НОВЫЕ АЛГОРИТМЫ ИЕРАХИЧЕСКОЙ КЛАСТЕРИЗАЦИИ В СРАВНЕНИИ С КЛАССИЧЕСКИМИ МЕТОДАМИ ИЕРАХИЧЕСКОЙ КЛАСТЕРИЗАЦИИ

Валиева З. И., магистр КФУ, г. Казань, Россия
Миссаров М.Д., д.ф.-м.н., профессор КФУ,
г. Казань, Россия

Аннотация.

Предложены два новых метода аггломеративной иерархической кластеризации, основанные на новых способах расчета расстояния между кластерами. Качество кластеризации оценивается с использованием 3 внутренних показателей качества: коэффициента силуэта, индекса Данна и индекса Дэвиса-Болдуина. Эксперименты с использованием метода Монте-Карло проводились для разного количества точек, разного количества кластеров и разных размерностей данных. Новые алгоритмы в большинстве экспериментов показывают средние характеристики показателей качества, превышающие средние характеристики этих показателей для классических алгоритмов монотонной иерархической кластеризации.

Ключевые слова: иерархическая кластеризация, расстояния между кластерами, индексы качества, метод Монте-Карло.

Кластерный анализ относится к области машинного обучения без учителя. Методы кластерного анализа используются в экономике, финансах, социологии, медицине, биоинформатике и других областях [1]- [3].

Пусть T - пространство с некоторой функцией расстояния (несходства) d. Для заданного множества данных X, состоящего из n точек $x_i \in T, i=1,2,\ldots,n$ с матрицей расстояний D_n, кластерным разбиением называется разбиение X на непересекающиеся подмножества $\{C_1, C_2, \ldots, C_k\}$, где k – желаемое количество кластеров. В аггломеративной иерархической кластеризации кластеры строятся с помощью итерационной процедуры. На первом шаге процедуры все точки

объявляются кластерами. Затем находится пара ближайших точек и они сливаются в кластер. На каждой итерации вместо пары ближайших кластеров U и V образуется новый кластер $U=U \cup V$. Существует несколько вариантов определения расстояния межу новым кластером и оставшимися кластерами. Далее процедура повторяется до тех пор, пока не останется ровно k кластеров. Метод иерархической кластеризации определяется способом вычисления расстояния между кластерами . В этой работе мы исследуем 4 классических расстояния и 2 новых.

Пусть U и V обозначают некоторые кластеры. Метод одиночной связи определяется с помощью расстояния

$$D_{sl}(U,V)=\min_{x\in U, y\in V} d(x,y) \;;$$

метод полной связи определяется расстоянием

$$D_{cl}(U,V)=\max_{x\in U, y\in V} d(x,y) \;;$$

метод средней связи – расстоянием

$$D_{av}(U,V)=\frac{1}{|U|\vee V\vee ¿*\sum_{x\in U}\sum_{y\in V} d(x,y)¿} \;;$$

метод Уорда корректно определяется в Евклидовом пространстве с помощью расстояния

$$D_W(U,V)=|U|\vee V\vee \frac{¿}{|U|+¿V\vee ¿*d^2}(c_U,c_V),¿¿$$

где d(x, y) обозначает Евклидово расстояние между точками x и y, c_U и c_V обозначают центроиды кластеров U и V

$$c_U=\sum_{x\in U}\frac{x}{¿U\vee ¿}, c_V=\sum_{y\in V}\frac{y}{¿V\vee ¿}¿¿ \;.$$

Мы вводим два новых расстояния:

1) Полусумма минимального и максимального попарных расстояний между точками кластеров U и V :

$$D_h(U,V)=\frac{1}{2}\left(\min_{x\in U, y\in V} d(x,y)+\max_{x\in U, y\in V} d(x,y)\right) \;;$$

2) Медиана выборки, состоящей из всех попарных расстояний между точками из различных кластеров:

$$D_m(U,V)=median\{d(x,y):x\in U,y\in V\}.$$

Заметим, что 2-е расстояние отличается от медианного расстояния, которое определяется как расстояние между медианами кластеров (другое название - метод WPGMC).

Целью работы является проведение сравнительного анализа качества кластерных разбиений, получаемых 4-мя классическими методами иерархической кластеризации и 2-мя новыми. Для того, чтобы оценить качество различных алгоритмов иерархической кластеризации, мы используем такие индикаторы, как коэффициент силуэта, индекс Дэвиса-Болдуина и индекс Данна [2], [3].

Коэффициент силуэта является мерой сходства объектов внутри кластеров и их несходства с объектами других кластеров:

$$S_i=\frac{b_i-a_i}{max(a_i,b_i)},SI=\frac{1}{n}\sum_{i=1}^{n}S_i$$

где a_i обозначает среднее расстояние от i-ого элемента до других элементов кластера, которому принадлежит i-ый элемент, b_i обозначает среднее расстояние от i-ого элемента до всех элементов ближайшего соседнего кластера.

Индекс Дэвиса-Болдуина для кластерного разбиения $\{C_1,C_2,\dots,C_k\}$ определяется как

$$DBI=\frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{\Delta(C_i)+\Delta(C_j)}{\delta(C_i,C_j)}\right),$$

где $\Delta(C_i)$- внутри-кластерное расстояние кластера C_i, $\delta(C_i,C_j)$ –

меж-кластерное расстояние между кластерами C_i и C_j. Мы определяем $\Delta(C_i)$ для C_i как среднее Евклидово расстояние между точками этого кластера и его центроидом. Меж-кластерное расстояние $\delta(C_i,C_j)$ определяется как расстояние между центроидами кластеров C_i и C_j.

Индекс Данна используется для оценки отношения меры компактности и разделимости кластеров. Индекс Данна определяется по формуле

$$DI = \min_{i,j \in \{1,\ldots,k\}, i \neq j} \left\{ \frac{d(C_i, C_j)}{\max_{l \in \{1,\ldots,k\}} diam(C_l)} \right\},$$

где $d(C_i, C_j)$ обозначает минимальное расстояние между двумя точками из этих кластеров, диаметр кластера $diam(C_l)$ определяется как максимальное расстояние между точками кластера C_i .

Для проведения экспериментов использовались наборы точек в двумерном и трехмерном пространствах. Качество различных методов кластеризации на оценивалось с помощью коэффициента силуэта SI , индекса Дэвиса-Боулдина DBI и индекса Данна DI. Наборы точек с целочисленными координатами независимо и равномерно распределены в квадрате 2-мерного пространства [0,...,99]×[0,...,99] и в кубе 3-х мерного пространства [0,...,99]×[0,...,99]×[0,...,99] . Количество точек в наборе равнялось 100, 50 и 20. Для генерации данных мы использовали функцию np.random.randint из библиотеки NumPy. Размерность пространства данных обозначена как d, количество точек в наборе обозначено как n. Для заданных значений d и n было сгенерировано 1000 независимых наборов данных. Для каждого заданного набора и заданного количества кластеров были рассчитаны индексы SI, DBI и DI для каждого из 6 рассмотренных методов, а результаты 1000 экспериментов были усреднены. Для реализации метрик мы использовали библиотеку sklearn (scikit-learn) на Python [4]. sklearn.metrics - это модуль библиотеки sklearn, который содержит реализации различных метрик для оценки качества моделей машинного обучения [5]. Для реализации расчета коэффициента Silhouette мы использовали метрику silhouette_score библиотечного модуля sklearn [6]. Для расчета индекса Дэвиса-Боулдина используется метрика davies_bouldin_score [7]. Индекс Данна не имеет библиотечной реализации на языке программирования Python. Поэтому для вычисления индекса Данна, а также для построения алгоритмов иерархической кластеризации с использованием функций расстояния D_h и D_m мы написали пользовательский код на Python.

Количество требуемых кластеров в эксперименте варьируется от 2 до 9, количество точек в наборе данных равнялось 100, 50 и 20. Таким образом, общее количество вариантов эксперимента равно 24. Кроме того, эксперименты проводились отдельно для данных из 2-мерного и 3-мерного евклидова пространств. Результаты экспериментов зависят от количества точек (плотности точек), количества требуемых кластеров и размерности пространства данных. В каждом варианте эксперимента каждый метод кластеризации характеризуется набором средних значений из 3 индексов SI, DBI и DI. Мы ввели следующее ранжирование результатов для разных методов кластеризации: метод A превосходит метод B, если 2 или 3 средних показателя A больше соответствующих средних показателей B. Для двумерных данных метод медианы (m) превосходит метод одиночной связи (sl) в 21 варианте (из 24), метод полной связи (cl) - в 23 вариантах, метод средней связи (av) - в 20 вариантах, метод Уорда (W) - в 23 вариантах, метод полусуммы (h) - в 18 вариантах. h-метод превосходит sl-метод в 20 вариантах, cl-метод в 22 вариантах, av-метод в 13 вариантах, W-метод в 20 вариантах. Для трехмерных данных m-метод превосходит sl-метод в 15 вариантах, cl-метод в 23 вариантах, av-метод в 18 вариантах, W-метод в 23 вариантах, h-метод в 15 вариантах. h-метод превосходит sl-метод в 21 варианте, cl-метод в 23 вариантах, av-метод в 19 вариантах, W-метод в 22 вариантах. Для трехмерных данных и количества кластеров $k>5$ медианный m-метод и h-метод превосходят все классические sl-, cl-, av-, W-методы. Все эти факты свидетельствуют о том, что методы агломеративной иерархической кластеризации, основанные на новых расстояниях, в среднем превосходят 4 вышеперечисленных классических алгоритмов кластеризации на 2-мерных и 3-мерных евклидовых данных.

Литература

1. Лагутин М.Б. Наглядная математическая статистика. М.: Лаборатория знаний, 2024, 472 с.

2. Simovici D.A. Clustering. Theoretical and Practical Aspects , Singapor: World Scientific Publishing Company, 2022, 865 p.

3. Guojun G., Ma C., and Wu J. Data Clustering: Theory, Algorithms, and Applications, Philadelphia ASA-SIAM Series on Statistics and Applied Probability, SIAM, 2007, 466 p.

4. 2.3. Clusterization- scikit-learn. – URL: https://scikit-learn.ru/clustering/

5. sklearn.metrics - scikit-learn 1.5.0 documentation. – URL: https://scikit-learn.org/stable/api/sklearn.metrics.html

6. silhouette_score - scikit-learn 1.5.0 documentation. – URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score. html

7. davies_bouldin_score - scikit-learn 1.5.0 documentation. – URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_ score.html