© 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 [6], 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. В той же работе [6] предложено дальнейшее усовершенствование алгоритма A-Ward, которое акцентирует внимание на обработке зашумлённых данных. Обобщённая версия алгоритма для произвольной степени Минковского p и с использованием весовых коэффициентов признаков w, определяемых отдельно для каждого кластера, была названа A-Ward$_{p\beta}$, где β обозначает степень весовых коэффициентов w.

Численные эксперименты на синтетических и реальных данных показали высокую эффективность алгоритма A-Ward$_{p\beta}$ в том числе для случаев с большой зашумленностью признаков, поэтому алгоритм представляет интерес для применения в практических случаях. Тем не менее, ввод новых параметров p и β породил необходимость выработки методики для определения их значений. В статье [7] рассмотрен подход к определению p и β методом перебора, в котором критерием качества результата при заданных значениях служит эмпирическая характеристика Silhouette Width (SW)[8]. Такой подход поглощает преимущество в производительности, достигнутое введением предварительного шага аномального кластер-анализа, и требует больших временных затрат.

Специалистам по анализу данных хорошо известна идея, которая лежит в основе принципа кросс-валидации [**?**] и заключается в сохранении основных свойств выборки даже в случае исключения из неё некоторой части объектов. Та же идея предположительно может быть использована для определения рациональных значений p и β по подвыборкам. Таким образом, в данной статье будет экспериментально исследована возможность выбора параметров p и β для алгоритма A-Ward$_{p\beta}$ путём перехода к подвыборкам и с использованием характеристики SW в качестве критерия качества результата.

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.
6. *de Amorim R.C., Makarenkov V., Mirkin B.* A-Ward$_{p\beta}$: Effective hierarchical clustering using the Minkowski metric and a fast k-means initialisation. Information Sciences. 2016. V. 370-371 P. 343–354.
7. *de Amorim R.C., Shestakov A., Mirkin B., Makarenkov V.* The Minkowski central partition as a pointer to a suitable distance exponent and consensus partitioning. Pattern Recognition. 2017. V. 67 P. 62–72.
8. *Rousseeuw P.* Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987. V.20 P. 53–65.

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