© 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]

Кластер-анализ находит широкое применение в различных областях деятельности. Наиболее популярный в настоящее время алгоритм кластеризации k-means обладает существенными недостатками, часть из которых удаётся успешно устранить благодаря современным исследованиям, предлагающим модернизации алгоритма. Одним из эффективных алгоритмов, предложенных в последнее время является A-Ward$_{p\beta}$. Как любое новое решение, алгоритм A-Ward$_{p\beta}$ требует выработки рекомендаций по применению. Данная статья посвящена поиску способов определения параметров, задействованных в алгоритме. Ранее предлагаемые способы опираются на перебор всех возможных значений параметров в определённом интервале и требуют больших временных затрат. Для ускорения подбора параметров предлагается произвести переход к подвыборкам, существенно меньшего размера чем исходные данные и производить перебор для подвыборок. Авторы статьи предлагают рассмотреть различные способы формирования подвыборок и оценить эффективность этих способов с точки зрения затрачиваемого времени и качества результирующего разбиения.

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,

[1]Тут пишут про финансовую поддержку.

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.

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 резуль-

тата 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 [9] 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 позво-

ляет 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 [10]. 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 [10] мы будем обозначать ключевым словом `kovaleva`, за которым через подчёркивание следуют обозначения размерности (например $\texttt{1000} \times \texttt{15}$), 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 × 15_c7_m100_a0.5`	1000	15	7	100	0.5
`kovaleva_1000 × 15_c12_m60_a0.5`	1000	15	12	60	0.5
`kovaleva_1000 × 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.

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 [11] 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 [12, 13]. 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.

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
Пятикрантый выбор по 100 (5-100)	100	5
Пятикрантый выбор по 200 (5-200)	200	5

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. Результатом работы алгоритма являются усреднённые по l значения параметров $p^* = \overline{p_l^*}$ и $\beta^* = \overline{\beta_l^*}$.

2.2. Алгоритм A-Ward$_{p\beta}$

Описание алгоритма A-Ward$_{p\beta}$ следует начать с рассмотрения общего принципа работы иерархических алгоритмов. 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, усовершенствованием которого является алгоритм A-Ward$_{p\beta}$, относится к агломеративным. На каждой итерации происходит слияние ближайших кластеров, таким образом чтобы внутрикластерная дисперсия была минимальной. Исходя из принципа работы алгоритма Ward видно, что на начальных этапах происходит сравнение большого числа кластеров, примерно равного числу объектов, что требует продолжительного время вычисления. Решением этой проблемы является метод аномальных кластеров, выполняющий роль инициализации. Инициализация осуществляется в две стадии: на первой выявляются аномальные кластеры (алгоритм 2), а на второй происходит их стабилизация (алгоритм 3). Ниже описан алгоритм взвешенной аномальной кластеризации для общего случая степени Минковского p и степени весовых коэффициентов β.

Алгоритм 2 (Инициализация аномальными кластерами).

1. *Инициализация.* Задаться значениями параметров p и β. Глобальный центр данных c_Y вычислить как покомпонентный центр Минковского по всем объектам $y_i \in Y$.

2. *Текущий центр.* Задать пустой аномальный кластер $C_t = \varnothing$. Веса равномерно распределить по всем признакам $w_{kv} = 1/V$ при $k = 1, 2$ и $v = 1, \ldots, V$. Текущий центр аномального кластера c_t выбрать как объект, наиболее удалённый от глобального центра c_Y. Расстояние между объектом y_i и центром произвольного кластера c_k вычисляется по формуле:

$$d_{p\beta}(y_i, c_k) = \sum_{v=1}^{V} w_{kv}^{\beta} |y_{iv} - c_{kv}|^p \tag{2}$$

3. *Формирование аномального кластера.* В аномальный кластер добавить объекты, которые расположены ближе к текущему центру аномального кластера c_t, чем

к глобальному центру c_Y согласно расстоянию, определяемому формулой 2. Если изменений в разбиении нет, прейти к шагу 6.

4. *Обновление текущего центра.* Вычислить текущий центр аномального кластера c_t как покомпонентный центр Минковского по всем объектам в аномальном кластере $y_i \in C_t$.

5. *Обновление весов.* Вычислить веса признаков по следующей формуле:

$$w_{kv} = \frac{1}{\sum_{u=1}^{V} \left(\frac{D_{kv}}{D_{ku}} \right)^{\frac{1}{\beta-1}}}, \tag{3}$$

где $D_{kv} = \sum_{i \in C_k} |y_{iv} - c_{kv}|^\beta$ — разброс признака v в кластере C_k

6. *Сохранение параметров.* Включить текущий центр аномального кластера c_t в список центров `c_list`, а веса w в список весов `w_list`.

7. *Исключение аномального кластера.* Исключить из Y все объекты $y_i \in C_t$. Если $Y \neq \varnothing$, перейти к шагу 2.

8. *Результат.* Результатом работы алгоритма является разбиение S, а также списки центров кластеров `c_list` и весов `w_list`.

Характерная особенность структуры данных, генерируемой алгоритмом 2 состоит в том, что, во-первых, число получаемых кластеров всегда больше, чем их действительное количество, а во-вторых, что эта структура сгущается ближе к центру данных. Для смягчения второй особенности применяется вариация k-means, которая использует те же параметры степени Минковского p и степени весовых коэффициентов β, а также начальное состояние, порождённое алгоритмом аномальной инициализации. Эта вариация получила название *imwk*-means$_{p\beta}$ и описана ниже.

Алгоритм 3 (*imwk*-means$_{p\beta}$).

1. *Инициализация.* Установить текущее разбиение пустым $S = \varnothing$, а число кластеров K равным длине списка `c_list`, который был получен при аномальной инициализации.

2. *Формирование кластеров.* Каждый объект $y_i \in Y$ поместить в кластер, центр которого c_k находится ближе всего к этому объекту. Близость объекта к центру кластера определяется по формуле 2. Если нет изменений в разбиении S, перейти к шагу 5.

3. *Обновление центров.* Вычислить новые координаты центра c_k каждого кластера C_k как покомпонентный центр Минковского всех объектов этого кластера $y_i \in C_k$.

4. *Обновление весов.* Вычислить новые веса w_{kv} по формуле 3 для $k = 1, \ldots, K$ и $v = 1, \ldots, V$. Перейти к шагу 2

5. *Результат.* Результатом работы алгоритма является разбиение S, а также списки центров кластеров `c_list` и весов `w_list`.

Итоговое разбиение строится агломеративным методом, обеспечивая заданное число кластеров. Модификация алгоритма Word с учетом весов признаков и для произвольной степени Минковского описана ниже.

Алгоритм 4 (A-Ward$_{p\beta}$).

1. *Инициализация.* Параметры p и β остаются неизменными, которые были определены для *imwk*-means$_{p\beta}$. Начальное состояние соответствует конечному для *imwk*-means$_{p\beta}$: исходный список центров кластеров `c_list` и весов `w_list` является результатом работы предыдущего этапа.

2. *Объединение кластеров.* Выбрать два ближайших кластера C_a, $C_b \in S$ и объединить их в новый C_{ab}. Близость кластеров определяется по следующей формуле:

$$d_{Ward}(C_a, C_b) = \frac{N_a N_b}{N_a + N_b} \sum_{v=1}^{V} \left(\frac{w_{av} + w_{bv}}{2} \right)^{\beta} |c_{av} - c_{bv}|^p, \tag{4}$$

где N_a, N_b — количество объектов в кластерах C_a и C_b соответственно
V — число признаков у каждого объекта $y_i \in Y$
w_{av}, w_{bv} — веса v-го признака в кластере C_a и C_b соответственно
c_{av}, c_{bv} — v-ая координата центров кластеров C_a и C_b соответственно

3. *Обновление центра.* Вычислить новое значение центра C_{ab} как покомпонентный центр Минковского по всем объектам $y_i \in C_{ab}$.

4. *Обновление весов.* Вычислить новые веса w_{kv} по формуле 3 для $k = 1, \ldots, K$ и $v = 1, \ldots, V$.

5. *Условие остановки.* Уменьшить текущее число кластеров на единицу. Если текуще число кластеров все ещё больше единицы или требуемого числа кластеров, прейти к шагу 2 .

Алгоритм A-Ward$_{p\beta}$ хорошо зарекомендовал себя при тестах как на синтетических, так и на реальных данных. Особый интерес представляют возможности алгоритма при обработки зашумленных данных. Благодаря описанным нововведениям алгоритмом учитывается различные признаки с учётом их дисперсии в каждом кластере.

2.3. Генератор данных

Для генерации данных используется простой, но удобный подход, описанный в работе [10]. Используя единственный параметр, предложенный генератор позволяет регулировать разброс объектов внутри кластера и одновременно взаимное смешивание кластеров.

Структура получаемых данных представляют собой заданное количество гауссовых кластеров K, сформированных при фиксированном общем числе объектов N и признаков V и определённом минимальном количестве объектов в каждом кластере m. Остаток объектов $\delta = N - K \cdot m$ распределяется случайно и равновероятно по всем кластерам.

Кластеры порождается независимо гауссовым распределением. Центры кластеров генерируется как случайный V-мерный вектор с равной вероятностью внутри гиперкуба $[-a, a]^V$. При этом параметр $a \in [0, 1]$ отвечает за степень взаимного смешивания кластеров: чем больше значение параметра, тем сложнее разделить соседние кластеры. Ковариационная матрица кластеров генерируется со случайными диагональными элементами, значения которых равномерно распределены в диапазоне [0.025, 0.05].

2.4. Характеристика разбиения Silhouette Width

Эмпирическая характеристика Silhouette Width (SW) [8] позволяет оценить качество разбиения непосредственно, не опираясь на известное эталонное разбиение, поэтому она хорошо подходит для поиска параметров алгоритма A-Ward$_{p\beta}$ на реальных данных. Для воспроизведения этой ситуации SW будет использована в эксперименте.

Значение характеристики для разбиения S определяется как среднее значение SW для всех объектов $y_i \in Y$. Silhouette Width для одного объекта рассчитывается по следующей формуле:

$$SW(y_i) = \frac{b(y_i) - a(y_i)}{max\{a(y_i), b(y_i))\}} \tag{5}$$

где $a(y_i)$ — среднее расстояние между объектом $y_i \in C_k$ и всеми объектами, принадлежащими тому же кластеру C_k, что и y_i,

$b(y_i)$ — наименьшее среднее расстояние между объектом $y_i \in C_k$ и объектами, которые принадлежат другим кластерам.

Значения характеристики SW лежат в промежутке от -1 до 1. Разбиения, для которых характеристика SW ближе к 1 предпочтительнее тех, у которых SW меньше.

2.5. Индекс ARI

Индекс ARI (Adjusted Rand Index) [11] является популярным способом сравнения эталонного и заданного разбиения. В условиях проводимого эксперимента используются синтетические данные, для которых известно истинное разбиение, поэтому для оценки эффективности применения эмпирической характеристики SW можно задействовать ARI. Формула для вычисления индекса записывается следующим образом:

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}} \tag{6}$$

где $n_{ij} = |C_i \cap C_j|$ — число объектов, входящий одновременно в i-ый кластер в первом разбиении и в j-ый — во втором.

$a_i = \sum_{j=1}^{K} |C_i \cap C_j|$ — число объектов, ходящих в i-ый кластер в первом разбиении

$b_j = \sum_{i=1}^{K} |C_i \cap C_j|$ — число объектов, ходящих в j-ый кластер во втором разбиении

Как и характеристика SW, индекс ARI принимает значения от -1 до 1. ARI достигает 1 только в том случае, если два разбиения совпадают.

3. Результаты эксперимента

Результаты эксперимента приведены на рисунке 1, который состоит из трёх строк и трёх столбцов. Первая строка содержит диаграммы, на которых закрашенными столбцами изображено среднее значение индекса ARI, вычисленного относительно разбиения, полученного как результат работы A-Ward$_{p\beta}$, с параметрами p и β, определёнными по алгоритму 1 с критерием качества SW. Вторая строка отличается от первой тем, что в роли критерия качества вместо SW был использован индекс ARI относительно истинного разбиения. В третьей строке изображены диаграммы для суммарного времени работы A-Ward$_{p\beta}$ на заданных подвыбокрах без учёта вычисления критериев качества. Эксперименты проводились на ПК с процессором Intel®

Core™2 Quad Q9550, работающем на частоте 2.83 ГГц и с объёмом оперативной памяти 4 Гб.

Столбцы на рисунке 1 соответствуют трём вариантам количества кластеров в сгенерированных данных: 7, 12 или 19. Тонкими линиями показаны среднеквадратичное отклонения, вычисленные по 10 экспериментам. Цвет столбца определяет один из двух анализируемых подходов разрешения неопределённости, которые описаны в разделе 2.1. Красный соответствует первому подходу, а синий — второму.

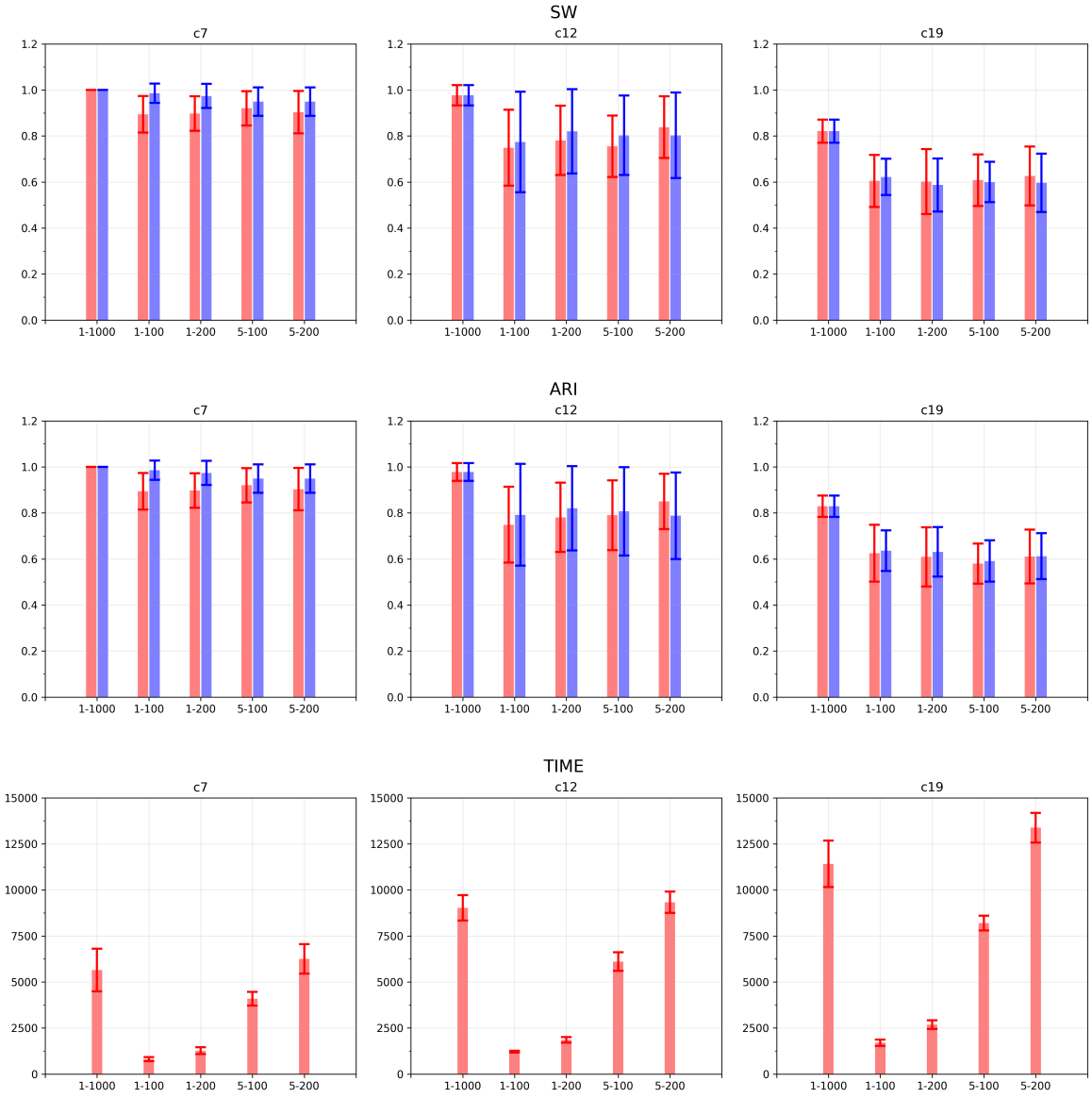

Рис. 1. Диаграммы результатов

Относительно сравнения используемых критериев качества можно сделать вывод, что характеристика SW показывает результаты, сопоставимые с результатами ARI. Наибольшее расхождение средних значений индекса ARI составило $0,044$ для схемы перехода к подвыборкам 1-200 при 19 кластерах и втором подходе разрешения

неопределённости. Таким образом, при выборе параметров A-Ward$_{p\beta}$ на реальных данных, для которых не известно истинное разбиение, применение характеристики SW оправдано.

При сравнении двух подходов разрешения неопределённости выясняется, что второй подход проявляет себя лучше, в среднем на $0,020$ единиц ARI. Исходя из этого, второй подход разрешения неопределённости можно рекомендовать для практического применения.

Существенна экономия времени наблюдается только для однократного формирования выборки по схемам 1-100 и 1-200. Пятикратное формирование не даёт значимого ускорения и в некоторых случаях работает медленнее, чем полная схема 1-1000. В целом, предлагаемое решение является эффективным лишь для небольшого числа кластеров и существенно ухудшает результат относительно анализа по полной выборке в противном случае. Любопытно также отметить, что значение индекса ARI разбиения, полученного по полной выборке 1-1000 для данных с семью кластерами, равно единице для всех 10 экспериментов, что вновь характеризует алгоритм A-Ward$_{p\beta}$ с лучшей стороны.

4. Заключение

Экспериментально проверено предположение о возможности перехода к подвыборкам при определении эффективных параметров алгоритма A-Ward$_{p\beta}$. В эксперименте использованы синтетические данные для трёх случаев числа кластеров: 7, 12 и 19. Стоит отметить, что эти случаи соответствуют трём типам отношения числа кластеров и признаков: число кластеров меньше числа признаков, примерно равно и больше. Для алгоритмов, в которых используется аномальный кластер-анализ, рассмотрение указанных характерных случаев необходимо для разносторонней оценки их возможностей. При выборе наилучшей пары параметров из 1681 комбинаций были проанализированы два подхода разрешения неопределённости относительно нескольких максимумов характеристики качества. Для оценки правомерности выбора параметров по характеристике SW было произведено сравнение качества результирующего разбиения со случаем использования индекса ARI. Рассмотрено четыре схемы формирования подвыборок с однократным и пятикратным выбором по 100 и 200 объектов.

В ходе эксперимента в очередной раз была подтверждена эффективность A-Ward$_{p\beta}$, что делает его привлекательным для дальнейших исследований, тем более что однозначного и полного ответа относительно быстрого выбора параметров p и β по результатам проведённой работы выработать не удалось. Для более разносторонней проработки темы следует провести аналогичные исследования для различных вариантов размерности данных и степени взаимного смешивания кластеров. Тем не менее, были получены следующие ценные выводы:

1) Применение эмпирической характеристики SW для выбора наилучшего разбиения оправдано;

2) При решении практических задач, для выбора пары параметров из определённого множества возможных вариантов следует отдать предпочтение набору с наименьшим отклонением от усреднённых значений;

3) Использование пятикраного формирования выборки теряет смысл ввиду больших временных затрат.

4) Для случая с небольшим числом кластеров схема формирования выборки 1-100 может быть рекомендована как средство выбора рациональных параметров алгоритма и обеспечивает существенный выигрыш по времени.

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. P. 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.
9. **Добавить какую-нибудь каноническую статью по кросс-валидации.**
10. *Kovaleva E., Mirkin B.* Bisecting K-Means and 1D Projection Divisive Clustering: A Unified Framework and Experimental Comparison. Journal of Classification. 2015. V. 32. Iss. 3. P. 414–442.
11. *Hubert L., Arabie P.* Comparing partitions. Journal of Classification. 1985. V. 2. Iss. 1. P. 193–218.
12. **Статья 1, в которой SW хорошо зарекомендовала себя**
13. **Статья 2, в которой SW хорошо зарекомендовала себя**

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