УДК 004.93'1

ВЫЯВЛЕНИЕ СЕТЕВОЙ РАЗВЕДКИ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Шараев Н.П., магистрант гр. 967241

Белорусский государственный университет информатики и радиоэлектроники[1]
г. Минск, Республика Беларусь

Петров С.Н. – канд. тех. наук

Аннотация. Сетевая разведка является первой стадией таргетированной или APT атаки, обнаружение которой позволит заблаговременно выполнить поиск возможных уязвимостей и предпринять меры по снижению рисков. Среди возможных унифицированных методов проведения сетевой разведки выделяются сканирование информационной сети и портов транспортного уровня. Процесс обнаружения данных типов сканирования основан на алгоритмах машинного обучения, в частности, методах классификации, кластеризации и ансамблирования. Обучающий датасет генерируется на базе сетевого трафика, в котором присутствуют отдельные пакет (сегменты) сетевой разведки.

Ключевые слова. Сетевая разведка, APT атака, машинное обучение.

В последнее время наблюдается тенденция перехода от массовых кибератак отдельных злоумышленников к масштабным атакам киберпреступных группировок на конкретные организации (таргетированные или APT атаки). Данные атаки в значительной мере опасны для организаций, что связано в первую очередь с созданием злоумышленниками вредоносного программного обеспечения с учетом специфики работы и сетевой инфраструктуры организации. В общем случае, APT атаки состоят из четырех этапов: подготовка, проникновение, распространение и достижение цели [1]. Обнаружить подобный тип атак на поздней стадии крайне сложно, а в отдельных случаях невозможно. По данной причине целесообразно провести обнаружение и анализ таргетированной атаки на этапе подготовки. На указанном этапе злоумышленники проводят процедуру сетевой разведки инфраструктуры организации. Сетевая разведка – это комплекс мероприятий, направленных на получения сведений об информационных системах, средствах защиты информации и используемом программном обеспечении [2].

Сетевая разведка может проводиться следующим образом:

– получение информации от whois-серверов (контактные данные владельца доменного имени, список DNS серверов и другое);

– получение информации от DNS серверов

– сканирование сети;

– сканирование портов транспортного уровня.

Наибольший интерес для обнаружения сетевой разведки представляют последние два пункта: сканирование информационной сети и портов транспортного уровня (так как невозможно повлиять на первые два пункта). Для выявление данных способов проведения сетевой разведки с помощью машинного обучения проведен анализ генерируемого ими сетевого трафика и выделены метрики, представленные в таблице 1 [3].

Таблица 1 – Анализируемые метрики

	Название метрики	Описание метрики
	count	Отношение количества отправленных сегментов (дейтаграмм) с одного IP адреса к общему количеству сегментов (дейтаграмм) с различных IP адресов.
	udp	Отношение количества отправленных дейтаграмм с одного IP адреса к общему количеству отправленных с этого же IP адреса сегментов (дейтаграмм).
	tcp	Отношение количества отправленных сегментов с одного IP адреса к общему количеству отправленных с этого же IP адреса сегментов (дейтаграмм).
	tcp_syn	Отношение количества отправленных с указанным флагом сегментов (SYN, ACK, FIN, NULL, XMAS, MAIMON, OTHER) с одного IP-адреса к общему количеству отправленных с этого же IP адреса сегментов.
	tcp_ack	
	tcp_fin	
	tcp_null	
	tcp_xmas	
	tcp_maimon	

	tcp_other	
	uniq_ports	Отношение количества уникальных портов, на которые были отправлены сегменты с одного IP адреса, к общему количеству отправленных с этого же IP адреса сегментов.

На основе указанных метрик сгенерировано два набора данных (датасета). Оба датасета представлены в формате JSON в виде словарей. Первый датасет состоит из 4025 событий, является лабораторным и генерируется автоматически на основании функции псевдослучаных чисел (random). Второй датасет состоит из 600 событий и является эмпирическим, то есть основанным на реальном трафике. Два датасета разработаны по той причине, что создание одного качественного датасета с большим количеством эмпирических событий сетевой разведки крайне сложно и требует значительного промежутка времени. Под качеством имеется ввиду наличие событий, полученных от различных средств проведения сетевой разведки. В настоящее время наиболее популярными утилитами для его проведения являются Nmap (Windows и Linux) и masscan (Linux), что недостаточно для создания разнообразия.

Визуализации обоих датасетов с помощью метода главных компонент (PCA) представлены на рисунках 1 и 2.

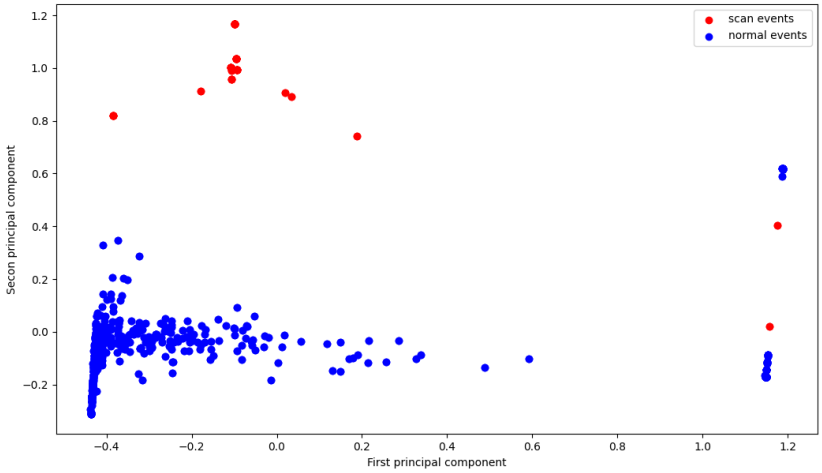

Рисунок 1 — Метод главных компонент для эмпирического датасета

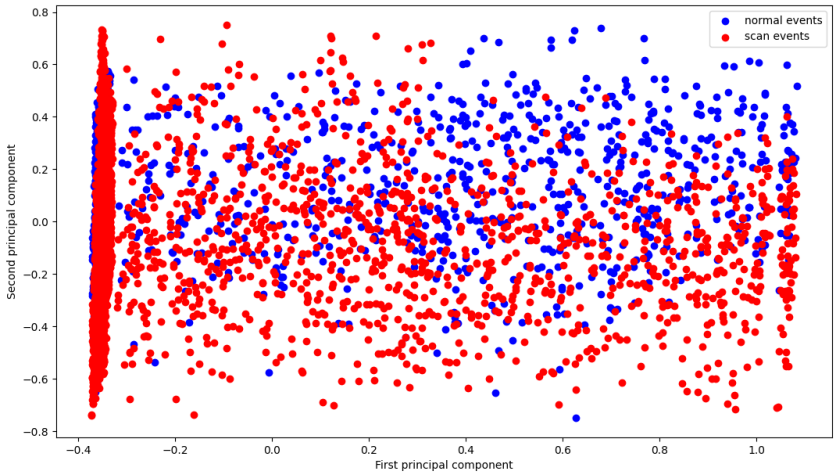

Рисунок 2 – Метод главных компонент для лабораторного датасета

Для обнаружения факта сетевой разведки выбраны методы машинного обучения, решающие задачи классификации и кластеризации. В основе алгоритмов классификации лежит тип машинного обучения с учителем, то есть создание правил по размеченным данным. Для целей настоящей работы используются следующие методы классификации: метод опорных векторов, метод логистической регрессии, гауссовский наивный байесовский классификатор, дерево принятий решений и многослойный персептрон. Данные алгоритмы были выбраны по причине простоты и скорости работы, а также высокой эффективности. В свою очередь решение задачи кластеризации

базируется на обучении без учителя (создание правил только по входным данным). В качестве методов кластеризации реализованы следующие алгоритмы: изолированный лес, метод обнаружение выбросов в гауссовском распределенном наборе данных, метод K-средних, пространственная кластеризация приложений с шумом на основе плотности (DBSCAN), метод сбалансированного итеративного сокращения и кластеризации с помощью иерархий (BIRCH). Указанные методы используются по причине их популярности, относительной простоты и скорости работы.

При обучении и тестировании эффективности методов машинного обучения оба датасета были дифференцированы на обучающую и тестовую выборку. Процентное соотношение для указанных выборок составляет 80% и 20% от общего количества событий соответственно. Для всех методов проведено обучение и измерение точности (accuracy). Точности обнаружения для лабораторного и эмпирического датасетов представлены в таблице 2.

Таблица 2 – Анализируемые методы машинного обучения

	Название метода	Сокращение	Тип метода	Точность лаб. датасета, %	Точность эмпир. датасета, %
	Метод опорных векторов	SVM	Классификация	92,42	100,00
	Метод логистической регрессии	LR		84,22	99,12
	Многослойный персептрон	MLP		92,52	100,00
	Гауссовский наивный байесовский классификатор	GaussianNB		76,40	99,12
	Дерево принятия решений	DecsicionTree		98,63	100,00
	Изолированный лес	IsolationForest	Кластеризация	61,87	87,72
	Метод обнаружения выбросов в гауссовском распределенном наборе данных	EllipticEnvelope		66,50	91,23
	Метод K-средних	KMeans		63,52	70,18
	Пространственная кластеризация приложений с шумом на основе плотности	DBSCAN		69,63	92,98
	Метод сбалансированного итеративного сокращения и кластеризации с помощью иерархий	BIRCH		65,26	98,25

Самыми перспективными методами обнаружения для классификации и кластеризации являются Дерево принятий решений и метод BIRCH соответственно. Тем не менее, эффективность указанных выше алгоритмов можно увеличить и стабилизирована путем использования ансамблей. Одинаковые алгоритмы (например, задачи классификации) могут быть объединены в один ансамбль, предназначенный для исправления ошибок друг друга. Так, несколько не очень эффективных методов обучения могут показать результат выше, чем каждый

метод в отдельности. При этом в ансамбль обычно объединяют алгоритмы максимально не стабильные и сильно зависящие от входных данных, в частности, регрессию и дерево принятия решений. Данная практика позволяет стабилизировать их результат несмотря на возможное наличие сильных аномалий в множестве входных объектов. Выделяют следующие виды ансамблей [4]:

– стекинг (stacking);
– бэггинг (bootstrap aggregating);
– бустинг (boosting).

Дополнительное применение алгоритмов бустинга (AdaBoost) или бэггинга для алгоритма Дерево принятия решения показало уменьшение точности (в среднем на 0,6%), что связано со стабилизацией результатов работы алгоритма. Применение указанных методов на алгоритм BIRCH не дало результатов по причине неустойчивости алгоритмы AdaBoost к выбросам. В дальнейшем планируется перейти на алгоритм градиентного бустинга (GBM).

Стоит отметить, что формально метод BIRCH является более подходящим для обнаружения факта сетевой разведки, так как нацелен на поиск аномалий в сетевом трафике. В то же время алгоритм Дерева принятия решений дает большую точность и позволяет описать условия обнаружения сетевой разведки языком программирования. В этой связи для дальнейших исследований предлагается выбрать лучший метод обнаружения признаков сетевой разведки используя практический подход, то есть разработать программное обеспечение, перехватывающее сетевой трафик и анализирующее его с использованием двух описанных алгоритмов.

Список использованных источников:

1. Левцов, В. Ю. Анатомия таргетированной атаки. Часть 1 / В.Ю. Левцов, П. Демидов // Системный администратор. – 2016. – №4 (161).

2. Шараев, Н. П. Обнаружение признаков сетевой разведки с использованием машинного обучения / Шараев Н. П., Петров С. Н. // Современные средства связи : материалы XXV Междунар. науч.-техн. конф., 22–23 окт. 2020 года, Минск / Белорусская государственная академия связи ; редкол.: А. О. Зеневич [и др.]. – Минск : БГАС, 2020. – С. 209-210.

3. Шараев, Н. П. Выявление и анализ признаков сетевой разведки методом машинного обучения / Шараев Н. П., Петров С. Н. // Управление информационными ресурсами: материалы XVII Междунар. науч.-практ. конф., 12 мар. 2021 года, Минск / Акад. упр. при Президенте Респ. Беларусь ; редкол. : А. С. Лаптенок. – Минск: Академия управления при Президенте Республики Беларусь, 2021. – С. 238-240.

4. Игнатов Д. И. Ансамблевый метод машинного обучения, основанный на рекомендации классификаторов / Д. И. Игнатов, Ю. С. Кашницкий // Интеллектуальные системы. Теория и приложения. – 2015. – Т. 19. – № 4. – С. 37-55.

UDC 004.93'1

IDENTIFICATION OF NETWORK INTELLIGENCE BY MACHINE LEARNING METHODS

Sharaev N.P., Master Student of the group 967241

Belarusian State University of Informatics and Radioelectronics
Minsk, Republic of Belarus

Petrov S.N. – PhD

Annotation. Network reconnaissance is the first stage of a targeted or APT attack, the detection of which will allow you to search for possible vulnerabilities in advance and take measures to mitigate the risks. Among the possible unified methods of conducting network reconnaissance, scanning the information network and ports of the transport layer stand out. The detection process for these scan types is based on machine learning algorithms, in particular, classification, clustering and ensemble methods. The training dataset is generated on the basis of network traffic, in which there are separate packet (s) of network intelligence.

Keywords. Network intelligence, APT attack, machine learning.