ИССЛЕДОВАНИЕ И ПЕРСПЕКТИВЫ ОБУЧЕНИЯ МОДЕЛЕЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ НА БОЛЬШИХ ДАННЫХ: КЛАСТЕРИЗАЦИЯ И ЕЕ РОЛЬ В ЭФФЕКТИВНОЙ ОБРАБОТКЕ ИНФОРМАЦИИ

Максим САВЕЛЬЕВ

Департамент Программной Инженерии и Автоматики, Группа TI-217, Факультет Вычислительной Техники и Микроэлектроники, Технический Университет Молдовы, Кишинёв, Республика Молдова

Автор: Максим Савельев, maxim.saveliev@isa.utm.md

Научный руководитель: **Дориан САРАНЧУК**, преподаватель, ТУМ

Аннотация: *Данная статья нацелена на исследование и анализ актуальных аспектов обучения моделей искусственного интеллекта (ИИ) и машинного обучения (МО) на больших данных, с углубленным фокусом на роли кластеризации в процессе эффективной обработки информации.*

Ключевые слова: *Искусственный интеллект, машинное обучение, большие данные, кластеризация, эффективная обработка информации, взаимосвязь данных, алгоритмы кластеризации, анализ данных*

Введение

В современном мире данные становятся всё более обширными, искусственный интеллект (ИИ) и машинное обучение (МО) играют важную роль в их обработке. Данная статья исследует, как ИИ и МО обучаются на больших данных, с фокусом на методе кластеризации для более эффективной работы с информацией.

“Big Data” - это технологии обработки структурированных и неструктурированных данных, которые постоянно увеличиваются в объеме.

Обработка больших объемов данных (от 100 Тбайт) позволяет находить более точные связи между данными, что в свою очередь упрощает аналитику и представление данных в понятном виде. В настоящее время объем информации может достигать сотен петабайт и даже эксабайт.

Основные характеристики Big Data

Для того чтобы данные могли быть отнесены к категории «big», необходимо, чтобы они соответствовали следующим характеристикам [1]:

- Объем (Volume): Информация измеряется в физической величине и занимает значительное пространство на цифровом носителе. Классифицируются как «big» массивы данных, превышающие 150 Гб в сутки.
- Скорость (Velocity): Данные регулярно обновляются, и для обработки в реальном времени требуются интеллектуальные технологии в рамках концепции больших данных.
- Разнообразие (Variety): Информация в этих массивах может иметь разнообразные форматы, быть частично или полностью структурированной, а также собираться бессистемно. Например, большие данные включают в себя тексты, видео, аудио, финансовые транзакции и другие форматы.

В современных системах рассматриваются два дополнительных фактора:

- Изменчивость (Variability): Потоки данных могут иметь пики и спады, сезонность и периодичность. Управление всплесками неструктурированной информации требует мощных технологий обработки.

- Значение данных (Value): Информация может иметь разную сложность для восприятия и обработки, что представляет сложности для интеллектуальных систем. Например, необходимо определить степень важности поступающей информации для её быстрой структуризации.

Этапы подготовки данных для передачи их моделям машинного обучения

Для эффективного анализа и обработки обширных объемов данных применяются инструменты, в частности "аналитические модели".Эти модели строят гипотезы на основе больших данных, ищут в них зависимости и закономерности, что позволяет получить всю самую полезную информацию для большинства бизнес-задач. При этом важна хорошая интерпретируемость построенной модели, так как это позволяет упростить её анализ без повторного её построения, что при работе с большими данными крайне важно. Для этого большие данные проходят через несколько этапов: чистка данных, работа с признаками и построение и обучение аналитической модели для предсказания целевой переменной.

На первом этапе, проводится чистка данных, в ходе которой осуществляется выявление параметров с наименьшей корреляцией к целевым значениям [1]. После этого производится удаление этих параметров, поскольку их наличие может внести нежелательные шумы в процесс обучения нашей модели. Также обычно выполняют стандартизацию, масштабирование и бинаризацию количественных характеристик, а также производят замену отсутствующих значений средними значениями.

Далее следует работа с признаками, где генерируются новые переменные для построения аналитических моделей. Этот этап позволяет учесть разнообразные аспекты данных, делая модели более информативными.

Ключевым этапом является разработка и обучение аналитической модели с целью прогнозирования целевой переменной. На данном этапе проводится проверка гипотез относительно взаимосвязей между целевой переменной и предикторами. Особое внимание уделяется интерпретируемости построенной модели, что играет решающую роль в упрощении анализа без необходимости повторного создания [2].

Машинное Обучение и Искусственный Интеллект

Машинное обучение - это подраздел искусственного интеллекта, фокусирующийся на разработке алгоритмов и моделей, которые позволяют компьютерам извлекать знания из данных и использовать их для автоматизированного принятия решений [3]. В отличие от традиционного программирования, где человек явно задает инструкции компьютеру, машинное обучение обучает компьютерные системы самостоятельно обнаруживать закономерности в данных и приспосабливаться к новой информации. Примерами машинного обучения могут служить классификация изображений, прогнозирование временных рядов и рекомендательные системы.

Искусственный интеллект стремится создать устройства и программы, способные выполнять задачи, обычно требующие человеческого интеллекта. Включая в себя машинное обучение, ИИ охватывает широкий спектр методов, включая логическое программирование, обработку естественного языка и компьютерное зрение. Он нацелен на создание систем, способных адаптироваться к изменяющимся условиям, обучаться на опыте и решать сложные проблемы.

Взаимосвязь между Большими Данными Искусственным Интеллектом и Машинным Обучением

На Рис. 1 представлен процесс обработки Big Data. Большие данные предоставляют сырой материал для анализа. Машинное обучение используется для создания моделей, способных извлекать полезные знания из этих данных. Искусственный интеллект, в свою очередь, обеспечивает интеллектуальные функции и автоматизацию процессов на основе полученных знаний.

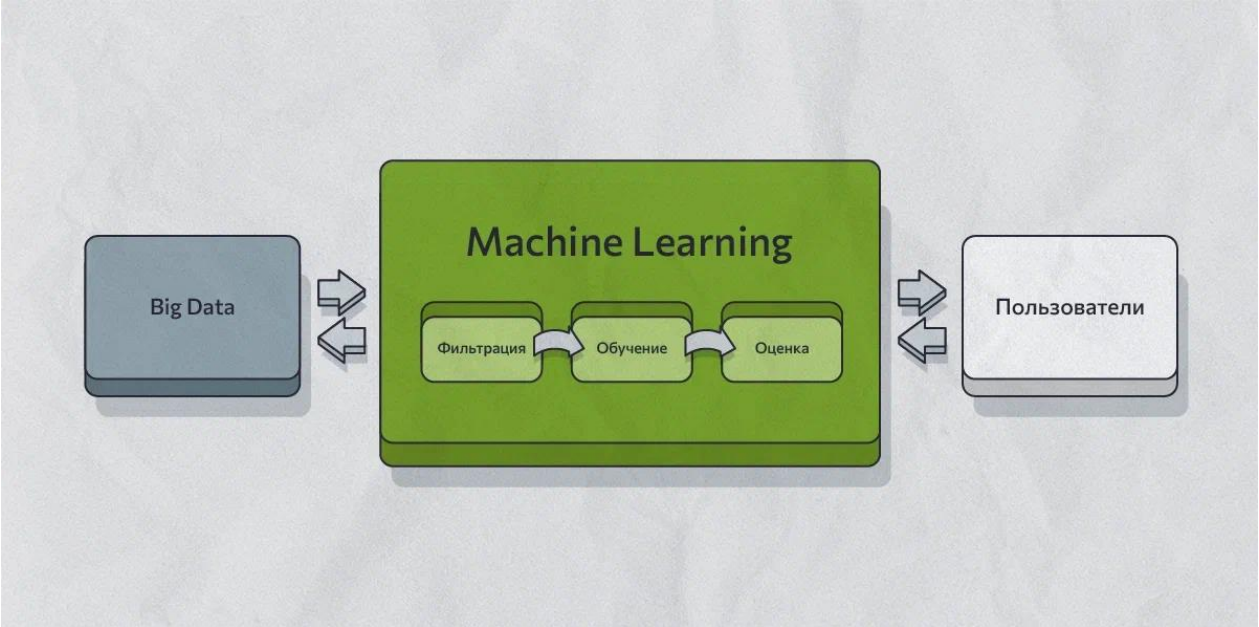

Рисунок 1. Этапы обработки Big Data в системах машинного обучения

Кластеризация в контексте больших данных и машинного обучения позволяет группировать данные по схожести, выделяя образующиеся паттерны и структуры. Искусственный интеллект, в свою очередь, может использовать результаты кластеризации для автоматического принятия решений и адаптации к изменяющейся среде.

Таким образом, кластеризация выступает как важное звено в этой взаимосвязи, обеспечивая более глубокий анализ данных, что в конечном итоге способствует улучшению и развитию искусственного интеллекта и машинного обучения.

Роль и значение кластеризации в обработке больших данных

Кластеризация - это метод машинного обучения, направленный на группировку схожих объектов внутри данных в кластеры или кластерные группы [4]. Целью кластеризации является максимизация схожести между элементами внутри одного кластера и минимизация схожести между элементами различных кластеров.

Роль кластеризации в обработке больших данных огромна, и ее значение можно выделить по нескольким направлениям:

1. Группировка по схожести: Кластеризация позволяет группировать данные, основываясь на их схожести и общих характеристиках.
2. Выделение структуры: Алгоритмы кластеризации выявляют внутренние структуры в больших наборах данных, делая их более понятными и интерпретируемыми.
3. Сжатие информации: Кластеризация позволяет сжимать информацию, представляя ее в виде отдельных кластеров.
4. Поиск аномалий: Выделение отдельных кластеров может помочь в выявлении аномалий и выбросов в данных, что имеет важное значение для обнаружения необычных событий или аномалий в больших объемах информации.
5. Улучшение процессов машинного обучения: Кластеризация может быть использована для предварительной обработки данных перед применением методов машинного обучения.

Таким образом, кластеризация в обработке больших данных не только помогает структурировать информацию, но и является важным инструментом для выделения паттернов, улучшения качества анализа и поддержки принятия решений.

Методы кластеризации

Существует множество методов кластеризации, каждый из которых подходит для определенных типов данных и задач [5]. Ниже представлен обзор некоторых популярных методов:

- K-средних (K-Means): Этот метод разделяет данные на K кластеров, где K представляет собой предварительно заданное число. Алгоритм минимизирует среднеквадратичное отклонение объектов внутри кластеров.
- Иерархическая кластеризация: Этот метод создает иерархию кластеров, начиная с каждого объекта как отдельного кластера и последовательно объединяя их в более крупные кластеры.
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Основанный на плотности, этот метод выделяет кластеры на основе плотности объектов, что позволяет обнаруживать кластеры произвольной формы.
- OPTICS (Ordering Points To Identify the Clustering Structure): Данный метод также основан на плотности, но, в отличие от DBSCAN, создает упорядоченный список объектов, что упрощает выделение кластеров различной плотности.
- Методы спектральной кластеризации: Основанные на матрицах схожести, эти методы используют собственные векторы для выделения структур в данных и выделения кластеров.

Выбор подходящего метода зависит от характеристик данных, структуры кластеров и конкретных целей анализа. Комбинирование различных методов и тщательный анализ результатов обычно приводят к наилучшим результатам в конкретных сценариях.

Метод K-средних (Алгоритм Ллойда)

Метод K-средних (K-Means) представляет собой итеративный алгоритм кластеризации, который разбивает набор данных на заранее заданное количество кластеров K [6]. Алгоритм начинает с инициализации K центроидов, которые могут быть выбраны случайным образом или иным образом. Затем каждый объект данных присваивается к ближайшему центроиду на основе выбранной метрики, такой как евклидово расстояние.

После присвоения объектов происходит перевычисление центроидов для каждого кластера. Новые центроиды представляют собой средние значения всех объектов в соответствующем кластере. Этот процесс присвоения и перевычисления повторяется до тех пор, пока на какой-то итерации не происходит изменения внутрикластерного расстояния.

Алгоритм завершается, когда стабилизируется разбиение на кластеры, и центроиды не изменяются на протяжении нескольких итераций Рис. 2. Важно отметить, что результаты могут зависеть от начального выбора центроидов, и алгоритм может сойтись к локальному минимуму. Поэтому рекомендуется запускать алгоритм несколько раз с разными начальными значениями и выбирать наилучший результат.

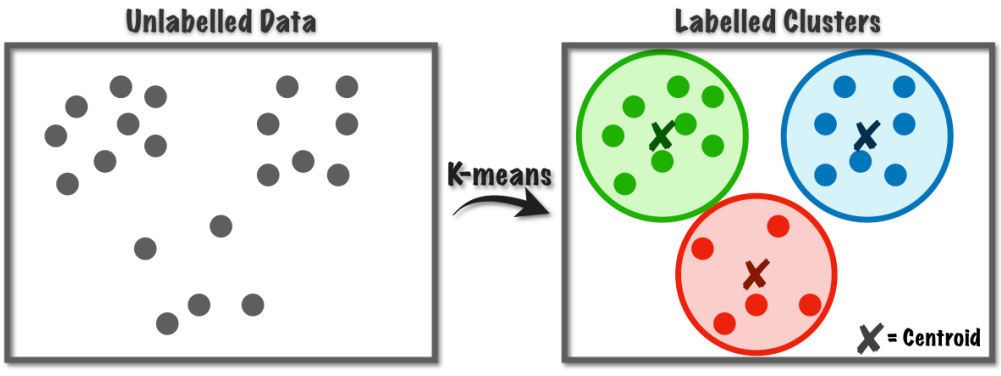

Рисунок 2. Процесс кластеризации данных по методу K-средних

Метод DBSCAN

Метод DBSCAN (Density-Based Spatial Clustering of Applications with Noise) представляет собой алгоритм кластеризации, который основывается на плотности данных. Он позволяет выявлять кластеры произвольной формы и обнаруживать выбросы в данных.

Основная идея DBSCAN заключается в следующих шагах. На Рис. 3. предствлен процесс отбора основных и соседних точек. Сначала случайным образом выбирается точка из нерассмотренных объектов данных. Если в окрестности этой точки находится минимальное количество соседей (определенное пользователем), то эта точка становится началом нового кластера. Затем происходит распространение по плотности: к кластеру добавляются соседи текущей точки, и процесс рекурсивно распространяется от соседей к их соседям до тех пор, пока не будут рассмотрены все доступные точки Рис. 4. Выбирается новая начальная точка из нерассмотренных объектов, и процесс повторяется.

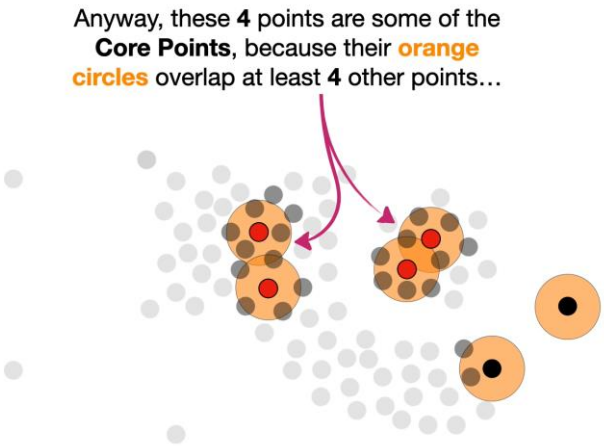

Рисунок 3. Выбор основных и соседних точек по методу DBSCAN [7]

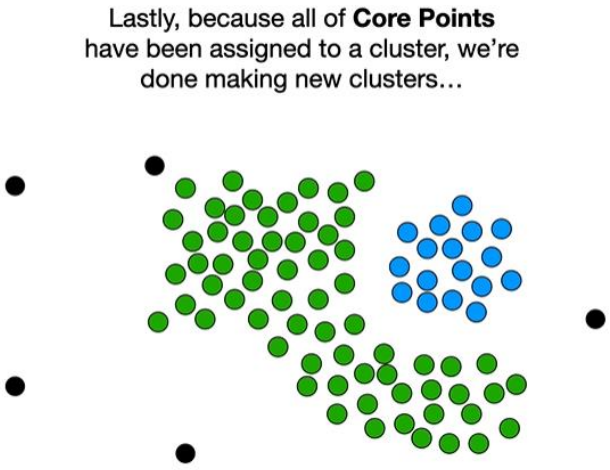

Рисунок 4. Результат кластеризации по методу DBSCAN [7]

Точки, которые не могут быть добавлены ни к одному кластеру, считаются выбросами. Основное преимущество DBSCAN в том, что он способен выделять кластеры произвольной формы, не требуя заранее заданного числа кластеров, и обнаруживать и игнорировать шумовые точки в данных. Однако, для достижения оптимальных результатов, необходимо тщательно настраивать параметры, такие как радиус окрестности и минимальное количество соседей, под конкретные условия использования.

Метод иерархической кластеризации

Иерархическая кластеризация представляет собой алгоритм, который создает иерархию кластеров, начиная с каждого объекта как отдельного кластера и последовательно объединяя их в более крупные кластеры. Этот метод основан на идее пошагового объединения близких кластеров, что позволяет представить данные в виде дерева (дендрограммы).

В начальной инициализации каждый объект данных рассматривается как отдельный кластер, формируя N кластеров, где N - число объектов. Затем вычисляется матрица расстояний между всеми парами кластеров. Далее выбираются два ближайших кластера для объединения, и процесс повторяется до тех пор, пока не останется один кластер, представляющий все данные. Пример полученной дендрограммы на Рис. 5.

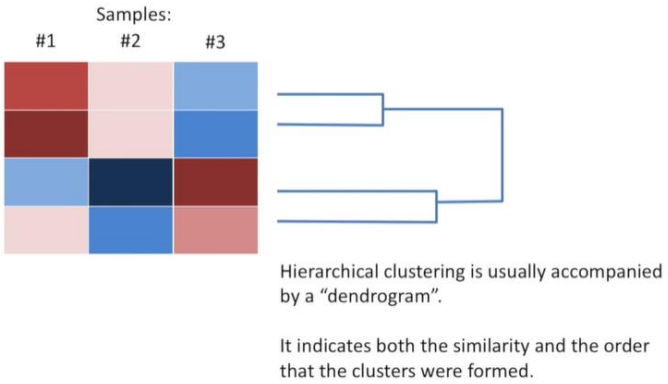

Рисунок 5. Этапы получения дендрограммы по методу иерархической кластеризации [7]

Процесс объединения кластеров строит дендрограмму для визуализации иерархической структуры. Иерархическая кластеризация не требует заранее заданного числа кластеров и подходит для данных с иерархической структурой. Однако она может быть вычислительно затратной, особенно при обработке больших объемов данных.

Заключение

В заключение данной статьи следует подчеркнуть важность исследования и применения методов машинного обучения и искусственного интеллекта на больших данных. Развитие технологий обработки и анализа данных в современном мире играет ключевую роль в принятии информированных решений в различных областях, начиная от бизнеса и заканчивая медициной и наукой.

В статье были расмотренны основные аспекты обучения моделей на больших данных, была подчеркнута важность роли кластеризации в эффективной обработке информации. Также были рассмотрены методы кластеризации, такие как K-средних, DBSCAN и иерархическая кластеризация, и выделен их вклад в группировку, структурирование данных, а также в выявление аномалий.

Данные алгоритмы и методы позволяют эффективно работать с большими данными, что становится критически важным аспектом для современных бизнес-процессов. В эпоху биг-дата объемы информации постоянно растут, и именно здесь машинное обучение выступает в роли ключевого инструмента для автоматизации и улучшения этого процесса. Автоматизация анализа больших данных позволяет бизнесу выявлять скрытые закономерности, высокоточно прогнозировать тренды и принимать обоснованные стратегические решения.

Библиография:

[1] ИТМО Обучение на больших данных - [Электронный ресурс] Режим доступа:
[2] https://neerc.ifmo.ru/wiki/index.php?title=%D0%9E%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_%D0%BD%D0%B0_%D0%B1%D0%BE%D0%BB%D1%8C%D1%88%D0%B8%D1%85_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85
[3] Enterprise AI - [Электронный ресурс] Режим доступа: https://www.techtarget.com/searchenterpriseai/tip/How-do-big-data-and-AI-work-together
[4] proglib - [Электронный ресурс] Режим доступа: https://proglib.io/p/kak-mashinnoe-obuchenie-uporyadochivaet-bolshie-dannye-2021-03-12
[5] Яндекс Практикум - [Электронный ресурс] Режим доступа: https://practicum.yandex.ru/blog/chto-takoe-klasterizaciya-i-klasternyi-analiz/
[6] ИТМО Кластеризация - [Электронный ресурс] Режим доступа: https://neerc.ifmo.ru/wiki/index.php?title=%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F
[7] Хабр - [Электронный ресурс] Режим доступа: https://habr.com/ru/companies/nix/articles/413269/
[8] StatQuest with Josh Starmer - [Электронный ресурс] Режим доступа: https://www.youtube.com/@statquest