

МЕТОДИКА ПРОГНОЗИРОВАНИЯ РЕСУРСА УЗЛОВ И ДЕТАЛЕЙ ЖЕЛЕЗНОДОРОЖНОГО ПОДВИЖНОГО СОСТАВА ПРИ ПОМОЩИ МАШИННОГО ОБУЧЕНИЯ. ПРОВЕРКА И ОБРАБОТКА ДАННЫХ



К.А. Сергеев



О.И. Мироненко

Данная статья освещает важный этап обработки данных с прибора Calipri-Ferberg CW-40 (CALIPRI), а именно проверку и предобработку. Этот этап критически влияет на качество машинного обучения, поскольку точность данных непосредственно влияет на аналитические и модельные результаты. Рассматриваются основные этапы проверки и предобработки данных.

Ключевые слова: разработки технологических процессов, ремонтное производство, оптимизация транспортных операций

EDN: NWDOIA

В статье «Методика прогнозирования ресурса узлов и деталей железнодорожного подвижного состава при помощи машинного обучения» [1] показана общая схема системы, выдающая оценку ожидаемого пробега колеса, показанная на рис. 1.

При обработке данных с прибора CALIPRI важным этапом является их проверка и предобработка. Этот этап играет ключевую роль в обеспечении высокого качества машинного обучения, так как точность и достоверность данных непосредственно влияют на результаты анализа и моделирования. В данной статье рассмотрены шаги проверки и предобработки данных с прибора CALIPRI, направленные на устране-

ние неточностей, выбросов и пропусков. Проверка и предобработка данных делится на следующие шаги.

Исследование и визуализация данных

Данные с прибора CALIPRI имеют следующий вид:

$$(x_i, y_i)_{i=1}^N,$$

где x_i — набор признаков;

y_i — набор целевых переменных;

N — фиксированное количество признаков и целевых переменных.

Сергеев Константин Александрович, доктор технических наук, доцент, доцент кафедры «Нетяговый подвижной состав» Российского университета транспорта (РУТ (МИИТ)). Область научных интересов: технологическая подготовка производства вагоноремонтных предприятий. Автор 141 научной работы, в том числе двух монографий.

Мироненко Олег Игоревич, кандидат технических наук, доцент кафедры «Нетяговый подвижной состав» Российского университета транспорта (РУТ (МИИТ)). Область научных интересов: динамика систем, организация вагонной отрасли, дистанционные образовательные технологии. Автор 29 научных работ, в том числе четырех учебных пособий.

Кривич Ольга Юрьевна, кандидат технических наук, доцент, начальник Учебного отдела учебно-методического многофункционального центра Российской открытой академии транспорта Российского университета транспорта (РОАТ РУТ (МИИТ)). Область научных интересов: повышение эффективности технологической подготовки вагоноремонтного производства, оценка потребительских свойств продукции железнодорожного транспорта. Автор 66 научных работ.

Мироненко Никита Олегович, студент Московского государственного технического университета им. Н.Э. Баумана (МГТУ им. Н.Э. Баумана). Область научных интересов: организация вагонной отрасли, искусственный интеллект, нейронные сети, предсказательное моделирование. Автор более 30 научных работ. Имеет один патент на изобретение.

Первым этапом при анализе данных является их исследование и визуализация. Это позволяет получить общее представление о структуре данных, их распределении и выявить наличие аномалий. Визуализация включает в себя построение гистограмм и диаграмм рассеяния, показанные на рис. 2. Эффективным инструментом для визуализации распределения данных является диаграмма размаха, которая широко используется для первичной оценки медианы, разброса данных через верхний и нижний квартили, а также для выявления возможных выбросов — аномальных значений [2].

Обработка пропусков

Пропуски данных возникают по разным причинам, таким как технические сбои, ошибки измерений или по другим факторам. Обработка пропусков включает в себя заполнение или удаление отсутствующих значений. Заполнение может быть выполнено с использованием статистических метрик: средних значений, медиан и т.д. Одним из методов заполнения пропусков в системе является линейная интерполяция, которая

позволяет заполнять пропущенные значения, если известны соседние значения.

Обработка выбросов

Выбросы представляют собой значения, сильно отличающиеся от остальных значений в выборке, которые могут быть результатом ошибок измерения, аномалий и т.д. [2]. Обработка выбросов включает в себя применение статистических методов для определения граничных значений, за пределами которых данные могут считаться аномальными. Выбросы удаляются, заменяются и преобразовываются таким образом, чтобы они не искажали результат анализа и прогнозирования. Для определения выбросов в системе используются статистические методы, такие как метод межквартильного размаха (IQR), стандартизированные значения (z-оценки) и др. Межквартильный размах выявляет выбросы путем вычисления разницы между третьим и первым квартилями выборки. Значения, находящиеся за пределами установленных граничных значений, считаются потенциальными выбросами. Обработка выбросов включает в себя различные под-



Рис. 1. Общая схема системы

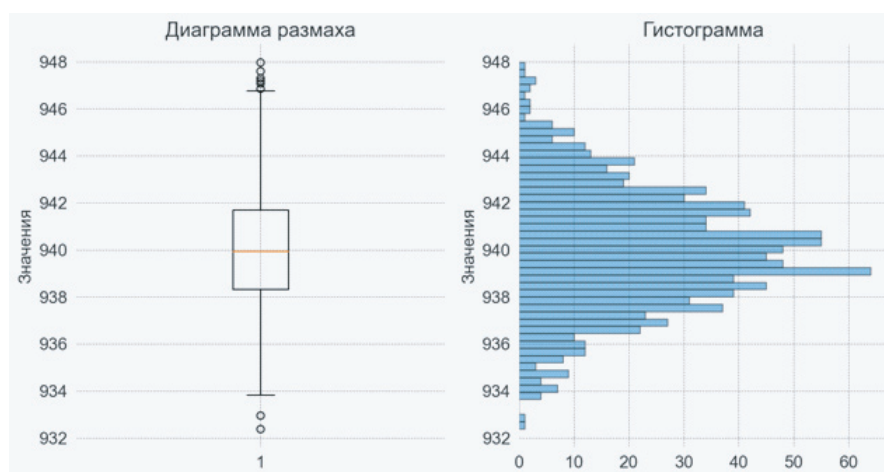


Рис. 2. Диаграмма размаха и гистограмма

ходы. Выбросы удаляются из выборки, если они являются результатом ошибок или аномалий и не имеют статистической значимости. Возможно преобразование данных, например, с использованием логарифмической шкалы, чтобы сделать выборку более нормально распределенной и уменьшить влияние выбросов на статистические оценки.

Нормализация и преобразование

Нормализация данных представляет собой процесс приведения признаков в наборе данных к стандартизированному диапазону значений, для получения наиболее точных результатов прогнозирования машинным обучением [3;4]. Цель нормализации заключается в обеспечении однородности масштабов признаков. На этапе нормализации данные преобразуются таким образом, что они располагаются в заданном интервале, позволяя избежать проблем, связанных с разнообразием масштабов признаков, обеспечивая при этом стабильность и эффективность работы алгоритмов машинного обучения [5]. Последовательное применение нормализации дает возможность достичь более стабильной сходимости алгоритмов, повышая качество и точность прогнозирования. В рассматриваемой системе используется следующая формула для нормализации значений:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}},$$

где x_{norm} — нормализованное значение;

x — исходное значение;

$x_{\min}$ — минимальное значение;

$x_{\max}$ — максимальное значение.

Дубликаты и уникальность

Наличие дубликатов в данных искажает результаты анализа и прогнозирования. Дубликаты возникают из-за ошибок ввода или технических сбоев и т.д. Удаление дубликатов позволяет сохранить только уникальные записи и предотвратить их влияние на обучение модели.

Выводы

Проверка и предобработка данных с прибора CALIPRI являются неотъемлемой частью перед применением алгоритмов машинного обучения. Правильность выполнения этих шагов влияет на качество и достоверность результатов анализа и прогнозирования. Эффективная обработка данных позволяет снизить влияние шума, улучшить стабильность моделей и повысить точность прогнозов, что в конечном итоге способствует более качественному и надежному прогнозированию. 

Литература

1. Методика прогнозирования ресурса узлов и деталей железнодорожного подвижного состава при помощи машинного обучения / К. А. Сергеев, О. И. Мироненко, М. В. Козлов, Н. О. Мироненко. - Текст : непосредственный // Транспортное машиностроение. - 2023. - № 11 (23). - С. 42-49.
2. Плас, Джейк Вандер. Python для сложных задач. Наука о данных и машинное обучение. Руководство / Плас Джейк Вандер. - Москва : Питер, 2018. - 759 с. - Текст : непосредственный.
3. Сальникова, К. В. Анализа массива данных с помощью инструмента визуализации «Ящик с усами» / К. В. Сальникова. - Текст : непосредственный // Universum: экономика и юриспруденция. - 2021. - №6(81). - С. 11-17. - DOI: 10.32743/UniLaw.2021.81.6.11778. - EDN: APSOIG.
4. T. Beysolow II. Applied Reinforcement Learning with Python: With OpenAI Gym, Tensorflow, and Keras. Apress, 2019. - ISBN 14842512615.
5. Шарден, Б. Крупномасштабное машинное обучение вместе с Python / Б. Шарден, Л. Массарон, А. Боскетти ; перевод А. В. Логунов. - Москва : ДМК Пресс, 2018. - 358 с. - ISBN 978-5-97060-506-6. - Текст: электронный. - URL: <https://www.iprbookshop.ru/124994.html> (дата обращения: 21.05.2024).