УДК: 004.855.5

ИСПОЛЬЗОВАНИЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБНАРУЖЕНИЯ ВРЕДОНОСНОГО ТРАФИКА

Турышев Артём Андреевич, Назаров Дмитрий Михайлович
Уральский государственный экономический университет, город Екатеринбург, Россия
artem.turyshev@vk.com

Аннотация. В статье описывается возможность применения машинного обучения к анализу данных сетевого трафика с целью классификации транзакций сетевого трафика по описанным классам. Автором был применен алгоритм случайного леса для реализации процесса классификации. Коэффициент ошибок алгоритма на 1000 деревьев составил 74,5%.

Ключевые слова. *Машинное обучение, сетевой трафик, прогнозирование, безопасность, классификация, кластерный анализ.*

USING MACHINE LEARNING TO DETECT HARMFUL TRAFFIC

Turyshev Artem Andreevich, Nazarov Dmitry Mikhailovich
Ural State University of Economics, Yekaterinburg, Russia

Annotation. This article describes the possibility of applying machine learning to the analysis of network traffic data in order to classify network traffic transactions into the described classes. The author has applied a random forest algorithm to implement the classification process. The algorithm error rate per 1000 trees was 74.5%.

Keywords. Machine learning, network traffic, forecasting, security, classification, cluster analysis.

За последние 20 лет количество пользователей сетей во всем мире увеличилось кратно. Это вызвало резкое увеличение трафика, которое пользователи сети генерируют в результате своей работы. При этом естественно выросло и число вредоносных программ, нарушающих конфиденциальность пользователей, собирающих данные, связанные с доступом к онлайн-магазинам и платежным аккаунтам, шифрующих файлы пользователей с целью получения выкупа. Большинство инцидентов такого плана фиксируется службами информационной безопасности соответствующих компаний и за это время накоплено достаточное число данных об этих инцидентах, которые связаны с характеристиками сетевого трафика в том числе. В условиях цифровой

трансформации и наличия достаточного количества данных для их анализа начинают применять аппарат машинного обучения.

Проблема применения алгоритмов машинного обучения для проведения анализа сетевого трафика становится важнейшим направлением в исследовании и поиске инцидентов информационной безопасности, как во всем мире, так и в отечественной практике.

Цель работы – исследовать и провести классификацию данных сетевого трафика, представленных в датасете NSL-KDD с использованием алгоритма машинного обучения – случайный лес и реализовать этот процесс в среде R

Многолетний опыт анализа сетевого трафика показывает, что полностью защитить все хосты в сети маловероятно, то есть неизбежно некоторые машины будут скомпрометированы вредоносным ПО. Поэтому специалисты в сфере информационной безопасности и системные администраторы должны выстраивать политику обеспечения безопасности хостов в корпоративных сетях опираясь на теории защиты сетей, вместо того, чтобы полагаться на случайное обнаружение зараженных машин в своих сетях.

Однако обнаружение вредоносных программ в сетевом трафике представляет собой ряд проблем:

во-первых, обнаружение вредоносного ПО во время заражения осложняется множеством векторов атак, которые могут использовать для заражения хоста. Электронная почта, USB-накопители и веб-интерфейс атаки – это лишь некоторые из распространенных механизмов, используемых авторами вредоносных программ для доставки своих полезных нагрузок;

во-вторых, однажды установленная на машине, вредоносная программа может выполнять различные вредоносные действия, такие как мошенничество, кража данных, DDoS-атаки или рассылка спама. Некоторые из этих поведений может быть неотличим от легального трафика, например, мошенничество или кража данных.

Модель защиты сетей сложна из-за широкого диапазона поверхностей атак и векторов потенциальных угроз. Как и при защите любой сложной системы, администраторы должны противостоять атакующим на многих направлениях и не полагаться на надежность какого-либо одного компонента решения.

Алгоритм случайного леса был предложен Лео Брейманом и Адель Катлер. Он сочетает в себе базовые принципы бэггинга со случайным выбором признаков, что позволяет увеличить разнообразие в моделях деревьев решений. После генерации деревьев (леса) модель объединяет прогнозы отдельных деревьев путем голосования. Бутстрэп-агрегирование или бэггинг, это метаалгоритм композиционного обучения машин, предназначенный для

улучшения стабильности и точности алгоритмов машинного обучения, используемых в статистической классификации и регрессии.

Результаты работы алгоритма для 1000 деревьев представлены на рисунке 1.

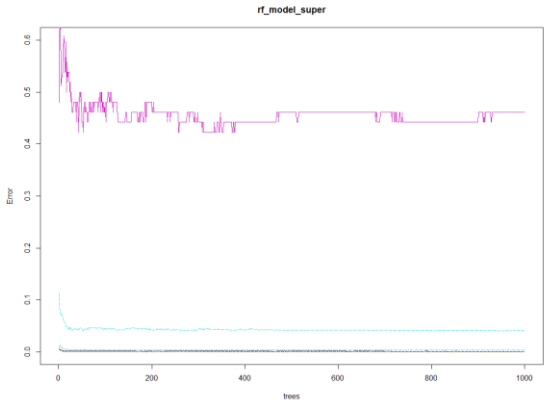

Рисунок 1 – Коэффициенты ошибок классов в ходе машинного обучения на 1000 деревьев.

Процент верных прогнозов составил примерно 74,5%. Примерно, до 150 дерева наблюдается сильный разброс коэффициента ошибок это обусловлено тем, что в процессе работы алгоритм учится на данных.

Таким образом, в ходе работы был изучен алгоритм машинного обучения – случайный лес, который был применен к набору данных о сетевых атаках для опций 500, 600, 700, 800, 900, 1000 деревьев. Результат работы алгоритма удовлетворительный всего74,5%, причем увеличение числа деревьев не привело к значимым изменениям коэффициента ошибок.

Библиографический список

1. G. Cormode, F. Korn, S. Muthukrishnan, and D. Srivastava. Diamonds in the rough: Finding hierarchical heavy hitters in multidimensional data. In Proc. SIGMOD, 2004.

2. M. Datar and S. Muthukrishnan. Computing rarity and similarity over data streams. In Proceedings ESA, 2002.

3. R. Duda, P. Hart, and D. Stork. Pattern Classification. Wiley Interscience, 2nd Edition, 2000.

4. J. Z. Kolter and M. A. Maloof, “Learning to detect and classify malicious executables in the wild,” Journal of Machine Learning Research, vol. 7, p. 2006, 2006.

5. E. Gandotra, D. Bansal, and S. Sofat, “Malware analysis and classification: A survey,” Journal of Information Security, vol. 5, pp. 56–64, 2014.