энергетического баланса в пользу большего использования местных и возобновляемых ресурсов [2].

В этой связи наилучшими направлениями политики энергетической диверсификации и повышения энергоэффективности являются большее использование местных источников энергии, а также общее повышение конкурентоспособности экономики путем реализации структурных реформ, в первую очередь в энергетическом секторе.

Список использованных источников

1. Цилибина, В.М. Энергоэффективность экономики: методология и практика / В.М. Цибилина ; Ин-т экономики НАН Беларуси. – Минск : Беларуская навука, 2021. – 215 с.

2. Михадюк, М. В. Экологические проблемы энергоснабжения в современных условиях / М. В. Михадюк, Е. И. Кузнецова // Экономический рост Республики Беларусь: глобализация, инновационность, устойчивость : программа XII Международной научно-практической конференции, Минск,16 мая 2019 г. / М-во образования Респ. Беларусь, Белорус. гос. экон. ун-т [и др.]. – Минск : БГЭУ, 2019. – С. 61–62.

УДК 004.032.26

А.Н. Мущук, Д.В. Шиман
Белорусский государственный технологический университет
Минск, Беларусь

ПОДГОТОВКА ДАННЫХ ДЛЯ ОБУЧЕНИЯ НЕЙРОННОЙ СЕТИ

Аннотация. *Статья посвящена обзору методов, сгруппированных по определенным признакам, применяемых при подготовке данных для обучения нейронной сети. А также даст краткие примеры ситуаций, когда тот или иной метод уместно применить.*

A.N. Mushchuk, D.V. Shiman
Belarusian State Technological University
Minsk, Belarus

DATA PREPARATION FOR NEURAL NETWORK TRAINING

Abstract. *The article is devoted to an overview of the methods grouped according to certain criteria used in preparing data for training a neural network. It will also give brief examples of situations where this or that method is appropriate to apply.*

Вопросы подготовки данных имеют весомое значение в машинном обучении. Процесс обучения, тестирование, и в целом работа нейронной сети основывается в первую очередь на информации. Если исходные данные некорректны, то логично предположить, что на выходе будут такие же некорректные результаты. Кроме того, даже хорошо обученная сеть, при обработке плохих входных значений, покажет на выходе не то, что должна [1].

Процесс подготовки данных можно условно разделить на две части: Предварительная обработка данных и сам Data mining [2]. В рамках доклада рассмотрим первый вариант.

Предварительная обработка включает в себя:

1) Очистка данных. Типичными проблемами, при которых требуется очистка являются Неполнота (частичное либо полное отсутствие некоторых данных), Шум (данные содержат некорректные значения), Несогласованность (данные противоречат друг другу, либо дублируются).

2) Преобразование данных. Подразумевает изменение данных под конкретные нужды. Как правило, речь идет не об изменении значений атрибутов, а скорее о приведении в формат, удобный для обработки. Включает в себя нормализацию, дискретизацию, уменьшение объема, в некоторых случаях имеет смысл скорректировать данные, путем их обрезки или округления.

Помимо них, также к предварительной обработке относятся интеграция и выборка данных, но, поскольку интеграция уместна не всегда, а выборка выполняется под каждую модель индивидуально, рассмотрим методы очистки и преобразования данных.

Начнем с неполноты данных. Дополнение отсутствующей информации один из самых сложных и важных процессов на этапе очистки данных, поскольку несогласованные или некорректные данные можно исправить или удалить без существенных потерь для набора, в то время, как отсутствующая информация - это те атрибуты, которые необходимы для корректной работы нейронной сети, но которых нет.

Если исходный набор имеет большой размер, и неполных записей немного, либо отсутствует информация, которая не критически важна для работы нейросети, то логичным и простейшим решением будет просто их не использовать, что избавит нас от траты большого количества времени, но, если речь идет о данных, скажем, идущих

последовательно (например, ежегодная статистика), или данных, связанных друг с другом (перечень расходов по категориям), то восстановление является необходимым.

В некоторых случаях, имеет смысл сделать фиктивную подстановку. Когда данные не восстанавливаются, а просто вместо них подставляется некоторое значение, говорящее о том, что они неполные (потому метод и называется фиктивной подстановкой). В основном это применяется для атрибутов, имеющих второстепенное значение.

Как правило, первое, что необходимо сделать для восстановления отсутствующих данных - это проверить актуальность, имеющейся информации. При использовании готовых наборов данных, может получиться так, что часть атрибутов попросту устарела, и, проведя их повторный поиск, можно будет получить все нужные значения, не прибегая к более сложным способам.

Следующим методом является подстановка среднего значения. Наиболее простой метод, однако из-за своей простоты, одновременно не является широко применяемым. Поскольку не учитывает контекст данных. Другими словами, чем от большего числа факторов зависит значение атрибута, тем Примером может служить заполнение атрибута среднесуточной температуры средним значением с предыдущих лет.

В некотором роде эволюцией подстановки среднего значения, является подстановка часто используемого элемента. Чаще всего данный метод используется для значений, которые описывают исследуемый предмет, но при этом, не могущие выражаться в количественной форме (но это не говорит, о том, что для их описания нельзя использовать числа). Примерами является цвет, возраст, наименование и т.д. Другими словами, это те элементы, по которым можно проводить разделение на группы или категории.

Последним способом заполнить недостающие данные является подстановка по регрессии. Данный метод является способом предсказать некоторое значение, по значениям, которые непосредственно на него влияют. Соответственно, предсказываемое значение называют зависимым (он же регрессант), а влияющее на него независимым (он же предиктор). Подстановка по регрессии, по сути, выдает некоторое значение, на основе общей тенденции его изменения. Данный метод применим, прежде всего, к данным, которые могут содержать в себе ошибки изменения. Простейшим примером будет статистика рождаемости. При одном и том же количестве человек, количество новорожденных каждый год будет отличаться. Кроме того, на рождаемость могут влиять такие предикторы как общее экономическое состояние страны, социальные явления и т.д. Но тем не

менее, зная все предикторы, ее можно относительно точно предсказывать.

Необходимо понимать, что все вышеописанные методы, за исключением обновления информации, не дадут абсолютно точных значений. Тем не менее, правильно их используя, можно дополнить данные с минимальной погрешностью, тем самым получив дополнительную информацию для обучения и работы нейронной сети.

Перейдем к очистке данных от некорректных значений. Под некорректными значениями, подразумеваются значения, выходящие за заданные интервалы, неверные данные и выбросы.

Начнем с последнего. Выбросы - это данные, отклоняющиеся от основной массы. К примеру, если взять среднюю скорость на дорогах Минска, то значения по МКАДу или проспекту Независимости будут сильно отличаться от значений по дорогам города.

Обнаружение выбросов проще всего сделать, в случае числовых значений, построением графика, а в случае категориальных – созданием гистограммы. В зависимости от специфики проекта, обработка будет сводиться либо к их удалению, либо к коррекции. В некоторых случаях наилучшим решением будет оставить их как есть [3].

Несогласованные данные чаще всего порождаются из-за невнимательности при заполнении. И, в отличии от предыдущих методов очистки, чаще всего этим вопросом занимаются встроенные в СУБД средства, либо специализированные системы для статистического анализа. Однако, если есть сомнения в полной очистке данных, имеет смысл написать скрипт, который по определенным условиям проанализирует информацию, и исправит ее.

Перейдем к преобразованию данных.

Первым рассмотренным методом будет нормализация. Цель ее в том, чтобы преобразовать различные числовые значения, имеющие разные диапазоны и единицы измерения к абстрактному единому виду, для удобства сравнения и нахождения схожести объектов в дальнейшем. Самым часто используемым способом нормализации является нормализация по методу минимакса. Этот способ подразумевает преобразование входных числовых значений в определенный диапазон (чаще всего [0,1). Так же для нормализации используются такие способы как Z-оценка, показывающая насколько конкретное значение отклоняется от среднего, и десятичное масштабирование. Которое в чем-то схоже с методом минимакса, но заключающееся не в сведении значений к определенному диапазону, а

в делении всех данных на порядок максимального из них. В результате будут получены данные в диапазоне от -1 до 1.

Дискретизация используется прежде всего при обработке данных, представляющих из себя некоторое непрерывное значение, либо когда данных слишком много и необходимо уменьшить их число. Поскольку в некоторых случаях избыточность будет вредить. Дискретизация проводится двумя способами:

1) Группирование разной ширины, подразумевающее разбиение диапазона значений на одинаковые группы. Например, разбиение товаров в магазине на группы по определенным ценовым диапазонам.

2) Группирование разной высоты, подразумевающее разбиение всех возможных значений в группы с одинаковым количеством экземпляров. Примером может служить набор данных с ежедневными доходами предприятия. Вместо того, чтобы обрабатывать информацию с каждого дня, разумнее будет разбить набор на определенные временные промежутки и работать уже с ними.

Уменьшение объема подразумевает не столько объединение, сколько приоритезацию данных. В рамках данного метода применяются три подхода: выборка записей, выборка атрибутов и агрегирование.

Выборка записей представляет из себя анализ всего набора данных, и выбор из них подмножества наиболее важных, либо наиболее типичных. В рамках обучения нейронных сетей, так или иначе, придется использовать данный метод. Поскольку часть исходной информации должна использоваться для проверки того, насколько хорошо обучилась модель. Выбор производится в зависимости от модели. Как правило, чем более непредсказуем результат, тем большее количество данных нужно отдать на обучение.

Выборка атрибутов подразумевает, что для ускорения обработки в процессе обучения, необходимо будет отказаться от части информации, размещенной в каждой строке набора данных.

Агрегирование, по сути, это разделение данных на группы, что, в некотором роде, роднит его с дискретизацией. Но если дискретизация предполагает разбиение на определенные интервалы, то при агрегировании данные объединяются для ускорения работы программы. Например, при обработке информации о пострадавших в дтп, можно объединить все случаи по моделям автомобилей, что впоследствии даст возможность дополнительно провести анализ безопасности каждой отдельной марки [4].

Резюмируя, прежде всего стоит сказать, что не менее половины времени, затраченного на проектирование и разработку нейронных

сетей, тратится на то, чтобы должным образом подготовить данные. Если не отнестись к данному вопросу должным образом, впоследствии придется потратить большое количество времени, на исправление ошибок, связанных с некорректной обработкой. Поэтому можно с уверенностью сказать, что знание методов подготовки, использование необходимых инструментов, а также умение правильно и уместно применять различные методики- один из самых важных навыков для любого специалиста по машинному обучению.

Список использованных источников

1. Data preparation. Средства Data mining для подготовки данных [Электронный ресурс] / Режим доступа: https://www.bigdataschool.ru/blog/data-preparation-operations – Дата доступа: 11.10.2022

2. Разбираемся, в чем разница между Data Mining и Data Extraction [Электронный ресурс] / Режим доступа: https://habr.com/ru/company/skillfactory/blog/524336/ – Дата доступа: 11.10.2022

3. Моем датасет: руководство по очистке данных в Python [Электронный ресурс] / – Режим доступа: https://proglib.io/p/moem-dataset-rukovodstvo-po-ochistke-dannyh-v-python-2020-03-27. – Дата доступа: 12.10.2022

4. Подготовка данных для расширенного машинного обучения [Электронный ресурс]/ – Режим доступа: https://learn.microsoft.com/ru-ru/azure/architecture/data-science-process/prepare-data. – Дата доступа: 12.10.2022

УДК: 339.137.2: 338.43

А.А. Наурызбаева
Таразский региональный университет им. М.Х. Дулати
Тараз, Казахстан

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ КОНСАЛТИНГА В ОБЕСПЕЧЕНИИ ПРОДОВОЛЬСТВЕННОЙ БЕЗОПАСНОСТИ КАЗАХСТАНА

Аннотация. *В данной статье приведены факторы, обусловливающие необходимость совершенствования информационных технологий в агробизнесе, что предполагает реализацию направлений развития предприятий АПК, и*