Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Ульяновский государственный технический университет»
Кафедра «Информационные системы и технологии»

Методы и технологии интеллектуальной обработки и анализа данных

Лабораторная работа №4

«Исследование алгоритма кластеризации»

Выполнила:
Кулагина П.С.
ИСТмд-11

Проверил:
Шишкин В.В.
к.т.н, доцент кафедры «ИВК»

г. Ульяновск
2025

1. Формулировка проблемы

Во многих практических задачах, где требуется выявить внутреннюю структуру данных, отсутствуют заранее размеченные классы. В таких случаях применяется кластеризация, позволяющая разделить выборку на группы объектов, сходных по совокупности признаков. Корректность результата при этом зависит от предварительной обработки данных и выбора параметров алгоритма, поэтому их влияние необходимо исследовать экспериментально.

В качестве примера для данного исследования используется набор данных о сотрудниках, включающий характеристики их рабочего опыта, условий труда и вовлечённости. Кластеризация в данном случае рассматривается как способ выявления типичных профилей сотрудников. При этом предметная область служит иллюстрацией применимости метода, тогда как основной фокус исследования направлен на оценку устойчивости и интерпретируемости выделенных кластеров.

2. Гипотеза

Предполагается, что результат кластеризации существенно зависит от способа подготовки данных. Если привести числовые признаки к единому масштабу и корректно преобразовать категориальные признаки в числовое представление, то алгоритм кластеризации будет формировать более чёткие и устойчивые группы. Если же кластеризацию выполнить без предварительной обработки признаков, разбиение окажется неустойчивым и слабо интерпретируемым.

3. План исследования

1. Изучить состав признаков исходного набора данных и определить переменные, не несущие аналитической нагрузки (идентификаторы, константы).
2. Выполнить два варианта подготовки данных:
    - Схема А: использовать только числовые признаки с

предварительной стандартизацией.

- Схема B (с предобработкой): выполнить стандартизацию числовых признаков и One-Hot кодирование категориальных.

3. Для каждого варианта подготовки данных применить алгоритм кластеризации K-Means.
4. Определить оптимальное количество кластеров на основе внутреннего критерия качества (коэффициент силуэта) и выбрать значение k для финального разбиения.
5. Сравнить результаты кластеризации по: значениям силуэта; устойчивости разбиения при повторных запусках; удобству интерпретации полученных кластеров.
6. Оценить, влияет ли предварительная обработка признаков на качество и читаемость кластерной структуры.

4. План эксперимента

План эксперимента строится на сравнении двух схем подготовки данных при неизменном алгоритме кластеризации. В первой схеме используются только числовые признаки, предварительно приведённые к единому масштабу. Во второй схеме используются числовые и категориальные признаки, где категориальные переменные преобразуются методом One-Hot Encoding. Для обеих схем дальнейшие процедуры идентичны: рассматривается ряд значений числа кластеров, и для каждого значения рассчитывается внутренний критерий качества (коэффициент силуэта; при необходимости дополнительно оценивается индекс Дэвиса—Булдина), что позволяет выбрать рабочий диапазон k без опоры на предметные метки.

После определения диапазона числа кластеров выполняется финальное разбиение для каждой схемы и проводится сопоставление полученных решений по трем осям: значению внутренней метрики, устойчивости к инициализации (повторные запуски с разными начальными центрами) и

интерпретируемости профилей кластеров на уровне признаков. Для наглядной проверки разделимости планируется использовать низкоразмерную проекцию признакового пространства (например, PCA), не влияющую на саму кластеризацию, а служащую лишь для визуальной валидации структуры. Итогом эксперимента станет вывод о том, влияет ли корректная предобработка на чёткость и стабильность найденных кластеров.

5. Реализация эксперимента

Для проведения кластерного анализа использовался набор данных Human Resources. Исходные данные содержали 35 признаков, часть которых дублировала идентификаторы или не содержала информативности для кластеризации. Признаки EmployeeNumber, EmployeeCount и StandardHours были удалены как постоянные или уникальные для каждой записи, то есть не влияющие на структуру данных. Признак Over18 также был исключён, так как во всей выборке содержал одно и то же значение. После очистки осталось 30 признаков, включающих числовые показатели (например, возраст, стаж, ежемесячный доход) и категориальные переменные (роль, отдел, образование и др.).

Далее были сформированы две альтернативные схемы подготовки данных:

1. Схема A (только числовые признаки).

В этом варианте использовались только числовые показатели (возраст, доход, стаж, уровень удовлетворённости и др.). Перед кластеризацией данные были стандартизированы с помощью *StandardScaler*. Для каждого значения k от 2 до 8 выполнялась кластеризация методом k-means, и вычислялась силуэт-метрика.

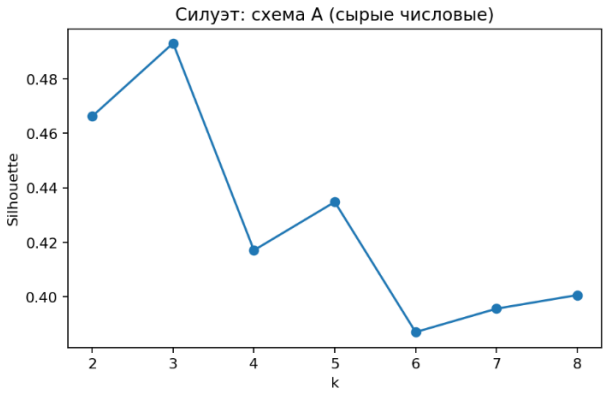


Рис 1. Значения силуэт-метрики в зависимости от числа кластеров (схема A)

График зависимости значения метрики от числа кластеров (*рис.1*) показал, что наилучший результат достигается при $k = 3$.

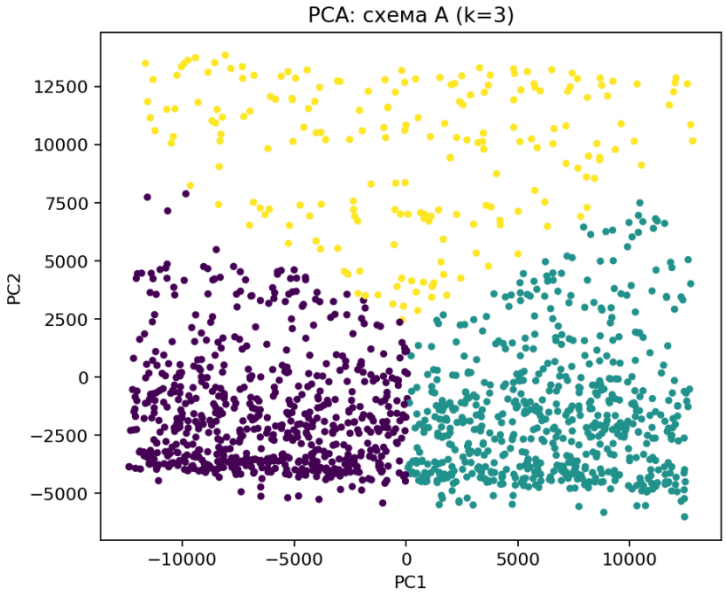
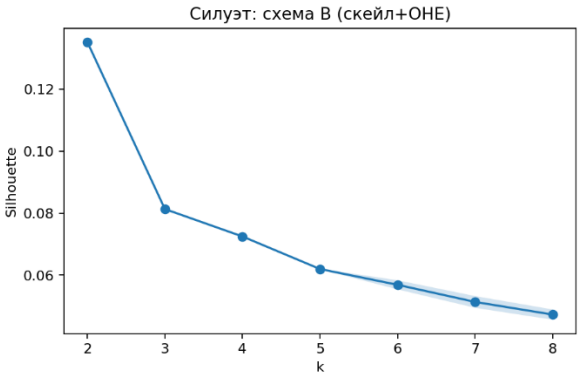


Рис 2. Размещение объектов на плоскости главных компонент при k=3 (схема A)

Визуализация данных в двух измерениях с использованием PCA (*рис.2*) подтверждает наличие трех отчётливо отделимых групп, что означает, что только на основе числовых признаков данные действительно образуют устойчивую кластерную структуру.

2. Схема B (числовые + категориальные признаки).



3.

Рис 3. Значения силуэт-метрики в зависимости от числа кластеров (схема B)

Во второй схеме использовались все признаки после преобразования категориальных переменных методом One-Hot Encoding. В результате размерность признакового пространства значительно увеличилась. Аналогичная процедура кластеризации и оценки силуэт-метрики показала, что значения метрики для всех k низкие (максимум около 0.13). Это видно на рисунке 3.

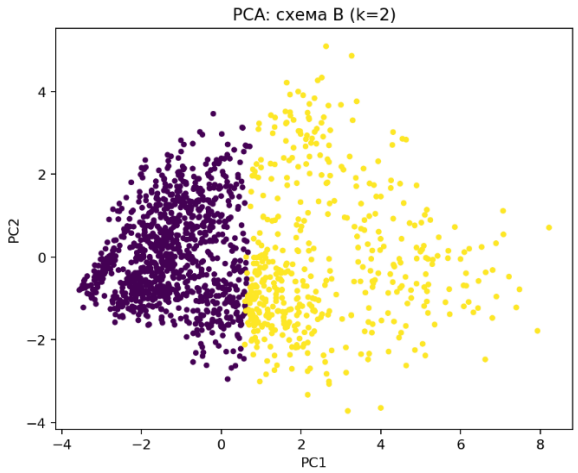


Рис 4. Размещение объектов на плоскости главных компонент при k=2 (схема B)

Визуализация распределения объектов после снижения размерности на рисунке 4 показывает плотное облако точек без выраженного разделения на группы. Таким образом, добавление категориальных признаков ухудшило качество кластеризации.

Основная причина заключается в том, что One-Hot Encoding значительно

увеличивает размерность пространства и изменяет баланс признаков: бинарные компоненты начинают доминировать при вычислении расстояний между объектами, из-за чего k-means теряет способность корректно выделять центры кластеров.

6. Вывод

Таким образом, гипотеза о зависимости качества кластеризации от выбранного способа подготовки данных подтверждена. Для кластерного анализа предпочтительно использовать числовые признаки, предварительно нормализованные, поскольку они сохраняют смысловые различия между объектами и обеспечивают более отчётливую кластерную структуру.

При этом важно отметить, что снижение качества кластеризации при добавлении категориальных признаков связано не с их природой, а с тем, что One-Hot Encoding значительно увеличивает размерность пространства и изменяет вклад признаков при вычислении расстояний. В реальных задачах могут использоваться альтернативные способы кодирования категориальных переменных (например, Target Encoding), позволяющие уменьшить это влияние и сохранять интерпретируемость признаков.