© 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.

1) Написать аннотацию

2) Проверить ссылки на литературу и добавить в библиографию недостающие

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

[1]

и исключения кластеров, наиболее удалённых от центра данных, называемых аномальными.

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 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

2.1. Описание эксперимента

Пусть задано множество Y из N объектов, каждый из которых обладает V признаками. Такое множество можно выразить в виде таблицы данных:

$$Y = \begin{pmatrix} y_1 \\ \cdots \\ y_N \end{pmatrix} = \begin{pmatrix} y_{11} & \cdots & y_{1V} \\ \cdots & \cdots & \cdots \\ y_{N1} & \cdots & y_{NV} \end{pmatrix} \tag{1}$$

Алгоритм A-Ward$_{p\beta}$ при заданных значениях параметров p и β позволяет получить кластерное разбиение $S = \{C_1, \ldots, C_K\}$ этого множества. Получаемое кластерное разбиение состоит из заранее определённого числа K непересекающихся кластеров C_k, объединение которых покрывает всё множество объектов Y. Принцип работы и формальное описание алгоритма A-Ward$_{p\beta}$ приведены в разделе 2.2.

Для произвольного множества Y до сих про не было сформулировано эффективной методики поиска параметров p и β. Как было отмечено во введении, в статье [7] был опробован способ выбора параметров путём многократного запуска алгоритма A-Ward$_{p\beta}$ при переборе значений параметров в диапазоне $[1, 5]$ с шагом 0.1. Указанный диапазон определён исходя из опыта применения алгоритма: с превышением значений параметров 5 не достигается существенного улучшения качества разбиения. Нижняя граница диапазона определена математическим смыслом параметров.

Описанный способ потребует $41^2 = 1681$ запуск алгоритма, что чрезвычайно затратно с точки зрения времени. Поэтому была предложена идея для определения значений параметров по случайным подвыборкам. Из всего множества Y выбирается L подмножеств $Y_l \subset Y$ с заданным числом элементов $n \ll N$. По каждому подмножеству Y_l оцениваются рациональные значения параметров p_l^* и β_l^*. Усреднённые по l значения $p^* = \overline{p_l^*}$ и $\beta^* = \overline{\beta_l^*}$ принимаются в качестве рационального выбора для всего множества Y. При правильном выборе соотношения между числом объектов в полной выборке N, числом объектов в подвыборке n, и числом подвыборок M, как ожидается, можно получить результат кластеризации близкий к результату, полученному с помощью оценки по всей выборке N, затратив при этом существенно меньшее время.

Для подтверждения вышеописанных предположений предлагается рассмотреть численный эксперимент на синтетических данных. Синтетическая генерация позволяет гибко изменять характеристики данных, например число признаков и количество кластеров, степень их взаимного смешивания, а также определять истинное разбиение. Для генерации данных применяется метод описанный в статье [**?**]. Подробно его принцип работы разобран в разделе 2.3. Сейчас стоит иметь ввиду, что результат работы данного генератора синтетических данных полностью определяется следующими пятью параметрами:

1) Число объектов N

2) Число признаков V

3) Количество кластеров K

4) Минимальное число объектов в кластере m

5) Степень взаимного смешивания кластеров $a \in [0, 1]$

Данные сгенерированные по методике, описанной в [?] мы будем обозначать ключевым словом `kovaleva`, за которым через подчёркивание следуют обозначения размерности (например `1000` $\times$ `15`), общего числа кластеров с префиксом `c`, минимального числа объектов в кластере с префиксом `m` и, наконец, степени взаимного смешивания с префиксом `a`. В таблице 1 приведены три типа данных которые будут использованы в ходе эксперимента с пояснениями в принятых обозначениях. Данные, сгенерированные по указанным значениям параметров, обладают важным свойством: они соответствуют трём случаям, взаимного отношения числа признаков и количества кластеров. В первом случае число кластеров меньше числа признаков, во втором случае примерно равно, а в третьем — больше.

Таблица 1. Параметры данных

Обозначение	N	V	K	m	a
`kovaleva_1000` $\times$ `15_c7_m100_a0.5`	1000	15	7	100	0.5
`kovaleva_1000` $\times$ `15_c12_m60_a0.5`	1000	15	12	60	0.5
`kovaleva_1000` $\times$ `15_c19_m35_a0.5`	1000	15	19	35	0.5

В ходе эксперимента будет рассмотрены четыре схем формирования подвыборки (без учёта полной выборки как эталонного случая). Первые две схемы представляют собой однократный выбор соответственно по 100 и 200 объектов, два оставшиеся — пятикратное формирование подвыборок по 100 и 200 объектов с последующим усреднением результата. В таблице 2 приведены все пять рассматриваемых схем.

Таблица 2. Схемы формирования подвыборки

Схема формирования подвыборки	Размер подвыборки n	Число подвыборок L
Полная выборка (1-1000)	1000	1
Однократный выбор по 100 (1-100)	100	1
Однократный выбор по 200 (1-200)	200	1
Пятикратный выбор по 100 (5-100)	100	5
Пятикратный выбор по 200 (5-200)	200	5

При применении алгоритма A-Ward$_{p\beta}$ на реальных данных для оценки качества результата и выбора наилучших значений p, β нет возможности использовать подлинное разбиение, поэтому популярный индекс ARI [?] не подходит для применения в роли целевой характеристики. В качестве замены этому индексу может быть использована эмпирическая величина Silhouette Width (SW)[8], хорошо зарекомендовавшая себя во многих приложениях [?, ?]. Интересно также установить насколько использование эмпирической величины SW ухудшает результат относительно результата, полученного с использованием подлинного разбиения и индекса ARI. Синтетическая

генерация данных позволит применить индекс ARI для оценки разбиений, получаемых при различных значениях параметров p, β, относительно истинного разбиения. Способы вычисления характеристики SW и индекса ARI описаны в разделах 2.4 и 2.5 соответственно.

При заданной характеристике качества может возникнуть ситуация неоднозначности при которой максимум достигается для нескольких пар параметров p и β. Для разрешения этой неоднозначности в рамках эксперимента предлагается рассмотреть два подхода:

1) Предпочтение отдаётся паре параметров p, β для которой величина $p^2 + \beta^2$ минимальна

2) Пусть $\{(p_t, \beta_t) : t = 1, \ldots, T\}$ — множество пар параметров для которых достигается максимум выбранной характеристики качества (SW или ARI). Рассчитаем поэлементные средние значения $\overline{p} = \frac{1}{T}\sum_{t=1}^{T} p_t$, $\overline{\beta} = \frac{1}{T}\sum_{t=1}^{T} \beta_t$. Наилучшим выбором признается та пара, для которой значение $(p_t - \overline{p})^2 + (\beta_t - \overline{\beta})^2$ минимально.

Ниже приведена формулировка алгоритма подбора параметров p, β, используемого в эксперименте. Описанный алгоритм применяется для синтетических данных из таблицы 1 с использованием в общей сложности пяти схем формирования подвыборки согласно таблице 2. При этом рассматриваются две характеристики качества разбиения: эмпирическая SW и индекс ARI, основанный на известном истинном разбиении. Для разрешения неопределённости выбора максимума характеристики рассматриваются два подхода, описанных в списке выше.

Алгоритм 1 (Подбор параметров p, β с переходом к подвыборкам).

1. Из множества Y случайным образом выбрать L подмножеств $Y_l \subset Y$ с заданным размером n. Пересечение выбранных подмножеств допускается.

2. Для всех значений параметров $p = 1, 1.1, \ldots, 5$ и $\beta = 1, 1.1, \ldots, 5$ выполнить алгоритм A-Ward$_{p\beta}$ применительно к каждому подмножеству Y_l и получить результирующее разбиение $S_l^{p\beta}$.

3. По каждому полученному разбиению $S_l^{p\beta}$ рассчитать значение характеристики Silhouette Width (или ARI): $SW_l^{p\beta} = SW(S_l^{p\beta})$.

4. Для каждого подмножества Y_l выбрать рациональные значения параметров p и β, соответствующие максимальным значениям характеристики Silhouette Width (или ARI): $(p_l^*, \beta_l^*) = \arg\max\{SW_l^{p\beta}\}$. В случае, если максимум достигается для нескольких пар параметров, рассмотреть два возможных подхода разрешения неоднозначности.

5. Усреднить полученные значения по l: $p^* = \overline{p_l^*}$ и $\beta^* = \overline{\beta_l^*}$. Получен результат алгоритма подбора параметров.

2.2. Алгоритм A-Ward$_{p\beta}$

Рассмотрение алгоритма A-Ward$_{p\beta}$ следует начинать с описания общего принципа работы иерархических алгоритмов и, в частности, Ward. В отличие от неиерархических алгоритмов, которые формируют плоскую структуру кластеров, как, например,

k-means, иерархические алгоритмы в качестве результата предоставляют дополнительную информацию относительно взаимосвязей между кластерами. Эта информация выражена в виде вложенной последовательности разбиений [6] и графически может быть изображена с помощью дендрограммы. На каждом уровне дендрограммы определённый объект $y_i \in Y$ принадлежит единственному кластеру C_k. В то же время этот объект может принадлежать на других уровнях другим кластерам, которые образованы разбиением кластера C_k или слиянием с другими кластерами.

Различают два вида иерархический алгоритмов: агломеративные (объединяющие) и дивизивные (разделяющие). Агломеративные алгоритмы работают по принципу "снизу верх". На начальном этапе работы таких алгоритмов каждый единичный объект выступает в роли кластера. Во время работы происходит попарное объединение кластеров до тех пор, пока не будет достигнуто заданное число кластеров. Дивизивные алгоритмы в противоположность агломеративным действуют "сверху вниз".

Алгоритм Ward относится к агломеративным. На каждой итерации происходит слияние ближайших кластеров, таким образом чтобы внутрикластерная дисперсия была минимальной. Близость кластеров определяется по формуле:

$$d_{Ward}(C_a, C_b) = \frac{N_a N_b}{N_a + N_b} \sum_{v=1}^{V} (c_{av} - c_{bv})^2, \tag{2}$$

где C_a, C_b — два произвольных кластера,
N_a, N_b — количество объектов в кластерах C_a и C_b соответственно
V — число признаков у каждого объекта $y_i \in Y$
c_{va}, c_{vb} — v-ая координата центров кластеров C_a и C_b соответственно

В традиционной формулировке алгоритмы Ward формулируется следующим образом:

Алгоритм 2 (Ward).

1. *Инициализация.* Задаться требуемым количеством кластеров K. Установить текущее число кластеров $k = N$ и текущее разбиение $S = \{C_1, \ldots, C_N\}$, где каждый кластер состоит из единственного объекта $C_k = \{y_k\}$. Центр кластера совпадает с объектом, входящим в этот кластер.

2. *Слияние.* Определить два ближайших кластера C_a, C_b, используя формулу 2. Произвести слияние найденных кластеров в результирующий кластер $C_{ab} = C_a \cup C_b$, содержащий одновременно объекты C_a и C_b.

3. *Обновление центра.* Установить центр кластера C_{ab} равным покомпонентному среднему всех объектов, входящих в C_{ab}.

4. *Условие остановки.* Уменьшить текущее количество кластеров k на единицу. Если $k > 1$ или $k > K$, перейти к шагу 2. В противном случае текущее разбиение S является результирующим.

2.3. Генератор данных типа kovaleva

2.4. Характеристика разбиения SW

2.5. Характеристика разбиения ARI

данные способ оценки что включено в сравнение (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