УДК 004.825

А.Н. Шушура, В.С. Аниканов
Донецкий государственный институт искусственного интеллекта, Украина

Кластеризация данных с использованием нечетких отношений

В статье на основе существующего алгоритма кластеризации данных на основе нечетких отношений разрабатывается его модифицированная версия. В рамках решения этой задачи проведено исследование применения нечетких отношений в задаче кластеризации, представлены модификации существующего алгоритма, проанализировано функционирование измененного алгоритма на контрольных примерах. Получил дальнейшее развитие метод кластеризации на основе нечетких отношений, что нашло выражение во введении значимости метрик при оценке близости объектов, и предложена процедура формирования единственного решения в условиях отсутствия данных для самостоятельного выбора уровня квазиэквивалентности. Результаты исследования могут быть использованы при кластеризации произвольных данных, не зависящих от рода задачи, учитывая при этом характер задачи путем введения коэффициентов.

В настоящее время растет спрос предприятий и организаций на системы управления и анализа, способные не просто накапливать данные, а добывать новые знания в соответствии со стандартами Data Mining. Одной из основных задач интеллектуального анализа является задача кластеризации.

Ведущими исследователями, сделавшими наибольший вклад в решение данной задачи, по праву считаются Фишер, Макнаотон, Кауфман, Роузеув. В текущем состоянии задача кластеризации имеет большое количество решений. Разносторонние случаи и подходы к ее решению рассмотрены в [1]. Существующие методы условно можно разделить на методы, не использующие нечеткую логику и использующие ее.

Недостатками методов, не использующих нечеткую логику, являются:

– использование для анализа центров кластеров, а не всех данных;
– отнесение объектов только к одному из кластеров, а не его частичное распределение по классам.

Существующие методы на основе нечеткой логики и нечетких отношений исправляют эти недостатки, но имеют ряд своих. К недостаткам существующих «нецентроидных» методов на основе нечеткой логики можно отнести использование меры близости, не учитывающей близость объектов между собой и всем множеством, а также отсутствие значимости метрик.

В данной статье проводится исследование и разработка модификации алгоритма кластеризации данных на основе нечетких отношений [1], способного учитывать разные меры близости объектов относительно всего множества и друг друга, а также учитывающего специфику метрик путем введения значимости каждой из них.

В рамках решения данной задачи необходимо:

– исследовать использование нечеткой логики для решения задач кластеризации;
– разработать алгоритм разбиения на кластеры на основе нечеткой логики, учитывающий близость объектов относительно всего кортежа данных и значимость метрик;
– создать алгоритм построения единого решения на основе сформированной шкалы в условиях отсутствия данных для выбора уровня α-квазиэквивалентности;

– провести анализ работы алгоритмов (сравнение результатов для полученной шкалы и единого решения).

Использование нечеткой логики и нечетких отношений [2], [3] для решения задачи кластеризации позволяет применять такой подход, который учитывал бы взаимосвязь между точками, а не точками и центрами кластеров. Кроме того, он позволяет учитывать значимость метрик и характер их влияния на конечный результат.

Представленный в данной статье алгоритм разбиения на классы базируется на алгоритме кластеризации на основе нечетких отношений [1].

Алгоритм в качестве меры близости двух объектов должен учитывать как близость объектов между собой, так и их расположение относительно всей группы данных. Следует также учитывать значимость метрик для учета их природного характера.

Рассмотрим алгоритм. Построим для каждого образца данных нормальную меру сходства по следующей формуле:

$$\mu_{x_q}(x_i) = 1 - d(x_q, x_i), q, i = \overline{1,Q}, \tag{1}$$

где d – расстояние, учитывающее значимость метрик s_t и их масштабируемость h_t, рассчитываемое по следующим формулам:

$$d(x_i, x_j) = \frac{\sqrt{\sum_{t=1}^{m} k_t \left(\frac{x_{it} - x_{jt}}{\max\limits_{q,r=1,Q} (x_{qt} - x_{rt})} \right)^2}}{\sum_{t=1}^{m} k_t}, i,\ j = \overline{1,Q}, \tag{2}$$

$$2k_t = \frac{s_t}{h_t}. \tag{3}$$

Данная модификация позволит масштабировать значение $\mu_{x_q}(x_i)$ не только относительно всех метрик, а по каждой из них с учетом значимости метрик как критерия близости классов.

Введем относительно каждого образца на основании нормальной меры сходства относительную меру сходства для пар образцов:

$$\xi_{x_q}(x_i, x_j) = 1 - |\, \mu_{x_q}(x_i) - \mu_{x_q}(x_j)\,|, i, j, q = \overline{1,Q}. \tag{4}$$

Построим меру сходства образцов данных на всем множестве, которая позволит учитывать как близость объектов между собой, так и их расположение относительно всей группы данных благодаря сравнению на всем множестве элементов:

$$\xi(a,b) = T(\xi_{x_1}(a,b,...,\xi_{x_Q}(a,b)) = \frac{r_1 \min\limits_{i=1,\overline{Q}} \xi_{x_i}(a,b) + r_2 \mu_{x_a}(x_b)}{2(r_1 + r_2)}, a, b \in X. \tag{5}$$

Данная модификация дает возможность учитывать близость объектов на всем множестве с учетом их собственной близости $\mu_{x_a}(x_b)$, что позволяет искать более близкие объекты, входящие в состав кластера. В качестве параметров r_1, r_2 могут выступать любые константы в зависимости от задачи. Для сравнения с существующим алгоритмом [1] в рассмотренных примерах примем их за 1.

Отношение, полученное по формуле (5) на всем множестве, является отношением α-толерантности.

Построим транзитивное замыкание отношения меры сходства образцов данных на множестве *X*, используя определение о транзитивном замыкании нечеткого отношения R = X o Y и договоренности об используемых в алгоритме *T*-норме и *T*-конорме [1], [3]. То есть каждый элемент построенного отношения будет иметь степень принадлежности, вычисляемую по формулам:

$$R_\xi^1 = R_\xi, \tag{6}$$

$$R_\xi^q = R_\xi^{q-1} \circ R_\xi, \tag{7}$$

$$R_\xi^{|X|} = R_\xi^Q, r_{ij}^{|X|} = r_{ij}^Q, \quad q = \overrightarrow{2,Q}. \tag{8}$$

Данное отношение $R_\xi^{|X|}$ есть отношение α-квазиэквивалентности.

Обобщенная блок-схема работы описанного алгоритма приведена на рис. 1. В результате работы алгоритма мы получим отношение α-квазиэквивалентности, которое позволит нам построить шкалы α-квазиэквивалентности на множестве элементов, что и является множеством решения [1].

Построим для отношения α-квазиэквивалентности $R_\xi^{|X|}$ шкалу α-квазиэквивалентности как множество различных элементов отношения $R_\xi^{|X|}$.

В результате работы алгоритма получим шкалу α-квазиэквивалентности, которая позволяет выбрать решения в зависимости от степени близости – уровня α-квазиэквивалентности. Таким образом, алгоритм дает множество решений, правильное из которых может быть выбрано в зависимости от ситуации и характера объектов.

В случаях, когда нет критерия выбора единственного решения, для его формирования, не зависящего от уровня α-квазиэквивалентности, предлагается алгоритм, осуществляющий нахождения решения на шкале α-квазиэквивалентности.

Исходными данными являются объекты и шкала α-квазиэквивалентности.

Помещаем первый объект в первый класс. Далее в цикле выбираем все последующие объекты для определения их класса. Для каждого объекта проверяем его на принадлежность каждому из предшествующих классов. В качестве критерия отнесения объекта к классу считаем условие:

$$S_{eq} >= k * S_{neq}, \tag{9}$$

где S_{eq} – сумма отношений α-уровней на общую сумму объектов, которые принадлежат классу, S_{neq} – сумма отношений остальных объектов, k – параметр (выбирается эмпирическим путем в зависимости от рода задачи).

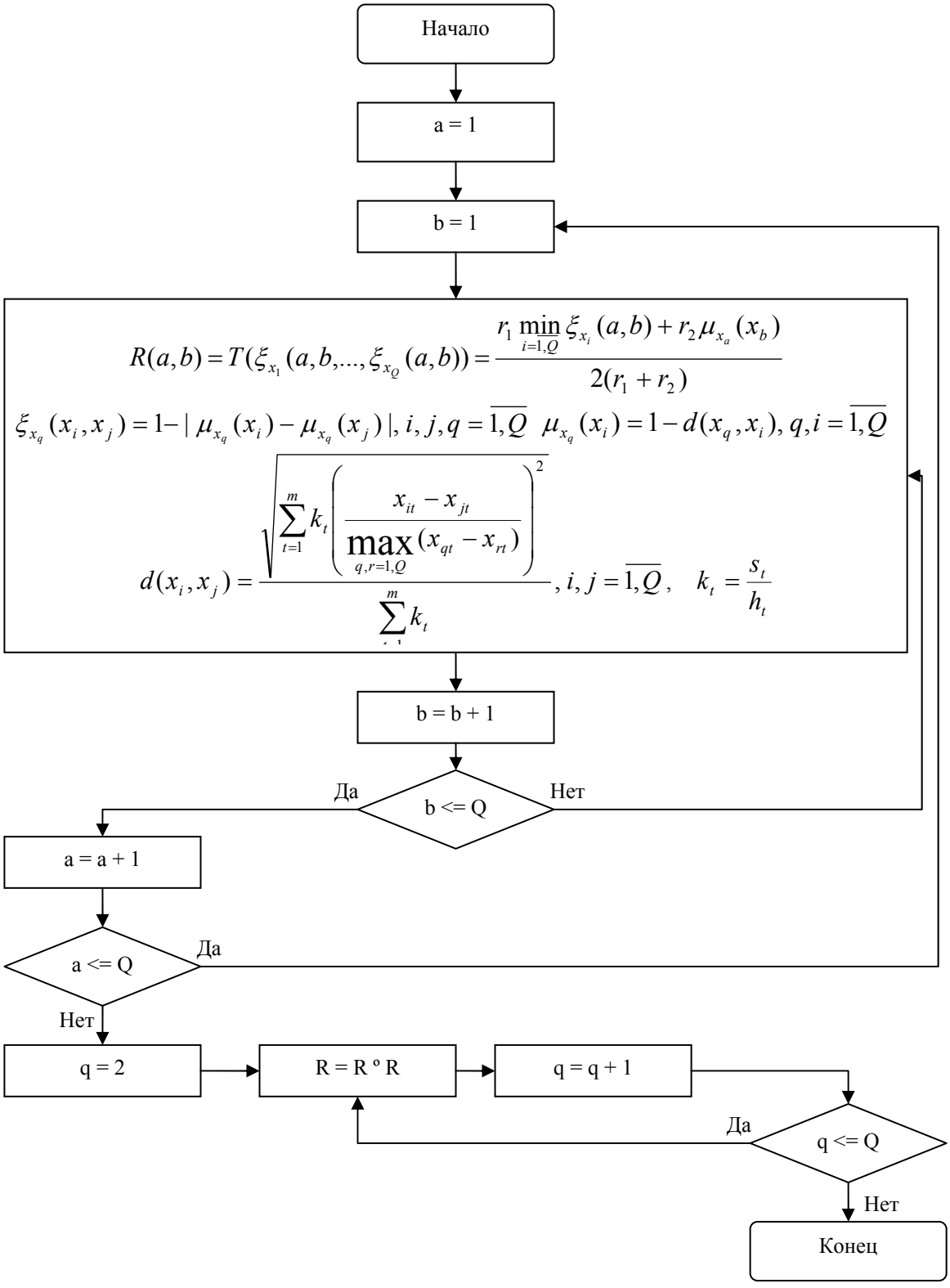

Рисунок 1 – Алгоритм построения шкалы α-квазиэквивалентности на множестве элементов X

В рамках анализа результатов работы алгоритмов был рассмотрен ряд тестовых примеров. Ниже приведен один из них, взятый для сравнения из [1].

Исходными данными являются точки в декартовом пространстве (табл. 1).

Таблица 1

№ точки	абсцисса	ордината
0	0,11	1
1	0,09	1,1
2	0,1	0,9
3	0,12	0,9
4	0,4	2
5	0,5	2
6	0,5	3
7	0,4	3
8	0,42	2,5
9	0,48	2,5
10	0,45	2,8
11	0,45	2,2
12	0,35	2,7
13	3,4	0
14	3,6	0
15	3,6	1
16	3,4	1
17	3,5	0,5

При решении задачи была получена следующая шкала α-квазиэквивалентности (табл. 2).

Таблица 2

№	α-уровень	Решение
1	1	class 1: 0 class 2: 1 class 3: 2 class 4: 3 class 5: 4 class 6: 5 class 7: 6 class 8: 7 class 9: 8 class 10: 9 class 11: 10 class 12: 11 class 13: 12 class 14: 13 class 15: 14 class 16: 15 class 17: 16 class 18: 17
2	0,997150997150997	class 1: 0 class 2: 1 class 3: 2 3 class 4: 4 class 5: 5 class 6: 6 class 7: 7 class 8: 8 class 9: 9 class 10: 10 class 11: 11 class 12: 12 class 13: 13 class 14: 14 class 15: 15 class 16: 16 class 17: 17
3	0,991452991452991	class 1: 0 class 2: 1 class 3: 2 3 class 4: 4 class 5: 5 class 6: 6 class 7: 7 class 8: 8 9 class 9: 10 class 10: 11 class 11: 12 class 12: 13 class 13: 14 class 14: 15 class 15: 16 class 16: 17
4	0,985754985754986	class 1: 0 class 2: 1 class 3: 2 3 class 4: 4 5 class 5: 6 7 class 6: 8 9 class 7: 10 class 8: 11 class 9: 12 class 10: 13 class 11: 14 class 12: 15 class 13: 16 class 14: 17
5	0,983272567976937	class 1: 0 2 3 class 2: 1 class 3: 4 5 class 4: 6 7 class 5: 8 9 class 6: 10 class 7: 11 class 8: 12 class 9: 13 class 10: 14 class 11: 15 class 12: 16 class 13: 17
6	0,983272567976937	class 1: 0 2 3 class 2: 1 class 3: 4 5 class 4: 6 7 class 5: 8 9 class 6: 10 class 7: 11 class 8: 12 class 9: 13 class 10: 14 class 11: 15 class 12: 16 class 13: 17
7	0,983091582125716	class 1: 0 1 2 3 class 2: 4 5 class 3: 6 7 class 4: 8 9 class 5: 10 class 6: 11 class 7: 12 class 8: 13 class 9: 14 class 10: 15 class 11: 16 class 12: 17
8	0,978075169131361	class 1: 0 1 2 3 class 2: 4 5 class 3: 6 7 class 4: 8 9 class 5: 10 12 class 6: 11 class 7: 13 class 8: 14 class 9: 15 class 10: 16 class 11: 17
9	0,971509971509972	class 1: 0 1 2 3 class 2: 4 5 class 3: 6 7 class 4: 8 9 class 5: 10 12 class 6: 11 class 7: 13 14 class 8: 15 16 class 9: 17
10	0,965914207962535	class 1: 0 1 2 3 class 2: 4 5 11 class 3: 6 7 10 12 class 4: 8 9 class 5: 13 14 class 6: 15 16 class 7: 17
11	0,965207154151708	class 1: 0 1 2 3 class 2: 4 5 11 class 3: 6 7 8 9 10 12 class 4: 13 14 class 5: 15 16 class 6: 17
12	0,949817703930812	class 1: 0 1 2 3 class 2: 4 5 6 7 8 9 10 11 12 class 3: 13 14 class 4: 15 16 class 5: 17
13	0,91545791062858	class 1: 0 1 2 3 class 2: 4 5 6 7 8 9 10 11 12 class 3: 13 14 15 16 17
14	0,843634833353529	class 1: 0 1 2 3 4 5 6 7 8 9 10 11 12 class 2: 13 14 15 16 17
15	0,554540843494298	class 1: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

Рассмотрим случаи минимального и максимального α-уровня.

Если α-уровень максимален, то близость объектов может быть только к самому себе и, следовательно, все классы обособлены (рис. 2).

Если α-уровень минимален, то любые объекты близки друг к другу, а значит, все множество является одним классом (рис. 3).

Анализируя шкалу, необходимо заметить, что она похожа на рассмотренную в аналоге [1], хотя и существуют значительные отличия (пример: данный алгоритм группирует объекты 3 и 4 в один класс, а 1 и 2 в разные при большом α-уровне, что наглядно видно на рисунке, а исходный алгоритм не имеет такого решения). Данное отличие позволяет получить более точные решения за счет учета близости объектов друг к другу. В ходе проведения анализа рассматриваются уровни значимости метрик, равные между собой (для демонстрации схожести как частного случая разработанного алгоритма). В случаях же неравных уровней значимости результаты будут отличны и ощутимы только на примере практического решения, так как плохо демонстрируются в декартовом пространстве.

Для рассмотренного примера построено решение с помощью единого алгоритма (рис. 4), который анализирует частоты близости объектов во всевозможных решениях шкалы α-квазиэквивалентности. Решение, полученное с его помощью, показывает довольно приемлемые результаты, хотя и есть некоторые неточности.

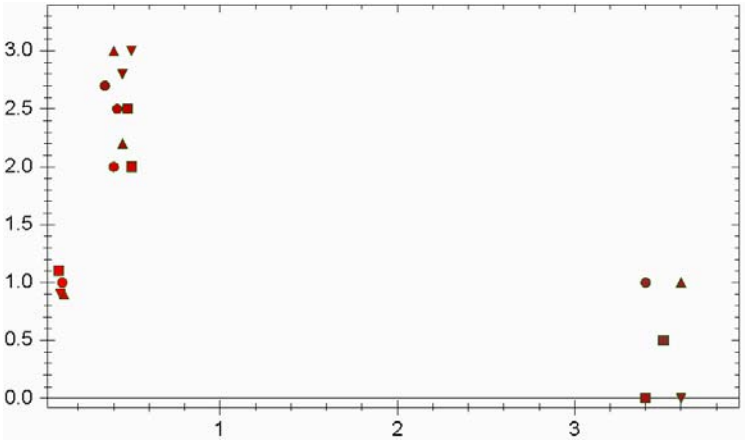

Рисунок 2 – Решение при α-уровне, равном максимуму (1)

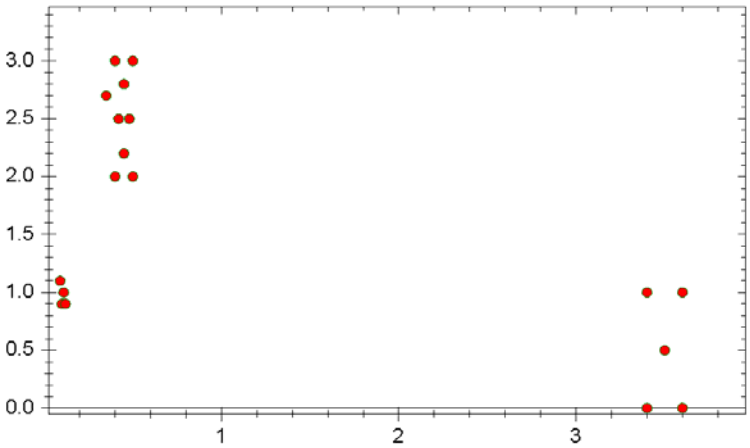

Рисунок 3 – Решение при α-уровне, равном минимуму (0,554540843494298)

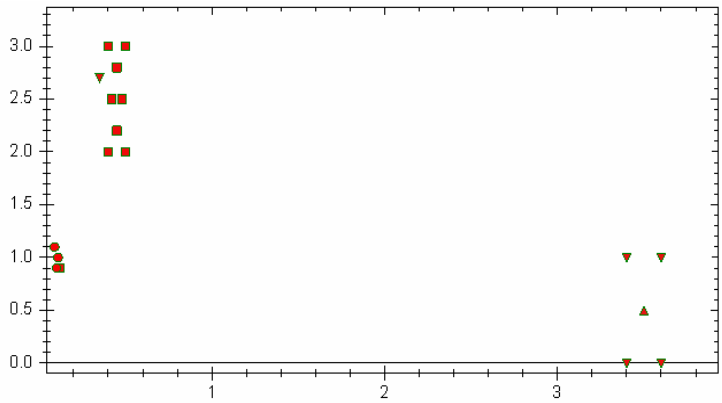

Рисунок 4 – Результат алгоритма единого решения

Выводы

Разработанные модификации позволяют улучшить алгоритм разбиения множества на нечеткие кластеры. Они дают возможность создать шкалу α-квазиэквивалентности, сгруппированную на больших значениях уровня α-квазиэквивалентности, учитывать при построении кластеров и близость объектов друг к другу, и их расположение относительно всех данных («похожесть», близость на множестве). Разработанный алгоритм позволяет учитывать характер задачи и вносить изменения за счет значимости метрик и параметров, что не было ранее реализовано ни в одном из алгоритмов на основе нечетких отношений.

Выбранный алгоритм поиска единого решения заменяет выбор уровня α-квазиэквивалентности на поиск обобщенного варианта разбиения, что способствует более качественному формированию кластеров в условиях отсутствия данных для самостоятельного выбора уровня. Следует отметить, что более приемлемым решением остается шкала квазиэквивалентности, а прибегать к решению с помощью единого алгоритма необходимо только в случае отсутствия данных для выбора α-уровня.

Литература

1. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. – СПб.: Петербург, 2004. – 336 с.
2. Кузьмин В.Б. Построение нечетких групповых отношений. – М.: Наука, 1988.
3. Леоненков А.В. Нечеткое моделирование в среде MATLAB и fuzzyTECH. – СПб.: БХВ-Петербург, 2003. – 276 с.

О.М. Шушура, В.С. Аніканов

Кластеризація даних з використанням нечітких відношень

У статті на основі існуючого алгоритму кластеризації даних на основі нечітких відношень розробляється його модифікована версія. У рамках вирішення цієї задачі проведено дослідження застосування нечітких відношень у задачі кластеризації, представлено модифікації існуючого алгоритму, проаналізовано функціонування зміненого алгоритму на контрольних прикладах. Одержав подальший розвиток метод кластеризації на основі нечітких відношень, що знайшло вираження у введенні значимості метрик при оцінці близькості об'єктів, і запропоновано процедуру формування єдиного рішення в умовах відсутності даних для самостійного вибору рівня квазіеквівалентності. Результати дослідження можуть бути використані при кластеризації довільних даних, які не залежать від роду задачі, враховуючи при цьому характер задачі шляхом уведення коефіцієнтів.

Статья поступила в редакцию 19.07.2005.