УДК 004.67

М. А. Пеливан[*]

КЛАСТЕРИЗАЦИЯ МЕТОДОМ НАИБОЛЕЕ УДАЛЕННЫХ СОСЕДЕЙ ИЛИ МЕТОДОМ ПОЛНОЙ СВЯЗИ

В современном мире приходится работать с большими объемами данных, что вызывает множество трудностей. Кластеризация позволяет разделить данные на группы и иметь дело с меньшим объемом информации.

Кластеризация может обладать множеством целей, которые определяются в зависимости от особенностей поставленной задачи. К основным целям можно отнести: понимание данных – определение структуры множества данных путем разбиения его на группы схожих объектов; обнаружение новизны – выделение объектов, не подходящих ни к одному из кластеров; сжатие данных, когда рассматриваются не целые классы данных, а лишь типичных представителей классов.

Кластеризация используется во множестве различных областей деятельности человека. В медицине кластеризация симптомов и заболеваний заметно облегчает анализ и вид лечения. Верное рассмотрение кластеров симптомов шизофрении, паранойи и других болезней является определяющим при ведении успешной терапии. Так же кластеризация полезна в маркетинговых исследованиях и в археологии. Кластеризация является эффективной и полезной практически во всех случаях, когда требуется классифицировать огромное количество информации к удобным для последующей работы группам.

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 разбить выборку на непересекающиеся подмножества, называемые кластерами, 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 $\frac{\sum_{i<j, c(x_j)\neq c(x_j)} \rho(x_i, x_j)}{\sum_{i<j, c(x_j)\neq c(x_j)} 1} \to \max$ [1].

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

Основной задачей при кластеризации является определение расстояний между внутрикластерными объектами – степень их «схожести». Существует множество методов определения расстояний между объектами, основные из них представлены в табл. 1.

Метод полной связи (CompleteLinkage) или метод удаленных соседей (FurthestNeighbor) относится к иерархическим агломеративным методам. В данных методах используется классификация, осуществляемая путем последовательного группирования (агломерации) объектов в кластеры, в результате объединения все объекты оказываются иерархически организованными. Иерархические методы кластеризации являются простыми комбинаторными процедурами, отличающимися критерием объединения объектов в группы (кластеры). В процессе работы алгоритма происходит многократное использование выбранного критерия объединения, применяемого к матрице расстояний между всеми объектами. В начале объединяются объекты, расположенные на одном уровне сходства, наиболее близкие друг к другу.

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	$D_{\text{М}}(x) = \sqrt{(x-\mu) S^{-1} (x-\mu)^T}$, где S – матрица ковариаций

Впоследствии поочередно присоединяются оставшиеся несгруппированными объекты до тех пор, пока все они не будут объединены в один кластер. Метод полной связи является восходящим алгоритмом, т.е. вначале каждый объект помещается в отдельный кластер, а затем происходит объединение в более крупные кластеры. Таким образом, появляется система вложенных разбиений. Работа данных алгоритмов обычно отображается в виде ветвистого древовидного графика (дендрограммы) [2].

В методе полной связи расстояние между классами определяется как расстояние между самыми удаленными объектами, входящими в эти классы. Происходит объединение кластеров, расстояние между самыми отдаленными представителями которых имеет наименьшее значение. При использовании данного метода формируется большое количество компактных кластеров, включающих в себя наиболее похожие элементы.

На начальном этапе работы алгоритма происходит расчет матрицы расстояний между объектами. На каждом последующем этапе в матрице расстояний определяется минимальное значение, которое соответствует расстоянию между двумя ближайшими кластерами. Далее происходит образование нового кластера, состоящего из двух найденных кластеров. Происходит повторение действий до тех пор, пока не объединятся все кластеры [3].

Метод полной связи или метод удаленных соседей был реализован в среде Matlab. Ниже представлена функция, отвечающая за выбор объединяемых кластеров (рис. 1), а также результаты работы алгоритма (рис. 2).

```
function y=klast_ob(S1,N, Ztmp, Maxz)
for i=1:N
    if (i~=Ztmp(1))&(i~=Ztmp(2))
        if S1(i,Ztmp(1))<S1(i,Ztmp(2))
        S1(i,Ztmp(1))=S1(i,Ztmp(2));
        S1(Ztmp(1),i)=S1(i,Ztmp(2));
        end
    end
end
S1(:,Ztmp(2))=Maxz+1;
S1(Ztmp(2),:)=Maxz+1;
y=S1;
```

Рис. 1. Функция выбора объединяемых кластеров

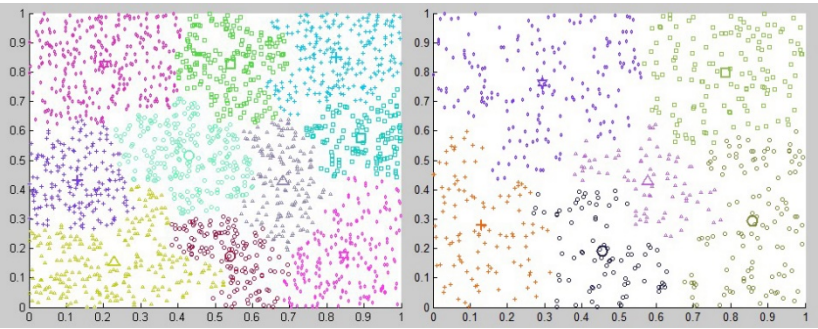

Рис. 2. Результаты работы алгоритма

Выводы. Таким образом, кластеризация является неотъемлемой частью современной обработки больших массивов информации, заметно облегчая работу с ней и охватывая множество областей применения. Выбор различных методов кластеризации зависит от конкретного случая и не может быть определен однозначно. Метод полной связи хорошо работает, когда объекты происходят из действительно различающихся групп. При удлиненной форме кластеров или их естественном «цепочечном» типе данный метод оказывается совершенно непригодным. Так же алгоритм хорошо справляется с задачей выделения аномальных объектов, не подходящих ни к одному из кластеров.

Список литературы

1. *Заде*, *Л. А*. Кластеризация и кластер / Л. А. Заде, С. Рао и др. – Москва : Мир, 1980. – 383 с.
2. *Петренко*, *С. В*. Синтез математической модели автоматизированной системы управления специального назначения с микроядерной архитектурой / С. В. Петренко, А. В. Яковлев, Ан. В. Яковлев // Вопросы современной науки и практики. Университет им. В. И. Вернадского. – 2009. – Вып. 1. – С. 160 – 169.
3. *Яковлев*, *А. В*. Использование формализма сетей Петри для моделирования распределенных систем с микроядерной архитектурой / А. В. Яковлев // Вопросы современной науки и практики. Университет им. В. И. Вернадского. – 2009. – Вып. 5. – С. 96 – 104.

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