УДК 004.891

Ю. С. Романова*
кандидат технических наук, доцент
Е. В. Пастухова*
кандидат технических наук, доцент
*Санкт-Петербургский государственный университет
аэрокосмического приборостроения

ЭФФЕКТИВНЫЕ ПОДХОДЫ К ПОДГОТОВКЕ ДАННЫХ ДЛЯ МАШИННОГО ОБУЧЕНИЯ

Рассматриваются различные методы подготовки данных для машинного обучения, включая очистку, масштабирование и работу с пропущенными значениями. Особое внимание уделяется техникам выбора признаков, агрегации данных и созданию новых признаков на основе существующих. Описанные эффективные стратегии обработки данных способны повысить качество моделей машинного обучения и улучшить их способность к обобщению.
Ключевые слова: dataset, машинное обучение, обработка данных.

Yu. S. Romanova*
PhD, Tech., Associate Professor
E. V. Pastukhova*
PhD, Tech., Associate Professor
*St. Petersburg State University of Aerospace Instrumentation

EFFECTIVE APPROACHES TO PREPARING A DATASET FOR MACHINE LEARNING

The article discusses various methods for preparing data for machine learning, including cleaning, scaling, and working with missing values. Particular attention is paid to techniques for feature selection, data aggregation, and the creation of new features based on existing ones. The considered effective data processing strategies can improve the quality of machine learning models and improve their ability to generalize.
Keywords: dataset, machine learning, data processing.

Введение

Предварительная подготовка данных играет ключевую роль в успешном применении алгоритмов машинного обучения. Необработанные или неправильно подготовленные данные могут привести к искаженным результатам модели, недостаточной точности и неправильным выводам. В свете этого исследование различных методов подготовки DataSet имеет высокую актуальность и значимость.

Материалы и методы

Подготовка данных помогает устранить шум, выбросы и пропущенные значения, что повышает стабильность и качество модели. Кроме того, масштабирование признаков позволяет моделям эффективнее работать с данными разного типа и масштаба. Выборка признаков, создание новых характери-

стик и агрегация данных способствуют созданию более информативных и универсальных моделей.

Первый этап обработки данных – **устранение шума и выбросов** – направлен на повышение качества и надежности моделей машинного обучения. Шум и выбросы могут привести к искажению статистических свойств данных и ухудшению обобщающей способности модели. Для их эффективного устранения применяются различные методы, каждый из которых имеет особенности и области применения [1].

Один из распространенных **методов устранения шума** – фильтрация данных, которая позволяет уменьшить влияние случайных колебаний и артефактов. Перечислим наиболее известные фильтры:

– медианный фильтр основан на замене значения каждого элемента данных медианным значением его окрестности. Применяется в обработке изображений, аудиосигналов и других типов данных, где важно сохранить структуру исходной информации;

– фильтр скользящего среднего вычисляет среднее значение элементов данных в окне заданной ширины и заменяет каждое значение на вычисленное, но он может сглаживать и полезные сигналы, поэтому его применение требует анализа влияния на конечные результаты;

– фильтр Гаусса основан на применении операции свертки с гауссовым ядром к исходным данным. Используется в обработке изображений и сигналов, когда необходимо сохранить детали и структуру данных;

– фильтр резонанса работает на основе принципа резонанса, усиливая или ослабляя определенные частоты в сигнале в зависимости от заданных параметров фильтра.

Для **обработки выбросов** применяются методы, основанные на статистических характеристиках данных. Например, метод межквартильного размаха (IQR) для определения границы выбросов и их последующего удаления, или метод, в котором значения, находящиеся за пределами порога в три стандартных отклонениях, считаются выбросами и могут быть заменены на среднее значение.

Методы машинного обучения также применяют для обнаружения и устранения шума и выбросов. Например, алгоритмы кластеризации, такие как K-means или DBSCAN, работают на принципе разделения данных на кластеры и обнаружении точек, которые не принадлежат ни одному кластеру, – выбросов.

Алгоритмы машинного обучения без учителя, такие как Isolation Forest, One-Class SVM и Local Outlier Factor, работают на основе предположения о том, что выбросы имеют отличные структурные характеристики от основной массы данных [2]. Isolation Forest (изолирующий лес) учитывает, что выбросы обычно имеют более короткие пути от корневого узла до них в дереве решений. Алгоритм строит несколько случайных деревьев, разбивая данные на подмножества. После построения леса выбросы можно определить по тому, сколько раз было нужно разделить их, чтобы добраться до них. Этот метод эффективен для обработки данных с большим количеством признаков и наличием различных типов выбросов. Он может использоваться как самостоятельный алгоритм для обнаружения аномалий или в сочетании с другими методами.

One-Class SVM (одноклассовая машина опорных векторов) стремится найти гиперплоскость, которая наилучшим образом разделяет «нормальные» данные от выбросов. Алгоритм обучается только на «нормальных» данных и затем пытается классифицировать новые точки данных как «нормальные» или выбросы. Метод используется в мониторинге кибербезопасности или обнаружении аномалий в процессах производства.

Local Outlier Factor (LOF, локальный фактор аномалии) вычисляет относительную плотность каждой точки данных по сравнению с ее соседями. Точки с низкой плотностью, отличающиеся от окружающих, могут быть классифицированы как выбросы. LOF широко используется для обнаружения аномалий в пространствах с высокой размерностью и сложной структурой данных.

Таким образом, эффективное устранение шума и выбросов в данных требует применения разнообразных методов, включая фильтрацию, статистические подходы и алгоритмы машинного обучения. Комбинирование различных методов позволяет добиться более точной и надежной предварительной обработки данных перед обучением модели.

Второй этап обработки данных – их **масштабирование –** помогает алгоритмам более эффективно и быстро сходиться к оптимальным решениям. Рассмотрим подробнее два основных метода масштабирования данных:

– нормализация (Min-Max Scaling) масштабирует значения признаков так, чтобы они находились в пределах определенного диапазона, обычно от 0 до 1;

– стандартизация (Z-score Scaling) центрирует данные относительно их среднего значения и масштабирует их так, чтобы они имели стандартное отклонение, равное единице.

Выбор между нормализацией и стандартизацией зависит от распределения данных, алгоритмов машинного обучения и требований задачи.

Работа с пропущенными значениями – важный этапом предварительной обработки данных перед анализом или применением алгоритмов машинного обучения. Пропущенные значения могут возникать из-за ошибок в сборе данных, недоступности информации и т. д. Обнаружить их можно с помощью методов Pandas в Python или аналогичных инструментов в других языках программирования. Простейший способ борьбы с пропущенными значениями – удаление строк или столбцов, содержащих пропуски, однако это может привести к потере значительного количества данных и искажению результатов анализа. Также можно заполнить пропущенные значения средними, медианными или модальными значениями соответствующих столбцов.

Для создания наиболее универсальных моделей машинного обучения необходимо среди множества признаков в данных отобрать наиболее значимые и информативные. Существует несколько **техник выбора признаков**, например случайный лес (Random Forest) или градиентный бустинг (Gradient Boosting), анализ корреляции между признаками, рекурсивное устранение признаков (Recursive Feature Elimination, RFE), взаимная информация (Mutual Information), линейная регрессия с L1 или L2 регуляризацией (Lasso и Ridge), статистические тесты (t-тест или анализ дисперсии (ANOVA)) и др. [3].

Техники агрегации данных используются для сведения данных к более высокому уровню или получения обобщенной информации. Это суммирование (Sum), усреднение (Mean или Average), определение медианы (Median) и моды (Mode), техника Min и Max, дисперсия и стандартное отклонение (Variance и Standard Deviation).

Выбор конкретного метода выбора признаков или агрегации данных зависит от типа данных, целей моделирования и характеристик признаков. Часто комбинируют несколько методов для достижения наилучших результатов.

Создание новых признаков на основе существующих – важный этап в анализе данных и построении моделей машинного обучения. Например, финансовые временные ряды часто имеют сезонность, цикличность и другие временные зависимости. Поэтому создание новых признаков на основе даты и времени может быть важным. Для преобразования ценовых данных можно использовать числовые методы. Например, разность между ценами на закрытие и открытие, скользящие средние цен и т. д. Эти признаки могут отражать тенденции и волатильность рынка. Признаки, связанные с волатильностью (например, стандартное отклонение цен) и объемом торгов (например, средний объем торгов за период), могут быть полезными для оценки рисков и предсказания движения ценовых рядов.

Можно создавать новые признаки, комбинируя различные финансовые показатели. Например, отношение капитализации к выручке, отношение долга к капиталу, коэффициенты доходности и т. д. Эти показатели помогают в анализе финансовых показателей компаний.

Технические индикаторы, такие как скользящие средние (SMA, EMA), индикаторы относительной силы (RSI), стохастический осциллятор и др., также могут быть использованы для создания новых признаков, отражающих технические аспекты торговли и трендов на рынке [4].

Методы уменьшения размерности, такие как PCA (Principal Component Analysis) или t-SNE (t-Distributed Stochastic Neighbor Embedding), создают новые признаки путем проекции исходных признаков в пространство меньшей размерности.

Заключение

Правильный подход к обработке данных и выбор наилучших методов анализа – ключевые факторы для успешного и надежного анализа данных и построения модели прогнозирования. Данное исследование поможет исследователям оптимизировать процесс подготовки данных и повысить результативность своих моделей.

Библиографический список

1. *Дьяченко Р. А., Косолапов П. А., Гура Д. А.* К вопросу об увеличении производительности машинного обучения на этапе выборки данных при решении задач классификации // Вестник Воронежского государственного университета. 2022. № 4. С. 146–155.

2. *Бабич В. Н., Пупцев Р. И.* Разработка методики предобработки данных и подготовка многомерного обучающего массива для алгоритмов прогнозиро-

вания, использующих методы машинного обучения и нейронные сети // Подготовка профессиональных кадров в магистратуре для цифровой экономики (ПКМ-2023). СПб.: СПбГУТ им. проф. М. А. Бонч-Бруевича, 2023. С. 243–247.

3. *Asaad Ja., Avksentieva E.* A Survey on Machine Learning Techniques for Software Engineering // Journal of Information Technologies and Computing Systems. 2023. № 4. P. 86–95.

4. Методы машинного обучения в задачах контроля криптовалютных транзакций / В. Г. Феклин, В. И. Соловьев, С. А. Корчагин, А. В. Царегородцев // Вопросы кибербезопасности. 2023. № 4(56). С. 2–11.