УДК 004.896

І. Р. Арсенюк

ЗМЕНШЕННЯ КІЛЬКОСТІ ІНФОРМАТИВНИХ ОЗНАК ДЛЯ ЗАДАЧІ ДЕТЕКТУВАННЯ КОМП'ЮТЕРНИХ АТАК

Вінницький національний технічний університет

Анотація

Метою роботи є зменшення кількості інформативних ознак для детектування комп'ютерних атак. Таке зменшення повинно підвищити швидкість та точність детектування.

Ключові слова: комп'ютерні атаки, детектування комп'ютерних атак, інформативні ознаки, вибір ознак, NSL-KDD.

Abstract

The purpose of the work is to reduce the number of informative features for detecting computer attacks. Such a decrease should improve both the speed and accuracy of detecting.

Keywords: computer attacks, computer attacks detecting, informative features, feature selection, NSL-KDD.

Вступ

Комп'ютерні мережі (КМ) сьогодні широко використовуються фактично в усіх напрямках життєдіяльності людини. Ні для кого не секрет, що неналежна організація системи захисту КМ може мати дуже сумні наслідки для компаній та закладів, що їх експлуатують. Натепер відомо багато різноманітних прикладів втрат різними компаніями десятків тисяч і навіть мільйонів доларів внаслідок вдалої реалізації атак їх КМ [1]. Відомо, що для детектування комп'ютерних атак часто використовують аналіз ряду ознак комп'ютерного трафіку. Для підвищення ефективності такого детектування дуже важливо не лише забезпечити відбір з доступної множини ознак найінформативніших, а й визначити таке їх сполучення, яке дасть змогу найточніше, найповніше та найшвидше здійснювати детектування, вказуючи наявність та прогнозований вид атаки. Таким чином, важливість відбору ознак для задачі детектування комп'ютерних атак не викликає сумнівів.

Обґрунтування вибору бази зразків комп'ютерних атак та її основні особливості

На даний момент є ряд різноманітних баз, що описують різні комп'ютерні атаки. Для проведення експериментів було розглянуто, найпопулярнішу на даний момент, базу сигнатур NSL-KDD [2, 3]. Ця база з'явилася за ініціативою американської асоціації перспективних оборонних наукових досліджень DARPA, шляхом аналізу та вдосконалення бази KDD-99 і містить широкий спектр зразків вторгнень, змодельованих у середовищі, що імітує мережу Військово-повітряних сил США.

База NSL-KDD вигідно відрізняється тим, що, зокрема: не містить надлишкових записів у навчальному наборі, у її тестовому наборі відсутнє дублювання записів, тестовий набір містить атаки, що відсутні у навчальному наборі (що дозволяє перевірити здатність класифікатора виявляти нові, не відомі йому раніше, види атак).

Варто зазначити, що навчальний набір бази містить 125973 зразки (рядки, записи) TCP/IP – з'єднань, що несуть інформацію про 21 різну атаку, а тестовий – 22544 зразки, що несуть інформацію про 37 різних атак.

Усі атаки в NSL-KDD поділені на чотири групи [4]:

1. Відмова в обслуговуванні (DoS, Denial of Service Attack) – зловмисник перевантажує систему запитами для того, щоб система не могла обробляти легальні запити. До групи DoS атак відносяться pod, land, back, smurf, teardrop, neptune, apache2, udpstorm, processtable та worm.

2. Неавторизований доступ до ядра (U2R, Users to Root Attack) – використання вразливостей для несанкціонованого доступу. Наприклад, переповнення буфера. До групи U2R атак відносяться buffer_overflow, loadmodule, perl, rootkit, sqlattack, xterm та ps.

3. Отримання доступу (R2L, Remote to Local Attack) – зловмисник надсилає у систему пакети, використовуючи вразливості, для отримання локального доступу. До групи R2L відносяться

guess_passwd, ftp_write, imap, phf, multihop, warezclient, warezmaster, spy, xlock, xsnoop, snmpguess, snmpgetattack, httptunnel, sendmail, named та mailbomb.

4. Сканування (Probe, Probing Attack) – сканування системи для отримання інформації про уразливість у ній. До групи Probe відносяться ipsweep, nmap, portsweep, satan, mscan та saint.

Кожен зразок NSL-KDD має 41 різні кількісні та якісні ознаки (атрибути) $f_1 - f_{41}$ та 42-а ознака f_{42} характеризує стан трафіку відповідного запису, який може бути або нормальним, або аномальним. Взагалі, усі ознаки можна поділити на три групи: до першої відносяться основні характеристики з'єднання (табл. 1); до другої – статистичні характеристики трафіку, що обчислюються з використанням двохсекундного часового вікна або протягом більшого часового проміжку (табл. 2); а до складу третьої – ознаки усередині окремого з'єднання (табл. 3).

Таблиця 1 – Основні ознаки.

Ознака	Тип	Опис
duration	continuous	тривалість з'єднання
protocol_type	discrete	тип протоколу (UDP, TCP, ICMP)
service	discrete	мережний протокол на вузлі призначення
flag	discrete	прапорці з'єднання (нормальне з'єднання або є помилка у з'єднанні)
src_bytes	continuous	кількість біт даних, переданих від джерела до вузла призначення
dst_bytes	continuous	кількість біт даних, переданих від вузла призначення до джерела
land	discrete	1, якщо з'єднання від того ж вузла/порта або до того ж вузла/порта; 0, у протилежному випадку
wrong_fragment	discrete	кількість неправильних фрагментів
urgent	discrete	кількість термінових пакетів

Таблиця 2 – Статистичні ознаки.

Ознака	Тип	Опис
count	continuous	кількість з'єднань з тим же вузлом, як і поточне з'єднання за останні 2 сек.
srv_count	continuous	кількість з'єднань з тим же мережним протоколом, як і поточне з'єднання за останні 2 сек.
serror_rate	continuous	відсоток з'єднань, що мають помилки SYN
srv_serror_rate	continuous	відсоток з'єднань з помилкою у пакеті SYN
rerror_rate	continuous	відсоток з'єднань, що мають помилки REJ
srv_rerror_rate	continuous	відсоток з'єднань з помилкою у пакеті REJ
same_srv_rate	continuous	відсоток з'єднань, що мали однаковий сервіс
diff_srv_rate	continuous	відсоток з'єднань на різні сервіси
srv_diff_host_rate	continuous	відсоток з'єднань від інших хостів

Таблиця 3 – Ознаки окремого з'єднання

Ознака	Тип	Опис
hot	continuous	кількість «гарячих» індикаторів
num_failed_logins	continuous	кількість невдалих спроб входу
logged_in	discrete	1 – якщо вдалий вхід, 0 – якщо ні
num_compromised	continuous	кількість умов, що ставлять пыдзагрозу
root_shell	discrete	1 – якщо коренева оболонка отримана
su_attempted	discrete	1 – якщо була команда "su root", 0 – якщо ні
num_root	continuous	кількість доступів через root
num_file_creations	continuous	кількість операцій створення файла
num_shells	continuous	кількість оболонок запиту
num_access_files	continuous	кількість операції з контролю доступу
num_outbound_cmds	continuous	кількість низхідних команд у сеансі ftp;
is_hot_login	discrete	1 – якщо вхід належіть до «гарячого» листа, 0 – якщо ні
is_guest_login	discrete	1 – якщо вхід здійснено гостем; 0 – якщо ні

Підхід до відбору інформативних атрибутів

Повертаючись до нашої задачі (відбору мінімальних множин найінформативніших ознак) варто зауважити, що основна її складність полягає в тому, що, навіть розуміючи сутність та значимість ко-

жної ознаки неможливо відібрати оптимальне сполучення заданої кількості ознак, тобто таке сполучення, на якому б класифікатор, що покладено в основу системи детектування атак, демонстрував би найбільшу повноту та найвищій відсоток правильно класифікованих станів трафіку. Зрозуміло, що нескладно визначити кореляцію та значимість кожної ознаки, проте відбір перших *n* ранжованих ознак, абсолютно не гарантує, що взявши довільно інші *n* ознак, ми отримаємо завжди гірші показники детектування. Повний перебір можливих підмножин ознак заданої потужності дав би відповідь на наше питання, проте час пошуку такого рішення, є неприйнятним. Наприклад, для відбору 7 з 41 ознак треба перевірити $C_{41}^{7} = \frac{41!}{(41-7)!7!} = 22481940$ варіантів; для відбору 10 ознак – 1121099408, а для 15 – 63432274896. Враховуючи також, що оскільки, у кінцевому підсумку, нас цікавить не конкретний розмір підмножини ознак (а такий їх розмір, який з одної сторони був би мінімально можливим, а з іншого, давав би максимально можливу і прийнятну повноту та точність детектування атак) – кількість досліджуваних (на деякому класифікаторі) підмножин суттєво зростає.

Ідея полягає в тому, що для того, щоб не потрапити у певний локальний максимум під час визначення шуканих підмножин ознак F_{sn}, F_{sD}, F_{sU}, F_{sR}, F_{sP} для визначення, відповідно, нормального трафіку, або атак DoS, U2R, R2L та Probe відповідно (рис. 1) можна випадковим чином сформувати за допомогою генератора G невеликий відсоток (від загальної їх кількості) різних підмножин (F_{si}) та оцінити для них точність, повноту результатів для деякого, простого та швидкого класифікатора (Cl). Після такого відбору підмножин селектор Sel подає їх на більш потужні класифікатори, або їх комбінації Cl_n, Cl_D, Cl_U, Cl_R, Cl_P, а далі на модуль прийняття остаточного рішення (DM), який дає відповідь Q щодо належності зразків до певного класу.

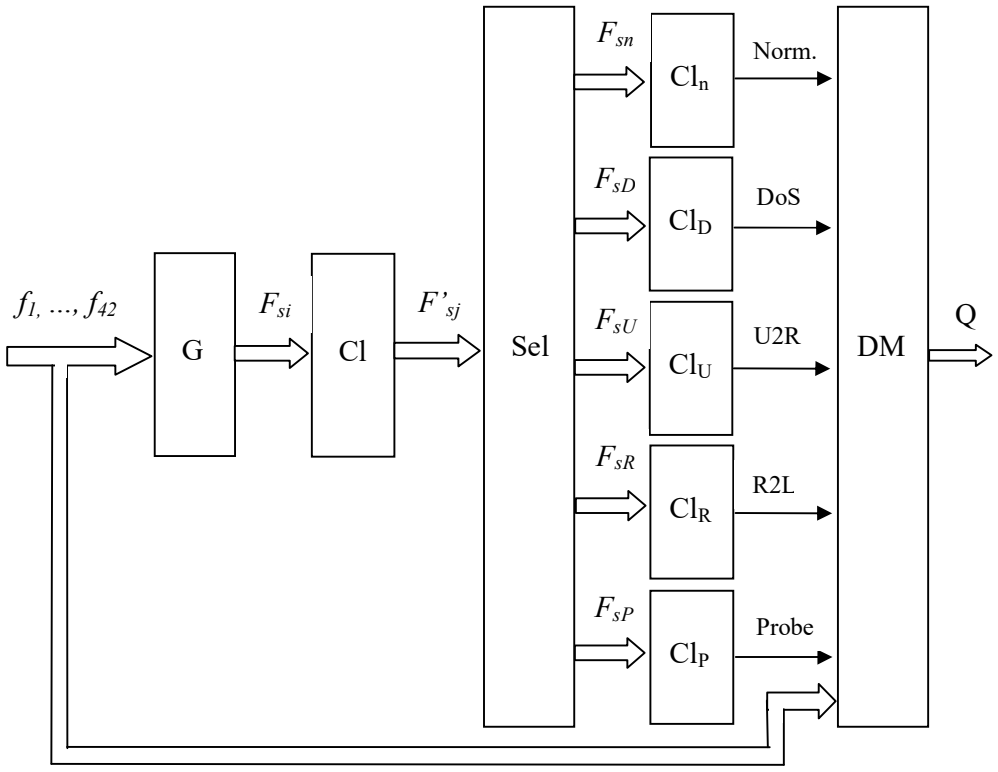

Рисунок 1 – Узагальнена структура детектора атак

Як показали експериментальні дослідження, досить гарні швидкісні та точнісні показники має алгоритм побудови дерев рішень J48 (C4.5), тому його і будемо використовувати для оцінювання корисності сформованих підмножин шуканих ознак.

Для такого оцінювання можна використовувати ряд, наприклад, таких показників:

1) кількість та відсоток правильно класифікованих зразків (*TI*);
2) кількість та відсоток неправильно класифікованих зразків (*FI*);
3) ймовірність істинно-позитивних результатів (True Positive Rate, TPR) – ймовірність правильної класифікації атак;

4) кількість помилково-негативних результатів (FN) – кількість зразків на яких класифікатор ідентифікує нормальний трафік, в той час, як реально, даний зразок характеризує атаку певної групи a_i: $FN = \sum_{i=1}^{4} C_{na_i}$, де C_{na_i} – кількість зразків, де будь-яку з чотирьох груп атак було детектовано як нормальний трафік.

5) ймовірність помилково-негативних результатів (False Negative Rate, FPR) – ймовірність отримання оповіщення про відсутність атаки, якщо ця атака є. Це помилки 1-го роду: *FNR=FN/All,* де *All* – загальна кількість зразків з'єднань.

6) кількість помилково-позитивних результатів (FP), тобто кількість зразків на яких класифікатор ідентифікує атаку певної групи a_i, в той час, як реально, трафік є нормальним. $FP = \sum_{i=1}^{4} C_{a_i n}$, де $C_{a_i n}$ – кількість зразків, де нормальний трафік було детектовано як будь-яку з чотирьох груп атак.

7) ймовірність помилково-позитивних результатів (False Positive Rate, FPR) – ймовірність отримання оповіщення про певну атаку, якщо ця атака відсутня. Це помилки 2-го роду *FPR=FP/All*.

8) ймовірність істинно-негативних результатів (True Negative Rate, TNR) – ймовірність правильної класифікації нормальних з'єднань;

9) кількість результатів неправильного визначення групи атаки: $FF = \sum_{i,j=1}^{4} C_{a_i a_j}$, $i \neq j$, де $C_{a_i a_j}$ – кількість зразків, де класифікатор вказує на атаку групи a_i, в той час, як реально має місце атака групи a_j, при цьому $i \neq j$.

10) міра точності (показує, яку долю об'єктів, що класифікатор передбачив як такі, що відносяться до деякого класу передбачено правильно): $Prec_q = \frac{C_{qq}}{C_{qq} + FN}$, де q – клас трафіку (може бути або нормальний, або відповідати одній з чотирьох групп атак), C_{qq} – кількість зразків, де класифікатор не помилився;

11) міра повноти (показує, яку долю об'єктів, що реально відносяться до деякого класу класифікатор передбачив правильно): $Rec_q = \frac{C_{qq}}{C_{qq} + FP}$;

12) F-міра (середнє гармонійне точності та повноти) $F_q = \frac{2 \cdot Prec_q \cdot Rec_q}{Prec_q + Rec_q}$, дозволяє зв'язати точність та повноту.

Для дослідження ефективності підмножин ознак було написано програму на Java з використанням API до програмного продукту WEKA 3.9.2 (Waikato Environment for Knowledge Analysis – безкоштовне середовище для інтелектуального аналізу даних і машинного навчання WEKA, написане на Java в університеті Вайкато, Нова Зеландія).

Результати експериментальних даних для різної кількості ознак наведено у таблиці 4.

Таблиця 4

Кількість ознак, тип	Ознаки	TI	FNR	FPR	FF (%)	Fq	Prec	Rec
41	1 – 41	17781 (78.872%)	4041 (17.925%)	439 (1.947%)	283 (1.255%)			
41, normal			439	4041		0,805	0,696	0,955
41, DoS			944	242		0,913	0,963	0,868
41, U2R			65	2		0,056	0,500	0,030
41, R2L			2832	6		0,196	0,983	0,109
41, Probe			483	472		0,802	0,804	0,800
5	2, 4, 22, 23, 30	17410 (77,226%)	4662 (20,679%)	228 (1,011%)	224 (1,082%)			
5, normal	3, 4, 6, 10, 11		138	3322		0,847	0,742	0,985
5, DoS	2, 4, 22, 23, 30		1092	169		0,906	0,973	0,848
5, U2R	1, 3, 14, 26, 29		42	10		0,490	0,714	0,373
5, R2L	3, 6, 7, 30, 37		2168	43		0,477	0,959	0,318
5, Probe	2, 4, 22, 23, 30		569	303		0,809	0,859	0,765

Таблиця 4 – Продовження

6	5, 7, 30, 32, 33, 40	17202 (76,304%)	3823 (16,958%)	538 (2,386%)	981 (4,351%)			
6, normal	1, 2, 9, 10, 11, 12		606	2868		0,840	0,760	0,938
6, DoS	5, 7, 30, 32, 33, 40		932	793		0,878	0,887	0,870
6, U2R	1, 4, 12, 14, 15, 41		43	8		0,485	0,750	0,358
6, R2L	3, 6, 7, 21, 22, 37		2264	49		0,441	0,949	0,288
6, Probe	4, 6, 22, 23, 30, 37		622	261		0,803	0,873	0,743
7	2, 3, 4, 20, 23, 30, 36	17730 (78,646%)	3988 (17,690%)	355 (1,575%)	471 (2,089%)			
7, normal	1, 3, 4, 5, 21, 37, 40		179	3535		0,837	0,729	0,981
7, DoS	1, 2, 4, 24, 26, 29, 30		1173	269		0,893	0,957	0,836
7, U2R	1, 3, 10, 11, 17, 22, 39		34	14		0,579	0,702	0,493
7, R2L	3, 6, 10, 21, 25, 29, 38		979	103		0,535	0,921	0,377
7, Probe	2, 5, 21, 23, 26, 30, 40		251	487		0,855	0,817	0,896
8	2, 4, 5, 6, 24, 26, 29, 30	17234 (76,446%)	4220 (18,719%)	313 (1,388%)	777 (3,447%)			
8, normal	3, 4, 5, 6, 15, 18, 27, 38		124	3499		0,841	0,733	0,987
8, DoS	2, 9, 14, 24, 26, 27, 30, 32		866	217		0,921	0,967	0,879
8, U2R	1, 3, 10, 14, 15, 16, 36, 39		42	6		0,510	0,806	0,373
8, R2L	3, 12, 18, 29, 30, 33, 36, 41		2003	36		0,535	0,970	0,370
8, Probe	1, 8, 30, 31, 33, 34, 37		594	176		0,826	0,912	0,755
9	1, 2, 4, 15, 18, 23, 30, 31, 36	17462 (77,457%)	3980 (17,654%)	316 (1,402%)	786 (3,487%)			
9, normal	2, 3, 7, 8, 18, 20, 21, 27, 36		366	2476		0,868	0,791	0,962
9, DoS	2, 3, 4, 13, 19, 27, 28, 30, 36		736	285		0,926	0,858	0,897
9, U2R	1, 7, 12, 14, 17, 27, 30, 38		38	8		0,558	0,784	0,433
9, R2L	3, 6, 14, 15, 27, 29, 30, 37, 40		2090	140		0,494	0,886	0,342
9, Probe	1, 2, 4, 15, 18, 23, 30, 31, 36		635	245		0,802	0,879	0,738
11	2, 4, 7, 8, 10, 21, 22, 23, 27, 28, 30	17728 (78,637%)	3803 (16,869%)	276 (1,224%)	737 (3,269%)			
11, normal	1, 3, 4, 5, 10, 12, 16, 22, 27, 37, 40		205	3440		0,839	0,734	0,979
11, DoS	1, 2, 5, 8, 15, 22, 23, 24, 26, 30		832	380		0,913	0,943	0,884
11, U2R	3, 12, 13, 14, 15, 17, 22, 23, 30, 31, 34		40	5		0,545	0,844	0,403
11, R2L	1, 3, 5, 6, 10, 23, 24, 28, 33, 37, 41		1787	44		0,603	0,969	0,438
11, Probe	1, 2, 4, 12, 15, 18, 23, 30, 32, 34, 40		504	468		0,798	0,804	0,792
15	1, 2, 3, 4, 5, 8, 11, 14, 18, 19, 22, 29, 30, 37, 38	18344 (81,370%)	3335 (14,793%)	359 (1,592%)	506 (2,244%)			
15, normal	1, 2, 3, 7, 11, 13, 15, 21, 23, 27, 28, 33, 34, 37, 38		346	3158		0,842	0,748	0,964
15, DoS	2, 3, 4, 6, 8, 10, 13, 15, 16, 22, 23, 29, 30, 37, 40		744	225		0,930	0,966	0,896
15, U2R	3, 4, 5, 10, 11, 12, 17, 21, 22, 25, 30, 37, 38, 40		40	9		0,524	0,750	0,402
15, R2L	1, 2, 3, 7, 11, 17, 21, 22, 23, 24, 25, 26, 29, 31, 32		1936	25		0,559	0,980	0,391
15, Probe	1, 2, 3, 4, 5, 8, 11, 14, 18, 19, 22, 29, 30, 37, 38		429	213		0,861	0,903	0,823
17	1, 2, 3, 4, 7, 8, 10, 11, 13, 14, 15, 20, 22, 27, 30, 36, 37	18036 (80,004%)	3767 (16,710%)	625 (2,772%)	116 (0,515%)			
17, normal	1, 2, 3, 4, 7, 10, 14, 18, 20, 21, 23, 25, 28, 30, 34, 38, 40		312	3427		0,834	0,733	0,968
17, DoS	2, 5, 6, 11, 13, 17, 18, 19, 23, 26, 27, 29, 30, 32, 33, 34		703	100		0,942	0,985	0,902
17, U2R	1, 2, 3, 4, 5, 8, 10, 14, 15, 16, 18, 22, 23, 26, 28, 36, 38		26	12		0,683	0,774	0,612
17, R2L	1, 3, 5, 6, 8, 9, 15, 16, 19, 20, 21, 22, 24, 26, 31, 37, 41		1958	66		0,547	0,949	0,384
17, Probe	1, 2, 3, 4, 6, 8, 11, 12, 14, 15, 17, 22, 29, 30, 31, 37, 40		325	272		0,875	0,885	0,866
19	1, 2, 3, 5, 11, 12, 13, 16, 20, 22, 23, 24, 26, 29, 30, 32, 34, 40, 41	18473 (81.942%)	3198 (14.186%)	400 (1.774%)	473 (2.098%)			
19, normal	2, 3, 4, 6, 7, 8, 9, 15, 16, 19, 20, 22, 27, 30, 32, 33, 34, 38, 40		324	2974		0,851	0,759	0,967
19, DoS	2, 3, 5, 6, 7, 9, 12, 13, 14, 17, 22, 23, 26, 27, 29, 30, 34, 36, 37		551	406		0,933	0,942	0,923
19, U2R	2, 3, 5, 6, 8, 9, 10, 12, 14, 17, 18, 20, 22, 23, 25, 28, 30, 35, 37		27	10		0,684	0,800	0,597
19, R2L	1, 2, 3, 4, 5, 10, 11, 13, 18, 19, 20, 21, 23, 25, 28, 29, 37, 38, 39		1688	32		0,634	0,979	0,469
19, Probe	1, 2, 3, 4, 5, 7, 9, 11, 13, 18, 20, 22, 26, 27, 28, 29, 30, 33, 40		338	273		0,872	0,884	0,860

Таблиця 4 – Продовження

20	2, 3, 5, 6, 8, 9, 12, 13, 16, 17, 19, 21, 23, 30, 31, 32, 34, 36, 37, 41	18397 (81.605%)	3550 (15,747%)	284 (1.260%)	313 (1.388%)			
20, normal	1, 2, 3, 4, 5, 6, 8, 10, 13, 17, 19, 21, 22, 23, 28, 35, 36, 39, 40, 41		248	3371		0,839	0,737	0,974
20, DoS	1, 2, 5, 6, 11, 12, 13, 17, 18, 19, 22, 26, 28, 30, 31, 33, 34, 37, 39, 41		674	245		0,934	0,964	0,906
20, U2R	1, 2, 3, 4, 5, 6, 7, 10, 11, 12, 16, 17, 18, 20, 22, 24, 26, 30, 31, 37		40	4		0,551	0,871	0,403
20, R2L	2, 3, 5, 6, 10, 11, 12, 13, 14, 15, 16, 17, 18, 21, 27, 29, 30, 34, 36, 37		3008	12		0,557	0,946	0,394
20, Probe	1, 2, 3, 5, 6, 12, 14, 15, 16, 17, 21, 22, 26, 27, 28, 29, 30, 37, 39, 41		422	182		0,869	0,917	0,826
25	1, 2, 3, 4, 5, 6, 8, 9, 10, 13, 14, 17, 18, 21, 22, 23, 24, 26, 28, 29, 30, 36, 39, 40, 41	18444 (81.813%)	3406 (15.108%)	310 (1.375%)				
25, normal	2, 3, 4, 5, 6, 7, 8, 10, 11, 13, 14, 16, 17, 18, 19, 20, 21, 22, 25, 26, 28, 33, 34, 36, 40		348	2988		0,849	0,758	0,964
25, DoS	2, 3, 5, 6, 7, 11, 12, 13, 14, 17, 18, 19, 20, 21, 22, 23, 28, 29, 30, 31, 32, 34, 35, 36, 40		395	235		0,956	0,966	0,945
25, U2R	1, 6, 7, 9, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20, 21, 24, 26, 27, 32, 33, 34, 35, 36, 39, 40		40	3		0,557	0,900	0,403
25, R2L	1, 2, 3, 4, 5, 8, 11, 16, 17, 18, 21, 26, 28, 29, 30, 31, 32, 34, 35, 36, 37, 38, 39, 40, 41		2021	16		0,532	0,986	0,346
25, Probe	1, 3, 4, 7, 8, 11, 13, 14, 16, 18, 19, 20, 23, 24, 25, 28, 29, 30, 31, 32, 34, 36, 38, 39, 40		419	344		0,840	0,853	0,827

Аналізуючи отримані результати, бачимо, що зменшення кількості інформативних ознак дозволило підвищити точність визначення класу трафіку відносно усієї множини ознак. Найкращими підмножинами, у випадку мінімізації кількості помилок 1 роду, виявилися такі:

DoS – 2, 3, 5, 6, 7, 11, 12, 13, 14, 17, 18, 19, 20, 21, 22, 23, 28, 29, 30, 31, 32, 34, 35, 36, 40;

U2R – 1, 2, 3, 4, 5, 8, 10, 14, 15, 16, 18, 22, 23, 26, 28, 36, 38;

R2L – 3, 6, 10, 21, 25, 29, 38;

Probe – 2, 5, 21, 23, 26, 30, 40.

Для порівняння отриманих результатів наведемо аналогічну таблицю (табл. 5), для ознак, визначених, наприклад, у роботах [4 – 9].

Таблиця 5

Кількість ознак,	Ознаки	TI	FNR	FPR	FF	Fq	Prec	Rec
20	1, 5, 7, 10, 11, 12, 13, 14, 18, 20, 21, 22, 23, 24, 26, 27, 30, 31, 37 [4]	15921 (70.62%)	5241 (23.248%)	721 (3.198%)	661 (2.932%)			
20,normal			721	5241		0,751	0,632	0,926
20, DoS			1538	772		0,830	0,879	0,785
20, U2R			62	2		0,135	0,714	0,075
20, R2L			3167	6		0,007	0,647	0,003
20, Probe			1135	602		0,597	0,681	0,531
25	1, 2, 4, 5, 6, 8, 10, 11, 13, 14, 16, 22, 23, 24, 27, 29, 31, 32, 33, 34, 35, 36, 37, 40, 41 [4]	17367 (77.036%)	4169 (18.493%)	249 (1.105%)	759 (3.367%)			
25,normal			249	4169		0,811	0,694	0,974
25, DoS			1087	685		0,873	0,899	0,848
25, U2R			65	0		0,058	1,000	0,030
25, R2L			2613	20		0,300	0,966	0,178
25, Probe			1163	303		0,632	0,806	0,520
25	1, 2, 5, 8, 10, 11, 12, 13, 14, 18, 20, 21, 22, 23, 24, 26, 27, 28, 29, 31, 35, 36, 37, 40, 41 [4]	17065 (75.696%)	4997 (22.166%)	301 (1.335%)	181 (0.803%)			
25,normal			301	4997		0,780	0,653	0,969
25, DoS			1517	139		0,872	0,976	0,788
25, U2R			67	1				
25, R2L			2774	3		0,225	0,993	0,127
25, Probe			820	339		0,734	0,825	0,661
17	3, 4, 5, 6, 8, 12, 17, 23, 25, 26, 29, 30, 31, 37, 38, 39 [5]	16634 (73.785%)	4282 (18.994%)	832 (3.691%)	796 (3.531%)			
17,normal			832	4282		0,776	0,675	0,914
17,DoS			1618	515		0,839	0,915	0,774
17,U2R			67	0				
17,R2L			3002	23		0,104	0,884	0,055
17,Probe			391	1090		0,732	0,650	0,838

Таблиця 5 – Продовження

15	3, 4, 5, 6, 12, 23, 25, 29, 30, 33, 34, 35, 38, 39 [6]	15865 (70.373%)	5155 (22.866%)	825 (3.660%)	699 (3.101%)			
15,normal			825	5155		0,748	0,633	0,915
15, DoS			1767	964		0,798	0,849	0,753
15, U2R			67	0				
15, R2L			2893	10		0,164	0,966	0,090
15, Probe			1127	550		0,607	0,702	0,534
25	2, 3, 4, 5, 6, 8, 10, 13, 23, 24, 25, 26, 27, 29, 30, 32, 33, 34, 35, 36, 37, 38, 39, 40 [7]	17979 (79.751%)	3730 (16.545%)	443 (1.965%)	392 (1.739%)			
25,normal			443	3730		0,816	0,713	0,954
25,DoS			944	402		0,902	0,939	0,868
25,U2R			67	0				
25,R2L			2626	6		0,296	0,989	0,174
25,Probe			485	427		0,809	0,819	0,800
25	2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 25, 26, 27, 28, 29, 30, 34, 35, 37, 38, 39, 40, 41 [7]	17508 (77.661%)	4263 (18.910%)	242 (1.073%)	531 (2.356%)			
25,normal			242	4263		0,808	0,690	0,975
25,DoS			898	565		0,896	0,917	0,875
25,U2R			65	3		0,060	0,400	0,030
25,R2L			2773	9		0,226	0,978	0,127
25,Probe			1058	196		0,685	0,874	0,563
25	2, 3, 4, 5, 6, 12, 23, 24, 25, 26, 27, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41 [7]	17345 (76.938%)	4061 (18.014%)	423 (1.876%)	715 (3.172%)			
25,normal			423	4061		0,806	0,696	0,956
25,DoS			1053	412		0,893	0,937	0,853
25,U2R			66	1		0,029	0,500	0,015
25,R2L			2656	24		0,029	0,956	0,164
25,Probe			1001	701		0,625	0,669	0,586
19	3, 4, 5, 6, 8, 12, 23, 25, 26, 29, 30, 31, 33, 34, 35, 37, 38, 39 [8]	15945 (70.994%)	5286 (23.447%)	842 (3.735%)	411 (1.823%)			
19,normal			842	5286		0,743	0,627	0,913
19, DoS			1628	458		0,842	0,924	0,773
19, U2R			66	1		0,029	0,500	0,015
19, R2L			3064	17		0,069	0,870	0,036
19, Probe			939	777		0,633	0,656	0,612
12, DoS	1, 2, 3, 4, 5, 6, 12, 23, 24, 31, 32, 37 [9]		1289	849		0,846	0,874	0,820
5, U2R	1, 2, 3, 10, 16 [9]		52	1		0,361	0,937	0,224
6, R2L	1, 2, 3, 4, 5, 10, 22 [9]		3009	12		0,101	0,934	0,053
12,Probe	1, 2, 3, 4, 12, 16, 25, 27, 28, 29, 30, 40 [9]		494	352		0,820	0,846	0,796

Аналізуючи дані таблиці, бачимо, що наш пошук множин інформативних ознак виявився не гіршим за ряд інших робіт.

Висновки

Запропоновано підхід щодо зменшення розмірності інформативних ознак TCP/IP – з'єднань бази NSL-KDD. В основу підходу покладено ідею визначення для кожного типу атаки та нормального трафіку ознак, які найбільш вдало їх характеризують. У загальному плані, задаючи той чи інший критерій пріоритетності, можна відібрати шукані відповідні підмножини ознак. Після визначення таких підмножин, пропонується використати потужніші класифікатори, а також, модуль, який, у випадку неоднозначності відповідей ансамблю класифікаторів, прийматиме остаточне рішення щодо відсутності або наявності атаки та належності останньої до певної групи. Для підвищення швидкодії пошуку підмножин пропонується задіяти обчислення з використання графічних процесорів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Шелухин О. И. Обнаружение вторжений в компьютерные сети [сетевые аномалии]: учебное пособие для вузов по направлению 210700 "Инфокоммуникационные технологии и системы связи" / О. И. Шелухин, Д. Ж. Сакалема, А. С. Филинова; ред. О. И. Шелухин. – М. : Горячая Линия-Телеком, 2013. – 220 с.

2. NSL-KDD data set for network-based intrusion detection systems [Електронний ресурс], Режим доступу: http://nsl.cs.unb.ca/NSL-KDD/, (дата звернення 10.03.2018) – Назва з екрана.

3. NSL-KDD data set [Електронний ресурс], Режим доступу: https://github.com/defcom17/NSL_KDD, (дата звернення 10.03.2018) – Назва з екрана.

4. Имамвердиев Я. Н. Обнаружение аномалий в сетевом трафике на основе информативных признаков / Я. Н. Имамвердиев, Л. В. Сухостат / Радіоелектроніка, інформатика, управління. 2017. № 3 – С. 113 – 120.

5. Chidananda M. P., Manjunatha A. S., Jaiswal A., Madhu B. R. Building Efficient Classifiers For Intrusion Detection With Reduction of Features. International Journal of Applied Engineering Research V. 11, № 6, 2016. [Электронный ресурс]: Режим доступу: https://www.ripublication.com/ijaer16/ijaerv11n6_143.pdf (дата звернення 10.03.2018) – Назва з екрана.

6. Nupur N. Majethiya, Dipak C. Patel. Efficient Intrusion Detection System with Reduced Dimensionality. IJETTCS, V. 4, № 2, March-April 2015. [Электронный ресурс]: Режим доступу: http://www.ijettcs.org/Volume4Issue2/IJETTCS-2015-04-06-77.pdf (дата звернення 10.03.2018) – Назва з екрана.

7. Singh N., Kaur A., Tech M. Feature selection for artificial neural network based intrusion detection system. International Journal For Technological Research In Engineering V. 2, Issue 11, July 2015. [Электронный ресурс]: Режим доступу: http://www.ijtre.com/images/scripts/2015021144.pdf (дата звернення 10.03.2018) – Назва з екрана.

8. Shrivas A. K., Singhai S. K., Hota H. S. An Efficient Decision Tree Model for Classification of Attacks with Feature Selection. International Journal of Computer Applications V. 84, № 14, December 2013/ [Электронный ресурс]: Режим доступу: https://pdfs.semanticscholar.org/a6d3/711e1c72b31006465f380d89fcf1760e2121.pdf (дата звернення 10.03.2018) – Назва з екрана.

9. Chou T. S., Yen K. K., Luo J." Network Intrusion Detection Design Using Feature Selection of Soft Computing Paradigms", International Journal Of Computational Intelligence, V. 2, № 1, 2008/ [Электронный ресурс]: Режим доступу: https://waset.org/publications/3936/network-intrusion-detection-design-using-feature-selection-of-soft-computing-paradigms (дата звернення 10.03.2018) – Назва з екрана.

Арсенюк Ігор Ростиславович – к. т. н., доцент, доцент кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця.

Igor R. Arsenyuk – Cand. Sc., Assistant Professor of the Chair of Computer Science, Vinnytsia National Technical University, Vinnytsia.