АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ

З. Б. Мингликулов

Центр разработки программных продуктов и аппаратно-программных комплексов
при Ташкентском университете информационных технологий,
100084, Республика Узбекистан

УДК 519.8

Предложен алгоритм кластеризации, основанный на нечетко-логическом выводе. Приведен сравнительный анализ результатов решения модельной задачи предлагаемым алгоритмом и нечетким алгоритмом c-means.

Ключевые слова: нечеткое множество, кластеризация, алгоритм, степень доверенности, нечетко-логические выводы.

Clustering algorithm based on fuzzy-logical conclusions is proposed in this paper. Comparative analysis of the results of solving of model problem by the proposed algorithm and c-means fuzzy algorithm is presented.

Keywords: fuzzy set, clustering, algorithm, the degree reliability, fuzzy-logical conclusions.

Введение. Одним из направлений обработки данных различной структуры и свойств является кластеризация. Существует множество методов кластеризации, которые можно классифицировать как четкие и нечеткие. Четкие методы кластеризации разбивают исходное множество объектов X_i на несколько непересекающихся подмножеств. При этом любой объект из X_i принадлежит только одному кластеру. Нечеткие методы кластеризации позволяют одному и тому же объекту принадлежать одновременно нескольким (или даже всем) кластерам, но с различной степенью принадлежности. Нечеткая кластеризация во многих ситуациях более "естественна", чем четкая, например, для объектов, расположенных на границе кластеров [1, 2, 5, 6].

Кластеризация — это разбиение элементов некоторого множества на группы на основе их схожести. Задача кластеризации состоит в разбиении объектов из X_i на несколько подмножеств (кластеров), в которых объекты более схожи между собой, чем с объектами из других кластеров. В метрическом пространстве "схожесть" обычно определяют через расстояние. Большинство алгоритмов кластеризации не опирается на традиционные для статистических методов допущения; они могут использоваться в условиях почти полного отсутствия информации о законах распределения данных [4].

Методы кластеризации также классифицируются по тому, определено ли количество кластеров заранее или нет. В последнем случае количество кластеров определяется в ходе выполнения алгоритма на основе распределения исходных данных.

1. Постановка задачи. Алгоритмы кластеризации оперируют с объектами. С каждым объектом отождествляется вектор характеристик $X_i = (x_i^1, ..., x_i^m)$.

Компоненты x_i^k, $i = 1, .., n$, $j = 1, .., m$ являются отдельными характеристиками объекта. Количество характеристик d определяет размерность пространства характеристик.

Множество, состоящее из всех векторов характеристик, обозначается $M = (X_1, ..., X_n)$, где $X_i = (x_i^1, .., x_i^m)$.

Кластер представляет собой подмножество "близких" друг к другу объектов из M. Расстояние $D(X_{i_1}, X_{i_2})$ между объектами X_{i_1} и X_{i_2} определяется на основе выбранной метрики в пространстве характеристик.

Четкая (непересекающаяся) кластеризация — кластеризация, в которой каждый объект X_i из M относится только к одному кластеру.

При анализе результатов кластеризации необходимо учитывать особенности использованных алгоритмов.

Большую популярность в последнее время получили нечеткие алгоритмы, среди которых особенно широко известен алгоритм "нечетких средних" — FCMA (Fuzzy C-Means Algorithm). Следует отметить, что главным недостатком k- и c-средних алгоритмов является необходимость априорного задания требуемого числа кластеров, а также других численных параметров, от величины которых существенно зависят результаты кластеризации. Из этого следует, что при использовании алгоритмов кластеризации необходимо иметь дополнительные критерии качества разделения объектов на кластеры, позволяющие численно оценить результат применения тех или иных параметров [1].

Основные известные алгоритмы кластеризации налагают ограничения на геометрию получаемых кластеров, в частности, требуя возможности охвата каждого кластера отдельным выпуклым множеством. Такое ограничение налагается предположениями таких алгоритмов о существовании центров кластеров (K-Means) или функции плотности вероятности для каждого кластера с соответствующим значением математического ожидания и дисперсией. В результате эти алгоритмы не в состоянии адекватно разбить на кластеры невыпуклые множества, тем более вложенные структуры.

Эту проблему решает описываемый ниже алгоритм кластеризации на основе нечетких отношений, позволяющий группировать в кластеры элементы, между которыми есть последовательность "близких" друг к другу элементов, что также соответствует интуитивному представлению о группировке.

С этой целью описывается подход к кластеризации конечного набора элементов произвольного метрического пространства на основании разбиения множества на классы эквивалентности по нечеткому отношению.

На основании метрики определяется нечеткое отношение, обладающее свойствами четкой рефлексивности и нормальной α-симметричности. Строится транзитивное замыкание отношения, позволяющее определить для каждого значения α в диапазоне от 0 до 1 отношение эквивалентности на исходном множестве. По построению отношения два элемента входят в один класс эквивалентности тогда и только тогда, когда между ними есть последовательность попарно "близких" друг к другу элементов.

2. Описание алгоритмов кластеризации. Алгоритм, состоящий из нескольких этапов, начинается с определения списков объектов (формирование экспериментальных данных):

$$X_i = x_i^k,\ i = \overline{1,n},\quad k = \overline{1,m}.$$

На следующем этапе производится их нормирование:

$$U_i^k = l\frac{x_i^k - x^{\min}}{x^{\max} - x^{\min}};\ \ i = \overline{1,n},\ k = \overline{1,m}.$$

Выполняется оператор фаззификации:

$$\mu^j\left(U_i^k\right) = \frac{1}{1+\frac{U_i^k - b_j}{c_j}};$$

b_j, c_j — параметры; $j = \overline{1,l}$; l — термы.

Вычисляются оператор фаззификации, а также степень истинности по каждому правилу:

$$\mu^*\left(U_i^k\right) = \max_j \mu^j\left(U_i^k\right); \quad SP_i = \prod_{k=1}^{m} \mu^*\left(U_i^k\right).$$

Выбор нечетких правил:

Нормирование SP_i

$$\eta^i = l\frac{SP_i - SP^{\min}}{SP^{\max} - SP^{\min}}.$$

Фаззификация

$$\mu^j\left(\eta^i\right) = \frac{1}{1+\frac{\eta^i - b_j}{c_j}}; \quad \mu^*\left(\eta^i\right) = \max_j \mu^j\left(\eta^i\right).$$

Пусть X — метрическое пространство и определенная на нем метрика, $(SP_1, ..., SP_n) \subset X$ — последовательность элементов из X.

Предполагаем, что

$$\forall i \in \{1, \ldots, n\} \exists j \in \{1, \ldots, n\} : SP_i \neq SP_j. \tag{1}$$

Из условия (1) следует, что при $\forall i \in \{1, \ldots, n\}$ справедливо неравенство:

$$\max\{d(SP_i, SP_k) | k \in \{1, \ldots, n\}\} > 0. \tag{2}$$

Таким образом, для каждого индекса элемента i мы можем определить функцию $\xi_i(j)$, описывающую меру сходства j-го элемента последовательности с i-м элементом.

Определение 1.

$$\xi_i : \{1, ..., n\} \to [0, 1], \quad \xi_i(j) := 1 - \frac{d(SP_i, SP_j)}{\max\{d(SP_i, SP_k) | k \in \{1, ..., n\}\}}. \tag{3}$$

Для каждого индекса i определим функцию $\varsigma_i(k, l)$, описывающую меру сходства k-го и l-го элементов относительно i-го элемента [5].

Определение 2.

$$\varsigma_i : \{1, ..., n\}^2 \to [0, 1], \quad \varsigma_i(k, l) := 1 - |\xi_i(SP_k) - \xi_i(SP_l)|. \tag{4}$$

Определим теперь функцию $\theta(i, j)$, описывающую меру сходства любых двух элементов последовательности относительно всех элементов последовательности.

Определение 3.

$$\theta : \{1, \ldots, n\}^2 \to [0, \ldots, 1], \quad \theta(i, j) := \min\{\varsigma_k(i, j) | k \in \{1, \ldots, n\}\}. \tag{5}$$

Определение 4. Для $t = 1, 2, \ldots, n$ определим рекурсивную функцию $\theta^{(t)} : \{1, ..., n\}^2! \to [0, 1]$:

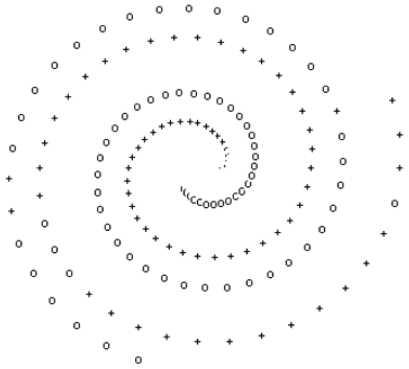

Рис. 1. Результаты решения спиральной задачи на основе предложенного метода

$$\begin{cases} \theta^{(1)}(i,j) := \theta(i,j), \\ \theta^{(t)}(i,j) := \max\left\{\min\left\{\theta^{(t-1)}(i,s), \theta^{(t-1)}(s,j)\right\} | s \in \{1,...,n\}\right\}. \end{cases} \tag{6}$$

Определение 5. Для $\alpha \in [0,1]$ определим на множестве $\{SP_1, ..., SP_n\}$ бинарное отношение $R_\alpha \subset \{SP_1, ..., SP_n\}^2$ следующим образом:

$$(SP_i, SP_j) \in R_\alpha :\Leftrightarrow \theta^{(n)}(i,j) \geq \alpha.$$

Покажем, что отношение R_α является отношением эквивалентности.

Рефлексивность. Из $\mu^{(t)}(i,i) = 1, \forall i,t$ следует, что $\theta^{(n)}(i,i) = 1 \geq \alpha \Rightarrow (SP_i, SP_i) \in R_\alpha$.

Симметричность. Пусть $(SP_i, SP_i) \in R_\alpha \Rightarrow \theta^{(n)}(i,j) \geq \alpha$. В силу $\mu^{(t)}(i,j) = \mu^{(t)}(j,i), \forall i,j,t$ следует, что $\theta^{(n)}(j,i) = \theta^{(n)}(i,j) \geq \alpha \Rightarrow (SP_j, SP_i) \in R_\alpha$.

Транзитивность. Пусть $(SP_j, SP_i) \in R_\alpha$, $(SP_j, SP_i) \in R_\alpha \Rightarrow \theta^{(n)}(i,j), \theta^{(n)}(j,m) \geq \alpha$. Из $\theta^{(n)}(i,m) \geq \alpha$ следует, что $\theta^{(n)}(i,m) \geq \alpha \Rightarrow (SP_i, SP_m) \in R_\alpha$.

Таким образом, отношение эквивалентности R_α разбивает множество $\{SP_1, ..., SP_n\}$ на непересекающиеся классы эквивалентности. Два элемента SP_i, SP_j входят в один класс эквивалентности тогда и только тогда, когда значение функции $\theta^{(n)}$ от этих элементов так велико, что на основании $\theta(i,j_1), ..., \theta(j_t,j) \geq \alpha \Rightarrow \theta^{(t)}(i,j) \geq \alpha$, $\theta(i,j_1), ..., \theta(j_m,j) \geq \alpha$ эквивалентно существованию последовательности пар элементов $(SP_i, SP_j), (SP_j, SP_j), ..., (SP_i, SP_i)$, на которых значение функции θ велико. По определению θ означает близость элементов каждой пары друг другу, т. е. два элемента входят в один класс эквивалентности тогда и только тогда, когда между ними есть последовательность попарно близких друг к другу элементов.

3. Вычислительный эксперимент. Работа алгоритма была проверена на задаче о спиралях. Задача состоит из разделения 150 объектов на два класса (кластера) на основании двух признаков. Поставленная задача является сложной, при этом объекты одного класса расположены очень близко к объектам второго класса (кластеры размыты между собой). На основании сравнительного анализа эта задача была решена и с помощью алгоритма c-средней (c-means) кластеризации. Полученные результаты на основе предложенного алгоритма приведены на рис. 1, а результаты на основе c-среднего алгоритма приведены на рис. 2.

При анализе результатов было выявлено, что при кластеризации на основе предложенного алгоритма, алгоритм классифицировал 7 объектов в другой кластер, что дает 4,7 % ошибки. Из графика, приведенного на рис. 2, видно, что решение задачи на основе c-среднего алгоритма дает 56 % ошибки (см. таблицу).

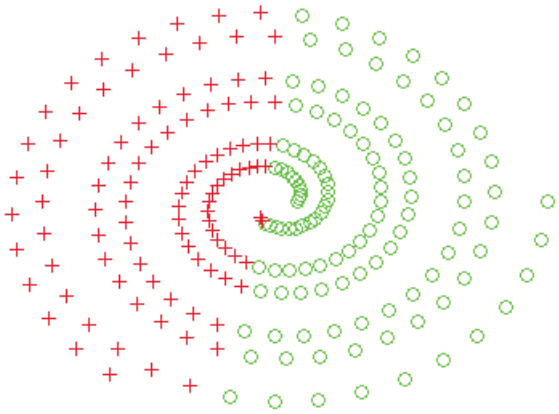

Рис. 2. Результаты решения спиральной задачи на основе с-среднего алгоритма

Таблица

Точность кластеризации алгоритмов

Сведения об объектах	Предлагаемый алгоритм	C-means алгоритм
Объекты	150	
Признаки	2	
Кластеры	2	
Точность	95,3 %	44 %

Заключение. Предложенный алгоритм кластеризации на основе нечетких отношений адекватно разбивает на кластеры невыпуклые множества, а также вложенные структуры. Применяя его в комбинации с нейронными сетями, можно найти решение задачи кластеризации, близкое к оптимальному.

Программа для кластеризации, на базе которой реализованы вышеизложенные методы применительно к нескольким задачам, приведена в [3].

Список литературы

1. Ротштейн А. П. Интеллектуальные технологии идентификации: нечеткая логика, генетические алгоритмы, нейронные сети. Винница: УНИВЕРСУМ-Винница, 1999.

2. Бекмуратов Т. Ф., Мухамедиева Д. Т. Методы и алгоритм синтеза нечетко-нейронных моделей принятия решений. Saarbrucken, Germany: изд-во “Palmarium Academic Publishing”, 2013.

3. Мухамедиева Д. Т., Мингликулов З. Б. Программа построения нейронечеткой модели идентификации // Гос. Патент. ведомство. Агентство по правовой охране программ для ЭВМ и базы данных. № DGU 02316. 2011.

4. Чубукова И. А. Data Mining. М.: Интернет-Университет информационных технологий; БИНОМ. Лаборатория знаний, 2006.

5. Мухамедиева Д. Т. Алгоритм кластеризации правил систем нечеткого вывода // Естеств. и техн. науки. 2013. № 2. С. 248–252.

6. Мухамедиева Д. Т. Решение задач многокритериальной оптимизации при наличии неопределенности нестатического характера // Акт. пробл. совр. науки. 2013. № 2. С. 237–239.

Мингликулов Зафар Бозорович — старш. науч. сотр. Центра разработки программных продуктов и аппаратно-программных комплексов при Ташкентском университете информационных технологий; тел. (+99871) 262-79-11; e-mail: mingliqulov@gmail.com

Дата поступления — 8.11.2013