Анализ подходов к оценке устойчивости кластеризации объектов на примере импульсов акустической эмиссии

П. А. Казначеев
Институт физики Земли им. О.Ю. Шмидта Российской академии наук
p_a_k@mail.ru

Д. В. Краюшкин
Институт физики Земли им. О.Ю. Шмидта Российской академии наук
krayushkindenv@yandex.ru

Г. С. Индаков
Институт физики Земли им. О.Ю. Шмидта Российской академии наук, МГУ имени М.В. Ломоносова
indakov.gs16@physics.msu.ru

З.-Ю. Я. Майбук
Институт физики Земли им. О.Ю. Шмидта Российской академии наук
rim@ifz.ru

Аннотация. **Для разделения групп импульсов акустической эмиссии разного происхождения, когда локация не позволяет различить их в пространстве или когда локация невозможна, применяется иерархическая кластеризация импульсов по мере сходства между собой. Мера сходства рассчитывается на основе взаимнокорреляционной функции пары волновых форм. Рассмотрены подходы к анализу устойчивости такой кластеризации, выделено три группы подходов, оперирующих или исходными волновыми формами, или параметрами импульсов. Рассмотрен пример данных одного эксперимента, в котором выделено три малочисленных несбалансированных кластера. Показано, что сравнительную степень устойчивости исходных кластеров с наименьшими вычислительными затратами можно оценить с помощью оценки быстроты обучения вероятностного классификатора.**

Ключевые слова: акустическая эмиссия, волновые формы, кластеризация, машинное обучение, разрушающие испытания, физика землетрясений

I. Введение

Разрушение материалов в лабораторных условиях контролируется с помощью регистрации акустической эмиссии (АЭ). Горные породы являются гетерофазными поликристаллическими материалами, разрушение которых происходит с некоторыми особенностями. Задача отслеживания эволюции разрушений горных пород важна как в фундаментальном (вулканология, физика землетрясений, метаморфизм), так и в прикладном аспекте (исследование устойчивости природно-технических комплексов).

Современные концепции развития разрушений основываются на представлении о масштабно-иерархическом характере структуры горных пород и присутствующих в них нарушений. Сейсмологические данные показывают, что подготовка землетрясений сопровождается постепенным перераспределением процесса разрушения по масштабам, которое развивается от малых размеров разрушений к крупным; обратная ситуация характерна для релаксации сейсмической активности в афтершоковом процессе [1]. В лабораторных условиях развитие разрушения в горных породах в лаборатории тоже связано с перераспределением разрушения, что связывается с появлением крупных трещин за счет взаимодействия мелких трещин [2, 3]. Особый интерес представляет исследование вариаций АЭ при совместном (не обязательно одновременном) механическом нагружении и высокотемпературном нагреве [4].

Нагрев горных пород до высоких температур может происходить в различных условиях. В природе это сопряжено, в первую очередь, с концепцией тектоники плит, с вулканической активностью при магматических интрузиях и с геотермальными системами [5]. Горные породы могут также испытывать высокотемпературное воздействие при разработке месторождений полезных ископаемых (в шахтах и карьерах при добыче твердых полезных ископаемых, в скважинах и коллекторах при добыче углеводородного сырья), в инженерных сооружениях при пожарах, в местах захоронения радиоактивных отходов (прежде всего, в гранитных массивах). В связи с этим изучение процесса термического разрушения горных пород является важным направлением развития фундаментальной и прикладной геофизической науки.

Одной из основополагающих задач, связанных с изучением распространения упругих колебаний в среде, является анализ зарегистрированных волновых форм. В рамках этой задачи активно развиваются методы выделения групп (кластеров) волновых форм и соответствующих им источников, что важно, в частности, для локализации землетрясений в сейсмологии [6] и обработки событий АЭ при

Работа выполнена в рамках государственного задания Института физики Земли им. О.Ю. Шмидта РАН

использовании методов акустического контроля [7]. Наиболее эффективным методом идентификации схожих сигналов является кросскорреляционный анализ. При этом оценивается максимум взаимной корреляционной функции волновой формы идентифицируемого сигнала с волновой формой заданного сигнала (шаблоном) как степень их сходства [6, 8].

Главная проблема такой кластеризации – очень большая вычислительная сложность. Количество операций и объем памяти составляют величину порядка $\theta(M^2 \cdot N^2)$, где M – число импульсов, N – длина импульса. Поэтому рассмотрение других методов важно для увеличения эффективности обработки данных.

Подобно многим численным задачам, для задач машинного обучения (кластеризации, классификации и пр.) можно оценить устойчивость решения по тому, как оно изменяется, если изменить обучающие данные [9] или параметры обучения [10]. Важно рассмотреть различные подходы к оценке устойчивости, учитывающие вычислительную сложность задачи.

II. Подходы к оценке устойчивости

В нашем случае устойчивость выделения исходных кластеров может быть сформулирована как суперпозиция устойчивости решения двух последовательных подзадач – вычисления попарной меры сходства волновых форм импульсов АЭ и кластеризации по полученной мере сходства.

Сами волновые формы АЭ являются вектором данных, который нельзя варьировать единственным параметром. Физически обоснованным является добавление ко всей волновой форме некоторого возмущающего временного ряда как шума или помехи, но это сложно параметризовать. Другая проблема – вычислительная сложность такой оценки устойчивости. Даже для грубой оценки методом Монте-Карло, путем добавления возмущения, корреляции возмущенных волновых форм и последующей кластеризации, потребуются колоссальные временные затраты.

С другой стороны, сама кластеризация производится аналогично подходу из сейсмологии [8], где строится матрица сходства на основе корреляционной функции и проводится бикластеризация на основе классической иерархической кластеризации – мера сходства воспринимается как расстояние (анализируемая метрика) в иерархической кластеризации. При классической иерархической кластеризации это расстояние вычисляется на основе некоторой меры над параметрами (над объектами в пространстве параметров). В нашем случае появляется некоторая обратная задача – мы знаем кластеры на основе иерархической кластеризации над некоторой матрицей расстояний, но расстояния вычислены не по параметрам. Если можно было бы приписать некоторые условные (виртуальные) параметры каждому импульсу АЭ, то матрица сходства для более или менее устойчивой кластеризации была бы, соответственно, более или менее разреженная, а сами импульсы образовывали в пространстве параметров более или менее компактные группы. Отметим, что компактность не обязательно подразумевает изолированность групп друг от друга.

Восстановление по матрице сходства объектов в пространстве виртуальных параметров - некорректная задача. Но при регистрации АЭ кроме волновых форм, определяются различные параметры импульсов, например, амплитуда, энергия, длительность и пр. [7]. Тогда можно попробовать оценить группируемость и устойчивость исходных (определенных по корреляции) кластеров импульсов АЭ в пространстве этих параметров.

Первый подход – кластеризация импульсов в пространстве параметров. При этом можно применить ту же иерархическую кластеризацию, которая используется для матрицы сходства по корреляционному методу [8].

Второй подход – оценить устойчивость модели классификатора. Для кластеров с лучшей группируемостью импульсов (и, следовательно, лучшей устойчивостью исходной кластеризации) модель классификатора при одинаковых гиперпараметрах будет оптимально обученной (см. пояснения ниже). Т. е. подход позволяет сравнить различные исходные кластеры по своей группируемости и устойчивости.

В свою очередь, для второго подхода можно предложить два варианта реализации:

1). Варьирование подобно [9] входных данных, т.е. параметров импульсов АЭ, например, путем добавления случайной величины, связанной с погрешностью измерения этого параметра (условно - «зашумленный набор импульсов»). Тогда модель классификатора, обученная на исходной разметке, может быть проверена на зашумленном наборе. Но это нельзя реализовать без повторной трудоемкой обработки временных форм, поскольку часть параметров определяется косвенно. Например, погрешность измерения времени маленькая, а амплитуды большая, поэтому погрешность измерения длительности будет определяться не погрешностью измерения времени, а погрешностью порога по амплитуде, и будет разной для каждой волновой формы.

2). Варьировать подобно [10] параметры обучения модели машинного обучения. Недообученная модель будет неустойчивой при изменении положения в пространстве параметров объектов любых классов, а переобученная – объектов только целевого класса. Тогда полученный классификатор можно оценить на устойчивость несколькими способами:

- Визуально, построив области классификации в пространстве параметров. Этот способ, хотя и нагляден, но не формализуем, и очень трудоемок, если проводить визуализацию для ряда гиперпараметров.
- Оценить некоторую меру близости импульсов из исходных кластеров к границам областей классификации. В этом способе необходимо сначала предложить и протестировать адекватную метрику меры близости, что непросто в виду разнородности (разной размерности) параметров, например, амплитуды и время.
- Оценить, как быстро модель становится переобученной из недообученной (т.е. обретает обобщающую способность) при варьировании

параметров обучения (т. е. гиперпараметров – например, увеличении глубины дерева решений), при условии, что можно применить вероятностный подход (т. е. с использованием элемента случайности).

Третий способ может дать конкретную количественную оценку качества модели, при этом не требует введения новых метрик. Практически всегда для этого используется разбиение набора данных на обучающую и тестовую выборки. Для минимизации систематической погрешности разбиения используют разные методики, например, кросс-валидацию с перемешиванием данных. Граница недообученная/переобученная модель оценивается по расхождению кривых ошибок на обучающей и тестовой выборках при изменении гиперпараметра.

Применимость этого способа зависит от качества данных. Данные могут быть сильно несбалансированными, когда количество некластеризованных импульсов намного больше количества импульсов в исходных кластерах, а сами кластеры могут быть малочисленными, т. е. число объектов в них очень мало, менее одного-двух десятков. В этом случае проблемой является представительное разбиение данных на выборки обучение/тест Даже при перемешивании состав выборок будет очень разным.

В таком случае предлагается следующий подход. Для задачи классификации мы не определяем границу недообученной и переобученной модели, а анализируем только ошибку на обучающей выборке, в которую входят все импульсы. Подвыборки при этом могут формироваться только при обучении самой модели, т.е. внутри неё, для оценки изменчивости решения.

И тогда тот класс (соответствующий одному из исходных кластеров), для которого ошибка на обучающей выборке в зависимости от гиперпараметров быстрее выходит на асимптоту, является лучше группирующийся, даже если объекты класса не образуют компактной группы в пространстве параметров. Т. е. лучше группирующийся класс – это такой, как можно больше объектов которого может быть охвачено областью, задаваемой как можно меньшим числом условий (неравенств, с точки зрения дерева решений). Исходный кластер, который соответствует лучше группирующемуся классу, является более устойчивым. При этом не важны гиперпараметры модели и конкретная граница между недообученной и переобученной моделями, поскольку сам поиск модели оптимального классификатора не является целью.

Для применения этого подхода нужно, чтобы либо при построении модели был элемент случайности, либо она выдавала вероятность отнесения объекта к классу. Дерево решений является детерминированным алгоритмом. Ансамбль деревьев может быть сформирован с позиций вероятностного подхода.

Самый очевидный вариант ансамбля со случайностью – это случайный лес. Но случайность в нем заключается в случайном формировании подвыборок данных, что возвращает нас к проблеме крайней малости относительного количества объектов отдельных классов – часть объектов из класса может не попасть в выборки. Априорное принудительное взвешивание объектов (присвоение разной вероятности попадания объекта в выборку) может потребовать большого количества деревьев решений.

Другой вариант ансамбля с вероятностным подходом – это бустинг. В бустинге результат отнесения объекта к классу является суперпозицией отдельных деревьев, взвешенных со своим весом. Каждое дерево является одновременно новой итерацией работы модели, его вес формируется в зависимости от его ошибки.

Адаптивный бустинг на каждой итерации делает подвыборку из исходных данных, и неудачно классифицированные данные получают больший вес как вероятность попадания в подвыборку на следующей итерации. За счет этого адаптивный бустинг обычно быстрее сходится – требует меньшего числа деревьев, чем случайный лес [11]. Т.е. адаптивный бустинг сам подстраивает подвыборки под редкие объекты.

Более того, некоторые авторы показывают, что адаптивный бустинг, вероятно, обладает свойством «самоусреднения», сглаживания интерполирующих границ, как и случайный лес, но при значительно меньшем количестве суб-моделей (например, деревьев решений) [11]. Т. е. то, что случайный лес может сделать за счет построения очень большого количества деревьев решений для случайных подвыборок данных, адаптивный бустинг делает на меньшем количестве деревьев. Отсюда также следует, что результат классификации адаптивного бустинга сохраняет больший элемент случайности в усреднении, поскольку решение определяется только результатом классификации исходных объектов (верно - неверно), без учета производных параметров (например, скорости сходимости, как у градиентного бустинга).

III. Применение подходов к анализу АЭ

Данные подходы были применены при анализе результатов эксперимента по регистрации АЭ при высокотемпературном нагреве образца песчаника, предварительно подвергнутого механическому нагружению. Механическое нагружение осуществлялось в одноосном режиме до нагрузки, близкой к разрушающей, с проявлением на поверхности признаков крупных трещин. После этого осуществлялось термическое воздействие до 650°C с регистрацией импульсов термоакустической эмиссии. Регистрация и обработка импульсов АЭ проводились в системе ALine 32D компании Интерюнис. Методика и результаты экспериментов подробнее описаны в [12].

Была проведена кластеризация импульсов на основе их сходства, определенного методом взаимной корреляции волновых форм, в программе ALine Stat. Выделено три кластера, каждый из которых содержит порядка 10 импульсов АЭ и может соответствовать определенному источнику, предположительно – отдельной крупной трещине. Отметим, что обработка около 3000 импульсов АЭ заняла более 8 часов на стандартном персональном компьютере, а обработка более 5000 импульсов АЭ становилась невозможной из-за нехватки памяти.

Распределение импульсов из кластеров во времени и по амплитудам в процессе нагрева различно для каждого кластера. Это может говорить как о разных моментах и пороге активации соответствующей трещины, так и о разной активности роста трещин. В пространстве параметров импульсы из кластеров не образуют изолированных групп (рис. 1).

Отдельно проводился контрольный эксперимент с термическим воздействием на образец песчаника, не подвергавшегося предварительному механическому нагружению, для которого не удалось выделить кластеры импульсов АЭ.

Первоначально была сделана попытка кластеризации в пространстве параметров. Методы кластеризации, использующие алгоритмы на основе вычисления плотности объектов, такие как метод k-средних, DBSCAN и т. п., не подходят для этой задачи, поскольку импульсы из исходных кластеров распределены на плотном фоне остальных некластеризованных импульсов. Поэтому была использована иерархическая кластеризация в агломеративном варианте, с формированием дендрограммы в направлении «снизу вверх», от «листьев» к «корню». Использовалась евклидова норма и три метода оценки связности между кластерами – метод Уорда, пригодный для выделения близкорасположенных кластеров, и, для сравнения с ним, медианный и центроидный методы [13]. Но такая кластеризация не позволила выделить группы импульсов, подобные исходным кластерам.

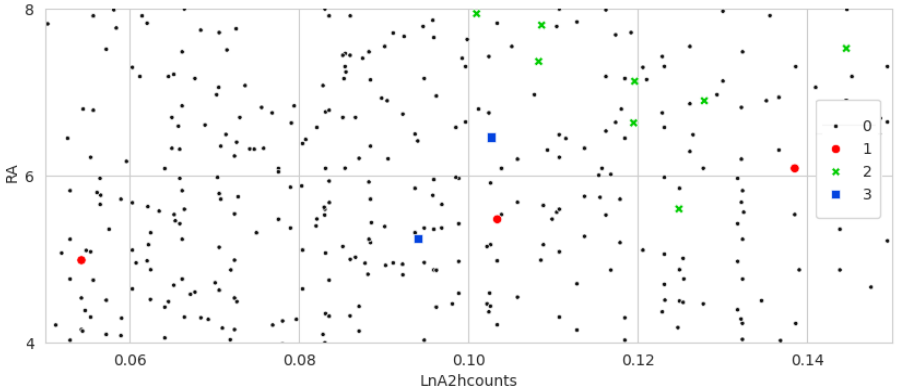

Рис. 1. Пример фрагмента распределения импульсов АЭ в пространстве избранных параметров «RA» – «LnA2hcounts». 1-3 – импульсы АЭ из 1-3 исходных кластеров, 0 – некластеризованные импульсы

Затем была сделана попытка оценить обучаемость классификатора. Анализ ошибки обучения вероятностного классификатора на основе ансамбля решающих деревьев в зависимости от гиперпараметров позволил определить, что третий кластер легче всего, а второй кластер – сложнее всего выделяются в пространстве параметров. Предположительно, исходное выделение третьего кластера наиболее устойчиво.

IV. Заключение

В работе рассмотрены различные подходы к анализу устойчивости решения задачи кластеризации как численной задачи машинного обучения, их формализуемость и вычислительная сложность.

На данных конкретного эксперимента по термическому воздействию на образец предварительно нагруженного песчаника проанализирована устойчивость кластеризации импульсов зарегистрированной акустической эмиссии. Кластеризация проводилась иерархическим методом по мере сходства, определенной на основе взаимной корреляции волновых форм импульсов. Было выделено три кластера импульсов, предположительно связанных с развитием отдельных крупных трещин.

Иерархическая кластеризация по параметрам импульсов (амплитуда, энергия, средняя частота, скорость нарастания и затухания) не позволила выделить группы импульсов, сколь-нибудь похожие на исходные кластеры. Анализ ошибки обучения вероятностного классификатора на основе ансамбля решающих деревьев в зависимости от гиперпараметров позволил определить, что третий кластер легче, а второй кластер – сложнее всего выделяется в пространстве параметров. Предположительно, выделение третьего кластера по корреляции волновых форм наиболее устойчиво.

Это позволяет сделать практический вывод, что импульсы из третьего кластера вероятнее всего связаны с крупной трещиной, которая активируется при определенной температуре (около 600°C) и при дальнейшем нагреве начинает активно расти с увеличением энергии событий АЭ более чем в 2 раза.

Список литературы

[1] Смирнов В.Б., Пономарев А.В., Бернар П., Патонин А.В. Закономерности переходных режимов сейсмического процесса по данным лабораторного и натурного моделирования // Физика земли. 2010. № 2. С. 17–49.

[2] Соболев Г.А. Основы прогноза землетрясений. М.: Наука, 1993. 313 с.

[3] Соболев Г.А., Пономарев А.В. Акустическая эмиссия и стадии подготовки разрушения в лабораторном эксперименте // Вулканология и сейсмология. 1999. № 4-5. С. 50-62.

[4] Шкуратник В.Л., Вознесенский А.С., Винников В.А. Термостимулированная акустическая эмиссия в геоматериалах. М.: изд. "Горная книга", 2015. 241 с.

[5] Heuze F.E. High-temperature mechanical, physical and thermal properties of granitic rocks – A review // International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts. 1983. V. 20, no. 1. Pp. 3-10.

[6] Gibbons S.J., Ringdal F. The detection of low magnitude seismic events using array-based waveform correlation // Geophysical Journal International. 2006. V. 165, no. 1. Pp.149–166.

[7] Иванов В.И., Барат В.А. Акустико-эмиссионная диагностика. М.: Спектр, 2017. 368 с.

[8] Septyana T., Nugraha A.D., Widiyantoro S., Supriyono. Multi-channel waveform clustering: a first look at microseismic multiplets from coalbed methane stimulation // Geosci. Lett. 2019. V. 6, no. 7. P. 1-11.

[9] Bousquet O., Elisseeff A. Stability and Generalization // Journ. Mach. Learn. Res. 2002. Vol. 2. P. 499-562.

[10] Luxburg U. Clustering Stability. An Overview // Foundations and Trends in Machine Learning. 2010. Vol. 2, No. 3. P. 235-274.

[11] Wyner A.J., Olson M., Bleich J., Mease D. Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers // Journ. Mach. Learn. Res. 2015. Vol. 18. P. 1-33.

[12] Казначеев П.А., Майбук З.-Ю.Я., Пономарев А.В., Смирнов В.Б., Бондаренко Н.Б. Исследование термоакустической эмиссии как возможный способ реконструкции истории pT-условий при формировании горных пород // Современная тектонофизика. Методы и результаты: Мат. пятой мол. тектонофиз. шк.-сем., Москва, 09–12 окт. 2017 года. М.: ИФЗ РАН, 2017. С. 240-246.

[13] Müllner D. Modern hierarchical, agglomerative clustering algorithms // arXiv. 2011. Art. ID 1109.2378. https://arxiv.org/abs/1109.2378