ПРИМЕНЕНИЕ МЕТОДОВ ОЦЕНКИ ЭНТРОПИИ К АНАЛИЗУ ПАКЕТОВ ПЕРЕДАЧИ ДАННЫХ

Капусто Р. А., Палуха В. Ю.

Белорусский государственный университет, Минск, Беларусь,
e-mail: reginakapusto@gmail.com, palukha@bsu.by

Перед аналитиками сетевого трафика ставится задача определить, является ли трафик зашифрованным. Анализ проводится с помощью сетевого оборудования, не способного на сложные математические преобразования, поэтому критерий зашифрованности должен быть основан на некой статистике, которая не требует особенных затрат времени и памяти. В данной статье рассмотрим оценку энтропии как возможный способ решения поставленной задачи, а также сравним различные оценки.

Так как надёжная криптосистема должна обеспечивать неотличимость шифртекста от случайных данных [1], то задача сводится к решению проблемы: удовлетворяет ли наблюдаемый поток данных модели равномерно распределённой случайной последовательности (РРСП).

Пусть имеется случайная последовательность $\{x_i : i = 1,...,n\}$ из распределения вероятностей $\{p_k : k = 1,...,N\}$. Энтропия Шеннона данной последовательности вычисляется по формуле [2]:

$$H(P) = -\sum_{k=1}^{N} p_k \log p_k. \quad (1)$$

Поскольку истинные значения вероятностей неизвестны, можно вычислить оценку энтропии последовательности. В работе рассмотрены несколько методов: подстановочный метод [3], метод Миллера-Мэдоу [4], байесовский метод [4], метод Грассбергера [5] и метод усадки Джеймса-Штейна [4].

Одним из подходов к статистическому оцениванию энтропии является построение частотных оценок вероятностей $\{\hat{p}_k\}$ и подстановка полученных оценок в функционал энтропии вместо истинных значений вероятностей $\{p_k\}$. По подстановочному методу построение частотных оценок для вероятностей производится по следующим формулам:

$$\hat{p}_k = \frac{\nu_k}{n}, \quad \nu_k = \sum_{i=1}^{n} I\{x_i = \omega_k\}, \quad I\{x_i = \omega_k\} = \begin{cases} 1, x_i = \omega_k; \\ 0, x_i \neq \omega_k. \end{cases} \quad (2)$$

Методом Миллера-Мэдоу называют подстановочный метод, скорректированный с помощью константы,

$$\hat{H}_{MM} = \hat{H}_{plug-in} + \frac{\hat{N} - 1}{2n} \log e, \quad (3)$$

где $\hat{H}_{plug-in}$ – оценка, полученная подстановочным методом, $\hat{N}$ – оценка количества исходов с ненулевыми вероятностями.

Метод Байеса основан на корректировке оценок вероятностей и является ещё одной модификацией подстановочного метода. Метод повторяет шаги плаг-ин оценки, только вместо формулы (2) используется следующая оценка:

$$\hat{p}_k^{Bayes} = \frac{\nu_k + a_k}{n + A}, \tag{4}$$

где a_k – поправочные коэффициенты, $A = \sum_{k=1}^{N} a_k, \quad k = 1,...,N$. Таким образом, оценка принимает следующий вид:

$$\hat{H}^{Bayes} = -\sum_{k=1}^{N} \hat{p}_k^{Bayes} \log \hat{p}_k^{Bayes}. \tag{5}$$

Метод Грассбергера также построен на основе подстановочного метода. Формула оценки следующая:

$$\hat{H}_G = \log n - \frac{1}{n}\sum_{k=1}^{N} \nu_k G_{\nu_k}, \quad G_k = \psi(k) + (-1)^k \int_0^1 \frac{x^{k-1}}{x+1} dx, \tag{6}$$

где $\psi(x) = \frac{d \ln \Gamma(x)}{dx}$ – дигамма функция, ν_k вычисляется по формуле (2).

Усадка Джеймса-Штейна основана на усреднении двух абсолютно разных моделей: многомерной модели с низким смещением и высоким отклонением и модели более низкой размерности с большим смещением, но меньшей дисперсией. Интенсивность усреднения определяется относительным весом входящих в состав моделей. Коэффициент λ для выпуклой комбинации

$$\hat{p}_k^{Shrink} = \lambda t_k + (1-\lambda)\,\hat{p}_k^{ML} \tag{7}$$

является коэффициентом интенсивности усадки. Он принимает значения от 0 (нет усадки) до 1 (полная усадка). В формуле 7) $\hat{p}_k^{ML}$ – оценки вероятностей подстановочного метода, вычисленные по формулам (2), t_k – цель усадки.

Формула для вычисления коэффициента интенсивности усадки принимает вид

$$\hat{\lambda} = \frac{1 - \sum_{k=1}^{N} \left(\hat{p}_k^{ML}\right)^2}{(n-1)\sum_{k=1}^{N} \left(t_k - \hat{p}_k^{ML}\right)^2}. \tag{8}$$

Оценка энтропии по методу усадки Джеймса-Штейна производится по формуле

$$\hat{H}^{Shrink} = -\sum_{k=1}^{N} \hat{p}_k^{Shrink} \log \hat{p}_k^{Shrink}. \tag{9}$$

Для подстановочного метода в статье [6] приведено математические ожидание оценки энтропии (используется натуральный логарифм):

$$E\hat{H}_{plug-in} = \ln n - e^{-\lambda} \sum_{k=1}^{+\infty} \frac{\ln(k+1)\lambda^k}{k!}, \quad \lambda = \frac{n}{N}. \tag{10}$$

Для этого метода вычислено математическое ожидание оценки энтропии в случае, если наблюдаемая последовательность является РРСП.

Для остальных методов эталонные значения оценки получены в ходе проведенных исследований следующим образом: сгенерированы по 1000 последовательностей различных длин генератором псевдослучайных чисел языка программирования Python, для каждой из которых вычислена оценка энтропии соответствующим методом; вычислены средние значения каждой из оценок по всем последовательностям. Далее полученные эталонные значения сравнивались с оценками энтропии наблюдаемых последовательностей сетевого трафика. Сетевой трафик записывался с помощью утилиты tcpdump и обрабатывался с помощью библиотеки scapy языка программирования Python.

В таблице 1 приведены результаты исследования пакетов протокола STUN со средней длиной 120 байт, при вычислениях использовался натуральный логарифм.

Табл. 1. Результаты исследования

Название метода	Значение оценки (незашифрованный текст)	Значение оценки (зашифрованный текст)	Математическое ожидание оценки (РРСП)
Подстановочный метод	1.8535014496458992	4.400838841524933	4.4914873843681145
Метод Миллера-Мэдоу	2.0096551506363407	4.779234898245982	4.886314881448898
Байесовский метод	1.8785170119897572	4.438500763516566	4.5263590144603505
Метод Грассбергера	2.184542425517303	5.180435301287537	5.302180066988605
Метод усадки Джеймса-Штейна	2.2668124555381692	5.485662336998431	5.531250633909393

В ходе исследования подтвердилось предположение о том, что незашифрованные данные не являются случайными, а зашифрованные по своим значениям энтропии близки к таковым.

Проведены численные эксперименты, иллюстрирующие применимость рассмотренного подхода для определения случайности данных на примере пакетов стандартных протоколов, а также пакетов с зашифрованным содержанием.

Литература

1. Криптология / Ю. С. Харин [и др.]. – Минск: БГУ, 2013. – 512 с.

2. Shannon, C. E. A mathematical theory of communication / C. E. Shannon // Bell. System Tech. – 1948. – J. 21. – P. 379–423.

3. Башарин, Г. П. О статистической оценке энтропии последовательности независимых случайных величин / Г. П. Бащарин // Теория вероятн. и ее примен. – 1959. – Том 4, выпуск 3. – С. 361–364.

4. Hausser, J. Entropy Interference and the James-Stein Estimator, With Application to Nonlinear Gene Association Networks / J. Hausser, K. Strimmer // Journal of Machine Learning Research. – 2009. – J. 10. – P. 1469–1484.

5. Grassberger, P. Entropy Estimates from Insufficient Sampling [Electronic resource] / P. Grassberger. – Mode of access: https://arxiv.org/pdf/physics/0307138.pdf. – Date of access: 31.03.2021.

6. Палуха, В. Ю. Статистические тесты на основе оценок энтропии для проверки гипотез о равномерном распределении случайной последовательности / В. Ю. Палуха // Весці НАН Беларусі. Серыя фізіка-матэматычных навук. – 2017. – № 1. – С. 79–88.