УДК 004.5

ПРИМЕНЕНИЕ ЭНТРОПИИ ДЛЯ ОБНАРУЖЕНИЯ АНОМАЛИЙ СЕТЕВОГО ТРАФИКА

М. С. Абрамович, Л. В. Гарматная

Белорусский государственный университет, Научно-исследовательский институт прикладных проблем математики и информатики, пр. Независимости, 4, к.427, 220030, г. Минск, Беларусь, abramovichms@bsu.by

Приводится концепция нормального и аномального сетевого трафика. Рассматриваются энтропии Шеннона, Цаллиса и Реньи. Построены информативные признаки для классификации сетевого трафика. Разработан алгоритм классификации аномального и нормального трафика на основе энтропии. Исследована точность классификации сетевого трафика методами машинного обучения с использованием различных видов энтропии.

Ключевые слова: сетевой трафик; аномалии трафика; энтропия; обнаружение аномалий трафика; машинное обучение; эффективность классификации.

Введение

В связи с активным развитием и использованием инфокоммуникационных технологий возникает необходимость быстрой и надежной передачи данных пользователей. По сети передается конфиденциальная и ценная информация, связанная с электронной коммерцией, банковским делом и бизнесом. В связи с этим возникает необходимость анализа сетевого трафика с целью быстрого и эффективного выявления сетевых угроз.

Компьютерные системы безопасности сосредоточены на защите данных, обеспечивая конфиденциальность, целостность и доступность. Основной целью защиты ресурсов является обнаружение состояний сети, которые не удовлетворяют вышеупомянутым требованиям. В этом случае третьи лица могут вмешиваться в работу сети, использовать персональные данные пользователей в корыстных целях и осуществлять иные действия, направленные на получение информации.

Одним из способов выявления сетевых угроз является обнаружение аномалий трафика. Под обнаружением аномалий понимается поиск закономерностей в данных, которые не соответствуют ожидаемому поведению. Установление наличия аномалий в сетевом трафике может свидетельствовать о получении конфиденциальной информации неавторизированным пользователем.

Обнаружение аномалий с помощью энтропии – это способ обнаружения аномалий в сетевом трафике путем анализа неопределенности или случайности данных. Обнаружение аномалий на основе вычисления энтропии зарекомендовало себя как эффективный метод обнаружения аномалий сетевого трафика. Данный подход позволяет обнаружить различные типы аномалий, включая сканирование портов и атак типа «отказ в обслуживании».

В статье предложены алгоритмы определение аномалий сетевого трафика с использованием различных видов энтропии (Шеннона, Цаллиса и Реньи), построения признаков для классификации трафика, а также проведен сравнительный анализ эффективности методов машинного обучения для классификации сетевого трафика.

Аномалии сетевого трафика

Аномалии сетевого трафика представляют собой необычные и существенные изменения в трафике сети. Под аномалиями сетевого трафика подразумеваются изменения объема трафика, канала, структуры распределения IP-адресов источника и (или) номера портов назначения и некоторые другие. Причинами возникновения аномалий сетевого трафика могут быть как законные, так и неправомерные действия. Законные действия включают временные изменения в потребительском спросе, внезапные скопления пользователей, изменения в таблице

маршрутизации и т.п. Неправомерные действия включают DDoS-атаки, сканирование портов, вирусы, черви и т.д.

Для нахождения подходящего решения для обнаружения сетевых аномалий необходимо определить концепцию «нормальности» трафика. Идея нормальности обычно вводится с помощью формальной модели, которая выражает отношения между переменными, участвующими в динамике системы. Событие или объект распознаются как аномальные, если степень их отклонения относительно поведения системы, заданного эталонной моделью, достаточно высока. Примером может являться некоторая система обнаружения аномалий, которая использует контролируемый подход. Данную систему можно представить в виде пары S = (M, D), где M – модель нормального поведения системы, D – мера близости, которая позволяет вычислить, с учетом записи активности, степень отклонения действий по отношению к модели M. Таким образом, каждая система имеет в основном два модуля: модуль моделирования и модуль обнаружения. Одна из систем обучается для получения модели нормальности M. Полученная модель впоследствии используется модулем обнаружения для оценки новых событий, объектов или трафика как аномальных или выбросов. Измерение отклонений позволяет классифицировать события или объекты как аномальные или выбросы.

Энтропии Шеннона, Цаллиса и Реньи

В рамках теории информации энтропия рассматривается как мера неопределенности, связанная со случайной величиной. Чем более случайной является величина, тем выше ее энтропия, и наоборот, чем более предсказуемой является переменная, тем ниже энтропия. Для дискретной случайной величины X с распределением вероятности $p(X = x_i)$, энтропия Шеннона определяется следующим образом [1]:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_a p(x_i) \quad (1)$$

где – это признак, который может принимать значения $\{x_1, \dots, x_n\}$,
$p(x_i)$ – функция плотности вероятности, отражающая вероятность того, что случайная величина X примет значение x_i.

Энтропия может быть интерпретирована как ожидаемое значение $-\log_a p(x)$. Для обнаружения сетевых аномалий часто используется оценка выборочных вероятностей на основе количества наблюдений x_i в заданном временном окне. Величина энтропии зависит от случайности данных и значения n. Для определения уровня случайности рекомендуется использовать нормализованные формы энтропии.

Энтропия Шеннона предполагает компромисс между основной массой распределения и хвостом. Для контроля этого компромисса были предложены два параметризованных обобщения энтропии Шеннона: энтропия Реньи и энтропия Цаллиса. Если параметр α (или q) имеет положительное значение, он выявляет основную массу (концентрацию событий, которые происходят часто), если значение отрицательное – относится к хвосту (дисперсия, вызванная редкими событиями).

Оба параметризованные виды энтропии (Реньи и Цаллиса) происходят из обобщения среднего Колмогорова-Нагумо [4]:

$$\langle X \rangle_\phi = \phi^{-1}(\sum_{i=1}^{n} p(x_i)\phi(x_i)), \quad (2)$$

где ϕ – функция, которая удовлетворяет постулату аддитивности;
ϕ^{-1}- обратная функция.

Формула энтропии Реньи имеет вид:

$$H_{R_\alpha}(X) = \frac{1}{1-\alpha} \log_a (\sum_{i=1}^{n} p(x_i)^\alpha). \quad (3)$$

Энтропию Реньи можно получить из энтропии Шеннона с помощью следующего преобразования [1]:

$$H_{R_\alpha}(X) = \phi^{-1}(\sum_{i=1}^{n} p(x_i)\phi(-\log_2 p(x_i))). \quad (4)$$

Формула энтропии Цаллиса записывается в виде [4]:

$$H_{T_\alpha}(X) = \frac{1}{1-\alpha}(\sum_{i=1}^{n} p(x_i)^\alpha - 1). \quad (5)$$

Обнаружение аномалий сетевого трафика

Для контроля работы сети осуществляется анализ трафика в фиксированных временных интервалах. Для ограничения области поиска аномалий данные фильтруются по адресу доставки, протоколу и т.д. Далее вычисляются значения энтропии Реньи или Цаллиса с учетом положительных и отрицательных значений α для распределения характеристик трафика, представленных в табл. 1.

Таблица 1

Распределение вероятности характеристик сетевого трафика

Характеристика	Вероятностная функция масс
ip-адрес (порт) отправителя (получателя)	$\frac{\text{количество } x_i\ ip-\text{адресов(портов)}\ \text{отправителя(получателя)}}{\text{количество } ip-\text{адресов(портов)}\ \text{отправителя(получателя)}}$
продолжительность передачи данных	$\frac{\text{количество потоков с продолжительностью } x_i}{\text{количество потоков}}$
пакеты, байты	$\frac{\text{количество пакетов(байт) с } x_i\ ip-\text{адресом(портом)}\ \text{отправителя(получателя)}}{\text{количество пакетов(байт)}}$

Процесс обнаружения аномального трафика состоит их двух этапов: обучение и обнаружение. На этапе обучения строится профиль нормального сетевого трафика и осуществляется подготовка модели для классификации. На этапе обнаружения текущие наблюдения сравниваются с моделью.

Первоначально, на этапе обучения, строится динамический профиль с использованием минимальных и максимальных значений энтропии в пределах скользящего временного окна для каждой пары $(\text{характеристика}, \alpha)$.

На этапе обнаружения наблюдаемая энтропия сравнивается с минимальными и максимальными значениями во временном окне, полученными при построении модели, в соответствии со следующим правилом [3].

$$r_\alpha(x_i) = \frac{H_\alpha(x_i) - k \cdot min_\alpha}{k \cdot (max_\alpha - min_\alpha)}, k \in \langle 1..2 \rangle. \quad (6)$$

Значения $r_\alpha(x_i) < 0$ или $r_\alpha(x_i) > 1$ указывают на наличие аномальной сетевой активности. Обнаружение основано на относительном значении энтропии по отношению к расстоянию между минимальным и максимальным значениями.

Коэффициент k в формуле является настраиваемым параметром. Высокое значение k, например $k = 2$, ограничивает количество ложных сигналов (сигналов, в которых не было зафиксировано никакой аномалии), в то время как низкое значение k увеличивает частоту обнаружения (процент правильно обнаруженных аномалий).

Результаты экспериментов

Для проведения экспериментов использовался набор данных CIC-IDS-2017 [4].

Для данных, собранных в понедельник 3 июля 2017 года, характерно отсутствие сетевых аномалий.

Данные, собранные во вторник 4 июля 2017 года, характеризуются наличием 13835 записей аномальных наблюдений и 432074 записей нормальных наблюдений.

Классификация производилось с помощью алгоритмов классификации: деревья решений, случайный лес, логистическая регрессия.

Усредненные по интервалам результаты классификации трафика на аномальный и нормальный для различных видов энтропии представлены в табл. 2. Результаты классификации трафика на аномальный и нормальный с использованием формулы (6) представлены в табл. 3.

Таблица 2

Усредненная эффективность классификации

		Деревья решений	Случайный лес	Логистическая регрессия
Точность	Шеннон	0,81	0,83	0,82
	Реньи	0,78	0,81	0,81
	Цаллис	0,77	0,82	0,82
Доля ложных положительных результатов	Шеннон	0,23	0,11	0,12
	Реньи	0,29	0,13	0,14
	Цаллис	0,30	0,13	0,13

Как следует из табл. 2, наибольшая точность классификации нормального и аномального трафика была достигнута с использованием энтропии Шеннона и случайного леса.

Таблица 3

Усредненная эффективность классификации с использованием формулы (6)

	Шеннон	Реньи	Цаллис
Точность	0,56	0,53	0,53
Доля ложных положительных результатов	0,30	0,29	0,29

Как следует из табл. 3 аномальная сетевая активность может быть обнаружена при использовании всех видов энтропии и формулы (6). По полученным результатам можно сделать вывод об эффективности использования энтропии для классификации сетевого трафика на аномальный и нормальный.

На данный момент используется большое количество различных подходов к обнаружению аномалий сетевого трафика. Энтропия является перспективным и малоизученным показателем, который в дальнейшем может применяться для анализа трафика и других данных.

Библиографические ссылки

1. Харин, Ю. С., Бодягин И. А., Вечерко Е. В. Математические основы теории информации. Минск: БГУ, 2018. С. 13-15.

2. Pietra S. D, Pietra V. D., Lafferty J. M. Introducing features of random fields // IEEE Transactions Pattern Analysis and Machine Intelligence. 1997. Vol. 19, №4. 380 с.

3. Kaggle CIC-IDS-2017 Dataset [Электронный ресурс]. URL: https://www.kaggle.com/datasets/chethuhn/network-intrusion-dataset (дата обращения: 08.02.2024.

4. Berezinski P., Jasiul B., Szpyrka M. An Entropy-Based Network Anomaly Detection Method / Military Communication Institute, AGH University of Science and Technology. 2015. С. 2374, 2375, 2379-2380.