УДК 517

Шапошникова Евгения Анатольевна
студентка
Восточноукраинского национального университета имени Владимира Даля
Shaposhnikova Yevgeniya
Student of the
Volodymyr Dahl East Ukrainian National University

ИССЛЕДОВАНИЕ И АНАЛИЗ АЛГОРИТМА КЛАСТЕРИЗАЦИИ ДАННЫХ МЕТОДОМ К-СРЕДНИХ

STUDY AND ANALYSIS DATA CLUSTERING ALGORITHM BY METHOD K-MEANS

Аннотация. *В статье рассматривается исследование многомерной статистической процедуры по сбору данных, содержащей информацию о выборке объектов и затем упорядочивания объектов в сравнительно однородные группы.*

Ключевые слова: *кластеризация, кластер, массив, объекты, центроиды, анализ, вектор.*

Summary. *The paper considers the study of a multidimensional statistical procedure for data collection that contains information on the selection of objects and then the ordering of objects into relatively homogeneous groups.*

Key words: *clustering, cluster, array, objects, centroids, analysis, vector.*

Введение. На сегодняшний день современные вычислительные машины и компьютерные сети позволяют накапливать большие массивы информации для задач обработки и анализа. К сожалению, сама по себе машинная форма представления данных содержит информацию, необходимую человеку, в скрытом виде, и для ее изучения нужно использовать специальные методы анализа данных [1].

Огромный объем информации позволяет получить более точные расчеты и анализ, однако, он превращает поиск информации в сложную задачу.

Известный способ анализа данных — Кластеризация, может применяться во многих областях, где необходимо исследование экспериментальных или статистических данных.

Аналитику часто легче выделить группы схожих объектов, изучить их особенности и построить для каждой группы отдельную модель, чем создавать одну общую модель для всех данных [7]. Таким приемом постоянно пользуются в маркетинге, выделяя группы клиентов, покупателей, товаров и разрабатывая для каждой из них отдельную стратегию [7].

Этапы кластеризации. Основной задачей кластеризации является поиск независимых групп (кластеров) во всем множестве анализируемых данных. Кластерный анализ позволяет лучше понять данные. Так же, группировка подобных объектов позволяет сокращать их число, что приводит к упрощению анализа. Каждая группа будет содержать «подобные» объекты, а объекты разных групп должны максимально отличаться. Перечень групп четко не задан и определяется в процессе работы алгоритма [2].

Применение кластерного анализа сводится к следующим этапам:

1. Выбор массива данных для кластеризации.
2. Определение главных (центральных) переменных для оценки объектов в выборке.
3. Определение сходства между всеми объектами относительно центральных переменных.
4. Использование кластерного анализа для создания групп подобных объектов (кластеров).
5. Вывод результатов анализа.

Полученный результаты можно откорректировать, выбрав другой метод анализы для получения максимально точных результатов.

Анализ алгоритма к-средних. K-means — простой повторяющийся алгоритм кластеризации, который разбивает большой набор данных на группы относительно заданного числа кластеров k. Алгоритм достаточно просто реализовать и программировать, является относительно быстрым, подходит под анализ практически любых данных. Часто используется для анализа закономерностей в покупках. Алгоритм k-means исторически один из самых важных алгоритмов интеллектуального анализа данных.

Исторически сложилось так, что k-means был открыт несколькими исследователями различных дисциплин, в первую очередь Ллойдом (1957, 1982), Форджи (1965), Фридманом и Рубином (1967), и МакКуином (1967)[6].

Алгоритм k-means применяется массиву значений точек в d-мерном векторном пространстве. Таким образом, это кластеры набора d-мерных векторов, $D=\{x_j \mid i=1,\ldots,N\}$, где $x_j \in S_i$ обозначает j-ый объект или «точку данных». Как уже говорилось, k-means является алгоритмом кластеризации, который разделяет D на k кластеров точек. То есть, алгоритм k-means объединяет все точки данных в D так, что каждая точка x_j попадает в один и только один из k кластеров. Можно отследить, какая точка находится в каком кластере, назначив каждой точке номер кластера. Точки с таким же номером кластера находятся в одном и том же кластере, в то время как точки с различными номерами кластера находятся в разных кластерах [6].

Алгоритм действует по принципу минимизации суммарного квадратичного отклонения точек кластеров от центров этих кластеров (центроидов — точек, которые являются центрами кластеров). В алгоритмах кластеризации группировка точек происходит подбором подобных самим себе, похожим по большинству признаков. Алгоритм k-means использует меру близости — Евклидово расстояние.

$$1.\ V=\sum_{i=1}^{k}\sum_{x_j\in S_i}\left(x_j-\mu_i\right)^2 \qquad (1)$$

где k — число кластеров, x_j — каждый вектор представленный точкой, i = 1,2,..., k, S_i — полученные кластеры и μ_i — центры масс векторов x_j (центроиды).

Значение k является основным из входных данных алгоритма и задается исследователем каждый раз вручную или случайным образом [5].

Алгоритм k-means приводит к минимизации итогового квадрата Евклидова расстояния между каждым вектором x_j и подобной точкой кластера μ_i. Уравнение (1) является целевой функцией метода k-means.

Алгоритм k-means:

1. Случайным образом выбирается к-точек (центроиды) из первоначального множества точек.

2. Используя формулу (1) распределяем точки по кластерам, относительно центроидов.

3. Находим новое положение центроидов, вычислив центр каждого кластера.

4. Выполняем пункт 2 и 3 до тех пор, пока центроиды не перестанут менять свое положение или до определенного порога изменения положения центроидов.

Каждая итерация нуждается в $N*k$ сравнений, что ясно видно из алгоритма (рис. 2), что определяет сложность одной итерации.

Число итераций может зависеть от N и итерации меняются в зависимости от N. А это значит, что чем больше точек во множестве (N), тем дольше будет работать алгоритм.

Для сокращения времени работы алгоритма используют распараллеливание этапов распределения точек по кластерам.

Каждый процесс можно разбить на столько же потоков и тогда начальное множество данных разбивается на такое же количество частей, при этом каждый поток будет иметь дело со своим объемом данных, независимо он остальных.

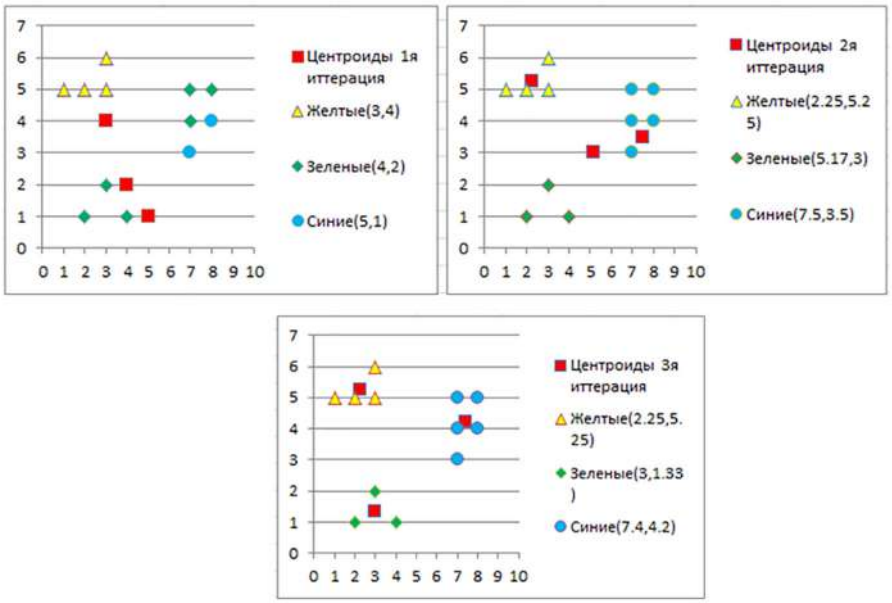

Рис. 1. Результат кластеризации методом k-средних
составлено автором на основе [5]

Рис. 2. Схема алгоритма k-средних составлено автором на основе [5]

При анализе больших объемов данных процесс изменения положения центроидов может занимать много времени и при этом центроиды отклоняются на не значительные расстояния, поэтому заранее устанавливают порог отклонения центроидов (крайняя точка, до которой ведется итерация) относительно предыдущего положения. Это позволяет сократить время для разбиения массива данных на группы.

После разбиения массива данных на группы используются другие методы Data Mining, для дальнейшего анализа и выяснения причин такого разбиения.

Выводы. Большим плюсом кластеризации данных является возможность сжимать большие объемы данных в компактный вид и отображать результаты в виде диаграмм, графиков или схем.

Так же кластерный анализ позволяет производить разбиение по нескольким параметрам, параметры задаются заранее перед проведением анализа. Метод к-средних рассматривает наборы данных различных моделей и, при этом, не имеет значения какую информацию необходимо проанализировать. Тем самым он отличается от других математико-статистических методов, которые ограничивают рассмотрение объектов из-за их природы. Такие ограничения затрудняют применять стандартные экономические подходы в прогнозировании конъюнктуры, особенно для объектов с разнородными показателями.

Литература

1. Барсегян А. А., Куприянов М. С., Степаненко В. В., Холод И. И. Методы и модели анализа данных: OLAP и Data Mining — СПб: БХВ-Петербург, 2004. — 336 с.

2. Сегаран Т. Программируем коллективный разум. — Пер. с англ. — СПб: Символ-Плюс, 2008. — 368 с.

3. Информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных — http://www.machinelearning.ru

4. Чубукова И. А. Курс лекций «Data Mining», Интернет-университет информационных технологий www.intuit.ru/department/database/datamining

5. Википедия Метод к-средних https://ru.wikipedia.org/wiki/Метод_k-средних.

6. Интеллектуальный анализ данных http://intellect-tver.ru/?p=265

7. Кластеризация: алгоритмы k-means и c-means https://habrahabr.ru/post/67078/