УДК 004.056, 004.067

ПРИНЦИПЫ ОБРАБОТКИ НАБОРА ДАННЫХ, СОДЕРЖАЩЕГО АТАКИ РАЗНЫХ ВИДОВ

И.С. Поздняк
ФГБОУ ВО Поволжский государственный университет телекоммуникаций и информатики, г. Самара

Ключевые слова: набор данных, атаки, обработка данных.

Модели на основе машинного обучения позволяют выявлять скрытые зависимости в данных, недоступные для стороннего наблюдателя. Их использование позволит определять атаки более гибко по сравнению с аналитически разработанными строгими фильтрами и, что более важно, в следствие схожести признаков многих атак, потенциально позволит фиксировать даже те атаки, обнаружение которых не закладывалось в систему архитектурно. Для обучения классификаторов в рассматриваемой работе понадобятся размеченные данные, включающие в себя ряд признаков трафика и метку принадлежности к атаке.

В наборе данных CSE-CIC-IDS2018 [1] представлены по несколько разновидностей атак одного типа. Поскольку в данной работе акцент установлен на бинарную классификацию, различные вариации однотипных атак будут объединены в один класс. Таким образом, будет решаться задача оценки принадлежности трафика к типовой атаке, а не к её конкретной реализации.

Для обработки данных и представления их в необходимом и удобном формате следует проделать ряд процедур, представленных ниже, которые разделены на два больших этапа:

1. Первичный анализ данных.
2. Предварительная обработка данных.

На первом этапе необходимо сначала оценить единство формы представления данных. В случае разделённых на разные файлы таблиц важно удостовериться, что таблицы единообразны: имеют одинаковое число столбцов и схожие данные, представленные в них. Это необходимо для унификации последующей обработки. Несоответствия, вне зависимости от причины их возникновения, усложнят работу с данными как с единым целым. Далее важно избавиться от пропущенных значений на этапе предобработки данных перед непосредственным обучением классификаторов, так как модели машинного обучения не могут обучаться на записях, содержащих пропущенные значения.

После чего, начинается работа с признаками объекта, с которыми работают алгоритмы машинного обучения. Эти признаки в зависимости от своей природы, делятся на два вида: числовые и категориальные. В случае,

если признак не принадлежит ни к категориальному, ни к числовому (например, является произвольной строкой), его следует привести к одному из этих видов на этапе предобработки либо отбросить. Большинство значений набора данных CSE-CIC-IDS2018 отображают числовые признаки. Но помимо них в таблицах указан номер порта назначения, протокол передачи данных, штамп времени инициализации потока и метка принадлежности потока к атаке. Этим данным следует уделить особое внимание, так как не все целесообразно использовать для дальнейшей работы, либо следует изменить их вид.

Далее следует произвести оценку корреляционных характеристик признаков объекта. Простейшим способом оценить линейную корреляцию является попарный расчет коэффициента корреляции для каждого представленного числового признака [2]. В виду большой степени родства многих признаков, представленных в CSE-CIC-IDS2018, в корреляционной таблице ожидаемо присутствует много сильно коррелирующих признаков. Их нужно уменьшить, несмотря на то что они обладают некоторой, пусть и совсем небольшой, специфичностью и способны внести свой вклад в принятие решения.

Дополнительно с целью оптимизации набора данных можно попытаться определить признаки, не несущие в себе полезной информации. Таковыми оказались признаки, касающиеся величины «bulk rate» потоков трафика, и некоторые признаки, касающиеся вхождения флагов в пакеты, так же признак «port» аналогично не несет в себе полезной информации без дополнительных преобразований и многие другие признаки. В идеальной обучающей выборке количество записей, представляющих каждый класс, должно быть равным. Чем более несбалансированны представленные в выборке классы, тем более вероятно, что обученная на такой выборке модель всегда будет смещать свои предсказания в сторону доминирующего класса. Для улучшения качества обучения классификаторов в данной работе данные приводятся к идеальной классовой сбалансированности (50% записей с вредоносным трафиком и 50% записей с трафиком легитимным).

На этом первичный анализ можно считать законченным. Следующий этап будет сводиться к построению универсального конвейера, преобразующего данные в предпочтительную для машинного обучения форму и исправляющего выявленные проблемы. И на начальном этапе необходимо заполнить пропущенные значения. Для этого существуют различные варианты обработки. Выбор в пользу того или иного варианта следует делать в зависимости от природы пропуска. В простейшем случае, записи с пропущенными значениями в некоторых полях можно просто отбросить. Однако в нашем случае это не лучшее решение, потому что при приходе подобной записи на классификацию в IDS в реальном времени, классификатор не сможет её обработать. Поэтому автоматизируем процесс

заполнения пропусков нулями, отброса строк с некорректной меткой времени, замены бесконечных значений максимальными значениями соответствующих признаков.

Ряд алгоритмов машинного обучения лучше работает с признаками, значения которых лежат в узком диапазоне и не имеют большого разброса. Поэтому следует провести процесс нормализации. Ее не следует применять к категориальным данным, а значит признак «протокол» будет проигнорирован. Также не имеет особого смысла нормализировать признаки вхождения того или иного флага в поток, так как они и без этого принимают очень ограниченный диапазон значений. Метки классов тоже нормализировать не нужно. Описанные выше преобразования выполнялись в виде функций, для последующего объединения в конвейер предварительной обработки. Этот конвейер также представляет собой функцию, последовательно вызывающую другие, выполняя тем самым этапы предварительной обработки данных. В результате работы конвейера сырые данные будут переведены в вид, пригодный для непосредственного обучения классификаторов и/или дополнительных, более специфичных преобразований.

В дальнейшем предполагается обучить и оптимизировать различные модели машинного обучения для классификации сетевых атак.

Список использованных источников

1. Документация набора данных CSE-CIC-IDS2018 [Электронный ресурс]. – Режим доступа: http://www.unb.ca/cic/datasets/ids-2018.html

2. Харрисон, М. Машинное обучение: карманный справочник. Краткое руководство по методам структурированного машинного обучения на Python [Текст] / М. Харрисон; пер. В.А. Коваленко. - СПб.: Диалектика, 2020. – 320 с.

Поздняк Ирина Сергеевна, к.т.н., доцент каф. информационной безопасности, i.podnyak@psuti.ru.

УДК 621.396.96

АНАЛИЗ ПРИЗНАКОВ РАСПОЗНАВАНИЯ ТИПОВ ВОЗДУШНЫХ ЦЕЛЕЙ

Л.В. Симакова
«Самарский национальный исследовательский университет имени академика С.П. Королёва», г. Самара

Распознавание – это одно из важнейших направлений исследований в современной радиолокации. Решение задачи распознавания подразумевает получение радиолокационных характеристик выявленных объектов, оценку