Д. О. Ковалев

ОЦЕНКА КОЛИЧЕСТВА СООБЩЕНИЙ ИБ В АВТОМАТИЗИРОВАННЫХ СИСТЕМАХ КАК МЕТОД ВЫЯВЛЕНИЯ СЕТЕВЫХ АТАК

Постановка задачи

АС крупной организации часто насчитывает десятки средств защиты информации (СЗИ), каждое из которых может регистрировать сотни тысяч сообщений ИБ в день [1]. Большинство этих сообщений ИБ являются результатом нормальной сетевой активности, и лишь немногие свидетельствуют о наличии реальных атак на АС. Поэтому в настоящее время мониторинг ИБ АС и управление СЗИ в динамически меняющейся сетевой среде — очень важная и сложная задача. Для решения этой задачи используются системы мониторинга ИБ, которые осуществляют сбор сообщений ИБ, приведение их к единому виду, агрегацию и корреляцию для выявления аномальной сетевой активности [2].

Основным результатом работы любой системы мониторинга ИБ является набор сообщений о выявленных событиях ИБ. В случае, если против АС направлена сетевая атака, то количество сообщений ИБ резко увеличивается. Это справедливо как для известных сетевых атак, так и для атак нулевого дня, поскольку СЗИ будут регистрировать большое количество сообщений ИБ, связанных с подозрительной сетевой активностью. Сложность заключается в том, чтобы оценить допустимое количество сообщений ИБ, прежде чем сделать заключение о том, что атака действительно имеет место.

На рисунке 1 показан график количества сообщений ИБ при отсутствии атак на АС в рамках 10-минутных временных интервалов (значения фиксировались на протяжении одной недели). Из графика видно, что в разное время суток количество сообщений ИБ может значительно различаться. Также видно, что количество сообщений ИБ циклически повторяется изо дня в день с определенной погрешностью.

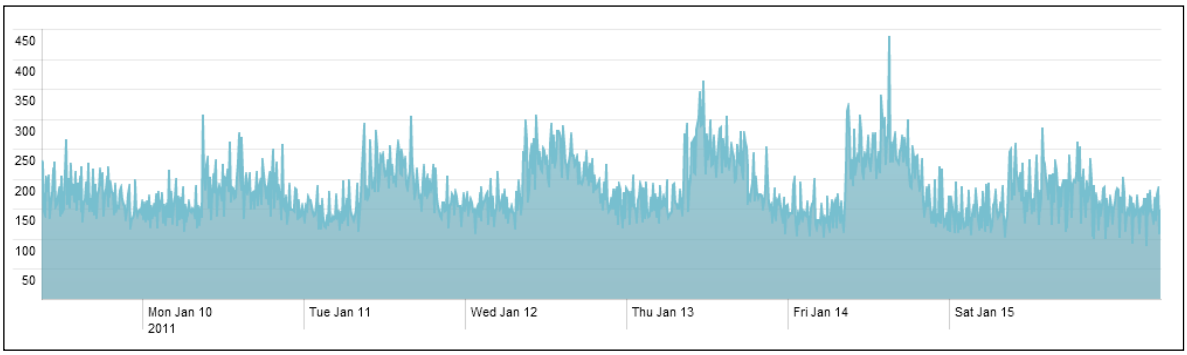

Рис. 1. Количество сообщений ИБ в рамках 10-минутных интервалов (Скриншот)

Таким образом, очевидно, что не существует единого порогового значения, которое можно задать заранее для идентификации атаки, и необходимо производить оценку количества сообщений ИБ.

Описание решения

Решение данной задачи требует создания адаптивных пороговых значений, которые могли бы динамически обучаться шаблонам сообщений ИБ и постоянно находиться в актуальном состоянии по мере обновления данных.

С точки зрения математической статистики знание распределения количества сообщений ИБ для различного времени суток при отсутствии атак в АС позволит достаточно точно определить пороговые значения в нужный момент времени. В таблице 1 приведена статистика количества сообщений ИБ для одного из 10-минутных интервалов, собранная на испытательном стенде в течение 8 недель.

Таблица 1. Количество сообщений ИБ в рамках 10-минутного интервала (шт.)

Неделя/День	1	2	3	4	5	6	7
1	173	154	173	175	162	178	163
2	164	180	172	180	184	180	173
3	164	173	180	174	184	187	184
4	173	184	176	184	192	174	180
5	210	175	175	195	176	185	178
6	185	157	184	185	140	197	185
7	174	180	163	178	184	179	181
8	173	180	175	174	165	173	161
Мат. ожидание(EX):		**176.5536**		**Дисперсия(DX):**		**121.0153**	

Для удобства работы со статистическими данными на основании таблицы 1 рассчитана нормированная статистика $\frac{X-EX}{\sqrt{DX}}$, приведенная в таблице 2.

Таблица 2. Нормированная статистика для количества сообщений ИБ (безразмерная величина)

Неделя/День	1	2	3	4	5	6	7
1	-0.3230	-2.0502	-0.3230	-0.1412	-1.3230	0.1315	-1.2321
2	-1.1412	0.3133	-0.4139	0.3133	0.6769	0.3133	-0.3230
3	-1.1412	-0.3230	0.3133	-0.2321	0.6769	0.9496	0.6769
4	-0.3230	0.6769	-0.0503	0.6769	1.4041	-0.2321	0.3133
5	3.0404	-0.1412	-0.1412	1.6768	-0.0503	0.7678	0.1315
6	0.7678	-1.7775	0.6769	0.7678	-3.3228	1.8586	0.7678
7	-0.2321	0.3133	-1.2321	0.1315	0.6769	0.2224	0.4042
8	-0.3230	0.3133	-0.1412	-0.2321	-1.0503	-0.3230	-1.4139
Мат. ожидание(EX):		**0.0000**		**Дисперсия(DX):**		**1.0000**	

Для проверки гипотезы о распределении количества сообщений ИБ в работе использовался непараметрический критерий Колмогорова. Выбор в пользу данного критерия обусловлен тем, что этот критерий достаточно прост, его можно применять для малых выборок и считается, что его мощность выше, чем у критерия x^2 [3]. Гистограмма распределения для нормированной статистики количества сообщений ИБ представлена на рисунке 3. Для построения гистограммы все множество значений выборки было разбито на $\sqrt{n} = \sqrt{56} \approx 8$ интервалов. Вид гистограммы позволяет выдвинуть гипотезу о том, что количество сообщений ИБ распределено по нормальному закону распределения.

Для приведенной статистики мера расхождения между теоретическими значениями нормального распределения и эмпирическими значениями выборки $\lambda = \sup_x |F_n(x)-F(x)|\sqrt{n} = 1.2848$, что меньше, чем $\lambda_\alpha \approx 1.63$, для уровня значимости $\lambda = 0.01$. Таким образом, гипотеза о распределении количества сообщений ИБ принимается при данном уровне значимости. Аналогичные результаты были получены для всех 10-минутных интервалов при уровне значимости $\lambda = 0.01$.

Рис. 2. Гистограмма распределения количества сообщений ИБ

Математическое ожидание и дисперсия распределения на соседних временных интервалах должны не очень сильно различаться, поэтому их оценки могут быть получены в результате интерполяции по значениям, хранящимся в специально для этого созданной таблице моментов (имеются в виду 1-й и 2-й моменты случайной величины). В ходе экспериментов было установлено, что метод квадратичной интерполяции позволяет получить наиболее точные значения оценок математического ожидания и дисперсии. Таблица моментов используется для отслеживания цикличности сообщений ИБ. В рамках данной работы таблица моментов содержит 24 значения для математического ожидания и 24 значения дисперсии, соответствующие каждому часу в сутках. В типовой организации размеры таблицы моментов будут выбираться исходя из цикличности сообщений ИБ. Значения, хранящиеся в таблице моментов, динамически обновляются посредством экспоненциального скользящего среднего после получения очередного количества сообщений ИБ. Метод экспоненциального скользящего среднего представляет собой взвешенное скользящее среднее, у которого веса уменьшаются экспоненциально с удаленностью рассчитываемой величины от текущего значения наблюдения [4]. Комбинация методов интерполяции значений таблицы моментов и расчета значений экспоненциальным скользящим средним позволяет оценивать как цикличность сообщений ИБ, так и долговременные тенденции.

Метод оценки количества сообщений ИБ реализуется последовательностью из четырех основных шагов, которые применяются каждый раз при получении очередного количества сообщений ИБ x_t:

1) интерполировать значения в таблице моментов и получить оценку параметров для нормального распределения F_t для t-го интервала времени;

2) обновить значения таблицы моментов в соответствии с полученным значением x_t;

3) оценить x_t путем расчета его стандартной оценки Z_t на основании параметров распределения F_t;

4) рассчитать значение показателя I_x на основании стандартной оценки Z_t.

Пусть x_t — это количество сообщений, полученное на t-м интервале времени, который соответствует циклу с, часу h ($1 \le h \le H$) и минуте m ($1 \le m \le M$), где $H = 24$ — количество часов в дне, $M = 60$ — количество минут в часе.

На первом шаге происходит получение оценок математического ожидания $\widehat{\mu}_{h,m}$ и дисперсии $\widehat{\sigma}^2_{h,m}$ нормального распределения на t-м интервале времени получаются в результате квадратичной интерполяции значений математического ожидания и дисперсии, хранящихся в таблице моментов $\{(E_h, D_h): h=1,\ldots,H\}$.

Пусть арифметическое среднее $M = 60$ математических ожиданий, полученных в результате интерполяции в рамках одного часа, равно соответствующему значению E_h из таблицы моментов. Аналогично арифметическое среднее $M = 60$ дисперсий, полученных в результате интерполяции в рамках одного часа, равно соответствующему значению D_h из таблицы моментов.

Тогда, если взять три последовательных часа (-1,0], (0,1], (1,2], то можно определить коэффициенты квадратической интерполяции (A,B,C):

$$\int_{-1}^{0}(At^2+Bt+C)dt=\frac{A}{3}-\frac{B}{2}+C=E_{-1}M$$

$$\int_{0}^{1}(At^2+Bt+C)dt=\frac{A}{3}+\frac{B}{2}+C=E_{0}M \quad .$$

$$\int_{0}^{2}(At^2+Bt+C)dt=\frac{7A}{3}+\frac{3B}{2}+C=E_{1}M$$

Решение данной системы уравнений относительно A,B,C дает:

$A = M(M_{-1} - 2E_0+E_1)/2,$

$B = M(E_0 - E_{-1}),$

$C = M(2E_{-1} + 5E_0 - E_1)/6.$

Пусть $m_1 = \frac{m-1}{M}$, $m_2 = \frac{m}{M}$ — две последовательные минуты, тогда значение оценки математического ожидания, полученное в результате процедуры интерполяции, соответствующей минуте m часа h, равно:

$$\widehat{\mu_{h,m}} = \int_{m_1}^{m2}(At^2+Bt+C)dt = \frac{A}{3M}(m_1^2+m_1m_2+m_2^2)+\frac{B}{2M}(m_1+m_2)+\frac{C}{M} .$$

Таким образом, выведено выражение для расчета оценки $\widehat{\mu_{h,m}}$ из сохраненных значений математических ожиданий в таблице моментов.

Аналогичным образом для оценки дисперсии:

$A' = M(D_{-1} - 2D_0 + D_1)/2,$

$B' = M(D_0 - D_{-1}),$

$C' = M(2D_{-1} + 5D_0 - D_1)/6,$

$$\widehat{\sigma^2_{h,m}} = \int_{m_1}^{m2}(A't^2+B't+C')dt = \frac{A'}{3M}(m_1^2+m_1m_2+m_2^2)+\frac{B'}{2M}(m_1+m_2)+\frac{C'}{M} .$$

Интерполяция коэффициентов (A, B, C) и (A′, B′, C′), использующихся для расчета оценок математического ожидания и дисперсии, производится раз в час. Интерполяция сглаживает значения как внутри часа, так и между часами, поскольку коэффициенты зависят от хранящихся в таблице моментов оценок для данного часа, а также для двух смежных с ним часов.

На втором шаге происходит обновление значений таблицы моментов. При этом сложная ситуация возникает в случае экстремальных значений количества сообщений ИБ. Большой разброс значений может привести к резкому увеличению математического ожидания и дисперсии, что, в свою очередь, означает длинные хвосты распределения и неоднозначность выявления атаки в будущем. Если просто перестать учитывать экстремальные значения в расчетах, то это приведет к недооценке математического ожидания и дисперсии, т. е. получится распределение с очень короткими хвостами и большим количеством ложных срабатываний. Таким образом, ситуации экстремальных значений количества сообщений ИБ необходимо обрабатывать особо. Возможны следующие варианты:

1) очень большое экстремальное значение $x_t > x_{0,9999}$ — в этом случае в расчетах вместо x_t используется произвольное значение x'_t из интервала: $x_{0,9999} < x'_t < x_{0,9999}$, где $x_{0,99}$, $x_{0,9999}$ — квантили распределения F_t уровней 0.99 и 0.9999 соответственно;

2) очень маленькое экстремальное значение $x_t < x_{0,0001}$ — в этом случае в расчетах вместо x_t используется произвольное значение x''_t из интервала: $x_{0,0001} < x''_t < x_{0,01}$, где $x_{0,01}$, $x_{0,0001}$ — квантили распределения F_t уровней 0.01 и 0.0001 соответственно;

3) отсутствие количества сообщений ИБ на интервале t — в этом случае в расчетах используется произвольное значение x'''_t из интервала: $x_{0,01} < x'''_t < x_{0,99}$, где $x_{0,01}$, $x_{0,99}$ — квантили распределения F_t уровней 0.01 и 0.99 соответственно.

Математическое ожидание и дисперсия сохраняются в таблицу моментов по истечении каждого часа. Обновленное значение математического ожидания для нормального распределения на t-м интервале времени, заканчивающемся на минуте m и часе h цикла c, рассчитывается как экспоненциальное скользящее среднее от двух величин: 1) $\widehat{\mu_{h,m}}$, полученной в результате интерполяции, 2) x_t, которое является количеством наблюдаемых сообщений ИБ (с учетом правил для экстремальных или отсутствующих значений, указанных выше):

$$\widehat{\mu_{h,m}}' = (1 - w_c)\,\widehat{\mu_{h,m}} + w_c x_t,$$

$$w_c = w + \frac{1 - w}{1 + c}.$$

где w — некоторый фиксированный вес, принимающий значения между 0 и 1; w_c — временный вес для текущего цикла, который снижается до постоянного веса w по мере прохождения циклов и позволяет быстрее определить значение математического ожидания при инициализации системы.

Фактически после прохождения большого количества циклов выражение принимает вид:

$$\widehat{\mu_{h,m}}' = (1 - w)\,\widehat{\mu_{h,m}} + w x_t.$$

Выражение для дисперсии выводится следующим образом:

так как $D_n = \frac{1}{n}\sum_1^n (X_i - \overline{X_n})^2$, где $\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$,

то

$$D_{n+1} = \frac{1}{n+1}\,[nV_n + (X_{n+1} - \overline{X_n})^2 + 2(\overline{X_n} - \overline{X_{n+1}})(X_{n+1} - \overline{X_n}) + (n+1)(\overline{X_n} - \overline{X_{n+1}})^2].$$

Поскольку $\overline{X_n} - \overline{X_{n+1}} = \frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n+1}\sum_{i=1}^{n+1} X_i = \frac{\overline{X_{n+1}} - X_{n+1}}{n}$,

то

$$D_{n+1} = \frac{n}{n+1}\,D_n + \frac{1}{n+1}(X_{n+1} - \overline{X_n})(X_{n+1} - \overline{X_{n+1}})^2].$$

Приняв $w_c = \frac{1}{n+1}$, получаем выражение для обновленного значения оценки дисперсии для t-го интервала времени:

$$\widehat{\sigma^2_{h,m}}' = (1 - w_c)\,\widehat{\sigma^2_{h,m}} + w_c(x_t - \widehat{\mu_{h,m}})(x_t - \widehat{\mu_{h,m}}').$$

Значения математического ожидания и дисперсии в таблице моментов не меняются в течение всего цикла. Для того чтобы в конце цикла обновить таблицу моментов, по мере поступления наблюдений рассчитываются временные значения математического ожидания E'_h и дисперсии D'_h. E'_h и D'_h инициализируются нулевыми значениями в начале часа, по мере поступления наблюдений x_t обновляются по следующим формулам:

$$E'_h = E'_{h,M};\ E'_{h,m} = \frac{(m-1)\,E'_{h,m-1} + \widehat{\mu_{h,m}}'}{m};\ E'_{h,0} = 0;$$

$$D'_h = D'_{h,M};\ D'_{h,m} = \frac{(m-1)\,D'_{h,m-1} + (\widehat{\sigma^2_{h,m}})'}{m};\ D'_{h,0} = 0.$$

В конце цикла полученные E'_h и D'_h замещают сохраненные в таблице моментов E_h и D_h соответственно и используются в следующем цикле.

На третьем шаге производится оценка x_t. Атаки в АС характеризуются повышенным количеством сообщений ИБ. Сниженное количество сообщений ИБ является индикатором того, что одно или несколько СЗИ вышло из строя. Таким образом, необходимо отслеживать оба типа ситуаций.

Поскольку нормальные распределения на различных временных интервалах будут иметь различные параметры, для того чтобы сравнивать наблюдения x_t между собой, их предварительно необходимо нормировать. Для этого рассчитывается нормированная статистика $Z_t = \frac{x_t - \widehat{\mu}_{h,m}}{\widehat{\sigma^2}_{h,m}}$.

На четвертом шаге производится расчет значения I_x по методу экспоненциального скользящего среднего от Z_t по следующей формуле:

$I_x = (1\text{-}w)I_{x-1} + wZ_t$,

где $Z_t = \frac{x_t - \widehat{\mu}_{h,m}}{\widehat{\sigma^2}_{h,m}}$ для веса w в интервале (0,1], $I_0 = 0$.

Показатель I_x позволяет учитывать как магнитуду, так и длительность изменений. Например, резкий скачок количества сообщений (высокая магнитуда, но короткая длительность) может вывести I_x за пороговые значения точно так же, как последовательность менее выраженных, но необычных скачков сообщений ИБ (малая магнитуда, но большая длительность) может свидетельствовать об атаке. Для того чтобы учитывать длительность изменений, используется вес w, принимающий значения в интервале (0,1]. Эмпирическим путем было показано, что значение $w = 0.25$ дает наиболее точные результаты.

Метод оценки количества сообщений ИБ реализуется алгоритмом для расчета показателя I_x, представляющего собой следующую последовательность шагов, которые предпринимаются каждый t-й интервал времени:

1) провести индексирование — определить минуту m, час h, день d для временной метки t;

2) рассчитать оценки математического ожидания $\widehat{\mu}_{h,m}$ и дисперсии $\widehat{\sigma^2}_{h,m}$ нормального распределения F_t для t-го интервала времени путем применения квадратичной интерполяции для часов h – 1, h и h + 1, используя сохраненные в таблице моментов данные;

3) проверить, что x_t не равно нулю. Если $x_t = 0$, считать $I_x = I_{x-1}$;

4) если x_t не равно нулю, рассчитать его нормированную оценку Z_t для распределения F_t и рассчитать показатель $I_x = (1\text{-}w)I_{x-1} + wZ_t$;

5) в случае получения очень больших или очень маленьких значений или отсутствия данных следовать следующим правилам:

- если x_t принимает очень большие значения, в расчетах вместо x_t используется произвольное значение x'_t из интервала: $x_{0,99} < x'_t < x_{0,9999}$;
- если x_t принимает очень маленькие значения, в расчетах вместо x_t используется произвольное значение x''_t из интервала: $x_{0,0001} < x''_t < x_{0,01}$;
- если x_t отсутствует, в расчетах вместо x_t используется произвольное значение x'''_t из интервала: $x_{0,01} < x'''_t < x_{0,99}$;

6) рассчитать обновленные оценки математического ожидания $\widehat{\mu}_{h,m}'$ и дисперсии $\widehat{\sigma^2}_{h,m}'$ для нормального распределения с учетом полученного значения x_t;

7) обновить временные значения моментов $E'_{h,m}$ и $D'_{h,m}$ для каждой минуты. В конце каждого цикла хранящиеся значения таблицы моментов E_h и D_h следует заменить обновленными значениями E'_h и D'_h и в дальнейшем использовать их в расчетах коэффициентов для квадратичной интерполяции.

Поскольку в начале самого первого цикла данные отсутствуют, в рамках первого цикла не рассчитываются значения I_x, а только происходит сбор данных для заполнения таблицы моментов.

Выводы

Для реализации метода оценки количества сообщений ИБ необходимо рассчитать оценки математического ожидания $\widehat{\mu}_t$ и дисперсии $\widehat{\sigma^2}_t$, которые получаются в результате динамического обучения в процессе работы системы мониторинга ИБ АС. Поскольку нормальные распределения на различных временных интервалах будут иметь различные параметры, для того чтобы сравнивать наблюдения x_t между собой, их предварительно необходимо нормировать. Для этого расчитывается

нормированная статистика $Z_t = \frac{x_t - \widehat{\mu_t}}{\widehat{\sigma^2_t}}$. В качестве пороговых значений для количества сообщений ИБ следует использовать квантили стандартного нормального распределения, поскольку это позволяет перейти к общепринятому подходу доверительного оценивания на основании уровней значимости.

Для оценки количества сообщений ИБ используется показатель I_x, который рассчитывается по методу экспоненциального скользящего среднего от Z_x по следующей формуле: $I_x = (1\text{-}w)I_{x-1} + wZ_t$. Показатель I_x позволяет учитывать как магнитуду, так и длительность изменений. Показатель I_x в совокупности с прочей информацией, хранящейся в базе данных системы мониторинга ИБ, может быть использован для создания упреждающей защиты путем расчета некоторой консолидированной оценки результатов мониторинга ИБ и адаптивного выбора на основании этой оценки необходимых настроек СЗИ.

СПИСОК ЛИТЕРАТУРЫ:

1. *Ковалев Д. О.* Управление информационной безопасностью // Аналитический банковский журнал. 2009. № 10 (173).
2. Security Information Management. Веб-сайт / Netforensics, 2003. URL: http://www.netforensics.com.
3. Критерий Колмогорова. FTP-сервер / Кафедра вычислительной техники факультета автоматики и вычислительной техники Томского политехнического университета. URL: ftp://ftp.ce.cctpu.edu.ru.
4. Экспоненциальное скользящее среднее. Веб-сайт / Форекс Арена, 2008. URL: http://www.forexarena.ru.