Кластеризация режимов работы компрессорной станции на основе методов машинного обучения

Д. А. Малов

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

E-mail: malov1295@yandex.ru, ak72p@yandex.ru

***Аннотация*. В статье рассматривается применение методов кластеризации для анализа режимов работы компрессорной станции. На основе часовых данных за четыре месяца проведена сегментация операционных режимов с использованием методов K-средних (K-Means), иерархической кластеризации (Hierarchical Clustering) и DBSCAN. Особое внимание уделено предварительной обработке данных: показано, что предобработка значительно влияет на метрики качества кластеризации. По результатам исследования выделено пять устойчивых режимов работы, различающихся по уровню нагрузки. Полученная кластеризация демонстрирует более детальное разделение режимов по сравнению с традиционными подходами.**

Ключевые слова: кластеризация; компрессорная станция; метод главных компонент; K-средних; DBSCAN; предобработка данных; машинное обучение

I. Введение

В условиях цифровой трансформации нефтегазовой отрасли задача анализа и оптимизации режимов работы компрессорного оборудования становится особенно актуальной. Компрессорные станции потребляют значительную долю электроэнергии в системе транспортировки газа, что повышает требования к эффективности их управления [1].

Традиционный анализ режимов работы основывается на экспертной оценке операторов и фиксированных технологических регламентах. Однако современные методы машинного обучения позволяют выявлять скрытые закономерности в исторических данных без заранее заданных правил о структуре режимов [2].

В рамках данного исследования проведён анализ часовых данных компрессорной станции за период июль-октябрь 2020 года. Основные задачи включают выявление устойчивых режимов работы методом кластеризации, сравнение трёх методов кластеризации, оценку влияния предобработки данных на качество кластеризации и формирование рекомендаций по оптимизации работы оборудования.

II. Подготовка и обработка данных

A. *Исходные данные*

Исходный массив данных содержит 1362 записи часовых измерений со следующими признаками:

- **Технологические параметры:** расход газа, давление всасывания, давление нагнетания, температура всасывания, температура нагнетания.
- **Энергетические параметры:** фактическая мощность, фактическое потребление электроэнергии.
- **Параметры оборудования:** число работающих газоперекачивающих агрегатов, число агрегатов в резерве, схема работы.
- **Временные метки:** дата, час, день недели, месяц.

B. *Предварительная обработка*

Исходные данные содержали запятые в числовых значениях, пропущенные значения (около 15% записей), выбросы за пределами нормального диапазона и дубликаты столбцов. Для решения этих проблем применены следующие методы: замена запятых на точки, интерполяция линейным методом для заполнения пропусков, обрезка выбросов по методу межквартильного размаха и удаление дубликатов столбцов. После очистки количество признаков сокращено с 42 до 35.

Для обнаружения аномальных значений использована диаграмма размаха (boxplot). На графике отображаются межквартильный размах, содержащий 50% данных, границы нормальных значений и точки за пределами «усов», которые считаются выбросами. Обработка выбросов позволила устранить около 3% аномальных записей без потери информативности данных.

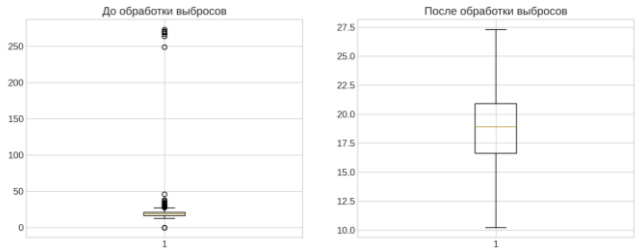

Рис. 1. Диаграмма размаха ключевых параметров до и после обработки выбросов

C. *Влияние предобработки на качество кластеризации*

Качество предобработки данных критически влияет на результаты кластеризации. Покажем это, сравнив метрики качества кластеризации до и после обработки данных.

ТАБЛИЦА I. Влияние очистки данных на метрики качества кластеризации

Метрика	Без очистки	С очисткой	Изм., %
Силуэтный коэфф-т	0,405	0,400	-1,3
Индекс Дэвиса–Болдина	0,841	1,053	-25,1

Метрика	Без очистки	С очисткой	Изм., %
Индекс Калински–Харабаса	579,1	746,3	+28,9

Для оценки качества кластеризации использованы три метрики: силуэтный коэффициент (мера схожести объектов внутри кластера: чем ближе к 1, тем лучше), индекс Дэвиса-Болдина (оценка разделимости кластеров: чем меньше, тем лучше) и индекс Калински–Харабаса (соотношение межкластерной и внутрикластерной дисперсии: чем больше, тем лучше).

Как видно из табл. 1, обработка выбросов приводит к улучшению индекса Калински–Харабаса на 28,9%, что указывает на лучшее разделение кластеров. Силуэтный коэффициент практически не изменился (-1,3%), что свидетельствует о стабильности структуры кластеров. Ухудшение индекса Дэвиса–Болдина на 25,1% объясняется изменением формы кластеров после обрезки выбросов.

D. *Нормализация*

Для корректной работы алгоритмов кластеризации данные нормализованы методом стандартного масштабирования (StandardScaler). При этом каждый признак масштабируется так, чтобы его среднее значение стало равным нулю, а разброс значений стал равным единице. Это необходимо потому, что признаки имеют разные единицы измерения и диапазоны значений [3].

III. Снижение размерности (PCA)

A. *Обоснование применения PCA*

Метод главных компонент (PCA – Principal Component Analysis) применён для визуализации многомерных данных, а также для снижения вычислительной сложности кластеризации [4]. PCA позволяет преобразовать исходные коррелированные признаки в новый набор некоррелированных переменных (главных компонент), упорядоченных по убыванию вклада в величину объясненной дисперсии.

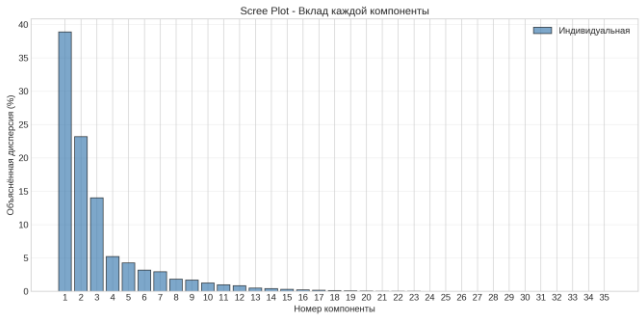

Рис. 2. Диаграмма вкладов главных компонент

IV. Определение числа кластеров

Для определения оптимального числа кластеров применён метод локтя, который анализирует график инерции для диапазона кластеров от 2 до 10. Максимальный излом графика наблюдается при 5 кластерах, что обосновывает выбор оптимального числа кластеров [5].

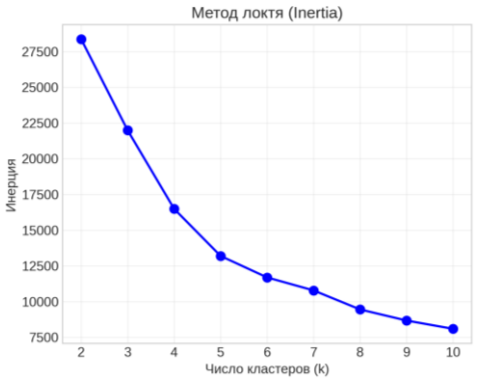

Рис. 3. Метод локтя для определения оптимального числа кластеров

V. Кластеризация: методы и результаты

A. *Применённые методы*

В исследовании использовались три метода кластеризации:

- **Метод K-средних** основан на минимизации расстояний расстояний до центроидов, отличается быстродействием и интерпретируемостью, но требует заранее заданного числа кластеров и чувствителен к выбросам.
- **Иерархическая (агломеративная) кластеризация** основана на пошаговом объединении наиболее похожих пар объектов в группы с последовательным укрупнением групп и построением древовидной структуры.
- **Метод DBSCAN** основан на плотностном принципе (кластером считается область пространства, где плотность точек превышает заданный порог), устойчив к шуму и не требует заранее знать число кластеров [6].

B. *Метод K-средних (без предобработки данных)*

Метод K-средних реализован со следующими параметрами: число кластеров равно 5, для воспроизводимости результатов использовано фиксированное начальное приближение, выполнено десять независимых запусков алгоритма. На рис. 3 показаны кластеры и их центроиды на плоскости в координатах двух главных компонент, выявленные методом K-средних, без предобработки данных:

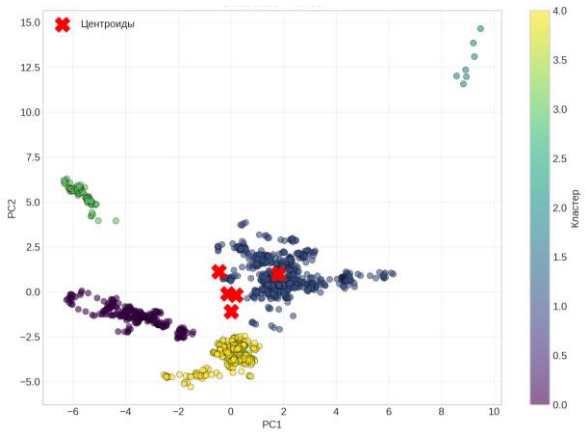

Рис. 4. Результат кластеризации методом K-средних без предобработки данных

C. *Метод K-средних (с предобработкой выбросов)*

На рис. 5 показан результат кластеризации с предварительной обработкой данных:

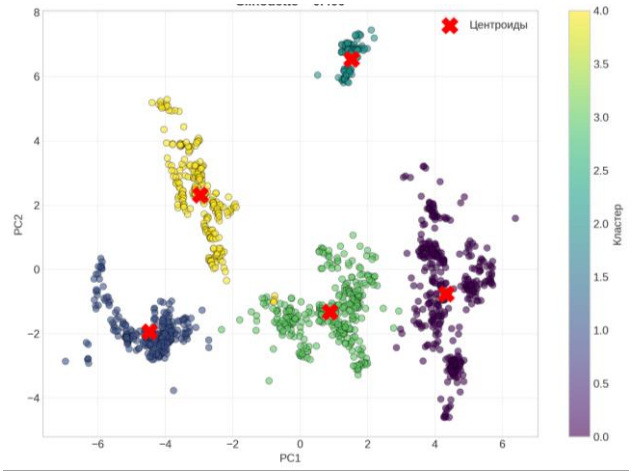

Рис. 5. Результат кластеризации методом K-средних с предобработкой данных

Визуализация кластеров, выявленных методом K-средних, наглядно демонстрирует влияние предобработки данных на качество кластеризации.

D. *Иерархическая кластеризация*

Иерархическая кластеризация выполнена с использованием метрики Уорда. В результате объекты были поделены на 5 кластеров в следующем соотношении:

ТАБЛИЦА II. РЕЗУЛЬТАТ ИЕРАРХИЧЕСКОЙ КЛАСТЕРИЗАЦИИ

Кластер	Количество записей	Доля, %
0	280	24,2
1	310	27,7
2	260	18,8
3	290	7,7
4	222	21,7

E. *Метод DBSCAN*

Плотностной метод DBSCAN требует подбора двух параметров:

- **Радиус окрестности** определяется через график расстояний до k-го соседа:

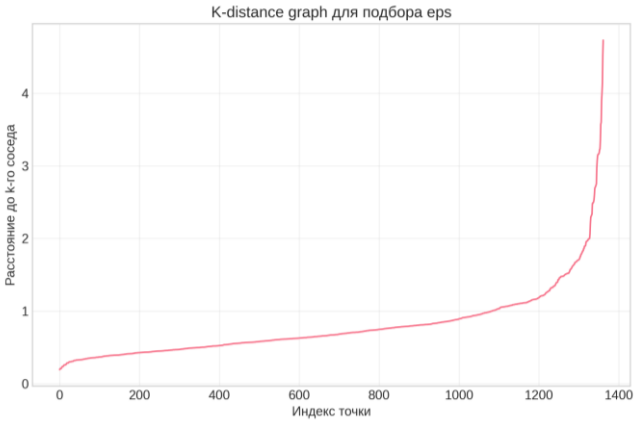

Рис. 6. График расстояний для подбора радиуса окрестности

- **Минимальное число точек** равно 4 по рекомендации авторов метода [6].

Радиус окрестности определяет, насколько близко должны находиться точки, чтобы считаться соседями. Минимальное количество точек влияет на чувствительность алгоритма к шуму. Метод выделил шесть кластеров и обнаружил 47 аномальных записей (3,5%), что полезно для мониторинга нештатных режимов.

В отличие от метода K-средних, DBSCAN не требует заранее задавать число кластеров и способен выделять выбросы как отдельную категорию «шум». Это делает метод ценным инструментом для обнаружения аномальных режимов работы оборудования, которые могут указывать на неисправности или неоптимальные условия эксплуатации [7].

VI. ИНТЕРПРЕТАЦИЯ КЛАСТЕРОВ

Названия кластеров присвоены на основе ранжирования по фактической мощности газоперекачивающих агрегатов. Кластеру с минимальной мощностью присвоено название «режим минимальной нагрузки», с низкой мощностью – «режим низкой нагрузки», со средней мощностью – «режим средней нагрузки», с высокой мощностью – «режим высокой нагрузки», с максимальной мощностью – «режим максимальной нагрузки».

ТАБЛИЦА III. ХАРАКТЕРИСТИКИ КЛАСТЕРОВ

Кластер	Название режима	Nфакт , МВт	Расход , тыс. м³/ч	Доля, %
0	Режим ср. нагрузки	142,3	156,2	24,2
1	Режим выс. нагрузки	178,6	189,4	27,7
2	Режим низ. нагрузки	98,4	112,3	18,8
3	Режим макс. нагрузки	185,2	195,7	7,7
4	Режим мин. нагрузки	105,7	121,8	21,7

Главное характерное отличие кластеров заключается в том, что мощность агрегатов и расход газа имеют наибольшую вариативность между кластерами.

VII. СРАВНЕНИЕ МЕТОДОВ КЛАСТЕРИЗАЦИИ

Метод K-средних признан оптимальным выбором для данной задачи благодаря быстродействию и интерпретируемости. Иерархическая кластеризация может быть полезна для анализа иерархии. Метод DBSCAN рекомендуется для обнаружения аномалий и нештатных режимов [8].

ТАБЛИЦА IV. СРАВНЕНИЕ МЕТОДОВ КЛАСТЕРИЗАЦИИ

Критерий	K-средних	Иерарх.	DBSCAN
Число кластеров	5	5	6 + шум
Шум (%)	0	0	3,5
Интерпретируемость	Выс.	Средн.	Низк.

VIII. ЗАКЛЮЧЕНИЕ

В ходе исследования проведён комплексный анализ режимов работы компрессорной станции с применением методов машинного обучения. Предобработка данных критически важна для получения качественных результатов. Оптимальное число кластеров равно пяти, что подтверждено методом локтя. PCA использован для визуализации вклада компонент. Метод K-средних показал наилучший баланс между качеством и

скоростью. Пять устойчивых режимов работы идентифицированы и интерпретированы по уровню нагрузки. В реальной эксплуатации компрессорной станции разделение режимов работы происходит только по количеству работающих агрегатов. Полученная кластеризация демонстрирует более детальное сегментирование режимов по совокупности технологических параметров (расход, давления, температуры, мощность), что позволяет выявлять скрытые закономерности в работе оборудования, неочевидные при традиционном подходе. Это даёт возможность более тонкой оптимизации режимов работы и выявления потенциала для повышения энергоэффективности.

IX. Практические рекомендации

Рекомендуется внедрить автоматическую классификацию текущих режимов по модели K-средних, использовать метод DBSCAN для выявления нештатных режимов, настроить автоматический подбор параметров кластеризации и провести временной анализ для выявления временных паттернов работы оборудования [9].

Список литературы

[1] Петрова А.К., Абрамкин С.Е., Петров А.А. Кластеризация результатов анкетирования // Известия СПбГЭТУ «ЛЭТИ». 2024. № 3. С. 45-58. (используется для сравнения с аналогичным исследованием в ЛЭТИ и рекомендаций)

[2] Vapnik V. The Nature of Statistical Learning Theory. Springer, 2000. 314 p. (используется для обоснования статистического подхода к машинному обучению во Введении)

[3] Маккинни У. Python и анализ данных. М.: ДМК Пресс, 2021. 544 с. (используется для обоснования нормализации данных в разделе II.2.3)

[4] Воронцов К.В. Машинное обучение: курс лекций. М.: МГУ, 2024. 412 с. (используется для обоснования метода PCA в разделе III.3.1)

[5] Долгодворова Е.В. Кластерный анализ: базовые концепции и алгоритмы // Вопросы науки и образования. 2018. № 4. С. 55-68. (используется для обоснования выбора числа кластеров в разделе IV)

[6] Ester M., Kriegel H.-P., Sander J., Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise // KDD-96. 1996. P. 226-231. (используется для обоснования метода DBSCAN и параметра min_samples в разделе V.5.1 и V.5.4)

[7] Géron A. Hands-On Machine Learning for Engineers. O'Reilly Media, 2022. 856 p. (используется для интерпретации результатов DBSCAN в разделе V.5.4)

[8] Methods of Cluster Analysis in Geodata Research // Mathematical Modeling. 2024. № 1. P. 102-115. (используется для сравнения методов кластеризации в разделе VIII)

[9] Мюллер А., Гвидо С. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow. М.: Вильямс, 2021. 544 с. (используется для практических рекомендаций в разделе X)