УДК 539.42 : 519.67

КЛАСТЕРНЫЙ АНАЛИЗ КЛЮЧЕВЫХ ПРИЗНАКОВ АКУСТОЭМИССИОННЫХ СИГНАЛОВ ОБРАЗЦОВ ГОРНЫХ ПОРОД. ЧАСТЬ 1. ОБЗОР АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ

М.Е. Чешев, Д.С. Кульков

Проведен анализ алгоритмов кластеризации библиотеки машинного обучения scikit-learn на синтетических наборах данных, имитирующих различные варианты кластеризации. В результате анализа были выбраны два алгоритма: алгоритм, основанный на анализе спектра матрицы схожести (Spectral clustering) и плотностный алгоритм кластеризации пространственных данных с присутствием шума – density-based spatial clustering of applications with noise (DBSCAN). Применение этих алгоритмов для выделения полезных сигналов акустической эмиссии на образце гранита и сравнение скорости их работы показало, что алгоритм DBSCAN работает в среднем на два порядка быстрее, чем алгоритм Spectral clustering. Область выделения полезных сигналов акустической эмиссии обеими алгоритмами отличается: P_1≈50–100, P_2≈40–100 для DBSCAN и P_1≈5–100, P_2≈50–100 для Spectral clustering. Полученный таким образом оптимальный алгоритм кластеризации используется в дальнейшем для выделения полезных сигналов акустической эмиссии образцов горных пород.

Ключевые слова: алгоритмы; кластеры; кластерный анализ; набор данных; ключевые признаки; акустическая эмиссия.

ТОО ТЕКТЕРИНИН ҮЛГҮЛӨРҮНҮН АКУСТИКАЛЫК ЭМИССИЯ СИГНАЛДАРЫНЫН МААНИЛҮҮ БЕЛГИЛЕРИНЕ КЛАСТЕРДИК ТАЛДОО ЖҮРГҮЗҮҮ. 1-БӨЛҮК. КЛАСТЕРЛӨӨ АЛГОРИТМДЕРИНЕ СЕРЕП САЛУУ

М.Е. Чешев, Д.С. Кульков

Бул макалада ар түрдүү кластерлөө варианттарын имитациялоочу синтетикалык маалыматтар топтомунда, scikit-learn машиналык үйрөнүү китепканасынын кластерлөө алгоритмдерине талдоо жүргүзүлдү. Талдоо жүргүзүүнүн жыйынтыгында эки алгоритм тандалып алынды: окшоштук матрицасынын спектрине талдоо жүргүзүү ыкмасына негизделген алгоритм (Spectralclustering) жана чуу аралашкан мейкиндик маалыматтарын класстерлөөнүн тыгыздык алгоритми (DBSCAN). Бул алгоритмдерди гранит үлгүсүнүн акустикалык эмиссиясынын пайдалуу сигналдарын бөлүп чыгуу үчүн колдонуу жана алардын иштөө ылдамдыгын салыштыруу иштери төмөнкүдөй натыйжа берди: DBSCAN алгоритми SpectralClustering алгоритмине караганда орточо эсеп менен эки жүз эсе тезирээк болду. Бул эки алгоритмди колдонуу аркылуу бөлүнүп чыккан пайдалуу акустикалык эмиссия сигналдарынын аймактары да бири-биринен айырмаланат: DBSCAN үчүн P1≈50-100, P2≈40-100 жана SpectralClustering алгоритми үчүн P1≈5-100, P2≈50-100. Мындай жол менен талдалып алынган оптималдуу кластердик алгоритми тоотектеринин үлгүлөрүнүн пайдалуу акустикалык эмиссия сигналдарын бөлүп чыгуу үчүн колдонулат.

Түйүндүү сөздөр: алгоритмдер; кластерлер; кластердик анализ; маалымат топтому; түйүндүү өзгөчөлүктөр; аку; стикалык эмиссия.

CLUSTER ANALYSIS OF KEY FEATURES OF ACOUSTIC EMISSION SIGNALS IN ROCK SPECIMENS. PART 1: OVERVIEW OF CLUSTERING ALGORITHMS

M.E. Cheshev, D.S. Kulkov

The authors analyzed application of clustering algorithms of scikit-learn machine learning library on synthetic data sets that simulate various clusterization. As a result of analysis we selected two clustering algorithms: algorithm based on analysis of similarity matrix spectrum (Spectral clustering) and density algorithm based on clustering of spatial data with presence of noise (DBSCAN). The application of these algorithms for extracting useful acoustic emission signals of granite specimen and comparing the speed of their work showed that DBSCAN algorithm works on average two orders of magnitude faster than Spectral Clustering algorithm. Range of useful acoustic emission signals is different for both of algorithms – P1≈50–100, P2≈40–100 for DBSCAN and P1≈5–100, P2≈50–100 for Spectral clustering. The optimal clustering algorithm obtained in this way is then used to retrieve useful acoustic emission signals of rock specimens.

Keywords: algorithms; clusters; cluster analysis; data set; key features; acoustic emission.

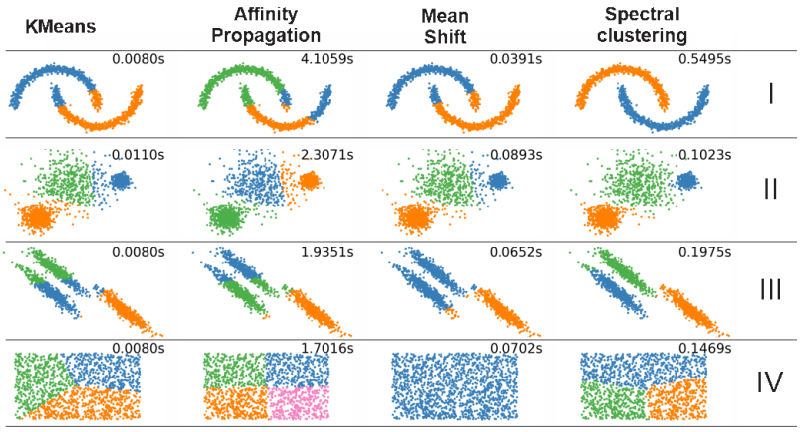

Рисунок 1 – Кластеризация данных из набора пакета scikit-learn алгоритмами KMeans, Affinity Propagation, Mean Shift, Spectral clustering

Введение. Исследование пластических и деформационных процессов, протекающих в геосреде, является актуальной задачей современной науки. Возникающие при этом акустические колебания непосредственно связаны с этими процессами, чем и обусловлен интерес к изучению акустической эмиссии. Акустическая эмиссия – излучение упругих волн, которые возникают при различных пластических процессах, протекающих в процессе перестройки внутренней структуры сред [1, 2]. Исследование акустической эмиссии на образцах горных пород, имитирующих геосреду, исследовалось в ряде работ [3–5]. Эксперименты, проводимые в НС РАН, показали, что волновые формы сигналов акустической эмиссии и шума обладают рядом ключевых признаков [6], по которым можно построить гистограмму распределения полезных и шумовых сигналов. Эти распределения могут иметь явно выраженные области группировки шумов и полезных сигналов. Поскольку при дальнейшем анализе необходимо руководствоваться повторяемостью выборки, ручное выделение этих областей каждый раз будет различным, и, как следствие этого, возникает необходимость автоматизировать выделение полезных сигналов методом кластерного анализа.

Сравнение имеющихся алгоритмов кластерного анализа в пакете scikit-learn. Существующий пакет машинного обучения scikit-learn языка python предлагает широкий набор алгоритмов кластерного анализа, а также датасеты [7] (наборы данных), имитирующих различные случаи кластеризации.

Наиболее распространенным методом выделения кластеров является алгоритм, основанный на минимизации квадратов взвешенных отклонений координат объектов от центров искомых кластеров – KMeans [8], вследствие простоты реализации и относительной быстроты работы. Кроме алгоритма KMeans, применяются и другие методы: Affinity propagation [9], реализующий объединение в кластеры с помощью матриц схожести, алгоритм сдвига среднего значения – Mean shift [10], алгоритм, основанный на анализе спектра матрицы схожести (Spectral clustering [11]), алгоритм агломеративной кластеризации [12], плотностный алгоритм кластеризации пространственных данных с присутствием шума – density-based spatial clustering of applications with noise (DBSCAN) [13], алгоритм построения дерева характерных признаков – characteristic feature tree (CFT) [14] и алгоритм ожидания-максимизации – expectation–maximization (EM) algorithm или алгоритм Гауссовой смеси (Gaussian Mixture) [15].

Для определения оптимального алгоритма построим следующие наборы данных: искривленные вытянутые кластеры (рисунок 1, группа I), шум с областями уплотнения (рисунок 1, группа II), вытянутые кластеры (рисунок 1, группа III), равномерно распределенный шум (рисунок 1, группа IV).

Как показано на рисунке 1, из построенных наборов данных явно выраженные кластеры (рисунки 1–2, I, II, III группы) были наиболее хорошо выделены методами Spectral Clustering и DBSCAN. Рассчитанные при этом скорости

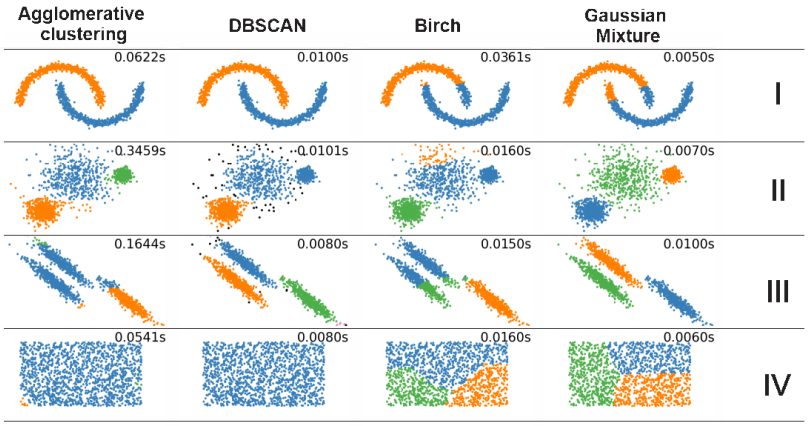

Рисунок 2 – Кластеризация данных из набора пакета scikit-learn алгоритмами Agglomerative clustering, DBSCAN, Birch, Gaussian Mixture

работы алгоритмов (таблица 1) показывают, что наиболее быстродействующим является DBSCAN.

Таблица 1 – Время работы алгоритмов Spectral Clustering и DBSCAN на наборах данных в пакете scikit-learn

	I	II	III	IV
DBSCAN	0,01	0,0101	0,008	0,008
Spectral Clust.	0,549	0,102	0,197	0,146

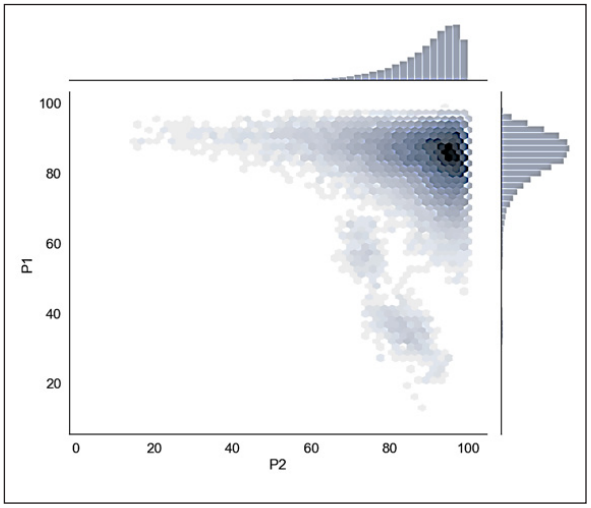

Рисунок 3 – Гистограмма распределения сигналов акустической эмиссии по ключевым признакам P1 и P2

Выделение кластеров на образце гранита алгоритмами Spectral Clustering и DBSCAN. Построим гистограммы распределения сигналов акустической эмиссии образца гранита (рисунок 3) по двум ключевым признакам – отношению мощности сигнала в пределах 40 мкс. к мощности всего сигнала (P_1) и отношению суммы амплитудно-частотного спектра в диапазоне от 200 кГц до 1.5 МГц к сумме амплитуд в спектре всего сигнала (P_2) [6].

На рисунке 3 видно, что большинство сигналов находится в области, примерно лежащей в пределах от $P_1 \approx 65–100$ и $P_2 \approx 65–100$, при этом полезные сигналы акустической эмиссии располагаются ближе к правому верхнему углу, примерно соответствуя $P_1 \approx 70–95$ и $P_2 \approx 85–100$.

Для выделения данной области применим выбранные алгоритмы кластерного анализа на образце гранита (рисунок 4). Рисунок показывает следующее: оба алгоритма выделили примерно одинаковые области, за исключением того, что алгоритм Spectral clustering включил в кластер полностью левую часть распределения, а алгоритм плотностной кластеризации ограничился значением $P_2 \approx 40$. Исходя из рисунка 3 и данных таблицы 1, получим, что алгоритм DBSCAN выделяет область, содержащую меньшее количество шумовых сигналов, и работает в среднем на два порядка быстрее.

Заключение. Проведенный анализ алгоритмов кластеризации показал, что для задач выделения полезных сигналов акустической эмиссии на образцах горных пород наиболее подходят методы Spectral clustering и DBSCAN. Исходя

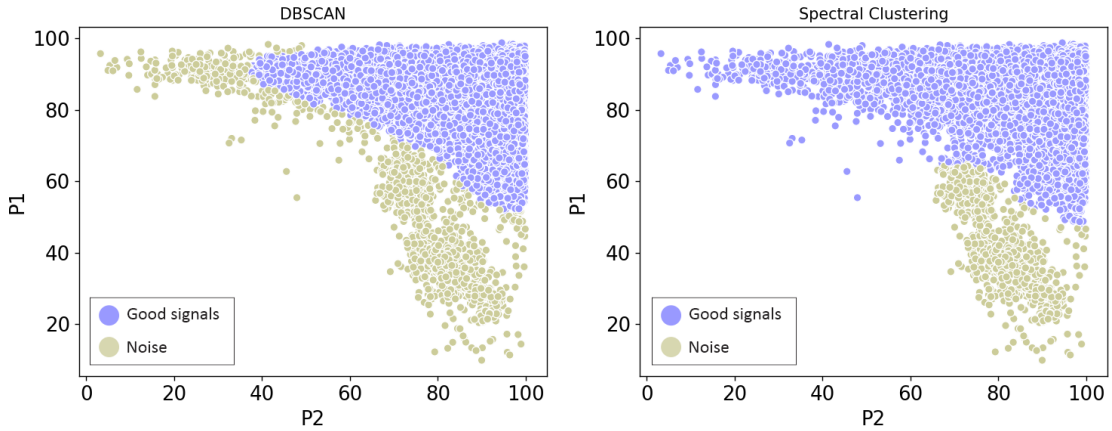

Рисунок 4 – Выделение Области полезных сигналов акустической эмиссии методами DBSCAN и Spectral clustering

из оценки скорости работы, предпочтителен алгоритм плотностной кластеризации, поскольку он работает в среднем на два порядка быстрее для всех четырех выбранных наборов данных. Анализ выделения сигналов акустической эмиссии на образце гранита показал, что алгоритм плотностной кластеризации выделяет область в границах $P_1 \approx 50–100$, $P_2 \approx 40–100$, а алгоритм спектральной кластеризации $P_1 \approx 5–100$, $P_2 \approx 50–100$, т. е. область, наиболее соответствующая положению полезных сигналов акустической эмиссии, была получена алгоритмом DBSCAN. Исходя из полученных результатов, можно заключить, что для выделения областей полезных сигналов для дальнейшего анализа целесообразно использовать плотностный алгоритм кластеризации пространственных данных с присутствием шума (DBSCAN).

Работа выполнена в рамках государственного задания Федерального государственного бюджетного учреждения науки Научной станции Российской академии наук в г. Бишкек (тема № АААА-А19-119020190064-9).

Литература

1. *Марапулец Ю.В.* Отклик геоакустической эмиссии на активизацию деформационных процессов при подготовке землетрясений / Ю.В. Марапулец, Б.М. Шевцов, И.А. Ларионов, М.А. Мищенко, А.О. Щербина, А.А. Солодчук // Тихоокеанская геология. 2012. Т. 31. № 6. С. 59–67.
2. *Закупин А.С.* Изучение влияния электромагнитного поля на нагруженные образцы горных пород тензонометрическим и акустоэмиссионным методами / А.С. Закупин // Вестник КРСУ. 2011. Т. 11. № 4. С. 73–78.
3. *Шамина О.Г.* Модельные исследования неоднородных и трещиноватых сред / О.Г. Шамина, В.И. Понятовская. М.: ИФЗ РАН, 1993. 179 с.
4. *Sobolev G.A.* Development of block hierarchy and of acoustic emission in samples of rock under three dimensional compression / G.A. Sobolev, Kh.O. Asatryan, V.A. Mansurov // J. Earthquake Prediction Res. 1995. Vol. 4. № 1. P. 107–111.
5. *Виноградов С.Д.* Акустический метод в исследованиях по физике землетрясений / С.Д. Виноградов. М.: Наука, 1989. 177 с.
6. *Кульков Д.С.* Методика предварительной обработки сигналов акустической эмиссии при одноосном сжатии образцов горных пород / Д.С. Кульков, М.Е. Чешев // В кн.: Современные техника и технологии в научных исследованиях: сб. матер. XI межд. конф. (Бишкек, 24–26 апреля 2019). Бишкек: НС РАН, 2019. С. 100–103.
7. URL: https://scikit-learn.org/stable/datasets/index.html (дата обращения: 01.07.2019)
8. *Arthur D.* k-means++: the advantages of careful seeding / D. Arthur, S. Vassilvitskii // Proceedings of the eighteenth annual ACM-SIAM Symposium on Discrete Algorithms. New Orleans, LA, January 7–9, 2007. P. 1027–1035.
9. *Dueck D.* Non-metric affinity propagation for unsupervised image categorization / D. Dueck, B.J. Frey // IEEE 11th International Conference on Computer Vision. Rio de Janeiro. Brazil, 2007. P. 1–8.
10. *Comaniciu D.* Mean shift: A robust approach toward feature space analysis / D. Comaniciu, P. Meer //

Transactions on Pattern Analysis and Machine Intelligence. 2002. Vol. 24. P. 603–619.
11. *Arias-Castro E.* Spectral clustering based on local linear approximations. / E. Arias-Castro, G. Chen, G. Lerman // Electronic Journal of Statistics. 2011. Vol. 5. P. 1537–1587.
12. *Müller A.* Introduction to Machine Learning with Python / A.C. Müller, S. Guido // O'Reilly Media Inc., 2017. P. 376.
13. *Schubert E.* DBSCAN revisited, revisited: why and how you should (still) use DBSCAN / E. Schubert, J. Sander, M. Ester, H.P. Kriegel, X. Xu // ACM Transactions on Database Systems (TODS), 42(3), 19.
14. *T. Zhang R. R.* BIRCH: An efficient data clustering method for large databases / R. R. Zhang, M. Livny // Proceedings of the 1996 ACM SIGMOD international conference on Management of data. Montreal, Quebec. Canada, 1996. P. 103–114 .
15. *Hastie T.* The Elements of Statistical Learning (2nd ed.) / T. Hastie, R. Tibshirani, J. Friedman. New York: Springer, 2009. P. 736.