УДК 519.248

А.С. Герасимова, С.В. Дронов

Алгоритм нечеткой кластеризации, основанный на выделении «основных объектов» кластеров

A.S. Gerasimova, S.V. Dronov

A Fuzzy Clusterization Algorythm Based on Main Objects Marking

Рассматривается новый подход к задаче нечеткой кластеризации множества объектов. С целью оценки функции принадлежности объекта к каждому из строящихся кластеров мы многократно применяем кластерные алгоритмы к одному и тому же набору данных. После этого оценка производится в соответствии с результатом. Решается также задача визуализации полученной конструкции. Проблему единой (универсальной) нумерации кластеров мы решаем путем выделения в каждом из них основных элементов.

Ключевые слова: кластерное разбиение, нечеткие множества, визуализация.

We deal with a fuzzy clasterization problem and propose a new approach to it. By repeated applications of some cluster algorythms to the same data we can estimate membership functions for an object with a respect to each of the clusters due to the constructions, and then try to visualizate the result. A problem of the universal cluster enumeration arises. We propose to solve it by marking several objets in each of the clasters.

Key words: clusterization, fuzzy sets, visualization.

1. Вводные замечания. Постановка задачи. Кластеризация – это разбиение заданного набора объектов на подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. В настоящее время существует множество алгоритмов и способов решения данной задачи. Среди них выделяется недавно возникший класс методов нечеткой кластеризации. Применение методов этого класса позволяет формализовать различного рода неопределенности, которые всегда существуют при решении реальных задач.

Нечеткая кластеризация является синтезом идей кластерного анализа (см., например, [1]) и теории нечетких множеств, основы которой изложены в [2]. На основе понятий, вводимых и изучаемых этой теорией, можно вместо обычных разбиений объектов на кластеры рассмотреть их нечеткие варианты. Это осуществляется путем построения функций принадлежности объектов к каждому из образуемых кластеров, каждая из которых изменяется в интервале [0, 1], что позволяет оценить степень принадлежности объекта к тому или иному кластеру. Опишем такой процесс более подробно.

Предположим, что задано n объектов $X_1, ...X_n$. Допустим, нам заранее известно количество k кластеров, на которые требуется разбить эти объекты. Если мы решаем классическую задачу, то каждому из объектов X_j будет поставлен в соответствие номер кластера $f(j)$, к которому он отнесен. Если j-й объект оказался в i-м кластере, то этот факт можно закодировать цепочкой символов 0 или 1 длины k, где единственная единица стоит на i-м месте. Переходя на язык нечетких множеств, такие цепочки, связываемые с каждым объектом, можно интерпретировать как задание k просто устроенных функций принадлежности, – вероятность того, что объект относится к i-му кластеру, есть 1, к любому другому – 0.

В практических задачах чаще всего любое утверждение о попадании объекта в кластер не является абсолютно истинным. Поэтому естественным представляется следующее обобщение классической задачи.

Для каждого объекта требуется указать числа $\mu_{j,1}, ...\mu_{j,k}$, которые объявляются значениями k функций принадлежности, соответствующим каждому из строящихся кластеров. Число $\mu_{j,i}$ будет интерпретироваться нами как вероятность попадания j-го объекта в i-й кластер.

Подобные задачи уже решались, см., например, [3–7]. Описанные в этих работах методы обладают некоторыми недостатками, которых, на наш взгляд, лишен предлагаемый нами новый алгоритм. В частности, ни один из них никак не учитывает различие результатов кластеризации при подходе к задаче разбиения одного и того же набо-

ра объектов с разных точек зрения. А ведь практически очевидно, что неоднозначность отнесения объектов в тот или иной кластер при различных подходах и является источником неопределенности оптимального кластерного разбиения.

2. Описание алгоритма. Приступим к изложению предлагаемого нового алгоритма. Пусть каждый из объектов $X_1, ..., X_n$ задан p признаками $x^1, ...x^p$. Не теряя общности, можно считать, что рассматриваемые объекты тождественны своим наборам признаков:

$$X_i \ = \ (x_i^1, ..., x_i^p), \quad i = 1, ..., n.$$

Например, если все признаки числовые, то мы отождествим объекты с точками в p-мерном евклидовом пространстве. Если же среди признаков встречаются качественные, то пространство, в котором располагаются исследуемые объекты, устроено более сложно, но все равно можно считать, что оно имеет размерность p.

Считаем также, что требуемое число кластеров k известно. Для построения функций принадлежности предлагается следующий способ. Рассмотрим различные «правдоподобные» разбиения наших объектов на k кластеров. Это могут быть результаты применения разных кластерных алгоритмов или одного и того же алгоритма в разных исходных условиях (замена стартовой конфигурации, если результат может зависеть от нее, исключение или добавление каких-то дополнительных признаков или внешних критериев качества и т.д.). Проведя однократное разбиение объектов на кластеры, получим, что каждому из объектов приписан номер кластера, к которому он относится. По l-му построенному кластерному разбиению определим функцию f_l, заданную на номерах объектов, принимающую значения от 1 до количества кластеров k. Конкретно, если j-й объект попал в i-й кластер, то будем писать $f_l(j) = i$.

Если при применении N разных способов разбиения j-й объект оказался в i-м кластере $m_{i,j}$ раз, то положим

$$\mu_{j,i} \ = \ \frac{m_{i,j}}{N}, \quad i = 1, ..., k; \ j = 1, ..., n. \qquad (1)$$

Это число является естественной оценкой соответствующей вероятности попадания j-го объекта в i-й кластер, и может быть принято за значение i-й функции принадлежности на j-м объекте.

Самая главная трудность, которая почти наверняка возникнет при практическом построении таким образом определяемых функций принадлежности, связана с тем, что номера кластеров при применении того или иного кластерного алгоритма к одному и тому же набору объектов, задаются этими алгоритмами достаточно произвольно. Смена нумерации кластеров может происходить и в случае повторного применения одного и того же алгоритма, например, при задании разных стартовых конфигураций. При этом даже одинаковые разбиения могут формально отличаться – пусть, например, кластеры получились одинаковыми, но первый и второй кластер обменялись номерами. Поскольку мы собираемся многократно использовать разные способы разбиения и хотим их сравнивать, возникает проблема «универсальной нумерации кластеров». Рассмотрим один вариант такой нумерации более подробно.

Допустим, мы уверены, что объекты обязательно разобьются на k кластеров. Выделим k основных объектов, заведомо относящихся к разным кластерам, и обозначим их $\hat{X}_1, ..., \hat{X}_k$. Выбор таких объектов можно сделать, располагая некоторой априорной информацией о данных. Если такой информации нет, то их можно определить «на глаз», предварительно применив к данным какой-либо современный алгоритм визуализации на плоскости (см, например, [8]).

В случаях, когда визуализация рассматриваемых объектов оказывается неоправданно сложна, и n относительно невелико, можно предложить следующий способ выделения k основных объектов с применением попарных расстояний между ними. Пусть каким-то образом определено расстояние между объектами. Например, это обычная евклидова метрика, когда все признаки $x^1, ...x^p$ числовые. Если k=2, то можно выбрать в качестве основных те объекты, расстояние между которыми является самым большим. Если k=3, то основными признаем те объекты, которые образуют треугольник с наибольшим периметром. Вопрос о том, верно ли то, что при $k > 3$ основными следует признать те объекты, которые образуют между собой k-угольник с наибольшим периметром, оставим пока открытым.

Если по каким-либо причинам ни один из вышеуказанных способов не подходит, то выбор основных объектов можно совершить перебором. Для этого рассмотрим все возможные «правдоподобные» в указанном ранее смысле разбиения множества объектов на k кластеров, после чего выделим такие k объектов, которые никогда не оказывались в одном кластере. Эти объекты и выбираются за основные.

После того, как некоторое кластерное разбиение построено, зададим номера кластеров так, чтобы тот из них, в который оказался отнесенным основной объект $\hat{X}_i$, получил бы номер i. Если же два основных объекта все же оказались в одном кластере, то нам следует признать такое разбиение непригодным для рассмотрения и исключить его из дальнейших вычислений. Таким образом, нумерация кластеров окажется универсальной, по крайней мере, с точки зрения основных объектов – они всегда попадают в кластеры с одними и те-

ми же номерами.

Рассмотрим более детально вариант предлагаемого способа действий, если кластерный алгоритм, который мы применяем для построения разбиений, фиксирован, но его результат может зависеть от того «стартового» разбиения, которое выбирается на его входе.

Переберем все стартовые разбиения заданных объектов на k кластеров. Количество всех таких разбиений, т.е. количество всех различных способов разбиения n объектов на k непустых подмножеств, как известно, равно числу Стирлинга второго рода (см., например, [9]). За счет предложенной выше универсальной нумерации кластеров мы получаем возможность оценить вероятности попадания каждого из объектов в фиксированный кластер уже описанным способом. Построим таблицу, в которой на пересечении j-й строки, соответствующей объекту X_j и l-го столбца, соответствующего l-му разбиению, находится номер кластера $f_l(j)$, к которому отнесен этот объект в данном разбиении.

Из построенной таблицы находим числа $m_{i,j}$ как количества повторений значения i в j-й строчке. Далее, перебирая по очереди все эти разбиения и исключая из них непригодные (если такие получатся), строим функции принадлежности по формуле (1), понимая под N количество разбиений, признанных пригодными.

3. Визуализация нечетких кластеров. Пусть мы успешно построили соответствующие каждому объекту и каждому кластеру функции принадлежности. Нарисуем двумерную картинку объектов согласно одному из современных методов визуализации. Для определения степени четкости того из кластеров, который для нас служит основным, предложим алгоритм раскраски результата нечеткой кластеризации. Будем красить точки-объекты следующим образом. Если функция принадлежности рассматриваемого объекта к основному кластеру равна 1, то соответствующую ему точку окрасим в красный цвет, и по мере уменьшения значения цвет будем менять по шкале радуги. Если же функция принадлежности рассматриваемого объекта равна 0, то соответствующую ему точку окрасим в фиолетовый цвет. В остальных случаях можно предложить границы, при попадании функции принадлежности в которые для раскраски точек будут применяться определенные цвета.

Переход от красного цвета в фиолетовый наблюдается по мере уменьшения вероятности попадания окрашиваемого объекта в тот из кластеров, который объявлен основным. Отметим, что шкала таблицы 1 может корректироваться и при увеличении количества объектов может становиться более плавной за счет добавления новых оттенков того или иного цвета. При этом возможно построить, по крайней мере, $k+1$ раскрашивание. Из них k раскрашиваний получаются путем использования для раскраски функции принадлежности каждого из кластеров. В таком случае ясно, что на j-ой картинке j-й основной объект всегда окрашен в красный цвет, причем другие основные объекты при этом все фиолетовые.

Таблица 1

Раскраска результата

$\mu_{j,i}$	Цвет	$\mu_{j,i}$	Цвет
0	фиолетовый	(0,4; 0,6]	зеленый
(0; 0,2]	синий	(0,6; 0,8]	желтый
(0,2; 0,4]	голубой	(0,8; 1,0]	оранжевый
		1	красный

Для построения $(k+1)$-го раскрашивания в качестве функции принадлежности будем использовать наибольшее значение из k функций принадлежности каждого объекта. При этом если объект с большой вероятностью постоянно относится к одному и тому же кластеру, то ему будет соответствовать точка цвета, близкого к красному. Ясно, что «центры покраснения» при этом совпадают с основными объектами или находятся поблизости от них. Назовем это последнее раскрашивание итоговым.

По результатам итогового раскрашивания можно судить об устойчивости выбранного кластерного алгоритма по отношению к изучаемому набору данных. Если в нем кроме основных объектов красным цветом или близким к нему не обладает ни один объект, то устойчивости как таковой нет. Если же красный цвет преобладает на всей картине, то каждый объект с вероятностью, близкой к 1, всегда попадает в один и тот же кластер, что означает постоянство или устойчивость алгоритма. Таким образом, степень упомянутой устойчивости можно характеризовать степенью покраснения точек-объектов в итоговом раскрашивании.

4. Один искусственный пример. Для иллюстрации изложенной методики рассмотрим небольшой искусственный пример. Пусть изучаемые объекты задаются таблицей 2.

Допустим, что данные объекты следует разбить на 3 кластера. Будем использовать вариант алгоритма k-средних, зависящий от стартового разбиения при k=3 [1, глава 7]. Количество всех стартовых разбиений, вычисленное по фор-

Таблица 2

Данные искусственного примера

объект	x^1	x^2	объект	x^1	x^2
X_1	1	1	X_4	6	7
X_2	4	4	X_5	10	5
X_3	4	9			

мулам из [6], равно 25. Возьмем в качестве трех основных объектов X_1, X_3, X_5. Наш выбор объясняется «крайним» размещением объектов на их двумерной визуализации (см. рис.1).

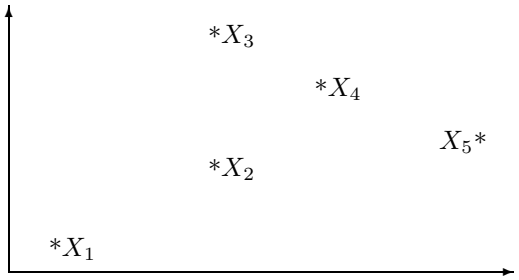

Рис. 1. Выделение основных объектов

Далее построим функции принадлежности для каждого объекта по формуле (1). Из 25 «стартовых» разбиений пригодными признаны 21. Результаты приведены в таблице 3.

Осуществим раскраску результата согласно методике, изложенной в п.3. Изобразим функции принадлежности каждого из кластеров на следующих трех рисунках.

Рис. 2. Раскрашивание. Первый кластер основной

Таблица 3

Функции принадлежности

Объект	$\mu_{.,1}$	$\mu_{.,2}$	$\mu_{.,3}$
X_1	1	0	0
X_2	0,82	0,09	0,09
X_3	0	1	0
X_1	0,05	0,71	0,24
X_1	0	0	1

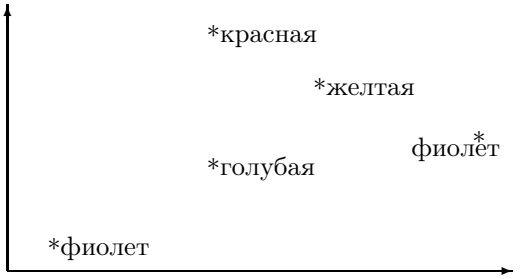

Рис. 3. Раскрашивание объектов со вторым основным кластером

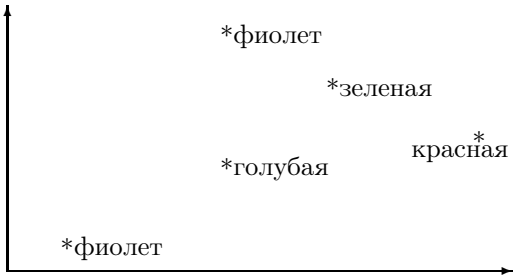

Рис. 4. Раскрашивание объектов с третьим основным кластером

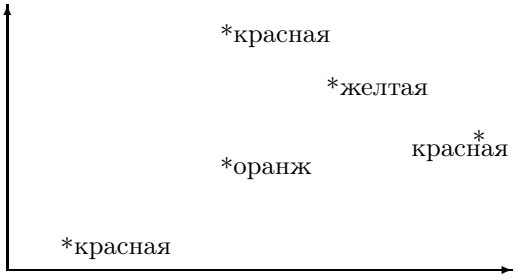

Рис. 5. Итоговое раскрашивание

Построим итоговое раскрашивание, используя для каждого объекта в качестве функции принадлежности наибольшее значение из функций принадлежности этого объекта каждому из трех кластеров (рисунок 5).

В этом примере 1-й, 3-й и 5-й объекты однозначно отнесены каждый в свой кластер, а 2-й и 4-й объекты не могут быть объективно отнесены в какой-либо из трех кластеров. Степень устойчивости данного алгоритма на данной структуре может быть, таким образом, оценена числом 3/5.

Число объектов в рассмотренном примере было весьма невелико. В реальных же задачах, как правило, число объектов гораздо больше, поэтому возможные ситуации будут более разнообразны, а шкала возможных степеней устойчивости – более плотной. Решение такого рода задач может привести к более интересным и разнообразно интерпретируемым результатам.

Библиографический список

1. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности. – М.: Финансы и статистика, 1989.

2. Zadeh, L. A. Fuzzy sets. // Information and Control. – 1965. – Vol. 8.

3. Бериков В.С., Лбов Г.С. Современные тенденции в кластерном анализе. // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы», 2008.

4. Вятченин Д.А. Нечеткие методы автоматической классификации. – Минск: Технопринт, 2004.

5. Рыжов А.П. Элементы теории нечетких множеств и её приложений. – М.: Диалог-МГУ, 2003.

6. Ahmed Ismail Shihab Fuzzy clustering algorithms and their application to medical image analysis. – PhD thesis, Imperial College of Science, Technology and Medicine University of London. – London, 2000.

7. Kosko B. Fuzzy systems as universal approximators. // IEEE Transactions on Computers. – 2004. – Vol. 43, No. 11.

8. Зиновьев А.Ю. Визуализация многомерных данных. – Красноярск: Изд-во КГТУ, 2000.

9. Яблонский С.В. Введение в дискретную математику. – М.:Наука, 1986.