© 2018 г. П.А. ЕРЕМЕЙКИН, студент НИУ ВШЭ
Б.Г. МИРКИН, док. техн. наук
(Национальный исследовательский университет
«Высшая школа экономики», Москва)

СОКРАЩЕНИЕ ВРЕМЕНИ ВЫЧИСЛЕНИЙ В АГЛОМЕРАТИВНОМ КЛАСТЕР-АНАЛИЗЕ ПЕРЕХОДОМ К ПОДВЫБОРКАМ[1]

Краткая аннотация статьи или заметки. Иногда не бывает. Хх хххххххххххх хххххх хххххххх х хххххххххххххх ххх хххххххх хххх. Хххххххххххххх ххх ххххх хххххх х хххххххххххххх ххх хххххххх хххх.

1. Введение

Методы кластеризации широко применяются для выявления структуры данных и поиска характерных групп объектов. По принадлежности заданного нового объекта к определённому кластеру можно сделать предположения о ключевых свойствах этого объекта. В общем случае под кластеризацией понимают поиск в заданном множестве непересекающихся однородных подмножеств, которые включают в себя подобные объекты [1].

Наиболее широко известный и часто применяемый метод кластеризации — k-means [2]. Этот метод состоит в попеременной минимизации квадратичного критерия по двум типам переменных: центрам кластеров и принадлежности объектов кластеру. Критерии минимизации аналогичного вида выражают смысл некоторых популярных иерархических алгоритмов, например, Ward [3] или Bisecting K-Means[4].

Несмотря на популярность алгоритма k-means, он обладает существенными недостатками. Во-первых, для работы необходимо явно задать число кластеров. В практических приложениях подлинное число кластеров как правило неизвестно. Во-вторых, k-means требует правильной инициализации начального состояния, от которого существенным образом зависит результат. И, наконец, алгоритм чувствителен к шуму в данных, то есть он не предусматривает никакого механизма учёта возможных погрешностей, которые зачастую возникают в реальных данных, полученных при помощи измерений.

За свою богатую историю алгоритм k-means получил множество усовершенствований, модификаций и новых применений. В частности, существенным вкладом в развитие алгоритма стала работа [5], в которой предложен так называемый метод аномальных кластеров. Метод аномальных кластеров позволяет рационально определить начальное состояние для алгоритма k-means путём поочерёдного выявления и исключения кластеров, наиболее удалённых от центра данных, называемых аномальными.

[1]

Иерархические алгоритмы, основанные на k-means, в основном наследуют его недостатки, но привносят важное свойство: в ходе работы они выявляют дерево вложенности кластеров, которое может быть естественным образом использовано в некоторых приложениях. Например, такое дерево может отражать филогенетическое родство при кластеризации биологических видов. Это интересное свойство побуждает исследователей искать пути устранения недостатков иерархических алгоритмов, использующих квадратичный критерий.

Как показали результаты экспериментов, проведённых в работе [?], метод аномальных кластеров в большинстве случаев порождает избыточное число кластеров, поэтому его можно использовать как предварительный шаг для агломеративного алгоритма Ward. Алгоритм Ward исходит из представления о том что на начальном этапе всякий единичный объект выступает в роли отдельного кластера и на каждом шаге происходит объединение двух ближайших кластеров пока их общее число не достигнет заданного значения. Несмотря на то, что Ward в каноническом виде не требует инициализации, применение описанного предварительного шага вызвано необходимостью повышения производительности. При большом количестве объектов, на первых итерациях алгоритма требуется выполнить большое число сравнений (квадратично зависящее от числа объектов), и следовательно, время работы алгоритма в этом случае недопустимо велико. Благодаря применению предварительной проработки кластерной структуры с помощью аномального анализа отпадает необходимость сравнивать большое число кластеров в которых содержится всего по одному или несколько объектов.

Описанная выше модификация получила название A-Ward. Есть еще A-Ward$_{p\beta}$, которая суть то-то. Она ооочень крутая, но вот фиг знает как выбрать эти самые $p\beta$. А вто тут то нам приходит на помощь статься [такая-то] в котрой предложен тупой перебор на основе SW. Но тупой перебор это тупо. Поэтому я предлагаю крутой эксперимент который покажет можем ли мы по части объектов определить свойства всей выборки.

объяснение проблемы и полезности ее решения (агломеративный кластеринг - заслуженный метод; но долгий; АУорд - способ ускорения вычислений за счет ускорения самой неприятной части; тем не менее - довольно долгий. Поэтому возникает идея - применять метод на подвыборке. Сложности реализации идеи: бла-бла)

2. Предлагаемое решение

(со всеми деталями, включая АУорд, Силуэт Видт и пр.)

3. Организация эксперимента

4. Результаты эксперимента

С пояснениями и выводами

5. Заключение

Что сделано и куда двигаться.

1) 3 разных числа кластеров соответствуют числу признаков $>, <, \approx$ признаков
2) SW работает не хуже ARI
3) Сокращение по времени есть

СПИСОК ЛИТЕРАТУРЫ

1. *Миркин Б.Г.* Введение в анализ данных. М.: Юрайт, 2015.
2. *Ball G.H., Hall D.J.* A clustering technique for summarizing multivariate data, Behavioral Science. 1967. V. 12 Iss. 2 P. 153–155.
3. *Joe H., Ward Jr.* Hierarchical grouping to optimize an objective function, Journal of the American Statistical Association. 1963. V. 58 Iss. 301 P. 236–244.
4. *Mirkin B.* Clustering: A Data Recovery Approach. London: CRC Press, 2012.
5. *Chiang M.M.-T., Mirkin B.* Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads, Journal of Classification. 2010. V. 27 Iss. 1. 3–40.

Еремейкин П.А., *Национальный исследовательский университет «Высшая школа экономики», студент, Москва*, eremeykin@gmail.com

Миркин Б.Г., *Национальный исследовательский университет «Высшая школа экономики», профессор, Москва*, bmirkin@hse.ru