Метод согласованной групповой классификации многопризнаковых объектов *

Н. А. Комарова[1], А. Б. Петровский[2]

Описана модификация метода групповой классификации многопризнаковых объектов, представленных в виде мультимножеств. Разработаны новые алгоритмы для построения согласованных групповых правил отнесения объектов к уточненным классам и выделения противоречиво классифицированных объектов.

Введение

Классификация рассматриваемых объектов по их свойствам, выраженных признаками (атрибутами), относится к числу основных задач многокритериального принятия решений. Решение задач индивидуального выбора осуществляется на основе предпочтений единственного лица, принимающего решение (ЛПР). При решении задач коллективного выбора требуется одновременно учитывать различные интересы многих участников, разнообразие и несовпадение их целей и способов выражения их предпочтений. Решение такого рода задач связано также с необходимостью обработки большого объема вербальных и числовых данных без применения необоснованных и необратимых преобразований исходной информации.

Известно достаточно много подходов к индивидуальной классификации объектов, характеризующихся многими количественными и/или качественными признаками [1, 4]. В то же время практически отсутствуют методы групповой классификации многопризнаковых объектов. Примером такой задачи может служить конкурсный отбор проектов, оцененных несколькими экспертами по многим критериям.

---

* Работа поддержана программами фундаментальных исследований президиума РАН «Фундаментальные проблемы информатики и информационных технологий» и ОНИТ РАН «Фундаментальные основы информационных технологий и систем», Российским фондом фундаментальных исследований (проекты 06-07-89352, 07-01-00515, 07-07-13546, 08-01-00247).

[1] *117312, Москва, проспект 60-лет Октября, 9, ИСА РАН, natashich@mail.ru*

[2] *117312, Москва, проспект 60-лет Октября, 9, ИСА РАН, pab@isa.ru*

В работе предложена модификация метода групповой классификации объектов, которые описываются многими разнородными признаками и могут существовать в нескольких экземплярах [5]. В методе используется аппарат теории метрических пространств мультимножеств, позволяющий строить согласованные групповые правила для классификации объектов, которые аппроксимируют различные индивидуальные правила экспертной сортировки объектов, а также позволяют находить объекты, имеющие противоречивые правила.

1. Групповая классификация многопризнаковых объектов

Процедура классификации объектов в рамках формальной логики может быть описана как совокупность (последовательность) решающих правил, которые представляются выражениями вида:

$$\text{ЕСЛИ } \langle\text{условия}\rangle\text{, ТО } \langle\text{решение}\rangle. \tag{1}$$

При прямой классификации терм ⟨условия⟩ включает названия объектов или перечень значений признаков, описывающих объекты класса. При непрямой классификации один или несколько термов ⟨условия⟩ конструируются как отношения между различными признаками и/или их значениями. Терм ⟨решение⟩ в обоих случаях означает, что объект принадлежит к определенному классу. Заметим, что подобным образом формируются и базы знаний экспертных систем продукционного типа.

Рассмотрим задачу групповой классификации объектов в следующей постановке. Задано множество объектов (вариантов, альтернатив) $A_1, \dots, A_n$, которые описываются m дискретными признаками $Q_1, \dots, Q_m$, имеющими количественные и/или качественные шкалы оценок. Каждая группа признаков $Q_s = \{q_s^{1}, \dots, q_s^{hs}\}$, $e_s = 1, \dots, h_s$, $s = 1, \dots, m$ отражает содержательное качество объектов. Например, q_s^{es} может быть значением показателя, определяющего какое-либо свойство объекта, или оценкой объекта по критерию, и тому подобное. Объекты $A_1, \dots, A_n$ предварительно рассортированы по нескольким классам $C_1, \dots, C_f$ путем прямой классификации. Принадлежность объекта A_i, $i = 1, \dots, n$ к некоторому классу C_t, $t = 1, \dots, f$ выражается атрибутом R, которое можно считать еще одним качественным признаком со шкалой значений $R = \{r_t\}$. Совокупность признаков $\{q_{i1}^{e1}, \dots, q_{im}^{em}; r_{it}\}$ представляет собой индивидуальное правило классификации вида (1), относящее объект A_i к классу C_t.

Любой объект A_i существует в k экземплярах, которые отличаются наборами признаков, его характеризующих. Например, когда объект независимо оценивается и сортируется по классам k экспертами или его параметры измеряются k различными способами. Тем самым имеется k различных и, возможно, противоречивых индивидуальных правил классификации объектов. Требуется построить одно или несколько агрегированных групповых правил вида (1), которые наилучшим (в некотором смысле) образом аппроксимируют совокупность всех индивидуальных правил классификации объектов и относят объекты в заданные классы с допустимой точностью.

Примером задачи групповой классификации многопризнаковых объектов является конкурсный отбор проектов. Так, при формировании одной из программ научных исследований на конкурс было представлено более 250 проектов. Требовалось провести экспертную оценку заявок и разделить их на группы, в той или иной степени отвечающие целям программы и им не соответствующие.

Каждая заявка оценивалась несколькими экспертами по многим качественным критериям: Q_1. Важность проекта для программы; Q_2. Перспективность проекта; Q_3. Новизна подхода к решению поставленных задач; Q_4. Квалификация исполнителей проекта; Q_5. Ресурсное обеспечение работ; Q_6. Возможность быстрого выхода результатов в практику. Критерии имели порядковые или номинальные шкалы оценок с развернутыми словесными формулировками градаций качества. Так, шкала критерия Q_1 «Важность проекта для программы» имела вид:

q_1^1 — проект обеспечивает достижение основных целей программы;

q_1^2 — проект способствует достижению одной из основных целей программы;

q_1^3 — проект не обеспечивает достижения целей программы.

Каждый эксперт, наряду с оценкой заявки по всем критериям, давал одну из следующих рекомендаций:

r_1 — принять проект;

r_2 — отклонить проект;

r_3 — отправить проект на доработку.

Экспертиза заявок осуществлялась экспертами независимо друг от друга без согласования их мнений. Рекомендации экспертов являются индивидуальными правилами предварительной классификации (сортировки) рассматриваемых заявок, которые могут совпадающих или различаться. Для отбора проектов нужно построить агрегированное групповое правило, которое обобщает индивидуальные решения всех экспертов и позволяет отнести проект в один из заданных классов.

2. Представление многопризнаковых объектов

Многопризнаковые объекты A_i, $i = 1, \ldots, n$ обычно принято представлять как векторы или кортежи $q_i = (q_{i1}^{e1}, \ldots, q_{im}^{em})$ в пространстве $Q = Q_1 \times \ldots \times Q_m$, где $Q_s = \{q_s^{es}\}$ — непрерывная или дискретная шкала s-го признака, $e_s = 1, \ldots, h_s$, $s = 1, \ldots, m$. Ситуация существенным образом усложняется, если одному и тому же объекту A_i может соответствовать не один, а несколько m-мерных векторов. В таких случаях объект A_i представляется в пространстве Q группой, состоящей из k векторов $\{q_i^{(1)}, \ldots, q_i^{(k)}\}$ вида $q_i^{(j)} = (q_{i1}^{e1(j)}, \ldots, q_{im}^{em(j)})$, $j = 1, \ldots, k$, которая должна рассматриваться как единое целое. Измеренные разными способами значения параметров или индивидуальные оценки экспертов, могут быть похожими, различающимися и даже противоречивыми, что может приводить к несравнимости m-мерных векторов $q_i^{(j)} = (q_{i1}^{e1(j)}, \ldots, q_{im}^{em(j)})$, характеризующих один и тот же объект A_i. Совокупность таких многомерных объектов может иметь в пространстве Q сложную структуру, достаточно трудную для анализа.

Эти трудности можно преодолеть, если воспользоваться иным способом представления многопризнаковых объектов, основанным на формализме мультимножеств, который позволяет одновременно учесть все комбинации значений количественных и качественных признаков, а также различное число значений каждого из этих признаков. Введем вместо прямого произведения m шкал признаков $Q = Q_1 \times \ldots \times Q_m$ множество $G = Q_1 \cup \ldots \cup Q_m$ — обобщенную шкалу признаков, состоящую из m групп признаков, и представим объект в таком символическом виде:

$$\boldsymbol{A}_i = \{k_{Ai}(q_1^{\,1})\circ q_1^{\,1}, \ldots, k_{Ai}(q_1^{\,h1})\circ q_1^{\,h1}, \ldots, k_{Ai}(q_m^{\,1})\circ q_m^{\,1}, \ldots, k_{Ai}(q_m^{\,hm})\circ q_m^{\,hm}\}, \quad (2)$$

где число $k_{Ai}(q_s^{\,es})$ указывает, сколько раз признак $q_s^{\,es} \in Q_s$ встречается в описании объекта $\boldsymbol{A}_i$, а знак $\circ$ обозначает кратность признака $q_s^{\,es}$. Множество G характеризует свойства совокупности объектов $\mathsf{A} = \{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n\}$. Такая запись объекта $\boldsymbol{A}_i$ представляет его как мультимножество или множество с повторяющимися элементами [3].

Определяются следующие операции над мультимножествами: объединение $\boldsymbol{A} \cup \boldsymbol{B}$, пересечение $\boldsymbol{A} \cap \boldsymbol{B}$, сложение $\boldsymbol{A} + \boldsymbol{B}$, вычитание $\boldsymbol{A} - \boldsymbol{B}$, симметрическая разность $\boldsymbol{A} \Delta \boldsymbol{B}$, дополнение $\overline{\boldsymbol{A}} = \boldsymbol{Z} - \boldsymbol{A}$, умножение на целое число $c \times \boldsymbol{A}$, $c > 0$, арифметическое умножение $\boldsymbol{A} \times \boldsymbol{B}$, арифметическая n-ая степень $\boldsymbol{A}^n$, прямое произведение $\boldsymbol{A} \times \boldsymbol{B}$, прямая n-ая степень $(\times\boldsymbol{A})^n$, $\boldsymbol{Z}$ — максимальное мультимножество с $k_Z(x) = \max_{A \in \mathsf{A}} k_A(x)$.

Новые типы операций над мультимножествами открывают новые возможности для агрегирования многопризнаковых объектов. Например, группа $\boldsymbol{X}_a$ объектов может быть получена как сумма $\boldsymbol{X}_a = \sum_i \boldsymbol{A}_i$, объединение $\boldsymbol{X}_a = \cup_i \boldsymbol{A}_i$ или пересечение $\boldsymbol{X}_a = \cap_i \boldsymbol{A}_i$ мультимножеств $\boldsymbol{A}_i$, описывающих объекты A_i, либо как линейная комбинация различных мультимножеств вида $\boldsymbol{X}_a = \sum_i c_i \times \boldsymbol{A}_i$, $\boldsymbol{X}_a = \cup_i c_i \times \boldsymbol{A}_i$ или $\boldsymbol{X}_a = \cap_i c_i \times \boldsymbol{A}_i$. Заметим кстати, что и само мультимножество $\boldsymbol{A}_i$ вида (2) формально можно представить как сумму $\boldsymbol{A}_i = \sum_j \{q_i^{(j)}\}$ k различных множеств (m-мерных векторов или кортежей) $q_i^{(j)}$, характеризующих один и тот же объект A_i.

На семействе мультимножеств $\mathsf{A} = \{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n\}$ можно ввести новые классы метрических пространств мультимножеств (A, d), которые задаются следующими метриками (псевдометриками):

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

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

где p — целое число, m — мера мультимножества, заданная на алгебре мультимножеств $\mathsf{L}(\boldsymbol{Z})$ [3].

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

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

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

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

Основная метрика $d_{1p}(\boldsymbol{A},\boldsymbol{B})$ является метрикой типа Хемминга, при $p = 1$ используемой во многих приложениях. Полностью усредненная метрика $d_{2p}(\boldsymbol{A},\boldsymbol{B})$ характеризует различие между двумя мультимножествами $\boldsymbol{A}$ и $\boldsymbol{B}$, отнесенное к расстоянию, максимально возможному в исходном пространстве. Локально усредненная метрика $d_{3p}(\boldsymbol{A},\boldsymbol{B})$ задает различие, отнесенное к максимально возможной «общей части» $\boldsymbol{A} \cup \boldsymbol{B}$ только этих двух мультимножеств в исходном пространстве.

3. Метод согласованной групповой классификации

Метод групповой классификации многопризнаковых объектов МАСКА (МногоАспектная Согласованная Классификация Альтернатив) предназначен для построения согласованных групповых правил классификации объектов $A_1, \ldots, A_n$, оцененных k экспертами по m критериям, агрегирующих индивидуальные правила сортировки объектов, и выявления противоречивых индивидуальных правил классификации. Сопоставим каждому многопризнаковому объекту мультимножество

$$\boldsymbol{A}_i = \{k_{Ai}(q_1^{\,1})\circ q_1^{\,1}, \ldots, k_{Ai}(q_1^{\,h1})\circ q_1^{\,h1}, \ldots, k_{Ai}(q_m^{\,1})\circ q_m^{\,1}, \ldots, k_{Ai}(q_m^{\,hm})\circ q_m^{\,hm}, k_{Ai}(r_1)\circ r_1, \ldots, k_{Ai}(r_f)\circ r_f\} \quad (3)$$

над доменом $G = Q_1 \cup \ldots \cup Q_m \cup R$, где $k_{Ai}(q_s^{\,es})$ и $k_{Ai}(r_t)$ равны числу экспертов, давших объекту A_i оценку $q_s^{\,es}$ и рекомендацию r_t. Принадлежность объекта A_i к некоторому классу C_t определяется, например, правилом большинства голосов, в соответствии с которым объект A_i считается принадлежащим к классу C_t, если $k_{Ai}(r_t) > k_{Ai}(r_p)$ для всех $p \neq t$.

Представление каждого объекта как мультимножества (3) есть, по сути, групповое правило классификации вида (1), объединяющее k индивидуальных правил сортировки. Очевидно, что такое правило может содержать различные, в том числе и противоречивые индивидуальные правила экспертов. Кроме того, могут быть противоречивыми и групповые правила классификации для разных объектов. Нужно найти одно или несколько агрегированных правил классификации всей совокупности объектов вида (1), которые наилучшим образом были согласованы с групповыми правилами классификации и позволяли бы выявлять противоречивые правила.

Для простоты положим, что результатом классификации должно быть разложение совокупности объектов только на два класса: C_a («хорошие») и C_b («плохие»). При необходимости рассортировать объекты на большее число классов, можно сначала разбить совокупность объектов на две группы, затем одну из них или обе группы — на подгруппы, и так далее. Сопоставим каждому классу решений C_a и C_b мультимножества $\boldsymbol{X}_a$ и $\boldsymbol{X}_b$, которые представляют собой соответственно совокупности хороших и плохих объектов, сформированных путем сложения соответствующих им мультимножеств

$$\boldsymbol{X}_t = \textstyle\sum_{i \in It} \boldsymbol{A}_i = \{k_{Xt}(q_1^{\,1})\text{o}q_1^{\,1}, \ldots, k_{Xt}(q_1^{\,h1})\text{o}q_1^{\,h1}, \ldots, k_{Xt}(q_m^{\,1})\text{o}q_m^{\,1}, \ldots,$$

$$k_{Xt}(q_m^{\,hm})\text{o}q_m^{\,hm}, \mathrm{k_{Xt}(r_a)or_a}, \mathrm{k_{Xt}(r_b)or_b}\},$$

$$k_{Xt}(q_s^{\,es}) = \textstyle\sum_{i \in It} k_{Ai}(q_s^{\,es}),\ k_{Xt}(r_t) = \sum_{i \in It} k_{Ai}(r_t),\ t = a,b, \quad (4)$$

где подмножества индексов $I_a \cup I_b = \{1, \ldots, n\}$, $I_a \mid I_b = \varnothing$. При суммировании мультимножеств учитываются значения всех признаков, характеризующих все объекты, входящие в группу.

Основная идея нахождения агрегированного правила, аппроксимирующего большое число противоречивых правил сортировки многопризнаковых объектов заключается в следующем [5]. Над тем же самым доменом G образуются новые мультимножества

$$\boldsymbol{R}_a = \{k_{Xa}(r_a)\circ r_a,\ k_{Xb}(r_a)\circ r_a\},\ \boldsymbol{R}_b = \{k_{Xa}(r_b)\circ r_b,\ k_{Xb}(r_b)\circ r_b\},$$

$$\boldsymbol{Q}_{as} = \textstyle\sum_{j\,\in\,Jas}\boldsymbol{Q}_j,\ \boldsymbol{Q}_{bs} = \textstyle\sum_{j\,\in\,Jbs}\boldsymbol{Q}_j,\ \boldsymbol{Q}_j = \{k_{Xa}(q_s^j)\circ q_s^j,\ k_{Xb}(q_s^j)\circ q_s^j\},$$

которые назовем «категориальными» и «содержательными», где подмножества индексов $J_{as} \cup J_{bs} = \{1, \ldots, h_s\}$, $J_{as} \cap J_{bs} = \varnothing$, $s = 1, \ldots, m$.

«Категориальные» мультимножества $\boldsymbol{R}_a$ и $\boldsymbol{R}_b$ соответствуют наилучшей из всех возможных декомпозиций рассматриваемой совокупности объектов на классы C_a и C_b для имеющегося набора первичных правил сортировки. Очевидно, что и расстояние $d^* = d(\boldsymbol{R}_a, \boldsymbol{R}_b)$ между мультимножествами $\boldsymbol{R}_a$ и $\boldsymbol{R}_b$ будет предельно возможным расстоянием между объектами, входящими в разные классы. При идеальных предварительных сортировках объектов, в которых отсутствуют противоречия между индивидуальными правилами экспертов, расстояние d^* будет равно соответственно $d_{1p}{}^* = [kn]^{1/p}$, $d_{2p}{}^* = [1/h]^{1/p}$, $d_{3p}{}^* = 1$, где $h = h_1 + \ldots + h_m + f$ — общее число значений всех признаков, описывающих объекты.

Аналогичным образом для каждой s-й группы признаков Q_s нужно сформировать такие пары «содержательных» мультимножеств $\boldsymbol{Q}_{sa}{}^*$, и $\boldsymbol{Q}_{sb}{}^*$, которые находились бы на максимально возможном расстоянии и с достаточной точностью совпадали бы с разложением $\{\boldsymbol{R}_a, \boldsymbol{R}_b\}$. Задача нахождения классифицирующих признаков $q_s^j \in Q_{as}{}^*$, $q_s^j \in Q_{bs}{}^*$ сводится, таким образом, к решению m оптимизационных задач вида

$$d(\boldsymbol{Q}_{as},\boldsymbol{Q}_{bs}) \to \max d(\boldsymbol{Q}_{as},\boldsymbol{Q}_{bs}) = d(\boldsymbol{Q}_{as}{}^*,\boldsymbol{Q}_{bs}{}^*). \tag{5}$$

Решение каждой из указанных задач задает наилучшее разложение $\{\boldsymbol{Q}_{sa}{}^*,\boldsymbol{Q}_{sb}{}^*\}$ имеющейся совокупности объектов по s-й группе признаков. Классифицирующие признаки для различных групп признаков можно упорядочить по величине $L_s = d(\boldsymbol{Q}_{as}{}^*,\boldsymbol{Q}_{bs}{}^*) / d(\boldsymbol{R}_a,\boldsymbol{R}_b)$, которая определяет точность аппроксимации и в некотором смысле характеризует относительную значимость s-й группы признаков Q_s в агрегированном групповом правиле классификации объектов.

Разные комбинации классифицирующих признаков, входящих в «содержательные» и «категориальные» мультимножества, дадут разные варианты агрегированных групповых правил для классификации совокупности многопризнаковых объектов:

$$\text{IF } \langle(q_u^j \in Q_{au}{}^*) \text{ AND } (q_v^j \in Q_{av}{}^*) \text{ AND } \ldots \text{ AND } (q_w^j \in Q_{aw}{}^*)\rangle,$$

$$\text{THEN } \langle\text{Object } A_i \in C_a\rangle, \tag{6}$$

$$\text{IF } \langle(q_u^j \in Q_{bu}{}^*) \text{ AND } (q_v^j \in Q_{bv}{}^*) \text{ AND } \ldots \text{ AND } (q_w^j \in Q_{bw}{}^*)\rangle,$$

$$\text{THEN } \langle \text{Object } A_i \in C_b \rangle. \tag{7}$$

Среди объектов, отобранных по агрегированному правилу в заданный класс C_a или C_b, будут как правильно, так и неправильно классифицированные объекты. Желательно найти такие согласованные групповые правила, которые обеспечат максимальную разность между числом правильно и числом неправильно классифицированных объектов. Комбинации соответствующих классифицирующих признаков включаются в согласованные групповые правила отнесения объектов к уточненным классам $C_a \setminus C_{ac}$ (безусловно хорошие) и $C_b \setminus C_{bc}$ (безусловно плохие), которые записываются в виде:

$$\text{IF } \langle (\textstyle\sum_{q \in Qau^*} k_{Ai}(q) > \sum_{q \in Qbu^*} k_{Ai}(q)) \text{ AND } (\sum_{q \in Qav^*} k_{Ai}(q) > \sum_{q \in Qbv^*} k_{Ai}(q))$$

$$\text{AND} \ldots \text{AND } (k_{Ai}(r_a) > k_{Ai}(r_b))\rangle, \text{ THEN } \langle \text{Object } A_i \in C_a \setminus C_{ac}\rangle, \tag{8}$$

$$\text{IF } \langle (\textstyle\sum_{q \in Qau^*} k_{Ai}(q) < \sum_{q \in Qbu^*} k_{Ai}(q)) \text{ AND } (\sum_{q \in Qav^*} k_{Ai}(q) < \sum_{q \in Qbv^*} k_{Ai}(q))$$

$$\text{AND} \ldots \text{AND } (k_{Ai}(r_a) < k_{Ai}(r_b))\rangle, \text{ THEN } \langle \text{Object } A_i \in C_b \setminus C_{bc}\rangle. \tag{9}$$

Одновременно формируется класс противоречиво классифицированных объектов $C_c = C_{ac} \cup C_{bc}$, которые имеют противоречивые индивидуальные правила сортировки

$$\text{IF } \langle [(\textstyle\sum_{q \in Qau^*} k_{Ai}(q) > \sum_{q \in Qbu^*} k_{Ai}(q)) \text{ AND} \ldots \text{AND } (k_{Ai}(r_a) < k_{Ai}(r_b))]$$

$$\text{OR } [(\textstyle\sum_{q \in Qau^*} k_{Ai}(q) < \sum_{q \in Qbu^*} k_{Ai}(q)) \text{ AND} \ldots \text{AND } (k_{Ai}(r_a) > k_{Ai}(r_b))]\rangle,$$

$$\text{THEN } \langle \text{Object } A_i \in C_c\rangle. \tag{10}$$

Эти объекты нуждаются в дополнительном анализе.

Таким образом, процедура построения согласованных групповых правил для классификации многопризнаковых объектов, которые агрегируют большое число индивидуальных правил сортировки, включает в себя следующие алгоритмы:

алгоритм 1 — поиск классифицирующих признаков и построение агрегированных групповых правил отнесения объектов к классам C_a и C_b;

алгоритм 2 — построение согласованных групповых правил отнесения объектов к уточненным классам $C_a \setminus C_{ac}$, $C_b \setminus C_{bc}$ и к классу противоречиво классифицированных объектов $C_c = C_{ac} \cup C_{bc}$.

4. Построение агрегированных правил групповой классификации объектов

Проиллюстрируем работу метода МАСКА на примере решения упомянутой в разделе 1 практической задачи конкурсного отбора проектов для формирования научной программы. Классифицируемыми многопризнаковыми объектами являются 259 заявок, которые оценивались 3 экспертами по шести критериям Q_1, …, Q_6 с вербальными шкалами оценок. Требуется найти согласованные групповые правила для отнесения заявок к классам $C_a \setminus C_{ac}$ (безусловно принятые), $C_b \setminus C_{bc}$ (безусловно отклоненные), C_c (противоречиво классифицированные).

Построение агрегированных групповых правил отнесения заявок к классам C_a и C_b осуществляется алгоритмом 1, который состоит из следующих шагов.

- Построить таблицу решений $\mathrm{K} = ||k_{Ai}(x_j)||_{k \times h}$ (таблица 1), строки которой представляют собой мультимножества $\boldsymbol{A}_i$ вида (3).

Таблица 1

Таблица решений

Объекты	Q_1			Q_2			Q_3			Q_4				Q_5				Q_6			R	
	q_1^1	q_1^2	q_1^3	q_2^1	q_2^2	q_2^3	q_3^1	q_3^2	q_3^3	q_4^1	q_4^2	q_4^3	q_4^4	q_5^1	q_5^2	q_5^3	q_5^4	q_6^1	q_6^2	q_6^3	r_a	r_b
$\boldsymbol{A}_1$	1	2	0	2	1	0	3	0	0	2	1	0	0	0	2	1	0	2	1	0	3	0
$\boldsymbol{A}_{175}$	1	1	1	0	2	1	1	2	0	0	2	1	0	0	1	2	0	0	0	3	2	1
$\boldsymbol{A}_{176}$	1	1	1	0	2	1	1	2	0	0	2	1	0	0	1	2	0	0	0	3	1	2
$\boldsymbol{A}_{259}$	0	2	1	0	1	2	0	3	0	0	1	1	1	0	0	2	1	0	3	0	0	3

Сформировать путем сложения мультимножеств $\boldsymbol{A}_i$ группы объектов A_i, относящихся к заданным классам C_a (принятые заявки) и C_b (отклоненные заявки), которым соответствуют мультимножествам $\boldsymbol{X}_a$ и $\boldsymbol{X}_b$. Отнесение заявки в один из двух классов производится по правилу большинства голосов на основе индивидуальных рекомендаций экспертов по признаку R без учета оценок по критериям Q_1, …, Q_6. Заметим, что проекты, получившие одинаковые оценки по каждому из критериев, но отличающиеся оценками по признаку R, например, проекты $\boldsymbol{A}_{175}$ и $\boldsymbol{A}_{176}$, попадают в разные классы. Построить агрегированную таблицу решений $\mathrm{K}' =||k_{Xt}(x_j)||_{2 \times h}$ (таблица 2), строки которой представляют собой мультимножества $\boldsymbol{X}_a$ и $\boldsymbol{X}_b$.

Таблица 2

Агрегированная таблица решений

Классы	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	R
	q_1^1 q_1^2 q_1^3	q_2^1 q_2^2 q_2^3	q_3^1 q_3^2 q_3^3	q_4^1 q_4^2 q_4^3 q_4^4	q_5^1 q_5^2 q_5^3 q_5^4	q_6^1 q_6^2 q_6^3	r_a r_b
$\boldsymbol{X}_a$	144 360 21	81 324 120	99 336 90	219 297 9 0	72 435 18 0	126 300 99	510 15
$\boldsymbol{X}_b$	45 156 51	27 93 132	36 111 105	51 132 63 6	60 147 30 15	45 135 72	78 174

- Найти классифицирующие признаки $q_s^j \in Q_{as}$*, $q_s^j \in Q_{bs}$* по каждой s-й группе признаков Q_s, которые являются решением оптимизационной задачи (5) с метрикой типа Хемминга. Проранжировать классифицирующие признаки по убыванию уровня аппроксимации L_s (таблица 3).

Таблица 3

Метрики *d* и уровни аппроксимации L_s

	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	R
d	333	297	303	393	327	273	591
Ls	0,563	0,503	0,517	0,665	0,553	0,462	1,000

- Построить агрегированные групповые правила классификации объектов (6) и (7):

$$\text{IF } \langle (q_4^j \in \{q_4^1, q_4^2\}) \text{ AND } (q_1^j \in \{q_1^1, q_1^2\}) \text{ AND } (q_5^j \in \{q_5^1, q_5^2\})$$
$$\text{AND } (q_3^j \in \{q_3^1, q_3^2\}) \text{ AND } (q_2^j \in \{q_2^1, q_2^2\}) \text{ THEN } \langle \text{Object } A_i \in C_a \rangle,$$
$$\text{IF } \langle (q_4^j \in \{q_4^3, q_4^4\}) \text{ AND } (q_1^j \in \{q_1^3\}) \text{ AND } (q_5^j \in \{q_5^3, q_5^4\})$$
$$\text{AND } (q_3^j \in \{q_3^3\}) \text{ AND } (q_2^j \in \{q_2^3\}) \text{ THEN } \langle \text{Object } A_i \in C_b \rangle.$$

Варианты группового правила для отбора лучших проектов, агрегирующего индивидуальные правила сортировки, выглядят так:

– исполнители проекта являются одним из лучших коллективов или обладают опытом и квалификацией, достаточными для проведения работ (оценки по критериям q_4^1 или q_4^2; уровень аппроксимации $L_s \geq 0{,}66$);
– проект обеспечивает достижение всех или одной из основных целей программы; исполнители проекта являются одним из лучших коллективов, имеют все необходимые материально-технические ресурсы или обладают опытом, квалификацией и материально-техническими ресурсами, достаточными для проведения работ (оценки по критериям q_4^1 или q_4^2; и q_1^1 или q_1^2; и q_5^1 или q_5^2; уровень аппроксимации $L_s \geq 0{,}55$)
Последнее правило совпадает с полученным на практике [2].

5. Построение согласованных правил групповой классификации объектов

Построение согласованных групповых правил отнесения заявок к уточненным классам $C_a \setminus C_{ac}$ (безусловно принятые), $C_b \setminus C_{bc}$ (безусловно отклоненные) и к классу противоречиво классифицированных $C_c = C_{ac} \cup C_{bc}$

производится путем последовательного выявления критериев Q_s*, которые обеспечивают максимальную разность между числом правильно и числом неправильно классифицированных заявок. Укажем основные шаги алгоритма 2 применительно к классу лучших проектов C_a.

- По каждому критерию Q_s*, входящему в групповое правило отнесения заявок к классу C_a, найти правильно классифицированные заявки, которые удовлетворяют условию

$$(\Sigma_{q \in Qau^*} k_{Ai}(q) > \Sigma_{q \in Qbu^*} k_{Ai}(q)) \text{ AND } (k_{Ai}(r_a) > k_{Ai}(r_b)),$$

и неправильно классифицированные заявки, которые удовлетворяют условию

$$(\Sigma_{q \in Qau^*} k_{Ai}(q) > \Sigma_{q \in Qbu^*} k_{Ai}(q)) \text{ AND } (k_{Ai}(r_a) < k_{Ai}(r_b)).$$

Результаты анализа совокупности принятых заявок даны в таблице 4.

Таблица 4

Итоги классификации заявок по одному критерию

условие	прав. класс.	неправ. класс.	разность
Q_1	173	55	118
$\boldsymbol{Q_2}$	**172**	**24**	**118**
Q_3	172	27	145
Q_4	174	57	117
Q_5	174	62	112

- Выбрать критерий Q_u*, обеспечивающий максимальную разность между числом правильно и числом неправильно классифицированных заявок (выделен жирным шрифтом в таблице 4). В нашем случае — это критерий Q_2*, который включается в согласованное групповое правило отнесения заявок к уточненному классу $C_a \setminus C_{ac}$.
- Рассмотреть все комбинации критерия Q_u* с остальными критериями, входящими в групповое правило отнесения заявок к классу C_a, и найти правильно и неправильно классифицированные заявки, которые удовлетворяют условиям

$$(\Sigma_{q \in Qau^*} k_{Ai}(q) > \Sigma_{q \in Qbu^*} k_{Ai}(q)) \text{ AND } (\Sigma_{q \in Qav^*} k_{Ai}(q) > \Sigma_{q \in Qbv^*} k_{Ai}(q)) \text{ AND } (k_{Ai}(r_a) > k_{Ai}(r_b)),$$

$$(\Sigma_{q \in Qau^*} k_{Ai}(q) > \Sigma_{q \in Qbu^*} k_{Ai}(q)) \text{ AND } (\Sigma_{q \in Qav^*} k_{Ai}(q) > \Sigma_{q \in Qbv^*} k_{Ai}(q)) \text{ AND } (k_{Ai}(r_a) < k_{Ai}(r_b)).$$

Результаты анализа совокупности принятых заявок даны в таблице 5.

Таблица 5

Итоги классификации заявок по двум критериям

условие	прав. класс	неправ. класс	разность
$\boldsymbol{Q_2}$ AND Q_1	171	20	151
$\boldsymbol{Q_2}$ AND $\boldsymbol{Q_3}$	170	8	162
$\boldsymbol{Q_2}$ AND Q_4	172	21	151
$\boldsymbol{Q_2}$ AND Q_5	172	23	149

- Выбрать пару критериев Q_u*, Q_v*, обеспечивающих максимальную разность между числом правильно и числом неправильно классифицированных заявок (выделены жирным шрифтом в таблице 5). В нашем случае — это критерии Q_2* и Q_3*, которые включаются в согласованное групповое правила отнесения заявок к уточненному классу $C_a \setminus C_{ac}$.
- Рассмотреть все комбинации пары критериев Q_u*, Q_v* с остальными критериями, входящими в групповое правило отнесения заявок к классу C_a, и найти правильно и неправильно классифицированные заявки. Результаты анализа принятых заявок указаны в таблице 6.

Таблица 6

Итоги классификации заявок по нескольким критериям

условие	прав. класс	неправ. класс	разность
$\boldsymbol{Q_2}$ AND $\boldsymbol{Q_3}$ AND $\boldsymbol{Q_1}$	**169**	**4**	**165**
Q_2 AND Q_3 AND Q_4	170	8	162
Q_2 AND Q_3 AND Q_5	170	7	163
Q_2 AND Q_3 AND Q_1 AND Q_4	169	4	165
$\boldsymbol{Q_2}$ AND $\boldsymbol{Q_3}$ AND $\boldsymbol{Q_1}$ AND $\boldsymbol{Q_5}$	**169**	**3**	**166**

- Выбрать тройку, затем четверку и т. д. критериев Q_u*, Q_v*, Q_w*, … обеспечивающих максимальную разность между числом правильно и числом неправильно классифицированных заявок (выделены жирным шрифтом в таблице 6). В нашем случае — это критерии Q_2*, Q_3*, Q_1*, Q_5*, которые включаются в согласованное групповое правила отнесения заявок к уточненному классу $C_a \setminus C_{ac}$. Заметим, что добавление критерия Q_4* не меняют результаты анализа принятых заявок.
- Включить критерии Q_u*, Q_v*, Q_w*, …, обеспечивающие максимальную разность между числами правильно и неправильно классифицированных заявок, в согласованное групповое правило отнесения заявок к уточненному классу решений $C_a \setminus C_{ac}$ (8)

IF $\langle(\sum_{q \in Qa2*} k_{Ai}(q) > \sum_{q \in Qb2*} k_{Ai}(q))$ AND $(\sum_{q \in Qa3*} k_{Ai}(q) > \sum_{q \in Qb3*} k_{Ai}(q))$

$$\text{AND}(\textstyle\sum_{q \in Qa1}{*k_{Ai}(q)} > \sum_{q \in Qb1}{*k_{Ai}(q)}) \text{ AND } (\sum_{q \in Qa5}{*k_{Ai}(q)} > \sum_{q \in Qb5}{*k_{Ai}(q)})$$
$$\text{AND } (k_{Ai}(r_a) > k_{Ai}(r_b))\rangle, \text{THEN } \langle\text{Object } A_i \in C_a \setminus C_{ac}\rangle.$$

- Построить аналогичным образом с помощью алгоритма 2 согласованное групповое правило отнесения заявок к уточненному классу решений $C_b \setminus C_{bc}$ (9)

$$\text{IF } \langle(\textstyle\sum_{q \in Qa2}{*k_{Ai}(q)} < \sum_{q \in Qb2}{*k_{Ai}(q)}) \text{ AND } (\sum_{q \in Qa3}{*k_{Ai}(q)} < \sum_{q \in Qb3}{*k_{Ai}(q)})$$
$$\text{AND}(\textstyle\sum_{q \in Qa1}{*k_{Ai}(q)} < \sum_{q \in Qb1}{*k_{Ai}(q)}) \text{ AND } (\sum_{q \in Qa5}{*k_{Ai}(q)} < \sum_{q \in Qb5}{*k_{Ai}(q)})$$
$$\text{AND } (k_{Ai}(r_a) < k_{Ai}(r_b))\rangle, \text{THEN } \langle\text{Object } A_i \in C_b \setminus C_{bc}\rangle.$$

- Сформировать класс $C_c = C_{ac} \cup C_{bc}$ заявок, которые имеют противоречивые индивидуальные правила сортировки (10)

$$\text{IF } \langle[(\textstyle\sum_{q \in Qa2}{*k_{Ai}(q)} > \sum_{q \in Qb2}{*k_{Ai}(q)}) \text{ AND } (\sum_{q \in Qa3}{*k_{Ai}(q)} > \sum_{q \in Qb3}{*k_{Ai}(q)})$$
$$\text{AND}(\textstyle\sum_{q \in Qa1}{*k_{Ai}(q)} > \sum_{q \in Qb1}{*k_{Ai}(q)}) \text{ AND } (\sum_{q \in Qa5}{*k_{Ai}(q)} > \sum_{q \in Qb5}{*k_{Ai}(q)})$$
$$\text{AND } (k_{Ai}(r_a) < k_{Ai}(r_b))]$$
$$\text{OR } [\textstyle\sum_{q \in Qa2}{*k_{Ai}(q)} < \sum_{q \in Qb2}{*k_{Ai}(q)}) \text{ AND } (\sum_{q \in Qa3}{*k_{Ai}(q)} < \sum_{q \in Qb3}{*k_{Ai}(q)})$$
$$\text{AND}(\textstyle\sum_{q \in Qa1}{*k_{Ai}(q)} < \sum_{q \in Qb1}{*k_{Ai}(q)}) \text{ AND } (\sum_{q \in Qa5}{*k_{Ai}(q)} < \sum_{q \in Qb5}{*k_{Ai}(q)})$$
$$\text{AND } (k_{Ai}(r_a) < k_{Ai}(r_b))]\rangle, \text{THEN } \langle\text{Object } A_i \in C_c\rangle.$$

- Провести дополнительный анализ таких заявок для устранения противоречий и принятия окончательного решения.

Согласованное групповое правило для отбора лучших проектов, агрегирующее индивидуальные правила сортировки, записывается следующим образом:

– проект обеспечивает достижение всех или одной из основных целей программы; имеет высокую или достаточную перспективность; предлагает совершенно новые или модернизированные подходы к решению поставленных задач; исполнители проекта имеют все необходимые материально-технические ресурсы или имеют объемы ресурсов, достаточные для проведения работ (оценки по критериям q_1^1 или q_1^2; и q_2^1 или q_2^2; и q_3^1 или q_3^2; и q_5^1 или q_5^2).

Таблица 7

Результаты согласованной классификации заявок.

Проекты класса $C_a \setminus C_{ac}$:	A_1, A_1-A_{15}, A_{17}-A_{30}, A_{17}-A_{30}, A_{32}-A_{49}, A_{51}-A_{58}, A_{60}-A_{174}
Проекты класса C_{ac}:	A_{175}, A_{176}, A_{185}, A_{245}, A_{253}
Проекты класса $C_b \setminus C_{bc}$:	A_{177}-A_{184}, A_{186}-A_{244}, A_{246}-A_{252}, A_{254}-A_{259}
Проекты класса C_{bc}:	A_{175}, A_{176}, A_2, A_{16}, A_{31}, A_{50}, A_{59}.

В таблице 7 представлены уточненные классы $C_a \setminus C_{ac}$ (безусловно принятые заявки), $C_b \setminus C_{bc}$ (безусловно отклоненные заявки) и класс противоречиво классифицированных заявок $C_c = C_{ac} \cup C_{bc}$, которые сформированы в соответствии с приведенным выше согласованным групповым правилом для отбора лучших проектов и аналогичным правилом для отбора худших проектов.

Заключение

Проблемы классификации объектов, которые описываются многими количественными и/или качественными признаками, причем каждый из объектов может существовать в нескольких различающихся «экземплярах», являются достаточно трудными для решения и до настоящего момента плохо разработаны. Трудности имеют и содержательные основания (например, процедура «усреднения» качественных признаков некорректна), и формальные причины (например, большая размерность задачи). Преодолеть главные из перечисленных трудностей оказалось возможным при использовании нового теоретического инструментария, основанного на понятии мультимножеств.

В работе изложен новый метод согласованной групповой классификации многопризнаковых объектов, которая агрегирует большое число противоречивых индивидуальных правил сортировки. Предложенный подход был проверен на базе данных, моделирующей результаты экспертной оценки и конкурсного отбора проектов при формировании научно-технической программы [3].

Сопоставление полученных результатов классификации показало их согласованность. Агрегированное групповое правило классификации позволило также выделить наиболее важные для отбора проектов критерии и выявить расхождения в индивидуальных правилах сортировки проектов, применявшихся разными экспертами.

Литература

1. *Ларичев О. И.* Вербальный анализ решений / Под ред. А. Б. Петровского. М.: Наука, 2006. С. 181.
2. *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.
3. *Петровский А. Б.* Пространства множеств и мультимножеств. М: URSS, 2003. С. 246.
4. *Doumpos M., Zopounidis C.* Multicriteria Decision Aid Classification Methods. Dordrecht: Kluwer Academic Publishers, 2002. P. 268.
5. *Petrovsky A. B.* Multi-Attribute Sorting of Qualitative Objects in Multiset Spaces // Multiple Criteria Decision Making in the New Millenium. Berlin: Springer-Verlag, 2001. P. 124–131.