Кластеризация результатов анкетирования

А. К. Петрова, С. Е. Абрамкин

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

ak72p@yandex.ru, seabramkin@etu.ru

А. А. Петров

Факультет фундаментальной математики и механики Московского государственного университета им. М. В. Ломоносова

fczapetrov@yandex.ru

***Аннотация*. В современном анализе данных часто возникает задача выявления скрытых закономерностей в результатах анкетирования, где встречаются как числовые, так и текстовые ответы. Для решения подобных задач применяются методы машинного обучения, в частности - кластеризация и обработка естественного языка. В статье рассмотрен анализ оценки обратной связи по образовательному курсу с применением этих методов.**

Ключевые слова: кластеризация; анкетирование; сегментация; векторизация; метод главных компонент; машинное обучение

I. Введение

В условиях цифровизации образования, задача структурирования и интерпретации обратной связи обучающихся становится особенно актуальной. Анкетирование является основным инструментом изучения мнений и предпочтений студентов, однако разнообразие форм ответов (от числовых оценок до развернутых текстов) усложняет анализ. Для повышения эффективности выделения паттернов в таких данных применяются современные методы машинного обучения, в частности – кластеризация, что позволяет не только структурировать аудиторию, но и выявить сегменты для последующей персонализации образовательных продуктов

В последние годы вопросы кластеризации анкетных данных активно обсуждаются в научном сообществе. В работах [1–3] подробно рассматриваются различные алгоритмы кластеризации и их применение к социальным и образовательным данным. Особое внимание уделяется проблеме выбора метрик схожести для смешанных данных, а также вопросам интерпретируемости результатов. Современные исследования подчеркивают важность предварительной обработки текстовых данных.

В рамках данного проекта проведён анализ оценки обратной связи по курсу Маркетинг с использованием комплексного подхода: были обработаны как числовые, так и текстовые ответы. Основные задачи – выявить группы респондентов с помощью различных методов кластеризации и визуализировать результаты в удобном для интерпретации виде.

II. Подготовка и обработка данных

A. Разделение на числовые и текстовые признаки

На первом этапе данные были разделены на две части: числовые и текстовые. К числовым – отнесены все столбцы, кроме четырёх вопросов с открытыми ответами, которые были выделены в отдельную переменную. Это позволило индивидуально обработать каждый тип данных наиболее подходящим способом.

B. Векторизация текстовых данных

Для преобразования текстовых ответов в числовой формат использовался метод TF-IDF. Это позволило преобразовать текстовые ответы в числовые векторы и учесть не только частоту встречаемости слов, но и их значимость в контексте всех ответов. В результате текстовые столбцы были представлены в виде разреженных матриц признаков, которые затем объединялись с числовыми данными в единую матрицу признаков для последующего анализа.

C. Объединение и нормализация данных

Далее числовые и текстовые признаки были объединены в общий датасет. Для корректной работы алгоритмов машинного обучения данные были нормализованы с помощью StandardScaler (без вычитания среднего, чтобы корректно обрабатывать разреженные данные).

D. Снижение размерности (PCA)

Для визуализации и уменьшения размерности данных применяется метод главных компонент (PCA), позволяющий свести исходные признаки к трём наиболее информативным и построить 3D-графики распределения респондентов по кластерам.

В оценке образовательного курса Маркетинг главными признаками оказались текстовые столбцы: «Какие наиболее ценные знания вы получили на курсе?» «Какие знания были наименее ценными?», «Что можно улучшить?».

III. Кластеризация: методы и результаты

A. Применённые методы

В исследовании использовались три метода кластеризации [4–10]:

- **K-Means** – быстрый и интерпретируемый метод, требующий заранее заданного числа кластеров. Позволяет выделить основные группы респондентов по тематике интересов
- **Агломеративная (иерархическая) кластеризация** – позволяет строить дендрограммы и выявлять иерархию групп что полезно для уточнения структуры данных и поиска вложенных подгрупп.
- **DBSCAN** – плотностной метод, автоматически определяющий число кластеров и выделяющий шумовые точки (аномалии), эффективен при наличии кластеров сложной формы, но требует тщательного подбора параметров.

B. Результаты

K-Means

Метод K-Means позволил разбить данные на четыре кластера, для каждого из которых были определены центры. Визуализация результатов выполнена в 3D-пространстве главных компонент. Для каждого кластера был проведён анализ наиболее часто встречающихся слов в текстовых ответах.

Иерархическая кластеризация

Агломеративная кластеризация также разбила данные на четыре группы. Для визуализации центры кластеров были вычислены как средние значения координат точек в каждом кластере. Сравнение с K-Means показало схожие, но не идентичные результаты.

DBSCAN

Метод DBSCAN автоматически определяет число кластеров и выделяет шумовые точки (отмеченные меткой -1). Этот подход хорошо работает при наличии кластеров разной формы, однако требует тщательного подбора параметров (eps и min_samples).

Результаты кластеризации тремя методами представлены на рис. 1 и в табл. 1 (графики получены в среде программирования Engee).

C. Сравнение результатов (табл. 1)

- **K-Means** показал интерпретируемость и позволил выделить основные группы респондентов по тематике интересов.
- **Иерархическая кластеризация** выявила схожие, но не идентичные группы, что позволило уточнить структуру данных.
- **DBSCAN** оказался полезен для обнаружения аномалий и «неопределившихся» респондентов, однако требовал тщательного подбора параметров.

ТАБЛИЦА I. СРАВНЕНИЕ МЕТОДОВ КЛАСТЕРИЗАЦИИ

	Критерии сравнения	
	Преимущества	***Недостатки***
K-Means	Быстрый, простая интерпретация	Требует заранее знать число кластеров
Иерархиче ская	Позволяет строить дендрограмму	Высокая вычислительная сложность
DBSCAN	Устойчив к шуму, не требует заранее зант числа кластеро	Чувствителен к параметрам

D. Интерпретация кластеров

Для каждого кластера был проведён анализ наиболее часто встречающихся слов и фраз в текстовых ответах. Это позволило выявить ключевые темы для разных групп респондентов, технические и организационные аспекты образовательного процесса.

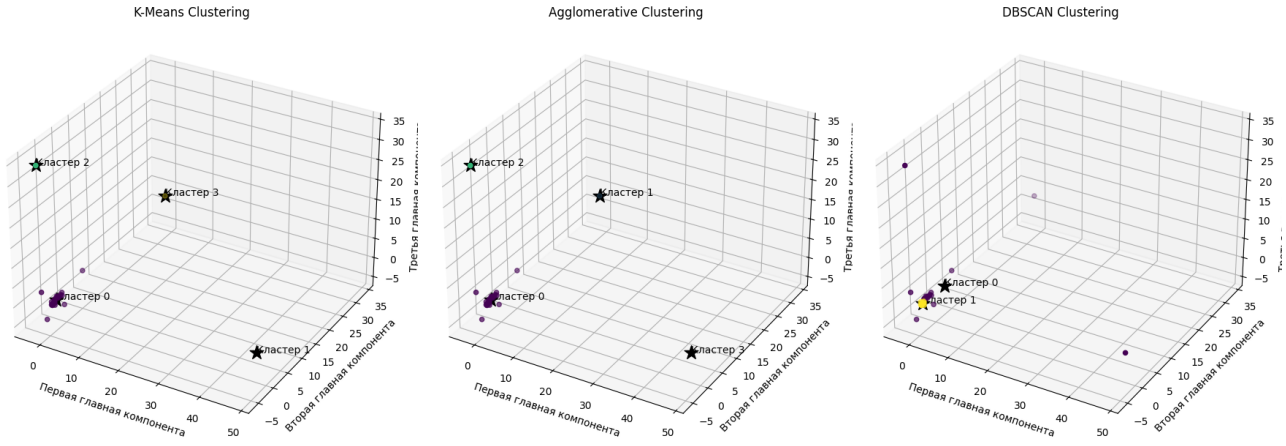

Рис. 1. Результаты кластеризации

В оценке образовательного курса Маркетинг наиболее часто встречающимися словами оказались (из результатов удалены неинформативные слова, такие как предлоги и т. п.):

Кластер 0: ('команде'), ('работа');

Кластер 1: ('преподавателя');

Кластер 2: ('научились'), ('каналы'), ('коммуникации');

Кластер 3: ('анализ'), ('конкурентов), ('стратегии').

Нас основе полученных кластеров можно сделать сегментацию обучающихся, следующим образом:

Сегмент 1: Фокус на командной работе;

Сегмент 2: Фокус на отношении преподавателя;

Сегмент 3: Фокус на прикладные навыки, в частности по разработке каналов коммуникаций;

Сегмент 4: Фокус на стратегический анализ как на ключевой ценности.

E. Практические выводы и рекомендации

1. **K-Means** оказался оптимальным методом для обеих задач благодаря наглядности и простоте интерпретации.
2. Для повышения качества анализа рекомендуется:

- использовать лемматизацию и очистку текстов,
- автоматизировать подбор параметров кластеризации (например, через GridSearch),
- применять интерактивные средства визуализации (Plotly) для глубокого анализа структуры данных.

3. Полученные кластеры могут быть использованы для персонализации образовательных программ.

F. Дополнительные аспекты и современные подходы

Исследования показывают, что методы кластеризации могут быть классифицированы по принципу используемой метрики: на основе расстояния (например, K-Means), на основе модели (например, EM-алгоритм для гауссовских смесей), а также на основе плотности (DBSCAN). Выбор метода зависит от структуры данных, их размерности и наличия выбросов.

В последние годы активно развиваются гибридные подходы, сочетающие кластеризацию с методами понижения размерности и автоматической оптимизацией числа кластеров по внутренним метрикам. Это позволяет повысить достоверность и интерпретируемость результатов, особенно при анализе больших и разнородных массивов анкетных данных.

IV. Заключение

Комплексный анализ анкетных данных с учётом как числовых, так и текстовых признаков позволяет глубже понять структуру аудитории, выявить скрытые паттерны и принять обоснованные решения для дальнейшего развития образовательного продукта. Применение современных методов машинного обучения и обработки естественного языка открывает широкие возможности для автоматизации и повышения эффективности аналитики в самых разных сферах.

Список литературы

[1] Васюкова Е.О. Реализация алгоритма кластеризации FOREL в математической среде MATLAB // Вестник новых технологий. 2020. № 2. С. 77–83.

[2] Сабиров В. Игра в цифры. М.: Манн, Иванов и Фербер, 2020. 304 с.

[3] Wheelan C. Naked Statistics: Стриптиз без формул. М.: Манн, Иванов и Фербер, 2017. 352 с.

[4] Воронцов К.В. Машинное обучение: курс лекций. М.: МГУ, 2024. 412 с.

[5] Долгодворова Е.В. Кластерный анализ: базовые концепции и алгоритмы // Вопросы науки и образования. 2018. № 7 (19). С. 73–76.

[6] Кластеризация: суть и задачи // GeekBrains. 2023. URL: https://geekbrains.ru/posts/clustering (дата обращения: 24.04.2025).

[7] Методы кластерного анализа: содержательная группировка данных. М.: Editverse, 2024. 176 с.

[8] Croll A., Yoskovitz B. Lean Analytics: Используй данные для создания прибыльного стартапа. М.: Альпина Паблишер, 2019. 320 с.

[9] Csera L., Feher T. Business Value in the Ocean of Data. Budapest: Data Science Press, 2018. 256 p.

[10] Методы кластеризации в исследовании массивов геоданных // Математическое моделирование. 2024. № 1. С. 102–115.