся на одном и том же расстоянии $d(Y_u, Y_v)$.

Шаг 3.2. Найти кластер C_r^* , представленный мультимножеством Y_r^* , которое минимизирует

компактность кластера, выражаемую функционалом

$$f(Y_r^*) = \min_l \sum_{i,p \in C_{rl}} d(A_i, A_p) / N_{rl}, \quad (8)$$

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

Эксперименты показали, что добавление к критерию (7) дополнительных критериев близости объектов позволяет существенно улучшить результаты, получаемые при иерархической кластеризации для всех способов образования кластеров.

Рассмотрим в качестве иллюстративного примера задачу номинальной классификации без учителя. Требуется разделить на два класса совокупность объектов $A = \{A_1, \dots, A_{10}\}$, описываемых признаками $Q_1, \dots, Q_8$, каждый из которых имеет одну и ту же пятибалльную шкалу $X = \{x_1, x_2, x_3, x_4, x_5\}$ с качественными (символьными или вербальными) оценками. Нужно также дать интерпретацию полученным классам.

Представим эти объекты как мультимножества $A_i = \{k_{A_i}(x_j) \circ x_j | x_j \in X\}$. Их описания даны в таблице данных «Объекты-Признаки» L (Табл. 2). Для классификации таких многопризнаковых объектов воспользуемся упрощенным алгоритмом иерархической кластеризации, который включает следующие шаги.

Представим эти объекты как мультимножества вида (1) $A_i = \{k_{A_i}(x_j) \circ x_j | x_j \in X\}$. Их описания даны в таблице данных L «Объекты-Признаки» (Табл. 2). Для классификации таких многопризнаковых объектов воспользуемся упрощенным алгоритмом иерархической кластеризации, который включает следующие шаги.

1⁰. Рассматриваем каждый объект как самостоятельный кластер. Объекту/кластеру соответствуют мультимножества $Y_i = A_i$, $i=1, \dots, 10$.

2⁰. Выберем в качестве меры различия между объектами/кластерами одну метрику (3) вида $d_{11}(Y_p, Y_q) = D_{pq} = \sum_j |k_{Y_p}(x_j) - k_{Y_q}(x_j)|$, $j=1, \dots, 5$, считая все признаки x_j одинаково важными ($w_j=1$). Вычислим расстояния между всеми парами объектов/кластеров.

3⁰. Самыми близкими будут объекты A_1 и A_6 , расстояние между которыми в пространстве мультимножеств (A, d_{11}) равно $d_{11}(A_1, A_6) = 0$, так как мультимножества A_1 и A_6 совпадают. Объединяем объекты A_1 и A_6 в кластер C_{11} , которому будет соответствовать мультимножество Y_{11} , сформированное с помощью операции сложения мультимножеств:

$$Y_{11} = A_1 + A_6 = \{0^\circ x_1, 0^\circ x_2, 0^\circ x_3, 4^\circ x_4, 4^\circ x_5\} + \\ + \{0^\circ x_1, 0^\circ x_2, 0^\circ x_3, 4^\circ x_4, 4^\circ x_5\} = \\ = \{0^\circ x_1, 0^\circ x_2, 0^\circ x_3, 8^\circ x_4, 8^\circ x_5\}.$$

Объекты A_1 и A_6 удаляются из дальнейшего рассмотрения. Число объектов/кластеров уменьшается на 1.

4⁰. Вычислим расстояния между всеми парами оставшихся объектов и остальными парами «объект/новый кластер». На этом шаге наиболее близкими будут объекты A_4 и A_8 , находящиеся на расстоянии $d_{11}(A_4, A_8) = |0-0| + |1-1| + |1-2| + |3-3| + |3-2| = 2$. Объекты A_4 и A_8 формируют кластер C_{12} , которому соответствует мультимножество

$$Y_{12} = A_4 + A_8 = \{0^\circ x_1, 1^\circ x_2, 1^\circ x_3, 3^\circ x_4, 3^\circ x_5\} + \\ + \{0^\circ x_1, 1^\circ x_2, 2^\circ x_3, 3^\circ x_4, 2^\circ x_5\} = \\ = \{0^\circ x_1, 2^\circ x_2, 3^\circ x_3, 6^\circ x_4, 5^\circ x_5\},$$

и удаляются из дальнейшего рассмотрения.

5⁰. Вычислим расстояния между всеми парами оставшихся объектов и образованными кластерами. На этом шаге самыми близкими будут две пары объектов A_2, A_7 и A_7, A_9 , которые находятся на одинаковых расстояниях в пространстве мультимножеств: $d_{11}(A_2, A_7) = 4$ и $d_{11}(A_7, A_9) = 4$. Выберем для формирования нового кластера C_{13} пару объектов A_7, A_9 и удалим их из дальнейшего рассмотрения. Кластеру C_{13} будет соответствовать мультимножество

$$Y_{13} = A_7 + A_9 = \{3^\circ x_1, 5^\circ x_2, 7^\circ x_3, 1^\circ x_4, 0^\circ x_5\}.$$

6⁰. Вычисляя последовательно шаг за шагом расстояния между всеми парами объектов/кластеров и выбирая на каждом шаге наиболее близкие пары, получим кластеры C_{14} , C_{15} , C_{16} , C_{17} , C_{18} , представленные следующими мультимножествами:

$$Y_{14} = A_2 + A_3 = \{7^\circ x_1, 4^\circ x_2, 2^\circ x_3, 3^\circ x_4, 0^\circ x_5\}, \\ d_{11}(A_2, A_3) = 8;$$

$$Y_{15} = Y_{11} + A_5 = \{0^\circ x_1, 0^\circ x_2, 0^\circ x_3, 15^\circ x_4, 9^\circ x_5\}, \\ d_{11}(Y_{11}, A_5) = 8;$$

$$Y_{16} = Y_{12} + A_{10} = \{0^\circ x_1, 2^\circ x_2, 4^\circ x_3, 8^\circ x_4, 10^\circ x_5\}, \\ d_{11}(Y_{12}, A_{10}) = 8;$$

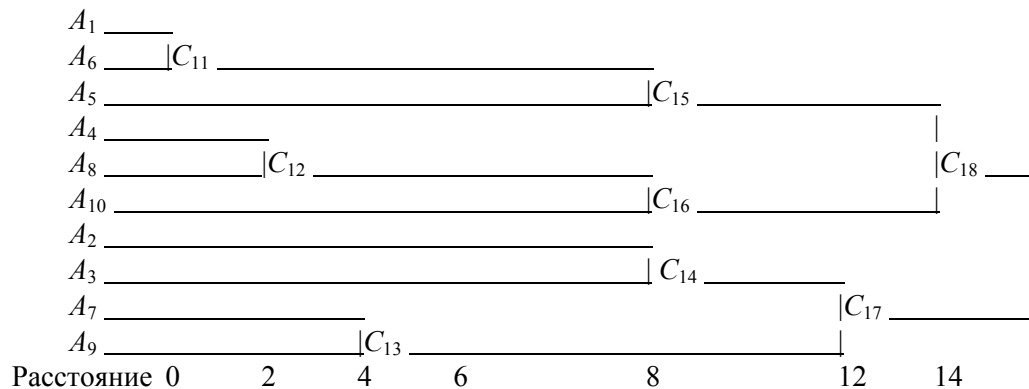
$$Y_{17} = Y_{13} + Y_{14} = \{10^\circ x_1, 9^\circ x_2, 9^\circ x_3, 4^\circ x_4, 0^\circ x_5\}, \\ d_{11}(Y_{13}, Y_{14}) = 12;$$

$$Y_{18} = Y_{15} + Y_{16} = \{0^\circ x_1, 2^\circ x_2, 4^\circ x_3, 23^\circ x_4, 19^\circ x_5\}, \\ d_{11}(Y_{15}, Y_{16}) = 14.$$

7⁰. Процедура оканчивается, когда останется два кластера C_{17} и C_{18} , которым соответствуют мультимножества Y_{17} и Y_{18} . В кластер C_{17} входят объекты A_2, A_3, A_7, A_9 , которые имеют, в основном, «низкие» оценки x_1, x_2, x_3 , а в кластер C_{18} – объекты $A_1, A_6, A_5, A_4, A_8, A_{10}$, имеющие, главным образом, «высокие» оценки x_4, x_5 . В результате процедуры классификации будет построено дерево, изображенное на рисунке.

Если на шаге 5⁰ сформировать другой кластер, объединяющий пару объектов A_2 и A_7 , то итоговое разбиение совокупности объектов $A = \{A_1, \dots, A_{10}\}$ будет состоять из тех же самых кластеров C_{17} и C_{18} . Отметим, что полученный в нашем случае конечный результат классификации многопризнаковых объектов полностью совпадает с приведенным в книге [15], где многопризнаковые объекты были представлены как векторы числовых значений признаков (Табл. 1), а агрегирование в кластеры проводилось с помощью так называемого алгоритма суммирования. Оба алгоритма успешно идентифицировали похожие группы объектов.

Изложенный выше алгоритм иерархической кластеризации для решения задачи номинальной классификации объектов, описываемых многими



Результирующее дерево алгоритма иерархической кластеризации

качественными признаками, без особых затруднений распространяется на случаи, где объекты могут присутствовать в нескольких версиях (копиях) и задано несколько классов решений.

В методах неиерархической кластеризации анализа число кластеров g считается фиксированным и заданным заранее. Обобщенная схема неиерархической кластеризации объектов, описываемых многими качественными признаками и представленных мультимножествами, включает следующие основные этапы.

Шаг 1. Выбрать некоторое начальное разбиение совокупности объектов $A = \{A_1, \dots, A_n\}$ на g кластеров $C_1, \dots, C_g$.

Шаг 2. Распределить все объекты $A_1, \dots, A_n$ по кластерам $C_1, \dots, C_g$ в соответствии с некоторым правилом. Например, вычислить расстояния $d(A_i, Y_t)$ между мультимножествами A_i ($i=1, \dots, n$), представляющими объекты A_i , и мультимножествами Y_t , представляющими кластеры C_t ($t=1, \dots, g$), и поместить объект A_i в ближайший из кластеров C_h , задаваемый условием $d(A_i, Y_h) = \min_t d(A_i, Y_t)$. Или найти для каждого кластера C_t его центр A_t° , например, как решение следующей задачи оптимизации:

$$J(A_t^\circ, Y_t) = \min_p \sum_i d(A_i, A_p), \quad (9)$$

и поместить каждый объект A_i в кластер C_r с ближайшим центром A_r° , который определяется условием $d(A_i, A_r^\circ) = \min_t d(A_i, A_t^\circ)$. Функционал $J(A_t^\circ, Y_t)$ (9) по своему смыслу близок критерию $f(Y_r^*)$ компактности кластера (8) и характеризует своего рода «кучность» распределения объектов в признаковом пространстве. Заметим, что в нашем случае центр кластера A_t° может совпадать с одним из реально имеющихся объектов A_i из совокупности A или быть так называемым «фантомным» объектом, который отсутствует в совокупности A , но сконструирован из признаков x_j как мультимножество A_t° вида (1) или (3).

Шаг 3. Если все объекты $A_1, \dots, A_n$ после их распределения по кластерам $C_1, \dots, C_g$ не изменят своей принадлежности к кластеру, заданной первоначальным разбиением, то вывести результат и остановиться. В противоположном случае вернуться к шагу 2.

Результат классификации объектов можно оценить качеством разбиения. Наилучшее разбиение может быть найдено, в частности, как решение следующей оптимизационной задачи:

$$\sum_t J(A_t^\circ, Y_t) \rightarrow \min \quad (10)$$

где $J(A_t^\circ, Y_t)$ определяется, например, выражением (9). В общем случае решение задачи (10) неоднозначно, поскольку функционал качества разбиения является функцией, имеющей много локальных экстремумов. Конечный результат будет зависеть, кроме того, и от первоначального (близко или далеко от оптимального) распределения объектов по классам.

Условие $\min d(Y_p, Y_q)$ должно заменяться условием $\max s(Y_p, Y_q)$, когда один из индексов сходства объектов используется в алгоритмах кластеризации.

При решении практических задач структуризации совокупности многопризнаковых объектов может быть полезным следующий прием. Вначале объекты классифицируются при помощи некоторого метода иерархической кластеризации, и формируются несколько возможных разбиений объектов. Затем совокупность этих разбиений анализируются методами неиерархической кластеризации, и находится наилучшее или наиболее приемлемое разбиение объектов.

Литература

1. Арлазаров В.Л., Логинов А.С., Славин О.А.. Характеристики программ оптического распознавания текста. // Программирование, 2002, №3, с.45-63.
2. Деза М.М., Лоран М.. Геометрия разрезов и метрик./Пер. с англ. – М.: МЦНМО, 2001.
3. Комарова Н.А., Петровский А.Б. Метод согласованной групповой классификации многопризнаковых объектов. // Поддержка принятия решений. Труды Института системного анализа Российской академии наук. Т. 35, с.19-32. – М.: Издательство ЛКИ, 2008.
4. Кнут Д.Э. Искусство программирования. Т.2. Получисленные алгоритмы./Пер. с англ. – М.: Мир, 1977.
5. Ларичев О.И. Вербальный анализ решений. – М.: Наука, 2006.
6. Ларичев О.И., Прохоров А.С., Петровский А.Б., Стернин М.Ю., Шепелев Г.И. Опыт планирования фундаментальных исследований на конкурсной основе. // Вестник АН СССР, 1989, № 7, с.51-61.
7. Петровский А.Б. Пространства множеств и мультимножеств. – М.: Едиториал УРСС, 2003.
8. Петровский А.Б. Кластерный анализ объектов с противоречивыми свойствами. // Труды Десятой национальной конференции по искусственному интеллекту с международным участием (КИИ-2006). В 3-х томах. Т.3, с.755-764. – М.: Физматлит, 2006.
9. А.Б.Петровский, И.П.Тихонов, А.В.Балышев, Н.А.Комарова. Построение согласованных групповых правил для конкурсного отбора научных проектов. // Третья международная конференция «Системный анализ и информационные технологии» (САИТ-2009).

- Труды конференции, с.337-348. – М.: ПолиПринтСервис, 2009.
10. Anderberg M.R. Cluster Analysis for Applications. – New York: Academic Press, 1973.
 11. Blizard W. Multiset theory. // Notre Dame Journal of Formal Logic, 1989, Vol. 30, N 1, p. 36-66
 12. Doumpos M., Zopounidis C. Multicriteria Decision Aid Classification Methods. – Dordrecht: Kluwer Academic Publishers, 2002.
 13. Furems E. Knowledge-based multi-attribute classification problems structuring. // Ruan D., Montero J., Lu J., Martinez L., D'hondt P., Kerre E. (eds.) Computational Intelligence in Decision And Control, p. 465-470. – Singapore: World Scientific Publisher, 2008.
 14. Greco S., Matarazzo B., Slowinski R. Rough sets methodology for sorting problems in presence of multiple attributes and criteria. // European Journal of Operational Research, 2002, Vol. 138, N 2, p. 247-259.
 15. Hartigan J.A. Clustering Algorithms. – New York: Wiley, 1975.
 16. Hwang C.L., Lin M.J. Group Decision Making under Multiple Criteria. – Berlin: Springer-Verlag, 1987.
 17. Köksalan M., Ulu C. An interactive approach for placing alternatives in preference classes. // European Journal of Operational Research, 2003, Vol. 144, N. 2, p. 429-439.
 18. Larichev O.I., Olson D.L. Multiple Criteria Analysis in Strategic Siting Problems. – Boston: Kluwer Academic Publishers, 2001.
 19. Miyamoto S. Cluster analysis as a tool of interpretation of complex systems. // Working Paper WP-87-41. – Laxenburg, Austria: IIASA, 1987.
 20. Petrovsky A.B. Structuring techniques in multiset spaces. // Fandel G., Gal T. (eds.) Multiple Criteria Decision Making, p. 174-184. – Berlin: Springer-Verlag, 1997.
 21. Petrovsky A.B. Multi-attribute sorting of qualitative objects in multiset spaces. // Koksalan M., Zionts S. (eds.) Multiple Criteria Decision Making in the New Millenium, p. 124-131. – Berlin: Springer-Verlag, 2001.
 22. Petrovsky A.B. Multi-attribute classification of credit cardholders: multiset approach. // International Journal of Management and Decision Making, 2006, Vol. 7, N 2/3, p.166-179.
 23. Roubens M. Ordinal multiattribute sorting and ordering in the presence of interacting points of view. // Bouyssou D., Jacquet-Lagrèze E., Perny P., Slowinski R., Vanderpooten D., Vincke P. (eds.) Aiding Decisions with Multiple Criteria: Essays in Honor of Bernard Roy, p. 229-246. – Dordrecht: Kluwer Academic Publishers, 2001.
 24. Roy B. Multicriteria Methodology for Decision Aiding. – Dordrecht: Kluwer Academic Publishers, 1996.
 25. Roy B., Bouyssou D. Aide Multicritère à la Décision: Méthodes et Cas. – Paris: Economica, 1993.
 26. Saaty T. Multicriteria Decision Making: The Analytic Hierarchy Process. – Pittsburgh: RWS Publications, 1990.
 27. Vincke P. Multicriteria Decision Aid. – Chichester: Wiley, 1992.
 28. Yager R.R. On the Theory of Bags. // International Journal of General Systems, 1986, Vol. 13, p. 23-37.
 29. Zadeh L.A. From computing with numbers to computing with words – from manipulation of measurements to manipulation of perceptions. // IEEE Transactions on Circuits and Systems, 1999, Vol. 45 N 1, p. 105-119.

Петровский Алексей Борисович. Заведующий лабораторией методов и систем поддержки принятия решений Учреждения Российской академии наук Института системного анализа РАН. Окончил Московский государственный университет им. М.В.Ломоносова в 1967 г. Доктор технических наук. Автор более 130 научных работ, в том числе 3 монографий и 1 учебника. Области научных интересов: дискретная математика, теория мультимножеств, многокритериальный анализ решений, системы поддержки принятия решений, информационные технологии, системный анализ, научно-техническая политика, прогнозирование, планирование и организация научных исследований.