Математические структуры и моделирование 2024. № 3 (71). С. 94–105

УДК 004.896
DOI 10.24147/2222-8772.2024.3.94-105

ПРИМЕНЕНИЕ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ КОНТРОЛЯ КАЧЕСТВА ВИБРОДИАГНОСТИЧЕСКИХ ДАННЫХ

Т.Е. Болдовская
к.т.н., доцент, e-mail: boldovskaya73@gmail.com
И.В. Берсенев
студент, e-mail: ilyabersenev2002@mail.ru

Омский государственный технический университет, Омск, Россия

Аннотация. Статья нацелена на определение наиболее эффективной модели машинного обучения для кластеризации данных вибродиагностики. Исследование включает анализ различных моделей и методов, таких как k-means, Agglomerative Clustering, TimeSeriesKMeans и CatBoost. Цель состоит в выборе метода, способного наилучшим образом выявить структуру данных и улучшить понимание особенностей вибрационных сигналов. Результаты исследования могут быть полезны для разработки эффективных систем мониторинга и диагностики оборудования, а также для повышения надёжности и производительности технических систем.

Ключевые слова: временные ряды, кластеризация данных, k-means, Agglomerative Clustering, временные данные.

Введение

Контроль качества вибродиагностических данных играет важную роль в эффективном обслуживании и эксплуатации грузовых вагонов. Вибродиагностические данные используются для оценки состояния колёсных пар, выявления дефектов и определения их типа и степени. Однако, точность и достоверность этих оценок могут быть подвержены различным факторам, таким как шум, воздействие внешних условий и повреждение оборудования.

Помимо этого правильная оценка состояния колёсных пар является критическим фактором для безопасности и эффективности работы грузовых вагонов. Недостоверные оценки повреждений могут привести к необходимости неожиданных ремонтных работ или даже к аварийным ситуациям.

Исходный массив вибродиагностических данных для исследования был предоставлен Омским заводом транспортного машиностроения (группа компаний «Энергосервис»). В качестве инструментов для практической работы с методами машинного обучения были использованы: язык Python, компоненты библиотек scikit-learn, Keras, NumPy, pandas.

1. Описание задачи

В работе поставлена задача кластеризации и классификации дефектов подшипников колёсной пары грузовых вагонов: результаты диагностики колёсных пар и их статистические показатели (248 признаков, 8108 колёс) нужно отнести к одному из выявленных кластеров, содержащих один или более дефект.

Данная задача является примером задачи машинного обучения «без учителя». Модель обучается находить закономерности и структуры в данных без предварительной маркировки [1]. В качестве основных метрик будут использоваться силуэт (Silhouette score) и Davies-Bouldin Index.

Силуэтная оценка измеряет, насколько каждый объект в кластере похож на свой собственный кластер по сравнению с другими кластерами. Она принимает значения от -1 до 1. Значение ближе к 1 указывает на то, что объекты хорошо сгруппированы внутри своих кластеров и отделены от других кластеров. Значение ближе к -1 означает, что объекты могли бы быть лучше распределены в других кластерах.

Davies-Bouldin Index измеряет среднее расстояние между каждым кластером и его самым похожим кластером, делённое на сумму внутрикластерных дисперсий. Низкое значение этого индекса указывает на лучшее разделение кластеров.

1.1. Описание данных. На стендах данные снимаются с помощью датчиков вибрации, установленных на колёсных парах. Собранные данные представляют собой временные ряды, отражающие колебания и вибрации, возникающие в колёсных парах в процессе прокатки. Для более детального анализа и выявления характеристик повреждений в колёсных парах применяется метод быстрого преобразования Фурье. Полученные данные содержат информацию о каждом колесе колёсной пары. При диагностике на выходе имеются две таблицы со значениями графика широкополосного спектра и статистические показатели о нём. Примеры графиков для пары колёс представлены на рис. 1.

На графике указана зависимость частоты вращения от силы звукового давления широкополосных спектров левого и правого колеса колёсной пары. При диагностике на выходе имеются две таблицы с значениями графика широкополосного спектра и статистические показатели о нём.

Так как данные записываются в виде строк для каждой части колёсной пары отдельно и содержат полную информацию о диагностике, включая серийный номер и часть колёсной пары, то перед использованием этих данных в машинном обучении их необходимо предварительно обработать.

1.2. Предварительная обработка данных. Предобработка данных является важным этапом перед их использованием в задачах машинного обучения. В случае с данными, собранными с помощью датчиков вибрации на колёсных парах грузовых вагонов, необходимо выполнить несколько действий. Неинформативные данные, например серийный номер или другая идентифицирующая информация, были удалены из исследуемого набора.

Для обеспечения их согласованности, данные, заданные в различных масштабах и единицах измерения, первоначально нормализовались. Нормализация включала

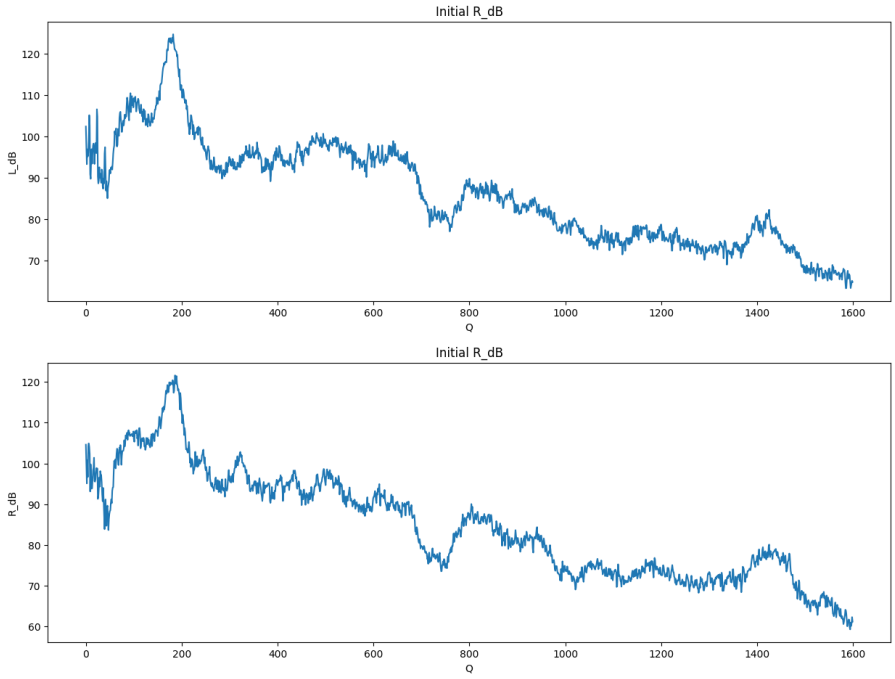

Рис. 1. Графики широкополосных спектров колёсной пары

масштабирование данных в определённом диапазоне (от 0 до 1) или их стандартизацию. Нормализация данных была проведена с помощью StandardScaler из библиотеки scikitlearn для масштабирования данных. StandardScaler используется для нормализации данных, приводя их к стандартному нормальному распределению. Этот метод применяет формулу для каждого признака:

$$z = (x - u)/s,$$

где z – новое значение признака после масштабирования, x – исходное значение признака, u – среднее значение признака в наборе данных, s – стандартное отклонение признака в наборе данных.

2. Ход работы

Статистические показатели были собраны по таблице значений графиков широкополосных спектров колёсной пары.

2.1. Статистика. Статистические показатели – ключевой инструмент анализа данных, предоставляющий информацию о характеристиках набора данных [2]. В работе по контролю качества вибродиагностических данных они используются для следующих целей:

1. Описательная статистика: минимум, максимум, среднее и медиана позволяют понять типичные значения и вариации в сигнале, выявить аномалии и выбросы.

2. Выявление особых структур: количество пиков и впадин помогает обнаружить характеристики, связанные с повреждениями или другими интересующими особенностями.

3. Измерение изменчивости: дисперсия и стандартное отклонение позволяют оценить уровень шума и стабильности сигнала.

4. Извлечение временных и амплитудных характеристик: сумма квадратов производной, показатель тренда и пересечения с определёнными метками предоставляют информацию о динамике и амплитуде сигнала.

5. Анализ спектральных характеристик: амплитудный спектр помогает обнаружить частотные особенности, связанные с интересующими характеристиками сигнала.

На рис. 2 представлена таблица статистических данных, собранных по данным рядам.

	Minimum	Maximum	Mean	Median	Standard Deviation	Peaks Count	Valleys Count	Vari
L_2	27.588340	51.626405	39.292452	39.316460	7.045081	411	411	49.66
R_2	27.628810	51.740060	39.395097	39.332460	6.995967	416	417	48.97
L_3	27.880530	52.453225	39.485587	39.426960	6.986542	425	426	48.84
R_3	28.011120	52.531610	39.574484	39.572995	6.936167	424	425	48.14
L_4	27.698975	51.562440	38.857676	39.099120	7.004503	429	430	49.09
...	...	...	...	...	...	...	...	
R_4052	59.990025	119.707300	85.400657	84.232835	13.811417	445	446	190.87
L_4053	57.290675	118.851000	78.122366	78.706885	15.427675	433	434	238.16
R_4053	65.537600	122.426650	86.326581	84.989275	12.508519	434	434	156.56
L_4054	57.053455	120.265400	81.045512	83.811695	16.513639	429	429	272.87
R_4054	61.394815	119.483050	84.559787	81.787510	13.075549	428	428	171.07

Рис. 2. Статистические показатели по данным рядам

Таблица дефектов показана на рис. 3.

	def-3/0_level	def-3/0_s	def-3/0_m	def-3/0_w	def-3/1_level	def-3/1_s	def-3/1_m	def-3/1_w	def-3/2_level
id									
L_1	0	0	0	0	0	0	0	0	0
R_1	0	0	0	0	0	0	0	0	0
L_2	0	0	0	0	0	0	0	0	0
R_2	0	0	0	0	0	0	0	0	0
L_3	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...
R_4052	0	0	0	0	0	0	0	0	0
L_4053	0	0	0	0	0	0	0	0	0
R_4053	0	0	0	0	0	0	0	0	0
L_4054	0	0	0	0	0	0	0	0	0
R_4054	0	0	0	0	0	0	0	0	0

Рис. 3. Дефекты в бинарном представлении, их уровень и степень

Одной из важных характеристик, полученных при диагностике, является принадлежность классов дефектов к каждому широкополосному спектру, их уровень и степень.

3. Применение моделей машинного обучения

Кластеризацию первично обработанных данных было решено провести в несколько этапов с разными моделями и методами, затем по оценке кластеризации выбрать наилучшее разбиение.

3.1. Первичная кластеризация. На начальном этапе была проведена кластеризация наших широкополосных спектров на основе их поведения. Для этого использовался алгоритм кластеризации TimeSeriesKMeans.

TimeSeriesKMeans – это алгоритм кластеризации временных рядов, который расширяет классический алгоритм k-means для работы с временными данными. Он предназначен для группировки временных рядов на основе их схожести в течение определённого отрезка времени [3].

Основная идея заключается в том, чтобы выявить группы рядов, которые имеют сходные формы или паттерны в течение определённого времени. Этот алгоритм учитывает не только значения временных рядов, но и их динамику и изменение во времени. Процесс работы TimeSeriesKMeans включает в себя начальную инициализацию центров кластеров, а затем итеративное перераспределение временных рядов между кластерами до тех пор, пока изменения стабилизируются или достигнут предела числа итераций [3].

Оптимальное число кластеров определялось с помощью процедуры «метод локтя» (Elbow Method), Инерции (Inertia) и оценки «силуэта» (Silhouette score).

График Distortion метрики Инерция (Inertia) отражает сумму квадратов расстояний между точками данных и центроидами их кластеров при разных значениях числа кластеров (k). При увеличении числа кластеров искажение обычно уменьшается, так как кластеры становятся более плотными. Однако на графике можно заметить точку, где снижение искажения замедляется, что создаёт «локоть» на графике. Эта точка указывает на оптимальное число кластеров для модели. Меньшее значение инерции указывает на более компактные кластеры и, следовательно, лучшее качество кластеризации.

График Silhouette отображает силуэтную оценку для различных чисел кластеров. Оценка силуэта показывает, насколько объекты в кластере схожи с другими объектами внутри того же кластера по сравнению с объектами из других кластеров, принимая значения от -1 до 1. Значение ближе к 1 указывает на хорошее группирование объектов внутри кластеров и хорошее разделение между кластерами. Значение ближе к -1 говорит о том, что объекты могли бы быть лучше распределены в других кластерах. Оптимальное число кластеров можно выбрать, основываясь на максимальной силуэтной оценке (рис. 4).

Анализ графиков показал плавное снижение графика при использовании «метода локтя», что не совсем информативно для исследования, и оптимальное число кластеров находится в пределах от 3 до 8. График силуэта выявил наибольшее значение оценки при 3 кластерах, но для нашего исследования это малое число разбиений, поэтому было принято решение остановиться на 7 кластерах, так как это следующий пик на графике.

На рис. 5 представлена полученная реализация разбиения на кластеры. Анализ

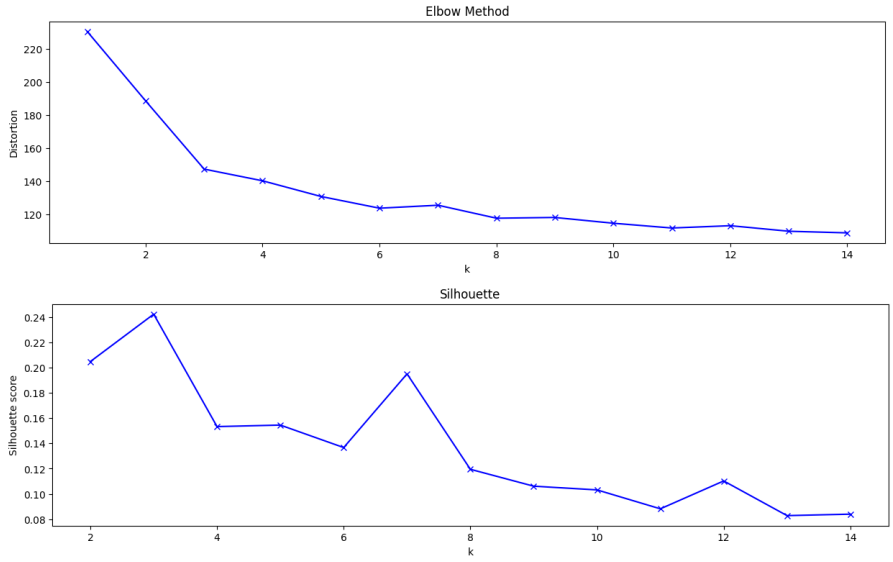

Рис. 4. Результаты оценки методами Elbow Method и Silhouette

полученной визуализации показал, что кластеры имеют различия и корректно описывают поведения спектров.

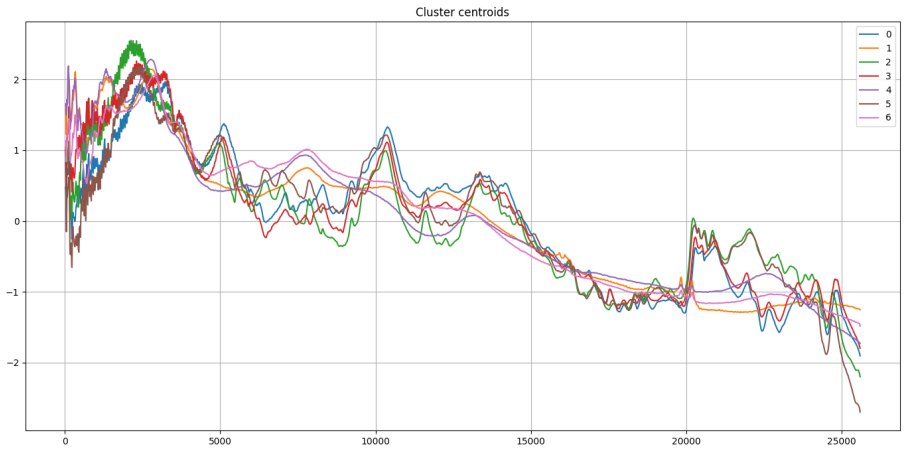

Рис. 5. Визуализация кластеров

Далее полученные данные были собраны в одну статистическую таблицу и для каждого графика был добавлен ID соответствующего кластера.

Для дальнейшей кластеризации по бинарным признакам было применено OneHot кодирование для кластеров. На рис. 6 показана часть таблицы после кодирования.

3.2. Вторичная кластеризация. Для вторичной кластеризации на основе статистических показателей были применены два основных метода кластеризации: k-means и агломеративная кластеризация.

k-means – это метод, разбивающий данные на заранее заданное количество кластеров, минимизируя сумму квадратов расстояний от каждой точки до центроида своего кластера [4].

Агломеративная кластеризация (Agglomerative Clustering) – это метод, который строит иерархию кластеров путём последовательного объединения ближайших кластеров [4].

Для оценки числа кластеров по методу k-means применялись методы силуэта и инерции на данных без применения масштабирования. Визуализация оценки кластеров представлена на рис. 7.

cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5	cluster_6	cluster_7	cluster_8	cluster_9
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1	0
...	...	...	...	...	...	...	...	...	...
0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0
0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0
0	1	0	0	0	0	0	0	0	0

Рис. 6. Метки кластеров после OneHot кодирования

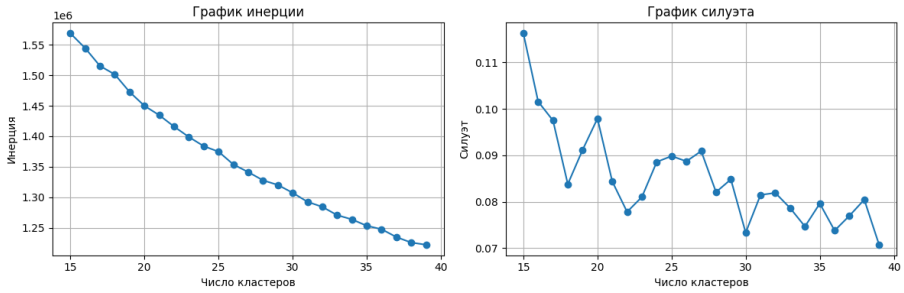

Рис. 7. Визуализация оценки кластеризации метриками инерции и силуэта с неотмасштабированными данными

Как видно по графику силуэта, оценки довольно близки к 0 и график нисходящий, что плохо сказывается на кластеризации. Визуализация оценки кластеров после проведения масштабирования представлена на рис. 8.

После масштабирования данных оценка силуэта значительно возросла – график стал восходящим, значения приблизились к 1. По графику силуэта виден пик на 22 кластерах.

В качестве параметров кластеризации указываем 22 кластера (рис. 9).

Для визуализации каждого полученного нами кластера создаётся список из ID

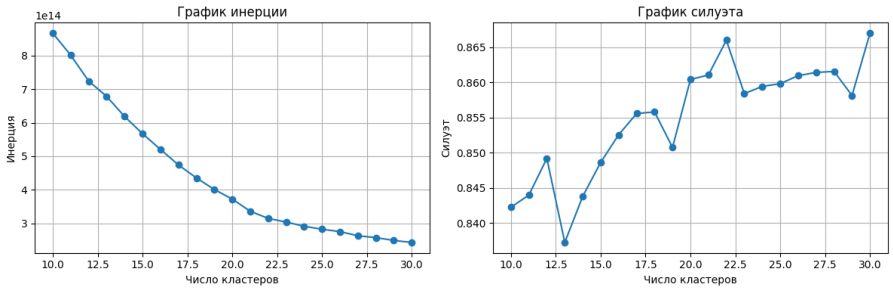

Рис. 8. Визуализация оценки кластеризации метриками инерции и силуэта с масштабированными данными

```
# Создаем объект KMeans с количеством кластеров
num_clusters = 22
kmeans = KMeans(n_clusters=num_clusters, random_state=42)

# Производим кластеризацию на основе масштабированных числовых данных
kmeans.fit(scaled_data)

# Добавляем метки кластеров в DataFrame
res_clastering['Second_cluster'] = kmeans.labels_
```

Рис. 9. Визуализация оценки кластеризации метриками инерции и силуэта с неотмасштабированными данными

колёс для каждого кластера и выбираются случайные N кластеров для отображения.

На рис. 10 представлена визуализация 4-го и 21-го кластеров моделью k-means.

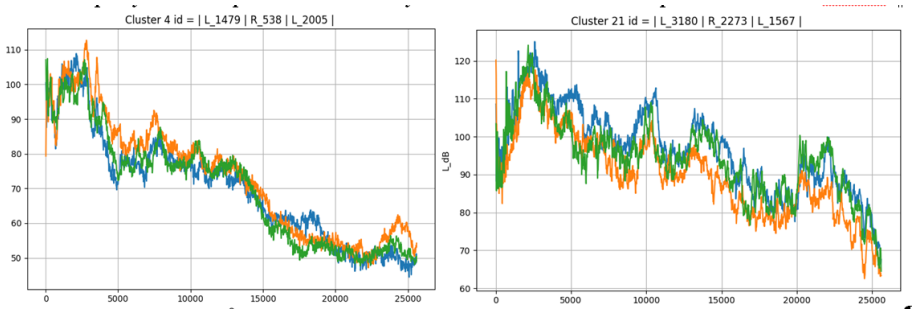

Рис. 10. Визуализация 4-го и 21-го кластеров моделью k-means

По графикам видно, что кластеризация по статистическим показателям довольно хорошо справилась со своей задачей.

Далее была проведена кластеризация методом Agglomerative Clustering. Оценка силуэта, представленная на рис. 11, в этом случае показала меньшее число кластеров, а именно 18, что несомненно оказывает влияние на качество кластеризации.

На рис. 12 представлена визуализация 7-го и 13-го кластеров моделью Agglomerative Clustering.

Рис. 11. Визуализация оценки кластеризации Agglomerative Clustering метрикой силуэта

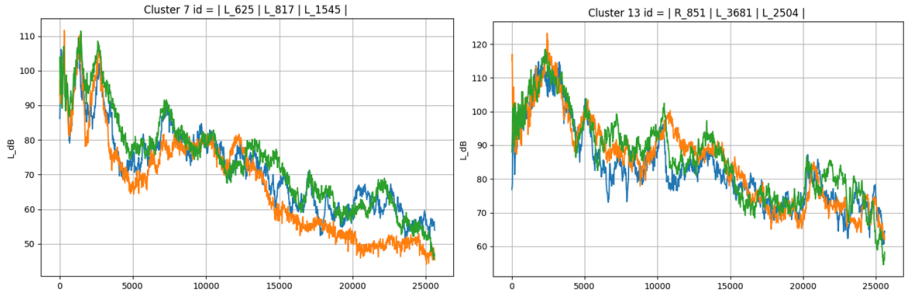

Рис. 12. Визуализация 7-го и 13-го кластеров моделью Agglomerative Clustering

Данная модель справилась хуже: оранжевый график на второй половине немного расходится с другими двумя и число кластеров получилось меньше, чем при использовании модели k-means.

3.3. Кластеризация по всем данным. Чтобы оценить, как хорошо справилась кластеризация по статистическим показателям, было решено провести кластеризацию по всем имеющимся данным, а именно объединить таблицы со статистикой и значениями графиков в одну.

Была применена модель k-means, так как она показала более лучшие результаты на прошлых итерациях. Для оценки была выбрана другая метрика – Davies-Bouldin Index.

Davies-Bouldin Index измеряет среднее расстояние между каждым кластером и его самым похожим кластером, делённое на сумму внутрикластерных дисперсий [5]. Низкое значение этого индекса указывает на лучшее разделение кластеров. Davies-Bouldin Index вычисляется по формуле:

$$DB = \frac{1}{k}\sum_{1}^{k}\max_{j\neq i}\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}.$$

Результаты оценки методом k-means метрикой Davies-Bouldin Index представлены на рис. 13.

```
Количество кластеров: 15, Davies-Bouldin Index: 2.0991405747194265
Количество кластеров: 16, Davies-Bouldin Index: 2.185810579352497
Количество кластеров: 17, Davies-Bouldin Index: 2.217394763817129
Количество кластеров: 18, Davies-Bouldin Index: 2.277211578958847
Количество кластеров: 19, Davies-Bouldin Index: 2.3574947003631306
Количество кластеров: 20, Davies-Bouldin Index: 2.3790367356201774
Количество кластеров: 21, Davies-Bouldin Index: 2.457389562830074
Количество кластеров: 22, Davies-Bouldin Index: 2.3620052082750793
Количество кластеров: 23, Davies-Bouldin Index: 2.4915251897120116
Количество кластеров: 24, Davies-Bouldin Index: 2.4960292905588766
Количество кластеров: 25, Davies-Bouldin Index: 2.465408883081157
Количество кластеров: 26, Davies-Bouldin Index: 2.5113393402354447
```

Рис. 13. Оценка кластеризации k-means метрикой Davies-Bouldin Index

При увеличении числа кластеров Davies-Bouldin Index также возрастает, что сказывается на ухудшении кластеризации. Но при количестве кластеров, равных 22, Davies-Bouldin Index уменьшается на 0.1 по сравнению с предыдущим значением. Это означает, что кластеризация на 22 кластерах имеет лучший результат, чем на 21 и 23.

На рис. 14 приведена визуализация 16-го кластера моделью k-means.

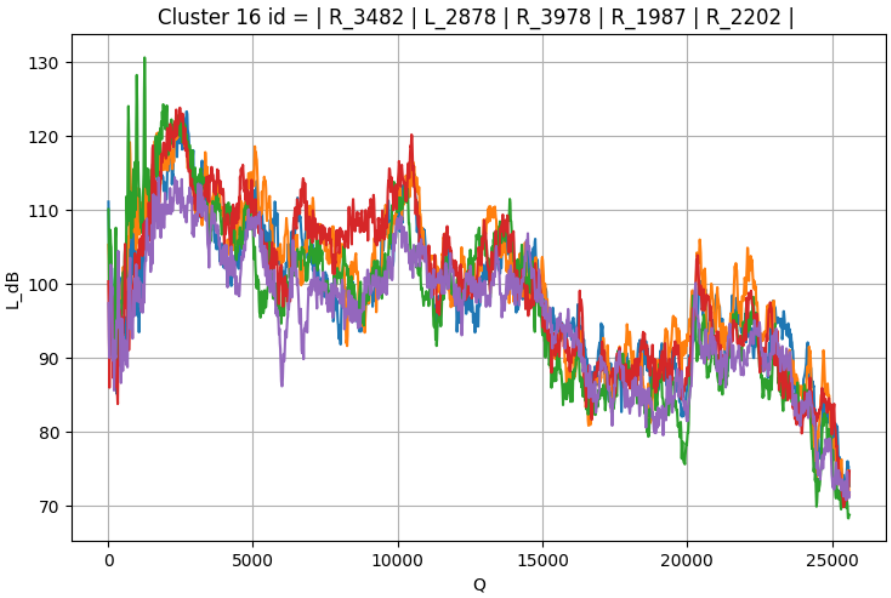

Рис. 14. Оценка кластеризации k-means. Визуализация 16-го кластера моделью k-means

Все графики близко расположены друг к другу для всех полученных кластеров, что свидетельствует о высокой точности кластеризации.

4. Классификация

Полученные значения кластеризации добавляются к таблице графиков. Эти значения будут целевыми для классификации, наряду с несколькими статистическими. В качестве модели было решено взять CatBoost.

CatBoost позволяет использовать категориальные признаки напрямую, без необходимости их предварительной обработки, что делает эту модель идеальным выбором для задач с небольшим объёмом данных. Кроме того, она обеспечивает повышенную точность за счёт снижения риска переобучения и поддерживает обучение на нескольких GPU [6].

Были проведены эксперименты с различными пропорциями в разделении наблюдений на тренировочную и тестовую выборки. Лучшим для данной задачи оказалось разделение 85 % на 15 %.

На рис. 15 показаны результаты оценки классификации.

```
Train score: 0.87631973145
Test score: 0.74913485317
```

Рис. 15. Оценка кластеризации k-means

Данная модель классификации показывает довольно высокий результат по точности: 87,63 % на тренировочных и 74,91 % на тестовых данных.

5. Заключение

Эта работа направлена на поиск оптимальной модели машинного обучения для решения задачи кластеризации вибродиагностических данных. В рамках исследования были рассмотрены различные модели и методы, включая k-means, Agglomerative Clustering, TimeSeriesKMeans и CatBoost.

Результаты оценки различными метриками кластеризации моделями машинного обучения приведены в табл. 1.

Таблица 1. Результаты оценки кластеризации

Модель	Inertia	Silhouette score	Davies-Bouldin Index
TimeSeriesKMeans (на знач. графиков спектра)	130	19 %	–
k-means (на статистич. данных)	3,25	86,7 %	2,1893
Agglomerative Clustering (на статистич. данных)	–	85,2 %	2,1624
k-means (На всех данных)	2,95	88,4 %	2,3620
Agglomerative Clustering (на всех данных)	–	83,6 %	2,3646

Модель k-means с задачей кластеризации справилась лучше, чем Agglomerative Clustering, как на статистических данных, так и на всех данных в совокупности (статистические параметры и значения графиков широкополосных спектров). Модель

TimeSeriesKMeans показала себя хуже остальных, но не маловажной для статистики.

В ходе работы было выяснено, что разбиение кластеризации на два этапа: по поведению графиков и по статистике – справилось не хуже кластеризации по всем имеющимся данным в целом, а также, что масштабирование данных играет важную роль в результатах кластеризации и является неотъемлемой её частью.

Литература

1. Введение в машинное обучение. URL: `https://habr.com/ru/post/448892/t/004793/Crypto-1.pdf` (дата обращения: 02.05.2024).
2. Перчев Н.Г. Статистические методы анализа данных: М.: Изд-во Моск. гос. ун-та, 2013. 215 с.
3. Анализ временных рядов, применение нейросетей. URL: `https://habr.com/ru/articles/693562/` (дата обращения: 03.05.2024).
4. Официальная документация Scikit-learn. URL: `URL:https://scikit-learn.org/stable/index.html` (дата обращения: 03.05.2024).
5. Николенко С.И., Кудрин А.А., Архангельская Е.О. Глубокое обучение. СПб.: Питер, 2019. 481 с.
6. Официальная документация CatBoost. URL: `https://catboost.ai/en/docs` (дата обращения: 03.05.2024).

APPLICATION OF MACHINE LEARNING MODELS FOR QUALITY CONTROL OF VIBRODIAGNOSTIC DATA

T.E. Boldovskaya
Ph.D. (Techn.), Associate Professor, e-mail: boldovskaya73@gmail.com
I.V. Bersenev
student, e-mail: ilyabersenev2002@mail.ru

Omsk State Technical University, Omsk, Russia

Abstract. Article aims to determine the most effective machine learning model for clustering vibration diagnostics data. The research includes analysis of various models and methods such as k-means, Agglomerative Clustering, TimeSeriesKMeans and CatBoost. The goal is to select a method that can best identify the data structure and improve understanding of the characteristics of vibration signals. The results of the study can be useful for the development of effective monitoring and diagnostic systems for equipment, as well as for improving the reliability and performance of technical systems.

Keywords: times series, data clustering, k-means, Agglomerative Clustering, time data.

Дата поступления в редакцию: 31.05.2024