УДК 004.056.53

Р.В. Грищук[1], В.М. Мамарєв[2]

[1] *Житомирський військовий інститут ім. С.П. Корольова НАУ, Житомир*
[2] *Київський оперативний центр, Київ*

МЕТОД ОЦІНЮВАННЯ ІНФОРМАТИВНОСТІ ПАРАМЕТРІВ ПОТОКУ ВХІДНИХ ДАНИХ ДЛЯ МЕРЕЖЕВИХ СИСТЕМ ВИЯВЛЕННЯ АТАК

В статті проведено аналіз підходів до побудови інформативної системи ознак. В результаті проведення експертного оцінювання яких було обрано найбільш доцільний, для побудови інформативної системи ознак, інформаційний підхід і на його основі розроблено метод оцінювання інформативності параметрів потоку вхідних даних для мережевих систем виявлення атак. Запропонований метод дозволяє одержувати кількісні оцінки інформативності параметрів потоку вхідних даних для мережевих систем виявлення атак, що в подальшому можуть бути використані на етапі формуванні шаблонів поведінки інформаційно-телекомунікаційних систем.

Ключові слова: *система захисту інформації, система виявлення атак, потік вхідних даних.*

Вступ

Постановка проблеми. Глобальність розповсюдження інформаційно-телекомунікаційних систем (ІТС), вільний доступ користувачів до ресурсів й латентність кіберзлочинів призводять до різкого зростання кількості порушень в галузі безпеки інформації.

Незважаючи на впровадження сучасних систем захисту інформації, а саме систем виявлення атак (СВА), статистичні дані свідчать про низький рівень захищеності інформації в ІТС. Так, відносно низька ефективність функціонування сучасних СВА зумовлена існуючим на сьогодні протиріччям, що проявляється в дисбалансі між стрімким зростанням інтенсивності комутації потоків вхідних даних в ІТС та високою обчислювальною складністю методів їх обробки. Тому якісне функціонування СВА вимагає або значних витрат ресурсів ІТС, або розробки нових підходів, які дозволять проводити обробку потоків вхідних даних в реальному часі. Для швидкоплинних процесів зміни стану системи вимога функціонування СВА в реальному часі може бути виконана шляхом скорочення множини параметрів, котрі визначають шаблони поведінки системи.

Аналіз останніх досліджень і публікацій [1-4] показав, що формування шаблонів поведінки систем здійснюється з використанням евристичних правил визначення множини контрольованих параметрів мережевого трафіку ІТС [1, 2]. Тобто розробник на етапі створення системи здійснює вибір множини контрольованих параметрів спираючись лише на власний досвід. Недоліками такого формування шаблонів поведінки ІТС є суб'єктивність вибору параметрів і повна відсутність математичного обґрунтування прийнятого рішення. Застосування інших методів обмежується в основному дослідженням можливості використання для розв'язання вказаної задачі методів нейронних мереж та SVM [3, 4]. До основних недоліків даних методів слід віднести локальну стійкість та їх слабку верифікованість.

Таким чином, з проведеного аналізу встановлено, що існує нагальна потреба у розробці верифікованого і глобально стійкого методу формування шаблонів поведінки ІТС, першим кроком на шляху реалізації якого є оцінка ваги і доцільності включення параметрів до зазначених шаблонів.

Метою статті є розробка методу оцінювання інформативності параметрів потоку вхідних даних для мережевих систем виявлення атак.

Основні матеріали дослідження. Представлені на ринку СВА базуються на двох засадницьких принципах: виявлення аномалій та виявлення зловживань. У обох випадках вхідними даними для роботи виступають сформовані на етапі розробки шаблони активності системи. Задача виявлення атаки у цих випадках зводиться до розпізнавання шаблону активності системи і фіксації факту атаки.

У роботі [5] доведено, що розпізнавання складних образів доцільно проводити на основі їх опису у просторі ознак. При цьому вибір інформативної системи ознак – найважливіша задача теорії розпізнавання, з розв'язанням якої зазвичай пов'язані питання спрощення системи розпізнавання і підвищення якості її роботи.

На сьогодні відомо два підходи до побудови інформативної системи ознак.

Перший - полягає у визначені малого числа ознак високої інформативності. Недоліком даного підходу вважається використання евристичних та емпіричних методів відбору ознак. Тобто його використання не дозволяє провести порівняльну оцінку двох систем, побудованих на базі одного підходу.

Другий - має на меті відбір з усієї множини ознак за заданим критерієм мінімально можливого числа корисних для розпізнавання. Його перевагою є можливість функціонально пов'язати критерій інформативності з ймовірністю похибки розпізнавання.

Але так чи інакше, обидва підходи до побудови інформативної системи ознак потребують визначення критеріїв оцінки інформативності.

На даний час розроблено різноманітні критерії оцінки інформативності параметрів, що базуються на методах математичної статистики і теорії інформації. Серед них слід виділити евристичний, інформаційний, статистичний, ймовірнісний і нейромережевий підходи (рис. 1).

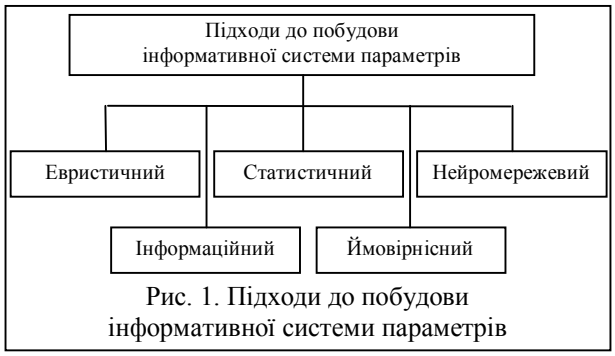

Рис. 1. Підходи до побудови інформативної системи параметрів

У роботі [5] дані підходи до відбору інформативних параметрів були проаналізовані за рядом критеріїв.

Виходячи з результатів експертного оцінювання [6], високий рівень математичного обґрунтування і середня відносна складність реалізації робить інформаційний підхід до вибору параметрів найбільш доцільним для розв'язання задачі скорочення розмірності потоку вхідних даних СВА.

Згідно обраного підходу, невизначеність стану (ентропія H) системи X, яка може приймати скінченну множину станів $x_1, x_2, ..., x_n$, з ймовірностями $p_1, p_2, ..., p_n$, визначається за формулою:

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i , \qquad (1)$$

де p_i – ймовірність того, що система X переходить в стан x_i.

Згідно [7] умовна ентропія системи Y за умови, що система X знаходиться в стані x_i, розраховується за формулою:

$$H(X \mid y_j) = -\sum_{j=1}^{m} p(x_i \mid y_j) \log p(x_i \mid y_j). \qquad (2)$$

При цьому зменшення ентропії системи X в результаті отримання інформації про систему Y може бути визначене як:

$$I_{Y \to X} = H(X) - H(X \mid y_j). \qquad (3)$$

Метод визначення інформативності параметрів потоку вхідних даних для мережевих СВА описується рівняннями (1) – (3) і полягає в такому.

На першому кроці здійснюється нормування параметрів потоку вхідних даних. Потреба нормування випливає з різної фізичної розмірності параметрів потоку вхідних даних. Правило нормування обираються в залежності від фізичних розмірностей параметрів. При цьому акцент робиться на виборі того методу нормування, що забезпечує найменший розкид нормованих параметрів.

Другий крок полягає у розрахунку таблиці розподілу частот двовимірної випадкової величини - контрольованого параметру Y та станів системи X (табл. 1).

Таблиця 1

Розподіл частот
двовимірної випадкової величини

Y	X					
	x_1	x_2	x_3	…	x_n	Total
y_1	k_{11}	k_{21}	k_{31}	…	k_{n1}	$k_{\Sigma 1}$
y_2	k_{12}	k_{22}	k_{32}	…	k_{n2}	$k_{\Sigma 2}$
…	…	…	…	…	…	…
y_m	k_{1m}	k_{2m}	k_{3m}	…	k_{nm}	$k_{\Sigma m}$
Total	$k_{1\Sigma}$	$k_{2\Sigma}$	$k_{3\Sigma}$	…	$k_{n\Sigma}$	$k_{\Sigma\Sigma}$

Для дискретної двовимірної випадкової величини X, Y, що має розмірність n, m, частотою елементу x_i, y_j є число k_{ij}, що відображає кількість реалізацій значення x_i, y_j у величині X, Y. А величини $k_{\Sigma 1}, k_{1\Sigma}$ є відповідними законами розподілу складових X та Y двомірної випадкової величини X, Y.

Третій крок методу полягає у розрахунку закону розподілу дискретної двомірної випадкової величини - контрольованого параметру Y та стану системи X (табл. 2).

Таблиця 2

Закон розподілу
дискретної двомірної випадкової величини

Y	X					
	x_1	x_2	x_3	…	x_n	Total
y_1	$P(x_1,y_1)$	$P(x_2,y_1)$	$P(x_3,y_1)$	…	$P(x_n,y_1)$	$P(y_1)$
y_2	$P(x_1,y_2)$	$P(x_2,y_2)$	$P(x_3,y_2)$	…	$P(x_n,y_2)$	$P(y_2)$
…	…	…	…	…	…	…
y_m	$P(x_1,y_m)$	$P(x_2,y_m)$	$P(x_3,y_m)$	…	$P(x_n,y_m)$	$P(y_m)$
Total	$P(x_1)$	$P(x_2)$	$P(x_3)$	…	$P(x_n)$	

Згідно теорії ймовірностей,, таблиця 2, що представляє закон розподілу, формується наступним чином. Перший рядок таблиці містить всі можливі значення складової X, а перший стовбець – всі можливі значення складової Y. Ймовірність $P(x_i, y_j)$ того, що дискретна двомірна випадкова величина прийме значення (x_i, y_j), розміщується на перетині «стовпця x_i» та «рядка y_j» і розраховується за формулою:

$$p(x_i, y_j) = \frac{k_{ij}}{k_{\Sigma\Sigma}}. \qquad (4)$$

Четвертий крок полягає у розрахунку таблиці розподілу умовних ймовірностей контрольованого параметру для кожного зі станів системи (табл. 3).

Таблиця 3

Розподіл умовних ймовірностей
двомірної випадкової величини

Y	X					
	x_1	x_2	x_3	…	x_n	Total
y_1	$P(x_1\|y_1)$	$P(x_2\|y_1)$	$P(x_3\|y_1)$	…	$P(x_n\|y_1)$	$P(X\|y_1)$
y_2	$P(x_1\|y_2)$	$P(x_2\|y_2)$	$P(x_3\|y_2)$	…	$P(x_n\|y_2)$	$P(X\|y_2)$
…	…	…	…	…	…	…
y_m	$P(x_1\|y_m)$	$P(x_2\|y_m)$	$P(x_3\|y_m)$	…	$P(x_n\|y_m)$	$P(X\|y_m)$
Total	$P(x_1\|Y)$	$P(x_2\|Y)$	$P(x_3\|Y)$	…	$P(x_n\|Y)$	

Так, для дискретної двомірної випадкової величини X, Y, складові якої набувають значень $x_1, x_2, ..., x_n$ та $y_1, y_2, ..., y_n$ відповідно, умовний закон розподілу X за умови, що відбулася подія $Y = y_j$, розраховується за формулою [7]:

$$p(x_i \mid y_j) = \frac{p(x_i, y_j)}{p(y_j)}. \qquad (5)$$

На п'ятому кроці методу здійснюється розрахунок ентропії:

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i. \qquad (6)$$

Шостий крок методу полягає у розрахунку умовної ентропії:

$$H(X \mid y_j) = -\sum_{j=1}^{m} p(x_i \mid y_j) \log p(x_i \mid y_j). \qquad (7)$$

На сьомому, заключному кроці розраховується зменшення ентропії системи:

$$I_{Y \to X} = H(X) - H(X \mid y_j). \qquad (8)$$

Верифікуємо метод та оцінимо його достовірність. Як потік вхідних даних оберемо базу шаблонів поведінки KDD99 (KDD99) [8]. Створена лабораторією Лінкольна, вона вперше була представлена на V-ій міжнародній конференції Knowledge Discovery and Data Mining і використовується розробниками академічних зразків СВА для проведення навчання і тесту-

вання створених ними систем. Вона являє собою модель функціонування локальної обчислювальної мережі ВВС США з доданими до нормальних шаблонів активності шаблонами типових атак. Включені до KDD99 шаблони атак представляють чотири категорії з множинами їх типів (табл. 4).

Таблиця 4

Шаблони атак

Атаки	Типи представлених в KDD99 «відомих» атак	Типи представлених в KDD99 «невідомих» атак
DoS	back, land, neptune (synflood), smurf, teardrop, pod (pingofdeath)	apache2, mailbomb, processtable, udpstorm
R2L	ftp_write, guess_passwd, imap, phf, multihop, spy, warezclient, warezmaster	httptunnel, worm, name, sendmail, xlock, xsnoop, snmpguess
U2R	bufferoverflow, perl, rootkit, loadmodule	Ps, sqlattack, xterm
PROBE	ipsweep, nmap, satan, portsweep	Mscan, saint

Шаблони поведінки в KDD99 записані у вигляді рядків і описані 41-м параметром мережевого з'єднання (табл. 5) та міткою стану системи (для навчального набору даних).

Таблиця 5

Параметри мережевого з'єднання

Базові параметри	**Параметри контенту**
1 duration	10 hot
2 protocol_type	11 num_failed_logins
3 service	12 logged_in
4 flag	13 num_compromised
5 src_bytes	14 root_shell
6 dst_bytes	15 su_attempted
7 land	16 num_root
8 wrong_fragment	17 num_file_creations
9 urgent	18 num_shells
Параметри з'єднання	
19 num_access_files	
20 num_outbound_cmds	
21 is_host_login	
22 is_guest_login	
23 count	
24 srv_count	
25 serror_rate	
26 srv_serror_rate	
27 rerror_rate	
28 srv_rerror_rate	
29 same_srv_rate	
30 diff_srv_rate	
31 srv_diff_host_rate	
32 dst_host_count	
33 dst_host_srv_count	
34 dst_host_same_srv_rate	
35 dst_host_diff_srv_rate	
36 dst_host_same_src_port_rate	
37 dst_host_srv_diff_host_rate	
38 dst_host_serror_rate	
39 dst_host_srv_serror_rate	
40 dst_host_rerror_rate	
41 dst_host_srv_rerror_rate	

1. Здійснимо нормування параметрів потоку вхідних даних. В залежності від їх фізичної розмірності застосуємо лінійне перетворення і метод нормування за супремумом.

2. Сформуємо таблицю розподілу частот двовимірної випадкової величини - контрольованого параметру $Y_{\text{тип\_протоколу}}$ та станів системи X (табл. 6).

Таблиця 6

Розподіл частот
двовимірної випадкової величини

Yтип протоколу	X					
	Normal	DoS	U2R	R2L	Probe	Total
ICMP	12763	2808150	0	0	12632	2833545
TCP	768670	1074241	1126	49	26512	1870598
UDP	191348	979	0	3	1958	194288
Total	972781	3883370	1126	52	41102	4898431

3. Згідно (4), розрахуємо закон розподілу дискретної двомірної випадкової величини - контрольованого параметру та стану системи X (табл. 7).

Таблиця 7

Закон розподілу
дискретної двомірної випадкової величини

Yтип протоколу	X					
	P_{Normal}	P_{DoS}	P_{U2R}	P_{R2L}	P_{Probe}	P_{Total}
P_{ICMP}	0.0026	0.5733	0	0	0.0026	0.5785
P_{TCP}	0.1569	0.2193	0.00023	0.000010003	0.0054	0.3819
P_{UDP}	0.0390	0.0002	0	0.000000612	0.0004	0.0396
P_{Total}	0.1986	0.7928	0.00023	0.000010615	0.0084	

4. Розрахуємо, згідно (5), таблицю розподілу умовних ймовірностей контрольованого параметру $Y_{\text{тип\_протоколу}}$ для кожного зі станів системи (табл. 8).

Таблиця 8

Розподіл умовних ймовірностей
контрольованого параметру

Yтип протоколу	X				
	P_{Normal}	P_{DoS}	P_{U2R}	P_{R2L}	P_{Probe}
P_{ICMP}	0.0045	0.9910	0	0	0.0045
P_{TCP}	0.4109	0.5743	0.0006	0.00002619	0.142
P_{UDP}	0.984	0.0050	0	0.00001544	0.0101

5. Використовуючи результати розрахунків закону розподілу дискретної двомірної випадкової величини (таблиця 7), згідно (6), обчислимо ентропію системи, яка для розглянутого випадку, розраховується за формулою:

$$H_{normal}(X) = -p_{normal}\log p_{normal} - p_{DoS}\log p_{DoS} - \\ -p_{R2L}\log p_{R2L} - p_{U2R}\log p_{U2R} - p_{Probe}\log p_{Probe}. \quad (9)$$

й дорівнює

$$H_{normal}(X) = -0.1986 * \log(0.1986) -$$

$$-0.7928*\log(0.7928)-0.00023*\log(0.00023)- -0.000010615*\log(0.000010615)- -0.0084*\log(0.0084)=0.7896. \quad (10)$$

6. За результатами розрахунків розподілу умовних ймовірностей контрольованого параметру для кожного зі станів системи (таблиця 8), згідно (7), обчислимо умовну ентропію системи:

$$H(X\mid y_{\text{тип\_протоколу}})=p_{\text{icmp}}\{H_{(\text{normal}\mid\text{icmp})}+ +H_{(\text{DoS}\mid\text{icmp})}+H_{(\text{R2L}\mid\text{icmp})}+H_{(\text{U2R}\mid\text{icmp})}+ +H_{(\text{Probe}\mid\text{icmp})}\}+p_{\text{tcp}}\{H_{(\text{normal}\mid\text{tcp})}+H_{(\text{DoS}\mid\text{tcp})}+ +H_{(\text{R2L}\mid\text{tcp})}+H_{(\text{U2R}\mid\text{tcp})}+H_{(\text{Probe}\mid\text{tcp})}\}+ +p_{\text{udp}}\{H_{(\text{normal}\mid\text{tcp})}+H_{(\text{DoS}\mid\text{tcp})}+H_{(\text{R2L}\mid\text{tcp})}+ +H_{(\text{U2R}\mid\text{tcp})}+H_{(\text{Probe}\mid\text{tcp})}\}. \quad (11)$$

$$H(X\mid y_{\text{тип\_протоколу}})= =0.5785*\{-0.0045*\log(0.0045)- -0.991*\log(0.991)-0.0045*\log(0.0045)\}+ +0.3819*\{-0.419*\log(0.419)- -0.5743*\log(0.5743)-0.0006*\log(0.0006)- -0.000026195*\log(0.000026195)\}+ +0.0397*\{-0.984*\log(0.984)- -0.3581*\log(0.3581)-0.0008*\log(0.0008)- -0.00003681*\log(0.00003681)- -0.027*\log(0.027)\}=0.4656. \quad (12)$$

7. Згідно (8), розрахуємо зменшення ентропії:

$$I_{Y\to X}=H(X)_{\text{normal}}-H(X\mid y_{\text{тип\_протоколу}}), \quad (13)$$

$$I_{Y\to X}=0,7896-0,4656=0,324. \quad (14)$$

Висновки

Отже, запропонований метод оцінювання інформативності параметрів потоку вхідних даних для мережевих систем виявлення атак може бути використаний для формування шаблонів поведінки ІТС.

Основною перевагою запропонованого методу є простота алгоритмізації та програмної реалізації. Використання методу в процесі формування шаблонів поведінки системи дозволить підвищити ефективність функціонування мережевих СВА за критерієм «швидкодія» та організувати роботу цих систем в реальному часі.

Список літератури

1. *Лукацкий А.В. Обнаружение атак / А.В. Лукацкий. – СПб: БХВ-Петербург, 2001. – 624 с.*
2. *Scarfone K. "Guide to Intrusion Detection and Prevention Systems (IDPS)" / Scarfone K., Mell P. // Computer Security Resource Center (National Institute of Standards and Technology). – Gaithersburg, Maryland, 2010.*
3. *Lunt A. "IDES: An Intelligent System for Detecting Intruders" / Lunt A., Teresa F. // Threats, and Countermeasures. - Rome, Italy, 1990. – P. 110–121.*
4. *Whitehurst R.A. "Expert Systems in Intrusion Detection: A Case Study." //Computer Science Laboratory, SRI International, Menlo Park, CA, November 1987.*
5. *Дубровин В.И. Интеллектуальные средства диагностики прогнозирования надежности авиадвигателей / Дубровин В.И., Субботин С.А., Богуслаев А.В., Яценко В.К. : [монография]. – Запорожье: ОАО «Мотор-Сич», 2003.– 279 с.*
6. *Мамарєв В.М. // Науково-технічне обґрунтування підходу щодо оцінювання інформативності параметрів потоку вхідних даних для мережевих систем виявлення атак. Вип.4 / Житомирський військовий інститут імені С.П. Корольова Національного авіаційного університету. – Житомир: ЖВІ НАУ, 2011. – С. 118-125.*
7. *Вентцель Е.С. Теория вероятностей / Вентцель Е.С. – М.: Наука, Главная редакция физико-математической литературы, 1969. – 576 с.*
8. *KDD 99 [Електронний ресурс]. - Режим доступу: http://kdd.ics.uci.edu/databases/kddcup99.html.*

Надійшла до редколегії 27.03.2012

Рецензент: д-р техн. наук, проф. В.О. Хорошко, Державний університет інформаційно-комунікаційних технологій, Київ.

МЕТОД ОЦЕНКИ ИНФОРМАТИВНОСТИ ПАРАМЕТРОВ ПОТОКА ВХОДЯЩИХ ДАННЫХ ДЛЯ СЕТЕВЫХ СИСТЕМ ОБНАРУЖЕНИЯ АТАК

Р.В. Грищук, В.Н. Мамарев

В статье проведен анализ подходов к построению информативной системы признаков. В результате проведения экспертной оценки, которых был избран наиболее целесообразный, для построения информативной системы признаков, информационный подход и на его основе разработан метод оценки информативности параметров потока входящих данных для сетевых систем обнаружения атак. Предложенный метод позволяет получать количественные оценки информативности параметров потока входящих данных для сетевых систем обнаружения атак, которые в дальнейшем могут быть использованы на этапе формировании шаблонов поведения информационно-телекоммуникационных систем.

Ключевые слова: *система защиты информации, система обнаружения атак, поток входящих данных.*

METHOD OF PARAMETERS' INFORMATION CONTENT ASSESSMENT OF THE INPUT DATA FLOW FOR THE NETWORK INTRUSION DETECTION SYSTEMS

R.V. Grischuk, V.N. Mamarev

The article analyzes the approaches to the informative features system construction. As a result of expert assessment an informational approach has been selected for the construction of informative features system and the method of assessment of the information content of the input data flow for the network intrusion detection systems elaborated on its basis. The proposed method allows obtaining quantitative estimations of the information content of the input data flow for the networking systems of attacks detection, which can be used subsequently at the stage of forming the behavior patterns of the information and telecommunication systems.

Keywords: *information protection system, intrusion detection system, incoming data flow.*