УДК 519.237.8+519.216.5 Вестник СПбГУ. Прикладная математика. Информатика... 2019. Т. 15. Вып. 1
MSC 62H30

Марковский момент остановки агломеративного процесса кластеризации в евклидовом пространстве

А. В. Орехов

Санкт-Петербургский государственный университет, Российская Федерация,
199034, Санкт-Петербург, Университетская наб., 7–9

Для цитирования: *Орехов А. В.* Марковский момент остановки агломеративного процесса кластеризации в евклидовом пространстве // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2019. Т. 15. Вып. 1. С. 76–92. https://doi.org/10.21638/11702/spbu10.2019.106

При обработке больших массивов эмпирической информации или данных большой размерности кластерный анализ является одним из основных методов предварительной типологизации. Это обусловливает в том числе необходимость получения формальных правил для вычисления количества кластеров. В настоящее время наиболее распространенным методом определения предпочтительного числа кластеров является визуальный анализ дендрограмм, но такой подход сугубо эвристический. Выбор множества кластеров и момент завершения алгоритма кластеризации зависят друг от друга. Кластерный анализ данных из n-мерного евклидова пространства методом «одиночной связи» можно рассматривать как дискретный случайный процесс. Последовательности «минимальных расстояний» задают траектории этого процесса. Аппроксимационно-оценочный критерий» (approximation-estimating test) позволяет определить марковский момент, когда характер возрастания такой последовательности изменяется с линейного на параболический, что, в свою очередь, может быть признаком завершения агломеративного процесса кластеризации. Расчет количества кластеров является актуальной проблемой во многих случаях автоматической типологизации эмпирических данных, например в медицине при цитометрическом исследовании крови, автоматическом анализе текстов и в ряде других случаев, когда количество кластеров заранее неизвестно.

Ключевые слова: кластерный анализ, метод наименьших квадратов, марковский момент.

Введение. Под кластерным анализом понимают алгоритмическую типологизацию элементов некоторого множества (выборочной совокупности) X по «мере» их сходства друг с другом. Произвольный алгоритм кластеризации является отображением

$$\mathcal{A}\colon \begin{cases} X \longrightarrow \mathbb{N}, \\ \overline{x}_i \longmapsto k, \end{cases}$$

ставящим в соответствие любому элементу $\overline{x}_i$ из выборки X единственное натуральное число k, являющееся номером кластера, которому принадлежит $\overline{x}_i$. Процесс кластеризации разбивает выборку X на попарно дизъюнктные подмножества X_h, называемые *кластерами*:

$$X = \bigcup_{h=1}^{m} X_h,$$

где для $\forall\ h, l \mid 1 \leqslant h, l \leqslant m\colon X_h \cap X_l = \varnothing$.

Следовательно, отображение $\mathcal{A}$ задает на X отношение эквивалентности; в качестве независимых представителей классов эквивалентности выбирают элементы, называемые *центроидами*. В n-мерном евклидовом пространстве $\mathbb{E}^n$ координаты центроидов равны среднему арифметическому соответствующих координат всех элементов (векторов), входящих в кластер (класс эквивалентности). Если отождествить каждый вектор из $\mathbb{E}^n$ с материальной точкой единичной массы, то центроиды можно рассматривать как центры масс.

Важной проблемой кластерного анализа является расчет предпочтительного числа классов эквивалентности. С решением этого вопроса связано нахождение момента завершения самого процесса. Данная связь предполагает, что правило определения числа кластеров и критерий завершения алгоритма кластеризации зависят друг от друга, а иногда и совпадают. Решение о количестве классов эквивалентности принимается или во время самого процесса, или еще до его начала (например, при использовании метода k-средних). В большинстве случаев определение числа кластеров во время выполнения процесса кластеризации основано на визуальном анализе дендрограмм, по которым можно сделать вывод об их предпочтительном количестве [1–3]. Но такой подход является эвристическим, а суть эвристических методов состоит в том, что они основываются на некоторых правдоподобных предположениях, а не на строгих выводах.

В настоящее время проблема истинного числа кластеров не решена. В книге, посвященной использованию статистических методов в археологических исследованиях, Бакстер (Baxter) утверждает, что для установления их предпочтительного количества наиболее распространенным подходом будет использование неформальных и субъективных критериев, основанных на экспертной оценке [4]. Согласен с ним и Эверитт (Everitt), который отмечает, что отсутствие единого мнения по данному вопросу делает комментарий Бакстера (Baxter) наиболее точным [1]. Тем не менее, особенно при обработке больших массивов эмпирических данных или данных большой размерности, кластерный анализ является одним из основных методов предварительной типологизации, а это обусловливает необходимость вывода формальных критериев завершения процесса и правил вычисления количества кластеров.

В подавляющем большинстве современных работ, в которых изучаются и решаются эти проблемы, авторы рассматривают не общий, а различные частные случаи кластеризации. Прежде всего следует выделить статью [5], в которой описан алгоритм, основанный на поиске и оценке скачков так называемых индексных функций. Главным недостатком этого метода является его большая вычислительная сложность. Развивая идеи, изложенные в [5], О. Н. Граничин с соавторами предложили применять для нахождения числа кластеров рандомизированные алгоритмы аппроксимации скачков индексных функций [6, 7].

Еще один способ решения этой задачи основан на оценке плотности распределения элементов выборочной совокупности (см., например, [8, 9]). В статье [9] значительное внимание уделяется не только проблеме определения предпочтительного числа кластеров, но и робастности самого процесса. Аналогичные вопросы изучаются в работах [10, 11].

Кроме проблемы количества классов эквивалентности, в кластерном анализе большое значение имеет оценка качества результатов типологизации и интеллектуального анализа данных (англ. *data mining*). Возможным подходом к изучению таких проблем может стать исследование робастности и устойчивости процесса кластеризации [9–12].

Методы «k-средних» и «одиночной связи». Сравним два алгоритма кластерного анализа данных, расположенных в n-мерном евклидовом пространстве $\mathbb{E}^n$. Наиболее популярный из современных методов кластеризации числовых данных — метод k-средних (англ. *k-means*), был изобретен в середине XX в. Штайнхаусом (Steinhaus) и Ллойдом (Lloyd) [13, 14]. Этот алгоритм стремится минимизировать суммарное квадратичное отклонение элементов классов эквивалентности от их центров масс. Действие алгоритма k-средних начинается с того, что выборка X разбивается на заранее заданное число кластеров со случайно выбранными центроидами. Основная идея такого метода заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге. Затем элементы разбиваются на новые классы эквивалентности в соответствии с тем, какой из новых центроидов оказался ближе. Алгоритм завершается тогда, когда на очередной итерации не происходит изменение суммарного квадратичного отклонения элементов от центра масс. Метод k-средних реализуется за конечное число итераций, так как количество возможных разбиений конечного множества (выборки) X конечно и на каждом шаге суммарное квадратичное отклонение уменьшается, поэтому алгоритм сходится [1, 14–16].

Метод k-средних имеет три существенных недостатка. Во-первых, он гарантирует достижение не глобального минимума суммарного квадратичного отклонения, а только одного из локальных минимумов. Во-вторых, результат кластеризации зависит от выбора исходных центроидов, а их оптимальный выбор неизвестен. В-третьих, число кластеров надо указать заранее, а это означает, что можно задать «обучающую выборку», и практически кластеризация превращается в классификацию.

В качестве альтернативы методу k-средних для действительно автоматической кластеризации в $\mathbb{E}^n$ можно предложить иерархический агломеративный алгоритм «*одиночной связи*» (англ. *single linkage*) [1, 16].

Представим этот метод формально. Пусть $X = \{\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_m\}$ — выборочная совокупность, в которой любой вектор $\overline{x}_i$ из X принадлежит евклидову пространству $\mathbb{E}^n$, т. е. для $\forall\, \overline{x}_i = (x_i^1, x_i^2, \ldots, x_i^n)$ и для $\forall\, i, j \mid 1 \leqslant i \leqslant m,\ 1 \leqslant j \leqslant n\colon x_i^j \in \mathbb{R}$.

В пространстве $\mathbb{E}^n$ задана стандартная метрика $\rho \mid \forall\, \overline{x}, \overline{y} \in \mathbb{E}^n$:

$$\rho(\overline{x}, \overline{y}) = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}.$$

Если выборочная совокупность X содержит m элементов (векторов), то полагают, что X разбита на m классов эквивалентности (кластеров), содержащих по одному элементу — $X_1 = \overline{x}_1$, $X_2 = \overline{x}_2$, $\ldots$, $X_m = \overline{x}_m$:

$$X = \bigcup_{h=1}^{m} X_h.$$

При этом понятно, что кластеры, состоящие из единственного элемента, и их центроиды совпадают: $X_h = \widehat{X}_h$ для $\forall\, h \mid 1 \leqslant h \leqslant m$.

Итерации алгоритма $\mathcal{A}$, реализующего метод «одиночной связи», можно описать следующим образом.

Первым шагом 1-й итерации $\mathcal{A}_1$ алгоритма $\mathcal{A}$ является построение диагональной матрицы расстояний между X_h:

$$\begin{pmatrix} 0 & \rho(X_1,X_2) & \rho(X_1,X_3) & \dots & \rho(X_1,X_m) \\ & 0 & \rho(X_2,X_3) & \dots & \rho(X_2,X_m) \\ & & \ddots & & \\ & & & 0 & \rho(X_{m-1},X_m) \\ & & & & 0 \end{pmatrix}.$$

Затем определяется ее минимальный элемент

$$F_1 = \min(\rho(X_h, X_l)),$$

где $1 \leqslant h, l \leqslant m$; F_1 — минимальное расстояние при $\mathcal{A}_1$.

После чего X_h и X_l, для которых ρ минимально, объединяются в один класс эквивалентности, который обозначим как X_1, а его центроид — как $\widehat{X}_1$. Кластеры X_h и X_l (при $\mathcal{A}_1$ элементы $\overline{x}_h$ и $\overline{x}_l$) заменяются на центроид $\widehat{X}_1$. Таким образом, после $\mathcal{A}_1$ выборочная совокупность X оказывается разбитой на $m-1$ элемент.

Не умаляя общности, будем считать, что в начале g-й итерации $\mathcal{A}_g$ агломеративного алгоритма кластеризации $\mathcal{A}$ выборочная совокупность X разбита на p кластеров. Первым шагом $\mathcal{A}_g$ является построение диагональной матрицы расстояний

$$\begin{pmatrix} 0 & \rho(X_1,X_2) & \rho(X_1,X_3) & \dots & \rho(X_1,X_p) \\ & 0 & \rho(X_2,X_3) & \dots & \rho(X_2,X_p) \\ & & \ddots & & \\ & & & 0 & \rho(X_{p-1},X_p) \\ & & & & 0 \end{pmatrix}.$$

Затем так же, как и при $\mathcal{A}_1$, находится минимальный элемент этой матрицы

$$F_g = \min(\rho(X_h, X_l)),$$

где $1 \leqslant h, l \leqslant p$; F_g — минимальное расстояние при $\mathcal{A}_g$.

Элементы X_h и X_l, для которых расстояние ρ является минимальным, объединяются в кластер, его обозначим как X_g. Его центроид $\widehat{X}_g$ имеет координаты, равные среднему арифметическому соответствующих координат всех векторов из X_h или X_l, объединенных в X_g. В конце итерации $\mathcal{A}_g$ элементы X_h и X_l заменяются на $\widehat{X}_g$. Таким образом, после завершения $\mathcal{A}_g$ выборочная совокупность X оказывается разбитой на $p-1$ элемент.

Главное преимущество метода «одиночной связи» заключается в его математических свойствах: результаты, полученные при его помощи, инвариантны монотонным преобразованиям матрицы сходства, его применению не мешает наличие совпадающих данных, по сравнению с другими методами кластеризации он обладает высокой устойчивостью и особенно эффективен в евклидовых пространствах [16].

Множество минимальных расстояний. Если нет правила завершения процесса кластеризации, то после $m-1$ итерации метода «одиночной связи» выборочная совокупность X будет объединена в один кластер, что является абсурдным результатом.

Для определения предпочтительного числа кластеров построим статистический критерий завершения агломеративного процесса кластеризации в $\mathbb{E}^n$.

Множество минимальных расстояний, полученное после $m-1$ итерации описанного алгоритма, имеет вид $\{F_1, F_2, \ldots, F_{m-1}\}$ и линейно упорядочено относительно числовых значений своих элементов: $0 \leqslant F_1 \leqslant F_2 \leqslant \ldots \leqslant F_{m-1}$. Используем это множество при выводе формального правила завершения агломеративного процесса кластеризации, реализующего метод «одиночной связи» в n-мерном евклидовом пространстве $\mathbb{E}^n$.

Сначала в качестве иллюстрирующего примера рассмотрим множество X, состоящее из 33 упорядоченных пар: $X = \{(0,0);\ (2,4);\ (3,3);\ (1,2);\ (3,0);\ (3,1);\ (1,1);\ (12,18);\ (13,17);\ (11,15);\ (13,14);\ (14,16);\ (11,16);\ (12,15);\ (13,18);\ (12,5);\ (13,2);\ (14,4);\ (12,3);\ (13,1);\ (14,2);\ (24,19);\ (22,22);\ (21,24);\ (23,21);\ (24,20);\ (22,39);\ (23,38);\ (24,39);\ (21,37);\ (2,26);\ (24,6);\ (10,36)\}$, которые можно отождествить с точками ограниченной области на плоскости (рис. 1). В этом простейшем случае количество кластеров и их расположение можно определить визуально: пять кластеров и три изолированные точки.

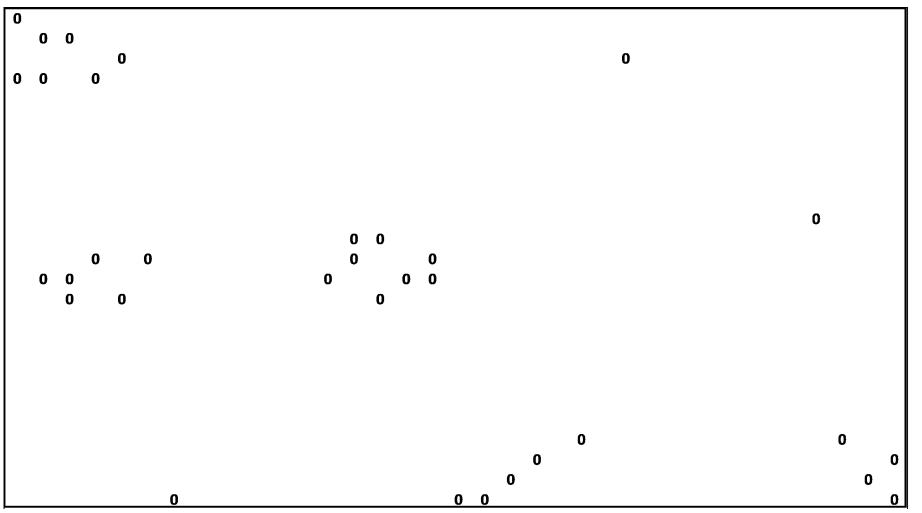

Рис. 1. Множество X (точка (0,0) находится в верхнем левом углу)

Элементы множества минимальных расстояний принимают следующие значения: $F_1 = 1.000$, $F_2 = 1.000$, $F_3 = 1.000$, $F_4 = 1.000$, $F_5 = 1.000$, $F_6 = 1.000$, $F_7 = 1.118$, $F_8 = 1.118$, $F_9 = 1.118$, $F_{10} = 1.414$, $F_{11} = 1.414$, $F_{12} = 1.414$, $F_{13} = 1.581$, $F_{14} = 1.803$, $F_{15} = 1.886$, $F_{16} = 2.134$, $F_{17} = 2.134$, $F_{18} = 2.236$, $F_{19} = 2.386$, $F_{20} = 2.500$, $F_{21} = 2.574$, $F_{22} = 2.603$, $F_{23} = 2.846$, $F_{24} = 2.864$, $F_{25} = 4.161$, $F_{26} = 11.214$, $F_{27} = 11.595$, $F_{28} = 12.701$, $F_{29} = 14.278$, $F_{30} = 17.322$, $F_{31} = 18.017$, $F_{32} = 28.475$.

При слиянии кластеров или присоединении к любому из них одной из изолированных точек должен произойти резкий скачок числового значения минимального расстояния (F_{25} на рис. 2), который, по здравому смыслу, совпадает с моментом завершения процесса кластеризации. На рис. 2 хорошо видно, что этот скачок лучше аппроксимировать не прямой, а параболой.

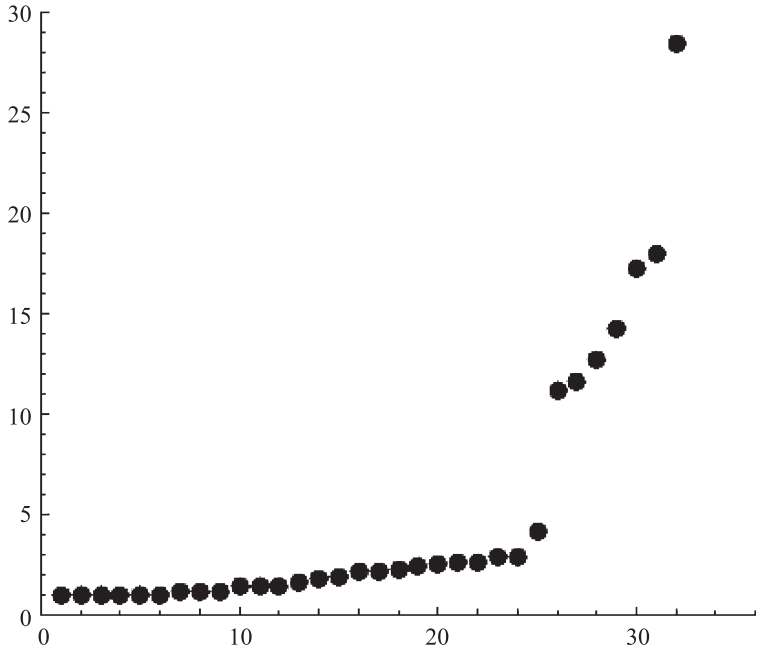

Рис. 2. График значений F_i (на оси абсцисс отложены номера итераций)

Кластерный анализ как случайный процесс. Пусть $T = \overline{1, m-1}$ — ограниченное подмножество натурального ряда, содержащее первое $m-1$ натуральное число. Тогда семейство $\xi = \{\xi_t, t \in T\}$ случайных величин $\xi_t = \xi_t(\omega)$, заданных для $\forall\ t \in T$ на одном и том же вероятностном пространстве $(\Omega, \mathcal{F}, \mathrm{P})$, называется *дискретным случайным процессом*.

Каждая случайная величина ξ_t порождает σ-алгебру, которую будем обозначать как $\mathcal{F}_{\xi_t}$. Тогда σ-алгеброй, порожденной случайным процессом $\xi = \{\xi_t, t \in T\}$, называется минимальная σ-алгебра, содержащая все $\mathcal{F}_{\xi_t}$, т. е.

$$\sigma(\xi) = \sigma\left(\bigcup_{t=1}^{m-1} \mathcal{F}_{\xi_t}\right).$$

Дискретный случайный процесс $\xi = \{\xi_t, t \in T\}$ можно представить как функцию двух переменных $\xi = \xi(t, \omega)$, где t — натуральный аргумент, ω — случайное событие. Если зафиксировать t, то, как указывалось выше, получим случайную величину ξ_t; если же зафиксировать случайное событие ω_0, то имеем функцию от натурального аргумента t, которая называется *траекторией* случайного процесса $\xi = \{\xi_t, t \in T\}$ и является случайной последовательностью $\xi_t(\omega_0)$.

Рассмотрим кластеризацию конечного множества X из евклидова пространства $\mathbb{E}^n$ как дискретный случайный процесс $\xi = \xi(t, \omega)$. Случайным событием $\omega \in \Omega$ будет извлечение выборки X из $\mathbb{E}^n$. Теоретически любая точка $\overline{x} \in \mathbb{E}^n$ может принадлежать выборочной совокупности X, поэтому σ-алгебра из вероятностного пространства $(\Omega, \mathcal{F}, \mathrm{P})$ содержит все $\mathbb{E}^n$, любое конечное множество X из пространства $\mathbb{E}^n$, все возможные счетные объединения таких множеств и дополнения к ним. Обозначим данную систему множеств как $\mathcal{S}(\mathbb{E}^n)$ и назовем *выборочной σ-алгеброй*, $\mathcal{F} = \mathcal{S}(\mathbb{E}^n)$. Те же рассуждения справедливы для любой σ-алгебры $\mathcal{F}_{\xi_t}$, потому

$$\sigma(\xi) = \mathcal{S}(\mathbb{E}^n).$$

Заметим, что эта σ-алгебра «беднее», чем борелевская $\mathcal{S}(\mathbb{E}^n) \subset \mathcal{B}(\mathbb{E}^n)$. Действительно, счетное объединение не более чем счетных множеств — счетно, поэтому $\mathcal{S}(\mathbb{E}^n)$ не содержит промежутков.

Рассмотрим бинарную задачу проверки статистических гипотез H_0 и H_1, где нулевая гипотеза H_0 — случайная последовательность $\xi_t(\omega_0)$ возрастает линейно, а альтернативная гипотеза H_1 — случайная последовательность $\xi_t(\omega_0)$ возрастает нелинейно (параболически). Для проверки статистической гипотезы необходимо построить критерий как строгое математическое правило, позволяющее ее принять или отвергнуть.

В евклидовом пространстве $\mathbb{E}^n$ при кластерном анализе выборочных данных методом «одиночной связи» одной из основных характеристик процесса будет множество минимальных расстояний. Естественно рассматривать его значение как случайную величину $\xi_t : \Omega \longrightarrow \mathbb{R}$, полагая, что t — номер итерации агломеративного алгоритма кластеризации $\mathcal{A}$. Для любого фиксированного случайного события $\omega_0 \in \Omega$ соответствующая траектория $\xi_t(\omega_0) = F_t$ — монотонно возрастающая случайная последовательность. Построим статистический критерий завершения процесса кластеризации как момент остановки τ [17].

На вероятностном пространстве $(\Omega, \mathcal{F}, \mathrm{P})$ семейство σ-алгебр $F = \{\mathcal{F}_t,\, t \in T\}$ называется *фильтрацией*, если для $\forall\ i, j \in T \,|\, i < j : \mathcal{F}_i \subset \mathcal{F}_j \subset \mathcal{F}$. При этом, если для $\forall\ t \in T : \mathcal{F}_t = \sigma(\xi_i,\, i < t)$, то фильтрация называется *естественной*.

Случайный процесс $\xi = \{\xi_t, t \in T\}$ называется *согласованным* с фильтрацией F, если для $\forall\ t \in T : \sigma(\xi_t) = \mathcal{F}_{\xi_t} \subset \mathcal{F}_t$. Очевидно, что любой случайный процесс согласован со своей естественной фильтрацией.

Отображение $\tau : \Omega \longrightarrow T$ называется *марковским моментом* относительно фильтрации F, если для $\forall\ t \in T$ прообраз множества $\{\tau \leqslant t\} \in \mathcal{F}_t$. Если к тому же вероятность $P(\tau < +\infty) = 1$, то τ называется *марковским моментом остановки* [18, 19].

Иначе говоря, пусть τ — момент наступления некоторого события в случайном процессе $\xi = \{\xi_t, t \in T\}$. Если для $\forall\ t_0 \in T$ можно однозначно сказать, наступило событие τ или нет, при условии, что известны значения ξ_t только в прошлом (слева от t_0), то тогда τ — марковский момент относительно естественной фильтрации F случайного процесса $\xi = \{\xi_t, t \in T\}$. А если наступление τ в конечный момент времени является достоверным событием, то τ — марковский момент остановки.

Аппроксимационно-оценочный критерий. Для определения момента, когда характер монотонного возрастания числовой последовательности изменяется с линейного на параболический, используем ранее построенный аппроксимационно-оценочный критерий [20, 21].

Сначала формально определим термины «линейное возрастание» и «параболическое возрастание» числовой последовательности. Узлами аппроксимации для числовой последовательности y_n являются упорядоченные пары (i, y_i), где i — натуральный аргумент, y_i — соответствующее значение последовательности y_n. Так как подстрочный индекс однозначно определяет натуральный аргумент, узел аппроксимации (i, y_i) будем отождествлять с элементом y_i.

Под квадратичной погрешностью аппроксимации для функции $f(x)$ будем понимать сумму квадратов разностей значений числовой последовательности в узлах аппроксимации и аппроксимирующей функции при соответствующем аргументе:

$$\delta_f^2 = \sum_{i=0}^{k-1} (f(i) - y_i)^2.$$

Функция $f(x)$ из класса X является аппроксимирующей для узлов $y_0, y_1, \ldots, y_{k-1}$ в смысле квадратичного приближения, если для $f(x)$ справедливо

$$\delta_f^2 = \min_{f \in X} \sum_{i=0}^{k-1} (f(i) - y_i)^2,$$

такой минимум всегда найдется, так как δ_f^2 — положительно определенная квадратичная форма.

Будем различать линейную аппроксимацию в классе функций вида $l(x) = ax + b$ и неполную параболическую аппроксимацию (без линейного члена) в классе функций $q(x) = cx^2 + d$. Квадратичные погрешности по k узлам для линейной и неполной параболической аппроксимаций будут соответственно равны

$$\delta_l^2(k) = \sum_{i=0}^{k-1} (a \cdot i + b - y_i)^2, \tag{1}$$

$$\delta_q^2(k) = \sum_{i=0}^{k-1} (c \cdot i^2 + d - y_i)^2. \tag{2}$$

Если в наших рассуждениях количество узлов аппроксимации несущественно или очевидно из контекста, то соответствующие квадратичные погрешности будем просто обозначать δ_l^2 и δ_q^2.

При сравнении δ_l^2 и δ_q^2 возможны три случая: $\delta_q^2 < \delta_l^2$, $\delta_q^2 > \delta_l^2$, $\delta_q^2 = \delta_l^2$.

Будем говорить, что последовательность y_n имеет *линейное возрастание* в узлах (точках) $y_0, y_1, \ldots, y_{k-1}$, если в этих значениях y_n монотонна и квадратичные погрешности линейной и неполной параболической аппроксимаций по этим узлам связаны неравенством $\delta_q^2 > \delta_l^2$. Если при тех же условиях справедливо неравенство $\delta_q^2 < \delta_l^2$, то последовательность y_n имеет *параболическое возрастание* в точках $y_0, y_1, \ldots, y_{k-1}$. Если же для узлов аппроксимации $y_0, y_1, \ldots, y_{k-1}$ выполняется равенство $\delta_q^2 = \delta_l^2$, то тогда точка y_{k-1} называется *критической*.

Вычислим по методу наименьших квадратов коэффициенты a, b линейной функции $ax + b$ и коэффициенты c, d для неполной квадратичной функции $cx^2 + d$, аппроксимирующих узлы $y_0, y_1, \ldots, y_{k-1}$ [20, 21]:

$$a = \frac{6}{k(k^2-1)} \sum_{i=0}^{k-1} (2i + 1 - k) y_i, \qquad b = \frac{2}{k(k+1)} \sum_{i=0}^{k-1} (2k - 1 - 3i) y_i, \tag{3}$$

$$c = \frac{30}{k(k-1)(2k-1)(8k^2-3k-11)} \sum_{i=0}^{k-1} (6i^2 - (k-1)(2k-1)) y_i, \tag{4}$$

$$d = \frac{6}{k(8k^2-3k-11)} \sum_{i=0}^{k-1} (3k(k-1) - 1 - 5i^2) y_i. \tag{5}$$

Чтобы определить момент, когда характер возрастания монотонной последовательности y_n изменяется с линейного на параболический, построим аппроксимационно-оценочный критерий δ^2.

Будем считать, по определению, что для узлов аппроксимации $y_0, y_1, \ldots, y_{k-1}$ критерий $\delta^2 = \delta^2(k_0) = \delta_l^2(k_0) - \delta_q^2(k_0)$. При этом положим, что всегда $y_0 = 0$. Выполнения этого условия легко добиться на любом шаге аппроксимации при помощи преобразования:

$$y_0 = y_j - y_j,\ y_1 = y_{j+1} - y_j,\ \ldots,\ y_{k-1} = y_{j+k-1} - y_j. \tag{6}$$

Вычислим, используя формулы (1)–(5), квадратичные погрешности линейной и неполной параболической аппроксимаций по четырем точкам y_0, y_1, y_2, y_3, а затем сравним их [20, 21]:

$$ax + b = \frac{1}{10}(-y_1 + y_2 + 3y_3)x + \frac{1}{10}(4y_1 + y_2 - 2y_3),$$

$$cx^2 + d = \frac{1}{98}(-5y_1 + y_2 + 11y_3)x^2 + \frac{1}{98}(42y_1 + 21y_2 - 14y_3),$$

$$\delta_l^2(4_0) = \sum_{k=0}^{3}\left[\frac{1}{10}(k(-y_1 + y_2 + 3y_3) + (4y_1 + y_2 - 2y_3)) - y_k\right]^2 =$$

$$= \frac{1}{10}(7y_1^2 + 7y_2^2 + 3y_3^2 - 4y_1y_2 - 2y_1y_3 - 8y_2y_3), \tag{7}$$

$$\delta_q^2(4_0) = \sum_{k=0}^{3}\left[\frac{1}{98}(k^2(-5y_1 + y_2 + 11y_3) + (42y_1 + 21y_2 - 14y_3)) - y_k\right]^2 =$$

$$= \frac{1}{98}(61y_1^2 + 73y_2^2 + 13y_3^2 - 44y_1y_2 + 6y_1y_3 - 60y_2y_3),$$

$$\delta^2(4_0) = \delta_l^2(4_0) - \delta_q^2(4_0) = \frac{1}{245}(19y_1^2 - 11y_2^2 + 41y_3^2 + 12y_1y_2 - 64y_1y_3 - 46y_2y_3).$$

Можно сказать, что вблизи элемента y_k характер возрастания числовой последовательности y_n изменился с линейного на параболический, если для узлов $y_0, y_1, \ldots, y_{k-1}$ линейная аппроксимация не хуже неполной параболической, т. е. справедливо неравенство $\delta^2 = \delta_l^2 - \delta_q^2 \leqslant 0$, а для набора точек $y_1, y_2, \ldots, y_k$, сдвинутых на один шаг дискретности, неполная параболическая аппроксимация стала точнее линейной, т. е. выполнилось неравенство $\delta^2 = \delta_l^2 - \delta_q^2 > 0$.

Для случайной последовательности минимальных расстояний $\xi_t(\omega_0) = F_t(X)$ при кластеризации выборочной совокупности $X \subset \mathbb{E}^n$ методом «одиночной связи» естественной фильтрацией, согласованной с процессом, будет выборочная σ-алгебра $\mathcal{S}(\mathbb{E}^n)$. Тогда, по определению, марковским моментом остановки агломеративного процесса кластеризации будет статистика

$$\tau = \min\{t \in T \mid \delta^2 > 0\}.$$

То есть марковским моментом остановки агломеративного процесса кластеризации является минимальное значение τ, при котором отвергается нулевая гипотеза H_0 (последовательность минимальных расстояний возрастает линейно) и принимается альтернативная гипотеза H_1 (последовательность минимальных расстояний возрастает параболически).

Чувствительность аппроксимационно-оценочного критерия. Для того чтобы окончательно сформулировать условие завершения описанного выше агломеративного процесса кластеризации, осталось рассмотреть «проблему чувствительности» аппроксимационно-оценочного критерия δ^2, которую можно связать с понятием «устойчивой кластеризации».

Предварительно решим «обратную задачу». А именно, пусть известны значения последовательности y_n в узлах y_0, y_1, y_2, и требуется определить, при каком значении в узле $(3, y_3)$ характер возрастания последовательности y_n изменился с линейного на параболический. Иными словами, надо определить, при каком числовом значении y_3 эта точка станет критической. Приравняем к нулю квадратичную форму (7) и, заменив y_3 на x, решим квадратное уравнение

$$41x^2 - (64y_1 + 46y_2)\,x + \left(19y_1^2 + 12y_2y_1 - 11y_2^2\right) = 0,$$

для которого

$$x_{1,2} = \frac{32y_1 + 23y_2 \pm 7\sqrt{5}\,(y_1 + 2y_2)}{41}.$$

Учитывая, что $0 \leqslant y_1 \leqslant y_2 \leqslant y_3$, окончательно получим

$$y_3 = \frac{32y_1 + 23y_2 + 7\sqrt{5}\,(y_1 + 2y_2)}{41}. \tag{8}$$

Вспомним введенное преобразование (6) и заметим, что если $y_j = y_{j+1} = y_{j+2}$, то тогда не только $y_0 = 0$, но и $y_1 = y_2 = 0$. Согласно (7), для любого $y_{j+3} > y_{j+2}$, даже если $y_3 = y_{j+3} - y_j > 0$ сколь угодно мало, квадратичная форма $\delta^2 > 0$.

Например, для рассмотренного выше множества минимальных расстояний $\{F_1, F_2, \ldots, F_{32}\}$ критерий $\delta^2(4_0)$ примет следующие значения:

$$\delta_4^2 = 0, \quad \delta_5^2 = 0, \quad \delta_6^2 = 0, \quad \delta_7^2 = 0.002,$$

символ δ_4^2 обозначает величину критерия по узлам F_1, F_2, F_3, F_4, символ δ_5^2 — по узлам F_2, F_3, F_4, F_5 и т. д.

Согласно принятым выше соглашениям, агломеративный алгоритм кластеризации множества X должен завершиться после итерации $\mathcal{A}_7$. Но в этом случае множество X будет разделено на 6 кластеров и 20 изолированных точек (рис. 3), что вряд ли можно считать удовлетворительным результатом.

Если ввести преобразование $y_i = F_i + q \cdot i$, то получим множество $\{y_1, y_2, \ldots, y_k\}$, которое назовем «*множеством тренда*», а q — «*коэффициентом тренда*». При применении критерия δ^2 не к набору $\{F_1, F_2, \ldots, F_{29}\}$, а к множеству $\{y_1, y_2, \ldots, y_{29}\}$ результат кластеризации качественно меняется.

Например, при $q = 0.2$ множество тренда и аппроксимационно-оценочный критерий принимают следующие значения: $y_1 = 1.0$, $y_2 = 1.2$, $y_3 = 1.4$, $y_4 = 1.6$, $y_5 = 1.8$, $y_6 = 2.0$, $y_7 = 2.318, y_8 = 2.518$, $y_9 = 2.718$, $y_{10} = 3.214$ и $\delta_4^2 = -0.016, \delta_5^2 = -0.016$, $\delta_6^2 = -0.016$, $\delta_7^2 = -0.005$, $\delta_8^2 = -0.025$, $\delta_9^2 = -0.039$, $\delta_{10}^2 = 0.020$ соответственно. При этом множество X разбивается на 7 кластеров и 16 изолированных точек (рис. 4).

Такой же результат кластеризации, но при других значениях $\{y_1, y_2, \ldots, y_{32}\}$ и $\{\delta_4^2, \delta_5^2, \ldots\}$ получается, когда $q = 0.3$. Если q изменяется в пределах от 0.4 до 1.1,

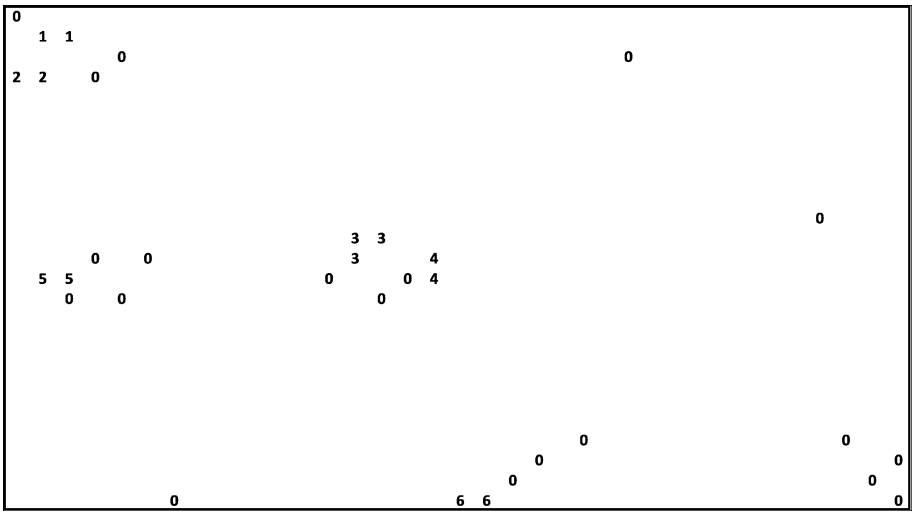

Рис. 3. Кластеризация множества X по узлам $\{F_1, F_2, \ldots, F_{32}\}$

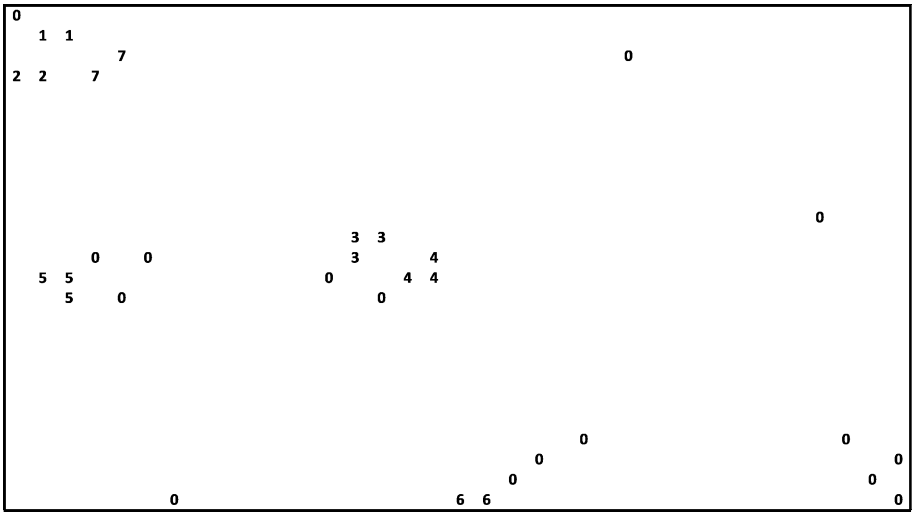

Рис. 4. Результаты кластеризации множества X при $q \in [0.2,\ 0.3]$

то множество X разбивается на 5 кластеров и 3 изолированные точки (рис. 5). При q в пределах от 1.2 до 8.1 множество X разделяется на 4 кластера и 3 изолированные точки (рис. 6), а при $q \geqslant 8.2$ множество X представляется как один кластер, состоящий из 33 точек.

Выполнение процесса кластеризации завершается при помощи аппроксимационно-оценочного критерия, который оценивает скачки монотонно возрастающей последовательности минимальных расстояний. Величина значимого скачка, достаточного для остановки процесса кластеризации, зависит от чувствительности критерия остановки, которая задается при помощи неотрицательного коэффициента q. Чем больше значение q, тем меньше чувствительность критерия остановки процесса кластеризации. Максимальной чувствительностью критерий остановки обладает при $q = 0$,

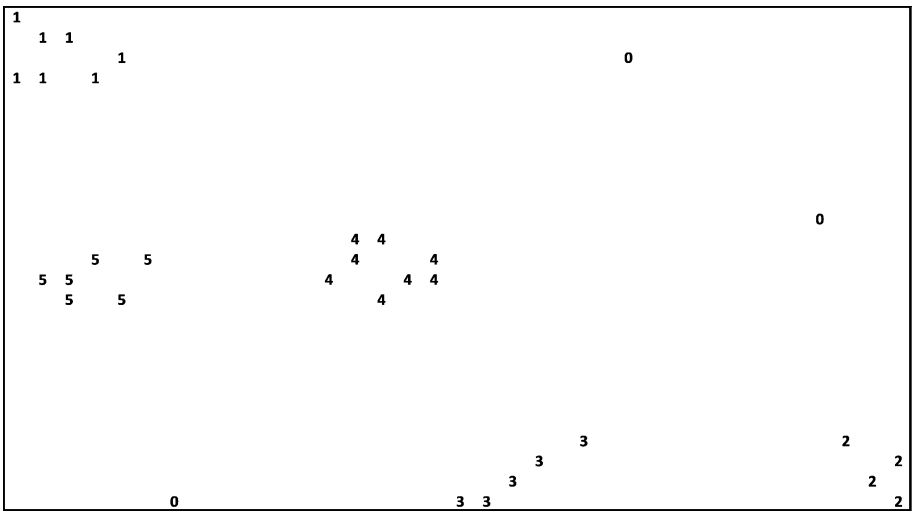

Рис. 5. Предпочтительное число кластеров при $q \in [0.4,\ 1.1]$

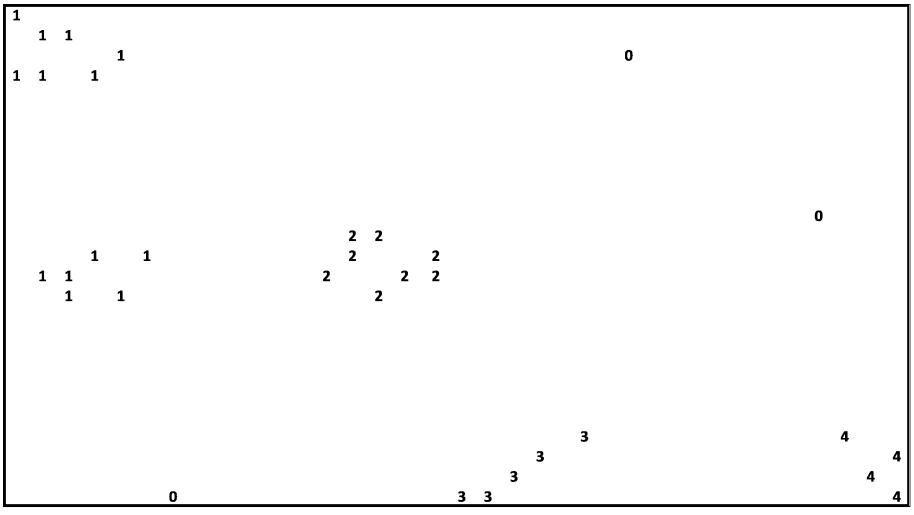

Рис. 6. Образование большого продолговатого кластера при $q \in [1.2,\ 8.1]$

в этом случае при кластеризации получится наибольшее число кластеров. Увеличивая q, можно уменьшить чувствительность критерия остановки так, что процесс будет продолжаться до тех пор, пока все m векторов не объединятся в один кластер. Действительно, если узлы аппроксимации y_0, y_1, y_2 изменяются как арифметическая прогрессия с разностью q, то формулу (8) можно записать в виде

$$y_3 = \frac{1}{41}\left(78 + 35\sqrt{5}\right) q \simeq 3.811q$$

и узлы аппроксимации в этом случае принимают значения 0, q, $2q$, $3.811q$. А это означает, что при увеличении коэффициента тренда q чувствительность критерия уменьшается и для достижения критического значения необходима бóльшая величина скачка изменения числового значения минимального расстояния.

Устойчивость кластеризации. Кластерный анализ обладает большой степенью субъективности, поэтому интерпретация его результатов во многом зависит от самого исследователя. Кроме нахождения приемлемого числа кластеров, важное значение имеет «устойчивость кластеризации». В работах [6, 7, 12] вместо строгого определения этого понятия вводится его интуитивное описание, например: «Устойчивость кластеризации показывает, насколько различными получаются результирующие разбиения на группы после многократного применения алгоритмов кластеризации для одних и тех же данных. Небольшое расхождение результатов интерпретируется как высокая устойчивость» [6, с. 87].

При использовании метода «одиночной связи» и аппроксимационно-оценочного критерия завершения процесса кластеризации в качестве количественной меры устойчивости можно рассматривать величину промежутка $Q_i = [\alpha_i, \beta_i]$ изменения коэффициента $q \in [\alpha_i, \beta_i]$, при котором для выборочной совокупности X получается один и тот же результат.

В этой связи необходимо вспомнить широко известную работу по кластерному анализу Олдендерфера (Aldenderfer) и Блэшфилда (Blashfield) [16], в которой они утверждают, что основной недостаток метода «одиночной связи» заключается в высокой вероятности возникновения «цепного эффекта» и образования больших продолговатых (вытянутых по одному или нескольким измерениям) кластеров. По мере приближения к окончанию процесса кластеризации образуется один большой кластер, к которому присоединяются ранее сформировавшиеся кластеры и изолированные точки. В качестве подтверждения этой мысли приводится соответствующая дендрограмма.

На рис. 3–6 можно наблюдать иллюстрацию этого процесса в численном эксперименте при кластеризации 33 точек из ограниченной плоской области (численный эксперимент проводился при помощи программы, написанной на языке Visual Basic в интегрированной среде разработки Visual Studio Community 2017). Сначала образуются подкластеры (как собственные подмножества) при значениях коэффициента q из промежутков $Q_1 = [0,\, 0.1]$ и $Q_2 = [0.2,\, 0.3]$, затем получается разбиение на приемлемое количество кластеров (в смысле визуальной оценки) при $Q_3 = [0.4,\, 1.1]$, потом происходит объединение двух из пяти кластеров в один «большой продолговатый» кластер (его элементы на рис. 6 обозначены цифрой 1) при $Q_4 = [1.2,\, 8.1]$ и, наконец, все точки собираются в один кластер из 33 элементов при $Q_5 = [8.2,\, \infty)$. В общем случае последовательность промежутков устойчивой кластеризации для различных значений параметра q обозначим как $Q_1, Q_2, \ldots, Q_{e-2}, Q_{e-1}, Q_e$, где Q_e — множество значений коэффициента q, при которых все m точек объединяются в один кластер.

Журнал корпорации Microsoft в 2015 г. опубликовал статью, посвященную программной реализации одной из модификаций метода k-средних [22]. В этой работе как пример производится кластеризация точек на евклидовой плоскости, при этом a priori задается разбиение на три кластера. Для тех же самых данных методом «одиночной связи» и при помощи аппроксимационно-оценочного критерия аналогичный результат, без априорного предположения о количества кластеров, был получен при $q \in Q_{e-2} = [0.3,\, 0.9]$, при $q \in Q_{e-1} = [1,\, 2.7]$ данные были разделены на два кластера и при $q \geqslant 2.8$ все точки объединились в один кластер.

Заключение. Статистический критерий завершения агломеративного процесса кластеризации, основанного на методе «одиночной связи» в евклидовом пространстве $\mathbb{E}^n$, можно сформулировать следующим образом.

Пусть $\{F_1, F_2, \ldots, F_k\}$ — линейно упорядоченное множество минимальных расстояний, а набор $\{y_1, y_2, \ldots, y_k\}$ — «множество тренда», полученное при помощи преобразования $y_i = F_i + q \cdot i$, где q — «коэффициент тренда», i — номер итерации агломеративного алгоритма кластеризации $\mathcal{A}$. Процесс кластеризации считается завершенным при k-й итерации, если для узлов $y_{k-4}, y_{k-3}, y_{k-2}, y_{k-1}$ справедливо неравенство $\delta^2 \leqslant 0$, а для набора точек $y_{k-3}, y_{k-2}, y_{k-1}, y_k$ выполнилось неравенство $\delta^2 > 0$, где

$$\delta^2 = \frac{1}{245}(19y_1^2 - 11y_2^2 + 41y_3^2 + 12y_1y_2 - 64y_1y_3 - 46y_2y_3).$$

Иначе говоря, марковский момент остановки алгоритма кластеризации $\mathcal{A}$ равен статистике

$$\tau(F_1, F_2, \ldots, F_k) = \min\{k \mid \delta^2 > 0\},$$

при этом отвергается нулевая гипотеза H_0 — значения элементов линейно упорядоченного множества тренда возрастают линейно и принимается альтернативная гипотеза H_1 — значения элементов линейно упорядоченного множества тренда возрастают параболически.

Автоматическое определение числа кластеров является актуальной проблемой во многих случаях предварительной типологизации эмпирических данных, например при цитометрическом исследовании крови [23], при автоматическом анализе текстов [24] и в других случаях, когда количество кластеров a priori неизвестно. Для решения этой задачи можно использовать алгоритм кластеризации, основанный на методе «одиночной связи», и аппроксимационно-оценочный критерий для завершения процесса. Кластеризация выборки X из n-мерного евклидова пространства $\mathbb{E}^n$ производится при различных величинах параметра q, который увеличивается от нуля до значения, при котором все точки X соберутся в один кластер. Окончательное решение о предпочтительном числе кластеров носит субъективный характер, но, на наш взгляд, наибольший интерес представляет разбиение при $q \in Q_{e-2}$.

Литература

1. *Everitt B. S.* Cluster analysis. Chichester, West Sussex, UK: John Wiley & Sons Ltd, 2011. 330 p.

2. *Duda R. O., Hart P. E., Stork D. G.* Pattern classification. 2nd ed. New York; Chichester: Wiley, 2001. 654 p.

3. *Calirnski T., Harabasz J.* A dendrite method for cluster analysis // Communications in Statistics. 1974. N 3. P. 1–27.

4. *Baxter M. J.* Exploratory multivariate analysis in archaeology. Edinburgh: Edinburgh University Press, 1994. 307 p.

5. *Sugar C. A., James G. M.* Finding the number of clusters in a dataset // Journal of the American Statistical Association. 2003. Vol. 98, N 463. P. 750–763.

6. *Граничин О. Н., Шалымов Д. С., Аврос Р., Волкович З.* Рандомизированный алгоритм нахождения количества кластеров // Автоматика и телемеханика. 2011. № 4. С. 86–98.

7. *Шалымов Д. С.* Рандомизированный метод определения количества кластеров на множестве данных // Науч.-технич. вестн. С.-Петерб. гос. ун-та информ. технологий, механики и оптики. 2009. № 5 (63). С. 111–116.

8. *Zhang G., Zhang C., Zhang H.* Improved K-means algorithm based on density Canopy // Knowledge-Based Systems. 2018. Vol. 145. P. 1–14.

9. *Jiali W., Yue Z., Xv L.* Automatic cluster number selection by finding density peaks // 2016 2nd IEEE Intern. Conference on Computer and Communications (ICCC). IEEE Proceedings. Chengdu, China, 2016. P. 13–18. doi: 10.1109 / CompComm.2016.7924655

10. *Cordeiro de Amorim R., Hennig C.* Recovering the number of clusters in data sets with noise features using feature rescaling factors // Information Sciences. 2015. Vol. 324. P. 126–145.

11. *Ложкинс А., Буре В. М.* Вероятностный подход к определению локально-оптимального числа кластеров // Вестн. С.-Петерб. ун-та. Прикладная математика. Информатика. Процессы управления. 2016. Т. 13. Вып. 1. С. 28–37.

12. *Шалымов Д. С.* Алгоритмы устойчивой кластеризации на основе индексных функций и функций устойчивости // Стохастическая оптимизация в информатике. 2008. Т. 4. № 1–1. С. 236–248.

13. *Steinhaus H.* Sur la division des corps materiels en parties // Bull. Acad. Polon. Sci. C1. III. 1956. Vol. IV. P. 801–804.

14. *Lloyd S.* Least squares quantization in PCM // IEEE Transactions on Information Theory. 1982. Vol. 28. Iss. 2. P. 129–137. doi: 10.1109/TIT.1982.1056489

15. *Hartigan J. A.* Clustering algorithms. New York; London; Sydney; Toronto: John Wiley & Sons Inc., 1975. 351 p.

16. *Aldenderfer M. S., Blashfield R. K.* Cluster analysis. Newburg Park: Sage Publications Inc., 1984. 88 p.

17. *Wald A.* Sequential analysis. New York: John Wiley & Sons Inc., 1947. 212 p.

18. *Sirjaev A. N.* Statistical sequential analysis: Optimal stopping rules. New York: American Mathematical Society, 1973. 174 p.

19. *Shiryaev A. N.* Optimal stopping rules. Berlin; Heidelberg: Springer, 2009. 220 p.

20. *Orekhov A. V.* Criterion for estimation of stress-deformed state of SD-materials // AIP Conference Proceedings. 2018. Vol. 1959. P. 070028. doi: 10.1063/1.5034703

21. *Орехов А. В.* Аппроксимационно-оценочные критерии напряженно-деформируемого состояния твердого тела // Вестн. С.-Петерб. ун-та. Прикладная математика. Информатика. Процессы управления. 2018. Т. 14. Вып. 3. С. 230–242. doi.org/10.21638/11702/spbu10.2018.304

22. *McCaffrey J.* Test run — k-means++ data clustering // MSDN Magazine. 2015. Vol. 30, N 8. P. 62–68.

23. *Зурочка А. В., Хайдуков С. В., Кудрявцев И. В., Черешнев В. А.* Проточная цитометрия в медицине и биологии. 2-е изд. Екатеринбург: Урал. отд. РАН, 2014. 574 с.

24. *Lappin S., Fox C.* The handbook of contemporary semantic theory. 2nd ed. Wiley-Blackwell: Wiley, 2015. 776 p.

Статья поступила в редакцию 28 февраля 2018 г.
Статья принята к печати 18 декабря 2018 г.

Контактная информация:

Орехов Андрей Владимирович — ст. преподаватель; A_V_Orehov@mail.ru

Markov moment for the agglomerative method of clustering in Euclidean space

A. V. Orekhov

St. Petersburg State University, 7–9, Universitetskaya nab., St. Petersburg, 199034, Russian Federation

For citation: Orekhov A. V. Markov moment for the agglomerative method of clustering in Euclidean space. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2019, vol. 15, iss. 1, pp. 76–92. https://doi.org/10.21638/11702/spbu10.2019.106 (In Russian)

When processing large arrays of empirical data or large-scale data, cluster analysis remains one of the primary methods of preliminary typology, which makes it necessary to obtain formal rules for calculating the number of clusters. The most common method for determining the preferred number of clusters is the visual analysis of dendrograms, but this approach is purely heuristic. The number of clusters and the end moment of the clustering algorithm depend on each other. Cluster analysis of data from n-dimensional Euclidean space using the "single linkage" method can consider as a discrete random process. Sequences of "minimum distances" define the trajectories of this process. The "approximation-estimating test" allows

us to establish the Markov moment when the growth rate of such a sequence changes from linear to parabolic, which, in turn, may be a sign of the completion of the agglomerative clustering process. The calculation of the number of clusters is the critical problem in many cases of the automatic typology of empirical data. For example, in medicine with cytometric analysis of blood, automated analysis of texts and in other instances when the number of clusters not known in advance.

Keywords: cluster analysis, least squares method, Markov moment.

References

1. Everitt B. S. *Cluster analysis.* Chichester, West Sussex, UK, John Wiley & Sons Ltd. Press, 2011, 330 p.

2. Duda R. O., Hart P. E., Stork D. G. *Pattern classification.* 2nd ed. New York, Chichester, Wiley Press, 2001. 654 p.

3. Calirnski T., Harabasz J. A dendrite method for cluster analysis. *Communications in Statistics*, 1974, no. 3, pp. 1–27.

4. Baxter M. J. *Exploratory multivariate analysis in archaeology.* Edinburgh, Edinburgh University Press, 1994, 307 p.

5. Sugar C. A., James G. M. Finding the number of clusters in a dataset. *Journal of the American Statistical Association*, 2003, vol. 98, no. 463, pp. 750–763.

6. Granichin O. N., Shalymov D. S., Avros R., Volkovich Z. Randomizirovannyy algoritm nakhozhdeniya kolichestva klasterov [A randomized algorithm for estimating the number of clusters]. *Avtomatika i telemekhanika* [*Automation and Remote Control*], 2011, no. 4, pp. 86–98. (In Russian)

7. Shalymov D. S. Randomizirovannyy metod opredeleniya kolichestva klasterov na mnozhestve dannykh. [Randomized method for determining the number of clusters on a data set]. *Nauchno-tekhnicheskiy vestnik Sankt-Peterburgskogo gosudarstvennogo universiteta informatsionnykh tekhnologiy, mekhaniki i optiki* [*Scientific and Technical Gazette of Saint Petersburg State University of Information Technologies, Mechanics and Optics*], 2009, no. 5 (63), pp. 111–116. (In Russian)

8. Zhang G., Zhang C., Zhang H. Improved k-means algorithm based on density Canopy. *Knowledge-Based Systems*, 2018, vol. 145, pp. 1–14.

9. Jiali W., Yue Z., Xv L. Automatic cluster number selection by finding density peaks. *2016 2nd IEEE International Conference on Computer and Communications (ICCC). IEEE Proceedings.* Chengdu, China, 2016, no. 7924655, pp. 13–18. doi: 10.1109 / CompComm.2016.7924655

10. Cordeiro de Amorim R., Hennig C. Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences*, 2015, vol. 324, pp. 126–145.

11. Lozkins A., Bure V. M. Veroyatnostnyy podkhod k opredeleniyu lokal'no-optimal'nogo chisla klasterov [A probabilistic approach to determining the locally optimal number of clusters]. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2016, vol. 13, iss. 1, pp. 28–37. (In Russian)

12. Shalymov D. S. Algoritmy ustoychivoy klasterizatsii na osnove indeksnykh funktsiy i funktsiy ustoychivosti [Algorithms for stable clustering based on index functions and stability functions]. *Stokhasticheskaya optimizatsiya v informatike* [*Stochastic optimization in computer science*], 2008, vol. 4, no. 1-1, pp. 236–248. (In Russian)

13. Steinhaus H. Sur la division des corps materiels en parties. *Bull. Acad. Polon. Sci. C1. III*, 1956, vol. IV, pp. 801–804.

14. Lloyd S. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 1982, vol. 28, iss. 2, pp. 129–137. doi: 10.1109/TIT.1982.1056489

15. Hartigan J. A. *Clustering algorithms.* New York, London, Sydney, Toronto, John Wiley & Sons Inc. Press, 1975, 351 p.

16. Aldenderfer M. S., Blashfield R. K. *Cluster analysis.* Newburg Park, Sage Publications Inc. Press, 1984, 88 p.

17. Wald A. *Sequential analysis.* New York, John Wiley & Sons Inc. Press, 1947, 212 p.

18. Sirjaev A. N. *Statistical sequential analysis: Optimal stopping rules.* New York, American Mathematical Society Publ., 1973, 174 p.

19. Shiryaev A. N. *Optimal stopping rules.* Berlin, Heidelberg, Springer Press, 2009, 220 p.

20. Orekhov A. V. Criterion for estimation of stress-deformed state of SD-materials. *AIP Conference Proceedings*, 2018, vol. 1959, pp. 070028. doi: 10.1063/1.5034703

21. Orekhov A. V. Approksimatsionno-otsenochnyye kriterii napryazhenno-deformiruyemogo sostoyaniya tverdogo tela [Approximation-evaluation tests for a stress-strain state of deformable solids].

Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes, 2018, vol. 14, iss. 3, pp. 230–242. doi.org/10.21638/11702/spbu10.2018.304 (In Russian)

22. McCaffrey J. Test run — k-means++ data clustering. *MSDN Magazine*, 2015, vol. 30, no. 8, pp. 62–68.

23. Zurochka A. V., Khaydukov S. V., Kudryavtsev I. V., Chereshnev V. A. *Protochnaya tsitometriya v meditsine i biologii*. 2-e izd. [*Flow cytometry in medicine and biology*. 2nd ed.]. Yekaterinburg, Ural Branch of the Russian Academy of Sciences Publ., 2014, 574 p. (In Russian)

24. Lappin S., Fox C. *The handbook of contemporary semantic theory*. 2nd ed. Wiley-Blackwell, Wiley Press, 2015, 776 p.

Received: February 28, 2018.
Accepted: December 18, 2018.

Author's information:

Andrey V. Orekhov — Senior Lecturer; A_V_Orehov@mail.ru