А. Н. Бородулина, Ю. Е. Карякин
Тюменский государственный университет, г. Тюмень
УДК 550.8.05

АЛГОРИТМ ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ ДАННЫХ

Аннотация. В статье представлен универсальный алгоритм предварительной обработки данных, содержащий семь этапов. В рамках каждого этапа рассматривается его содержание, решаемые задачи и методы их решения.

Ключевые слова: предварительная обработка данных; методы предобработки данных; препроцессинг данных; машинное обучение.

Введение. Качественные данные — это непременное условие правильной работы любой модели машинного обучения. Но, как правило, собираемые данные непригодны для дальнейшего их использования без подготовки: в них присутствуют выбросы, пропуски, различные форматы заполнения и т. д., что делает необходимым проведение процедуры предварительной обработки данных.

Благодаря наличию готовых решений, технологии машинного обучения стали универсальным инструментом, который может применяться широким кругом пользователей, что делает особенно актуальной не решенную до сих пор [1; 195] проблему наличия общепринятой методики подготовки данных.

Как в научной литературе, так и на практике, данной теме уделено большое внимание: широко представлены как универсальные алгоритмы [1-5], так и специализированные, предназначенные для определенного типа данных или предметной области [6-8].

Исходя из вышесказанного, целью нашей работы является наиболее полное формулирование алгоритма процесса обработки данных. Для этого нам необходимо решить следующие задачи: рассмотреть существующие схемы препроцессинга данных, выявить их сходства и различия, и обобщить данный опыт.

Материалы и методы. Несмотря на то, существует большое разнообразие подходов к препроцессингу, под данными, как правило, если прямо не указано обратное, понимаются структурированные

(табличные) данные, что в значительной степени предопределяет выбор методов для их обработки Во всех рассмотренных источниках алгоритм предполагал под собой линейное прохождение ряда этапов (от трех до пяти), содержащих в себе определенные операции. На основе изученной информации можно сказать, что базовыми этапами препроцессинга можно назвать являются встречающиеся в том или ином виде во всех рассмотренных работах очистка данных, уменьшение размерности и трансформация признаков. Вместе с тем, различными исследователями выделяются иные этапы, такие, как например, этапы оценки качества данных [8; 222], интеграции данных [7;45], разделения данных [2; 27] и т. д.

Результаты. Наиболее полную схему препроцессинга данных можно представить следующим образом:

1. Оценка качества данных. На данном этапе предполагается решение следующих задач: унификация гетерогенных данных (устранение несоответствий в именах атрибутов, единицах измерения, периодичности значений, форматах данных) и оценка нормальности данных, выявление выбросов и пропусков.

2. Очистка данных — это наиболее важный этап предварительной обработки данных, который включает в себя удаление дубликатов и устранение пропусков, аномалий и шумов. Если доля пропущенных и аномальных данных сравнительно невелика и не окажет влияние на дальнейший анализ, целесообразным является их удаление. В противном случае, возникает необходимость их заполнения и корректировки. Для заполнения пропусков могут применяться следующие тактики: задание значения по умолчанию, использование средних или медианных значений, статистической корреляционной модели, методов машинного обучения (например, байесовские формальные методы и индукцию дерева решений). Для работы с шумом и аномалиями применяют биннинг, поиск регрессии, анализ выбросов.

3. Интеграция данных. Данный этап необходим, если после очистки осталось недостаточное количество данных для выполнения поставленной задачи и предполагает поиск и добавление информации из сторонних источников.

4. Сокращение объема данных включает в себя как сжатие с использованием инструментария кодирования данных, так и удаление излишней информации, что благоприятно сказывается на точности модели и объеме хранилища данных. Для уменьшения количества наблюдений могут использоваться различные методы выборки, для отбора признаков широко используются статистические методы, например, корреляция, на основе которой признаки ранжируются и отбираются наиболее релевантные.

5. Трансформация подразумевает преобразования, «направленные на улучшение качества признаков (и данных) или трансформацию признаков таким образом, чтобы они были применимы для машинного обучения» [3]. На этом этапе может изменяться тип переменной (путем дискретизации, порядкового преобразования или методом быстрого кодирования (One-Hot encoding). Происходит масштабирование данных при помощи нормализации и стандартизации, а также изменение распределения значений вероятности.

6. Конструирование признаков — это процесс создания новых входных переменных из доступных данных [1; 200] на основе логики и знаний из предметной области или математических преобразований, основная цель которого — разложение сложной переменной для более прозрачного представления, либо добавление расширенного контекста к отдельному наблюдению.

7. Балансирование данных. В ряде случаев встречается ситуация, когда один или несколько классов являются важными для модели, но представлены существенно меньшим количеством записей, чем остальные. При несбалансированности данных модель не сможет правильно обучиться, поэтому для решения данной проблемы используются следующие методы: усечения данных (такие, как CNN («сокращенное правило ближайшего соседа») и Tomek links), дополнения (например, SMOTE, создающий новые экземпляры на основе разности между образцом и его ближайшим соседом), либо их комбинация, что является наиболее целесообразным, поскольку не приводит к перемешиванию классов и построению сложной модели.

Заключение. Предварительная обработка данных является очень сложной и многогранной задачей, правильное решение которой оказывает определяющее влияние на корректность будущей модели. Несмотря на то, что состав и содержание этапов предобработки данных в значительной степени зависит от целевой задачи, решаемой в ходе анализа данных, некоторые общепринятые подходы возможно объединить в единую методику. Предложенный алгоритм является более подробным, поскольку объединил в себе как универсальные, так и специфические методы препроцессинга. Это позволит повысить качество данных, подаваемых на вход модели, и, следовательно, эффективность ее работы, что будет полезным для широкого круга лиц при подготовке моделей машинного обучения.

СПИСОК ЛИТЕРАТУРЫ

1. Татур М. М. «Сырые» данные и некоторые рецепты их «приготовления» / М. М. Татур, В. М. Проровский, Д. В. Куприянова, И. Н. Носырев. — Текст : электронный // Информационные системы и технологии : материалы Международного научного конгресса по информатике. — Минск: Белорусский государственный университет. — URL: https://www.elibrary.ru/download/elibrary_49732158_13093697.pdf (дата обращения: 24.05.2023).
2. Акимов А. А. Предварительная обработка данных для машинного обучения / А. А. Акимов, Д. Р. Валитов, А. И. Кубряк. — Текст : электронный // Научное обозрение. Технические науки. — 2022. — № 2. — URL: https://www.elibrary.ru/download/elibrary_48411861_16222054.pdf (дата обращения: 24.05.2023).
3. Кацер Ю. Д. Методы предварительной обработки данных для машинного обучения / Ю. Д. Кацер. — Текст : электронный // vc.ru: сайт. 2022. — URL: https://vc.ru/u/1167333-yuriy-katser/425314-metody-predvaritelnoy-obrabotki-dannyh-dlya-mashinnogo-obucheniya (дата обращения: 24.05.2023).
4. Кириллов Д. С. Способы предварительной обработки данных для прогнозирования с применением нейронных сетей / Д. С. Кириллов, Д. Д. Молостов. — Текст : непосредственный // Фундаментальная и прикладная наука: актуальные вопросы теории и практики : сборник статей Междунар. научно-практ. конф. — Пенза, 2023. — С. 45-47.

5. Юданова В. В. О предварительной подготовке больших данных для процессов data Minning / В. В. Юданова. — Текст : электронный // Наукосфера. — 2022. — № 10-2. — URL: https://www.elibrary.ru/ (дата обращения: 24.05.2023).
6. Бардамова М. Б. Методы предобработки несбалансированных данных / М. Б. Бардамова. — Текст : электронный // Сборник избранных статей научной сессии ТУСУР. — 2018. — № 1-3. — URL: https://www.elibrary.ru/ (дата обращения: 24.05.2023).
7. Копырин А. С. Алгоритм препроцессинга и унификации временных рядов на основе машинного обучения для структурирования данных / А. С. Копырин, И. Л. Макарова. — Текст : непосредственный // Программные системы и вычислительные методы. — 2020. — № 3. — С. 40-50.
8. Седова А. И. Методы предобработки массивов текстовых данных для их последующего анализа / А. И. Седова, А. А. Матушкова. — Текст : непосредственный // Оригинальные исследования. — 2022. — Т. 12, № 6. — С. 220-226.