ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2024, том 520, с. 29–34

МАТЕМАТИКА

УДК 004.622

МЕТОД ТУННЕЛЬНОЙ КЛАСТЕРИЗАЦИИ

© 2024 г. Ф. Т. Алескеров[1,2,]*, А. Л. Мячин[1,2,], В. И. Якуба[1,2,]*****

Представлено академиком РАН Д. А. Новиковым

Получено 05.04.2024 г.
После доработки 23.07.2024 г.
Принято к публикации 30.10.2024 г.

Предлагается новый метод быстрого поиска закономерностей в числовых данных большой размерности, названный “туннельной кластеризацией”. Основными преимуществами нового метода являются: относительно невысокая вычислительная сложность; эндогенное определение состава и количества кластеров; высокая степень интерпретируемости конечных результатов. Приведено описание трех различных вариаций: с фиксированными гиперпараметрами, адаптивными, а также комбинированный подход. Рассмотрены три основных свойства туннельной кластеризации. Практическое применение приведено как на синтетических (100.000 объектов), так и на классических тестовых данных.

Ключевые слова: кластер, кластеризация, кластерный анализ, туннельная кластеризация, степень перехода.

DOI: 10.31857/S2686954324060052, **EDN:** KLEIQU

1. ВВЕДЕНИЕ

Современный этап развития методов сбора и хранения информации приводит к быстрому накоплению огромных массивов данных, обработка которых требует соответствующих методов анализа. Приведем небольшую статистику: согласно [1], количество пользователей сети интернет в 2024 году превысило 5,35 млрд человек (примерно 66,2% населения Земли), а пользователями социальных сетей являются 5,04 млрд человек (62,3% населения). Количество пользователей в России также увеличивается с каждым годом. К примеру, в [2] представлены следующие значения: суммарное число визитов социальной сети ok.ru составило 520,3 млн при среднем времени пребывания пользователя на сайте около 10 минут; а социальной сети vk.com – 1,2 млрд (при среднем времени почти 11 минут).

Для поиска закономерностей среди собранных данных возможно использование нескольких основных методов, среди которых: классификация [3], построение регрессий [4], поиск аномалий [5], нейросетевые методы [6], анализ паттернов [7, 8], поиск ассоциативных правил [9], кластерный анализ [10] и отдельные методы машинного обучения [11]. Условное разбиение данных методов возможно различными способами: к примеру, обучение с учителем и без учителя; предсказательные и описательные методы. Настоящая работа посвящена развитию методов кластерного анализа, относящихся к группе “обучение без учителя”, тип решаемых задач – описательные.

Приведенный в работе метод, названный “туннельной кластеризацией” в связи с особенностями его алгоритмической реализации, имеет низкую вычислительную сложность, что позволяет обрабатывать большие массивы данных. Ручная настройка гиперпараметров позволяет регулировать конечное число кластеров, а приведенная мера оценки качества кластеризации – автоматизировать процесс нахождения оптимального (с точки зрения плотности и среднего межкластерного расстояния) разбиения.

Модели и методы кластер-анализа были предметом рассмотрения в большом количестве публикаций (см. [12, 13]). Даже краткий обзор этих методов привести здесь не представляется возможным.

2. ТУННЕЛЬНАЯ КЛАСТЕРИЗАЦИЯ

2.1. Формальная постановка задачи

Приведем описание разновидности метода, примененного ранее в НИУ ВШЭ и ИПУ РАН при решении множества прикладных задач (см. [14]).

[1] *Национальный исследовательский университет “Высшая школа экономики”, Москва, Россия*

[2] *ФГБУН Институт проблем управления им. В. А. Трапезникова РАН, Москва, Россия*

* *E-mail: alesk@hse.ru*

** *E-mail: amyachin@hse.ru*

*** *E-mail: yakuba@ipu.ru*

Отметим, что полное описание различных вариаций туннельной кластеризации, описание свойств, а также совместное использование с понятием “степени перехода” впервые описывается в литературе. Данный метод имеет классическую (для кластеризации) постановку задачи. Исследуется некоторое множество объектов X, каждый объект x_i которого описан n признаками ($x_i = (x_{i1}, \ldots, x_{in})$). Задано некоторое множество меток (номеров, имен) Y, а также некоторая выборка $X = \{x_1, x_2, \ldots, x_m\}$. На основе меры близости $d(x_i, x_j)$ необходимо разбить X на подмножества (кластеры) $c : c_l \cap c_k = \varnothing$. Таким образом, необходима функция $\beta : X \to Y$, которая $\forall x_i \in X$ ставит в соответствие некоторую метку (номер, имя) $y_i \in Y$. При этом множество, состоящее из всех полученных кластеров, обозначим через C.

Для туннельной кластеризации необходимого определить формирующий центроид x (один из объектов множества X) каждого кластера для дальнейшего объединения оставшихся на основе отличий значений признаков от значений центроида не более чем на ε. Предполагается использование сопоставимых шкал с положительными значениями либо нормирование исходных значений признаков. При этом, возможны два различных типа кластеризации:

1) на основе абсолютных значений признаков ($x_c \in R^n : x_{cj} = x_{ij}$);

2) на основе тангенсов углов наклона. При условии, что оси в системе параллельных координат [15], в которой происходит визуализация как исходных данных, так и результатов туннельной кластеризации, равноудалены друг от друга, тангенсы углов наклона кусочно-линейных функций, характеризующих объекты множества X зависят от разности $x_j - x_{j-1}$. Таким образом, получаем $x_c \in R^{n-1} : x_{cj} = x_{ij} - x_{ij-1}$.

Отметим здесь основную отличительную черту этой модификации метода — основное внимание уделяется качественному сходству совокупности признаков объекта, а не близости их абсолютных значений.

На первом этапе случайным образом выбирается $x_i \in X$ и задается центроидом, т.е. $x_c = x_i$. Далее, необходимо выбрать $\varepsilon \in (0; 1)$ (при нормировке шкал измерений $[0; 1]$) и одну из возможных реализаций:

1. Фиксированная ε-окрестность. Задается верхняя и нижняя границы кластера, как $x_c^b = (x_{c1} + \varepsilon, x_{c2} + \varepsilon, \ldots, x_{cj} + \varepsilon, \ldots, x_{cn} + \varepsilon)$ и $x_c^a = (x_{c1} - \varepsilon, x_{c2} - \varepsilon, \ldots, x_{cj} - \varepsilon, \ldots, x_{cn} - \varepsilon)$ соответственно. Для любых $x_i : x_c - \varepsilon \leqslant x_i \leqslant x_c + \varepsilon$ (или $x_c^a \leqslant x_i \leqslant x_c^b$), объекту x_i приписывается номер кластера с центроидом x_c;

2. Адаптивная ε-окрестность. Используется в случаях, когда отклонения значений по разным признакам могут отличаться. В таком случае, объекту x_i приписывается номер кластера с центроидом x_c только если $x_c - (\varepsilon|x_i| + \Delta) \leqslant x_i \leqslant x_c + (\varepsilon|x_i| + \Delta)$ при заданном параметре Δ.

3. Комбинированная ε-окрестность. Фактически, является некоторой комбинацией методов 1,2. При данной реализации туннельной кластеризации объекту x_i приписывается номер кластера с центроидом x_c только если $x_c - (\varepsilon \max_i |x_i| + \Delta) \leqslant x_i \leqslant x_c + (\varepsilon \max_i |x_i| + \Delta)$.

Однако, в таком случае, конечный результат разбиения множества объектов X на некоторые кластеры с зависит от случайно заданного начального центроида x_c. С целью получения наилучшего результата введем оценку качества кластеризации

$$P(C) = \frac{\sum_{1 \leqslant \eta_{c1}, \eta_{c2} \leqslant \alpha} (f(x_{av}^{c_1}, x_{av}^{c_2}))}{\alpha^2} - \frac{\sum_{1 \leqslant \eta_c \leqslant \alpha} \left(\frac{\sum_{x_c \in c} (f(x_c, x_{av}))}{|c|} \right)}{\alpha},$$

где $|c|$ — размерность кластера c; x_{av} — объект, значения которого являются среднеарифметическими из всех входящих в данный кластер объектов x_i; $f(x_c, x_{av}) = \sum_j (x_{cj} - x_{avj})^2$ — квадрат евклидова расстояния; α — число кластеров при конкретном разбиении; η_c — номер кластера c.

Конечное разбиение множества объектов X выбирается при $P(C) \to \max$ (при запуске q прогонов; q является гиперпараметром). При этом, фактически, при максимизации первой дроби полученные x_{av} максимально удалены друг от друга. При минимизации второй — все объекты кластеров расположены максимально плотно.

Можно также оценить вычислительную сложность туннельной кластеризации: $q \cdot |X| \cdot \sum_{i=0}^{|C|} (n - |c_i|)$, при $|C| \in [1, |X|]$. Туннельная кластеризация обладает следующими свойствами:

1. При $x_{ij} \in [0; 1]$ $\forall i = \overline{1, |X|}$, $j = \overline{1, n}$, если $\varepsilon = 1$, то $|C| = 1$;

2. $\forall \varepsilon < \min(|x_{ij} - x_{lj}|)$, $|C| = |X|$;

3. $\forall \varepsilon < \varepsilon^*$, $|C| < |C^*|$.

Замечание. Одной из возможных проблем, возникающих при анализе данных большой размерности, является возникновение кластеров, незначительно отличающихся друг от друга. При небольшом числе кластеров возможна их ручная настройка и смена реализующего алгоритма (при туннельной кластеризации — фиксированной, адаптивной и комбинированной ε-окрестности). Другим возможным подходом является применение допол-

нительных метрик оценки качества кластеризации. В связи с этим, при использовании туннельной кластеризации предлагается введение понятия “степени перехода”.

Определение. Степень перехода — минимальное число признаков, позволяющее объекту $x_i \in X$ при использовании туннельной кластеризации сменить принадлежность кластера (от c к c^*). Далее, степень перехода конкретного объекта x_i к кластеру c^* (при центроиде x_{c^*}) будем обозначать $\gamma(x_i, x_{c^*})$.

Понятие “степень перехода” имеет вполне конкретное практическое применение и может быть использовано с целью получения рекомендаций относительно точности исследуемых данных. Предположим, что на основе туннельной кластеризации были получены кластеры c^1, c^2, …, c^z, образующие центроиды которых являются соответствующие объекты x_{c^1}, x_{c^2}, …, x_{c^z}. При близких значениях $C \cdot |X|$ и $\sum_{i=1}^{|X|} \gamma(x_i, x_{c^*})$ целесообразным представляется либо огрубление данных, либо их дополнение, в связи с возможной схожестью объектов разных кластеров.

Далее, продемонстрируем использование туннельной кластеризации на нескольких практических примерах.

2.2. Туннельная кластеризация: практическое применение

Проведем несколько экспериментов, позволяющих оценить практическую реализацию туннельной кластеризации. Все расчеты выполнены для туннельной кластеризации с фиксированной ε-окрестностью при значении $q = 50$.

а) Синтетический набор данных. Для начала воспользуемся данными, состоящими из 100.000 строк, сгенерированных по 4 признакам. Все значения находятся в интервале от 0 до 1. На рис. 1 представлены распределение всех объектов по кластерам, визуализация исследуемых признаков (слева), а также частотное распределение объектов по кластерам (справа).

Рассмотрим четыре полученных кластера на рис. 2. Разбиение 100.000 объектов по кластерам получено на основе авторского комплекса программ на языке Python. Время работы составило менее 5 минут, что существенно меньше классических методов кластерного анализа. Отметим, что при равных условиях, такие методы, как k-means и DBSCAN требуют существенно больше вычислительных мощностей. На скорость работы реализующего алгоритма влияет низкая вычислительная сложность используемого метода, а также возможность сопоставления сразу векторов объектов (вместо рассмотрения отдельно взятых признаков).

Следует отметить, что в данном случае, как видно на рис. 1 (справа), количество полученных кластеров довольно большое, что обусловлено спецификой данных (получены равномерным распределением значений), а также относительно небольшим значением ε.

Далее, поскольку в приведенном наборе данных использовался генератор случайных чисел, необходимо изучить его устойчивость к первоначальной инициализации псевдослучайной последовательности. С этой целью повторим эксперимент еще 8 раз и приведем полученные результаты. На рис. 3 представлены 10 первых кластеров, полученных на основе различных случайно сгенерированных выборок, состоящих из 100.000 объектов каждая.

Поскольку первый объект для составления кластера выбирается случайным образом, каждому из 9 наборов синтетических данных проводится по 10 дополнительных запусков, после чего сравниваются конечные результаты (при одинаковом значе-

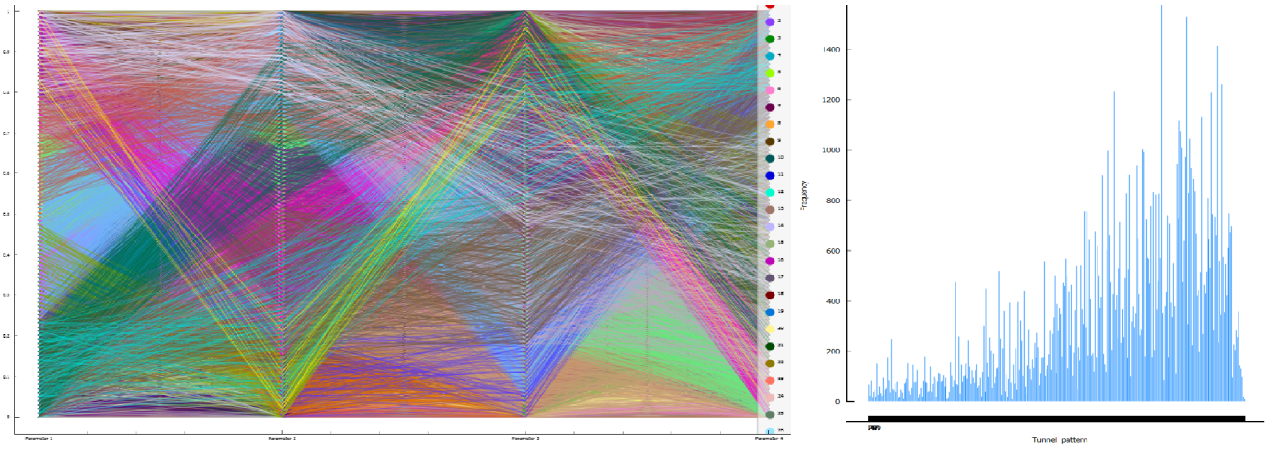

Рис 1. Визуализация в 4-мерной системе параллельных координат исходных данных и конечных результатов туннельной кластеризации.

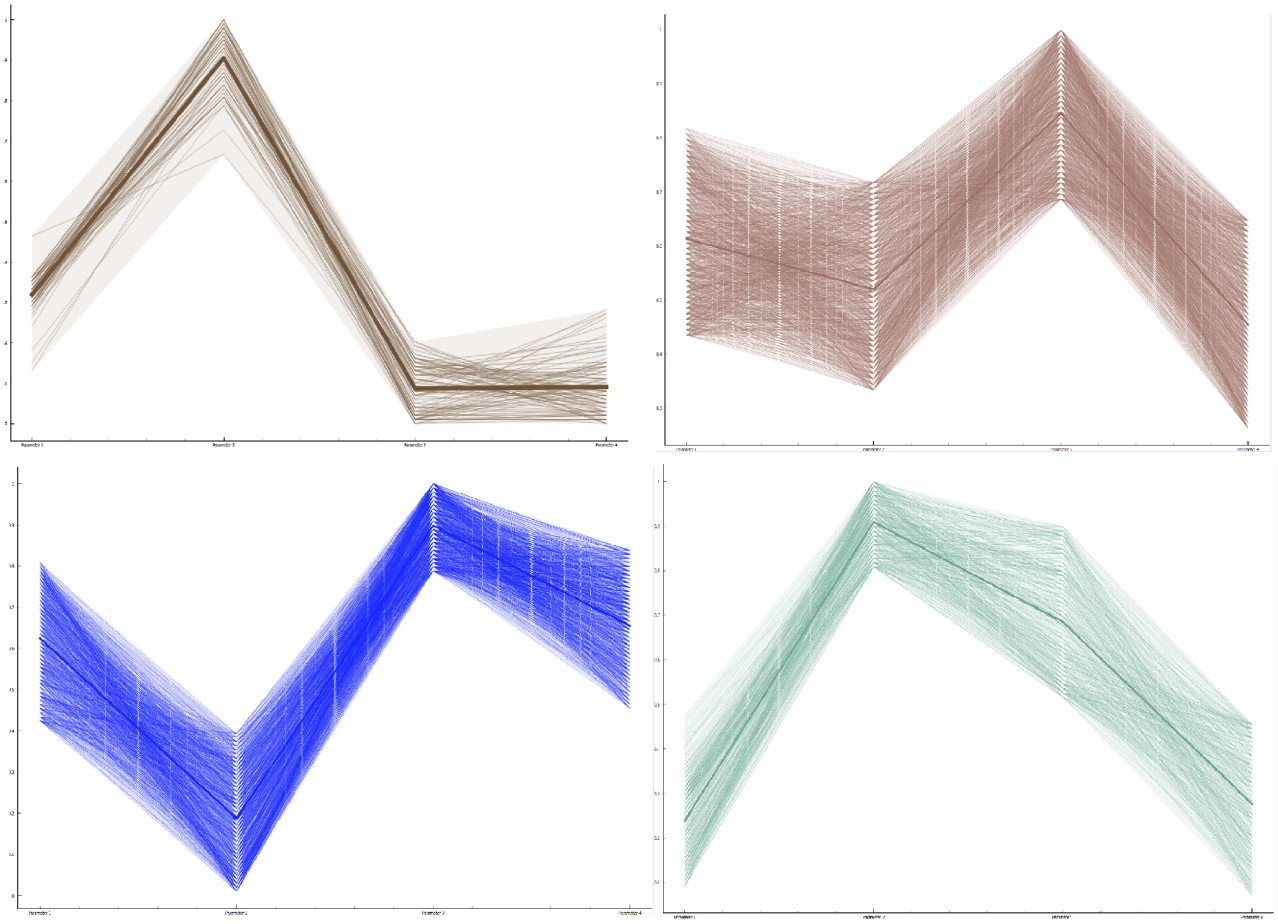

Рис 2. Примеры кластеров, полученных на основе синтетических данных с использованием туннельной кластеризации.

нии ε). Наименьший процент совпадений у четвертого запуска пятого набора данных (97,375%).

б) Классические тестовые данные. Далее, воспользуемся классическими наборами данных: Iris Data [16] и Wine Data [17]. Первый набор представляет собой 150 цветков ирисов трех возможных типов, рассматриваемых по четырем признакам (длина и ширина лепестка, длина и ширина чашелистика). Второй – химический анализ трех видов итальянских вин (всего 178 объектов и 14 признаков, таких, как: содержание алкоголя, яблочной кислоты, магния, фенолов, проантоцианидины, интенсивность цвета и оттенок, зольность и щелочность золы, флавоноиды и нефлавоноидные фенолы, пролин, а также OD280/OD315 разбавленных вин). Рассмотрим полученные на основе туннельной кластеризации результаты.

При использовании туннельной кластеризации были получены три кластера для Iris Data (что соответствует необходимым результатам). Точность кластеров: “Iris-Setosa” – 100%; “Iris Versicolor” – 96%; “Iris Virginica” – 88%.

Далее, перейдем ко второму набору классических тестовых данных Wine Data. При калибровке гиперпараметра ε с целью получения трех кластеров (поскольку известно, что исследуются три типа вин), получаем следующую точность: кластер 1 – 100%, кластер 2 – 89%, кластер 3 – 98%. В данном случае относительно высокая точность результатов достигается в том числе за счет округления исходных данных, на что указывают близкие значения $C \cdot |X|$ и $\sum_{i=1}^{|X|} \gamma(x_i, x_{c^*})$.

Таким образом, туннельная кластеризация демонстрирует свою эффективность при работе на нескольких практических примерах. Дальнейшие исследования направлены на дополнительное изучение подборки гиперпараметра ε при различных реализациях туннельной кластеризации (с фиксированной, адаптивной и комбинированной ε-окрестностью), а также на изучение отдельных случаев использования степени перехода.

3. ЗАКЛЮЧЕНИЕ

Анализ данных большой размерности связан с рядом сложностей, в т.ч. хранением, структурированием и извлечением полезной информации. Увеличение сложности и объема данных приводит

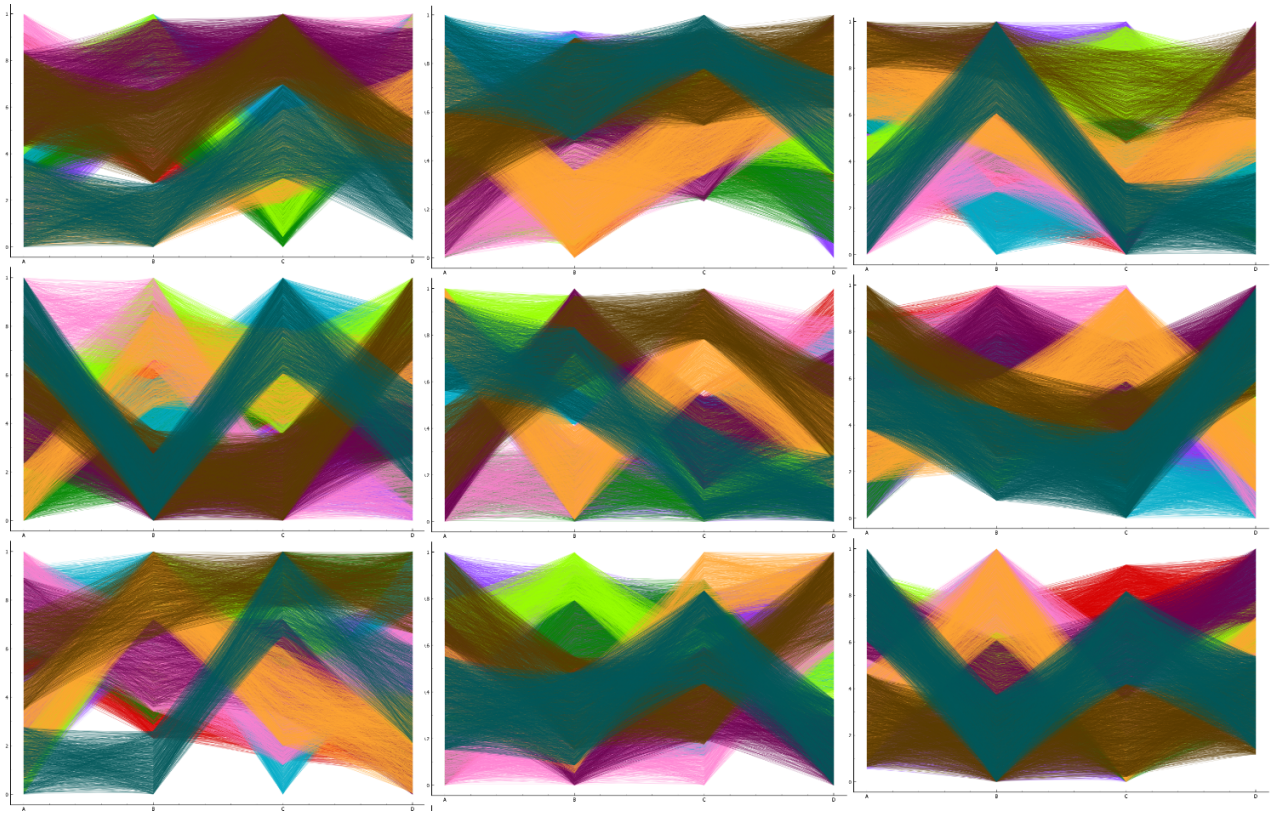

Рис 3. Примеры кластеров, полученных на основе синтетических данных с использованием туннельной кластеризации.

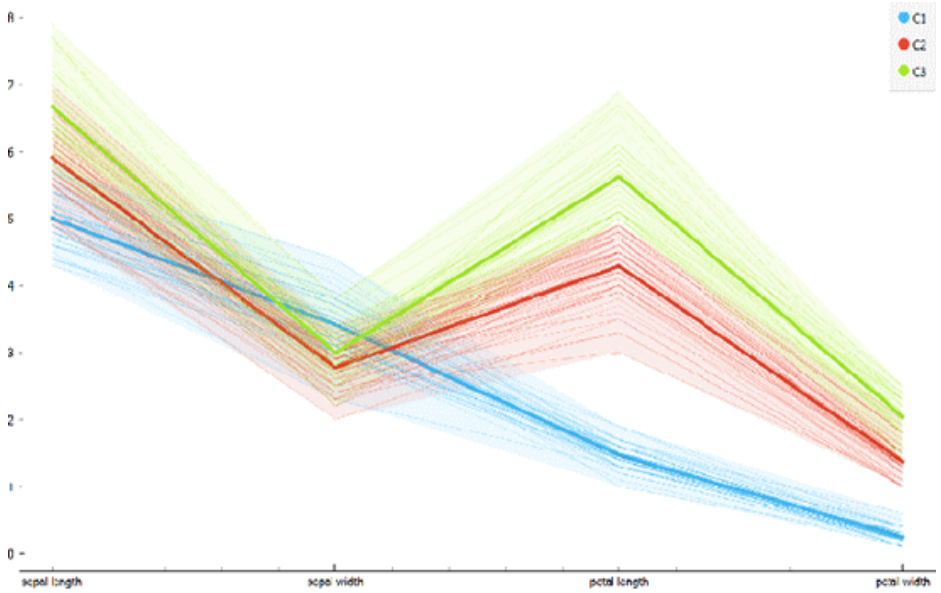

Рис 4. Результаты применения туннельной кластеризации на классических тестовых данных Iris Data.

к необходимости развития новых методов их исследования. В связи с этим, в работе предложен метод выявления закономерностей в данных, названный “туннельной кластеризацией”. Три рассмотренные практические реализации (с фиксированной, адаптивной и комбинированной ε-окрестностью) позволяют применять предложенный метод к данным различной специфики. Поскольку вычислительная сложность является низкой (при фиксированном значении ε достигается линейная сложность), туннельная кластеризация может быть использована на данных большой размерности. В частности, туннельная кластеризация использовалась для анализа покупок 1.5 млн покупателей в течение 15 меся-

цев по 400 тыс. наименований товаров для одной из крупных торговых сетей, а также для анализа более 40 млн транзакций крупного российского банка.

Предложенное к рассмотрению понятие “степени перехода” позволяет лучше понять структуру исследуемых данных, и, в отдельных случаях, необходимость их округления. Три рассмотренных примера практического использования тунельной кластеризации (как на синтетических, так и на классических тестовых данных) демонстрируют эффективность практического применения предлагаемого к рассмотрению метода.

ИСТОЧНИК ФИНАНСИРОВАНИЯ

Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ, а также при поддержке Лаборатории теории выбора и анализа решений Института проблем управления им. В. А. Трапезникова РАН. Исследование частично выполнено за счет гранта Российского научного фонда № 24-61-00030, https://rscf.ru/project/24-61-00030/.

СПИСОК ЛИТЕРАТУРЫ

1. Digital 2023: Global Overview Report. https://datareportal.com/reports/digital-2024-global-overview-report (дата обращения: 04.06.2024).
2. SimilarWeb. https://www.similarweb.com/ru/ (дата обращения 04.06.2024).
3. *Cormack R. M.* A review of classification // Journal of the Royal Statistical Society: Series A (General). 1971. V. 134. №. 3. P. 321–353.
4. *Draper N. R., Smith H.* Applied regression analysis. John Wiley & Sons, 1998.
5. *Chandola V., Banerjee A., Kumar V.* Anomaly detection: A survey // ACM computing surveys (CSUR). 2009. V. 41. №. 3. P. 1–58.
6. *Cheng B., Titterington D. M.* Neural networks: A review from a statistical perspective // Statistical science. 1994. P. 2–30.
7. *Myachin A. L.* Pattern analysis in parallel coordinates based on pairwise comparison of parameters // Automation and Remote Control. 2019. V. 80. P. 112–123.
8. *Shawe-Taylor J., Cristianini N.* Kernel methods for pattern analysis. Cambridge university press, 2004.
9. *Agrawal R., Imieliński T., Swami A.* Mining association rules between sets of items in large databases // Proceedings of the 1993 ACM SIGMOD international conference on Management of data. 1993. P. 207–216.
10. *Anderberg M. R.* Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks. Academic press, 2014.
11. *Mahesh B.* Machine learning algorithms – a review // International Journal of Science and Research (IJSR). [Internet]. 2020. V. 9. №. 1. P. 381–386.
12. *Mirkin B.* Clustering for data mining: a data recovery approach. Chapman and Hall/CRC, 2005.
13. *Romesburg C.* Cluster analysis for researchers. Lulu. com, 2004.
14. *Aleskerov F., Emre Alper C.* A Clustering Approach to Some Monetary Facts: A Long-Run Analysis of Cross-Country Data // The Japanese Economic Review. 2000. V. 51. №. 4. P. 555–567.
15. *Inselberg A.* The plane with parallel coordinates // The visual computer. 1985. V. 1. P. 69–91.
16. *Fisher R. A.* The use of multiple measurements in taxonomic problems // Annals of eugenics. 1936. V. 7. №. 2. P. 179–188.
17. Machine Learning Repository. https://archive.ics.uci.edu/dataset/109/wine (дата обращения: 04.06.2024)

TUNNEL CLUSTERING METHOD

F. T. Aleskerov[a,b], A. L. Myachin[a,b], V. I. Yakuba[a,b]

[a] *National Research University Higher School of Economics, Moscow, Russia*

[b] *V. A. Trapeznikov Institute of Control Science of Russian Academy of Science, Moscow, Russia*

Presented by Academician of the RAS D. A. Novikov

We propose a novel method for rapid pattern analysis in high-dimensional numerical data, termed “tunnel clustering”. The main advantages of this method are its relatively low computational complexity, endogenous determination of cluster composition and number, and a high degree of interpretability of the final results. We present descriptions of three different variations: one with fixed hyperparameters, an adaptive version, and a combined approach. Three fundamental properties of tunnel clustering are examined. Practical applications are demonstrated on both synthetic datasets containing 100,000 objects and on classical benchmark datasets.

Keywords: cluster, clustering, cluster analysis, tunnel clustering, transition degree.