В.А. Нестеренко
Россия, г. Ростов-на-Дону, Южный федеральный университет

ИСПОЛЬЗОВАНИЕ ВЕСОВЫХ ФУНКЦИЙ ПРИ ОПРЕДЕЛЕНИИ СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК ПОТОКА ПАКЕТОВ В СЕТИ

Системы обнаружения нарушений в сети, основанные на выявлении аномалий, определяют отклонение текущего состояния от базового профиля, характеризующего нормальное состояние системы. Основным достоинством этих методов является возможность выявления новых, неизвестных ранее видов атак. Для построения базового профиля системы используется набор данных, соответствующий состоянию сети свободному от аномалий, или применяются статистические методы. Использование статистических методов для построения базового профиля системы предполагает, что аномальные события составляют небольшую долю в сетевом потоке и не дают заметного вклада при вычислении статистических характеристик для большого числа пакетов. Таким образом, статистические методы обнаружения нарушений в сети основаны на сравнении характеристик потока пакетов вычисленных на относительно небольшом промежутке времени (локальные характеристики), с соответствующими характеристиками за продолжительный промежуток времени (глобальные характеристики) [1]. В качестве статистических характеристик обычно используются выборочные моменты, энтропия, критерий согласия Пирсона χ^2 и т.п. [2-4]. Если локальные характеристики сильно отличаются от соответствующих глобальных характеристик, то это свидетельствует об аномальном поведении потока пакетов и вполне вероятна попытка сканирования сети или сетевой атаки. Таким образом, возникает задача построения эффективных методов вычислений локальных и глобальных статистических характеристик. В данной статье предлагается набор весовых функций для практической реализации вычисления статистических характеристик потока событий в течение некоторого заданного интервала времени.

Будем считать, что величина X_i, $1 \le i \le N$ представляет некоторое событие из потока событий. В качестве статистической характеристики потока событий можно использовать среднее арифметическое функции $f(X)$:

$$w(N) = \frac{1}{N} \cdot \sum_{i=1}^{N} f(X_i).$$

Общее количество событий N определяется интервалом времени, в течение которого ведётся наблюдение за потоком. При нахождении статистических характеристик среднее значение необходимо вычислять не для всего потока из N событий, а только для n последних событий. С этой целью введём весовую функцию $F(z)$ и значения локальных характеристик $W(N)$ будем вычислять по формуле:

$$W(N) = \sum_{i=1}^{N} F\left((N-i)/\tau_n\right) \cdot f(X_i) \qquad (1)$$

Значение аргумента N у величины $W(N)$ означает, что соответствующая характеристика вычисляется вблизи N-го события потока, а размер выборки, для которой находится эта величина, определяется видом весовой функцией и значением параметра τ_n. Использование весовой функции подходящего вида позволяет выделить из всей последовательности событий подпоследовательность заданного размера n. Простым примером такой весовой функции может служить функция:

$$F_\theta(z)=\frac{1}{n}\cdot\theta(1-z/\tau_n),$$

где $\tau_n = n$, а $\theta(z)$ - обычная тета-функция. В этом случае из формулы (1) для арифметического среднего получаем

$$W(N)=\frac{1}{n}\cdot\sum_{i=N-n+1}^{N} f(X_i),$$

Использование весовой функции $F_\theta(z)$ имеет один недостаток: при вычислении среднего значения (1) в потоке необходимо хранить значения X_i для всех n последних событий.

В работе [2] в качестве весовой функции предлагается использовать функцию

$$F_0(z/\tau)=\frac{1}{k}\exp(-z/\tau).$$

Такой выбор весовой функции позволяет при вычислении статистических характеристик $W(N)$ использовать простые рекуррентные соотношения:

$$W(N)=\frac{1}{k}\left(f(X_N)+\exp(-1/\tau)\cdot W(N-1)\right). \qquad (2)$$

В тех случаях, когда интервал усреднения состоит из большого числа (несколько сотен или тысяч) событий, использование таких рекуррентных соотношений даёт значительный выигрыш при вычислении среднего значения $W(N)$.

Отличие весовой функции $F_0(z)$ от функции $F_\theta(z)$ заключается в том, что при вычислении статистических характеристик с использованием функции $F_0(z)$ события при малых значениях z дают относительно больший вклад по сравнению с остальными событиями выбранного интервала. В работах [5-6] предлагается использовать для нахождения статистических характеристик потока событий весовую функцию

$$F_s(z/\tau_s)=\frac{1}{k_s}\cdot\sum_{j=0}^{s}\frac{(z/\tau_s)^j}{j!}\cdot\exp(-z/\tau_s). \qquad (3)$$

Функция $F_0(z)$ является частным случаем $F_s(z)$ при значении параметра $s=0$.

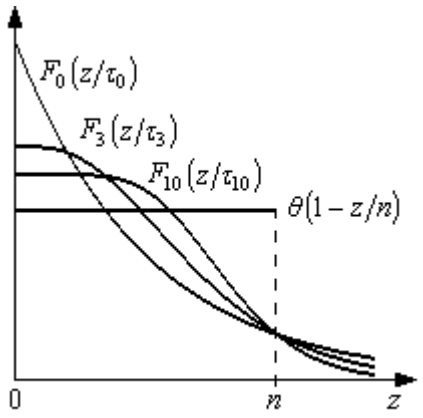

Рис 1

Из графиков приведённых на рис.1 видно, что с увеличением значения s функция $F_s(z)$ становится "более похожей" на тета-функцию, выравнивается относительный вклад разных событий на интервале усреднения. Из результатов работы [5] следует, что в предположении $N >> n$ и $\tau_s >> 1$ нормирующий множитель k_s задаётся выражением

$$k_s = (s+1)\cdot\tau_s$$

а значение параметра τ_s хорошо апроксимируется формулой

$$\tau_s = \frac{n}{1.1\,s+2}.$$

Здесь параметр τ_s определяет интервал усреднения в формуле (1), n – число событий на интервале усреднения.

Выбор вида весовой функции $F_s(z)$ обусловлен тем обстоятельством, что использование выражения (3) позволяет получить рекуррентные соотношения, аналогичные (2) для вычисления $W(N)$. Учитывая тот факт, что величины $X_1,...,X_N$ характеризуют события, происходящие в последовательные моменты времени $t_1,...,t_N$, рекуррентные соотношения, подобные (2), позволяют реализовать вычисления величин $W(N)$, $W(N+1)$, ... в режиме реального времени, по мере поступления новых пакетов и получения соответствующих характеристик X_N, X_{N+1}, ... для потока в сети. Использование рекуррентных соотношений более эффективно, с точки зрения программной реализации вычислений статистических характеристик (1) потока событий, если интервал усреднения содержит большое число событий.

Ключевым моментом при выборе весовой функции $F(z)$ является возможность получения рекуррентных соотношений, аналогичных (2). Для этого следует использовать функции удовлетворяющие соотношениям

$$\varphi_k(z_1+z_2)=\sum_j \varphi_j(z_1)\cdot\varphi_j(z_2).$$

К таким функциям относятся показательные, степенные, тригонометрические и некоторые другие функции. Это обстоятельство позволяет использовать в качестве весовой функции любую функцию, допускающую представление с достаточной степенью точности в виде частичной суммы ряда Фурье:

$$F_m(z)=\frac{1}{2}a_0+\sum_{j=1}^{m}\left(a_j\cos(jz)+b_j\sin(jz)\right).$$

В качестве примера рассмотрим периодическую функцию

$$\mu(z)=\begin{cases}1 & при \quad 2i\cdot T \le z < (2i+1)\cdot T,\\ 0 & при \quad (2i+1)\cdot T \le z < (2i+2)\cdot T,\end{cases} \qquad (4)$$

где $i = 0, 1, 2, \ldots$.

При значениях аргумента $0 \le z < 2T$ функция $\mu(z)$ совпадает с тета-функцией $\theta(1-z/T)$ и может быть использована в качестве весовой функции для нахождения статистических характеристик (1) на интервале усреднения содержащем $n = T$ событий для потока, состоящего из $T < N \le 2T$ событий.

Используя выражение для частичной суммы ряда Фурье функции $\mu(z)$, введём весовую функцию:

$$F_\mu\left(\frac{z}{n}\right)=\frac{1}{n}\left(\frac{1}{2}+\sum_{j=0}^{m}\frac{\sin\left(\pi(2j+1)\frac{z}{n}\right)}{\frac{\pi}{2}(2j+1)}\right). \quad (5)$$

Подставляя функцию $F_\mu(z)$ в формулу для арифметического среднего (1) и выделяя вклад последнего события в величину $W(N)$, можно получить рекуррентные соотношения, аналогичные (2).

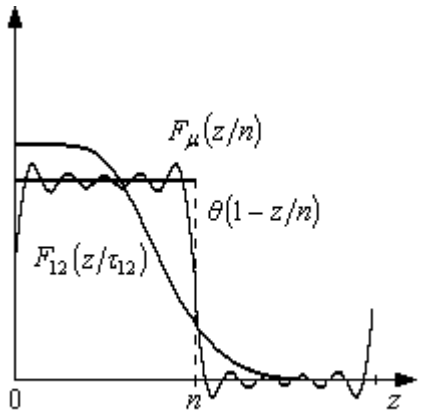

Рис 2

На рис. 2 приведены графики весовых функций $F_\theta(z)$ (тета-функция); $F_s(z)$ (комбинация степенной и показательной функции (2) при значении параметра $s=12$); $F_\mu(z)$ (функция на основе частичной суммы ряда Фурье при значении $m=5$). Значения параметров $s=12$ и $m=5$ выбраны так, чтобы для весовых функций $F_s(z)$ и $F_\mu(z)$ на каждом шаге рекурсии вычислялось одинаковое количество объектов.

Весовая функция $F_\mu(z)$ при значениях аргумента $0\le z<2T$ близка к тета-функции и может быть использована в качестве весовой функции для нахождения статистических характеристик (1) на интервале усреднения, содержащем $n=T$ событий для потока, состоящего из $T<N\le 2T$ событий. При использовании рекуррентных соотношений вычисления начинаются с произвольного события N_0, в течение интервала $(N_0, N_0+T]$ набирается необходимое число событий для интервала усреднения T. Затем при вычислении средних на интервале $(N_0+T, N_0+2T]$ весовая функция $F_\mu(z)$ даёт результат, близкий результату тета-функции $F_\theta(z)$. Для учёта следующих событий потока $N>N_0+2T$ функция $F_\mu(z)$ становится неприемлемой, так как в характеристику $W(N)$ начинает давать вклад импульс второго периода $(2T<z\le 3T)$ весовой функции $F_\mu(z)$. Эту ситуацию можно исправить и применять периодическую (с периодом $2T=2n$) весовую

функцию $F_\mu(z)$ для обработки последовательности событий произвольной длины. Для этого следует использовать два набора рекуррентных соотношений для вычисления арифметического среднего $W(N)$, начала рекурсий для этих наборов должны быть сдвинуты относительно друг друга на величину T . Как только использование одного набора становится некорректным из-за большого числа событий в потоке $N > N_0 + 2T$, то следует переключиться на другой набор рекуррентных соотношений, а для текущего набора рекурсию следует начать заново. Другими словами: следует переключать с одного набора рекуррентных соотношений для случая использования весовой функции $F_\mu(z)$ при вычислении (1) на другой набор через каждые T событий.

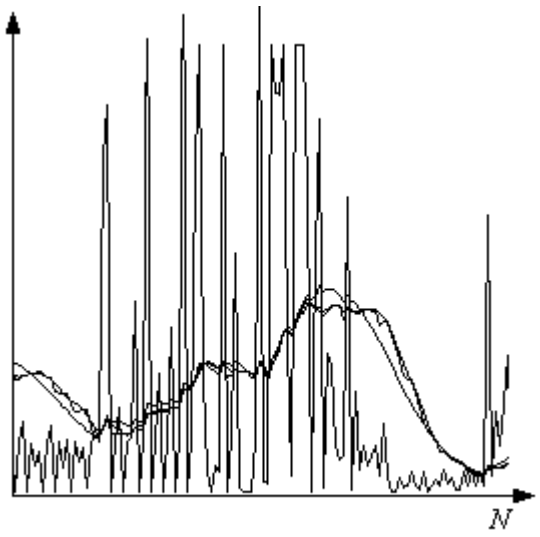

Рис 3

На рис. 3 приведены результаты вычислений усреднённых значений для величин X_i, представляющих некоторое событие для потока пакетов: $W(N) = \sum_{i=1}^{N} F((N-i)/\tau_n) \cdot X_i$. В качестве характеристики X используется временной промежуток между двумя соседними пакетами $\Delta_i = t_i - t_{i-1}$, интервал усреднения принят равным $n = 30$. На горизонтальной оси графика рис. 3 отложены последовательные события в потоке, вдоль вертикальной оси – значения величины $X_i = \Delta_i$. Усреднённые значения вычисляются на интервале $N - n + 1 \le i \le N$ с использованием трёх весовых функций $F_\theta(z)$, $F_s(z)$ и $F_\mu(z)$. Из рис.3 видно, что графики поведения усреднённых характеристик для разных весовых функций практически совпадают.

Заметное различие результатов использования весовых функций $F_\theta(z)$, $F_s(z)$ и $F_\mu(z)$ появляется тогда, когда при вычислении статистических характеристик потока усреднение производится не для последних n событий, а для совокупности событий отстоящих от последнего события на заданную величину n_0 в прошлое:

$$W(N-n_0)=\frac{1}{n}\cdot\sum_{i=N-n_0-n+1}^{N-n_0} f(X_i). \quad (6)$$

Такая ситуация может возникнуть при необходимости сравнения текущих характеристик потока пакетов в сети с соответствующими характеристиками в прошлом: например при сравнении текущего трафика с загруженностью сети несколько часов назад.

В этом случае также можно воспользоваться весовыми функциями, рассмотренными в данной работе. Для этого при вычислении статистических характеристик $W(N-n_0)$ вначале используем формулу (1) для вычисления усреднённых значений на интервале $[N-n_0-n+1, N]$, затем на интервале $[N-n_0+1, N]$ и вычитаем полученные результаты один из другого:

$$W(N-n_0)=\sum_{i=1}^{N}\left(F\left((N-i)/\tau_{n_0+n}\right)-F\left((N-i)/\tau_{n_0}\right)\right)\cdot f(X_i). \quad (7)$$

Использование тета-функции в качестве весовой $F_\theta(z)$ для этого случая приводит к результату, соответствующему выражению (6).

В заключение данной статьи следует сказать, что использование предлагаемых весовых функций (3) (5) и соответствующих рекуррентных соотношений позволяет реализовать эффективные вычисления статистических характеристик потока пакетов в сети (1) (7) в течение заданного интервала времени.

При использовании интервала усреднения, состоящего из небольшого числа событий, лучше использовать тета-функцию $F_\theta(z)$ в качестве весовой функции, так как хранение в памяти n характеристик последних событий и прямое вычисление суммы $\sum_{i=N-n+1}^{N} f(X_i)$ может быть более эффективно с вычислительной точки зрения, чем использование рекуррентных соотношений.

Если число событий на интервале усреднения велико, то использование весовых функций $F_s(z)$ и $F_\mu(z)$ становится более эффективным, так как в этом случае вычисление нескольких коэффициентов в рекуррентных соотношениях будет более эффективным, чем хранение и обработка нескольких сотен или тысяч событий в потоке пакетов.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Roland Kwitt.* A Statistical Anomaly Detection Approach for Detecting Network Attacks. 14th December 2004/ 6QM Workshop, Salzburg.

2. *L.Feinstein and D.Schnackenberg.* Statistical Approaches to DDoS Attack Detection and Response. Proceedings of the DARPA Information Survivability Conference and Expostion (DISCEX'03), April 2003.

3. *Vinay A.Mahadik, Xiaoyong Wu and Douglas S.Reeves.* Detection of Denial-of-QoS Attacks Based On χ^2 Statistic And EWMA Control Charts. http://arqos.csc.ncsu.edu/papers/2002-02-usenixsec-diffservattack.pdf, NC State University, Raleigh.

4. *Nong Ye and Qiang Chen.* An Anomaly Detection Technique Based on a Chi-Square Statistic for Detecting Intrusions into Information Systems. Quality and Reliability Eng. Int'l, Vol 17, No. 2, pp. 105–112, 2001.

5. *Нестеренко В.А.* Определение локальных статистических характеристик потока пакетов в сети // Изв. вузов. Сев.-Кавк. регион. Естеств. науки, 2006, в. S5, С 20-26.

6. *Нестеренко В.А.* Статистические методы обнаружения нарушений безопасности в сети. // Информационные процессы, 2006, Т. 6, в. 3, – С 208–217.