УДК 004.056.5

О. С. Терновой, А. С. Шатохин

Использование байесовского классификатора для получения обучающих выборок, позволяющих определять вредоносный трафик на коротких интервалах

O. S. Ternovoy, A. S. Shatokhin

Using a Bayesian Classifier for Training Samples Allowing to Determine the Malicious Traffic in Short Intervals

Представлен оригинальный алгоритм по классификации вредоносного и благонадежного трафика. В основе алгоритма лежит использование теоремы Байеса и байесовского классификатора. Применение данного алгоритма позволяет получать обучающие выборки, которые могут быть использованы для обучения нейронных сетей и других различных классификаторов, а также для фильтрации нежелательного трафика.

Ключевые слова: DDOS атака, бот сеть, байесовский классификатор, статистический анализ.

The paper presents an original algorithm for the classifying bad traffic and legitimate traffic. Algorithm is based on the use of Bayes' theorem and Bayesian classifier. The use of this algorithm allows to obtain training samples, which can be used for training neural networks, and various other classifiers, as well as to filter out unwanted traffic.

Key words: DDOS attack, bot network, Bayesian classifier, statistical analysis.

Введение. Постановка задачи. DDOS атака — распределенная атака, направленная на отказ в обслуживании. В результате атаки такого типа атакуемый сетевой ресурс получает лавинообразное количество запросов, которые не успевает обработать. Источником вредоносных запросов являются так называемые зомби сети, состоящие большей частью из компьютеров обычных пользователей, в силу каких-то причин зараженных вредоносным программным обеспечением [1].

Для создания фильтров, позволяющих отсечь вредоносный трафик, применяются разнообразные методы, в основе которых лежат математическая статистика, определение поведенческих факторов, качественный и количественный анализ поступающего трафика и т. д. Одним из самых перспективных методов определения вредоносного трафика являются методы, основанные на статистических классификаторах и нейронных сетях. Эти методы показывают хорошие результаты, определяя вредоносный трафик с высокой точностью [2, 3]. Но для их успешной работы необходимо иметь две актуальные обучающие выборки, соответствующие вредоносному и благонадежному трафику. Обе эти выборки не могут быть получены до начала атаки. Причем если невозможность получения выборки вредоносного трафика очевидна, то вопросы актуальности выборки с благонадежным трафиком не всегда понятны. Невозможно использовать выборку недельной или месячной давности, так как за этот период сетевая картина может измениться. Использовать в качестве выборки последние *n* значений, предшествующих началу атаки, не всегда возможно, так как есть вероятность, что в эту выборку попадут неблагонадежные запросы и в процессе обучения неблагонадежный трафик будет отнесен к легитимному.

Весь трафик, который приходит после начала атаки, представляет собой смесь трафика из двух групп — целевого и неблагонадежного. Именно из этого трафика необходимо выделить выборку вредоносного трафика. Во многих системах предотвращения DDOS атак эта выборка создается в ручном или полуавтоматическом режимах. После начала атаки администратор ожидает накопления отрицательной статистики и только после этого начинает помечать трафик как неблагонадежный. Причем если структура нежелательных запросов меняется, администратору необходимо вновь пометить приходящие запросы как неблагонадежные. Минусами такого полуавтоматического режима являются низкая скорость реакции, достаточно большая доля погрешности.

Для решения этих проблем необходим подход, который бы позволил в автоматическом режиме создавать обучающие выборки, а также поддерживать их актуальность во время проведения атаки.

Ход исследования. Каждый сетевой клиент имеет определенный набор свойств. Для различных сетевых сервисов эти наборы могут быть разными, но, как правило, в них входят IP адрес клиента, тип запроса, длительность сессии, скорость поступления запросов, время запроса, целевой ресурс и т. д.

Пусть множество A (a_1, a_2, a_3,…a_n) — это набор всех возможных свойств для всех сетевых клиентов. Множество B (b_1, b_2, b_3,…b_m) — это множество сете-

вых клиентов какого-то конкретного ресурса. Каждый сетевой клиент обладает набором индивидуальных свойств. Например, клиент b_1 имеет свойства A1 (a_4, a_8, a_{11}, a_{14}), клиент b_1 — свойства A2 (a_3, a_8, a_{10}, a_{14}) и т. д. Эти свойства представляют собой набор подмножеств множества A. Пересечение всех этих подмножеств характеризует клиентов сетевого ресурса, по которым они могут быть классифицированы. Точно так же неблагонадежные клиенты будут иметь свой набор свойств, по которому они могут быть классифицированы.

Так, например, в прошлом был популярен прием фильтрации злонамеренных запросов к web-серверу по типу загружаемых данных [4]. Дело в том, что благонадежные клиенты при загрузке web-страницы также загружают сопутствующие данные: картинки, каскадные страницы стилей, скрипты, размещенные во внешних файлах, и т. д. Зомби-компьютеры большей частью обращаются только к интересующим их скриптам и игнорируют эти данные.

Естественно, что злоумышленники пытаются имитировать свойства легитимных клиентов. И на сегодняшний момент многие зомби-компьютеры также пытаются загружать все сопутствующие данные, делая таким образом отслеживание этого свойства нерезультативным. Поэтому при классификации приоритет следует отдать только тем свойствам, которые не могут быть подделаны злоумышленником. При недостатке таких свойств необходимо ввести искусственные свойства. Например, таким свойством может являться успешное отгадывание картинки, которое позволяет определить автоматические запросы.

Наиболее точно определить момент начала атаки позволяет метод раннего обнаружения DDOS атак, учитывающий сезонные колебания [5]. Точное определение начала атаки позволяет считать предшествующий трафик благонадежным и отнести его к соответствующей выборке.

Трафик же, поступаемый после этого момента, будет состоять из благонадежного и вредоносного трафика.

Пусть множество T — это множество клиентских запросов, поступаемых до начала атаки. Множество клиентских запросов, поступаемых после начала атаки, — это объединение множеств H — вредоносные клиентские запросы — и множества T^* — благонадежные клиентские запросы. Таким образом, для получения выборки с вредоносным трафиком необходимо будет разделить трафик, получаемый после начала атаки, на две группы.

Для такого разделения автором создан специальный алгоритм, который позволяет с большой точностью разделить трафик на благонадежный и вредоносный и в дальнейшем использовать эти данные в виде обучающих выборок.

На первом шаге алгоритм производит предварительную кластеризацию. Так как число кластеров заранее известно — благонадежный и неблагонадежный трафик, то в данном случае оптимально воспользоваться методом кластеризации k-means (метод k средних). Данный метод позволяет проводить кластеризацию при заранее известном числе кластеров.

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2,$$

где ***k*** — число кластеров; ***Si*** — полученные кластеры; ***i=1,2...,k*** и μ_i — центры масс векторов $x_j \in S_i$.

Суть метода заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике.

Шаг завершается, когда на какой-то итерации не происходит изменения кластеров. Это происходит за конечное число итераций, так как количество возможных разбиений конечного множества конечно, а на каждом шаге суммарное квадратичное уклонение V уменьшается, поэтому зацикливание невозможно [6].

На втором шаге алгоритма происходят основные расчеты и окончательная кластеризация. Критериями успеха кластеризации будут являться два условия:

— размерность полученных кластеров T^* и H. Если в период атаки число запросов за определенный период составляет n, а число запросов за аналогичный период, предшествующий началу атаки, — m, то можно предположить, что количество вредоносных запросов будет $n - m$. Этот критерий позволяет отсечь вредоносные запросы, которые злоумышленник пытается выдать за легитимный трафик;

— максимальная схожесть благонадежного трафика, полученного в результате разделения всего трафика, поступаемого после начала атаки (множество T^*), с благонадежным трафиком, соответствующим началу атаки (множество T).

На этом шаге рассчитывается вероятность принадлежности каждого клиентского запроса к своему классу. После этого элементы в группах сортируются в порядке убывания вероятности. Элементы с наименьшей вероятностью переносятся в противоположные группы с учетом критерия размерности групп.

Для расчета принадлежности элементов множества T^* к вредоносному или благонадежному трафику используется «наивный байесовский классификатор».

Абстрактно, вероятностная модель для классификатора — это условная модель

$$p(C|F_1, \ldots, F_n)$$

над зависимой переменной класса C с малым количеством результатов или *классов*, зависимая от нескольких переменных $F_1 \ldots F_n$. Проблема заключается в том, что когда количество свойств n очень велико или когда свойство может принимать большое количество значений, тогда строить такую модель на вероятностных таблицах становится затруднительно. Поэтому мы переформулируем модель, чтобы сделать ее легко поддающейся обработке [7].

Используя теорему Байеса, запишем

$$p(C|F_1, \ldots, F_n) = \frac{p(C)\ p(F_1, \ldots, F_n|C)}{p(F_1, \ldots, F_n)}.$$

На практике мы заинтересованы лишь в числителе этой дроби, так как знаменатель не зависит от C и значения свойств F_n даны, так что знаменатель — константа.

Числитель эквивалентен совместной вероятности модели

$$p(C, F_1, \ldots, F_n),$$

которая может быть переписана следующим образом, используя повторные приложения определений условной вероятности:

$$\begin{aligned}
&p(C, F_1, \ldots, F_n) = \\
&= p(C)\ p(F_1, \ldots, F_n|C) = \\
&= p(C)\ p(F_1|C)\ p(F_2, \ldots, F_n|C, F_1) = \\
&= p(C)\ p(F_1|C)\ p(F_2|C, F_1) * \\
&\quad * p(F_3, \ldots, F_n|C, F_1, F_2) = \\
&= p(C)\ p(F_1|C)\ p(F_2|C, F_1)\ p(F_3|C, F_1, F_2) * \\
&\quad * p(F_4, \ldots, F_n|C, F_1, F_2, F_3)
\end{aligned}$$

и т. д. Теперь начинаем использовать «наивные» предположения условной независимости: предположим, что каждое свойство F_i условно независимо от любого другого свойства F_j при $j \neq i$. Это означает $p(F_i|C, F_j) = p(F_i|C)$.

Таким образом, совместная модель может быть выражена как $p(C, F_1, \ldots, F_n) = p(C) *$

$$\begin{aligned}
&* p(F_1|C)\ p(F_2|C)\ p(F_3|C) = \\
&= p(C) \prod_{i=1}^{n} p(F_i|C).
\end{aligned}$$

Это означает, что из предположения о независимости условное распределение по классовой переменной C может быть выражено так [7]:

$$p(C|F_1, \ldots, F_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(F_i|C).$$

Таким образом, для классификации трафика по двум классам имеем:

$P(T \mid D) = \frac{P(T)}{P(D)} \prod_{i=1}^{n} P(w_i \mid T)$ — для класса благонадежных пользователей;

$P(H \mid D) = \frac{P(H)}{P(D)} \prod_{i=1}^{n} P(w_i \mid H)$ — для класса неблагонадежных пользователей.

В качестве обучающих выборок используются множества T и H. По завершении этого шага элементы из множества T*, отнесенные к группе вредоносного трафика, меняются местами с элементами множества H с учетом указанных выше критериев.

Этот шаг повторяется до тех пор, пока все элементы множества T не будут помечены как благонадежные либо пока алгоритм не достигнет порогового значения интеракций.

В дальнейшем данные выборки можно использовать в качестве обучающих для нейронных сетей и классификаторов, например для байесовского классификатора.

Данный подход позволяет начать классифицировать трафик уже на коротких интервалах. Причем вредоносный трафик будет являться вредоносным независимо от размера временного периода. Таким образом, объединяя данные о вредоносном трафике для временных периодов (5 мин, 30 мин, 1 час и т. д.), можно повысить точность обучающих выборок.

Библиографический список

1. DDOS атаки [Электронный ресурс]. — URL: http://localname.ru/soft/ataki-tipa-otkaz-v-obsluzhivanii-dos-i-raspredelennyiy-otkaz-v-obsluzhivanii-ddos.html

2. Предотвращение атак с распределенным отказом в обслуживании (DDoS) [Электронный ресурс] / Официальный сайт компании Cisco. — URL: http://www.cisco.com/web/RU/products/ps5887/products_white_paper0900aecd8011e927_.html

3. Обнаружение DDoS атак нечеткой нейронной сетью [Электронный ресурс]. — URL: http://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=isu&paperid=67&option_lang=rus

4. Простой способ защиты от классического HTTP DDoS [Электронный ресурс]. — URL: http://habrahabr.ru/post/151420/

5. Терновой О. С. Раннее обнаружение DDOS атак методами статистического анализа // Перспективы развития информационных технологий. — Новосибирск, 2012.

6. k-means [Электронный ресурс]. — URL: http://ru.wikipedia.org/wiki/K-means

7. Наивный байесовский классификатор [Электронный ресурс]. — URL: http://ru.wikipedia.org/wiki/