Намычкин Сергей Дмитриевич, магистрант,
Поволжский государственный университет
телекоммуникаций и информатики

СИСТЕМА ИНТЕЛЛЕКТУАЛЬНОЙ КЛАСТЕРИЗАЦИИ ТЕКСТОВЫХ ДАННЫХ

Аннотация. Современный этап развития информационных технологий характеризуется стремительным ростом объемов данных, генерируемых в различных сферах деятельности: от бизнес-аналитики и маркетинга до медицины и промышленности. В связи с этим возрастает потребность в эффективных методах автоматизированной обработки информации, позволяющих выявлять скрытые закономерности, структурировать данные и принимать обоснованные решения.

Ключевые слова: Кластеризация, текстовые данные.

Текстовые данные – это неструктурированная информация, представленная в виде последовательностей символов, слов, предложений или документов. Отметим основные особенности текстовых данных: высокая размерность, семантическая неоднородность, шум, зависимость от языка.

Пример.

Отзывы пользователей в интернет-магазине содержат как структурированную оценку (рейтинг), так и неструктурированный текст с эмоциональной окраской.

Кластерный анализ текстов – это метод машинного обучения без учителя, предназначенный для автоматической группировки текстовых данных на основе их смыслового сходства. Основная цель-объединить тексты (документы, предложения, отзывы и т.д.) в группы (кластеры) так, чтобы элементы внутри одной группы были максимально похожи друг на друга, а между разными группами – значительно отличались.

Процесс решения задачи кластеризации коллекции текстовых документов, как правило, представляет собой последовательное выполнение четырех шагов: предварительная обработка данных; представление данных в векторном виде; применение метода машинного обучения; оценка качества кластеризации.

Задача кластеризация относится к классу задач обучения без учителя, то есть для группирования объектов в кластеры не нужно перед этим обучаться на данных, для которых заранее известен результат распределения объектов по классам. Классические методы кластерного анализа принимают на вход матрицу $N \times M$ числовых значений, в которой строки соответствуют объектам, которые необходимо сгруппировать в классы, а столбцы – их признакам. Такие алгоритмы кластеризации, как правило, основаны на подсчете расстояния между объектами в пространстве размерности M.

Теперь рассмотрим подробно самый популярный и простой в реализации алгоритм **k-means**. Этот алгоритм был открыть в различных дисциплинах Ллойдом (1957), Форджи (1965), Фридманом и Рубином (1967), а также МакКуином (1967).

Достоинства: простота, скорость работы.

Недостатки: требует задания числа кластеров, чувствителен к шуму и начальной инициализации.

Для текста эффективен после снижения размерности (PCA, LSA).

DBSCAN – это алгоритм кластеризации, который основан на плотности. Если задан набор точек в некотором пространстве, то алгоритм группирует вместе точки, которые тесно

расположены (точки со многими близкими соседями), при этом помечая как выбросы те точки, которые находятся одиноко в областях с малой плотностью (ближайшие соседи лежат далеко).

Иерархическая кластеризация – это итеративная процедура, используемая для создания иерархии кластеров.

Алгоритмы иерархической кластеризации используются для обнаружения основных закономерностей в наборе данных для проведения статистических исследований.

Тематическое моделирование – это подход анализа текстовых данных, направленный на выявление скрытых тематических структур в коллекции документов. Оно позволяет автоматически определить, какие темы присутствуют в наборе текстов, и какие слова характеризуют каждую тему.

Основная идея заключается в том, что каждый документ в коллекции можно представить как смесь различных тем. В свою очередь, каждая тема связана с определённым распределением слов. Таким образом, тематическое моделирование позволяет разложить тексты на «скрытые» темы и показать, какие слова наиболее характерны для каждой темы.

Оценка качества кластеризации – это важный этап анализа данных, который позволяет определить, насколько хорошо алгоритм кластеризации разделил данные на группы. Поскольку в кластеризации нет "правильных" меток (в отличие от классификации), используются *метрики внутренней и внешней оценки*, а также визуализация.

Метрики внутренней оценки (Internal Metrics) основаны на анализе структуры кластеров без внешней информации.

Метрики внешней оценки (External Metrics) используются, если есть истинные метки кластеров (ground truth).

Кластеризация данных – это мощный метод машинного обучения без учителя, который находит применение в самых разных областях. Вот основные сферы, где она используется: маркетинг и бизнес, биология и медицина, компьютерные науки и IT: финансы и банковское дело: логистика и транспорт.

Список литературы:

1. Алексеев, Д. П. Технологии интеллектуального анализа данных [Текст]: учебное пособие для вузов / Д. П. Алексеев, О.В. Щекочихин. – М: Высшее образование, 2022. – 256 с.

2. Барский, А.Б. Нейронные сети: распознавание, управление, принятие решений [Текст]: учебное пособие для вузов / А.Б. Барский. – М: Финансы и статистика, 2004. – 176 с.

3. Заенцев, И.В. Нейронные сети: основные модели [Текст]: учебное пособие для вузов / И.В. Заенцов. – Воронеж: Издательство Воронежского государственного университета, 1999. – 142 с.