УДК 004.67

Е. О. Васюкова[*]

РЕАЛИЗАЦИЯ АЛГОРИТМА КЛАСТЕРИЗАЦИИ FOREL В МАТЕМАТИЧЕСКОЙ СРЕДЕ MATLAB

Кластеризация или кластерный анализ – это задача разбиения заданной выборки объектов на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял их схожих объектов, а объекты различных кластеров значительно отличались [1].

Кластеризация может иметь различные цели, определяемые особенностью конкретной задачи. К основным целям относятся:

– аналитическая обработка данных – определение структуры множества данных путем разбиения его на группы схожих объектов;

– выявление атипичности – обособление объектов, которые не подходят ни к одному из кластеров;

– сокращение объема хранимых данных в случаях сверхбольшой выборки, оставив по одному наиболее типичному представителю от каждого кластера.

Процедура кластеризации позволяет разделить всю совокупность данных на группы, элементы которых схожи друг с другом по определенному признаку. Технология кластеризации применяется во многих областях. Например, в области медицины кластеризация заболеваний, лечения заболеваний или их симптомов приводит к широко используемым таксономиям. В археологии с помощью кластерного анализа исследователи пытаются установить схожести каменных орудий, предметов быта и т.д. В информационных системах используется для «интеллектуального» отбора результатов при поиске файлов, вебсайтов, других объектов, предоставляя пользователю возможность выбора заведомо более релевантного подмножества поиска и исключения заведомо менее релевантного [2].

Кластерный анализ отличается от классификации тем, что не накладывает ограничения на представление изучаемых объектов, позволяет проводить анализ показателей различных типов данных, не требует предварительных гипотез о наборе данных. При этом следует помнить, что переменные должны измеряться в сопоставимых шкалах.

Целевой функцией для определения схожести и различия кластеров является расстояние между точками на диаграмме рассеивания. Основной задачей кластеризации является определение расстояний между внутрикластерными объектами – мера их близости. Существует

[*] Работа выполнена под руководством канд. техн. наук, доцента ФГБОУ ВПО «ТГТУ» А. В. Яковлева.

множество способов определения расстояний между объектами, основные из них представлены в табл. 1 [1].

Формальная постановка задачи кластеризации: пусть X – множество объектов; Y – множество номеров (имен, меток) кластеров. Задана функция расстояния между объектами $\rho(x, x')$. Имеется конечная обучающая выборка объектов $X^m = \{x_1, \ldots, x_m\} \subset X$. Требуется разделить выборку на непересекающиеся подмножества – кластеры, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ: $\frac{\sum_{i<j, c(x_j)=c(x_j)} \rho(x_i, x_j)}{\sum_{i<j, c(x_j)=c(x_j)} 1} \to \min$, а объекты разных кластеров существенно отличались $\frac{\sum_{i<j, c(x_j)\neq c(x_j)} \rho(x_i, x_j)}{\sum_{i<j, c(x_j)\neq c(x_j)} 1} \to \max$. При этом каждому объекту $x_i \in X^m$ приписывается номер кластера y_i [3].

1. Метрики для определения расстояний между объектами

Метрика	Формализованное описание
Евклидово расстояние	$\rho(x, x') = \sqrt{\sum_i^n (x_i - x_i')^2}$
Квадрат евклидова расстояния	$\rho(x, x') = \sum_i^n (x_i - x_i')^2$
Расстояние Хемминга	$\rho(x, x') = \sum_i^n \lvert x_i - x_i' \rvert$
Расстояние Чебышева	$\rho(x, x') = \max\left(\lvert x_i - x_i' \rvert\right)$
Степенное расстояние	$\rho(x, x') = \sqrt{\sum_i^n (x_i - x_i')^p}$, где p – взвешивание разностей по отдельным координатам; r – прогрессивное взвешивание расстояний между объектами
Расстояние Махаланобиса	$D_{\text{M}}(x) = \sqrt{(x-\mu) S^{-1} (x-\mu)^T}$, где S – матрица ковариаций

Алгоритм кластеризации – это функция a: $X \to Y$, которая любому объекту $x \in X$ ставит в соответствие номер кластера $y \in Y$. Множество Y в определенных случаях известно заранее, однако чаще ставится задача определить оптимальное количество кластеров с точки зрения того или иного критерия качества кластеризации [3, 4].

Автором в пакете Matlab реализован алгоритм FOREL. Суть алгоритма заключается в том, что задаются некоторая точка $x_0 \in X$ и параметр R. Выделяются все точки выборки $x_i \in X^l$, попадающие внутрь сферы $\rho(x_i, x_0) \le R$, и точка x_0 переносится в центр тяжести выделенных точек. Эта процедура повторяется до тех пор, пока состав выделенных точек, а значит, и положение центра, не перестанет меняться. При этом сфера перемещается в место локального сгущения точек. Центр сферы x_0 при этом не является объектом выборки, потому и называется формальным элементом.

Последовательность шагов алгоритма представляет собой следующую совокупность:

1: инициализировать множество некластеризованных точек: $U := X^l$;
2: пока в выборке есть некластеризованные точки, $U \neq \varnothing$:
3: взять произвольную точку $x_0 \in U$ случайным образом;
4: повторять
5: образовать кластер-сферу с центром в x_0 и радиусом R: $K_0 := \{ x_i \in U \mid \rho(x_i, x_0) \le R \}$;
6: поместить центр сферы в центр масс кластера: $x_0 := \frac{1}{|K_0|} \sum_{x_i \in K_0} x_i$;
7: пока центр x_0 не стабилизируется;
8: пометить все точки K_0 как кластеризованные: $U := U \setminus K_0$;
9: применить алгоритм КНП к множеству центров всех найденных кластеров;
10: объект $x_i \in X^l$ приписать кластеру с ближайшим центром.

На рисунке 1 представлены результаты реализованного алгоритма FOREL.

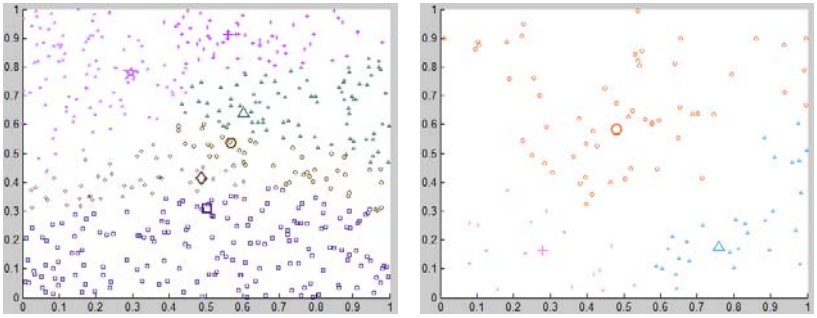

Рис. 1. Результат кластеризации по алгоритму FOREL

Использование алгоритма кластеризации FOREL в конкретной области принципиально неоднозначно, и тому есть несколько причин:

– количество кластеров заранее не известно и устанавливается в соответствии с некоторым субъективным критерием;

– результат кластеризации существенно зависит от исходного радиуса, выбор которого, как правило, также субъективен и определяется экспертом.

Список литературы

1. *Мандель*, *И. Д.* Кластерный анализ / И. Д. Мандель. – Москва : Финансы и Статистика, 1988. – 176 с.

2. *Райзин*, *Дж. Вэн*. Классификация и кластер / Дж. Вэн Райзин. – Москва : Мир, 1980. – 392 с.

3. *Петренко*, *С. В*. Синтез математической модели автоматизированной системы управления специального назначения с микроядерной архитектурой / С. В. Петренко, Ал. В. Яковлев, Ан. В. Яковлев // Вопросы современной науки и практики. Университет им. В. И. Вернадского. – 2009. – № 1. – С. 34 – 41.

4. *Петренко*, *С. В*. Использование формализма сетей Петри для моделирования распределенных систем с микроядерной архитектурой / С. В. Петренко, Ал. В. Яковлев, Ан. В. Яковлев // Вопросы современной науки и практики. Университет им. В. И. Вернадского. – 2009. – № 5. – С. 11 – 19.

Кафедра «Информационные системы и защита информации»
ФГБОУ ВПО «ТГТУ»