УДК 004.891.3

ИССЛЕДОВАНИЕ ЗАДАЧИ КЛАССИФИКАЦИИ ШИФРОВАННОГО ТРАФИКА НА ОСНОВЕ НАБОРА ДАННЫХ ISCXVPN2016

Старун И.Г. (Университет ИТМО)

Научный руководитель – Юганссон А.Н. (Университет ИТМО)

Аннотация. В работе на основе набора данных ISCXVPN2016 подробно рассмотрена задача классификации шифрованного трафика с помощью анализа временных характеристик потока с дальнейшей обработкой методами машинного обучения. Проведено экспериментальное сравнение двух сценариев классификации – с предварительным разделением на VPN и non-VPN и без него. По большинству ключевых метрик классификации первый вариант является более результативным. Даны рекомендации по выбору значения тайм-аута при обработке потока. В частности, наилучшие результаты достигаются при использовании более коротких значений тайм-аута.

Введение

Активное распространение шифрования в Интернете и корпоративных сетях сделало актуальной задачу классификации шифрованного сетевого трафика. Ее решение необходимо как для обеспечения информационной безопасности, так и для изучения поведения пользователей, поиска аномалий в их действиях и контроля производительности в приложениях.

Традиционно выделяют три подхода к классификации трафика – на основе анализа портов, путем анализа полезной нагрузки и с помощью машинного обучения. Но при шифровании значительная часть полезной нагрузки трафика теряется, что затрудняет задачу классификации. В таких условиях на передний план выходит не анализ отдельных пакетов и их содержимого, а изучение потока трафика и его временных характеристик, таких как количество пакетов в единицу времени или длительность потока. С помощью машинного обучения могут быть достигнуты высокие показатели качества классификации на основе временных признаков.

Основная часть

Для исследования задачи классификации шифрованного трафика методами машинного обучения особую популярность имеет набор данных ISCXVPN2016, включающий в себя 14 категорий шифрованного трафика, полученных соединением через VPN и без него.

Этот трафик в дальнейшем анализируется как двунаправленный поток. Чтобы получить признаки для обучения моделей, он разбивается на отрезки одинаковой

временной продолжительности (timeout), по которым затем рассчитываются значения временных характеристик. Наиболее популярные значения тайм-аута – 15, 30, 60 и 120 секунд.

Как правило, рассматривают два основных сценария классификации на основе анализа потока:

1. Трафик предварительно классифицируется на VPN и non-VPN, а затем проводится раздельная классификация этих двух видов трафика по типам приложений и протоколов.
2. Весь датасет сразу классифицируется по типам приложений и протоколов без предварительного разделения на VPN и non-VPN.

Выводы

В результате исследования двух сценариев можно сделать следующие выводы:

- предварительное разделение на VPN и non-VPN позволяет получить значительно более качественную классификацию по всем ключевым метрикам – Accuracy, Recall, Precision;
- работа второго сценария в среднем занимает в 4 раза меньше времени, чем работа первого.

Изучение различных значений timeout при разбивке потока показало следующие тенденции:

- чем меньше значение тайм-аута, тем выше качество классификации;
- при увеличении значения тайм-аута растет и скорость классификации. Время работы модели при тайм-ауте в 15 секунд в 1,5–2,5 раза превышает аналогичный показатель при тайм-ауте в 120 секунд.