ВЫЯВЛЕНИЕ И АНАЛИЗ ПРИЗНАКОВ СЕТЕВОЙ РАЗВЕДКИ МЕТОДОМ МАШИННОГО ОБУЧЕНИЯ

Шараев Н. П., Петров С. Н.
Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники» (г. Минск)

В настоящее время количество таргетированных атак (APT-атак), направленных на информационные системы организаций непрерывно растет. Вместе с тем повышается значимость превентивных мер противодействия указанному типу атак, в частности, защиты информации о функционирующих в организациях информационных системах. Основным способом защиты указанной информации является противодействие сетевой разведке. Сетевая разведка – это комплекс мероприятий, направленных на получение сведений об информационных системах, средствах защиты информации и используемом программном обеспечении. В связи с тем, что сетевая разведка является первым звеном APT-атаки и предваряет собой активные действия, ее обнаружение позволяет заблаговременно выполнить поиск возможных уязвимостей и предпринять меры по снижению рисков.

Для обнаружения признаков сетевой разведки с помощью машинного обучения из трафика были выделены следующие метрики (таблица).

На основе указанных метрик был создан предварительный датасет, состоящий из 600 событий в виде словаря в формате JSON. Датасет сформирован эмпирическим способом, то есть путем проведения сетевой разведки на тестовой информационной системе и снятием с нее дампа сетевого трафика. Для работы с массивами данных выбран язык программирования Python, включая библиотеку numpy для извлечения признаков из набора данных.

Для последующего обнаружения признаков сетевой разведки, датасет был визуализирован в виде графика в двумерном пространстве. Для этого проводилось уменьшение размерности входного вектора датасета с одиннадцати пространств (в соответствии с числом метрик)

до двух, с использованием метода главных компонент (PCA). В результате уменьшения количества пространств можно наблюдать дифференциацию выборки на два условных подмножества. Одно подмножество состоит из легитимных событий и обращений к тестовой информационной системе, а второе – из событий, связанных с сетевой разведкой. Тем не менее, часть событий находится в промежуточном диапазоне, что затрудняет их определение.

Таблица

Анализируемые метрики

№	Название метрики	Описание метрики
1	count	Отношение количества отправленных сегментов (дейтаграмм) с одного IP-адреса к общему количеству сегментов (дейтаграмм) с различных IP-адресов.
2	udp	Отношение количества отправленных дейтаграмм с одного IP-адреса к общему количеству отправленных с этого же IP-адреса сегментов (дейтаграмм).
3	tcp	Отношение количества отправленных сегментов с одного IP-адреса к общему количеству отправленных с этого же IP-адреса сегментов (дейтаграмм).
4-10	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	
11	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.

Для разделения событий, находящихся в промежуточном диапазоне, предлагается использовать методы классификации, основанные на обучении с учителем, и кластеризации. С точки зрения формального подходка к обеспечению информационной безопасности методы кластеризации являются более подходящими, так как нацелены на поиск аномалий в сетевом трафике (например, метод К-средних). В то же время алгоритмы классификации дают большую точность и позволяют формально описать условия обнаружения сетевой разведки. Дополнительно, для повышения точности указанных методов можно применить

алгоритмы бустинга (boosting) или бэггинга (bagging), предназначенные для последовательного построения композиции алгоритмов машинного обучения.

Полученные в результате исследования данные, в частности, датасет, можно применить при построении модели нейронной сети, а также при ее непосредственном обучении и оценке точности.