Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования

Ульяновский государственный Технический университет

Кафедра «Вычислительная техника»

Дисциплина «Системы искусственного интеллекта»

**Лабораторная работа №5**

**«Алгоритмы кластеризации данных»**

Выполнила:

студентка группы ИВТАСбд-31

Архипова Е.Ю

Проверил работу:

Хайруллин И. Д.

Ульяновск 2025

**Общее задание**

1. Произвести масштабирование признаков (scaling).
2. С использованием библиотеки scikit-learn написать программу с использованием алгоритмов кластеризации данных, позволяющую разделить исходную выборку на классы, соответствующие предложенной вариантом задаче (http://scikit-learn.org/stable/modules/clustering.html).
3. Провести эксперименты и определить наилучший алгоритм кластеризации, параметры алгоритма. Необходимо использовать не менее 3-х алгоритмов. Данные экспериментов необходимо представить в отчете (графики, ход проведения эксперимента, выводы).

**Теоретические данные**

**Кластеризация** — это процесс группировки объектов на основе сходства между ними. В контексте машинного обучения и статистического анализа, кластеризация используется для выявления внутренней структуры в наборе данных, где объекты группируются вместе на основе их подобия друг к другу. Главная цель кластеризации — выявить группы объектов, которые похожи друг на друга больше, чем на объекты из других групп.

Основные принципы кластеризации:

**Неявная структура данных:** кластеризация позволяет исследовать структуру данных без предварительного знания о её природе. Это делает её полезной для открытия новых закономерностей и паттернов в данных.

**Группировка объектов:** в результате кластеризации данные разделяются на группы (или кластеры), где каждый кластер содержит объекты, которые близки друг к другу по некоторому критерию сходства.

**Безопасность кластеров:** Кластеры формируются таким образом, чтобы максимизировать внутрикластерное сходство и минимизировать межкластерное сходство. Это означает, что объекты внутри одного кластера будут более похожи друг на друга, чем на объекты из других кластеров.

**Различные методы кластеризации:** существует множество методов кластеризации, включая K-means, DBSCAN, Spectral Clustering, Hierarchical Clustering и многие другие. Каждый метод имеет свои особенности и подходит для разных типов данных и задач.

**KMeans** — это метод непараметрической кластеризации, который делит набор данных на K кластеров таким образом, что каждый объект в кластере наиболее близок к центру этого кластера. Центры кластеров инициализируются случайным образом, и затем в процессе обучения объекты присваиваются кластерам на основе расстояния до центров. Цикл обучения повторяется до достижения критерия остановки, например, максимального числа итераций или стабилизации центров кластеров. KMeans эффективен для больших наборов данных и позволяет легко интерпретировать результаты, так как каждый кластер имеет свой центр.

**Agglomerative Clustering (агломеративная иерархическая кластеризация) —** это метод иерархической кластеризации, который начинается с того, что каждая точка данных рассматривается как отдельный кластер. Затем итеративно происходит последовательное объединение ближайших кластеров до тех пор, пока не останется заданное число кластеров или пока не будет достигнуто другое условие остановки.

**Gaussian Mixture Model (GMM) —** это вероятностная модель, которая описывает данные как смесь нескольких нормальных (гауссовских) распределений с разными параметрами (средними и ковариациями). В контексте кластеризации GMM рассматривает, что каждая точка данных принадлежит к одному из нескольких кластеров с некоторой вероятностью, и пытается определить параметры этих гауссианов и вероятности принадлежности с помощью алгоритма Expectation-Maximization (EM).

**Adjusted Rand Index (ARI)** — это скор, который измеряет сходство между двумя разбиениями данных на кластеры: полученным алгоритмом и истинной разметкой. ARI основан на подсчёте пар объектов, которые либо оказались вместе в одном кластере, либо в разных кластерах в обоих разбиениях.

Важный момент — ARI **корректируется на случайное совпадение**. Это значит, что если кластеры были бы построены случайным образом, ARI будет близок к 0.

Значение ARI лежит в диапазоне от −1 до 1:

1 — полное совпадение кластеров с истинными классами,

0 — совпадение на уровне случайного разбиения,

отрицательные значения означают, что совпадение хуже случайного.

ARI хорошо подходит, когда надо учитывать, насколько парные связи объектов между кластерами совпадают с эталоном, и при этом учитывает случайность.

**Adjusted Mutual Information (AMI)** — это мера взаимной информации между двумя разбиениями, скорректированная с учётом случайных совпадений. Взаимная информация показывает, сколько информации (в битах или энтропии) одно разбиение даёт о другом.

AMI учитывает взаимную зависимость кластеров и истинных классов, но, как и ARI, корректирует значение с учётом того, что случайные разбиения могут иметь некоторую взаимную информацию.

Значение AMI также варьируется от 0 до 1:

* + 1 — полное совпадение разбиений,
  + 0 — совпадение на уровне случайности.

AMI часто используется, когда хочется оценить общую «информационную близость» между кластеризацией и эталоном, особенно если число кластеров и классов разное.

**Описание набора данных**

Заданный по варианту набор данных о губках. Это атлантико-средиземноморские морские губки, относящиеся к O.Hadromerida (Demospongiae.Porifera).

27 атрибутов являются нечисловыми и номинальными. 15 атрибутов являются булевыми и принимают значения (NO SI). 3 атрибута являются числовыми и принимают натуральные числа.

**Описание реализации**

В нашем проекте для кластеризации данных о видах морских губок использовался комплексный подход, включающий предобработку, снижение размерности и сравнительный анализ трёх алгоритмов кластеризации: KMeans, AgglomerativeClustering и GaussianMixture. Работа началась с загрузки исходного датасета, где для удобства и чистоты данных были заменены пропуски и проведена категоризация признаков на булевые и числовые. Булевые признаки, выраженные как 'SI' и 'NO', были преобразованы в бинарный формат, а числовые заполнены медианными значениями с помощью импьютера, что позволило избежать искажений при дальнейшем анализе.

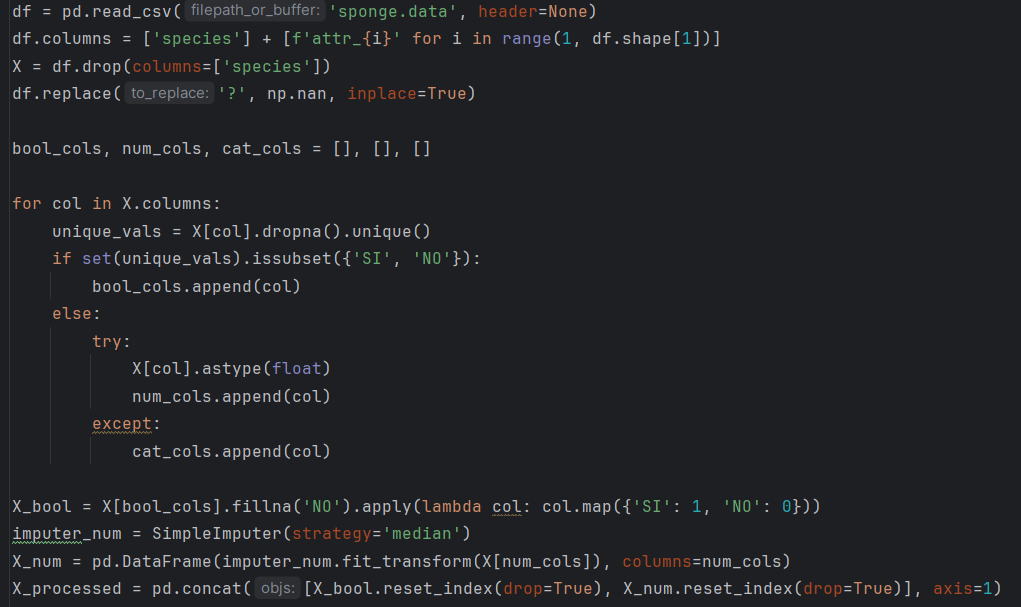
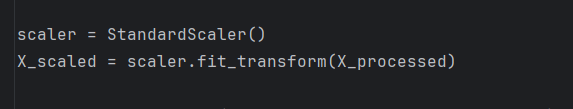


Рис 1. Загрузка и подготовка данных

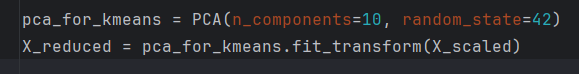
Далее данные были стандартизированы с помощью StandardScaler, чтобы нивелировать разницу в масштабах признаков и подготовить их для более корректной работы алгоритмов.



**Рис 2.** StandardScaler

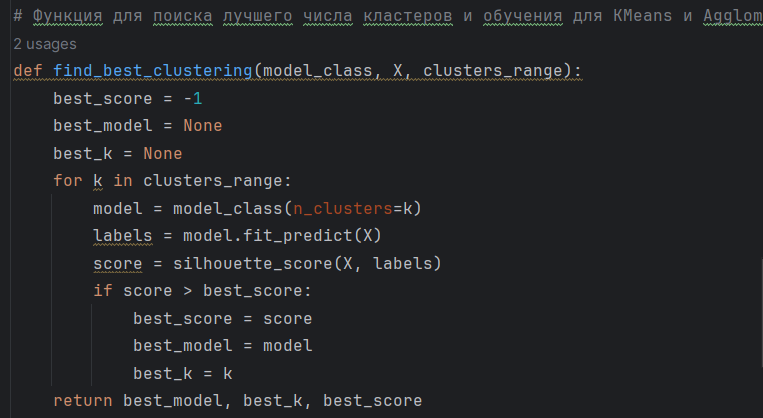
Поскольку исходное пространство признаков могло содержать избыточную и коррелированную информацию, применили метод главных компонент (PCA), уменьшив размерность до 10 компонент. Это улучшило вычислительную эффективность и качество кластеризации, снизив шумовые эффекты. PCA — это способ выделить из множества переменных (признаков) такие новые переменные (компоненты), которые максимально сохраняют информацию о вариациях в данных. То есть первые несколько главных компонент содержат почти всю важную информацию, а остальные — уже менее значимы и могут быть отброшены без большой потери качества.

Почему мы это сделали? Большое количество признаков усложняет и замедляет работу алгоритмов кластеризации, а также может привести к «проклятию размерности», когда из-за высокой размерности данные становятся разреженными и сложными для анализа.



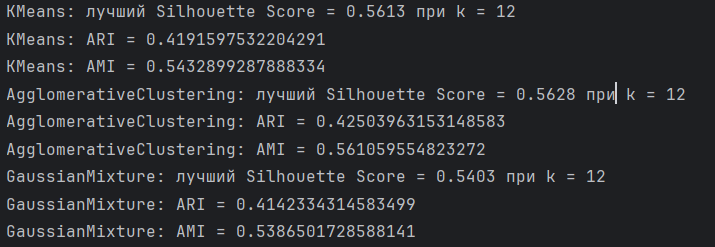
**Рис 3.** PCA

Основная часть эксперимента заключалась в подборе оптимального количества кластеров для каждого метода. Для KMeans и иерархического кластерного анализа AgglomerativeClustering реализовали функцию, которая перебирает k в диапазоне от 2 до 12, обучает модель и вычисляет метрику качества — силуэтный коэффициент (Silhouette Score). Лучшее значение силуэта при k=12 было зафиксировано для обоих алгоритмов. Аналогичный подбор провели для GaussianMixture, который также показал максимум при 12 кластерах.



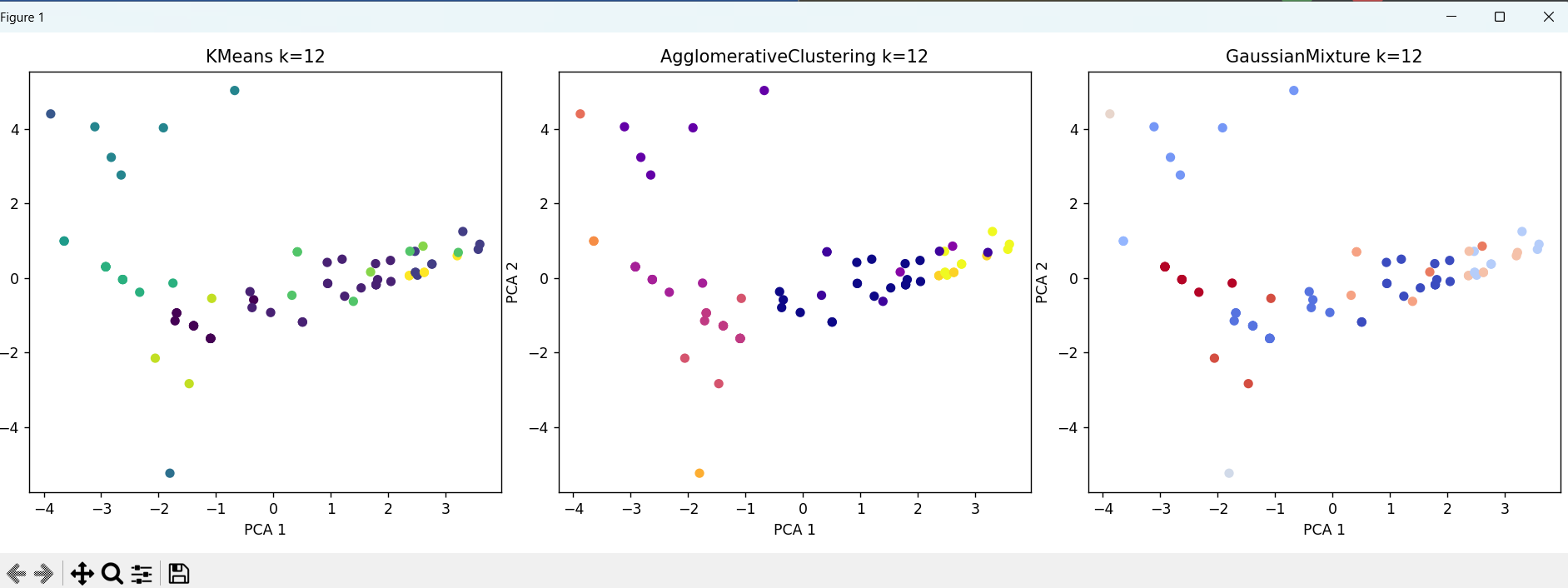
**Рис 4.** Функция для поиска лучшего числа кластеров

После выбора оптимального числа кластеров для каждого метода дополнительно оценили качество кластеризации с помощью метрик соответствия истинным классам: Adjusted Rand Index (ARI) и Adjusted Mutual Information (AMI).



**Рис 5.** Вывод

AgglomerativeClustering показал незначительно лучшие показатели ARI (0.425) и AMI (0.561), что свидетельствует о его чуть более точном разделении данных с учётом известной разметки, по сравнению с KMeans и GaussianMixture. При этом силуэтный балл, отражающий компактность и раздельность кластеров, был примерно одинаков для KMeans и AgglomerativeClustering (около 0.56), а GaussianMixture отстал с 0.54. Это связано с тем, что GMM использует вероятностный подход, который учитывает более мягкую принадлежность точек к кластерам, что иногда снижает чёткость границ кластеров.



**Рис 5.** Графики

Визуализация результатов на двумерном пространстве, полученном с помощью PCA, показала, что все три метода смогли выявить структуру данных с достаточно хорошей дифференциацией групп. При этом наблюдается, что иерархический метод формирует более чёткие кластеры, что подтверждается и метриками.

Таким образом, в нашем исследовании реализована комплексная процедура от предобработки до выбора и оценки моделей кластеризации, что позволило объективно сравнить три популярных метода и выявить, что AgglomerativeClustering в данном случае даёт несколько более качественную сегментацию, сохраняя при этом высокую интерпретируемость и стабильность результатов.