ORES
Научно-практический электронный журнал
Оригинальные исследования (ОРИС), *ISSN 2222-0402*
www.ores.su
original-research.ru

УДК 004.622

МЕТОДЫ ПРЕДОБРАБОТКИ МАССИВОВ ТЕКСТОВЫХ ДАННЫХ ДЛЯ ИХ ПОСЛЕДУЮЩЕГО АНАЛИЗА

Седова Алена Игоревна
Магистрант 2 курса,
кафедры прикладной информатики и математических методов в экономике (ПИММЭ) Института экономики и управления (ИУЭ) Волгоградского государственного университета (ВолГУ),
Россия, Волгоград,
E-mail: piem-201_524818@volsu.ru

Матушкова Анастасия Алексеевна,
Магистрант 2 курса,
кафедры ПИММЭ ИЭУ ВолГУ,
Россия, Волгоград,
E-mail: piem-201_625542@volsu.ru

Аннотация

В статье приведены методы предварительной обработки данных. Рассмотрены основные проблемы, которые встречаются при предобработке и пути их устранения. Обосновано приведение данных к формату пригодному для дальнейшего анализа массивов текстовых данных.

Ключевые слова: предобработка, анализ данных, очистка данных, методы предобработки данных

METHODS FOR PRE-PROCESSING TEXT DATA ARRAYS FOR THEIR FURTHER ANALYSIS

Alena I. Sedova
2nd year master's student,
IEM VolSU,
Russia, Volgograd,
E-mail: piem-201_524818@volsu.ru

Anastasia A. Matushkova
2nd year master's student,
IEM VolSU,
Russia, Volgograd,
E-mail: piem-201_625542@volsu.ru

ABSTRACT

The article presents methods of data preprocessing. The main problems encountered during preprocessing and ways to eliminate them are considered. The reduction of data to a format suitable for further analysis of text data arrays is substantiated.

Keywords: preprocessing, data analysis, data cleaning, data preprocessing methods

Качество данных – это параметр, обозначающий полноту, точность, своевременность и способность данных к интерпретации. Различают данные высокого и низкого качества.

Данные высокого качества отличаются полнотой, точностью, своевременностью. Такие данные хорошо поддаются интерпретации и обеспечивают получение качественного результата.

Данные низкого качества, или грязные данные характеризуются отсутствующими, неточными или бесполезными значениями. Такие данные требуют предварительной обработки [1, с. 371].

Чтобы применять количественные методы анализа данных, в частности, методы искусственного интеллекта необходимы данные высокого качества. Как правило, получаемые в науке и практике массивы данных характеризуются плохим качеством (такими как выбросы, пустые данные, линейная зависимость между рядами, данные в буквенном формате, существенная разница в порядках чисел в данных и т.п.), что плохо совместимо с анализом данных. Поэтому данные должны быть качественно предварительно обработаны [2, с. 247].

ЭВМ любят обрабатывать красивую и аккуратную информацию — они считывают данные как 1 и 0. Когда, вычисление структурированных данных, таких как целые числа и проценты, очень просто. Однако неструктурированные данные в виде текста и изображений должны быть предварительно очищены и отформатированы перед анализом [5, с. 214].

Предварительная обработка данных является одним из наиболее важных этапов интеллектуального анализа данных, который занимается подготовкой и преобразованием набора данных и в то же время стремится сделать обнаружение знаний более эффективным.

Предобработка данных – комплекс методов и алгоритмов, применяемых с целью подготовки данных к решению конкретной задачи и приведению их в соответствие с требованиями, определяемыми спецификой задачи [3, с. 152].

Предварительная обработка данных является одной из самых важных задач, выполнять которую необходимо перед тем, как применять методы анализа данных.

Данные, как правило, неполны, зашумлены и противоречивы. Низкое качество собранных данных приводит к низкому качеству моделей, построенных на таких данных. Чтобы решить эти проблемы, предварительная обработка данных предоставляет операции, которые могут организовать данные в надлежащую форму для лучшего понимания процесса интеллектуального анализа данных [6, с. 242].

Специалистам при работе с данными всегда необходимо применять некоторые методы предварительной обработки, чтобы сделать данные более удобными для использования. Эти методы облегчат использование в алгоритмах машинного обучения, уменьшат сложность, чтобы предотвратить переобучение, и приведут к лучшей модели.

Наборы данных могут быть переданы с помощью «features», из которых они состоят. Это может быть размер, возраст, время, цвет и т.д. Они отображаются в виде столбцов в

наборах данных и также известны как атрибуты, переменные, показатели и поля. Существует два типа показателей для описания данных: количественные (представленные целыми числами, дробями и процентами) и категориальные (текст, изображение, видео и т.д.) [13, с. 157].

Если пропустить этап предварительной обработки данных, это повлияет на нашу работу позже при применении этого набора данных к модели машинного обучения. Большинство моделей не могут обрабатывать пропущенные значения. На некоторые из них влияют выбросы, высокая размерность и зашумленные данные, поэтому после предварительной обработки, набор данных станет более полным и точным. Этот этап имеет решающее значение для внесения необходимых корректировок в данные перед применением машинного обучения [18].

Принято выделять несколько этапов предварительной обработки данных:

1.Этап оценки качества данных

Практически в любом наборе данных имеются проблемы, которые следует учитывать:

Несоответствие типов данных: данные, собранные из множества разных источников, могут поступать в разных форматах. Хотя конечной целью предобработки является переформатирование данных для машин, начинать все равно следует с данными одного формата [7, с. 53].

Смешанные значения данных: возможно, в разных источниках используются разные описания характеристик, например, мужчина или мужской. Все эти описания значений должны быть унифицированы [20, с. 273].

Выбросы данных. Объекты или наблюдения, которые значительно выделяются в наборе данных. Выбросы могут оказать огромное влияние на результаты анализа.

Отсутствующие данные. Отсутствующие поля данных, пробелы в тексте или оставшиеся без ответа вопросы. Это связано с человеческим фактором или неполными данными. Чтобы позаботиться о недостающих данных, придется выполнить очистку данных [12, с. 63].

2.Этап очистки данных

Очистка данных — это процесс добавления отсутствующих данных и исправления, или удаления неправильных, или нерелевантных данных. Этот этап является наиболее важным в предварительной обработке, потому что он гарантирует, что данные готовы к дальнейшим нуждам [4, с. 314].

Очистка данных исправит все несогласованные данные, которые были обнаружены при оценке качества данных. В зависимости от типа данных, с которыми вы работаете, существует ряд возможных инструментов очистки, которые вам понадобятся для обработки ваших данных:

Отсутствующие данные

Есть несколько способов это исправить, рассмотрим два из них:

- Игнорировать. Если строка содержит пропуски, то ее можно просто удалить. Это рекомендуется только для больших наборов данных, когда несколько проигнорированных объектов не повредят дальнейшему анализу.
- Вручную заполнить недостающие данные: это может быть утомительно, но необходимо при работе с небольшими наборами данных [8, с. 38].

Зашумленные данные

Очистка данных также включает исправление «зашумленных» данных. Это данные, которые не содержат никакой полезной информации, а также ненужные данных,

нерелевантные данные и данные, которые сложнее сгруппировать. Для решения таких проблем используют робастные методы (методы, направленные на выявление выбросов, снижение их влияния или исключение их из выборки). Такие методы отличаются устойчивостью к сильным возмущениям. В результате применения этих методов используется один из следующих вариантов действия: удаление или замена на ближайшее значение.

После очистки данных вы можете обнаружить, что у вас недостаточно данных для выполнения поставленной задачи. На этом этапе вы также можете выполнить обработку или обогащение данных, чтобы добавить новые наборы данных и, прежде чем добавлять их к исходным данным, необходимо применить к ним этапы оценки качества и очистки [10, с. 283].

3.Этап преобразования данных

На предыдущем этапе данные уже начали меняться, но на этапе преобразования данных начнется процесс трансформирования данных в надлежащий формат, который понадобится для анализа и других последующих процессов [16, с. 197].

Обычно это происходит при:

- Агрегации. Агрегация данных объединяет все данные вместе в едином формате.
- Нормализации. Нормализация масштабирует данные до упорядоченного диапазона, чтобы можно было более точно сравнивать их [21, с. 204].
- Выборе признаков. Выбор признаков — это процесс принятия решения о том, какие переменные (признаки, характеристики, категории и т. д.) наиболее важны для анализа. Эти функции будут использоваться для обучения моделей машинного обучения. Важно помнить, что чем больше функций решено использовать, тем дольше будет процесс обучения, а иногда и менее точны ваши результаты, поскольку некоторые характеристики функций могут перекрываться или отсутствовать в данных [19, с. 326].
- Дискредитации. Дискредитация объединяет данные в меньшие интервалы. Обычно происходит после очистки данных [14, с. 407].
- Генерации иерархии концепций. Генерация иерархии концепций может добавить иерархию внутри объектов и между ними, которой не было в исходных данных [9, с. 421].

4. Этап уменьшения объема данных

Чем больше массив данных с которым мы работаем, тем сложнее будет их анализировать, даже после их очистки и преобразования. В зависимости от поставленной задачи данных может быть больше, чем нужно. Особенно при работе с анализом текста большая часть текста является лишней или не имеет отношения к потребностям исследователя. Сокращение данных не только упрощает и делает анализ более точным, но и сокращает объем хранилища данных. Это также помогает определить наиболее важные особенности рассматриваемого процесса [17, с. 132].

Выбор атрибута: Подобно дискредитации, выбор атрибута может поместить данные в пулы меньшего размера. По сути, он объединяет теги или функции, так что такие теги, как мужчина/женщина и профессор, могут быть объединены в мужчина-профессор/женщина-профессор [15, с. 31].

Уменьшение количества: Это поможет с хранением и передачей данных.

Уменьшение размерности: Это, опять же, уменьшает объем данных, используемых для облегчения анализа и последующих процессов. Такие алгоритмы, как K-ближайшие соседи, используют распознавание образов, чтобы объединять схожие данные и делать их более управляемыми [11, с. 112].

Таким образом, этап предварительной обработки данных имеет решающее значение для определения правильных входных данных для алгоритмов машинного обучения. В результате устранения всех вышеперечисленных проблем, встречающихся при предобработке и очистке данных, изначально сырой и неподходящий для применения методов анализа данных массив становится пригодным к обработке методами машинного обучения.

Список литературы:

1. Барсегян, А. А. Анализ данных и процессов: учеб. пособие [Текст] / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д.Тесс, С. И. Елизаров; 3-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2009.— 512 с.
2. Ветошкин Н. В. Методы интеллектуального анализа данных / Н. В. Ветошкин— Екатеринбург: ООО «Издательский Дом «Ажур», 2020. — 474с.
3. Волк, В. К. Базы данных. Проектирование, программирование, управление и администрирование: учебник для вузов / В. К. Волк. — 2-е изд., стер. — Санкт-Петербург: Лань, 2021. — 244 с.
4. Гласснер, Э. Глубокое обучение без математики / Э. Гласснер ; перевод с английского В. А. Яроцкого. — Москва : ДМК Пресс, [б. г.]. — Том 1 : Основы — 2019. — 578 с.
5. Груздев, А. В. Прогнозное моделирование в IBM SPSS Statistics, R и Python: метод деревьев решений и случайный лес: руководство / А. В. Груздев. — Москва: ДМК Пресс, 2018. — 642 с.
6. Гудфеллоу, Я. Глубокое обучение / Я. Гудфеллоу, И. Бенджио, А. Курвилль; перевод с английского А. А. Слинкина. — 2-е изд. — Москва : ДМК Пресс, 2018. — 652 с.
7. Лопатин, В. М. Информатика для инженеров: учебное пособие для вузов / В. М. Лопатин. — 2-е изд., стер. — Санкт-Петербург : Лань, 2021. — 172 с.
8. Луньков А.Д., Харламов А.В. Интеллектуальный анализ данных: Учебно-методическое пособие. Часть 1. // Саратов: Саратовский государственный университет имени Н.Г.Чернышевского. — 94 с.
9. Лю, Б. Интеллектуальный анализ данных в Интернете, изучение гиперссылок, содержимого и данных об использовании. Второе издание. Спрингер, 2011. — 622 с.
10. Макленнен Дж., Танг Ч., Криват Б. Microsoft SQL Server 2008: Datamining — интеллектуальный анализ данных. Пер. с англ. СПб.: БХВ-Петербург, 2009. — 720 с.
11. Макшанов, А. В. Технологии интеллектуального анализа данных : учебное пособие / А. В. Макшанов, А. Е. Журавлев. — 2-е изд., стер. — Санкт-Петербург: Лань, 2022. — 212 с.
12. Мойзес, Б. Б. Статистические методы контроля качества и обработка экспериментальных данных: учебное пособие / Б. Б. Мойзес, И. В. Плотникова, Л. А. Редько. — Томск: ТПУ, 2016. — 119 с.
13. Москвитин, А. А. Данные, информация, знания: методология, теория, технологии : монография / А. А. Москвитин. — Санкт-Петербург: Лань, 2022. — 236 с.

14. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям (+CD): Учебное пособие. 2-е изд., испр. // СПб.: Питер, 2013.— 704 с.
15. Рафаэлович В. Data mining, или Интеллектуальный анализ данных / В. Рафаэлович —Москва: И-трейд, 2016. — 110 с.
16. Рашка, С. Python и машинное обучение: крайне необходимое пособие по новейшей предсказательной аналитике, обязательное для более глубокого понимания методологии машинного обучения : руководство / С. Рашка ; перевод с английского А. В. Логунова. — Москва : ДМК Пресс, 2017. — 418 с.
17. Советов, Б. Я. Информационные технологии: теоретические основы: учебное пособие / Б. Я. Советов, В. В. Цехановский. — 2-е изд., стер. — Санкт-Петербург: Лань, 2022. — 444 с.
18. Технологии обработки информации // E-biblio : [сайт]. — URL: http://www.e-biblio.ru/book/bib/01_informatika/technolog_obrabotki_inform (дата обращения: 18.05.2022)
19. Ферлитш, Э. Шаблоны и практика глубокого обучения / Э. Ферлитш ; перевод с английского А. В. Логунова. — Москва: ДМК Пресс, 2022. — 538 с.
20. Хан, Дж. Камбер, М. Интеллектуальный анализ данных: концепции и методы. – Massachusetts: Morgan Kaufmann Publishers, 2011. — 740 с.
21. Хултен, Д. Разработка интеллектуальных систем: руководство / Д. Хултен ; перевод с английского В. С. Яценкова. — Москва: ДМК Пресс, 2019. — 284 с.

References

1. Barseghyan, A. A. Analysis of data and processes: textbook. allowance [Text] / A.V. A. Barseghyan, M. S. Kupriyanov, I. I. Kholod, M. D. Tess, S. I. Elizarov; 3rd ed., revised. and additional - St. Petersburg: BHV-Petersburg, 2009. - 512 p.
2. Vetoshkin N. V. Methods of data mining / N. V. Vetoshkin - Yekaterinburg: Publishing House Azhur LLC, 2020. - 474p.
3. Volk, V. K. Databases. Design, programming, management and administration: a textbook for universities / V. K. Volk. - 2nd ed., erased. - St. Petersburg: Lan, 2021. - 244 p.
4. Glassner, E. Deep learning without mathematics / E. Glassner; translation from English by V. A. Yarotsky. - Moscow: DMK Press, [b. G.]. - Volume 1: Basics - 2019. - 578 p.
5. Gruzdev, A. V. Predictive Modeling in IBM SPSS Statistics, R and Python: Decision Tree Method and Random Forest: A Guide / A. V. Gruzdev. - Moscow: DMK Press, 2018. - 642 p.
6. Goodfellow, Y. Deep learning / Y. Goodfellow, I. Bengio, A. Courville; translation from English by A. A. Slinkin. - 2nd ed. - Moscow: DMK Press, 2018. - 652 p.
7. Lopatin, V. M. Informatics for engineers: textbook for universities / V. M. Lopatin. - 2nd ed., erased. - St. Petersburg: Lan, 2021. - 172 p.
8. Lunkov A.D., Kharlamov A.V. Data Mining: Educational and Methodological Guide. Part 1. // Saratov: Saratov State University named after N.G. Chernyshevsky. - 94 p.
9. Liu, B., 2011. Web Data Mining Exploring Hyperlinks, Contents, and Usage Data. Second Edition. Springer, 622 p.

10. Maklennen Dzh., Tang Ch., Krivat B. Microsoft SQL Server 2008: Datamining — data mining. Per. from English. [Microsoft SQL Server 2008: Datamining] SPb.: BKhV-Peterburg, 2009.- 720 p.
11. Makshanov, A. V. Data mining technologies: textbook / A. V. Makshanov, A. E. Zhuravlev. - 2nd ed., erased. - St. Petersburg: Lan, 2022. - 212 p.
12. Moizes, B. B. Statistical methods of quality control and processing of experimental data: textbook / B. B. Moizes, I. V. Plotnikova, L. A. Redko. - Tomsk: TPU, 2016. - 119 p.
13. Moskvitin, A. A. Data, information, knowledge: methodology, theory, technologies: monograph / A. A. Moskvitin. - St. Petersburg: Lan, 2022. - 236 p.
14. Paklin N.B., Oreshkov V.I. Business Intelligence: From Data to Knowledge (+CD): Textbook. 2nd ed., rev. // St. Petersburg: Peter, 2013. - 704 p.
15. Rafaelovich V. Data mining, or data mining / V. Rafaelovich - Moscow: I-trade, 2016. - 110 p.
16. Raschka, S. Python and Machine Learning: A Much-needed Handbook of the Latest Predictive Analytics, Mandatory for a Deeper Understanding of Machine Learning Methodology: A Guide / S. Raschka; translation from English by A. V. Logunov. - Moscow: DMK Press, 2017. - 418 p.
17. Sovetov, B. Ya. Information technologies: theoretical foundations: textbook / B. Ya. Sovetov, V. V. Tsekhanovsky. - 2nd ed., erased. - St. Petersburg: Lan, 2022. - 444 p.
18. Information processing technologies // E-biblio: [website]. — URL: http://www.e-biblio.ru/book/bib/01_informatika/technolog_obrabotki_inform (date of access: 05/18/2022)
19. Ferlitsch, E. Patterns and practice of deep learning / E. Ferlitsch; translation from English by A. V. Logunov. - Moscow: DMK Press, 2022. - 538 p.
20. Khan, J. Kamber, M. Data Mining: Observations and Methods. Massachusetts: Morgan Kaufmann Publishers, 2011. - 740p.
21. Hulten, D. Development of intelligent systems: a guide / D. Hulten; translation from English by V. S. Yatsenkov. - Moscow: DMK Press, 2019. - 284 p.