Инженерный вестник Дона, №6 (2022)
ivdon.ru/ru/magazine/archive/n6y2022/7736

Подготовка данных для кластеризации событий в журналах информационной безопасности

Д.Н. Сидорова

Новосибирский государственный технический университет

Аннотация: В статье показано, что подготовка данных для использования в дальнейшем в алгоритмах играет важную роль и этому стоит уделить внимание. Рассмотрены задачи обработки исходных данных: выборка данных, очистка данных, генерация признаков, интеграция, форматирование. Исследование данных состоит в изучении следующих шагов: обобщение данных, группировка данных, исследование отношений между разными атрибутами. Приведен алгоритм действий подготовки данных в рамках событий журнала информационной безопасности для дальнейшей кластеризации.
Ключевые слова: данные, кластеризация данных, события, журнал информационной безопасности, алгоритм, Data Mining, Data Preparation, dataset, Machine Learning

Из всех этапов анализа подготовка данных представляется наименее проблемным шагом, но на самом деле требует наибольшего количества ресурсов и времени для завершения. Во многих рассмотренных статьях [1-3] этап подготовки данных указан, как просто этап, без подробного описания, без указания сложных моментов в начале. Зачастую указано, что данные описываются в виде таблиц и все на этом, на самом деле все намного сложнее и трудозатратнее. Данные нередко собираются из различных источников, любой из которых может представлять их в своем виде либо в определенном формате. Поэтому их необходимо подготовить для процесса анализа.

Информация, которая получена из витрины данных или корпоративного хранилища из исходных данных, зачастую имеет нечеткую структуру. Но машинное обучение не работает самостоятельно и независимо, как считают большинство пользователей. Для адекватной деятельности этого инструмента, как и любого ИТ-средства, нужны верно определенные начальные данные и инструкции. Не бывает так, что, загрузив все большие накопленные данные различных форматов в алгоритм Machine Learning,

можно получить корректные результаты на выходе. А также начальные данные часто ненадежны и изменены: они могут содержать аномальные значения (выбросы); в них могут находиться значения, которые выходят за рамки возможных значений (шумы); и отсутствовать значения (пропуски).

Притом, нередко появляется задача подготовительной работы начальных данных. К примеру, может стоять задача установления тональности клиентских отзывов, для этого необходимо сперва разделить текст на смысловые выражения (токены), преобразовать слова («оцифровка»), перевоплотить их в числовые векторы. Из-за своеобразия местности, а именно, по причине наличия холмов или подвальных помещений, попадаются в географических данных ошибки установления координат и опечатки в адресах. В числовой последовательности наблюдаются значения, которые выходят за рамки допустимого диапазона, к примеру, цифра 11 в десятибалльной шкале оценок. Кроме того, числовые значения начальных данных могут очень колебаться по абсолютным величинам: от нескольких сотых процентов до десятков тысяч единиц. Такие погрешности изменят показатели моделирования и не разрешат получить модель машинного обучения с удовлетворительным качеством.

Стандарты Data Mining не просто так представляют подготовку данных в отдельный этап [4]. Data Preparation – это процесс манипулирования необработанными данными в форме, которая может быть легко и точно проанализирована. Он является кропотливым итеративным, также занимает до 80 % всех затрат времени и ресурсов в жизненном цикле. В него входят задачи обработки исходных («сырых») данных, которые представлены ниже:

1. Выбор данных – выбор признаков (функций или предикторов) и объектов с учетом их актуальности для задач Data Mining, качества и технических ограничений (размер и тип);

2. Очистка данных – удаление опечаток, некорректных значений (например, числа в строковом параметре и т.п.), отсутствующих значений (Missing values или NA), устранение дубликатов и разных описаний одного и того же объекта, восстановление уникальности, целостность и логические отношения;

3. Генерация признаков – получение признаков и преобразование их в векторы для модели машинного обучения, а также преобразование для повышения точности алгоритмов машинного обучения;

4. Интеграция – объединение данных из различных источников (информационных систем, таблиц, протоколов и т. д.), в том числе, их агрегация, когда новые значения рассчитываются путем суммирования информации из множества существующих записей;

5. Форматирование – это синтаксические изменения, не меняющие смысла данных, но требуемые инструментами моделирования, такие как сортировка в определенном порядке или удаление ненужных знаков препинания в текстовых полях, обрезка «длинных» слов, округление действительных чисел до целого, и т.п.

Подготовка данных включает такие процессы, как:

- Получение;
- Очистка;
- Нормализация;
- Превращение в оптимизированный набор данных.

Обычно это табличная форма, подходящая для методов, которые были намечены на шаге проектировки. Перед тем, как использовать методы машинного обучения, нужно конвертировать данные в табличное представление, более распространенное в Machine Learning и Data Mining [5-6]. Получив файл с «сырыми» данными, к примеру, в формате CSV, специалист поначалу просматривает его, чтоб осознать характер записей

(строк), также смысл, тип и спектр значений признаков (столбцов). После специалист по данным создает выборку (dataset, набор данных) – выбирает данные, связанные потенциально с проверяемой гипотезой машинного обучения. Почти все трудности могут появиться при возникновении недействительных, многосмысленных либо недостающих значений, повторении полей либо данных, несоответствующих приемлемому интервалу. Исследование данных состоит из предварительного изучения, которое необходимо для понимания типа и смысла полученной информации [7-8]. Вместе с данными, которые были собраны при определении проблемы, такая категоризация описывает, какой способ изучения данных идеальнее всего подойдет для определения модели. Изучение состоит из следующих шагов:

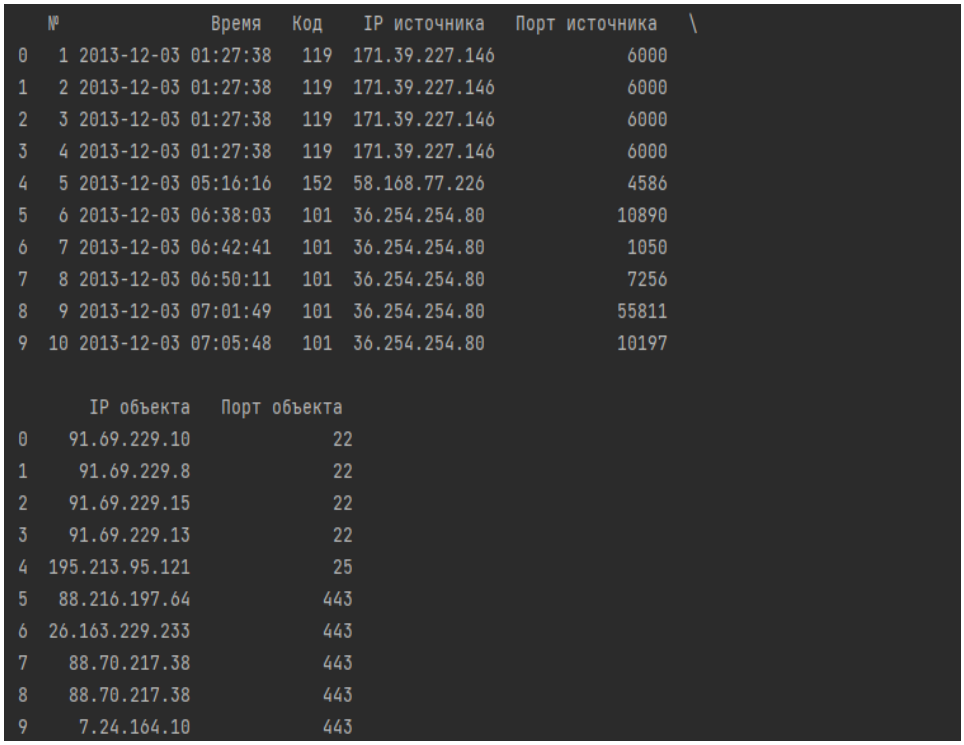
- Обобщение данных;
- Группировка данных;
- Исследование отношений между разными атрибутами;
- Определение моделей и тенденций;
- Построение моделей регрессионного анализа;
- Построение моделей классификации.

Для наших данных необходимы первые три шага. Как правило, анализ данных требует обобщающих утверждений об изучаемых данных. Обобщение – это процесс уменьшения количества данных, подлежащих интерпретации, без потери важной информации.

Кластерный анализ – метод анализа данных, используемый для поиска групп, объединенных общими атрибутами (также называется группировкой) [9]. Далее выполняется очистка данных: конвертация типов данных, агрегация признаков, заполняются отсутствующие значения, исправляются шумы и выбросы. Нормализация значений применяется к числовым переменным, чтобы привести их в один и тот же диапазон и использовать

вместе в одной модели Machine Learning. Как правило, нормализация данных означает преобразование исходных числовых значений в новые значения в диапазоне от 0 до 1 на основе начального минимума и максимума [10].

Несмотря на работу по правильному сохранению данных для анализа, для каждой конкретной задачи всё равно могут требоваться корректировки значений. Основные манипуляции по подготовке данных проводились относительно значений Даты-Времени, а также IP-адресов источника и объекта. Результат вывода исходной таблицы с датой, приведённой в стандартный формат для работы, представлен на рис. 1.



	№	Время	Код	IP источника	Порт источника	\
0	1	2013-12-03 01:27:38	119	171.39.227.146	6000	
1	2	2013-12-03 01:27:38	119	171.39.227.146	6000	
2	3	2013-12-03 01:27:38	119	171.39.227.146	6000	
3	4	2013-12-03 01:27:38	119	171.39.227.146	6000	
4	5	2013-12-03 05:16:16	152	58.168.77.226	4586	
5	6	2013-12-03 06:38:03	101	36.254.254.80	10890	
6	7	2013-12-03 06:42:41	101	36.254.254.80	1050	
7	8	2013-12-03 06:50:11	101	36.254.254.80	7256	
8	9	2013-12-03 07:01:49	101	36.254.254.80	55811	
9	10	2013-12-03 07:05:48	101	36.254.254.80	10197	

	IP объекта	Порт объекта
0	91.69.229.10	22
1	91.69.229.8	22
2	91.69.229.15	22
3	91.69.229.13	22
4	195.213.95.121	25
5	88.216.197.64	443
6	26.163.229.233	443
7	88.70.217.38	443
8	88.70.217.38	443
9	7.24.164.10	443

Рис. 1 – Вывод исходной таблицы

Далее, для приведения всех значений к типу целочисленных, значения даты и времени кодируются. IP-адреса очищаются от разделителей и так же приводятся к целочисленному типу. Результат вывода таблицы, приведённой к единому типу данных представлен на рис. 2.

	№	Время	Код	IP источника	Порт источника	IP объекта	\
0	1	1386034058000000000	119	17139227146	6000	916922910	
1	2	1386034058000000000	119	17139227146	6000	91692298	
2	3	1386034058000000000	119	17139227146	6000	916922915	
3	4	1386034058000000000	119	17139227146	6000	916922913	
4	5	1386047776000000000	152	5816877226	4586	19521395121	
5	6	1386052683000000000	101	3625425480	10890	8821619764	
6	7	1386052961000000000	101	3625425480	1050	26163229233	
7	8	1386053411000000000	101	3625425480	7256	887021738	
8	9	1386054109000000000	101	3625425480	55811	887021738	
9	10	1386054348000000000	101	3625425480	10197	72416410	

	Порт объекта
0	22
1	22
2	22
3	22
4	25
5	443
6	443
7	443
8	443
9	443

Рис. 2 – Результат вывода таблицы после корректировки

Полученные данные обладают несопоставимыми значениями. Для их использования требуется привести их к общему масштабу. Для этих целей в программе была использована нормализация на стандартное отклонение:

$$Z = \frac{x - \mu}{\sigma}$$

где μ – среднее; σ – стандартное отклонение.

Результат вывода датасета, содержащего нормализованные значения, представлен на рис. 3.

```
[[-1.41080419 -0.75295138  0.9863355   0.22699433 -0.51688237 -1.63570403]
 [-1.41080419 -0.75295138  0.9863355   0.22699433 -0.52976015 -1.63570403]
 [-1.41080419 -0.75295138  0.9863355   0.22699433 -0.51688237 -1.63570403]
 ...
 [ 1.53734217  1.30301168  1.28320691 -0.30461439  2.48453954 -0.60258201]
 [ 1.55235991  1.30301168  1.28320691 -0.30461439  2.48453954  1.25489695]
 [ 1.61824377  0.76196877  1.24912142 -0.30220083 -0.20942293  0.24023581]]
```

Рис. 3 – Результат вывода с нормализованными значениями

Полученный набор данных готов для обработки алгоритмами кластеризации.

Литература

1. Max Landauer, Florian Skopik, Markus Wurzenberger, Andreas Rauber, System log clustering approaches for cyber security applications: A survey, Computers & Security, Volume 92, 2020, 101739, ISSN 0167-4048. URL: sciencedirect.com/science/article/pii/S0167404820300250?via%3Dihub.

2. Чернов А.В., Бутакова М.А., Шевчук П.С. Кластеризация данных методом растущего нейронного газа // Инженерный вестник Дона, 2020, №7. URL: ivdon.ru/magazine/archive/ N7y2020/6537.

3. Долгодворова Е.В. Кластерный анализ: базовые концепции и алгоритмы // Вопросы науки и образования, 2018, №7 (19). URL: cyberleninka.ru/article/n/klasternyy-analiz-bazovye-kontseptsii-i-algoritmy.

4. Bartschat A, Reischl M, Mikut R. Data mining tools // WIREs Data Mining Knowl Discov. 2019;9:e1309. URL: wires.onlinelibrary.wiley.com/doi/10.1002/widm.1309.

5. Kavitha G., Raj L. Educational Data Mining and Learning Analytics – Educational Assistance for Teaching and Learning // Internal Journal of Computer and Organization Trends (IJCOT). 2017. Vol. 7, issue 2. Pp. 21-25.

6. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / пер. с анг. А.А. Слинкина. – М.: ДМК Пресс, 2015. – 400 с.: ил.

7. Шумейко А.А., Сотник С.Л. Интеллектуальных анализ данных (Введение в Data Mining): учеб. пособ. – Днепропетровск: Белая Е.А., 2012. – 212 с.

8. Mirmozaffari, M., Boskabadi, A., Azeem, G., Massah, R., Boskabadi, E., Dolatsara, H.A. and Liravian, A. Machine learning Clustering Algorithms Based on the DEA Optimization Approach for Banking System in Developing Countries // European Journal of Engineering and Technology Research. 2020. Vol. 5, issue 6. Pp. 651–658.

9. Гранков М.В., Аль-Габри В.М., Горлова М.Ю. Анализ и кластеризация основных факторов, влияющих на успеваемость учебных групп вуза // Инженерный вестник Дона, 2016, №4. URL: ivdon.ru/magazine/archive/ /n4y2016/3775.

10. Dalwinder Singh, Birmohan Singh, Investigating the impact of data normalization on classification performance, Applied Soft Computing, Volume 97, Part B, 2020, 105524, ISSN 1568-4946. URL: sciencedirect.com/science/article/abs/pii/S1568494619302947?via%3Dihub.

References

1. Max Landauer, Florian Skopik, Markus Wurzenberger, Andreas Rauber, System log clustering approaches for cyber security applications: A survey, Computers & Security, Volume 92, 2020, 101739. URL: sciencedirect.com/science/article/pii/S0167404820300250?via%3Dihub.

2. Chernov A.V., Butakova M.A., Shevchuk P.S. Inzhenernyj vestnik Dona, 2020, №7. URL: ivdon.ru/magazine/archive/ N7y2020/6537.

3. Dolgodvorova YE.V. Voprosy nauki i obrazovaniya, 2018, №7 (19). URL: cyberleninka.ru/article/n/klasternyy-analiz-bazovye-kontseptsii-i-algoritmy.

4. Bartschat A, Reischl M, Mikut R. Data mining tools. WIREs Data Mining Knowl Discov. 2019;9:e1309. URL: wires.onlinelibrary.wiley.com/doi/10.1002/widm.1309.

5. Kavitha G., Raj L. Educational Data Mining and Learning Analytics Educational Assistance for Teaching and Learning. Internal Journal of Computer and Organization Trends (IJCOT). 2017. Vol. 7, issue 2. Pp. 21-25.

6. Flakh P. Mashinnoye obucheniye. Nauka i iskusstvo postroyeniya algoritmov, kotor·yye izvlekayut znaniya iz dannykh [Machine learning. The science and art of building algorithms that extract knowledge from data]. per. from eng. A.A. Slinkin. M.: DMK Press, 2015. 400 p.: ill.

7. Shumeyko A.A., Sotnik S.L. Intellektual'nykh analiz dannykh (Vvedeniye v Data Mining) [Data Mining (Introduction to Data Mining)]: Dnepropetrovsk: Belaya E.A., 2012. 212 p.

8. Mirmozaffari, M., Boskabadi, A., Azeem, G., Massah, R., Boskabadi, E., Dolatsara, H.A. and Liravian, A. European Journal of Engineering and Technology Research. 2020. Vol. 5, issue 6. Pp. 651–658.

9. Grankov M.V., Al'-Gabri V.M., Gorlova M.YU. Inzhenernyj vestnik Dona, 2016, №4. URL: ivdon.ru/magazine/archive/ /n4y2016/3775.

10. Dalwinder Singh, Birmohan Singh, Investigating the impact of data normalization on classification performance, Applied Soft Computing, Volume 97, Part B, 2020, 105524, ISSN 1568-4946. URL: sciencedirect.com/science/article/abs/pii/S1568494619302947?via%3Dihub.