УДК 550.8.05

«СЫРЫЕ» ДАННЫЕ И НЕКОТОРЫЕ РЕЦЕПТЫ ИХ «ПРИГОТОВЛЕНИЯ»

М.М. Татур[1], В.М. Проровский[1], Д.В. Куприянова[1], И.Н.Носырев[2]

[1]*Белорусский государственный университет информатики и радиоэлектроники, ул. П. Бровки, 6, г. Минск, 220013, Беларусь,*
tatur@bsuir.by, slawapro@gmail.com, diankupriyanova@gmail.com
[2]*Объединенный институт машиностроения Национальной академии наук Беларуси, ул. Академическая, 12, г. Минск, 220072, Беларусь,* nosureviluha@mail.ru

Подготовка данных для их обработки формальными алгоритмами анализа (классификации, кластеризации, регрессии и др.) имеет важное значение, поскольку существенно влияет на результат принимаемых решений. В работе рассматриваются типовые операции препроцессинга данных на примере набора данных о пожарах. Наряду с тривиальными операциями по очистке и форматированию, этап подготовки данных включает неформальные процедуры, которые требуют участия как специалистов по анализу данных, так и экспертов из предметной области. Показывается, как некоторые признаки необходимо преобразовывать из разряда порядковых, номинальных, необрабатываемых, в числовые значения, а также придавать вес в принятии решений.

***Ключевые слова*:** интеллектуальный анализ; машинное обучение; подготовка данных; статистические данные о пожарах.

«RAW» DATA AND SOME RECIPES FOR THEIR «COOKING»

M.M. Tatur[a], V.M. Prorovsky[a], D.V. Kupriyanova[a], I.N. Nosarau[b]

[a]*Belarusian State University of Informatics and Radioelectronics, P.Brouki st., 6, Minsk, 220013, Belarus,*
tatur@bsuir.by, slawapro@gmail.com, diankupriyanova@gmail.com
[b]*Joint Institute of Mechanical Engineering of the National Academy of Sciences of Belarus, Akademicheskaya str., 12, Minsk, 220072, Belarus,* nosureviluha@mail.ru

Preparation of data for their processing by formal analysis algorithms (classification, clustering, regression, etc.) is important, since it significantly affects the result of decisions made. The paper considers typical data preprocessing operations using the fire data set as an example. Along with the trivial cleansing and formatting, the data preparation phase includes informal procedures that require the participation of both data scientists and subject matter experts. It is shown how some signs need to be converted from the category of ordinal, nominal, unprocessed, into numerical values, and also to give weight in decision making.

Keywords: data mining; machine learning; data preparation; fire statistics.

Введение

В большинстве научных статей, посвященных интеллектуальному анализу данных и машинному обучению, рассматриваются результаты обработки тестовых или полученных (измеренных) наборов данных различными *формальными* алгоритмами. Как правило, предметом исследования выступает оценка эффективности примененных методов и обоснование выбора наиболее наилучшего метода, алгоритма и их параметров [1-4]. Не преуменьшая указанной стороны анализа данных, следует обращать внимание на этап подготовки данных (Data preparation в методологии CRISP-DM) [5], в котором значительная часть является *неформальной.*

В академической среде имеется устоявшееся мнение, что этап подготовки данных занимает примерно 60-80% времени и ресурсов, отведенных на анализ. В этой связи, возникает ряд вопросов, из которых назовем лишь три, на наш взгляд, наиболее значимые.

1. Существует ли общепринятая методика препроцессинга исходных (сырых) данных?
2. Какие этапы подготовки следует выделять и степени их формализации?
3. Насколько качество подготовки данных способно повлиять на конечные результаты анализа и, в конечном итоге, на принятие решений?

По первому вопросу, можно привести ряд публикаций [6-9], из которых следует, что общепринятой методики подготовки данных не существует. Даже при использовании современных библиотек алгоритмов автоматического машинного обучения (AutoML), осуществляющих предварительную обработку сырых данных этот процесс реализуется в каждом случае по-своему [10]. При ручной обработке дата-аналитик решает на основании своего опыта какие методы и в каком объеме использовать.

Второй вопрос будет более детально рассмотрен в настоящей работе.

Ответ на третий вопрос связан с проведением специальных исследований, детальным анализом результатов экспериментов и выходит за рамки данной публикации.

1. Пример исходных данных для обработки

Ниже, в таблице 1, приведен пример исходных данных о пожарах, содержащий признаки и данные, сходные с реальными.

Таблица 1 - Образец исходных данных о пожарах

	Дата	Адрес (текст)	Причина (текст)	Площадь пожара (число)	Объект пожара (текст)	Расстояние (число)	Собственник объекта (текст)
	1	2	3	4	5	6	7
1	15.01.2021	ул. Лиственная, 2/58	неосторожное обращение с огнем	43	жилой дом	15	Иванов В.В.
2	24.03.2021	ул. Шмидта, 5	поджог	60	склад	7,5	ООО “Пистон”
3	13.05.2021	дддлоолол		ошибка	жилой дом	-	Петров И.И.
4	18/07/2021	пр-д. Жемчуж-ный, 2в	неосторожное обращение с огнем	13	склад		УП “Промстрой”
5	01/09/2021		поджог	85	жилой дом	12	Сидоров С.Г.
…							
n	24.03.2021	ул. Шмидта, 5	поджог	60	склад	7,5	ООО “Пистон”

Число записей (строк) в данных исчисляется тысячами и даже десятками тысяч, в зависимости от выбранного для анализа исторического периода.

Число информативных признаков (столбцов) в данном примере для наглядности ограничено семью, в то время как реальные данные содержат десятки, детализирующие способы реагирования, причиненный ущерб и другую информацию о пожаре.

Несложно заметить, что данные в таблице имеют как числовое (площадь, расстояние), так порядковое (дата), так и номинальные (адрес, причина, объект, собственник) значения.

Заметим, что на текущий момент изложения, задача анализа данных не сформулирована.

Для начала, перечислим операции, которые обычно относят к этапу подготовки данных, с кратким пояснением их сути.

Эти операции включают:

- Очистку данных – выявление и исправление ошибок или ошибок в данных.

- Отбор признаков – определение признаков, которые наиболее важны для анализа.
- Преобразование данных – изменение масштаба или распределения переменных.
- Конструирование признаков – получение новых переменных из доступных данных.
- Уменьшение размерности – создание компактных проекций данных.

Данный перечень не претендует на полноту, а названия операций и их содержание также могут отличаться в изложении различных авторов.

2. Операции по предварительной обработке данных

2.1 Очистка данных

Включает исправление систематических проблем или ошибок в беспорядочных данных. Наиболее полезная очистка данных требует глубоких знаний предметной области и может включать выявление и устранение конкретных ошибочных наблюдений. Имеется много причин, по которым в данных могут быть неверные значения. Например, опечатки, повреждения, дублирование и т. д. Изучение предметной области может позволить идентифицировать явно ошибочные наблюдения, поскольку они отличаются от типовых. Например, отрицательное расстояние от пожарной части до места пожара.

После выявления беспорядочных, зашумленных, искаженных или ошибочных наблюдений их можно устранить. Это может быть удаление строк или столбцов, или замена наблюдений новыми значениями.

К общим операциям по очистке данных, можно отнести:

- Применение статистических методов для определения нормальности данных и выявления выбросов.
- Выявление дубликатов строк данных и их удаление
- Замена пустых значений как отсутствующих
- Замена пропущенных значений с использованием статистики или обученных моделей.

В таблице 1 строки 2 и n являются идентичными, и для дальнейшей работы одна из них должна быть удалена. В колонке 3 имеются пропуски, а в колонках 4 и 6 имеются не только пропуски, но и текстовые значения, которые не могут быть преобразованы в числа напрямую. В столбце 1 даты представлены в разных форматах. В качестве примера выявления выбросов в данных о пожарах можно привести реальное применение методов S-ESD (Seasonal Extreme Studentized Deviate) при построении прогнозной модели возникновения пожаров [11]. Кроме этого, исследование

выбросов в контексте интеллектуального анализа данных может выявить полезную, неизвестную ранее информацию.

Очистка данных это операция, которая обычно выполняется в первую очередь, перед другими операциями подготовки данных

2.2 Отбор признаков

Заключается в выборе подмножества входных признаков, наиболее релевантных прогнозируемой целевой переменной. Нерелевантные и избыточные входные переменные могут отвлекать или вводить в заблуждение алгоритмы обучения, что может привести к снижению эффективности прогнозирования. Более эффективно разрабатывать те модели, которые используют меньше признаков, необходимых для прогнозирования, т.е. наиболее простые из работающих достаточно точно моделей.

Методы выбора признаков условно подразделяются на использующие целевую переменную (supervised), и не использующие (unsupervised). Контролируемые методы могут быть далее разделены на модели, которые автоматически выбирают признаки в процессе подбора модели (встроенные), те, которые явно выбирают признаки, приводящие к наилучшей производительности модели (обертки), и те, которые оценивают каждую входную переменную и позволяют выбрать их несколько из набора (фильтры).

Распространено использование статистических операций (например, корреляция) для оценки входных признаков. Далее признаки ранжируются по их оценкам и набор с наилучшими показателями используется в качестве входных данных для модели. Выбор статистической операции зависит от типа данных входных переменных. Различают общие варианты выбора признаков, например:

- Категориальные входные данные для целевой переменной классификации.
- Числовые входные данные для целевой переменной классификации.
- Числовые входные данные для целевой переменной регрессии.

В случае, если типы данных входных переменных отличаются, используют разные методы фильтрации. Как альтернативу можно использовать метод-оболочку, например, метод устранения рекурсивных признаков (RFE), который не зависит от типа входной переменной.

В нашем примере очевидным для исключения столбцами является 7. А вот данные из столбца 2, несмотря на имеющиеся ошибки могут оказаться полезными. Используя вспомогательные сервисы вроде Nominatim от OpenStreetMap возможно получение для дальнейшего применения геокоординат по текстовой части информации.

2.3 Преобразование данных.

Эти операции используются для изменения типа переменных или распределения данных. Они представляют собой достаточно большой набор различных методов, применяемых как к входным, так и к выходным переменным. Данные могут иметь один из нескольких типов:

Числовые	Категориальные
Integer – целые числа без дробной части. Float – числа с плавающей запятой.	Ordinal – метки с ранговым порядком. Nominal – метки без ранжирования. Boolean – значения True и False.

При необходимости возможно преобразовать числовую переменную в порядковую переменную (дискретизация). Для большинства задач классификации категориальные переменные могут быть закодированы в виде целого числа или логических переменных.

- Дискретизация – кодирование числовой переменной как порядковой переменной.
- Порядковое преобразование – кодирование категориальной переменной в целочисленную переменную.
- Быстрое кодирование (One-Hot encoding) – кодирование категориальной переменной в двоичные переменные.

Масштабирование числовых данных, т.е. их обезличивание, в отношении единиц измерения. Операция формализуемая или почти формализуемая.

Существует ряд способов масштабирования, среди которых наиболее часто применяются такие как нормализация и стандартизация. Нормализация – это преобразование значений отдельного признака в диапазон значений с плавающей точкой от 0 до 1. Стандартизация преобразовывает значения признака путем вычитания среднего значения (так называемое центрирование) и деления на стандартное отклонение, чтобы сдвинуть распределение так, чтобы среднее значение равнялось нулю, а стандартное отклонение равнялось единице.

Для некоторых алгоритмов масштаб числовых значений признака не влияет. В первую очередь это деревья решений и ансамбли деревьев, такие как случайный лес.

Выбор необходимости и способа масштабирования определяется дата-аналитиком, исходя из конкретной решаемой задачи.

Распределение значений вероятности для числовых переменных может быть изменено. Если распределение почти нормальное, но искажено или смещено, его можно сделать более нормальным с помощью степенно-

го преобразования [12]. Квантильные преобразования можно использовать для принудительного распределения вероятностей (равномерной или нормальной) для переменной с необычным естественным распределением.

При преобразовании данных операции обычно выполняются отдельно для каждой переменной.

В нашем наборе данных столбцы 3 и 5 являются категориальными признаками и с помощью порядкового преобразования или быстрого кодирования могут быть преобразованы к формату, который далее может быть использован для обучения конкретных моделей.

2.4 Конструирование признаков.

Процесс создания новых входных переменных из доступных данных называется конструированием признаков. Эти операции сильно зависят от состава и типов данных. При этом необходимо сотрудничество эксперта в предметной области. Эти условия затрудняют унификацию общих методов. Есть ряд приемов, которые можно использовать повторно:

- Добавление логической переменной флага для некоторого состояния.
- Добавление групповой или глобальной сводной статистики, например, среднего значения.
- Добавление новых переменных для каждого компонента составной переменной, такой как дата-время.

Например, при построении обобщенной линейной модели обстановки с пожарами [13] выполнено разложение признака с типом «дата» на 21 признак, значения весов которых приведены в таблице 2.

Таблица 2 – Признаки и их веса

№	Показатель	Вес	№	Показатель	Вес
1	Разность дат (Текущая дата - Дата)	0,726	8	Дата: полугодие = 2	0,085
2	Дата: день недели = 7	0,509	9	Дата: месяц года	0,085
3	Дата: день недели = 1	0,436	10	Дата: полугодие = 1	0,030
4	Дата: день месяца	0,245	11	Дата: день недели = 6	0,022
5	Дата: месяц квартала = 1	0,182	12	Дата: день недели = 5	0,018
6	Дата: день недели = 2	0,168	13	Остальные показатели	0,000
7	Дата: квартал = 2	0,108			

Подход, основанный на статистике, заключается в создании копий числовых входных переменных, которые были изменены с помощью простой математической операции, такой как возведение их в степень или умножение на другие входные переменные, называемые полиномиальными признаками.

Полиномиальное преобразование заключается в создании копий числовых входных переменных, возведенных в степень.

Основная цель конструирования признаков состоит в том, чтобы добавить более широкий контекст к отдельному наблюдению или разложить сложную переменную для более прозрачного представления о входных данных. Конструирование признаков некоторые авторы относят к разновидности преобразования данных.

2.5 Уменьшение размерности

Количество входных признаков для набора данных рассматривается как размерность данных. Например, две входные переменные образуют двумерную область, где каждая строка данных определяет точку в этом пространстве. Набор данных может содержать любое количество входных переменных для создания многомерного пространства. Проблема в том, что чем больше измерений имеет это пространство, тем больше вероятность того, что набор данных представляет собой избыточную и, вероятно, нерепрезентативную выборку.

Это заставляет дата-аналитика выбирать наиболее информативные признаки. Альтернативой выбору признаков является создание проекции данных в пространство более низкого измерения, которое по-прежнему сохраняет наиболее важные свойства исходных данных. Этот процесс называется уменьшением размерности. В отличие от выбора признаков, переменные в прогнозируемых данных не связаны напрямую с исходными входными переменными, что затрудняет интерпретацию прогноза. Наиболее распространенным подходом к уменьшению размерности является использование методов матричной факторизации: анализ главных компонент, сингулярное разложение.

Основное преимущество этих методов заключается в том, что они устраняют линейные зависимости между входными переменными, например, коррелированными переменными. [9].

Заключение

Перечень рассмотренных операций предварительной подготовки данных может быть расширен, например, такой процедурой как оценка репрезентативности выборки для анализа данных и принятия решений. Данный вопрос тесно связан с такими параметрами выборки как количество наблюдений, размерность пространства признаков, распределение объектов (образов) в этом пространстве, наличие шумов, выбросов и др. Т.е. оценка репрезентативности тесно связана с другими операциями подготовки данных и, в некоторых случая, может быть представлена их композицией.

Также надо принимать во внимание целевую задачу, решаемую в ходе предстоящего анализа данных, которая в свою очередь будет существенно определять состав и содержание операций предварительной обработки.

Библиографические ссылки

1. Borges L. Comparison of data mining techniques and tools for data classification / L. Borges, M.Viriato, B.Jorge // ACM International Conference Proceeding Series. 2013. P. 113–116., DOI:10.1145/2494444.2494451.
2. Tapak L. Real-data comparison of data mining methods in prediction of diabetes in Iran / L.Tapak, H.Mahjub, O.Hamidi, J.Poorolajal // Healthc Inform Res. 2013. Vol. 19(3). P. 177–185., DOI: 10.4258/hir.2013.19.3.177.
3. Garg S. Comparative Analysis of Data Mining Techniques on Educational Dataset / S. Garg, A. K. Sharma // International Journal of Computer Applications. 2013. Vol. 74(5). P. 0975 – 8887.
4. Trifonov R. Analysis of data mining evaluation methods' efficiency / R. Trifonov, D. Gotseva, V. Angelov // International Journal of Development Research. 2017. Vol 11(7). P. 16880–16884.
5. Azevedo A. KDD, SEMMA and CRISP-DM: A parallel overview [Electronic resource] / A. Azevedo, M. Santos // IADIS Multi Conf. on Computer Science and Information Systems, Amsterdam, 22–27 July 2008 / Intern. Assoc. for Development if the Inform. Soc.; Associate Ed.: Luís Rodrigues and Patrícia Barbosa. Amsterdam, 2008. P. 182–185.
6. Shichao Z. Data Preparation for Data Mining / Z. Shichao, Z. Chengqi, Y. Qiang. // Applied Artificial Intelligence. 2003. Vol. 17. P. 375–381., DOI:10.1080/713827180.
7. Zheng, A. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists / A. Zheng, A.Casari. 1st Edition. O'Reilly Media, 2018. 218 p.
8. Kuhn M. Feature Engineering and Selection: A Practical Approach for Predictive Models / M.Kuhn, K.Johnson. 1st Edition. Chapman & Hall/CRC, 2019. 298 p.
9. Brownlee J. Data Preparation for Machine Learning. 1st Edition. 2020. 398 p.
10. Чистый AutoML для "грязных" данных: как и зачем автоматизировать предобработку таблиц в машинном обучении. URL: https://habr.com/ru/company/ods/blog/657525/. (дата обращения: 24.06.2022.)

11. Tatur M.M. Exploratory analysis of the fire statistics using automatic time series decomposition / M.M. Tatur, V.M. Prorovsky, A.G. Ivanitskiy, M. Kvassay // Information and Digital Technologies 2021: Proc. of the Intern. Conf., 22–24 June 2021, Zilina, Slovakia, ed. J. Rabcan [et al.]. Zilina: University of Zilina, 2021. P. 158–161.
12. Box G.E.P. An Analysis of Transformations / G.E.P. Box; D.R. Cox // Journal of the Royal Statistical Society. Series B (Methodological), Vol. 26, №. 2. 1964. P. 211–252.
13. Татур М.М. и др. Сравнение точности алгоритмов автоматического машинного обучения при прогнозировании обстановки с пожарами на объектах жилого сектора //Чрезвычайн. ситуации: предупреждение и ликвидация. 2021. № 22(50). С. 61–70.