КОМПЬЮТЕРНАЯ МАТЕМАТИКА

УДК 519: 339.1

М.Ф. КАСПШИЦКАЯ, А.А. ПРОВОТАРЬ

РЕШЕНИЕ НЕЧЕТКОЙ ЗАДАЧИ КЛАССИФИКАЦИИ: АЛГОРИТМ И РЕЗУЛЬТАТЫ ВЫЧИСЛИТЕЛЬНОГО ЭКСПЕРИМЕНТА

Рассмотрена одна многокритериальная задача классификации не связанных между собой объектов на основе значений множества признаков. Основное внимание уделено важному случаю, когда признаки выражены нечеткими числами. Формально задача представлена как многокритериальная оптимизационная комбинаторная задача в пространстве разбиений. Предложен приближенный метод ее решения и представлены результаты вычислительного эксперимента для нечеткой задачи.

Введение. В работе рассмотрена нечеткая задача классификации не связанных между собой объектов на основе значений множества признаков. Такие задачи возникают во многих областях деятельности человека, в частности, при классификации биологических объектов [1]. В настоящей работе эта задача представлена в виде оптимизационной комбинаторной задачи. Основное внимание уделено случаю, когда значения признаков являются нечеткими числами. Предложен метод решения и приводятся результаты вычислительного эксперимента.

Пусть задано множество A объектов. Каждый объект $a_i \in A$, $i = 1, \ldots, n$, характеризуется вектором $b_i = [\alpha_{i1}, \alpha_{i1}, \ldots, \alpha_{im},]$; эти векторы образуют прямоугольную матрицу $T = \|\alpha_{ij}\|_{i=1\,j=1}^{n\;\;m}$. Требуется разбить множество A на подмножества без общих элементов, т. е. $A = A_1 \cup A_2 \cup \ldots \cup A_s$ и $A_r \cap A_t = \varnothing$; r, $t = 1, \ldots, s$. Согласно терминологии работы [2], подмножества A_r называются кластерами. В противном случае, когда $A_r \cap A_t \neq \varnothing$, имеем случай классификации. Таким образом, задача кластеризации является частным случаем классификации. Как правило, значения признаков задаются числами 0 и 1. Однако такое упрощение не всегда оправдано, так как более естественными были бы значения признаков из отрезка [0,1]. Например, признак «цвет» может быть задан: 0 белый, 1 – черный, а все оттенки серого, от светло-серого до темно-серого, – дробными числами из отрезка [0,1]. Кроме того, значения признаков, как правило, в действительности являются числами нечеткими, что обусловлено рядом причин: изменчивость признака во времени, нечеткость его определения (особенно при визуальном методе); зависимость признака от среды и т. п.

Таким образом, представление значений признаков в нечеткой форме является более точным и положительно влияет на конечное решение, так как способ постановки задачи определяет и содержание, и методы, и результаты исследований.

Процесс формальной классификации детально описан в работе [1] на примере классификации моллюсков; эта методология может быть использована и в случае объектов другой природы.

При рассмотрении вычислительных аспектов, а также для исследования основной задачи кластеризации [1] целесообразно представить ее в виде комбинаторной оптимизационной задачи, формализованной на множестве разбиений n объектов на k непустых подмножеств (кластеров) [3, 4].

Рассматриваемая оптимизационная задача, согласно вышеизложенному, представляется в следующем виде.

Разбить множество A на минимальное число классов при минимальном количестве сомнительных и неохваченных элементов.

Отметим, что «сомнительные» элементы входят в два или несколько классов, а «неохваченные» – одиночные элементы, которые не входят ни в один из классов. Это так называемая основная формулировка задачи. Здесь имеются три критерия оптимизации без ограничительных условий.

В работе рассмотрена аналогичная задача кластеризации: определить разбиение множества A на минимальное количество классов без общих элементов и минимальным количеством неохваченных элементов.

Как будет ясно из последующего изложения, минимальность (или допустимость) количества неохваченных элементов можно обеспечить соответствующим изменением величины ε (ее смысл раскрыт далее).

Существенную роль при решении этой задачи играет понятие «расстояния» между элементами множества A, введенное с помощью данных матрицы T.

Приведем приближенный алгоритм решения этой задачи. Сделаем некоторые предварительные замечания. Как уже отмечалось, элементы матрицы $T = \left\|\alpha_{ij}\right\|_{i=1\,j=1}^{n\quad m}$ определяют «количество» свойства b_j в элементе a_i. Числа α_{ij} могут быть: 0 или 1; четкие числа; размытые числа, m-мерный вектор $\vec{a}_i$ будем представлять как дискретную функцию на множестве свойств $B = \left\|b_j\right\|_{j=1}^{m}$. Условно назовем ее конфигуративной функцией (полной) $R(a_i, B)$. Таким образом, конфигуративная функция определена на множестве A элементов, подлежащих классификации. «Значением» этой функции является m-значный вектор, координаты которого определяются значениями соответствующей строки матрицы T.

Будем рассматривать также конфигуративные функции сокращенные, т. е. зависящие не от всего множества свойств B, а только от некоторого подмножества $B_0 \subset B$; такие функции обозначим $K(a_i, B_0)$.

Пусть определено расстояние между конфигуративными функциями $K(a_i, B_0)$, $K(a_r, B_0)$. Это может быть сделано, например, следующим способом:

$$d(K(a_i, B_0), K(a_r, B_0)) = \frac{1}{|B_0|} \sum_{b_j \in B_0} \left| \alpha_{ij} - \alpha_{rj} \right|.$$

Кроме того, задан вектор

$$\vec{\varepsilon}(B_0) = \left[\varepsilon_1(B_0), \varepsilon_2(B_0), ..., \varepsilon_{|B_0|}(B_0) \right].$$

Тогда классификация элементов множества A согласно множества признаков B_0 и вектора $\vec{\varepsilon}(B_0)$ можно осуществить следующим образом: элементы a_i и a_r принадлежат одному классу, если

$$\left| a_{i1(B_0)} - a_{r1(B_0)} \right| \le \varepsilon_1(B_0), \left| a_{i2(B_0)} - a_{r2(B_0)} \right| \le \varepsilon_2(B_0), ...,$$
$$\left| a_{i|B_0|(B_0)} - a_{r|B_0|(B_0)} \right| \le \varepsilon_{|B_0|}(B_0). \qquad (1)$$

Таким образом, решение рассматриваемой задачи классификации зависит от вектора $\vec{\varepsilon}(B_0)$. Понятно, что если при определенных значениях координат вектора $\vec{\varepsilon}(B_0)$ задача классификации не имеет решения, то, изменяя соответствующим образом (увеличивая) значения координат вектора $\vec{\varepsilon}(B_0)$, можно достичь того, что задача будет иметь решение.

Процесс классификации согласно вектора $\vec{\varepsilon}(B_0)$ и неравенств (1) можно представить как итерационный: используя информацию вектора $\vec{\varepsilon}(B_0)$ и некоторого элемента $a^{(1)} \in A$, строим классификационную группу A_1; повторяем действия для некоторого элемента $a^{(2)} \in A \backslash A_1$; получаем классификационную группу A_2 и т. д.

Вычисления прекращаются, если исчерпаны все элементы множества A, кроме тех, которые оказались неохваченными.

Приведенные в этой работе результаты вычислительного эксперимента касаются, в основном, случая нечетких данных: данные матрицы T представлены нечеткими числами, т. е. α_{ij} являются значениями функции принадлежности

μ_{ij}, где $x \in U_{ij}$, U_{ij} – носитель нечеткой величины α_{ij}. Условия (1) в этом случае будут определяться на множестве нечетких чисел.

Используемые в вычислительном эксперименте числа являются числами L-R типа. Такие числа записываются в виде: $C = (m_c, \gamma_c, \delta_c)_{LR}$. Здесь m_c – действительное четкое число, называемое средним значением нечеткого числа C, $\mu(m_c) = 1$; γ_c – положительное действительное четкое число, называемое левосторонним отклонением; соответственно δ_c – положительное действительное число, называемое правосторонним отклонением.

Далее, согласно определениям действий над нечеткими числами, что вытекают из принципа обобщения [5], имеем:

разность $C_{LR} \Theta D_{LR} = (m_C - m_D, \gamma_C - \delta_D, \delta_C + \gamma_D)$;

абсолютная величина

$$\mu_{|C|}(x) = \begin{cases} \max(\mu_C(x), \mu_C(-x)), & \text{для} \quad x \geq 0 \\ 0, \quad \text{для} \quad x < 0 \end{cases};$$

сравнение двух $L - R$ чисел: $C \leq D$, если $m_C \leq m_D$; $\gamma_C \geq \delta_D$; $\delta_C \leq \delta_D$. В дальнейшем предполагаем, что используемые нечеткие числа – треугольные.

Алгоритм работы программы. Дана таблица данных: $T = \left\| \alpha_{ij} \right\|_{i=1\, j=1}^{n \quad m}$, где α_{ij} – количество свойств $b_j \in B$ $(j = 1, \ldots, m)$ в элементе $a_i \in A$.

Элементов (n) – 10, свойств (m) – 5, (n = 10; m = 5), функции принадлежности – треугольные ($C = (m_C, \gamma_C, \delta_C)_{LR}$). Здесь m_C – действительное четкое число, называющееся средним значением нечеткого числа C; γ_C – положительное действительное четкое число, называющееся левосторонним отклонением нечеткого числа C; δ_C – положительное действительное четкое число, называющееся правосторонним отклонением нечеткого числа C.

Кроме того, задан вектор $\vec{\varepsilon}(B_0) = \left[\varepsilon_1(B_0), \varepsilon_2(B_0), \ldots, \varepsilon_{|B_0|}(B_0)\right]$.

Шаги работы программы классификации

Шаг 1. Выбираем произвольный элемент табл. 1 данных $a^0 \in A$ (в нашем случае элемент a_1b_1) и строим на его основе множество A_0, включающее все те элементы, удовлетворяют условию:

$$| a_{i(B_0)} - a_{r(B_0)} | \leq \varepsilon_1(B_0), \ldots, | \alpha_{i(B_0)} - \alpha_{r(B_0)} | \leq \varepsilon_{|B_0|}(B_0) \ (*).$$

Поиск элементов начинается с первого элемента a_1b_1 и продолжается до елемента a_nb_1.

Шаг 2. Все элементы a_1b_1, ..., a_nb_1, удовлетворяюющие условию (*) проверяются на выполнение условия (*) с элементами, которые имеют другие свойства (b) a_ib_i, ..., a_ib_n.

Шаг 3. Среди полученной совокупности множеств выбираем наибольшее множество (определяется по количеству частичных множеств, принадлежащих данному множеству), которое и образует класс $\tilde{A}_i$.

Шаг 4. Элементы, образующие классификационную группу, исключаются из дальнейшего рассмотрения.

Шаг 5. Признаком окончания процесса шагов 1 – 4 будет пустота всех следующих (частичных) множеств. Объединение всех частичных множеств $A(a^0)$, созданных на основе елемента $a^0 \in A$, будет входить в определенный класс $\tilde{A}_0$, т. е. $A(a^0)$, $A(a^0) \in \tilde{A}_0$.

Задача формальной классификации проводилась по сочетаниям двух признаков.

Реализация программы выполнена на языке программирования С++.

ТАБЛИЦА 1. Данные в числовом представлении

$a_i \backslash b_j$	b_1	b_2	b_3	b_4	b_5
a_1	1.15; 0.5; 0.5	2.35; 0.3; 0.4	3.48; 1; 1	1.75; 0.1; 0.2	3.14; 0.5; 0.5
a_2	0.59; 1; 0.5	2.54; 1; 1	5.45; 1.5; 1	0.97; 0.3; 0.2	3.53; 0.5; 0.5
a_3	2.22; 1.5; 1	1.34; 0.5; 0.5	2.38; 0.5; 0.3	1.45; 0.5; 0.5	2.15; 0.5; 0.5
a_4	4.32; 1; 0.5	0.58; 0.1; 0.1	1.54; 0.5; 0.5	3.12; 1; 2	2.53; 0.5; 0.5
a_5	6.21; 0.3; 0.7	0.76; 0.2; 0.2	2.38; 1; 1	1.49; 0.5; 0.1	0.59; 0.5; 0.5
a_6	2.32; 0.7; 0.1	0.87; 0.3; 0.2	1.35; 0.5; 0.5	2.63; 0.3; 0.3	1.89; 0.1; 0.3
a_7	3.43; 0.5; 0.1	0.98; 0.5; 0.5	4.25; 0.1; 0.8	4.19; 0.2; 0.2	0.49; 0.3; 0.3
a_8	4.35; 0.6; 0.3	3.21; 1; 1	3.82; 1; 1.5	4.54; 0.5; 0.5	4.73; 1; 1
a_9	1.55; 1; 1	2.27; 0.5; 1	2.65; 1; 0.5	1.56; 0.1; 0.2	3.28; 1; 1.5
a_{10}	1.52; 0.1; 0.2	1.99; 0.7; 0.8	2.53; 0.7; 0.5	5.48; 0.2; 0.2	2.13; 1; 1

Результаты для четких данных получены лишь при некотором значении $\varepsilon>0$.

В табл. 2 – 6 приведены результаты работы программы для нечетких данных.

ТАБЛИЦА 2. При $\varepsilon = 0.5$

Группа	Элементы	Свойства
1	3; 6	1; 5
2	5; 7	2; 5
3	9; 10	2; 3

Классы созданные элементами: 3, 5, 6, 7, 9, 10.
Неохваченные элементы: 1, 2, 4, 8.

ТАБЛИЦА 3. При $\varepsilon = 1.0$

Группа	Элементы	Свойства
1	3; 6	1; 5
2	9; 10	1; 2
3	1; 2	2; 5

Классы созданные элементами: 1, 2, 3, 6, 9, 10.
Неохваченные элементы: 4, 5, 7, 8.

ТАБЛИЦА 4. Расширенная таблица (1) (n = 30; m = 5)

$a_i \backslash b_j$	b_1	b_2	b_3	b_4	b_5
1	2	3	4	5	6
a_1	1.15; 0.5; 0.5	2.35; 0.3; 0.4	3.48; 1; 1	1.75; 0.1; 0.2	3.14; 0.5; 0.5
a_2	0.59; 1; 0.5	2.54; 1; 1	5.45; 1.5; 1	0.97; 0.3; 0.2	3.53; 0.5; 0.5
a_3	2.22; 1.5; 1	1.34; 0.5; 0.5	2.38; 0.5; 0.3	1.45; 0.5; 0.5	2.15; 0.5; 0.5
a_4	4.32; 1; 0.5	0.58; 0.1; 0.1	1.54; 0.5; 0.5	3.12; 1; 2	2.53; 0.5; 0.5
a_5	6.21; 0.3; 0.7	0.76; 0.2; 0.2	2.38; 1; 1	1.49; 0.5; 0.1	0.59; 0.5; 0.5
a_6	2.32; 0.7; 0.1	0.87; 0.3; 0.2	1.35; 0.5; 0.5	2.63; 0.3; 0.3	1.89; 0.1; 0.3
a_7	3.43; 0.5; 0.1	0.98; 0.5; 0.5	4.25; 0.1; 0.8	4.19; 0.2; 0.2	0.49; 0.3; 0.3
a_8	4.35; 0.6; 0.3	3.21; 1; 1	3.82; 1; 1.5	4.54; 0.5; 0.5	4.73; 1; 1
a_9	1.55; 1; 1	2.27; 0.5; 1	2.65; 1; 0.5	1.56; 0.1; 0.2	3.28; 1; 1.5
a_{10}	1.52; 0.1; 0.2	1.99; 0.7; 0.8	2.53; 0.7; 0.5	5.48; 0.2; 0.2	2.13; 1; 1
a_{11}	2.21; 0.5; 0.5	0.98; 0.3; 0.4	4.27; 1; 1	1.56; 0.1; 0.2	1.02; 0.5; 0.5
a_{12}	3.19; 1; 0.5	1.25; 1; 1	3.94; 1.5; 1	3.47; 0.3; 0.2	0.89; 0.5; 0.5
a_{13}	1.98; 1.5; 1	0.34; 0.5; 0.5	3.64; 0.5; 0.3	2.56; 0.5; 0.5	2.45; 0.5; 0.5
a_{14}	5.45; 1; 0.5	2.51; 0.1; 0.1	2.58; 0.5; 0.5	3.12; 1; 2	4.09; 0.5; 0.5
a_{15}	4.29; 0.3; 0.7	2.13; 0.2; 0.2	4.32; 1; 1	4.32; 0.5; 0.1	3.24; 0.5; 0.5
a_{16}	1.19; 0.7; 0.1	0.59; 0.3; 0.2	5.03; 0.5; 0.5	1.34; 0.3; 0.3	1.78; 0.1; 0.3
a_{17}	2.51; 0.5; 0.1	0.32; 0.5; 0.5	1.89; 0.1; 0.8	1.65; 0.2; 0.2	2.93; 0.3; 0.3

a_{18}	8.36; 0.6; 0.3	0.56; 1; 1	3.45; 1; 1.5	4.23; 0.5; 0.5	4.23; 1; 1
a_{19}	3.83; 1; 1	1.28; 0.5; 1	2.72; 1; 0.5	3.26; 0.1; 0.2	2.45; 1; 1.5
a_{20}	1.23; 0.1; 0.2	2.35; 0.7; 0.8	4.05; 0.7; 0.5	2.84; 0.2; 0.2	1.45; 1; 1

Окончание табл. 4

1	2	3	4	5	6
a_{21}	3.65; 0.5; 0.5	3.08; 0.3; 0.4	1.89; 1; 1	5.09; 0.1; 0.2	3.45; 0.5; 0.5
a_{22}	3.56; 1; 0.5	2.56; 1; 1	2.94; 1.5; 1	3.76; 0.3; 0.2	1.18; 0.5; 0.5
a_{23}	2.81; 1.5; 1	1.45; 0.5; 0.5	1.99; 0.5; 0.3	4.48; 0.5; 0.5	2.34; 0.5; 0.5
a_{24}	1.94; 1; 0.5	0.76; 0.1; 0.1	4.45; 0.5; 0.5	2.45; 1; 2	1.12; 0.5; 0.5
a_{25}	2.76; 0.3; 0.7	0.68; 0.2; 0.2	3.39; 1; 1	1.77; 0.5; 0.1	0.99; 0.5; 0.5
a_{26}	1.58; 0.7; 0.1	1.09; 0.3; 0.2	3.09; 0.5; 0.5	2.65; 0.3; 0.3	4.11; 0.1; 0.3
a_{27}	3.02; 0.5; 0.1	3.05; 0.5; 0.5	2.76; 0.1; 0.8	2.81; 0.2; 0.2	2.44; 0.3; 0.3
a_{28}	2.56; 0.6; 0.3	2.05; 1; 1	4.12; 1; 1.5	3.47; 0.5; 0.5	2.87; 1; 1
a_{29}	4.65; 1; 1	1.98; 0.5; 1	2.09; 1; 0.5	1.43; 0.1; 0.2	1.58; 1; 1.5
a_{30}	5.01; 0.1; 0.2	0.59; 0.7; 0.8	3.99; 0.7; 0.5	2.76; 0.2; 0.2	3.17; 1; 1

ТАБЛИЦА 5. Результаты работы программы при $\varepsilon = 0.5$

Группа	**Элементы**	**Свойства**
1	7; 12; 18; 26	3; 4
2	3; 6; 23; 28	1; 5
3	9; 14; 22; 27	2; 3
4	5; 10; 11; 29	2; 5
5	13; 19	4; 5

Классы созданные элементами: 3, 5, 6, 7, 9, 10, 11, 12, 13, 14, 18, 19, 22, 23, 26, 27, 28, 29.

Неохваченные элементы: 1, 2, 4, 8, 15, 16, 17, 20, 21, 24, 25, 30.

ТАБЛИЦА 6. При замене всех γ и δ на 0.5 ($\gamma = 0.5$, $\delta = 0.5$), результаты работы программы при $\varepsilon = 0.5$

Группа	Элементы	Свойства
1	5; 7; 10; 12; 24; 29	2; 5
2	8; 13; 14; 19; 27; 30	4; 5
3	3; 6; 17; 23	1; 5
4	9; 16	1; 4

5	11; 25	2; 5
6	18; 26	3; 4

Классы созданные элементами: 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 16, 17, 18, 19, 23, 24, 25, 26, 27, 29, 30.

Неохваченные элементы: 1, 2, 4, 15, 20, 21, 22, 28.

Заключение. В результате реализации алгоритма решения нечеткой задачи классификации найден минимальный набор классификационных групп в которых отсутствуют сомнительные и неохваченные элементы. В случае четких данных, результаты достигнуты путем изменения величины ε. Таким образом, представление значений признаков в нечеткой форме является более точным и положительно влияет на конечное решение.

М.Ф. Каспшицька, О.О. Провотар

РІШЕННЯ НЕЧІТКОЇ ЗАДАЧІ КЛАСИФІКАЦІЇ:
АЛГОРИТМ ТА РЕЗУЛЬТАТИ ОБЧИСЛЮВАЛЬНОГО ЕКСПЕРИМЕНТУ

Розглянута одна багатокритеріальна задача класифікації не зв'язаних між собою об'єктів на основі значень множини ознак. Основна увага приділена важливому випадку, коли ознаки виражені нечіткими числами. Формально задача представлена як багатокритеріальна оптимізаційна комбінаторна задача в просторі розбиттів. Запропоновано наближений метод її розв'язання та наведені результати обчислювального експерименту для нечіткої задачі.

M.F. Kaspshizka, A.A. Provotar

SOLUTION OF FUZZY CLASSIFICATION PROBLEM:
ALGORITHMS AND RESULTS OF COMPUTATIONAL EXPERIMENTS

A multi-criteria optimization problem of unrelated objects based on the values of set of features is examined. Main attention is devoted to important case, when the features are expressed by fuzzy numbers. Formally, the problem is presented as a multi-criteria combinatorial optimization problem in a space of partitions. An approximate method for its solution is proposed and results of computational experiment for the fuzzy problem are presented.

1. *Дородницин А.А., Каспшицкая М.Ф., Сергиенко И.В.* Об одном подходе к формализации классификации // Кибернетика. – 1976. – № 6. – С. 132 – 140.
2. *Дюран Б., Оделл П.* Кластерный анализ. – М.: Статистика, 1977. – 240 с.
3. *Каспшицкая М.Ф., Парасюк И.Н.* О некоторых классах размытых задач классификации: формализация, методы решения // Компьютерная математика. – 2004. – № 1. – С. 73 – 90.
4. *Сергиенко И.В., Парасюк И.Н., Каспшицкая М.Ф.* Об одной нечеткой задаче многопараметрического выбора оптимальных решений // Кибернетика и системный анализ. – 2003. – № 2. – С. 3–14.
5. *Леоненков А.В.* Нечеткое моделирование в среде MATLAB u fuzzy TECH. – Санкт-Петербург: БХВ – Петербург, 2005. – 716 с.

Получено 18.09.2013

Об авторах:

Каспшицкая Мария Фадеевна,

кандидат физико-математических наук, старший научный сотрудник
Института кибернетики имени В.М. Глушкова НАН Украины,

Провотарь Алексей Алексеевич,
младший научный сотрудник
Института кибернетики имени В.М. Глушкова НАН Украины.