УДК 004.8

ОБРАБОТКА ДАННЫХ ДЛЯ МАШИННОГО ОБУЧЕНИЯ

М.А. ШЕВЦОВ
(*Представлено: канд. физ.-мат. наук, доц. Ю.Ф. ПАСТУХОВ*)

Предварительная обработка и очистка данных являются важными задачами, которые обычно необходимо выполнить, прежде чем набор данных можно будет эффективно использовать для машинного обучения. В статье рассматриваются способы повышения эффективности данных.

Введение. Необработанные данные зачастую искажены и ненадежны, и в них могут быть пропущены значения. Использование таких данных при моделировании может приводить к неверным результатам. Эти задачи являются частью процесса обработки и анализа данных группы и обычно подразумевают первоначальное изучение набора данных, используемого для определения и планирования необходимой предварительной обработки. Данная статья направлена на описание разных проблем и методов их решений при обработке данных для машинного бучения.

Основной раздел. Реальные данные собираются для последующей обработки из разных источников и процессов. Они могут содержать ошибки и повреждения, негативно влияющие на качество набора данных. Типичные проблемы с качеством данных:

–**Неполнота:** данные не содержат атрибутов, или в них пропущены значения.

–**Шум:** данные содержат ошибочные записи или выбросы.

–**Несогласованность:** данные содержат конфликтующие между собой записи или расхождения.

Качественные данные — это необходимое условие для создания качественных моделей прогнозирования. Чтобы избежать появления ситуации «мусор на входе, мусор на выходе» и повысить качество данных и, как следствие, эффективность модели, необходимо провести мониторинг работоспособности данных, как можно раньше обнаружить проблемы и решить, какие действия по предварительной обработке и очистке данных необходимы.

Стандартные методы мониторинга работоспособности данных:

– Количество записей.

– Количество атрибутов (или компонентов);

– Типы данных атрибута (номинальные, порядковые или непрерывные).

– Количество пропущенных значений.

– Правильность формата данных.

1) Несогласованные записи данных (проверьте допустимость диапазона значений).

При обнаружении проблем с данными необходимо выполнить обработку, которая зачастую включает очистку пропущенных значений, нормализацию данных, дискретизацию, обработку текста для удаления и/или замены внедренных символов, которые могут влиять на выравнивание данных, смешанные типы данных в общих полях и пр. Главные задачи предварительной обработки данных: **Очистка данных** — восполнение пропущенных значений, обнаружение и удаление искаженных данных и выбросов.

При работе с пропущенными значениями лучше сначала определить причину их появления в данных, что поможет решить проблему.

Методы обработки пропущенных значений:

– **Удаление:** удаление записей с пропущенными значениями.

– **Фиктивная подстановка** — замена пропущенных значений фиктивными, например, подстановка значения unknown (неизвестно) вместо категориальных или значения 0 вместо чисел.

– **Подстановка среднего значения:** пропущенные числовые данные можно заменить средним значением.

– **Подстановка часто используемого элемента:** пропущенные категориальные значения можно заменить наиболее часто используемым элементом.

– **Подстановка по регрессии:** использование регрессионного метода для замены пропущенных значений регрессионными.

2) Преобразование данных — нормализация данных для снижения измерений и искажений. Нормализация данных позволяет масштабировать числовые значения в указанном диапазоне. Ниже представлены распространенные методы нормализации данных.

– **Нормализация по методу минимакса:** линейное преобразование данных в диапазоне, например, от 0 до 1, где минимальное и максимальное масштабируемые значения соответствуют 0 и 1 соответственно.

– **Нормализация по Z-показателю:** масштабирование данных на основе среднего значения и стандартного отклонения: деление разницы между данными и средним значением на стандартное отклонение.

– **Десятичное масштабирование:** масштабирование данных путем удаления десятичного разделителя значения атрибута.

3) Уплотнение данных — создание выборки данных или атрибутов для упрощения обработки данных.

Существуют различные методы, с помощью которых вы можете уменьшить размер данных для упрощения обработки данных. В зависимости от размера данных и домена вы можете применить такие методы:

1. **Выборка записей:** создание выборки записей данных и выбор репрезентативного подмножества из общего набора данных.
2. **Выборка атрибутов:** выбор в данных набора важнейших атрибутов.
3. **Агрегирование:** разделение данных на группы и хранение числовых значений для каждой группы.

4) Дискретизация данных — преобразование непрерывных атрибутов в категориальные. Данные можно дискретизировать, преобразовав непрерывные значения в номинальные атрибуты или интервалы. Это можно сделать несколькими способами.

1. **Группирование равной ширины:** разделение диапазона всех возможных значений атрибута в группы (N) одинакового размера с последующим присвоением значений, относящихся к ячейке с соответствующим номером.
2. **Группирование равной высоты:** разделение всех возможных значений атрибута в группы (N), содержащие одинаковое количество экземпляров, с последующим присвоением значений, относящихся к ячейке с соответствующим номером.

5) Очистка текста — удаление внедренных символов, которые могут нарушать выравнивание данных, например, внедренных символов табуляции в файле с разделителем-табуляцией, внедренных новых линий, которые могут разбивать записи, и пр.

3. **Текстовые поля в табличных данных** могут содержать символы, сбивающие выравнивание столбцов или границы записей (или и то и другое вместе). Например, табуляции, внедренные в файл с разделителем-табуляцией, могут сбить выравнивание столбцов, а внедренные символы новой строки могут разорвать линии записей. Неправильная кодировка текста приводит при его чтении или записи к потере информации, появлению нечитаемых символов, например, нуль-символов, и может также помешать разбору текста. Чтобы очистить текстовые поля, исправить выравнивание и извлечь структурированные текстовые данные из неструктурированных или полу-структурированных, могут потребоваться тщательные разбор и редактирование текста.

Заключение. Предварительная обработка данных важная часть в процессе работы с алгоритмами машинного обучения, помогая подготовить данные для обучения в дальнейшем получая намного более качественные результаты чем с данными которые были поставлены в чистом виде.

ЛИТЕРАТУРА

1. microsoft.com [Электронный ресурс]. Режим доступа: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/prepare-data: Дата доступа: 19.09.2019.
2. kdnuggets.com [Электронный ресурс]. Режим доступа: https://www.kdnuggets.com/2018/12/six-steps-master-machine-learning-data-preparation.html: Дата доступа: 19.09.2019.
3. machinelearningmastery.com [Электронный ресурс]. Режим доступа: https://machinelearningmastery.com/how-to-prepare-data-for-machine-learning/: Дата доступа: 19.09.2019.