NETWORK AND APPLICATION SECURITY

УДК 004.056.57 (045)

СЕРГІЙ БОНДАРОВЕЦЬ,
ОКСАНА КОВАЛЬ,
СЕРГІЙ ГНАТЮК

СИСТЕМА ВИЯВЛЕННЯ АНОМАЛІЙ ДЛЯ ОПЕРАТОРА СТІЛЬНИКОВОГО ЗВ'ЯЗКУ ЗА КОНЦЕПЦІЄЮ BIG DATA

Постійне зростання використання інформаційних технологій у сучасному світі зумовило поступове збільшення обсягів даних, які циркулюють в інформаційно-телекомунікаційних системах, що в свою чергу породжує велику кількість нових загроз, які стає вже не так просто виявити. Стандартні методи виявлення загроз засновані на сигнатурному підході, який полягає у порівнянні трафіку, що надходить у мережу із базами даних відомих загроз. Проте такі методи стають неефективними, коли загроза є новою і її ще не встигли додати в базу. У такому випадку потрібно використовувати більш інтелектуальні методи, які здатні відстежувати будь-яку незвичну для конкретної системи активність – методи виявлення аномалій. Особливо гостро ця проблема постає для операторів стільникового зв'язку, які останнім часом дуже часто стикаються з різними видами шахрайства (витік міжнародного трафіка, фальшива тарифікація), які неможливо визначити у режимі реального часу. Тому доцільним є впровадження у мережу оператора інтелектуальної системи, що буде здатною обробляти великі масиви даних у реальному часі та попереджати про ймовірні загрози. Проте відомі загрози швидше виявлятимуться сигнатурним модулем, тому логічно включити в систему і його. Швидкодія такої системи буде забезпечуватись застосуванням методів та інструментів Big Data, які за рахунок використання розподіленої файлової системи та паралельних обчислень на багатьох серверах дозволять динамічно обробляти дані.

Ключові слова: виявлення аномалій, концепція Big Data, інформаційна безпека, аналіз даних, машинне навчання, стільниковий зв'язок, сигнатурне виявлення.

Постановка проблеми. Виявлення аномалій є однією з найважливіших концепцій аналізу даних. Інформаційний об'єкт вважається аномальним, якщо він суттєво відрізняється від звичайної поведінки даних у певній галузі. У загальному, це означає, що об'єкт є не таким як інші в конкретному масиві даних [1]. Важливо виявляти такі об'єкти під час аналізу даних, щоб розглядати їх під іншим кутом та використовуючи інші методи. У процесі виявлення аномалій дослідники стикаються з такими проблемами: визначення нормальної області, яку можливо представити в адекватному вигляді, часто є складною задачею; межа між нормальною та аномальною поведінкою не завжди є чіткою; точне визначення аномалії відрізняється в залежності від області застосування; наявність відповідних даних для тренування або перевірок; дані можуть містити шум; нормальна поведінка є динамічною та постійно еволюціонує [2, 6].

Методи виявлення аномалій широко застосовуються в наступних галузях: виявлення шахраїв у банківській та мобільній сфері; моніторинг стану апаратних засобів інформаційних систем; виявлення мережевих вторгнень; обробка зображень відео спостереження; виявлення підозрілих веб-сайтів тощо [7].

Якщо брати до уваги сферу стільникового зв'язку, то основними вимогами до поступово і неминуче зростаючих мобільних стільникових мереж є: висока пропускна спроможність; низькі витрати капіталу; низькі операційні витрати.

Ці вимоги продиктовані необхідністю високошвидкісного доступу до послуг зв'язку за помірні кошти. Тому технології радіодоступу і стільникові мережі постійно розвиваються і намагаються досягти більш ефективного використання радіоресурсів.

Постійне зростання кількості мобільних пристроїв охоплює багато аспектів безпеки, починаючи від захисту користувацької інформації і закінчуючи захистом провайдерів мобільного зв'язку від шахрайського використання їхніх послуг: клонування SIM-карт, маршрутизація зарубіжного трафіка через власні сервери зловмисників тощо.

Проте, незважаючи на зростаючу кількість подібних загроз, більшість мобільних операторів реагують на нові загрози вже після їх реалізації, а не діючи на випередження, що зумовлює необхідність використання більш сучасних систем виявлення загроз.

Аналіз існуючих досліджень та постановка завдання. Аналіз відомих методів виявлення аномалій проводився за такими критеріями (див. табл. 1) [1-11]: низька вимогливість до обчислювальних ресурсів (НВОР); відсутність потреби у певному розподілі даних (ВПРД); простота реалізації (ПР); мала кількість хибно-позитивних викидів (МКХПВ); можливість "навчання без вчителя" (НбВ).

Таблиця 1 – Багатокритеріальний аналіз сучасних методів виявлення аномалій

Метод	**Критерії**				
	НВОР	ВПРД	ПР	МКХПВ	НбВ
Нейронні мережі	+	+	+/–	+	–
Байесові мережі	+	+	+/–	+	–
Метод опорних векторів	+	+	+/–	+	–
Decision Tree	+	+	+	+	–
Використання відстані до k -го "найближчого сусіда"	–	+/–	+	+/–	+
Використання відносної щільності.	–	+/–	+	+/–	+
Кластеризація	+/–	+	+	–	+
Параметричні методи	+	–	–	+/–	+
Непараметричні методи	+	–	–	+/–	+
Складність Коломогорова	+/–	+	+/–	+/–	+
Ентропія	+/–	+	+/–	+/–	+
PCA	–	+/–	+/–	+/–	+

Відповідно до проведеного аналізу, однією з кращих "відправних точок" для побудови системи виявлення аномалій є метод Decision Tree. Тим не менш, недоліками, що об'єднують усі наведені вище методи є наступні:

– у той час, поки система виявлення аномалій навчається і будує нормальний профіль системи, сама система залишається в незахищеному стані;

– якщо злоякісна активність відповідає нормальному профілю системи, попередження про аномальну активність не відбудеться;

– хибно-позитивні спрацювання можуть виникати дуже часто;

– через скупчення великої кількості даних та абстрагування від конкретної інформації для переходу до математичного моделювання, оповіщення та попередження про аномалії можуть не містити достатньо інформації для подальшого аналізу [6, 10].

Для позбавлення визначених вище недоліків пропонується використовувати систему, яка поєднує в собі як виявлення нових аномалій, так і відстеження існуючих, використовуючи сигнатурні методики та наявні бази даних. Для підвищення швидкодії такої системи рекомендується використовувати методи й інструменти Big Data.

Отже, розробка системи виявлення аномалій, яка була б здатною швидко обробляти великі масиви даних у реальному часі та знаходити відхилення від нормальної поведінки мережі оператора є актуальною науково-практичною задачею, що має теоретичне та практичне значення. З огляду на це, **метою статті** є розробка гібридної системи виявлення аномалій для оператора стільникового зв'язку за концепцією Big Data. Для цього потрібно розв'язати такі **задачі**: 1) проаналізувати відомі методи виявлення аномалій, створити їх класифікацію та виділити основні переваги та недоліки; 2) розробити систему виявлення аномалій за концепцією Big Data; 3) експериментально дослідити модуль виявлення аномалій розробленої системи.

Теоретична частина розробки системи виявлення аномалій. Загальна структура розробленої системи зображена на рис. 1.

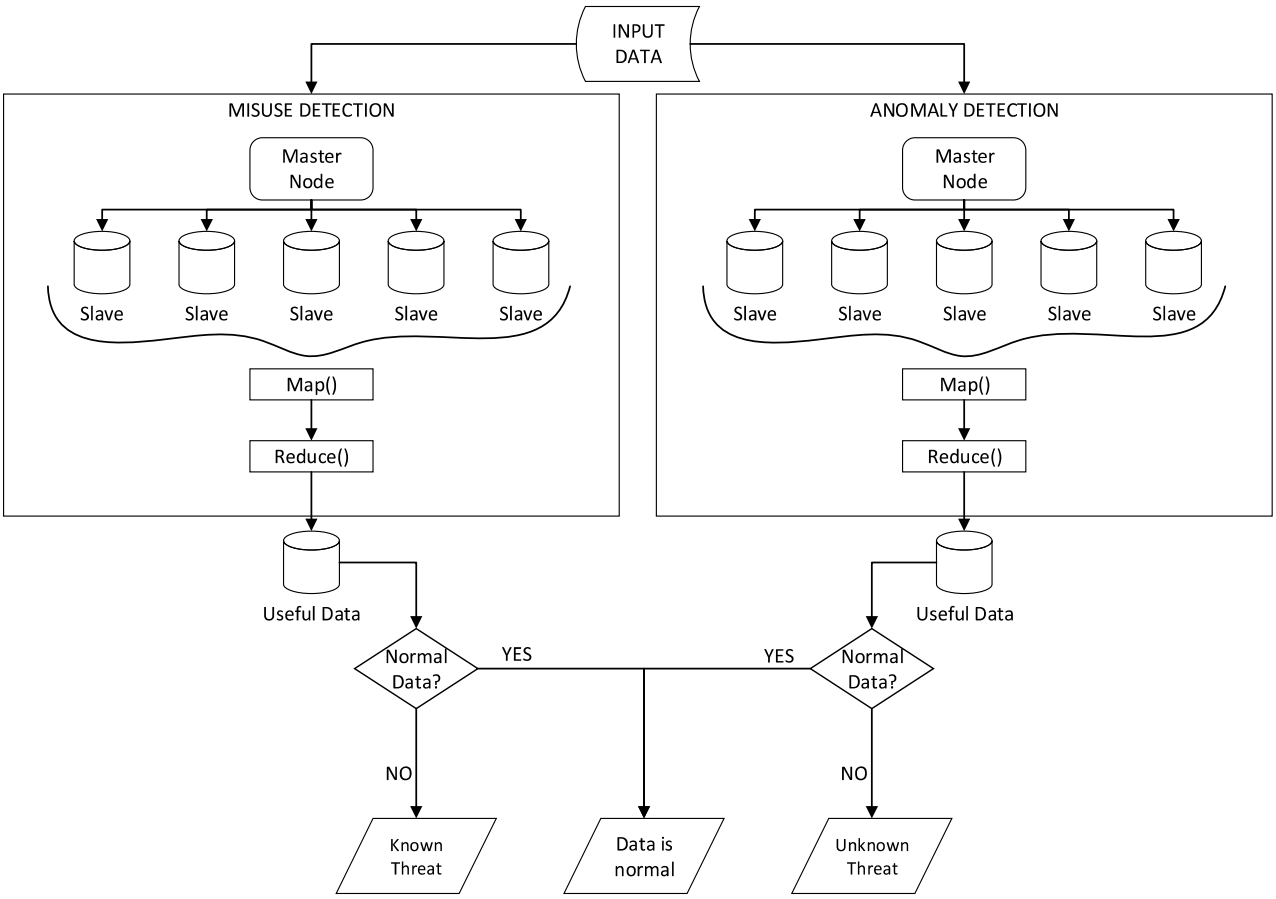

Рисунок 1 – Структура гібридної системи виявлення аномалій

Вхідні дані у паралельному режимі подаються на два модулі, у кожному з яких першим починає працювати головний вузол (Master Node), який розподіляє навантаження між робочими вузлами (Slave), на яких відбувається двокрокова реалізація методу MapReduce, на виході отримується корисна інформація, яка потім перевіряється умовами, і тоді результатом буде або рішення системи про нормальність даних, або класифікована загроза, або невідома активність. Логіка гібридної системи наведена у табл. 2.

Таблиця 2 – Загальна логіка гібридної системи

	Anomaly Detection	**Misuse Detection**	Пояснення
Спрацювання	0	0	*Дані нормальні*
	1	0	*Загроза виявлена*
	0	1	*Загроза виявлена і класифікована*
	1	1	*Загроза виявлена і класифікована*

Як модуль Misuse Detection доцільно використати open-source утиліту Snort, яка працює під управління операційних систем Windows і Linux.

Snort – це система виявлення вторгнень (СВВ), яка є надзвичайно потужним інструментом, навіть у порівнянні з комерційними СВВ. Багато користувачів у активній спільноті Snort діляться їхніми правилами безпеки, що може стати у нагоді, якщо потрібно мати найсучасніші правила [9].

Snort може бути запущений у 4 режимах:

– Sniffer mode (режим перехоплювача) – зчитування мережевого трафіку і виведення його на екран;

– Packet logger mode (режим збирання логів) – запис мережевого трафіку у файл;

– IDS mode (режим СВВ) – мережевий трафік, який відповідає правилам безпеки, записується;

– IPS mode (режим системи попередження вторгнень) – модифікований варіант попереднього режиму. Він приймає пакети від фаєрволу, порівнює їх з сигнатурними правилами і ставить мітку ”Відкинуто” у випадку, якщо пакети відповідають правилу.

Як модуль Anomaly Detection використаємо метод Decision Tree, основою якого є побудова так званого “дерева рішень”. Дерево складається з вузлів, які поділяються на внутрішні та термінальні, й гілок [1].

Внутрішні вузли розділяються на два дочірніх. Кожному внутрішньому вузлу відповідає одна з вхідних характеристик, а дочірні вузли містять кожне можливе значення для цієї характеристики.

Термінальні вузли містять мітку класу, з яким вони асоціюються, наприклад, спостереження, які відповідають конкретному термінальному вузлу. Для використання Decision Tree, як вхідних даних потрібно надати вектор характеристик. Якщо значення характеристики менше за визначене, тоді рішення переходить то лівого дочірнього вузла. Якщо більше – перехід до правого дочірнього вузла.

Процес продовжується, поки не буде досягнутий один з термінальних вузлів і мітка часу, яка відповідає термінальному вузлові, буде призначена шаблонові.

Індукційні алгоритми Decision Tree функціонують рекурсивно:

– спочатку обирається характеристика як головний вузол;

– для того, щоб створити найбільш ефективне (найменше) дерево, головний вузол повинен ефективно розподілити дані. Кожен розподіл має на меті зменшити набір значень (фактичних даних) до тих пір, поки вони всі не матимуть однакової класифікації. Найкращий розподіл забезпечує найбільше так зване підсилення інформації;

– дерево зростає шляхом рекурсивного розподілу кожного вузла, використовуючи характеристики, які забезпечують найкраще підсилення інформації до тих пір, поки термінальний вузол не стане відповідним.

Для обчислення посилення інформації для однієї характеристики використовується наступна послідовність дій:

1. Обчислюється ентропія для вузла A (див. рис. 2):

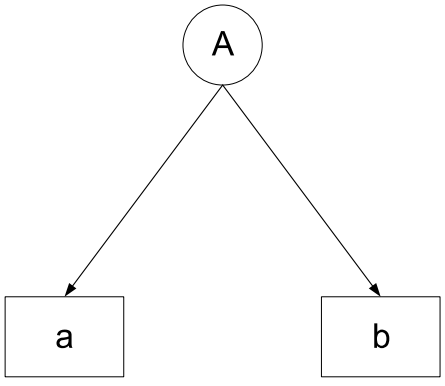

Рисунок 2 – Приклад моделі дерева

$$H(S) = -\left(\frac{M}{M+N}\right)\times\log_2\left(\frac{M}{M+N}\right)-\left(\frac{N}{N+M}\right)\times\log_2\left(\frac{N}{N+M}\right),$$

де M – кількість аномальних даних у вузлі А,

N – кількість нормальних даних у вузлі А,

$H(S)$ – значення ентропії перед розподілом.

2. Після того, як набір даних розподіляється на дві гілки за різними характеристиками, обраховується ентропія для вузла кожної гілки:

$$H_a = H(m,n);$$

$$H_a = -\left(\frac{m}{m+n}\right)\times\log_2\left(\frac{m}{m+n}\right)-\left(\frac{n}{n+m}\right)\times\log_2\left(\frac{n}{n+m}\right),$$

$$H_b = H(M-m, N-n),$$

$$H_b = -\left(\frac{M-m}{(M-m)+(N-n)}\right)\times\log_2\left(\frac{M-m}{(M-m)+(N-n)}\right)-$$

$$-\left(\frac{N-n}{(N-n)+(M-m)}\right)\times\log_2\left(\frac{N-n}{(N-n)+(M-m)}\right),$$

де m – кількість аномальних даних у вузлі *a,*

n – кількість нормальних даних у вузлі *a.*

3. Ентропія для вузлів кожної гілки пропорційно додається, щоб підрахувати загальну ентропію для розподілу:

$$H(S \mid A) = P_a \times H_a + P_b \times H_b,$$

$$H(S \mid A) = \left(\frac{m+n}{M+N}\right)\times H_a + \left(\frac{(M-m)+(N-n)}{(M+N)}\right)\times H_b,$$

де P_a – це відношення між кількістю елементів вузла *a* та кількістю елементів вузла *A*,

P_b – відношення між кількістю елементів вузла *b* та вузла *A*.

4. Отримане у результаті значення ентропії віднімається від її ж значення перед розподілом. Результатом цієї операції і буде інформаційне підсилення або зменшення ентропії:

$$I.G.(S, A) = H(S) - H(S \mid A).$$

Decision Tree є ”жадібним” алгоритмом і збільшує дерево згори донизу. На кожному вузлі обираються характеристики, які найкраще класифікують локальні елементи для тренування. Процес продовжується до тих пір, поки дерево найкращим чином не класифікує тренувальні елементи або поки не будуть використані усі характеристики [10].

Концепція Big Data. Для роботи з великими даними використовують цілу низку спеціально призначених методів. Одним із прикладів є MapReduce [3].

MapReduce – це програмний фреймоворк для розподіленого обчислення, що використовує метод ”розділяй і володарюй” для розбивання складних проблем великих даних на невеликі блоки роботи й обробляє їх у паралельному режимі.

MapReduce складається з двох кроків: крок ”Map” – дані з головного вузла розділяються на велику кількість менших підпроблем. Робочі вузли обробляють деякі підмножини менших проблем під контролем вузла JobTracker і зберігають результат у локальній файловій системі. Крок ”Reduce” – даний крок аналізує та виконує операцію зливання вхідних даних з попереднього кроку. Можлива наявність великої кількості задач Reduce з метою виконання процесів об’єднання у паралельному режимі, і ці задачі теж виконуються на робочих вузлах під контролем JobTracker.

Іншим методом є Hadoop. У складі Hadoop наявні: розподілена файлова система; платформи для аналізу та зберігання даних; рівень, що управляє паралельними обчисленнями; адміністрування конфігураціями [10].

Ще однією утилітою є Apache Spark. Apache Spark – це обчислювальний кластер, який забезпечує надзвичайно високу швидкість обробки даних та надійність. У ньому наявні програмні інтерфейси, що базуються на різних мовах програмування: Java, Python, Scala.

Він підтримує обчислення прямо в пам'яті (in-memory computing), що дозволяє отримувати доступ до даних та обробляти запити набагато швидше, ніж з використанням систем, заснованих на дисках (disk-based systems), таких як Hadoop [3].

У загальному, Spark є прогресивним та надзвичайно корисним оновленням до Hadoop, яке спрямоване на покращення можливості Hadoop аналізу у реальному часі.

Переваги Apache Spark перед конкурентами:

– найшвидше оброблення великих масивів даних.

– робочі процеси визначені у стилі, схожому на MapReduce, що спрощує його впровадження поруч із Hadoop.

– просте встановлення.

– Spark реалізований на Scala, сучасній об'єктно-орієнтованій мові програмування, із значною кількістю ресурсів.

– багато платформ даних підтримують Spark і стек його технологій (Map R, Cloudera, Databricks).

– надійність Spark може бути підтверджена рекомендацією Intel для використання його у рішеннях, пов'язаних з охороною здоров'я.

– однією з найбільш використовуваних функцій Spark є можливість об'єднувати набори даних із декількох несумісних джерел [8].

Експериментальне дослідження модуля виявлення аномалій розробленої системи.

Мета експерименту: перевірити результативність класифікації обраного методу виявлення аномалій.

Вхідні/вихідні дані експерименту: вхідними даними є 10% від набору даних KDDCup99, вихідними – класифіковані дані (нормальні чи аномальні).

KDDCup99 – це масив даних, що використовувався для Третього Міжнародного конкурсу з добування знань та майнінгу даних і містить набір параметрів, сукупність яких визначає чи поведінка мережі є нормальною, чи є однією з атак, присутніх у KDDCup99:

– DoS-атаки (спрямовані на спричинення збоїв у роботі апаратного забезпечення);

– U2R-атаки (спрямовані на здобуття доступу до користувацького облікового запису з подальшим доступом до запису адміністратора);

– R2L-атаки (атаки з віддалених робочих станцій);

– Probing-атаки (зондування мережі).

Середовище проведення експерименту: open-source застосунок Weka v.3.8.

Етапи проведення експерименту:

1. Завантаження вхідних даних у середовище.
2. Вибір класифікаційного алгоритму, у даному випадку – це J48 – реалізація на мові програмування Java алгоритму Decision Tree.
3. Побудова моделі Decision Tree.

Для перевірки точності обраного алгоритму був обраний режим крос-валідації із розбиванням вихідного масиву даних на 7 частин, 6 з яких використовуються для тренування, а 1 – для тестування.

4. Перегляд результатів експерименту:
   4.1. Кількість і відсоток правильно та помилково визначених даних (див. рис. 3)

```
Correctly Classified Instances      493820               99.9595 %
Incorrectly Classified Instances       200                0.0405 %
Kappa statistic                          0.9993
Mean absolute error                      0
Root mean squared error                  0.0057
Relative absolute error                  0.0962 %
Root relative squared error              3.5746 %
Total Number of Instances           494020
```

Рисунок 3 – Відсоток правильно та помилково визначених даних

4.2. Графічний вигляд побудованого дерева представлено на рис. 4.

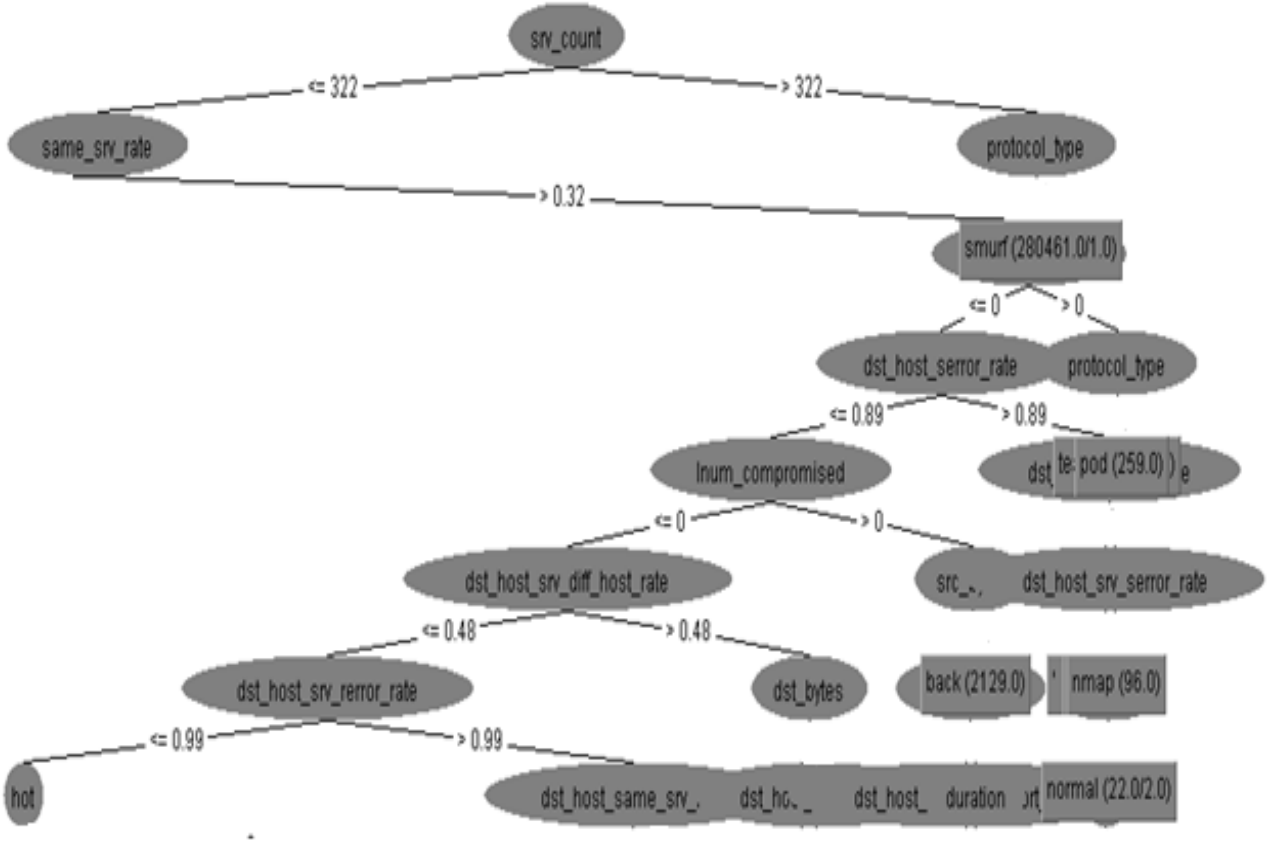

Рисунок 4 – Приклад відображення частини побудованого дерева

У результаті експерименту визначено, що відсоток правильно визначених даних становить 99.96%. Це доводить високу точність та стійкість до хибно-позитивних викидів обраного алгоритму, а також розглянуто модель побудованого дерева.

Висновки. Основним результатом дослідження є розроблена модель системи виявлення аномалій для оператора стільникового зв'язку на основі концепції Big Data. Під час виконання роботи було отримано такі результати:

1. Проаналізовано сучасні методи виявлення аномалій, що дозволило створити їх класифікацію та визначити основні недоліки.

2. Розроблена гібридна система виявлення аномалій, яка за рахунок використання методу Decision Tree, сигнатурного модуля Snort, технологій Big Data (HDFS, YARN, MapReduce, Spark) та бази даних KDDCup99 дозволяє виявляти аномалії в трафіку операторів стільникового зв'язку.

3. Експериментально досліджено модуль виявлення аномалій у застосунку Weka, що довело високу точність алгоритму. Практична цінність полягає у можливості інтеграції

розробленої системи виявлення аномалій у мережу операторів стільникового зв'язку та підвищенні відсотку виявлення загроз за рахунок використання сигнатурного модуля, що дозволяє виявляти відомі атаки.

СПИСОК ВИКОРИСТАННОЇ ЛІТЕРАТУРИ

[1] B. Abraham, and A. Chuang, "Outlier detection and time series modeling", *Technometrics,* vol. 31, iss. 2, pp. 241-248, May 1989.
doi: 10.2307/1268821.

[2] D. Barbara, Y. Li, J. Couto, J.-L. Lin, and S. Jajodia, "Bootstrapping a data mining intrusion detection system", in *Proc. of the 2003 ACM symposium on Applied computing (SAC '03)*, Melbourne, USA, pp. 421-425.
doi: 10.1145/952532.952616.

[3] H. Chen, R. Chiang, and V. Storey, "Business intelligence and analytics: From big data to big impact", *MIS Quarterly*, vol. 36, iss. 4, pp. 1165-1188, December 2012.

[4] P. Chan, M. Mahoney, and M. Arshad, "A machine learning approach to anomaly detection", Florida Institute of Technology, Melbourne, USA, Tech. Rep. CS-2003-06, March 2003.

[5] M. Mahoney, and P. Chan, "Learning nonstationary models of normal network traffic for detecting novel attacks", in *Proc. of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02)*, Edmonton, Canada, pp. 376-385.
doi: 10.1145/775047.775102.

[6] A. Nairac, T. Corbett-Clark, R. Ripley, N. Townsend, and L.Tarassenko, "Choosing an appropriate model for novelty detection", in *Proc. of the 5th IEEE International Conference on Artificial Neural Networks (Conf. Publ. No. 440)*, Cambridge, UK, pp. 117-122.
doi: 10.1049/cp:19970712.

[7] S. Papadimitriou, H. Kitagawa, P. Gibbons, and C. Faloutsos, "LOCI: Fast outlier detection using the local correlation integral", Carnegie Mellon University, Pittsburgh, USA, Tech. Rep. CMU-CS-02-188, November 2002.

[8] S. Ramaswamy, R. Rastogi, and K. Shim, " Efficient algorithms for mining outliers from large data sets", in *Proc. of the 2000 ACM SIGMOD international conference on Management of data (SIGMOD '00)*, Dallas, USA, pp. 427-438.
doi: 10.1145/335191.335437

[9] R. Rehman, *Intrusion Detection Systems with Snort: Advanced IDS Techniques Using Snort, Apache, MySQL, PHP, and ACID*, New Jersey, USA, Pearson Education LTD, 2003.

[10] A. Sebyala, T. Olukemi, and L. Sacks, "Active platform security through intrusion detection using naive bayesian network for anomaly detection", in *Proc. of the London communications symposium (2002)*, London, UK, pp. 1-5.

[11] J. Zhang, and H. Wang, "Detecting outlying subspaces for high-dimensional data: the new task, algorithms, and performance", *Knowledge and Information Systems*, vol. 10, iss. 3, pp. 333-355, October 2006.
doi: 10.1007/s10115-006-0020-z.

Стаття надійшла до редакції 22.03.2016.

REFERENCES

[1] B. Abraham, and A. Chuang, "Outlier detection and time series modeling", *Technometrics,* vol. 31, iss. 2, pp. 241-248, May 1989.
doi: 10.2307/1268821.

[2] D. Barbara, Y. Li, J. Couto, J.-L. Lin, and S. Jajodia, "Bootstrapping a data mining intrusion detection system", in *Proc. of the 2003 ACM symposium on Applied computing (SAC '03)*, Melbourne, USA, pp. 421-425.
doi: 10.1145/952532.952616.

[3] H. Chen, R. Chiang, and V. Storey, “Business intelligence and analytics: From big data to big impact”, *MIS Quarterly*, vol. 36, iss. 4, pp. 1165-1188, December 2012.
[4] P. Chan, M. Mahoney, and M. Arshad, “A machine learning approach to anomaly detection”, Florida Institute of Technology, Melbourne, USA, Tech. Rep. CS-2003-06, March 2003.
[5] M. Mahoney, and P. Chan, “Learning nonstationary models of normal network traffic for detecting novel attacks”, in *Proc. of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02)*, Edmonton, Canada, pp. 376-385. doi: 10.1145/775047.775102.
[6] A. Nairac, T. Corbett-Clark, R. Ripley, N. Townsend, and L.Tarassenko, “Choosing an appropriate model for novelty detection”, in *Proc. of the 5th IEEE International Conference on Artificial Neural Networks (Conf. Publ. No. 440)*, Cambridge, UK, pp. 117-122. doi: 10.1049/cp:19970712.
[7] S. Papadimitriou, H. Kitagawa, P. Gibbons, and C. Faloutsos, “LOCI: Fast outlier detection using the local correlation integral”, Carnegie Mellon University, Pittsburgh, USA, Tech. Rep. CMU-CS-02-188, November 2002.
[8] S. Ramaswamy, R. Rastogi, and K. Shim, “ Efficient algorithms for mining outliers from large data sets”, in *Proc. of the 2000 ACM SIGMOD international conference on Management of data (SIGMOD '00)*, Dallas, USA, pp. 427-438. doi: 10.1145/335191.335437
[9] R. Rehman, *Intrusion Detection Systems with Snort: Advanced IDS Techniques Using Snort, Apache, MySQL, PHP, and ACID*, New Jersey, USA, Pearson Education LTD, 2003.
[10] A. Sebyala, T. Olukemi, and L. Sacks, “Active platform security through intrusion detection using naive bayesian network for anomaly detection”, in *Proc. of the London communications symposium (2002)*, London, UK, pp. 1-5.
[11] J. Zhang, and H. Wang, “Detecting outlying subspaces for high-dimensional data: the new task, algorithms, and performance”, *Knowledge and Information Systems*, vol. 10, iss. 3, pp. 333-355, October 2006. doi: 10.1007/s10115-006-0020-z.

СЕРГЕЙ БОНДАРОВЕЦ,
ОКСАНА КОВАЛЬ,
СЕРГЕЙ ГНАТЮК

СИСТЕМА ОБНАРУЖЕНИЯ АНОМАЛИЙ ДЛЯ ОПЕРАТОРА СОТОВОЙ СВЯЗИ НА ОСНОВЕ КОНЦЕПЦИИ BIG DATA

Постоянный рост использования информационных технологий в современном мире обусловил постепенное увеличение объемов данных, циркулирующих в информационно-телекоммуникационных системах, что в свою очередь порождает большое количество новых угроз, которые становится уже не так просто найти. Стандартные методы обнаружения угроз основаны на сигнатурном методе, который заключается в сравнении трафика, который поступает в сеть с базами данных известных угроз. Однако такие методы становятся неэффективными, когда угроза является новой и её еще не успели добавить в базу. В таком случае нужно использовать более интеллектуальные методы, которые способны отслеживать любую необычную для конкретной системы активность – методы выявления аномалий. Особенно остро эта проблема стоит для операторов сотовой связи, которые в последнее время очень часто сталкиваются с различными видами мошенничества (утечка международного трафика, фальшивая тарификация), которые невозможно определить в режиме реального времени. Поэтому целесообразным является внедрение в сеть оператора интеллектуальной системы, которая будет способна обрабатывать большие массивы данных в реальном времени и предупреждать о возможных угрозах. Однако известные угрозы быстрее обнаруживаются сигнатурным модулем, поэтому логично включить в систему и его. Быстродействие такой системы будет обеспечиваться применением методов и инструментов Big Data, которые за

счет использования распределенной файловой системы и параллельных вычислений на многих серверах позволят динамично обрабатывать данные.

Ключевые слова: выявление аномалий, концепция Big Data, информационная безопасность, анализ данных, машинное обучение, сотовая связь, сигнатурное обнаружение.

SERHII BONDAROVETS,
OKSANA KOVAL,
SERHII HNATIUK

ANOMALY DETECTION SYSTEM FOR MOBILE CARRIER BASED ON BIG DATA CONCEPT

The continuous growth of information technologies in the modern world has caused a gradual increase in data circulating in the information and telecommunication systems, which in turn generates a large number of new threats, that is not so easy to detect. Standard methods of detection based on the signature method, which is comparing the traffic coming into the network with databases of known threats. However, these methods are ineffective when the threat is new and it has not yet been added to the database. In this case, it is necessary to use a more intelligent methods, which are able to monitor any unusual activity for a particular system – the methods of anomaly detection. Particularly, this problem is actual for mobile operators that have recently often face different types of fraud (leakage international traffic, false billing), which is impossible to determine in real time. Therefore, it is appropriate to implement in carrier's network intelligent system that is able to process large amounts of data in real time and warn about possible threats. However, known threats will be faster detected by signature module, so it is logical to include it in system. The performance of the system will be provided using the methods and tools of Big Data, concretely by using a distributed file system and parallel computing on multiple servers will dynamically process data. That anomaly detection system was developed in this paper.

Key words: anomaly detection, Big Data concept, information security, data analysis, machine learning, cellular communication, signature detection.

Бондаровець Сергій Сергійович, студент бакалаврату кафедри безпеки інформаційних технологій, Національний авіаційний університет, Київ, Україна.
E-mail: bondss29@gmail.com
Коваль Оксана Сергіївна, студентка бакалаврату кафедри безпеки інформаційних технологій, Національний авіаційний університет, Київ, Україна.
E-mail: oksanakoval@mail.ua
Гнатюк Сергій Олександрович, кандидат технічних наук, доцент, доцент кафедри безпеки інформаційних технологій, Національний авіаційний університет, Київ, Україна.
E-mail: s.gnatyuk@nau.edu.ua

Бондаровец Сергей Сергеевич, студент бакалаврата кафедры безопасности информационных технологий, Национальный авиационный университет, Киев, Украина.
Коваль Оксана Сергеевна, студентка бакалаврата кафедры безопасности информационных технологий, Национальный авиационный университет, Киев, Украина.
Гнатюк Сергей Александрович, кандитат технических наук, доцент, доцент кафедры информационных технологий, Национальный авиационный университет, Киев, Украина.

Serhii Bondarovets, bachelor student of IT-security academic department, National aviation university, Kyiv, Ukraine.
Oksana Koval, bachelor student of IT-security academic department, National aviation university, Kyiv, Ukraine.
Serhii Hnatiuk, candidate of technical sciences, associate professor, associate professor of IT-security academic department, National aviation university, Kyiv, Ukraine.