УДК 004.85

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ С УЧИТЕЛЕМ В РЕШЕНИИ ЗАДАЧ РАСПОЗНАВАНИЯ СЕТЕВОГО ТРАФИКА НА ПРИМЕРЕ РАСПОЗНАВАНИЯ ТРАФИКА GOOGLE

Сашин В.А.
(Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)
Научный руководитель – кандидат технических наук, доцент факультета БИТ Кузнецов А.Ю.
(Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

В работе проводится экспериментальное исследование применения методов машинного обучения с учителем для распознавания сетевого трафика на примере распознавания трафика Google. В ходе работы при помощи 9 методов машинного обучения был проанализирован набор данных, содержащий информацию о 3,5 миллионах сессиях (потоках трафика), и на основании выбранных критериев был определен метод, наиболее эффективный для решения поставленной задачи.

Введение. В настоящее время большинство межсетевых экранов (фаерволов), используемых для фильтрации сетевого трафика на предприятиях в целях предотвращения загрузки вредоносного ПО, нежелательного контента и несанкционированной передачи конфиденциальной информации, используют методы, не способные блокировать весь потенциально опасный трафик. Все больше приложений обходят ограничения фаерволов, например, шифруя свой трафик, маскируя его или в процессе работы изменяя используемые протоколы. Примером такого приложения является Skype – всемирно популярный мессенджер, который способен обходить ограничения фаерволов и может быть использован для несанкционированной передачи конфиденциальной информации. Система анализа трафика, способная распознать даже зашифрованный трафик приложений, обходящих ограничения межсетевых экранов, позволит не только закрыть один из каналов утечки информации, но и более эффективно фильтровать трафик организаций в целом: ограничить получение нежелательного или запрещенного контента, доступ к запрещенным ресурсам.

Основная часть. Для решения поставленной проблемы предлагается использовать методы машинного обучения, так как, по мнению ряда исследователей, они позволяют анализировать трафик на уровне сессий (потоков) без необходимости анализа содержимого каждого пакета и проверки IP-адресов и портов отправки и назначения.

Существует множество методов машинного обучения, поэтому для их сравнения и выбора наиболее подходящего для решения задачи распознавания трафика решено провести эксперимент – распознавание трафика Google методами машинного обучения с учителем на размеченном наборе данных, содержащем более 3,5 миллиона измерений (потоков трафика). Таким образом, была поставлена задача бинарной классификации. Для решения поставленной задачи необходимо сбалансировать соотношение классов, поэтому в рабочую выборку вошли 1,5 миллиона измерений. Зарубежные исследователи подчеркнули 21 полезную для классификации характеристику, именно они были использованы в ходе эксперимента.

В эксперименте применялись алгоритмы машинного обучения с учителем, условно разделенные на две группы. В первую группу вошли метод к-ближайших соседей, логистическая регрессия, метод опорных векторов, наивный байесовский классификатор, дерево принятия решений. Во вторую группу вошли такие ансамблевые методы, как случайный лес, а также методы градиентного бустинга: XGBoost, CatBoost, LightGBM. Выбор для исследования именно вышеприведенных методов обусловлен тем, что методы первой

группы достаточно просты, но при этом бывают эффективны и как правило применяются в первую очередь, а методы второй группы представляют особый интерес, поскольку именно они являются одними из самых популярных и эффективных методов на соревнованиях по анализу данных. Более того, перечисленные методы градиентного бустинга позволяют производить вычисления на графическом процессоре, а случайный лес – сразу на нескольких ядрах центрального процессора, что значительно повышает скорость обучения предсказательной модели.

Для сравнения эффективности применения методов машинного обучения с учителем решено использовать следующие методы: метод случайного поиска для подбора их оптимальных гиперпараметров и метод k-fold кросс-валидации для контроля качества обучившихся моделей.

Для оценки эффективности работы полученных в ходе исследования моделей на тестовой выборке были использованы такие метрики, как площадь под кривой ошибок (AUC-ROC), точность (precision), полнота (recall) и F-мера.

Выводы. Предложенный метод анализа потоков (сессий) сетевого трафика благодаря использованию методов машинного обучения позволит распознавать даже зашифрованный и способный обходить ограничения межсетевых экранов трафик, что позволит закрыть один из каналов утечки конфиденциальной информации и более эффективно фильтровать нежелательный контент.

В рамках работы был проведен эксперимент, позволивший выявить самый эффективный метод машинного обучения для решения поставленной задачи на примере распознавания трафика Google. По итогам эксперимента лучшим был признан метод градиентного бустинга XGBoost, показавший следующие результаты: площадь под кривой ошибок – 0,94, точность, полнота и F-мера – 0,86. Таким образом, экспериментальным путем была подтверждена эффективность использования методов машинного обучения для распознавания трафика и был выбран метод, наиболее подходящий для решения этой задачи.

Сашин В.А. (автор) Подпись

Кузнецов А.Ю. (научный руководитель) Подпись