УДК 004.852

АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ В ЗАДАЧАХ МАШИННОГО ОБУЧЕНИЯ

Бабаев А.М.
студент,
Институт сферы обслуживания и предпринимательства (филиал) ДГТУ в г. Шахты,
Шахты, Россия

Аннотация.
В статье рассмотрены наиболее популярные алгоритмы кластеризации, применяемые для решения задач машинного обучения и анализа данных: k-means, иерархическая кластеризация и mean shift. Описан процесс поиска кластеров с помощью данных методов. Проанализированы их достоинства, недостатки и области применения. Также каждый из алгоритмов был протестирован в задаче кластеризации социального графа сети Вконтакте.
Ключевые слова: машинное обучение, кластеризация, k-means, иерархическая кластеризация, mean shift.

CLUSTERIZATION ALGORITHMS IN TASKS OF MACHINE LEARNING

Babaev A.M.
student,
Institute of Service and Entrepreneurship (branch) of DSTU in Shakhty,
Shakhty, Russia

Abstract.
The article discusses the most popular clustering algorithms used to solve problems of machine learning and data analysis: k-means, hierarchical clustering and mean

shift. The process of searching for clusters using these methods is described. Their advantages, disadvantages and areas of application are analyzed. Each of the algorithms was also tested in the task of clustering the social graph of the Vkontakte network.
Keywords: machine learning, clustering, k-means, hierarchical clustering, mean shift.

Обучение без учителя – это термин, используемый в машинном обучении и обозначающий широкий спектр задач, а также решающие их модели, которые обучаются на неразмеченных данных и ищут шаблоны в них, не прибегая к известной целевой переменной [1]. Примерами проблем, решаемых алгоритмами обучения без учителя являются ранжирование выдачи поисковых сайтов и сегментация изображений. Наиболее обширным классом технологий обучения без учителя являются методы кластеризации.

Кластеризация – это задача распределения данных по кластерам на основе информации об их сходстве. При этом объекты, принадлежащие к одному кластеру, должны быть более похожи друг на друга, чем объекты соседних кластеров [2]. Наряду с классификацией и регрессионным анализом кластеризация является одной из основных задач интеллектуального анализа данных и включает в себя проблемы анализа текста, распознавания образов, сжатия данных. Существует большое количество алгоритмов кластеризации и их модификаций. В данной работе будут рассмотрены три наиболее широко применяемых из них: k-means, иерархическая кластеризация и mean shift.

K-means (кластеризация методом k-средних) – это наиболее популярный алгоритм кластеризации на основе центроида. Также относится к классу методов неиерархической кластеризации [3]. Для построения кластеров методом k-means необходимо выполнить следующую последовательность шагов:

1. Выбрать k – количество кластеров, которое требуется найти.

2. Случайным образом сгенерировать центры каждого кластера.

3. Связать с каждым кластером ближайшие к нему объекты на основе одной из метрик расстояния, в качестве которой чаще всего используется евклидово расстояние (формула 1):

$$d_i(x,y) = \sum_{j=1}^{m}(x_j - y_j)^2 \quad (1)$$

где i – индекс объекта в выборке; j – это индекс признака объекта в выборке. Таким образом, в k-means схожесть объектов определяется путем вычисления расстояния между ними.

4. Вычислить инерцию кластера (формула 2):

$$SSE = \sum_{i=1}^{n}\sum_{j=1}^{k}(x_i - \mu_j)^2 \quad (2)$$

где n – число объектов в выборке; k – число кластеров.

5. Новые центры будут рассчитываться как среднее значение координат точек, принадлежащих кластеру на предыдущем шаге.

6. Вернуться к шагу 3.

Пример кластеризации социального графа для социальной сети Вконтакте алгоритмом k-means приведен на рисунке 1. Алгоритм сумел выделить три наиболее явных кластера, каждый из которых содержит людей, живущих в трех разных городах.

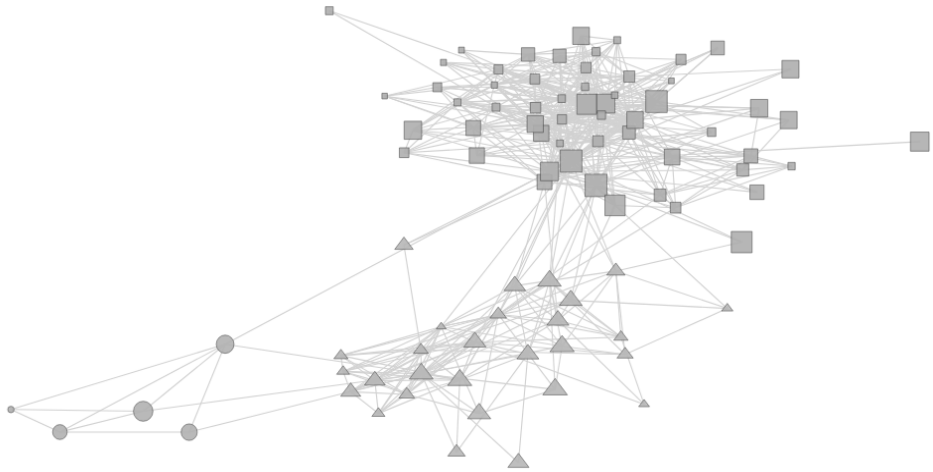

Рис. 1 – Кластеризация социального графа Вконтакте методом k-means

Достоинствами алгоритма являются простота реализации и высокая вычислительная эффективность. Однако при работе с данным алгоритмом возникает ряд проблем:

– высокая чувствительность к выбросам в данных;

– высокая чувствительность к генерации начальных центров кластеров;

– необходимо подбирать количество кластеров;

– низкая эффективность при работе с данными, имеющими несферические кластеры;

– чувствительность к шкале, по которой измеряются значения признаков (для повышения качества работы алгоритма необходимо выполнить нормализацию признаков).

Иерархическая кластеризация – это алгоритм, который также основан на настройке положения центров кластеров. Особенность алгоритма заключается в том, что в результате его работы будут получены не отдельные непересекающиеся кластеры, а иерархия кластеров в виде дерева. Таким образом, в отличие от алгоритма k-means, иерархическая кластеризация не требует указания количества кластеров. Иерархическая кластеризация позволяет не строить все уровни дерева кластеров, а остановиться, получив необходимое для исследования количество кластеров [4].

Существует два подхода к этому типу кластеризации:

– нисходящий – этот метод начинает работу с объединения всех точек данных в одном кластере. Затем он итеративно разбивает кластер на два кластера меньших размеров, пока каждый из них не будет содержать только одну выборку.

– восходящий – этот метод начинает работу с того, что каждый объект выборки представляет собой отдельный кластер. Затем алгоритм итеративно объединяет кластеры, расположенные близко друг к другу, пока не будет

получен один кластер, объединяющий все объекты выборки. Большинство алгоритмов иерархической кластеризации основано именно на этом подходе.

Кроме того, данный алгоритм менее чувствителен к выбросам и выбору метрики расстояния. Но для получения лучших результатов его следует применять к данным, имеющим иерархическую структуру.

Пример работы иерархической кластеризации приведен на рисунке 2. Визуально полученный результат выглядит неудовлетворительно из-за разделения среднего кластера. Однако, в данном случае алгоритм объединил людей в кластеры не по признаку местожительства, а по совпадению места учёбы. Таким образом, объекты, отмеченные треугольным маркером – это люди, проживающие в разных городах, но обучающиеся в одном учебном заведении.

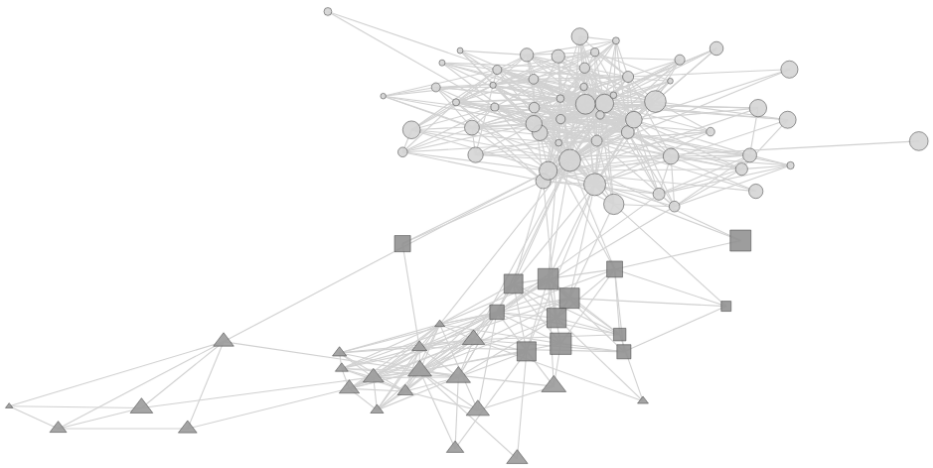

Рис. 2 – Результат иерархической кластеризации социального графа Вконтакте

Mean Shift (кластеризация методом сдвига среднего) – это подход, при котором центр каждого кластера перемещается в самую плотную область в его окрестности [5].

На начальном этапе работы алгоритма необходимо сгенерировать точки-кандидаты на роль центров кластеров. В случае небольшой выборки, кандидатом является каждый объект. При большом объеме выборке кандидаты

генерируются как узлы сетки, распределённые по всему пространству признаков.

Каждая точка-кандидат становится центром скользящего окна – окружности с фиксированным радиусом. На каждой итерации центр окна перемещается в область с большей плотностью точек. В качестве нового центра выбирается среднее значение всех точек в скользящем окне. Процесс прекращается, когда последующие смещения приводят к уменьшению количества объектов в пределах скользящего окна. При этом пересекающиеся скользящие окна необходимо фильтровать, выбирая область с самым большим количеством объектов. В конце работы алгоритма останется набор точек, которые и станут центрами новых кластеров, а точки данных будут группироваться в соответствии со скользящим окном, в котором они находятся.

В отличие от алгоритма k-means, mean shift позволяет находить кластеры произвольной формы. Кроме того, он не требует информации о количестве кластеров. Однако является более дорогостоящим с точки зрения вычислений в связи с долгим итеративным процессом смещения кандидатов. По этой причине работа алгоритма с многомерными данными является очень затруднительной.

На рисунке 3 приведен результат кластеризации социального графа Вконтакте. Алгоритм нашел два кластера, которые, как и случае k-means, можно обосновать географическим положением места жительства людей.

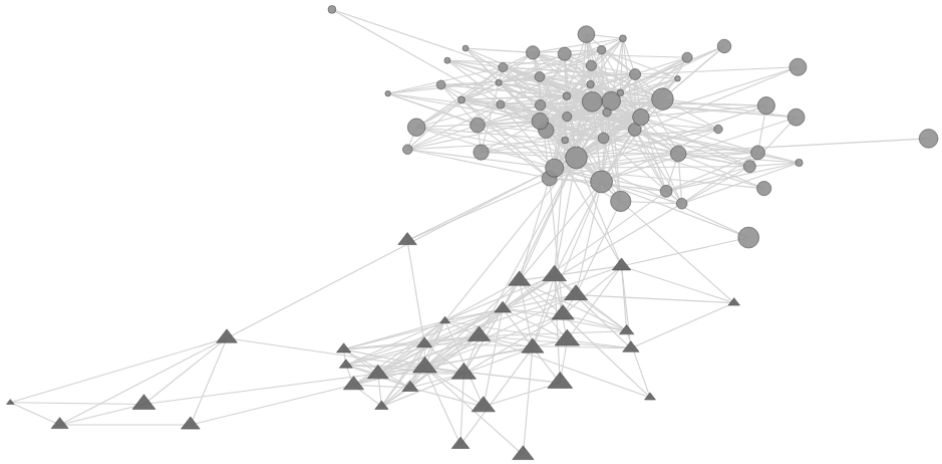

Рис. 3 – Результат кластеризации социального графа Вконтакте методом mean shift

В статье были проанализированы популярные алгоритмы кластеризации. Наиболее вычислительно эффективным и простым в реализации является метод k-means, но он подойдет только для очищенных данных со сферическими кластерами. Иерархическая кластеризация позволит восстановить иерархию кластеров данных, не заботясь о выборе метрики расстояния и количества кластеров для поиска. А для анализа данных с произвольной формой кластеров и небольшим количеством признаков лучшим алгоритмом станет mean shift. И также следует помнить, что однозначно лучшей кластеризации не существует, так как она определяется областью применения и целями исследования.

Библиографический список

1. Бринк Х. Машинное обучение / Х. Бринк, Д. Ричардс, М. Феверолф. – СПб.: Питер, 2017. – 336 с.

2. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / пер. с англ. А.А. Слинкина. – М.: ДМК Пресс, 2015. – 400 с.

3. Youguo L. A Clustering Method Based on K-Means Algorithm / L. Youguo, W. Haiyan // Physics Procedia. – 2012. – №25. – Pp. 1104-1109.

4. Murtagh F. Methods of Hierarchical Clustering / F. Murtagh, P. Contreras // ArXiv. – 2011. – Pp. 1-21. URL: https://arxiv.org/pdf/1105.0121.pdf (дата обращения: 30.11.2019)

5. Anand S. Semi-Supervised Kernel Mean Shift Clustering / S. Anand, S. Mittal, O. Tuzel, P. Meer // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2014. – № 36 (6). – Pp. 1201-1215.

Оригинальность 98%