Информационные процессы, Том 6, № 3, 2006, стр. 208–217.

Статистические методы обнаружения нарушений безопасности в сети

В.А.Нестеренко

Факультет математики, механики и компьютерных наук,
Ростовский государственный университет, Ростов-на-Дону, Россия

Поступила в редколлегию 31.05.2006

Аннотация—В предлагаемой статье рассматривается использование статистических методов обнаружения нарушений в сети. Предлагается набор весовых функций для реализации эффективных методов вычисления статистических характеристик и обсуждается выбор значений параметров используемых весовых функций. В качестве статистических характеристик потока пакетов в сети рассматриваются выборочное среднее, выборочная дисперсия и критерий согласия Пирсона χ^2, в статье также рассматриваются признаки наличия аномального поведения потока в сети и связь между этими признаками при использовании различных статистических характеристик.

1. ВВЕДЕНИЕ

Статистические методы обнаружения попыток нарушения безопасности в сети основаны на том обстоятельстве, что в этом случае могут изменяться некоторые статистические характеристики потока пакетов. Так, например, в случае Flood-атаки возрастает трафик, при сканировании сети увеличивается доля пакетов определённого типа (SYN, ACK, FIN, .) и т.п. В этом случае методы обнаружения нарушений основываются на сравнении текущих, локальных характеристик потока пакетов с усреднёнными за продолжительный промежуток времени, глобальными характеристиками [1, 2]. В качестве статистических характеристик обычно используются энтропия, критерий согласия χ^2 и т.п. [2, 3]. Если локальные характеристики сильно отличается 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 статистических характеристик в течение некоторого ограниченного интервала времени и определение аномального отклонения локальных характеристик от глобальных статистических характеристик потока.

В данной статье предлагается набор весовых функций для реализации эффективного метода вычисления локальных статистических характеристик и обсуждается выбор значений параметров используемых весовых функций. В качестве статистических характеристик потока пакетов в сети рассматриваются выборочное среднее, выборочная дисперсия и критерий согласия Пирсона χ^2, в статье также рассматриваются признаки наличия аномального поведения потока в сети и связь между этими признаками при использовании различных статистических характеристик.

2. ВЫЧИСЛЕНИЕ ЛОКАЛЬНЫХ ХАРАКТЕРИСТИК

Будем считать, что числовая величина X_i, $x_{min} \le X_i \le x_{max}$ представляет некоторое событие из потока событий произошедшее в момент времени t_i, $1 \le i \le N$. Весь набор событий

характеризуется средним значением $\bar{x}$ и дисперсией σ_x^2 величины X. В качестве статистической характеристики потока событий будем использовать среднее арифметическое функции $f(X)$ от величины X:

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

Общее количество событий N определяется интервалом времени, в течении которого ведётся наблюдение за потоком. При увеличении числа событий N среднее арифметическое $w(N)$ стремится к $\mathrm{M}[f(X)]$ - математическому ожиданию величины $f(X)$ и может быть использовано в качестве глобальной, долговременной характеристики потока. Для определения локальных характеристик среднее значение будем вычислять не для всего потока 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:

$$W(N) = \sum_{i=1}^{N} F(t_N - t_i) \cdot f(X_i) \tag{1}$$

Значение аргумента локальной характеристики $W(N)$ означает, что её значение вычисляется вблизи N-го события потока, а размер выборки, для которой находится эта величина, определяется видом весовой функции $F(z)$. Использование весовой функции подходящего вида позволяет выделить из всей последовательности подпоследовательность n событий. Простым примером такой весовой функции может служить тета-функция:

$$F(z) = \theta(n\Delta - z)/n$$

где Δ - среднее значение интервала времени между двумя последовательными событиями $\Delta \approx t_i - t_{i-1}$. В этом случае из формулы (1) получаем:

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

В данной статье предлагается использовать для нахождения локальных статистических характеристик потока событий весовую функцию следующего вида:

$$F_s(z) = \frac{1}{k_s}\sum_{j=0}^{s} \frac{(z/\tau)^j}{j!} exp(-z/\tau) \tag{2}$$

Функция $F_s(z)$ локализована вблизи нуля и довольно быстро (экспоненциально) убывает с ростом аргумента z. Параметр τ, присутствующий в определении весовой функции, задаёт временной интервал, на котором эффективно вычисляются локальные характеристики $W(N)$ (1). Коэффициент k_s в формуле (2) введён для обеспечения правильной нормировки функции $F_s(z)$:

$$\sum_{i=N-n+1}^{N} F_s(t_N - t_i) = 1$$

Предлагаемый в данной работе выбор весовой функции обусловлен тем обстоятельством, что формула (2) позволяет получить простые рекуррентные соотношения для вычисления

$W(N)$. Следуя работе [4] введём обозначения

$$W(N) = \sum_{j=0}^{s} A_j(N) \Big/ \sum_{j=0}^{s} K_j(N) \tag{3}$$

где

$$A_j(N) = \frac{1}{\tau^j j!} \sum_{i=1}^{N} f(X_i) \cdot (t_N - t_i)^j \cdot exp(-(t_N - t_i)/\tau)$$

$$K_j(N) = \frac{1}{\tau^j j!} \sum_{i=1}^{N} (t_N - t_i)^j \cdot exp(-(t_N - t_i)/\tau)$$

и выделяя вклад последнего события, получаем рекуррентные соотношения для вычисления величин $A_j(N)$ и $K_j(N)$:

$$A_j(1) = f(X_1) \cdot \delta_{j0}$$

$$A_j(N) = f(X_N) \cdot \delta_{j0} + e^{-\Delta_N/\tau} \sum_{l=0}^{j} \frac{(\Delta_N/\tau)^{j-l}}{(j-l)!} \cdot A_l(N-1) \tag{4}$$

$$K_j(1) = \delta_{j0}$$

$$K_j(N) = \delta_{j0} + e^{-\Delta_N/\tau} \sum_{l=0}^{j} \frac{(\Delta_N/\tau)^{j-l}}{(j-l)!} \cdot K_l(N-1) \tag{5}$$

где $\Delta_N = t_N - t_{N-1}$ - временной интервал между последним и предпоследним событиями в потоке. Учитывая тот факт, что величины $X_1, ..., X_N$ характеризуют события, происходящие в последовательные моменты времени $t_1, ..., t_N$, формулы (3)-(5) позволяют реализовать вычисления локальных характеристик $W(N)$, $W(N+1)$, ... в режиме реального времени, по мере поступления новых пакетов и получения числовых характеристик X_N, X_{N+1}, ... потока сети.

Для оценки величины параметра τ, входящего в определение весовой функции (2) и задающего временной интервал усреднения используем способ, предложенный в работе [4]. Рассмотрим статистику:

$$y(X_N, ..., X_1) = \sum_{l=0}^{N} \lambda_i X_i \text{ где } \lambda_i = F_s(t_N - t_i)$$

и потребуем, чтобы математическое ожидание и дисперсия статистики $y(X_N, ..., X_1)$ совпадали с соответствующими величинами для выборки из n событий. Другими словами с точки зрения значений математического ожидания и дисперсии использование функции (2) эквивалентно использованию $\theta(n\Delta - z)/n$ в качестве весовой функции. При этих условиях получаем соотношения

$$\bar{y} \,/\bar{x} = \sum_{i=1}^{N} \lambda_i = 1$$

$$\sigma_y^2 \,/\sigma_x^2 = \sum_{i=1}^{N} \lambda_i^2 = 1/n$$

которые позволяют найти значение параметра τ. Из результатов работы [4] следует, что в предположении $n << N$ и $\Delta << \tau$ результат вычисления τ хорошо апроксимируется выражением

$$\tau = \frac{n\Delta}{1.1s + 2} \tag{6}$$

Таким образом, использование весовой функции (2) при вычислении усреднённых значений W эквивалентно нахождению среднего значения функции $f(X)$ от n последних элементов в последовательности $X_1, ...X_N$. Величина n и параметр τ при фиксированном значении s связаны соотношением (6).

Выражение (2) представляет семейство функций $F_s(z)$ отличающихся выбором параметра s. Конкретное значение этого параметра определяет поведение функции $F_s(z)$ вблизи нуля:

$$\left(\frac{d^m}{dz^m}F_s(z)\right)\Big|_{z=0} = 0, \text{ при } 0 < m \leq s$$

На Рис.1 показано поведение функции $F_s(z)$ при разных значениях параметра s.

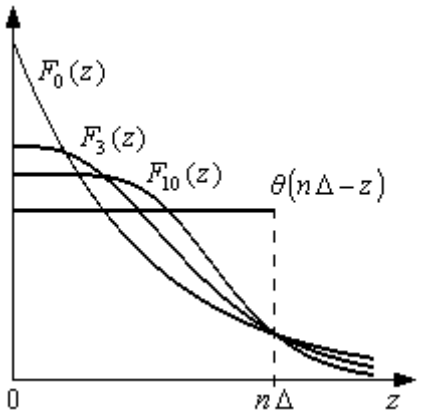

Рис. 1. Поведение весовой функции при разных значениях параметра s

Из приведённых графиков видно, что с увеличением значения s функция $F_s(z)$ становится "более похожей"на тета-функцию: выравнивается относительный вклад разных событий на временном интервале $T \approx n\Delta$ при вычислении локальных статистических характеристик. Используемая в работе [2] весовая функция является частным случаем функции (2) при $s = 0$.

Если оценивать эффективность вычисления статистических характеристик, то при выборке из небольшого числа событий $n \leq 100$ в качестве весовой функции следует использовать тета-функцию $F(z) = \theta(n\Delta - z)/n$. Однако в некоторых случаях (при определении характеристик усреднённых в течение продолжительного интервала времени - час, сутки и т.п.), когда число событий на интервале усреднения превышает несколько сотен, более эффективно использовать весовую функцию (2).

3. СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ ПОТОКА СОБЫТИЙ

Статистические характеристики потока событий, которые используется при определении локальных характеристик, задаются конкретным видом функции $f(X)$ в выражении (1). Если

функция $F(X)$ имеет вид $F(X) = X$, то величина

$$W_1(N) = \sum_{i=1}^{N} F(t_N - t_i) \cdot X_i$$

соответствует среднему арифметическому величины X для n последних событий в потоке; если $f(X) = X^m$, то

$$W_m(N) = \sum_{i=1}^{N} F(t_N - t_i) \cdot X_i^m$$

- начальный выборочный момент порядка m и т.п..

В случае использования критерия согласия χ^2 необходимо ввести классовые интервалы и подсчитывать количество событий, для которых величина X попадает в тот или иной интервал. С этой целью разобьём область возможных значений величины X на B частей: $[x_{min}, x_{max}) \to [x_0, x_1)[x_1, x_2)...[x_{B-1}, x_B)$, где $x_0 = x_{min}$, $x_B = x_{max}$. В дальнейшем полуинтервал $[x_{b-1}, x_b)$, $1 \le b \le B$ будем называть контейнером. Для учёта числа событий попадающих в контейнер с номером b (величина X_i удовлетворяет условию $x_{b-1} \le X_i < x_b$) определим функцию $\Phi_b(X)$:

$$\Phi_b(X) = \begin{cases} 1 & \text{если } X \in [x_{b-1}, x_b) \\ 0 & \text{если } X \notin [x_{b-1}, x_b) \end{cases}$$

и введём набор величин y_b, $1 \le b \le B$:

$$\begin{gathered} y_b(N) = \frac{1}{N} \sum_{i=1}^{N} \Phi_b(X_i) \\ \sum_{b=1}^{B} y_b(N) = 1 \end{gathered} \tag{7}$$

для учёта доли от общего числа событий попадающей в контейнер с номером b. Общее число событий N определяется интервалом времени, в течении которого ведётся наблюдение за потоком. При увеличении числа событий N частоты $y_b(N)$ стремятся к p_b - вероятностям попадания события в контейнер с заданным номером и могут быть использованы в качестве глобальных, долговременных характеристик потока. Для определения локальных характеристик будем учитывать содержимое контейнеров не для всего потока, а только для n последних событий. Значения локальных частот Y_b будем находить в соответствии с формулой (1)

$$\begin{gathered} Y_b(N) = \sum_{i=1}^{N} F(t_N - t_i) \Phi_b(X_i) \\ \sum_{b=1}^{B} Y_b(N) = 1 \end{gathered} \tag{8}$$

и использовать выражение (2) для весовой функции $F_s(z)$, рекуррентные соотношения (3)-(5) для вычисления частот Y_b.

Классовые интервалы $[x_{b-1}, x_b)$, локальные Y_b и глобальные y_b частотные характеристики можно также использовать для нахождениz других статистических характеристик потока пакетов в сети.

Для выявления аномалий потока в сети будем использовать в качестве статистических характеристик выборочное среднее числовой характеристики X:

$$\xi = \sum_{b=1}^{B} \tilde{x}_b \cdot Y_b \tag{9}$$

где $\tilde{x}_b = (x_{b-1} + x_b)/2$ - середина полуинтервала $[x_{b-1}, x_b)$,
выборочную дисперсию:

$$d^2 = \sum_{b=1}^{B} (\tilde{x}_b - \xi)^2 \cdot Y_b \tag{10}$$

и статистику χ^2:

$$\chi^2 = n \cdot \sum_{b=1}^{B} \frac{(Y_b - y_b)^2}{y_b} \tag{11}$$

Величина χ^2 подчиняется хорошо известному χ^2-распределению с $(B-1)$ степенями свободы.

4. КРИТЕРИИ АНОМАЛЬНОГО ПОВЕДЕНИЯ

Признаком появления аномалии в потоке будем считать значительное отклонение локальных статистических характеристик от соответствующих глобальных характеристик. Для каждой из используемых статистических характеристик (9)-(11) можно сформулировать свой критерий присутствия аномалии в потоке событий.

Для выборочного среднего (9) признаком аномалии будем считать превышение заданного порога при отклонении величины ξ от её среднего значения:

$$\left|\xi - \bar{\xi}\right| \geq k\sigma_x \tag{12}$$

где

$$\bar{\xi} = \sum_{b=1}^{B} \tilde{x}_b \cdot y_b$$

- математическое ожидание величины ξ, $y_b \approx p_b$ - вероятность попадания события в контейнер с номером b. Параметр k в формуле (12) задаёт границы интервала $[\bar{\xi} - k\sigma_x, \bar{\xi} + k\sigma_x]$, выход за пределы этого интервала мы воспринимаем как аномалию. Используя то обстоятельство, что распределение суммы независимых случайных величин близко к нормальному распределению, можно записать:

$$k \approx u_\alpha \,/\sqrt{n}$$

где u_α - α-значение нормального отклонения, α - вероятность случайного отклонения величины ξ за пределы (12) и n - количество событий участвующих в формировании локальных статистических характеристик. Таким образом, при выполнении условия (12) можно считать, что с вероятностью $1-\alpha$ это отклонение вызвано появлением аномалии. Так, например, при выбранных значениях числа событий $n = 30$, вероятности $\alpha = 0.001$ и использовании соответствующего табличного значения $u_\alpha = 3.30$ [5], величина параметра k составит $k \approx 0.60$.

Если в качестве характеристики потока событий мы используем статистику χ^2 (11), то признаком появления аномалии будем считать превышение величиной χ^2 установленного порогового значения:

$$\chi^2 \geq \mathrm{X}_0^2 \tag{13}$$

Критерии (12) и (13) аномального поведения потока событий не эквивалентны, факт появления аномалии по одному критерию может соответствовать нормальному поведению потока согласно другому критерию. Это связано с тем обстоятельством, что используемые критерии введены для разных статистических характеристик. В случае критерия (12) мы оцениваем отклонение выборочного среднего от математического ожидания, в случае (13) . отклонение плотности локальной функции распределения от плотности глобальной функции распределения величины X.

Для согласования используемых критериев рассмотрим ситуацию, в которой возникновение аномалии приводит к одновременному выполнению условий (12), (13) и установим связь между параметрами k и X_0^2. Будем считать, что в обычном состоянии при отсутствии аномалий частоты попадания событий в контейнеры имеют вид:

$$Y_b = y_b + \beta_b \text{ при } 1 \leq b \leq B$$

а аномалия заключается в том, что частота попадания в первый контейнер возрастает на регулярную величину δ:

$$\begin{aligned} Y_1 &= y_1 + \beta_1 + \delta \\ Y_b &= y_b + \beta_b - \frac{\delta}{(B-1)} \text{ при } 1 < b \leq B \end{aligned} \tag{14}$$

В реальности подобная ситуация может возникнуть при сканировании сети или flood-атаке, когда на фоне обычного сетевого трафика появляется множество пакетов с близкими характеристиками. В этом случае из условий (12) и (13), представления (14) для Y_b и свойств вариаций β_b

$$\begin{aligned} \mathrm{M}\left[\beta_b\right] &= 0 \\ \mathrm{M}\left[\sum_{b=1}^{B} \beta_b^2\right] &= \frac{B-1}{n} \end{aligned}$$

где $\mathrm{M}[z]$ - математическое ожидание величины z
находим связь между параметрами двух разных критериев наличия аномалий в сети:

$$\mathrm{X}_0^2 = (B-1)\left(1 + n\frac{k^2\sigma_x^2}{(\bar{x} - \tilde{x}_1)^2}\right) \tag{15}$$

При получении соотношения (15) было использовано допущение о равенстве вероятностей $p_b \approx 1/B$, $1 \leq b \leq B$, это легко реализовать посредством подбора границ полуинтервалов $[x_{b-1}, x_b)$ таким образом, чтобы вероятности попадания событий в разные контейнеры были равны.

Рассуждая аналогичным образом для случая использования d^2 (10) в качестве статистической характеристики потока событий находим значение для граничной величины S_0^2:

$$\mathrm{S}_0^2 = \sigma_x^2\left(1 + k\frac{\bar{x} - \tilde{x}_1}{\sigma_x}\right)\left(1 - k\frac{\sigma_x}{\bar{x} - \tilde{x}_1}\right) \tag{16}$$

Величина S_0^2 задаёт нижнюю границу, выход за пределы которой свидетельствует о наличии аномалии:

$$d^2 \le S_0^2 \tag{17}$$

Для того чтобы значение S_0^2 (16) соответствовало критерию присутствия аномалии (17) необходимо выполнение условия:

$$k > \frac{\bar{x} - \tilde{x}_1}{\sigma_x} - \frac{\sigma_x}{\bar{x} - \tilde{x}_1}$$

В случае использования в качестве числовой характеристики X временного интервала между двумя соседними пакетами $X_i = t_i - t_{i-1}$ и если плотность функция распределения реального потока пакетов в сети близка к плотности функции распределения для стационарного пуассоновского потока, то последнее условие обычно выполняется, так как для стационарного пуассоновского потока значения математического ожидания и среднеквадратичного отклонения равны.

5. ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ

Графики, приведённые на Рис2, иллюстрируют возможность практического использования различных критериев присутствия аномалий в потоке пакетов сети.

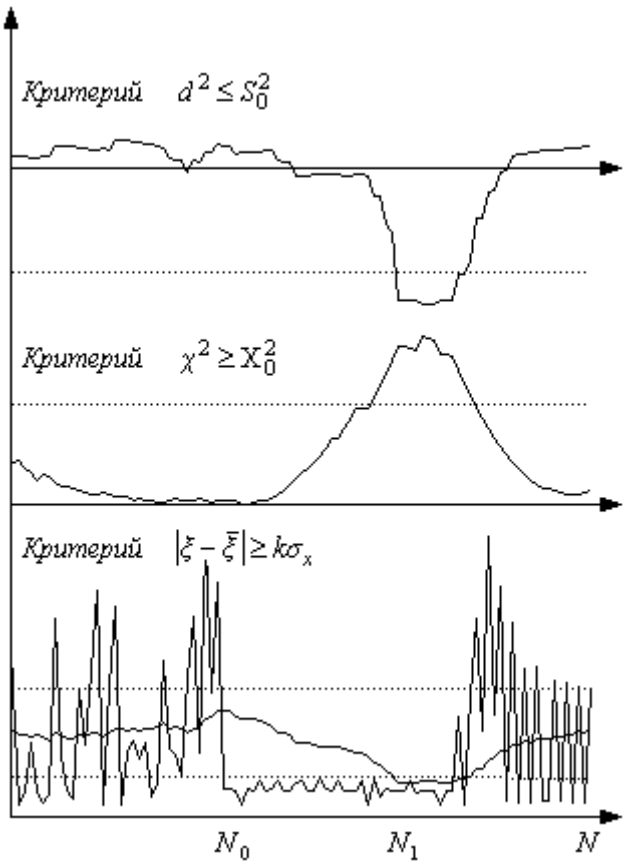

Рис. 2. Графики поведения локальных статистических характеристик

При построении графиков использованы реальные данные, полученные на одном из узлов

сети. В качестве числовой характеристики X используется временной интервал между двумя соседними пакетами: $X_i = t_i - t_{i-1}$. В приведённом примере средний интервал времени между пакетами и среднеквадратичное отклонение составляют $\bar{x} \approx \sigma_x \approx 85$ мсек. При обработке потока пакетов и нахождении статистических характеристик (9), (10) и (11) были выбраны значения $B = 5$ и $n = 30$.

Вдоль горизонтальных осей графиков отложены номера событий в сети (событие - поступление нового пакета), вертикальная ось соответствует промежутку времени между приходом двух пакетов для нижнего графика, значениям статистики χ^2 и $(d^2 - \sigma_x^2)$ для двух других графиков. На верхнем графике показано поведение выборочной дисперсии d^2 (10) минус σ_x^2, на среднем - поведение статистики χ^2 (11), на нижнем графике приведены значения числовых характеристик X_i поступающих пакетов и поведение выборочного среднего ξ (9). Пунктирные линии на графиках обозначают границы (12), (13) и (17), выход за указанные границы свидетельствует о наличии аномалии в сети. Начиная с пакета номер N_0, трафик резко возрастает, средняя частота поступления пакетов увеличивается в 5-6 раз. Из приведённых графиков видно, что в этом случае значения статистических характеристик d^2, χ^2 и ξ выходят за границы "коридора"допустимых значений и используемые критерии указывают на появление аномалии.

6. ЗАКЛЮЧЕНИЕ

Результаты, полученные в работе, могут быть использованы при создании системы обнаружения нарушений безопасности в сети. Использование контейнеров, глобальных y_b (7) и локальных Y_b (8) частотных характеристик позволяет получать различные статистические характеристики (9), (10), (11) потока пакетов в сети. Вычисление частотных характеристик Y_b (8) может быть реализовано с использованием весовых функций $F_s(z)$ (2) и рекуррентных соотношений (3)-(5). Использование рекуррентных соотношений будет особенно полезным при вычислении статистических характеристик на большом числе событий $n \approx T/\Delta$, так как в этом случае вычисление нескольких коэффициентов $A_j(N)$ и $K_j(N)$ в выражении (3) будет более эффективным чем обработка нескольких сотен или тысяч событий в потоке пакетов.

Предложенные критерии наличия аномалий в потоке пакетов (12), (13), (17) сформированы таким образом, что разные статистические характеристики одновременно указывают на присутствие аномалий определённого вида (14). Это позволяет связать параметры пороговых значений X_0^2, S_0^2 и k (15), (16) и "выровнять"чувствительность различных статистических методов обнаружения аномалий. Для аномалий других видов, отличных от вида (14), применение критериев соответствующих различным статистическим характеристикам будет приводить к разным результатам: появления аномалии по одному критерию может соответствовать нормальному поведению потока согласно другому критерию. Вопрос о классификации аномалий и применении тех или иных статистик для обнаружения аномалий конкретного вида требует дальнейшего изучения.

СПИСОК ЛИТЕРАТУРЫ

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

2. L.Feinstein, 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.

4. Нестеренко В.А. Известия Вузов. Северо-Кавказский регион. Естественные науки. 2006. Приложение. . 5.

5. Абрамовиц М. и Стиган И. Справочник по специальным функциям. Москва, Наука, 1979, стр.754, Таблица 26.1.

Статью представил к публикации член редколлегии В.И. Венец