УДК 004.93`14

МЕТОДЫ ОЦЕНКИ КАЧЕСТВА КЛАСТЕРИЗАЦИИ ДЛЯ НЕПРЕРЫВНЫХ СИГНАЛОВ

Тюрикова В. А., студент гр. ИВБ1-21, IV курс
Научный руководитель: Гусаров А. В., канд. техн. наук, доцент
Рыбинский государственный авиационный технический университет имени П. А. Соловьева, г. Рыбинск

Кластеризация является одной из центральных задач анализа данных и машинного обучения. Она используется для разбиения совокупности объектов на группы (кластеры) таким образом, чтобы объекты внутри одного кластера были максимально схожи, а объекты из разных кластеров – максимально различны. Если данные выборки представить, как точки в признаковом пространстве, то задача кластеризации сводится к определению "сгущений точек" [1].

В данной работе рассматривается задача кластеризации непрерывного сигнала, разбитого на 385 фрагментов фиксированной длины. Такие фрагменты можно интерпретировать как отдельные объекты для кластеризации. На рисунках 1 и 2 представлены 100-й и 200-й фрагменты непрерывного сигнала.

Основной целью является оценка качества кластеризации с применением различных методов, включая метрики и визуализацию.

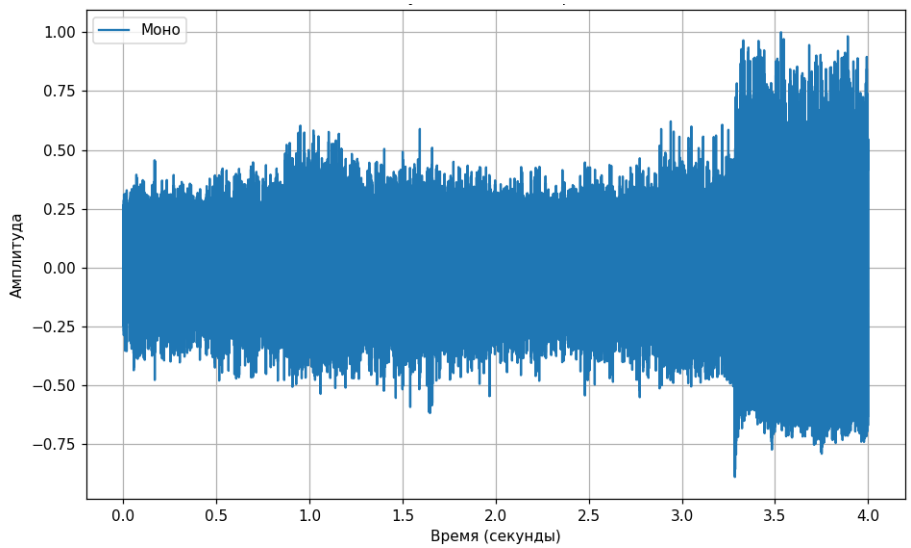

Рисунок 1 — Визуализация фрагмента 100

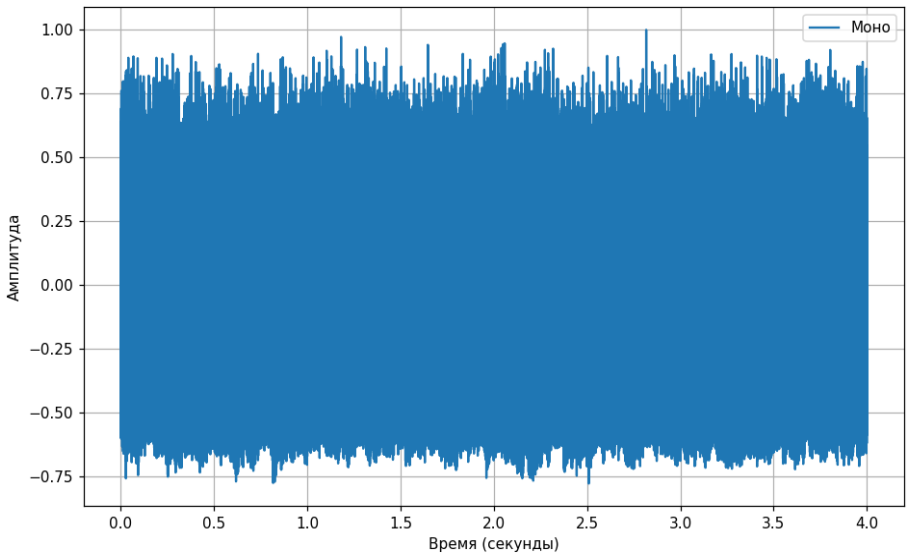

Рисунок 2 — Визуализация фрагмента 200

Основной задачей кластеризации является выявление скрытых закономерностей в данных без использования заранее заданных меток классов.

Существует множество различных методов кластеризации, каждый из которых имеет свои особенности и применимость в зависимости от типа данных и целей исследования. Важным аспектом является оценка качества полученных кластеров, что позволяет определить, насколько успешно решена поставленная задача и насколько адекватно представлена структура данных.

Для дальнейшей обработки необходимо выделить признаки для каждого получившегося объекта. Одним из самых распространенных методов анализа сигналов в частотной области является преобразование Фурье. На рисунках 3 и 4 представлены амплитудные спектры для 100-го и 200-го фрагментов непрерывного сигнала. Мы можем предположить, что объекты скорее всего имеют разную природу, и их можно будет отнести к разным классам.

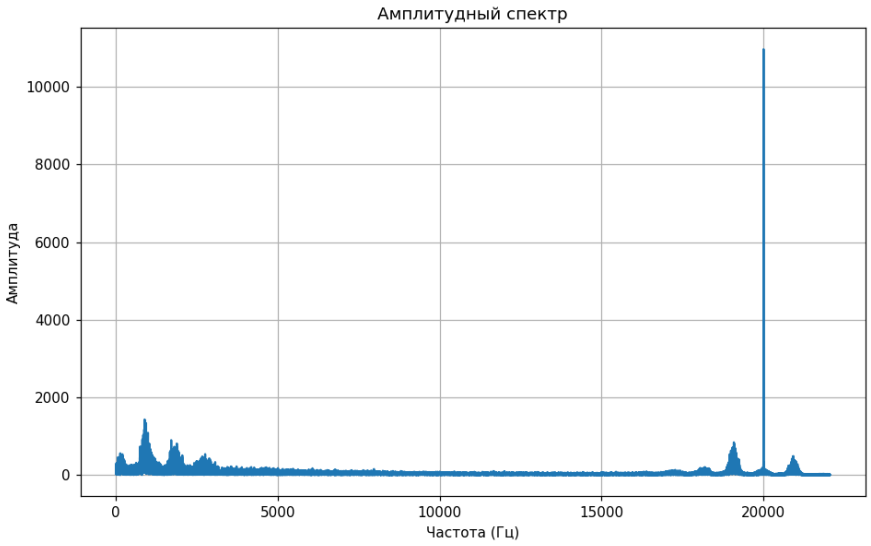

Рисунок 3 — Амплитудный спектр фрагмента 100

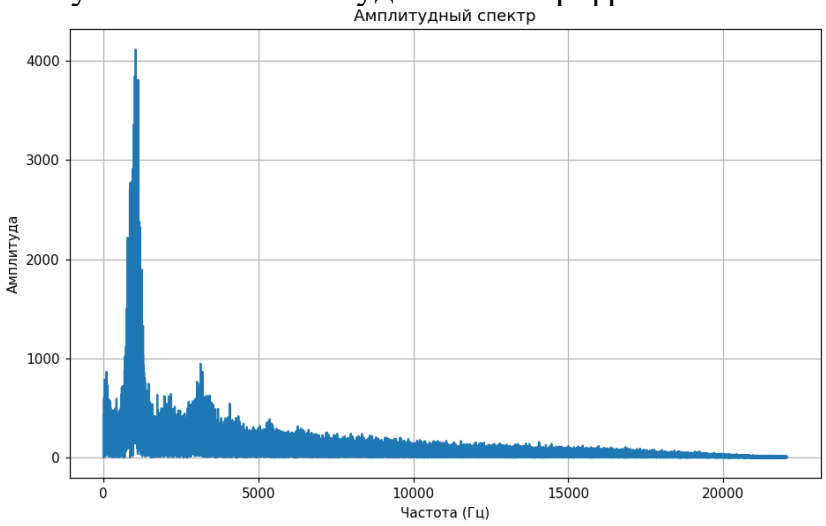

Рисунок 4 — Амплитудный спектр фрагмента 200

Полученные значения амплитуд для каждого фрагмента записываются в отдельный файл. В результате для каждого объекта формируется свой вектор признаков, содержащий значения амплитуд на разных частотах, что позволяет использовать их в алгоритмах кластеризации.

На основе полученных данных можно выполнить кластеризацию. Одним из эффективных методов кластеризации является метод K-Means.

В основе работы K-Means лежит принцип минимизации расстояния между объектами внутри одного кластера.

Алгоритм случайным образом выбирает k начальных точек, называемых центроидами. Эти точки служат временными центрами кластеров. Каждый объект в наборе данных назначается к кластеру, центроид которого находится ближе всего. Для расчета расстояния используется Евклидова метрика. После назначения объектов кластерам вычисляют новые центроиды, и каждый центроид перемещается в среднюю точку всех объектов, принадлежащих его кластеру. Данное действие повторяется до тех пор, пока центроиды не перестанут значительно изменяться, что свидетельствует о достижении сходимости [4].

В результате работы алгоритма каждый объект принадлежит конкретному кластеру.

Полученные кластеры можно визуализировать, но так как количество признаков большое, могут возникнуть сложности. Целесообразно использовать методы снижения размерности данных.

Цель состоит в том, чтобы упростить набор данных, сохранив при этом как можно больше релевантной информации. Это особенно полезно при работе с многомерными данными, когда количество признаков велико по сравнению с количеством выборок [2].

T-SNE – это метод снижения размерности и визуализации данных, который позволяет сохранить локальные структуры данных и обнаруживать нелинейные зависимости.

Основная идея заключается в том, чтобы преобразовать исходные данные таким образом, чтобы схожие объекты в исходном пространстве сохраняли свою схожесть и в новом, сниженном пространстве [3].

Для каждой точки вычисляются вероятности того, что она является соседом другой точки в исходном (высокоразмерном) пространстве. Далее алгоритм строит распределение вероятностей для целевого (низкоразмерного) пространства, таким образом, чтобы объекты, которые близки в исходном пространстве, имели высокие вероятности быть близкими и в сниженном пространстве. Процесс T-SNE заключается в минимизации расхождения Кульбака-Лейблера между двумя распределениями вероятностей: распределением вероятностей, которое вычисляется для исходных данных, и распределением вероятностей, которое строится для целевого пространства. Минимизация расхождения позволяет получить проекцию данных в сниженном пространстве таким образом, чтобы схожие объекты оставались близкими, а различные объекты – располагались на некотором расстоянии друг от друга.

На рисунке 5 представлен результат работы алгоритма для визуализации полученных кластеров. После применения алгоритма количество признаков сокращается до двух. Эти признаки являются осями на графике.

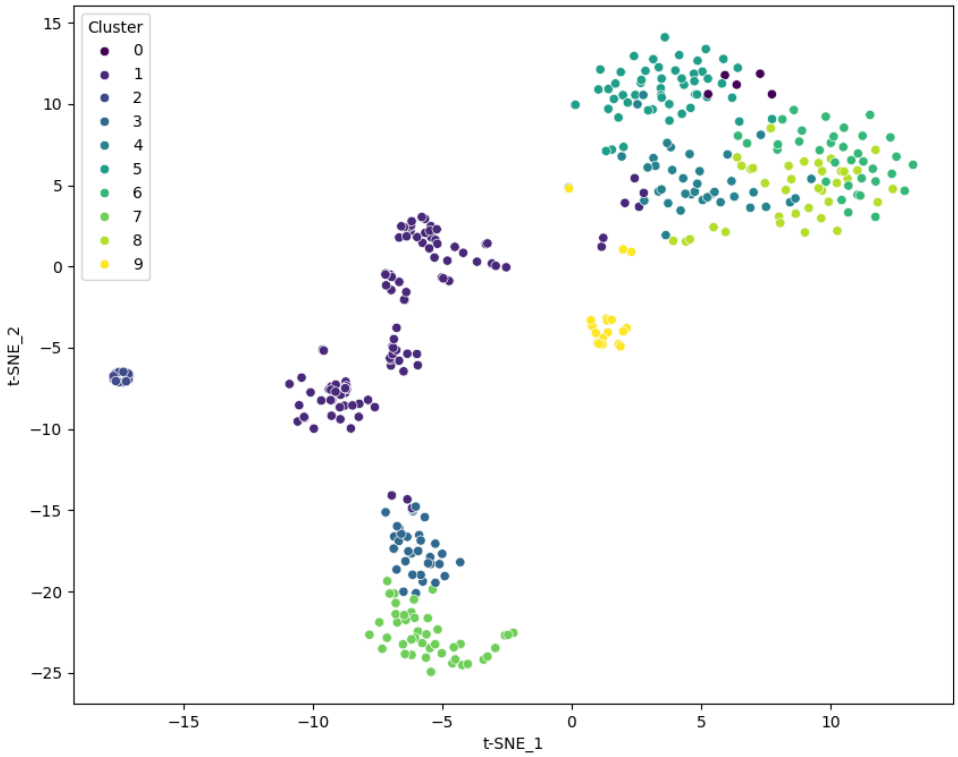

Рисунок 5 – Визуализация кластеров

Так как объекты формируются путем разбиения непрерывного сигнала, мы можем считать, что могут формироваться последовательности объектов, которые относятся к одному кластеру. На основе этого мы можем оценить качество кластеризации.

На рисунке 6 представлен график, отображающий соответствие объектов разным кластерам. Номера объектов упорядочены по времени.

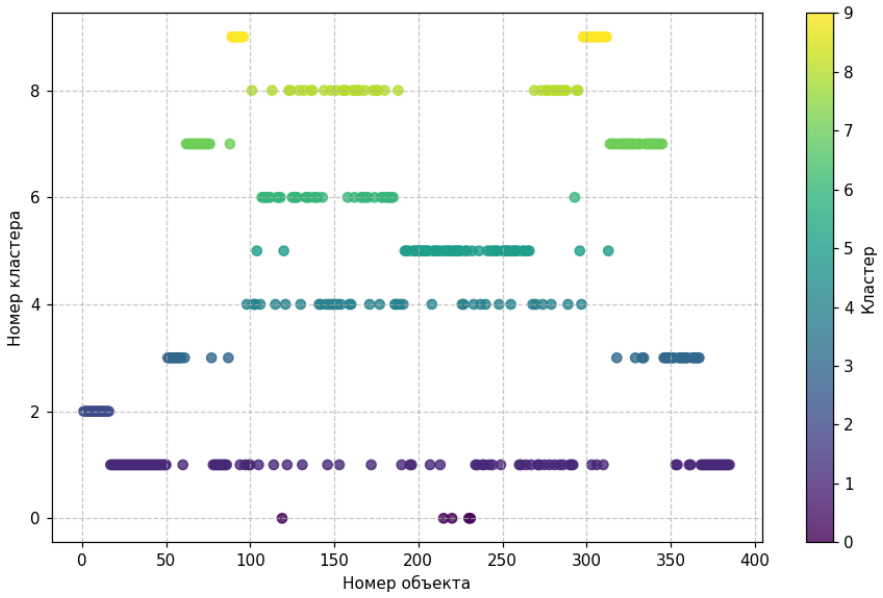

Рисунок 6 – Распределение объектов по кластерам

В сигнале есть постоянная составляющая, которая сохраняется на всём его протяжении и присутствует в нескольких кластерах. На графике видно, как с течением времени последовательно сменяются кластеры, что свидетельствует о корректно выполненной кластеризации.

Список литературы:

1. Интуит. Лекция 5: Задачи Data Mining. Классификация и кластеризация [Электронный ресурс]. – URL: https://intuit.ru/studies/courses/6/6/lecture/166?page=4 (дата обращения: 17.03.2025).

2. Analytics Vidhya. Beginners Guide To Learn Dimension Reduction Techniques [Электронный ресурс]. – URL: https://www.analyticsvidhya.com/blog/2015/07/dimension-reduction-methods/ (дата обращения: 17.03.2025).

3. Хабр. Применение преобразований PCA и t-SNE для снижения размерности данных [Электронный ресурс] / Отус. – URL: https://habr.com/ru/companies/otus/articles/757030/ (дата обращения: 17.03.2025).

4. Хабр. Алгоритм k-means и метод локтя: кластеризация данных с примерами на Python [Электронный ресурс] / SkillFactory. – URL:

https://habr.com/ru/companies/skillfactory/articles/877684/ (дата обращения: 17.03.2025).