УДК 519.172.1+519.237.8+519.216.5+004.85
MSC 05C05+62H30

Бинарные метрические деревья и иерархия вложенных кластеров

А. В. Орехов, Е. В. Васильев

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

Для цитирования: *Орехов А. В., Васильев Е. В.* Бинарные метрические деревья и иерархия вложенных кластеров // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2024. Т. 20. Вып. 4. С. 487–499.
https://doi.org/10.21638/spbu10.2024.405

Методы машинного обучения используют деревья данных для организации и хранения информации. Каждая из таких структур обладает определенными преимуществами и позволяет улучшить качество конкретного алгоритма. Если у всех узлов дерева не более двух потомков, то оно называется бинарным; главное его преимущество — высокая эффективность реализации алгоритмов поиска и сортировки. В связи с этим важно отметить, что дендрограммы иерархических агломеративных методов кластеризации также относятся к бинарным деревьям и отражают таксономию элементов множества данных. Любой кластер, не являющийся синглетоном, можно разделить на подкластеры, что позволяет сформировать иерархическую структуру в метрическом пространстве (метрическое дерево) с дополнительными свойствами, например, автоматически задать высоту дерева, считая, по определению, что число уровней, на которых располагаются его узлы, совпадает с количеством вариантов разбиения выборочного множества на кластеры, подкластеры, подкластеры подкластеров и т. д. Такую задачу можно решить, используя аппроксимационно-оценочные критерии, изменение чувствительности которых при помощи коэффициента тренда дает возможность получить различные варианты кластеризации. При проведении вычислительных экспериментов использовалось синтетическое множество точек на евклидовой плоскости и изучались результаты его разбиения на кластеры центроидным методом. Марковские моменты остановки процесса кластеризации определялись посредством параболического аппроксимационно-оценочного критерия, построенного по четырем точкам. Верификация результатов, полученных при численном моделировании, производилась за счет изменения величины шага коэффициента тренда.

Ключевые слова: метрическое дерево, агломеративная кластеризация, марковский момент, метод наименьших квадратов.

1. Введение. Эффективные структуры данных используются искусственным интеллектом для анализа и обработки больших объемов информации. Для алгоритмов машинного обучения существует несколько видов таких конструкций.

Массивы — простая структура, которая позволяет хранить данные одного типа и обеспечивает прямой доступ к своим элементам. Представление табличных данных возможно при помощи матриц, которые являются двухмерными массивами и отображают данные в структурированном виде. Матричные вычисления широко применяются в приложениях машинного обучения [1].

Связанные списки необходимы для обработки последовательных данных или построения конвейеров. В отличие от массивов они обеспечивают динамическое распределение памяти, что делает их пригодными для обработки данных различного

размера. Вставка и удаление элементов из связанных списков проста, благодаря этому они используются при работе с потоковыми данными, когда требуются обновления в режиме реального времени [1, 2].

Деревья решений — один из методов автоматического анализа данных. Их внутренние узлы реализуют решающие правила алгоритма, а листовые узлы — варианты окончательных результатов. Эти структуры хорошо интерпретируются и могут решать задачи как классификации, так и регрессии [1, 3].

Еще один важный тип структур данных — метрические деревья (metric trees) [4], которые предназначены для разметки данных в метрических пространствах. Исследования метрических древовидных структур данных получили расцвет в конце 1990-х годов, первая монография, посвященная им, была опубликована в 2006 г. [5]. Метрические деревья представимы в виде связных графов, не содержащих циклы, у любого внутреннего узла такого дерева есть некоторое число узлов-потомков и только один предок. Узел, у которого нет предков, называется корневым, каждый узел метрического дерева можно рассматривать как корневой для поддерева, «растущего» из него. Поддерево — часть древообразной структуры данных, которая может быть представлена в виде отдельного дерева.

Если все узлы дерева имеют не более двух потомков, то оно называется бинарным; главное его преимущество — высокая эффективность реализации алгоритмов поиска и сортировки, в частности поиска ближайших соседей [6, 7]. В связи с этим важно отметить, что при кластеризации точек метрического пространства дендрограммы иерархических агломеративных методов являются бинарными метрическими деревьями [8, 9]. Любой кластер, содержащий два и более элементов, можно разделить на подкластеры, которые на дендрограмме изображаются соответствующими поддеревьями, это позволяет сформировать в метрическом пространстве таксономию элементов множества данных с дополнительными свойствами, например, автоматически задать высоту дерева, считая, по определению, что число уровней, на которых располагаются его узлы, совпадает с количеством вариантов разбиения выборочного множества на кластеры, подкластеры, подкластеры подкластеров и т. д. Такую задачу можно решить, используя аппроксимационно-оценочные критерии.

2. Агломеративная кластеризация и эвристический «метод локтя». Под кластерным анализом понимают алгоритмическую типологизацию элементов некоторого множества (выборочной совокупности) $X = \{X_1, X_2, \ldots, X_N\}$ по «мере» их сходства друг с другом. При строгой кластеризации выборка X разбивается на непересекающиеся подмножества (кластеры), поэтому на X задается отношение эквивалентности, и отдельные кластеры можно рассматривать как классы эквивалентности [10, 11]. В качестве их независимых представителей обычно 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 — центры масс соответствующих кластеров [10]. Одна из основных проблем кластерного анализа — определение числа кластеров. В общем случае задача не решена [12, 13], а для иерархических методов ее решение связано с моментом остановки процесса.

Иерархические агломеративные алгоритмы кластеризации начинаются с обработки каждого элемента выборки X как одноэлементного кластера. Если не определен момент остановки процесса кластеризации, пары кластеров последовательно

объединяются до тех пор, пока все классы эквивалентности не будут объединены в один большой кластер, содержащий все элементы X.

Для визуального представления результатов агломеративной кластеризации используются дендрограммы, включающие некоторое количество уровней, представляющих собой каждый шаг процесса последовательного укрупнения кластеров. Взаимные связи между элементами множества X изображаются при помощи ребер. Внутренние узлы дендрограммы можно рассматривать как корневые для поддеревьев, содержащих элементы выборки X, и на каждом уровне дендрограммы их совокупность составляет исходное множество элементов для следующей итерации алгоритма кластеризации.

Объединение узлов дерева соответствует слиянию двух текущих кластеров, находящихся на минимальном «расстоянии» друг от друга, согласно выбранной мере сходства. Высота ребра вертикальной дендрограммы (длина ребра горизонтальной дендрограммы) зависит от минимального значения «расстояния» F_i между текущими кластерами на i-м шаге процесса кластеризации. Числовые значения минимальных «расстояний» образуют кортежи, которые имеют вид $F = \langle F_1, F_2, \ldots, F_{N-1} \rangle$. Для агломеративных методов (кроме центроидного) числовые значения компонент кортежа F всегда монотонно возрастают $F_1 \leqslant F_2 \leqslant \ldots \leqslant F_{N-1}$ [14, 15].

Если объединяются элементы X_i и X_j, «физически» принадлежащие одному кластеру, численные значения компонент кортежа F при увеличении подстрочного индекса возрастают медленно, и их монотонное изменение является почти линейным. В случае центроидного метода, если $F_{i-1} > F_i$, то при построении кортежа F применяется простейший фильтр, при котором компоненте F_i присваивается значение F_{i-1}. Подобный характер роста числовых значений компонент кортежа F сохраняется до тех пор, пока формируются кластеры из «близких» относительно друг друга элементов множества X, но как только начинается объединение сформировавшихся кластеров, очередная компонента F резко возрастает. В этот момент график числовых значений компонент F похож на «руку, согнутую в локтевом суставе», т. е. для определения момента остановки процесса агломеративной кластеризации можно применить «метод локтя».

Идея эвристического «метода локтя» впервые была высказана американским психологом Робертом Л. Торндайком в 1953 г. [16]. На это, например, указывают Олдендерфер и Блэшфилд [17]. Смысл такого метода заключается в следующем: если график некоторой переменной величины, описывающей, например, процесс кластеризации, напоминает руку, то «локоть» (точка резкого изгиба графика) является хорошим показателем того, что в данный момент получено предпочтительное число кластеров. Заметим, что если некоторая величина сначала возрастала линейно, то в точке «локтя» ее рост становится нелинейным.

3. Аппроксимационно-оценочные критерии. Если рассматривать агломеративную кластеризацию как квазидетерминированный случайный процесс [18, 19], то его траекториями будут кортежи минимальных расстояний F. Тогда при построении критериев завершения процесса кластеризации возможно использование квадратичных форм аппроксимационно-оценочных критериев [18, 20]. В этом случае определение количества кластеров основано не на эвристических выводах, а на последовательном статистическом анализе.

Рассмотрим бинарную задачу проверки статистических гипотез H_0 и H_1.

Нулевая гипотеза H_0 — последовательность y_t возрастает линейно, альтернативная гипотеза H_1 — последовательность y_t возрастает нелинейно. Для проверки

статистической гипотезы необходимо сформулировать критерий как строгое математическое правило, позволяющее ее принять или отвергнуть. В общем случае принятие решения в некоторый момент времени может быть основано только на известных значениях дискретного квазидетерминированного случайного процесса $\xi = \xi(t, \omega)$, где t — дискретное время, ω — случайное событие, принадлежащее некоторому вероятностному пространству $(\Omega, \mathcal{F}, \mathrm{P})$. Если использовать формальный подход, то изучаемые события должны быть измеримы в неубывающей последовательности σ-алгебр $\mathcal{F}_n \in \mathcal{F}$, порожденных процессом $\xi = \xi(t, \omega)$ [21]. Если τ — момент наступления некоторого события в случайном процессе $\xi = \xi(t, \omega)$ и для любого момента времени t_0 можно однозначно сказать, наступило τ или нет, при условии, что известны значения процесса $\xi = \xi(t, \omega)$ только в прошлом (слева от t_0), то тогда τ — марковский момент относительно неубывающей последовательности σ-алгебр $\mathcal{F}_n \in \mathcal{F}$, порожденных процессом $\xi = \xi(t, \omega)$ [22, 23].

В описываемом случае марковским моментом остановки квазидетерминированного случайного процесса $\xi = \xi(t, \omega)$ со случайным параметром $\omega \in \Omega$ и монотонно возрастающей траекторией y_t будет минимальное значение τ, при котором отвергается нулевая гипотеза H_0 и принимается альтернативная гипотеза H_1. Для проверки статистических гипотез H_0 и H_1 используем квадратичные формы аппроксимационно-оценочных критериев, которые строятся в виде разности квадратичной погрешности линейной аппроксимации числовой последовательности y_t и квадратичной погрешности аппроксимации этой же последовательности в различных классах нелинейных функций [10, 18].

Узлами аппроксимации для y_t являются упорядоченные пары (i, y_i), где i — натуральный аргумент, y_i — соответствующий элемент последовательности y_t. Так как подстрочный индекс однозначно определяет значение натурального аргумента, для обозначения узла аппроксимации вместо пары (i, y_i) можно просто использовать y_i и называть его натуральным узлом аппроксимации. Коэффициенты аппроксимирующих функций ищутся при помощи метода наименьших квадратов. Квадратичные формы аппроксимационно-оценочных критериев строятся локально, не по всем значениям последовательности y_t, а только по нескольким ее членам $y_{t_0-k}, \ldots, y_{t_0-2}, y_{t_0-1}$, расположенным в левой полуокрестности точки t_0.

При построении квадратичных форм аппроксимационно-оценочных критериев будем применять «метод скользящего окна». Он основан на том, что некоторое подмножество данных фиксированного размера перемещается по основному массиву с целью поиска оптимального решения. Будем рассматривать значения y_t в точках $y_0, y_1, \ldots, y_{k-1}$, полагая, что всегда $y_0 = 0$. Выполнения этого условия легко добиться на любом шаге поиска оптимального решения при помощи преобразования:

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

Существует несколько аппроксимационно-оценочных критериев, которые предназначены для определения момента, когда характер возрастания монотонной последовательности y_t изменяется от линейного типа к нелинейному, например параболическому или экспоненциальному. Квадратичная погрешность линейной аппроксимации по k натуральным узлам равна

$$\delta_l^2(k_0) = \sum_{i=0}^{k-1} (a \cdot i + b - y_i)^2 \, .$$

Тогда параболический аппроксимационно-оценочный критерий для узлов $y_0, y_1, ..., y_{k-1}$ может быть вычислен так:

$$\delta^2_{ql}(k_0) = \delta^2_l(k_0) - \delta^2_q(k_0),$$

где $\delta^2_q(k_0) = \sum_{i=0}^{k-1}(c \cdot i^2 + d - y_i)^2$ — квадратичная погрешность неполной (без линейного члена) параболической аппроксимации. Неизвестные коэффициенты a, b, c, d аппроксимирующих функций $a \cdot x + b$ и $c \cdot x^2 + d$ определяются стандартным способом при помощи метода наименьших квадратов [11, 18].

В общем случае аппроксимационно-оценочный критерий можно сформулировать следующим образом. Будем говорить, что вблизи элемента y_{k-1} тип возрастания последовательности y_t изменился с линейного на нелинейный, если для натуральных узлов $y_0, y_1, \ldots, y_{k-1}$ справедливо неравенство $\delta^2_{ql}(k_0) \leqslant 0$, а для набора узлов $y_1, y_2, \ldots, y_k$, сдвинутых на один шаг дискретности вправо, нелинейная аппроксимация стала точнее линейной, т. е. $\delta^2_{ql}(k_0) > 0$.

Иначе, в терминах последовательного статистического анализа марковским моментом остановки для квазидетерминированного случайного процесса $\xi = \xi(t, \omega)$ со случайным параметром $\omega \in \Omega$ и монотонно возрастающей траекторией y_t будет выражение

$$\tau = \min\{t \mid \delta^2_{ql}(k_0) > 0\},$$

при котором отвергается гипотеза H_0 и принимается альтернативная гипотеза H_1.

Для проведения вычислительных экспериментов воспользуемся параболическим аппроксимационно-оценочным критерием для четырех узлов аппроксимации [18]:

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

Кроме определения числа кластеров важное значение имеет конструкция «устойчивой кластеризации», например авторы работы [24] на с. 87 приводят следующее интуитивное описание этого понятия: «Устойчивость кластеризации показывает, насколько различными получаются результирующие разбиения на группы после многократного применения алгоритмов кластеризации для одних и тех же данных...». Очевидно, что если при многократной кластеризации одних и тех же данных получаются одинаковые результаты алгоритмической типологизации, то устойчивость является максимальной.

Введем преобразование кортежей минимальных расстояний $F = \langle F_1, F_2, \ldots, F_{N-1}\rangle$: $y_i = F_i + q \cdot i$ и получим кортеж $\langle y_1, y_2, \ldots, y_{N-1}\rangle$, который назовем «множеством тренда», а q — «коэффициентом тренда». При применении критерия $\delta^2_{lq}(4_0)$ не к кортежу $\langle F_1, F_2, \ldots, F_{N-1}\rangle$, а к множеству $\langle y_1, y_2, \ldots, y_{N-1}\rangle$, результаты агломеративной кластеризации будут качественно изменяться при различных значениях коэффициента q.

При использовании аппроксимационно-оценочных критериев в качестве правил остановки агломеративных методов кластеризации количественной мерой устойчивости будет величина отрезка $Q_i = [\alpha_i, \beta_i]$ изменения коэффициента $q \in [\alpha_i, \beta_i]$, при котором для выборочной совокупности X получается один и тот же результат.

Кластеризацию выборки X можно производить при разных величинах коэффициента тренда. При $q = 0$ получается максимальное число мелких кластеров, при увеличении q могут формироваться отрезки устойчивой кластеризации, которым соответствуют более крупные кластеры, пока q не достигнет такого значения, что все

элементы X объединятся в один кластер. В общем случае набор отрезков устойчивой кластеризации для различных значений параметра q можно обозначить как $Q_1, Q_2, \ldots, Q_{e-2}, Q_{e-1}, Q_e$, где последнее множество значений коэффициента тренда q является лучом, на котором все N элементов выборки X объединяются в один кластер.

4. *In silico*. При проведении вычислительных экспериментов и компьютерного моделирования использовалось синтетическое множество, содержащее 78 точек евклидовой плоскости: $X = \{(3, 3); (4, 4); (5, 5); (3, 5); (3, 5); (3, 9); (4, 10); (5, 11); (5, 9); (3, 11); (4, 22); (5, 23); (6, 24); (6, 22); (4, 24); (8, 15); (9, 16); (10, 17); (8, 17); (10, 15); (2, 26); (3, 27); (4, 28); (4, 26); (2, 28); (6, 30); (7, 31); (8, 32); (8, 30); (6, 32); (2, 34); (3, 35); (4, 36); (4, 34); (2, 36); (22, 28); (23, 29); (24, 30); (24, 28); (22, 30); (2, 13); (4, 20); (16, 28); (7, 7); (6, 7); (7, 8); (7, 8); (6, 7); (2, 20); (2, 19); (2, 21); (3, 20); (1, 20); (2, 26); (2, 25); (2, 27); (3, 26); (1, 26); (12, 15); (11, 15); (13, 15); (12, 16); (12, 14); (15, 17); (15, 16); (15, 18); (16, 17); (14, 17); (14, 17); (13, 17); (15, 17); (14, 18); (14, 16); (20, 24); (19, 24); (21, 24); (20, 25); (20, 23)\}$ и рассматривались варианты его разбиения на кластеры центроидным методом [25].

Марковские моменты остановки процесса кластеризации определялись при помощи параболического аппроксимационно-оценочного критерия $\delta^2_{lq}(4_0)$, построенного по четырем точкам (см. последнюю формулу на с. 491).

Синтетическое множество X изображено на рис. 1, *а*. График числовых значений кортежа F с моментами остановки показан на рис. 1, *б*.

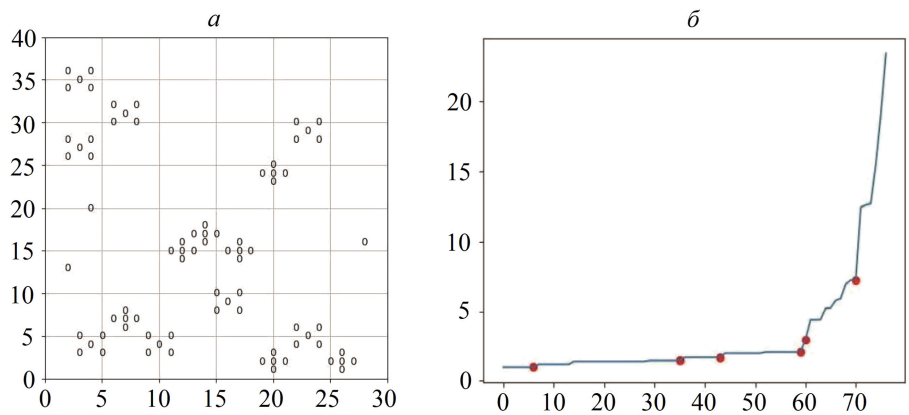

Рис. 1. Синтетическое множество X (*а*) и график числовых значений кортежа F (*б*)
На *б* точками отмечены моменты остановки для различных значений коэффициента q
(по оси абсцисс отложены номера итераций центроидного алгоритма,
по оси ординат — числовые значения компонент кортежа F).

Первоначально в эксперименте *in silico* коэффициент тренда изменялся с шагом 0.1, и было получено шесть отрезков устойчивой кластеризации: $Q_1 = [0, 1.6]$, $Q_2 = [1.7, 1.8]$, $Q_3 = [1.9, 2.3]$, $Q_4 = [2.4, 7.1]$, $Q_5 = [7.2, 12.0]$, $Q_6 = [12.1, 38.2]$. Если коэффициент q принадлежит лучу $Q_e = [38.3, +\infty)$, то все точки X объединяются в один кластер.

Результаты иерархической агломеративной кластеризации центроидным методом и соответствующие дендрограммы изображены на рис. 2–7.

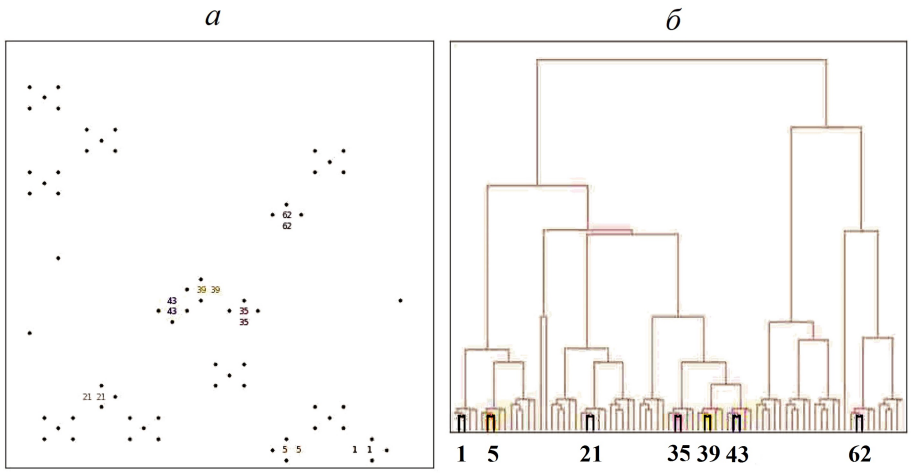

Рис. 2. Результаты кластеризации (*а*) и дендрограмма (*б*) для отрезка $Q_1 = [0, 1.6]$

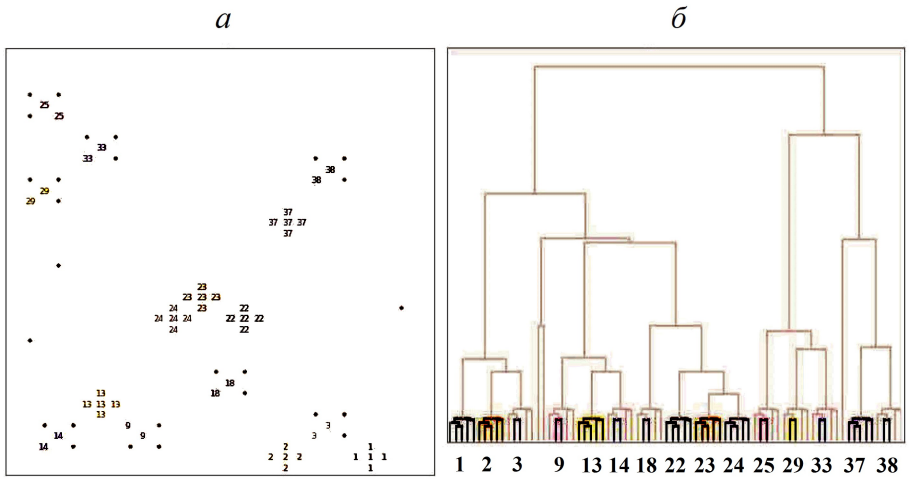

Рис. 3. Результаты кластеризации (*а*) и дендрограмма (*б*) для отрезка $Q_2 = [1.7, 1.8]$

При $q \in Q_1 = [0, 1.6]$ формируются семь двухэлементных кластеров; положим по построению, что это первый уровень (снизу) метрического дерева данных для множества X на евклидовой плоскости. Заметим, что синтетическое множество X содержит семь структур в виде правильных крестов, лучи которых направлены по горизонтали и вертикали (рис. 2, *а*), каждый такой крест содержит по четыре пары точек, находящихся на одинаковом расстоянии. Поэтому кластеры с метками **1, 5, 21, 35, 39, 43, 62** выбирались случайным образом. Будем считать, по определению, что эти кластеры представляют собой поддеревья первого уровня метрического дерева данных для множества X (рис. 2, *б*).

Второй уровень метрического дерева данных для X формируется при любом q из $Q_2 = [1.7, 1.8]$ (рис. 3). Поддеревья этого уровня соответствуют несколько более крупным кластерам с метками **1, 2, 3, 9, 13, 14, 18, 22, 23, 24, 25, 29, 33, 37, 38**.

Третий уровень метрического дерева данных для X формируется при любом q из

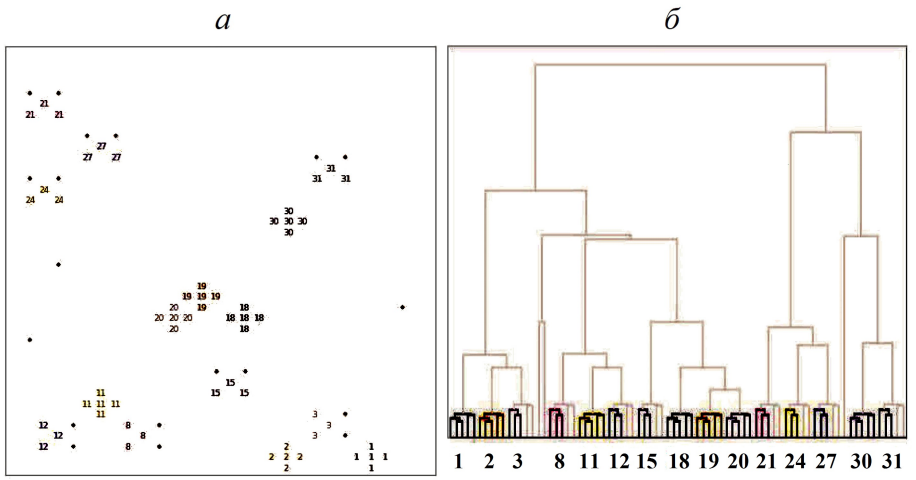

Рис. 4. Результаты кластеризации ($а$) и дендрограмма ($б$) для отрезка $Q_3 = [1.9, 2.3]$

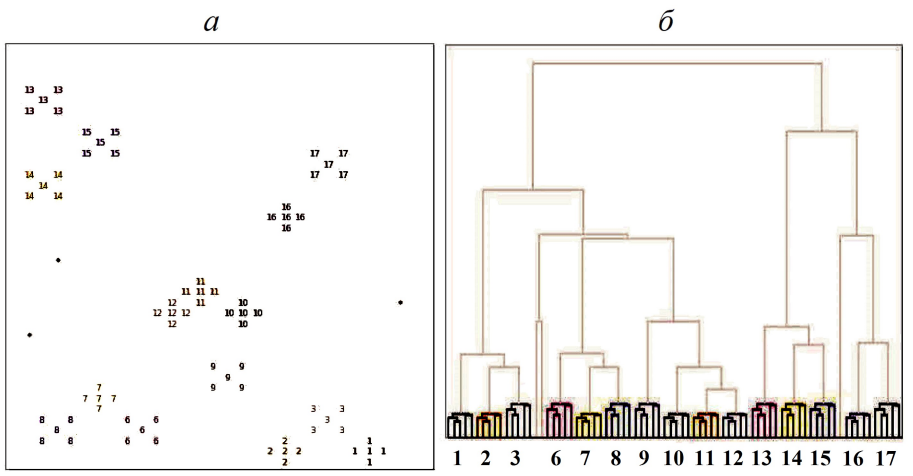

Рис. 5. Результаты кластеризации ($а$) и дендрограмма ($б$) для отрезка $Q_4 = [2.4, 7.1]$

$Q_3 = [1.9, 2.3]$ (рис. 4). Поддеревья этого уровня соответствуют кластерам с метками **1, 2, 3, 8, 11, 12, 15, 18, 19, 20, 21, 24, 27, 30, 31**.

При $q \in Q_4 = [2.4, 7.1]$ (рис. 5) формируется первый вариант «предпочтительного» числа кластеров, когда результат типологизации элементов синтетического множества X совпадает с адекватной визуальной оценкой. Полученным кластерам с метками **1, 2, 3, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17** соответствуют поддеревья четвертого уровня метрического дерева данных для X.

Если $q \in Q_5 = [7.2, 12.0]$ происходит объединение кластеров **11, 12** в один класс эквивалентности, элементы которого получают метку **11** (рис. 6), при этом формируется пятый уровень метрического дерева данных, содержащий кластеры с метками **1, 2, 3, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16**.

При $q \in Q_6 = [12.1, 38.2]$ (рис. 7) формируется второй вариант «предпочтительного» числа кластеров, когда результат кластеризации совпадает с визуальной оценкой.

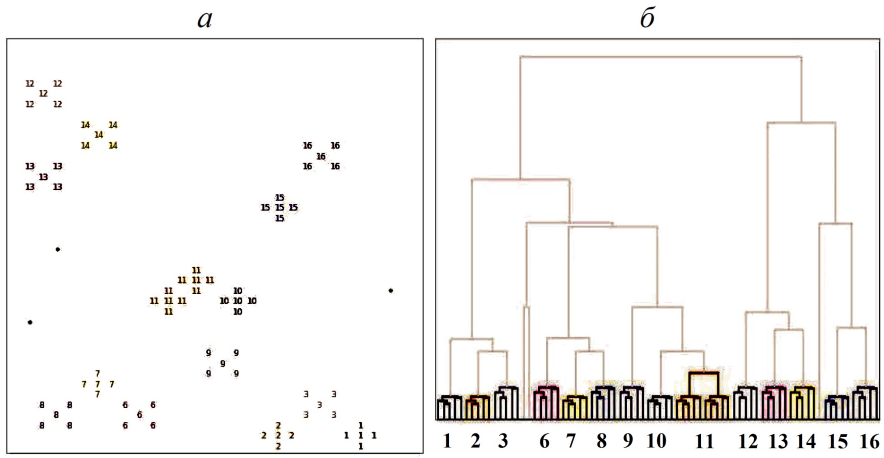

Рис. 6. Результаты кластеризации ($а$) и дендрограмма ($б$) для отрезка $Q_5 = [7.2, 12.0]$

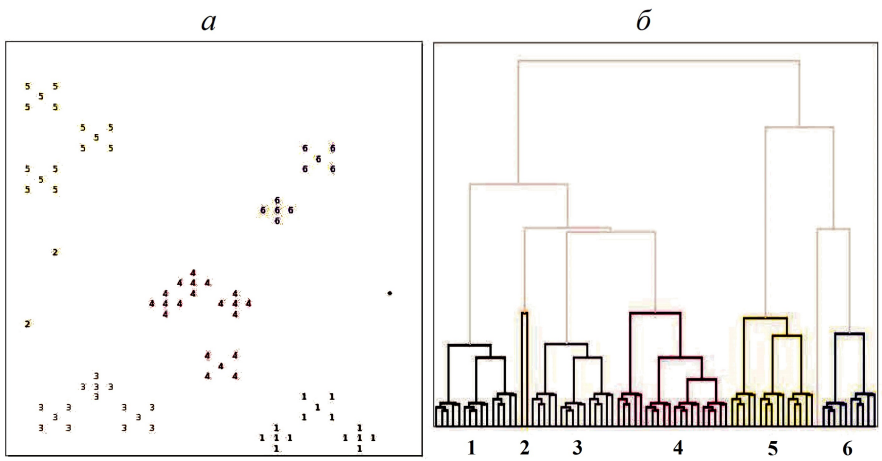

Рис. 7. Результаты кластеризации ($а$) и дендрограмма ($б$) для отрезка $Q_6 = [12.1, 38.2]$

Кластеры с метками **1, 2, 3, 4, 5, 6** формируют шестой уровень метрического дерева данных для X.

5. Заключение. Аппроксимационно-оценочные критерии — аналитическое обобщение эвристического «метода локтя». Сейчас популярным способом определения предпочтительного числа кластеров является «коэффициент силуэта» [26, 27], но в эксперименте удалось показать, что для набора данных «Ирисы Фишера» [28] аппроксимационно-оценочные критерии позволяют получить более точный результат, чем подход, предложенный Руссеу и Кауфманом [20, с. 10, 11].

Бинарное метрическое дерево — способ организации данных из евклидова пространства в виде иерархической структуры. Наибольшую эффективность оно придает алгоритмам поиска и сортировки. Например, поиск значения в неструктурированном наборе из тысячи элементов потребует до тысячи операций, тогда как в упорядоченном наборе может хватить всего несколько десятков операций. Пример, рассмотренный в п. 4, можно считать в определенном смысле «игрушечным», но даже он

иллюстрирует преимущество предлагаемого подхода к построению бинарного дерева данных в метрическом пространстве. Для множества X при помощи центроидного метода кластеризации была построена таксономия, которая без корневого узла содержит 6 уровней, соответствующих 6 различным вариантам разбиения X на кластеры, подкластеры, подкластеры подкластеров и т. д.

Кортеж минимальных расстояний F содержит 77 компонент со следующими числовыми значениями: $F = \{1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.2, 1.2, 1.2, 1.2, 1.2, 1.2, 1.2, 1.4, 1.4, 1.4, 1.4, 1.4, 1.4, 1.4, 1.4, 1.4, 1.4, 1.4, 1.4, 1.4, 1.4, 1.4, 1.47, 1.47, 1.47, 1.47, 1.47, 1.47, 1.47, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.08, 2.08, 2.08, 2.08, 2.08, 2.08, 2.08, 2.08, 2.98, 4.38, 4.38, 4.4, 5.2, 5.26, 5.85, 5.9, 6.9, 7.265, 7.28, 12.49, 12.64, 12.74, 15.552, 19.17, 23.47\}$. Компоненты F принимают 23 различных числовых значения, и если исходить из формальных правил построения дендрограммы, то она, как дерево данных без корневого узла будет иметь 22 уровня. Можно предположить, что такая структура будет мало информативной.

Это замечание легко пояснить: если мы предполагаем, например, что похожие тексты имеют сходные векторные представления, то они должны образовывать кластеры, содержащие тексты одного типа. В то же время проблема нечетких границ между разными типами текстов хорошо известна. Согласно Снелл-Хорнби [29], нет четких границ между типами текстов, а есть только определенный центр для каждого из типов. Следовательно, можно предположить, что при представлении текстовых данных нормализованными векторами в евклидовом пространстве подтемам некоторой выбранной темы соответствуют не отдельные фрагменты из корпуса текстов, а наборы подкластеров, содержащие множества текстовых фрагментов, которые можно отнести к соответствующим подтемам. Еще один пример — типологизация больных различными заболеваниями, когда к существенным признакам относится не только диагноз, но и целый набор дополнительных параметров: пол, возраст, длительность заболевания, группа крови и т. д. Понятно, что в такой ситуации построение таксономии для сотен или тысяч пациентов возможно только автоматическим способом и не для каждого человека, а для целых групп, выделенных по набору признаков.

Верификация результатов, полученных в вычислительном эксперименте, производилась за счет изменения величины шага коэффициента тренда. Как для шага 0.01, так и для шага 0.001 получились те же варианты устойчивого разбиения на кластеры множества X, что и для шага изменения коэффициента тренда q, равного 0.1. Небольшие изменения коснулись только границ отрезков Q_1, Q_2, Q_3, Q_4, Q_5 и Q_6.

Для шага 0.01 $Q_1 = [0, 1.64]$, $Q_2 = [1.65, 1.86]$, $Q_3 = [1.87, 2.32]$, $Q_4 = [2.33, 7.10]$, $Q_5 = [7.11, 12.06]$, $Q_6 = [12.07, 38.23]$. Для шага 0.001 $Q_1 = [0, 1.642]$, $Q_2 = [1.643, 1.869]$, $Q_3 = [1.870, 2.322]$, $Q_4 = [2.323, 7.104]$, $Q_5 = [7.105, 12.060]$, $Q_6 = [12.061, 38.226]$.

Построение дендрограммы иерархической агломеративной кластеризации как бинарного метрического дерева данных возможно не только центроидным методом и не только при помощи евклидовой метрики, так как выбор вариантов зависит от типа данных и специфики конкретной задачи.

Литература

1. *Mwangi D.* 5 common data structures and algorithms used in machine learning. URL: https://dzone.com/articles/5-common-data-structures-and-algorithms-used-in-ma (дата обращения: 19 сентября 2024 г.).

2. *Cormen T. H., Leiserson C. E., Rivest R. L., Stein C.* Introduction to algorithms. 4th ed. Cambridge, Massachusetts, US: MIT Press and McGraw-Hill, 2022. 1312 p.

3. *Quinlan J. R.* Induction of decision trees // Machine Learning. 1986. Vol. 1. P. 81–106. https://doi.org/10.1007/BF00116251

4. *Uhlmann J. K.* Satisfying general proximity/similarity queries with metric trees // Information Processing Letters. 1991. Vol. 40. Iss. 4. P. 175–179. https://doi.org/10.1016/0020-0190(91)90074-R

5. *Samet H.* Foundations of multidimensional and metric data structures (The Morgan Kaufmann series in data management systems). San Francisco, US: Morgan Kaufmann Publ. Inc., 2006. 1024 p.

6. *Bozkaya T., Ozsoyoglu Z. M.* Indexing large metric spaces for similarity search queries // ACM Trans. Database Systems. 1999. Vol. 24. N 3. P. 361–404. https://doi.org/10.1145/328939.328959

7. *Brin S.* Near neighbor search in large metric spaces // 21$^{\text{th}}$ International Conference on Very Large Data Bases (VLDB 1995). September 11–15, 1995. Zurich, Switzerland, 1995. P. 574–584.

8. *Жамбю М.* Иерархический кластер-анализ и соответствия / пер. с фр. М.: Финансы и статистика, 1988. 344 с.

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

10. *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/10.21638/11701/spbu10.2019.106

11. *Orekhov A. V.* Agglomerative method for texts clustering // Lecture Notes in Computer Science. 2019. Vol. 11551 LNCS. P. 19–32. https://doi.org/10.1007/978-3-030-17705-8_2

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

13. *Duda R. O., Hart P. E., Stor D. G.* Pattern classification. New York, Chichester: Wiley Press, 2001. 654 p.

14. *Lance G. N., Williams W. T.* A general theory of classificatory sorting strategies. 1. Hierarchical systems // The Computer Journal. 1967. Vol. 9. P. 373–380.

15. *Milligan G. W.* Ultrametric hierarchical clustering algorithms // Psychometrika. 1979. Vol. 44. P. 343–346.

16. *Thorndike R. L.* Who belongs in the family? // Psychometrika. 1953. Vol. 18. P. 267–276. https://doi.org/10.1007/BF02289263

17. *Олдендерфер М. С., Блэшфилд Р. К.* Кластерный анализ // Факторный, дискриминантный и кластерный анализ / пер. с англ.; под ред. И. С. Енюкова. М.: Финансы и статистика, 1989. С. 139–209.

18. *Orekhov A. V.* Quasi-deterministic processes with monotonic trajectories and unsupervised machine learning // Mathematics. 2021. Vol. 9. Art. N 2301. https://doi.org/10.3390/math9182301

19. *Левин Б. Р.* Теоретические основы статистической радиотехники. М.: Радио и связь, 1989. 656 с.

20. *Bodrunova S. S., Orekhov A. V., Blekanov I. S., Lyudkevich N. S., Tarasov N. A.* Topic detection based on sentence embeddings and agglomerative clustering with Markov moment // Future Internet. 2020. Vol. 12. Art. N 144. https://doi.org/10.3390/fi12090144

21. *Мазалов В. В.* Математическая теория игр и приложения. СПб.: Лань, 2017. 448 с.

22. *Булинский А. В., Ширяев А. Н.* Теория случайных процессов. М.: Физматлит, 2003. 400 с.

23. *Ширяев А. Н.* Статистический последовательный анализ. 2-е изд. М.: Наука, 1976. 272 с.

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

25. *Иерархическая кластеризация.* URL: https://neerc.ifmo.ru/wiki/index.php?title=Иерархическая_кластеризация (дата обращения: 19 сентября 2024 г.).

26. *Rousseeuw P. J.* Silhouettes: A graphical aid to the interpretation and validation of cluster analysis // Journal of Comput. Appl. Math. 1987. Vol. 20. P. 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

27. *Kaufman L., Rousseeuw P. J.* Finding groups in data: An introduction to cluster analysis. New York: John Wiley & Sons Inc., 1990. 342 p. https://doi.org/10.1002/9780470316801

28. *Fisher R. A.* The use of multiple measurements in taxonomic problems // Annals of Eugenics. 1936. Vol. 7. P. 179–188. http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x

29. *Snell-Hornby M.* Translation studies: An integrated approach. Amsterdam: John Benjamins Publ., 1988. 172 p.

Статья поступила в редакцию 4 июля 2024 г.
Статья принята к печати 4 октября 2024 г.

Контактная информация:

Орехов Андрей Владимирович — ст. преп.; https://orcid.org/0000-0001-7641-956X, a.orekhov@spbu.ru; a_v_orehov@mail.ru

Васильев Егор Владимирович — бакалавр; st097569@student.spbu.ru

Metric binary trees, and nested cluster hierarchy building

A. V. Orekhov, I. V. Vasiliev

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

For citation: Orekhov A. V., Vasiliev I. V. Metric binary trees, and nested cluster hierarchy building. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2024, vol. 20, iss. 4, pp. 487–499. https://doi.org/10.21638/spbu10.2024.405 (In Russian)

Machine learning methods use data trees to organize and store information. Each of these structures has certain advantages and allows improving the quality of a particular algorithm. If all tree nodes have no more than two descendants, then it is called binary; its main advantage is the high efficiency of implementing search and sorting algorithms. In this regard, it is important to note that dendrograms of hierarchical agglomerative clustering methods are also binary trees reflecting the taxonomy of elements of a data set. Any cluster that is not a singleton can be divided into subclusters, and this allows us to form a hierarchical structure in a metric space (metric tree) with additional properties. For example, automatically set the height of the tree, considering by definition that the number of levels on which its nodes are located coincides with the number of options for dividing the sample set into clusters, subclusters, subclusters of subclusters, etc. This problem can be solved using approximation-estimation tests, the change in sensitivity of which, using the trend coefficient, allows us to obtain various clustering options. When conducting computational experiments, a synthetic set of points on the Euclidean plane was used and were studied the results of centroid-based clustering. Markov moments of stopping the clustering process were determined using approximation-estimation test. Verification of the results obtained in numerical modeling was carried out by changing the step size of the trend coefficient.

Keywords: metric tree, agglomerative clustering, Markov moment, least squares method.

References

1. Mwangi D. *5 common data structures and algorithms used in machine learning.* Available at: https://dzone.com/articles/5-common-data-structures-and-algorithms-used-in-ma (accessed: September 19, 2024).

2. Cormen T. H., Leiserson C. E., Rivest R. L., Stein C. *Introduction to algorithms.* 4th ed. Cambridge, Massachusetts, US, MIT Press and McGraw-Hill Publ., 2022. 1312 p.

3. Quinlan J. R. Induction of decision trees. *Machine Learning*, 1986, vol. 1, pp. 81–106. https://doi.org/10.1007/BF00116251

4. Uhlmann J. K. Satisfying general proximity/similarity queries with metric trees. *Information Processing Letters*, 1991, vol. 40, iss. 4, pp. 175–179. https://doi.org/10.1016/0020-0190(91)90074-R

5. Samet H. *Foundations of multidimensional and metric data structures (The Morgan Kaufmann series in data management systems).* San Francisco, US, Morgan Kaufmann Publ. Inc., 2006, 1024 p.

6. Bozkaya T., Ozsoyoglu Z. M. Indexing large metric spaces for similarity search queries. *ACM Trans. Database Systems*, 1999, vol. 24, no. 3, pp. 361–404. https://doi.org/10.1145/328939.328959

7. Brin S. Near neighbor search in large metric spaces. *21th International Conference on Very Large Data Bases (VLDB 1995), September 11–15*, 1995. Zurich, Switzerland, 1995, pp. 574–584.

8. Jambue M. *Iyerarkhicheskiy klaster-analiz i sootvetstviya* [*Hierarchical cluster analysis and correspondences*]. Moscow, Finance and Statistics Publ., 1988, 344 p. (In Russian)

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

10. Orekhov A. V. Markovskii moment ostanovki aglomerativnogo protsessa klasterizatsii v evklidovom prostranstve [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/10.21638/11701/spbu10.2019.106 (In Russian)

11. Orekhov A. V. Agglomerative method for texts clustering. *Lecture Notes in Computer Science*, 2019, vol. 11551 LNCS, pp. 19–32. https://doi.org/10.1007/978-3-030-17705-8_2

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

13. Duda R. O., Hart P. E., Stor D. G. *Pattern classification.* New York, Chichester, Wiley Press, 2001, 654 p.

14. Lance G. N., Williams W. T. A general theory of classificatory sorting strategies. 1. Hierarchical systems. *The Computer Journal*, 1967, vol. 9, pp. 373–380.

15. Milligan G. W. Ultrametric hierarchical clustering algorithms. *Psychometrika*, 1979, vol. 44, pp. 343–346.

16. Thorndike R. L. Who belongs in the family? *Psychometrika*, 1953, vol. 18, pp. 267–276. https://doi.org/10.1007/BF02289263

17. Oldenderfer M. S., Blashfield R. K. Klasterniy analiz [Cluster analysis]. *Faktorniy, diskriminantniy i klasterniy analiz* [*Factor, discriminant and cluster analysis*]. Moscow, Finance and Statistics Publ., 1989, pp. 139–209. (In Russian)

18. Orekhov A. V. Quasi-deterministic processes with monotonic trajectories and unsupervised machine learning. *Mathematics*, 2021, vol. 9, art. no. 2301. https://doi.org/10.3390/math9182301

19. Levin B. R. *Teoreticheskiye osnovy statisticheskoy radiotekhniki* [*Theoretical foundations of statistical radio engineering*]. Moscow, Radio and Communication Publ., 1989, 656 p. (In Russian)

20. Bodrunova S. S., Orekhov A. V., Blekanov I. S., Lyudkevich N. S., Tarasov N. A. Topic detection based on sentence embeddings and agglomerative clustering with Markov moment. *Future Internet*, 2020, vol. 12, art. no. 144. https://doi.org/10.3390/fi12090144

21. Mazalov V. V. *Matematicheskaya teoriya igr i prilozheniya* [*Mathematical game theory and applications*]. St. Petersburg, Lan' Publ., 2017, 448 p. (In Russian)

22. Bulinsky A. V., Shiryaev A. N. *Teoriya sluchaynykh protsessov* [*Theory of random processes*]. Moscow, Fizmatlit Publ., 2003, 400 p. (In Russian)

23. Shiryaev A. N. *Statisticheskiy posledovatel'nyy analiz.* 2-ye izd. [*Statistical sequential analysis.* 2nd ed.]. Moscow, Nauka Publ., 1976, 272 p. (In Russian)

24. Granichin O. N., Shalymov D. S., Avros R., Volkovich Z. Randomizirovannyy algoritm nakhozhdeniya kolichestva klasterov [Randomized algorithm for finding the number of clusters]. *Automation and Telemechanics*, 2011, no. 4, pp. 86–98. (In Russian)

25. *Hierarchical clustering.* Available at: https://neerc.ifmo.ru/wiki/index.php?title=Иерархическая_кластеризация (accessed: September 19, 2024). (In Russian)

26. Rousseeuw P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Comput. Appl. Math.*, 1987, vol. 20, pp. 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

27. Kaufman L., Rousseeuw P. J. *Finding groups in data: An introduction to cluster analysis.* New York, John Wiley & Sons Inc. Publ., 1990, 342 p. https://doi.org/10.1002/9780470316801

28. Fisher R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 1936, vol. 7, pp. 179–188. http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x

29. Snell-Hornby M. *Translation studies: An integrated approach.* Amsterdam, John Benjamins Publ., 1988, 172 p.

Received: July 4, 2024.
Accepted: October 4, 2024.

A u t h o r s' i n f o r m a t i o n:

Andrey V. Orekhov — Senior Lecturer; https://orcid.org/0000-0001-7641-956X, a.orekhov@spbu.ru, a_v_orehov@mail.ru

Igor V. Vasiliev — Bachelor; st097569@student.spbu.ru