УДК 004.67

ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ДАННЫХ ДЛЯ МАШИННОГО ОБУЧЕНИЯ

Акимов А.А., Валитов Д.Р., Кубряк А.И.

Стерлитамакский филиал Башкирского государственного университета, Стерлитамак, e-mail: denisvalitof@yandex.ru

Данные – это не что иное, как актив в современном мире. Данные в настоящее время сильно искажены несоответствиями, шумом, неполной информацией и пропущенными значениями. Они агрегируются из разнообразных источников с использованием методов интеллектуального анализа данных и хранилищ. Предварительная обработка данных служит основой для достоверного анализа. Это необходимый шаг в построении анализа оперативных данных, учитывая недостатки в их качестве. Также предобработка относится к основному набору методов для повышения качества исходных данных, таких как очистка, нормализация, отбор необходимых признаков и экземпляров. Так как полученные данные представлены в необработанном виде, обучение модели с их использованием может оказаться недостижимым. Эта статья представляет собой обзор методов предварительной обработки для анализа используемых данных. Также в ней рассмотрены этапы предварительной обработки данных, изучены виды признаков в машинном обучении, произведен обзор категориальных переменных. Продемонстрированы пять этапов предобработки данных, рассмотрены виды признаков при обработке данных для машинного обучения, описаны категориальные признаки и приведены два вида примеров категориальных переменных.

Ключевые слова: данные, обработка, категориальные переменные, генерация признаков, очистка данных

DATA PREPROCESSING FOR MACHINE LEARNING

Akimov A.A., Valitov D.R., Kubryak A.I.

Sterlitamak branch of Bashkir State University, Sterlitamak, e-mail: denisvalitof@yandex.ru

Data is nothing but an asset in the modern world. The data is currently heavily distorted by inconsistencies, noise, incomplete information and missing values. They are aggregated from a variety of sources using data mining and storage methods. Preliminary data processing serves as the basis for reliable analysis. This is a necessary step in building an analysis of operational data, given the shortcomings in their quality. Also, preprocessing refers to the main set of methods to improve the quality of the source data, such as cleaning, normalization, selection of necessary features and instances. Since the data obtained is presented in raw form, training the model using them may not be achievable. This article is an overview of preprocessing methods for analyzing the data used. The stages of data preprocessing are also considered, the types of features in machine learning are studied, and a review of categorical variables is made. Five stages of data preprocessing are demonstrated, the types of features in data processing for machine learning are considered, categorical features are described and two types of examples of categorical variables are given.

Keywords: data, processing, categorical variables, feature generation, data cleaning

Предварительная обработка данных в машинном обучении – это важный шаг, который помогает повысить качество данных. Предобработка данных в машинном обучении относится к технике подготовки необработанных данных с целью сделать их пригодными для построения и обучения моделей машинного обучения. Иными словами, это метод интеллектуального анализа данных, который преобразует необработанные данные в понятный и читаемый формат.

Предварительная обработка данных является одним из основных этапов, от качества выполнения которого зависит получение качественных результатов процесса анализа данных. Без подготовки данных не обходится ни один нейросетевой метод. Как правило, при описании различных нейроархитектур предполагается, что данные для обучения уже представлены в том виде, в котором требует нейросеть, однако на практике дела обстоят совсем иначе, именно этап предобработки данных может занимать большую часть времени, отведенного на проект в целом. Результат обучения нейросети также может зависеть от того, в каком виде представлена информация для ее обучения. Таким образом, предварительная обработка данных позволяет повысить качество как интеллектуального анализа данных, так и самих данных.

Цель исследования – рассмотрение этапов предварительной обработки данных и основ разработки признаков, а также обзор конструирования категориальных признаков.

Материалы и методы исследования

При создании модели машинного обучения предварительная обработка данных служит первым шагом. Как правило, реальные данные являются неполными, непоследовательными, могут содержать ошибки или выбросы, а также в них могут отсутствовать

конкретные значения или атрибуты. Именно здесь используется предобработка данных: она помогает очищать, форматировать и упорядочивать необработанные данные, тем самым делая их готовыми к работе с моделями машинного обучения.

Предобработка данных состоит из пяти этапов:

− очистка данных, которая направлена на повышение качества данных за счет присваивания пропущенных значений и удаления выбросов;

− сокращение объема данных, которое уменьшает объем данных и, следовательно, снижает связанные с ними вычислительные мощности;

− масштабирование данных – направлено на преобразование исходных данных в аналогичные диапазоны для прогнозного моделирования;

− преобразование, целью которого является организация исходных данных в подходящие форматы для различных алгоритмов интеллектуального анализа данных;

− разделение, которое делит весь набор данных на различные подмножества для более глубокого анализа.

Существует два общих способа обработки отсутствующих значений при построении оперативных данных. Первый – просто отбросить выборки данных с пропущенными значениями, так как большинство алгоритмов получения данных не могут обрабатывать данные с пропущенными значениями. Такой метод применим только тогда, когда доля отсутствующих значений незначительна. Второй – применение методов интерполяции пропущенных значений для замены пропущенных данных значениями, полученными в результате расчетов.

Как показано на рисунке 1, распространенные методы интерполяции отсутствующих значений можно разделить на две группы, т.е. одномерные и многомерные методы.

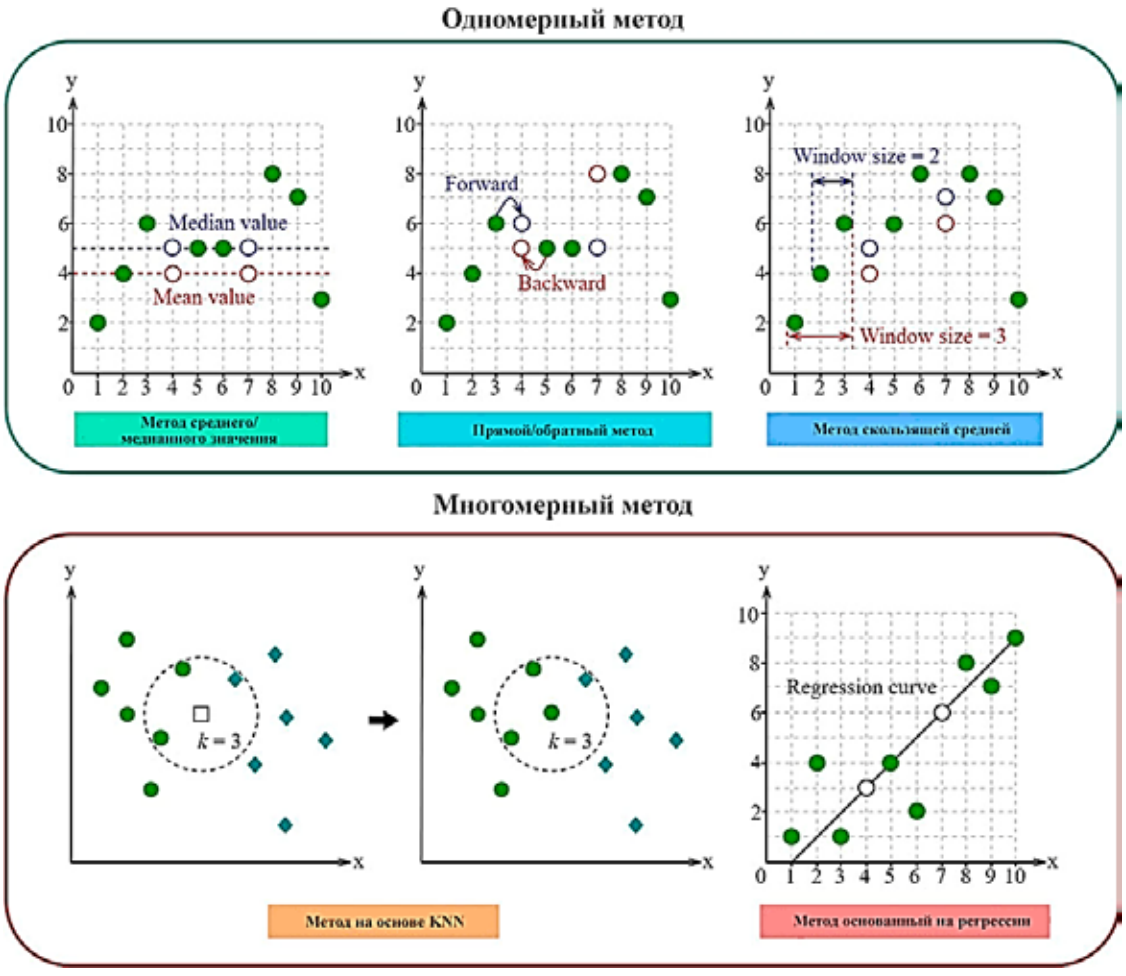

Рис. 1. Методы интерполяции пропущенных значений для построения данных

К первым относятся средняя интерполяция, прямая или обратная интерполяция и методы скользящего среднего. В этом случае недостающие значения выводятся на основе характеристик данных только одной переменной и поэтому называются одномерными методами. Метод интерполяции среднего или медианы заменяет недостающие значения средним или медианой этой переменной. Прямой или обратный метод просто заменяет отсутствующее значение предыдущим или следующим измерением данных. Эти два метода просты в реализации, но не учитывают временные корреляции на временных этапах и могут не сделать правильную замену данных [1].

В основном используются два метода обнаружения выбросов – статистические методы и методы, основанные на кластеризации.

Сокращение данных обычно проводится в двух направлениях, т.е. по строкам для сокращения выборки данных и по столбцам для сокращения переменных данных. Для сокращения данных по строкам могут применяться различные методы выборки данных, такие как случайная и стратифицированная выборка.

Существуют три основных метода сокращения переменных данных по столбцам. Первый заключается в использовании знаний о домене для прямого отбора интересующих переменных. Второй – использование статистических методов отбора признаков для выбора важных переменных для дальнейшего анализа. Третий – использование методов извлечения признаков для построения полезных признаков для анализа данных.

Масштабирование данных часто необходимо для обеспечения достоверности прогностического моделирования, особенно когда входные переменные имеют различные масштабы. Нормализация max-min (т.е., $x'=x-xmin/xmax-xmin$) и стандартизация z-score (т.е., $x'=x-\mu/\sigma$) – два наиболее широко используемых метода, где min(x) и max(x) означают минимум и максимум переменной x, значения переменной μ – среднее значение, а σ – стандартное отклонение.

Метод нормализации max-min чувствителен к выбросам данных, поскольку их присутствие может резко изменить диапазон данных. В отличие от него метод стандартизации z-score менее подвержен влиянию выбросов. Он обычно используется для реформирования переменной, чтобы она была нормально распределена со средним значением, равным нулю, и стандартным отклонением, равным единице. Теоретически, нормализация по z-score работает лучше всего, когда данные нормально распределены. Нормализация по методу max-min рекомендуется, когда эксплуатационные данные не соответствуют нормальному распределению и не содержат явных выбросов. Другой тип метода масштабирования данных может изменять структуры данных. Например, исходные данные могут быть отображены в новое пространство с помощью определенных математических функций, таких как логарифмическая, сигмоидальная функции или арктангенс. Такие методы часто используются для минимизации дифференциалов в переменных данных [2].

Преобразование данных в основном применяется для преобразования числовых данных в категориальные для обеспечения совместимости с алгоритмами интеллектуального анализа данных. Методы равной ширины и равной частоты широко используются благодаря своей простоте. Количество интервалов обычно предопределяется пользователем на основе знаний о предметной области. По сравнению с методом равной ширины метод равной частоты менее чувствителен к выбросам [3].

Преобразование данных также может применяться для преобразования категориальных переменных в числовые для облегчения разработки моделей прогнозирования. Для этой цели широко используется метод однократного кодирования, при котором для категориальной переменной с L уровнями создается матрица из L – 1 столбцов [4]. Все признаки могут быть следующих видов:

− бинарные, которые принимают два значения (да/нет, 0/1, true/false);

− номинальные, которые имеют конечное количество уровней. Также они могут быть упорядоченными и неупорядоченными;

− количественные значения в диапазоне от $-\infty$ до $+\infty$.

Признак – это переменная, которая описывает отдельную характеристику объекта. В табличном представлении выборки признаки – это столбцы таблицы, а объекты – строки. Входные, независимые переменные для модели машинного обучения называются предикторами, а выходные, зависимые – целевыми признаками.

Признаки могут извлекаться из данных любого типа, в том числе из текста, изображений и геоданных. При обработке текстовой информации сначала выполняется ее токенизация, а затем лемматизация и цифровизация – перевод слов в числовые вектора. В случае изображений часто анализируется не только содержание картинки как набора пикселей различного цвета, но и метаданные графического файла: дата

съемки, разрешение, модель камеры и т.д. Географические данные чаще всего представлены в виде адресов или пар [5].

Разработка признаков – очень важный аспект машинного обучения и науки о данных, и его нельзя игнорировать. Основная цель разработки признаков – получить наилучшие результаты от алгоритмов. В науке о данных производительность модели зависит от предварительной обработки и обработки данных. Если модель построена без обработки данных, то точность будет составлять около 70%. Применив генерацию признаков к той же модели, производительность можно повысить на несколько десятков процентов. Проще говоря, благодаря генерации признаков улучшается производительность модели.

Выбор признаков – это не что иное, как выбор необходимых независимых признаков. Выбор важных независимых признаков, которые имеют большую связь с зависимым признаком, поможет построить хорошую модель. Существует несколько методов отбора признаков [6].

В случае одномерного массива статистические тесты могут использоваться для отбора независимых признаков, которые имеют наиболее сильную связь с зависимым признаком. Метод SelectKBest может быть использован с набором различных статистических тестов для выбора определенного количества признаков (рис. 2). Признак, имеющий наивысший балл, будет более связан с зависимым признаком, и эти признаки будут выбраны для модели.

Метод ExtraTreesClassifier помогает определить важность каждого независимого признака с зависимым признаком. Важность признака дает оценку для каждого признака данных: чем выше оценка, тем важнее или релевантнее признак текущей выходной переменной (рис. 3).

Тепловая карта – это графическое представление двумерных данных (рис. 4). Здесь каждое значение данных представлено в матрице. Необходимо построить парный график между всеми независимыми и зависимыми функциями, в итоге получится отношение между двумя признаками. Если отношение между независимой и зависимой функциями меньше 0,2, тогда эта независимая функция выбирается для построения модели.

```
In [43]: from sklearn.feature_selection import SelectKBest
         from sklearn.feature_selection import chi2
         bestfeatures = SelectKBest(score_func=chi2, k=10)
         fit = bestfeatures.fit(x,y)
         dfscores = pd.DataFrame(fit.scores_)
         dfcolumns = pd.DataFrame(x.columns)
         featureScores = pd.concat([dfcolumns,dfscores],axis=1)
         featureScores.columns = ['Specs','Score']
         featureScores
```

```
In [47]: featureScores
```

Out[47]:

	Specs	Score
0	City_Code	0.122643
1	Region_Code	74.805013
2	Accomodation_Type	0.756115
3	Reco_Insurance_Type	3.965972
4	Upper_Age	2.612166
5	Lower_Age	1.572930
6	Is_Spouse	0.632296
7	Health Indicator	0.453390
8	Holding_Policy_Duration	7.662646
9	Holding_Policy_Type	0.836189
10	Reco_Policy_Cat	1894.032997
11	Reco_Policy_Premium	9575.065324
12	diff_age	0.039347

Рис. 2. Пример одномерного отбора

```
In [192]: feat_importances = pd.Series(model.feature_importances_, index=x.columns)
          feat_importances.nlargest(10).plot(kind='barh')
          plt.show()
```

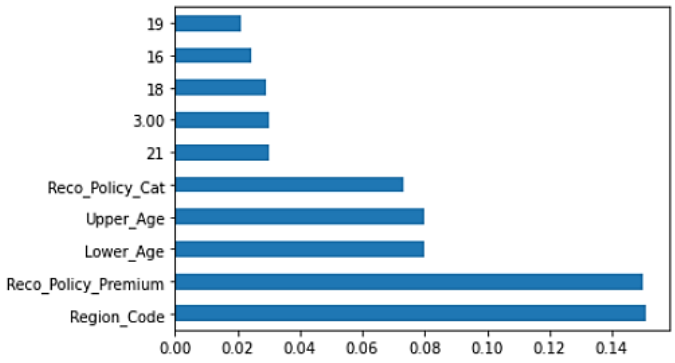

Рис. 3. Пример метода ExtraTreeClassifier

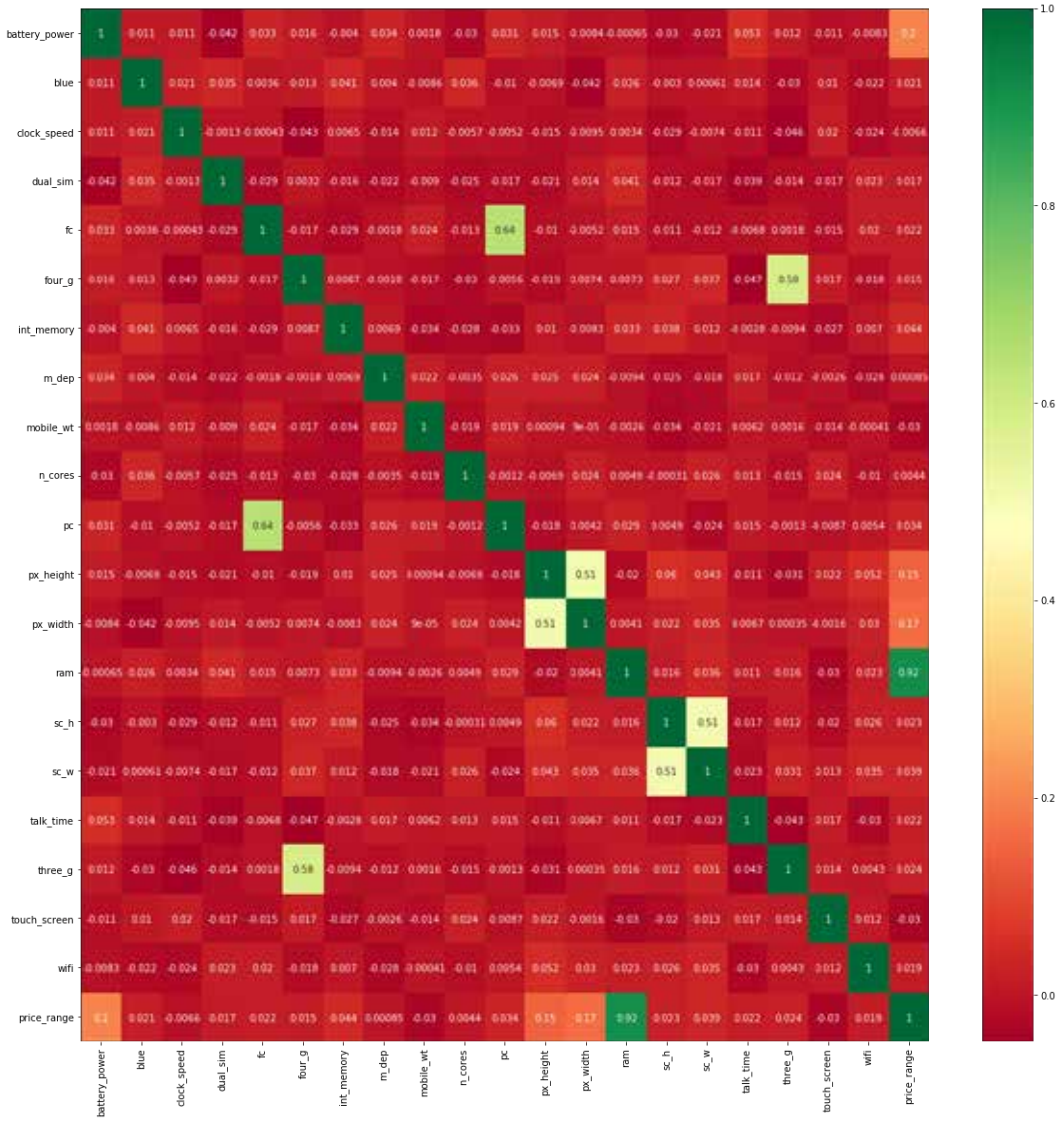

Рис. 4. Пример тепловой карты

Категориальные данные – это тип данных, который используется для группировки информации с похожими характеристиками, в то время как числовые данные – это тип данных, выражающих информацию в виде чисел. Примером категориальных данных может служить пол человека [7].

Большинство алгоритмов машинного обучения не могут работать с категориальными переменными, если их не преобразовать в числовые значения. Производительность многих алгоритмов даже зависит от того, как закодированы категориальные переменные.

Категориальные переменные можно разделить на две категории:

− номинальные, которые не имеют определенного порядка;

− порядковые, между значениями которых существует определенный порядок.

Заключение

Предобработка данных является важнейшим этапом построения моделей машинного обучения, она занимает большую часть времени, так как от подготовки данных зависит корректность будущей модели. В случае ошибки при первичном анализе возможны переобучение модели или утечка данных, которые могут нарушить корректность работы модели. Таким образом, предварительная обработка данных позволяет значительно повысить качество как самих данных, так и результата анализа.

Список литературы

1. Быков К.В. Особенности предобработки данных для применения машинного обучения // Молодой ученый. 2021. № 53 (395). С. 1-4. [Электронный ресурс]. URL: https://moluch.ru/archive/395/87491/ (дата обращения: 22.04.2022).

2. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. М.: ДМК Пресс, 2015. 400 с.

3. Близнюк Б., Васильева Л., Стрельников П., Ткачук Д. Современные методы обработки естественного языка // Вестник Харьковского национального университета им. Каразина. Серия «Математическое моделирование. Информационные технологии. Автоматизированные системы управления». 2017. № 36. С. 14-26.

4. Kumar D. Introduction to Data Preprocessing in Machine Learning. [Электронный ресурс]. URL: https://towardsdatascience.com/introduction-to-data-preprocessing-in-machine-learning-a9fa83a5dc9d (дата обращения: 22.04.2022).

5. Акимов А.А., Мустафина С.А. Обзор современных методов искусственного интеллекта по распознаванию девиантного поведения индивида // Вестник Технологического университета. 2020. Т. 23. № 8. С. 69-79.

6. Mustafina S., Akimov A., Plotnikova A. Comparison of semantic convolution neural networks on the example of crack segmentation in asphalt images. International Journal of Computing. 2021. Т. 19. № 3. С. 415-423.

7. Hamza A. Effective Data Preprocessing and Feature Engineering. [Электронный ресурс]. URL: https://becominghuman.ai/effective-data-preprocessing-and-feature-engineering-452d3a948262 (дата обращения: 22.04.2022).