Adabiyotlar ro‘yxati

1. Otto C.M. Textbook of Clinical Echocardiography. -5th ed.-Philadelphia: Elsevier, 2018.512 p.
2. Goodfellow I., Bengio Y., Courville A. Deep Learning. –Cambridge: MIT Press, 2016.–775 p.
3. Woo S., Park J., Lee J.-Y., Kweon I.S. CBAM: Convolutional Block Attention Module // Proceedings of the European Conference on Computer Vision (ECCV). – 2018. – P. 3–19.
4. Lang R.M., Badano L.P., Mor-Avi V. et al. Recommendations for cardiac chamber quantification by echocardiography // European Heart Journal. – 2015. – Vol. 16. – P. 233–271.
5. LeCun Y., Bengio Y., Hinton G. Deep learning // Nature. – 2015. – Vol. 521. – P. 436–444.
6. Litjens G. et al. A survey on deep learning in medical image analysis // Medical Image Analysis. — 2017. — Vol. 42. — P. 60–88.
7. Donahue J. et al. Long-term recurrent convolutional networks for visual recognition and description // IEEE CVPR. — 2015. — P. 2625–2634.
8. Zubayda Jumayeva, Nurbek Khushvaktov, Rustam Nizomov. BIO Web Conf., 173 (2025) 01030. DOI:https://doi.org/10.1051/bioconf/202517301030
9. Madani A. et al. Fast and accurate view classification of echocardiograms using deep learning // NPJ Digital Medicine. — 2018. — Vol. 1. — Article 6.
10. Nithish S., Maheshwari P., Venkatsubramaniam B. et al. Auto-LVEF: A novel method to determine ejection fraction from 2D echocardiograms // Communications in Computer and Information Science. — 2024. — Vol. 2093. — P. 107–122.

ГИБРИДНАЯ МОДЕЛЬ КЛАСТЕРИЗАЦИИ ТЕКСТОВ НА ОСНОВЕ FS-DBSCAN

Игнатьев Н.А., Абдуллаев К.Д.
Национальный университет Узбекистана имени Мирзо Улугбека n_ignatev@rambler.ru
abdullaevkomiljon1747@gmail.com

Аннотация: В работе рассматривается задача кластеризации текстовых данных в высокоразмерном пространстве признаков. Предлагается гибридная модель кластеризации на основе алгоритма FS-DBSCAN (Feature Stability DBSCAN). Метод ориентирован на анализ структуры данных с различной плотностью распределения и позволяет интерпретировать результаты кластеризации через устойчивость конфигураций статусов объектов. Для оценки сходства текстовых документов используется косинусная мера сходства. Эксперименты проведены на наборе текстовых данных XimTex. Полученные результаты показывают эффективность предложенного метода при обнаружении кластеров и выделении аномальных объектов.

Ключевые слова: кластеризация, DBSCAN, FS-DBSCAN, косинусная метрика, анализ текстов, устойчивость признаков.

A HYBRID MODEL FOR TEXT CLUSTERING BASED ON FS-DBSCAN

Ignatev N.A., Abdullaev K.D.
National University of Uzbekistan named after Mirzo Ulugbek n_ignatev@rambler.ru
abdullaevkomiljon1747@gmail.com

Annotation: This paper considers the problem of clustering textual data in a high-dimensional feature space. A hybrid clustering model based on the FS-DBSCAN (Feature Stability DBSCAN) algorithm

is proposed. The method is focused on analyzing the structure of data with different density distributions and allows interpreting clustering results through the stability of object status configurations. To evaluate the similarity of textual documents, the cosine similarity measure is used. Computational experiments were conducted on the XimTex dataset containing scientific text documents. The obtained results demonstrate that the proposed method effectively detects clusters and identifies anomalous objects in text data.

Keywords: clustering, DBSCAN, FS-DBSCAN, cosine similarity, text analysis, feature stability.

FS-DBSCAN ASOSIDA MATNLARNI KLASTERLASHNING GIBRID MODELI

Ignatev N.A., Abdullaev K.D.
n_ignatev@rambler.ru
abdullaevkomiljon1747@gmail.com Mirzo Ulug‘bek
nomidagi O‘zbekiston Milliy universiteti

Annotatsiya: Mazkur ishda yuqori o‘lchamli belgilar fazosida matnli ma’lumotlarni klasterlash masalasi ko‘rib chiqiladi. FS-DBSCAN (Feature Stability DBSCAN) algoritmiga asoslangan gibrid klasterlash modeli taklif etiladi. Ushbu metod turli zichlikka ega bo‘lgan ma’lumotlar strukturasini tahlil qilishga yo‘naltirilgan bo‘lib, klasterlash natijalarini obyektlar statuslari konfiguratsiyasining barqarorligi orqali talqin qilish imkonini beradi. Matnli hujjatlar o‘rtasidagi o‘xshashlikni baholash uchun kosinus o‘xshashlik metrikasi qo‘llaniladi. Hisoblash tajribalari XimTex nomli matnli ma’lumotlar to‘plamida o‘tkazildi. Olingan natijalar taklif etilgan metod matnli ma’lumotlar ichidagi klasterlarni aniqlash va anomal obyektlarni ajratishda samarali ekanligini ko‘rsatdi.

Kalit so‘zlar: klasterlash, DBSCAN, FS-DBSCAN, kosinus metrikasi, matnlarni tahlil qilish, belgilar barqarorligi.

Введение. Современные методы интеллектуального анализа данных широко применяются для обработки больших массивов информации. Одной из важнейших задач является кластеризация — разделение объектов на группы на основе их сходства. Особенно актуальной эта задача является при анализе текстовых данных, где объекты представлены в виде высоко размерных векторов признаков.

Алгоритмы плотностной кластеризации, такие как DBSCAN и HDBSCAN, позволяют обнаруживать кластеры произвольной формы [6,7] и эффективно работать с аномальным объектами. Однако эффективность этих алгоритмов во многом зависит от выбора параметров, прежде всего радиуса локальной области ε и числа ближайших соседей k.

Побудительными мотивами для разработки были попытки объяснить процесс принятия решении подбор параметра радиуса локальной области для присвоения статусов объектов (основной, грани и аномальными). Наличие статусов дало возможность для использования их в качестве категории объектов.

Таким образом возникает необходимость разработки методов, позволяющих более интерпретируемо выбирать параметры алгоритма и анализировать структуру данных с различной плотностью распределения.

В данной работе рассматривается модифицированный алгоритм кластеризации FS-DBSCAN [5], основанный на концепции устойчивости признаков (feature stability). Предложенный подход позволяет исследовать многообразие решений кластеризации и выявлять оптимальные параметры алгоритма.

Основная часть. Алгоритм DBSCAN является одним из наиболее известных методов плотностной кластеризации. Он выделяет кластеры на основе плотности распределения объектов в пространстве признаков.

Каждому объекту в алгоритме DBSCAN присваивается один из трех статусов:

- основной объект (core)

- граничный объект (border)
- аномальный объект (noise)

Основной объект имеет не менее k соседей в радиусе ε. Граничный объект принадлежит окрестности основного объекта, но сам не имеет достаточного числа соседей. Аномальные объекты не принадлежат ни одному кластеру.

Однако стандартный алгоритм DBSCAN имеет ряд недостатков:

- высокая чувствительность к выбору параметра ε
- невозможность корректной обработки кластеров с различной плотностью
- отсутствие механизма интерпретации выбора параметров.

Алгоритм FS-DBSCAN [5] решает эти проблемы за счет анализа устойчивости конфигураций статусов объектов при изменении радиуса локальной области.

Выбор метода обусловлен тем, что гарантировано можно получить единственное решение:

- Число кластеров определяется алгоритмическим путем
- Возможность оценить фоновые темы от настоящих по плотности распределения
- Фоновые темы нужно искать среди аномальных объектов
- Устойчивость признаков определяются через функции принадлежности к классам.
- Используется метод обобщении в двух классовой задаче.
- Один из классов искусственно и формируется через распределение Дирихле относительно статусов объектов (основной, граничный и аномальный)
- Оптимальное значение радиуса определяется через близость реального распределения с распределение Дирихле в искусственном классе.

Пусть задана выборка $E_\theta \equiv \{S_1, S_2, \dots, S_m\}$ где каждый объект описывается вектором признаков $X = \{x_1, x_2, \dots, x_m\}$. Для каждого объекта вычисляется расстояние до k ближайших соседей $\varepsilon(k)$. Затем радиус масштабируется коэффициентом λ. Таким образом рассматривается семейство радиусов $\lambda \cdot \varepsilon(k)$. Для каждого значения λ выполняется кластеризация и анализ распределения статусов объектов.

Для анализа структуры данных формируется новое пространство признаков Y, где каждый признак соответствует определенной конфигурации статусов объектов.

Для каждого признака вычисляется показатель устойчивости g. Значение устойчивости находится в интервале $0.5 \le g \le 1$. Если g близко к 1, признак хорошо различает классы. Если значение близко к 0.5, различие между классами отсутствует. Оптимальное значение параметра λ определяется из условия $(g - 0.5) \to min$. Это означает, что структура кластеризации максимально согласуется с заданной долей аномальных объектов. Подход, основанный на анализе устойчивости признаков, подробно рассматривается в работах Игнатьева [1,4].

По сравнению с другими алгоритмами кластеризации FS-DBSCAN имеет ряд преимуществ.

1. По сравнению с K-Means: K-Means требует заранее заданного числа кластеров и предполагает сферическую форму кластеров. FS-DBSCAN не требует задания числа кластеров и может обнаруживать кластеры произвольной формы.
2. По сравнению с DBSCAN: DBSCAN использует фиксированный радиус ε. FS-DBSCAN анализирует диапазон значений радиуса и выбирает оптимальное значение на основе устойчивости признаков
3. По сравнению с HDBSCAN: HDBSCAN использует иерархическую модель плотности. FS-DBSCAN ориентирован на интерпретируемость параметров и позволяет анализировать структуру кластеризации через устойчивость признаков.

Методы кластеризации активно используются в задачах обработки естественного языка и анализа текстовых корпусов [8,9].

Основные задачи:

- группировка документов по тематике
- обнаружение скрытых тем

- выявление аномальных текстов
- обнаружение машинно-сгенерированных текстов.

Современные методы анализа текстов также включают использование больших языковых моделей и методов обнаружения машинно-сгенерированного текста [2,3]. Каждый документ представляется в виде вектора признаков. Наиболее распространенные представления:

- TF-IDF
- word embeddings
- sentence embeddings.

Для измерения сходства между текстами используется косинусная метрика:

$\cos(x, y) = \frac{(x \cdot y)}{(||x|| \cdot ||y||)}$.

Косинусная метрика особенно эффективна для текстовых данных, поскольку учитывает направление векторов признаков, а не их абсолютную длину.

Вычислительный эксперимент

Для проверки предложенного метода использовался набор данных **XimTex (**из авторефератов диссертации ВАК Республики Узбекистан), содержащий текстовые документы. Каждый документ был представлен вектором признаков TF-IDF. Для оценки сходства документов использовалась косинусная метрика.

1 Таблица. Результаты фиксированном k и разной доле аномальным объектов.

	Anomaly %	(λ_min) lamda_min	(λ_max) lamda_max	Average radius	(λ) Extremum lamda	Using radius (r)	Clusters	Quantity of clusters	Quantity of anomal
40	15	4.41E-05	9.8911	0.005821726	1.04554	0.0061	[-1 0]	2	228
40	25	4.41E-05	9.8911	0.005821726	0.84893	0.0049	[-1 0]	2	381
40	35	4.41E-05	9.8911	0.005821726	0.70003	0.0041	[-1 0]	2	534
40	50	4.41E-05	9.8911	0.005821726	0.53789	0.0031	[-1 0]	2	762
40	65	4.41E-05	9.8911	0.005821726	0.40330	0.0023	[-1 0 1 2 3 4]	6	657
40	80	4.41E-05	9.8911	0.005821726	0.29658	0.0017	[-1 0 1]	3	209

2 Таблица. Результаты при фиксирование доле аномальных объектов и разных значения

k	Anomaly %	(λ_min) lamda_min	(λ_max) lamda_max	Average radius	(λ) Extremum lamda	Using radius (r)	Clusters	Quantity_clusters	Quantity_anomal
20	50	4.72E-05	10.506	0.0054402	0.56052	0.00305	[-1 0 1 2 3 4]		85
50	50	4.32E-05	9.7155	0.0059415	0.53089	0.00315	[-1 0]		763
90	50	4.09E-05	9.2456	0.0062798	0.51338	0.00322	[-1 0]		760
140	50	3.89E-05	8.8534	0.0065951	0.50558	0.00333	[-1 0]		762
200	50	3.71E-05	8.4821	0.0069218	0.49489	0.00342	[-1 0]		763
250	50	3.57E-05	8.2211	0.0071820	0.49613	0.00563	[-1 0]		762

В представленной таблице приведены результаты экспериментального исследования

алгоритма кластеризации для набора данных, содержащего 1523 объекта и 472 признака. Цель эксперимента состояла в анализе влияния двух основных параметров алгоритма: доли аномальных объектов (Anomaly %) и параметра соседства k.

В Таблица 1 параметр k был фиксирован (k = 40), а процент аномалий изменялся от 15% до 80%. Полученные результаты показывают, что с увеличением доли аномальных объектов радиус кластеризации уменьшается. Например, при 15% аномалий радиус составляет приблизительно

0.0061, тогда как при 80% он уменьшается до 0.0017. Это объясняется тем, что увеличение количества аномальных объектов снижает плотность данных, вследствие чего алгоритм выбирает меньший радиус для формирования кластеров. При этом наиболее выраженная кластерная структура наблюдается при 65% аномалий, где алгоритм обнаруживает 6 кластеров, тогда как при меньших значениях аномалий выделяется только 2 кластера.

В Таблица 2 экспериментов процент аномалий был фиксирован на уровне 50%, а параметр k изменялся от 20 до 250. Результаты показывают, что увеличение параметра k приводит к увеличению радиуса кластеризации. Это связано с тем, что при большем количестве соседей требуется более широкий радиус для охвата соответствующих объектов. При малых значениях k (например, k = 20) алгоритм выявляет более детализированную структуру данных и формирует 6 кластеров. Однако при увеличении k кластеры постепенно объединяются, и при $k \geq 50$ алгоритм обнаруживает только 2 крупных кластера.

Таким образом, проведённые эксперименты показывают, что параметры доли аномалий и размера локального соседства существенно влияют на структуру получаемых кластеров. Малые значения параметра k позволяют выявить более сложную внутреннюю структуру данных, тогда как большие значения приводят к укрупнению кластеров. Кроме того, увеличение доли аномалий приводит к уменьшению радиуса кластеризации и изменению распределения объектов между кластерами.

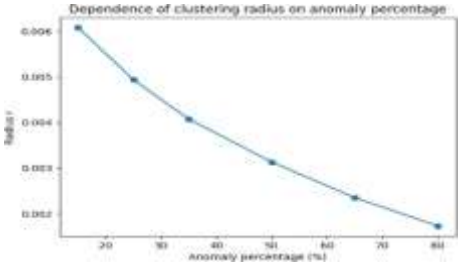

График 2

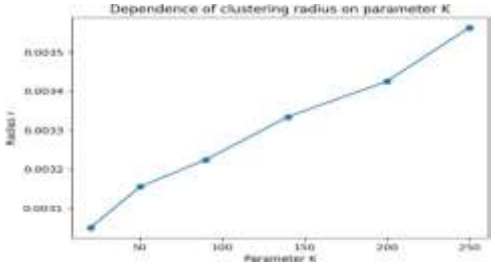

График 1

График 1. Показывает зависимость радиуса кластеризации от процента аномалий. Видно, что при увеличении доли аномальных объектов радиус уменьшается. Это свидетельствует о снижении плотности данных и необходимости более локального анализа.

График 2. Показывает зависимость радиуса кластеризации от параметра k. При увеличении числа соседей радиус также увеличивается, поскольку алгоритму требуется охватить больше объектов в локальной окрестности. Это подтверждает эффективность алгоритма FS-DBSCAN[5] для анализа текстовых данных.

Заключения. В работе предложена гибридная модель кластеризации текстовых данных на основе алгоритма FS-DBSCAN. Основной особенностью метода является использование устойчивости конфигураций статусов объектов для выбора параметров алгоритма.

Проведенный вычислительный эксперимент на наборе данных XimTex показал, что предложенный метод способен эффективно обнаруживать структуру текстовых данных и выделять аномальные объекты.

Использование косинусной метрики позволило корректно измерять сходство между текстовыми документами, что особенно важно для задач лингвистического анализа.

Предложенный подход может быть использован в различных задачах анализа текстов, включая тематическую кластеризацию документов, обнаружение аномальных текстов и анализ больших текстовых корпусов.

Дальнейшие исследования могут быть направлены на применение алгоритма FS-DBSCAN для анализа больших текстовых коллекций и интеграцию метода с современными моделями представления текстов.

Список литературы

1. Игнатьев Н. А. Выбор целевых признаков для классификации и кластерного анализа структур отношений объектов // International Journal of Open Information Technologies. – 2026. – Vol. 14, No. 3. – P. 35–42.

2. Gritsai G.M., Khabutdinov I.A., Grabovoy A.V. Stack More LLM's: Efficient Detection of Machine-Generated Texts via Perplexity Approximation // Doklady Mathematics. – 2024. – Vol. 110 (Suppl.1). – P. S203–S211. https://doi.org/10.1134/S1064562424602075

3. Gerasimenko N., Vatolin A., Ianina A. et al. SciRus: Tiny and Powerful Multilingual Encoder for Scientific Texts // Doklady Mathematics. – 2024. – Vol. 110 (Suppl.1). – P. S193–S202. https://doi.org/10.1134/S1064562424602178

4. Игнатьев Н.А., Абдуллаев К.Д. Разметка документов по семантическим ролям // Проблемы вычислительной и прикладной математики. – 2024. – №5(61). – С. 80–90.

5. Toshpulatov A.O. Problems of parameter selection in the DBSCAN algorithm in multidimensional data analysis // Zamonaviy matematikaning dolzarb muammolari va tatbiqlari. Proceedings of the Republican Scientific Conference. – Tashkent, 2026. – P. 1–2.

6. Ester M., Kriegel H.P., Sander J., Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD). – 1996. – P. 226–231.

7. Campello R., Moulavi D., Sander J. Density-Based Clustering Based on Hierarchical Density Estimates // Advances in Knowledge Discovery and Data Mining. – 2013.

8. Manning C., Raghavan P., Schütze H. Introduction to Information Retrieval. – Cambridge University Press, 2008.

9. Aggarwal C. Data Mining: The Textbook. – Springer, 2015.