МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ АВІАЦІЙНИЙ УНІВЕРСИТЕТ
КАФЕДРА БЕЗПЕКИ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

ДОПУСТИТИ ДО ЗАХИСТУ
Завідувач кафедри

_____

«____» _____ 20 р.

На правах рукопису
УДК 004.056.57(084.93)

ДИПЛОМНА РОБОТА

ВИПУСКНИКА ОСВІТНЬО-КВАЛІФІКАЦІЙНОГО РІВНЯ «Магістр»

Тема: Система виявлення аномалій для оператора стільникового зв’язку за концепцією Big Data

Автор: С.Ю. Тарасюк

Науковий керівник: к.т.н., доц.

Нормоконтролер: к.т.н., доц.

Київ 2022

ДЕРЖАВНИЙ УНІВЕСТИТЕТ ТЕЛЕКОМУНІКАЦІЙ

Інститут:
Кафедра:
Освітньо-кваліфікаційний рівень:
Напрям:

ЗАТВЕРДЖУЮ
Завідувач кафедри

_____

«» л 202 р.

З А В Д А Н Н Я

на виконання дипломної роботи

студента Тарасюка Сергія Юрійовича

1. Тема: *Система виявлення аномалій для оператора стільникового зв'язку за концепцією Big Data.*

затверджена наказом ректора від « » 202 р. № .

2. Термін виконання: *з* «» 202 р. до «» 202 р.

3. Вихідні дані: *проаналізувати існуючі інформаційно-аналітичні та інші системи виявлення аномалій; дослідити інформаційно-аналітичні системи, що використовуються операторами стільникового зв'язку; розробити і експериментально дослідити модель інформаційно-аналітичної системи виявлення аномалій для оператора стільникового зв'язку за концепцією Big Data.*

4. Зміст пояснювальної записки (перелік питань, що підлягають розробці):

1. *Аналіз сучасних систем виявлення аномалій.*
2. *Розробка моделі інформаційно-аналітичної системи для оператора стільникового зв'язку за концепцією Big Data.*
3. *Експериментальне дослідження розробленої системи.*

КАЛЕНДАРНИЙ ПЛАН

виконання дипломної роботи

№ п/п	Етапи виконання бакалаврської атестаційної роботи	Термін виконання етапів	Примітка
1.	*Уточнення завдання*	12	
2.	*Аналіз літературних джерел*		
3.	*Обґрунтування вибору рішення*		
4.	*Збір інформації*		
5.	*Аналіз існуючих інформаційно-аналітичних та інших систем виявлення аномалій*		
6.	*Дослідження інформаційно-аналітичних систем, що використовуються операторами стільникового зв'язку*		
7.	*Розробка моделі інформаційно-аналітичної системи для оператора стільникового зв'язку за концепцією Big Data*		
8.	*Проведення експериментального дослідження моделі розробленої системи.*		
9.	*Оформлення і друк пояснювальної записки*		
10.	*Оформлення презентації*		
11.	*Отримання рецензій*		
12.	*Захист в ЕК*		

Дипломник С.Ю. Тарасюк

(підпис, дата)

Науковий керівник

(підпис, дата)

РЕФЕРАТ

Дипломна робота складається зі вступу, трьох розділів, загальних висновків, списку використаних джерел, додатків, загальним обсягом 59 сторінок, має 25 рисунків, 10 таблиць, 7 сторінок додатків. Список використаних джерел містить 63 найменування та займає 6 сторінок.

Метою дипломної роботи є розробка системи виявлення аномалій для оператора стільникового зв'язку на основі концепції Big Data.

В дипломній роботі розглянуто питання особливостей існуючих методів виявлення аномалій, що дало змогу створити їх класифікацію та виділити основні переваги та недоліки. Розроблено систему виявлення аномалій на основі концепції Big Data та експериментально досліджено модуль виявлення аномалій розробленої системи.

Новизна роботи полягає у тому, що вперше запропоновано структуру системи 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.

Практична цінність роботи полягає у тому, що є можливість інтеграції розробленої системи виявлення аномалій у мережу операторів стільникового зв'язку; за рахунок використання сигнатурного модуля, що дозволяє виявляти відомі атаки, було підвищено відсоток виявлення загроз. Крім цього, проведений експеримент довів високу точність модуля виявлення аномалій.

Ключові слова: ВИЯВЛЕННЯ АНОМАЛІЙ, КОНЦЕПЦІЯ BIG DATA, ІНФОРМАЦІЙНА БЕЗПЕКА, АНАЛІЗ ДАНИХ, МАШИННЕ НАВЧАННЯ, СТІЛЬНИКОВИЙ ЗВ'ЯЗОК, СИГНАТУРНЕ ВИЯВЛЕННЯ.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ 7

ВСТУП 8

Розділ 1. АНАЛІЗ СУЧАСНИХ МЕТОДІВ ВИЯВЛЕННЯ АНОМАЛІЙ 10

1.1. Підходи і методи до виявлення аномалій 10

1.2. Методи виявлення аномалій в стільникових мережах 16

1.3. Переваги та недоліки існуючих підходів до виявлення аномалій та шляхи їх удосконалення. 26

1.4. Висновки до першого розділу 29

Розділ 2. РОЗРОБКА МОДЕЛІ СИСТЕМИ ВИЯВЛЕННЯ АНОМАЛІЙ 30

2.1. Застосування систем виявлення аномалій в мережах операторів . 30

2.2. Структурна схема системи виявлення аномалій 36

2.3. Концепція Big Data 45

2.4. Проблеми інформаційної безпеки Big Data в мережах оператора стільникового зв’язку 54

2.5. Удосконалення початкової моделі за допомогою концепції Big Data 58

2.6. Висновки до другого розділу 59

Розділ 3. ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ МОДУЛЯ ВИЯВЛЕННЯ АНОМАЛІЙ РОЗРОБЛЕНОЇ СИСТЕМИ 60

3.1. Методика експерименту 60

3.2. Середовище Weka v.3.8 60

3.3. Проведення експерименту 66

3.4. Висновки до третього розділу 69

ВИСНОВКИ 70

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 71

Додаток А ..... **Ошибка! Закладка не определена.**

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

ВНМ – випадкові нейронні мережі.

ЗБД – знання у базі даних.

ОГ – основа графа.

СВВ – система виявлення вторгнень.

ECL – enterprise control language (мова програмування, орієнтована на роботу з даними).

EM – expectation-maximization (максимізація очікувань).

ESD – extreme studentized deviate (екстремальне студентизоване відхилення).

HDFS – hadoop distributed file system (розподілена файлова система).

HPCC – high-performance computing cluster (високошвидкісний обчислюваний кластер).

IDS – intrusion detection system (система виявлення вторгнень).

IPS – intrusion prevention system (система запобігання вторгненням).

KDD – Knowledge Discovery and Data Mining (добування знань та аналіз даних).

LOF – local outlier factor (фактор локальної аномалії).

LSA – latent semantic analysis (прихований семантичний аналіз).

PCA – principal component analysis (аналіз головних компонент).

QoS – quality of service (якість послуг).

SHESD – seasonal hybrid extreme studentized deviate (сезонне гібридне екстремальне студентизоване відхилення).

SIM – subscriber identification module (модуль ідентифікації абонентів).

SQL – structured query language (структурована мова запитів).

SVM – support vector machine (метод опорних векторів).

XML – extensible markup language (розширювана мова розмітки).

YARN – yet another resource negotiator («ще один ресурсний посередник» – модуль управління ресурсами кластерів та планування завдань.

ВСТУП

Актуальність роботи. 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 методі, який полягає в тому, що порівнює трафік, який надходить в 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.

Метою дослідження є розробка системи виявлення аномалій для оператора стільникового зв'язку на основі концепції Big Data.

Для цього потрібно розв'язати наступні **задачі**:

1. Аналіз існуючих методів виявлення аномалій, створення їх класифікації та виділення основних переваг та недоліків.
2. Розробка системи виявлення аномалій на основі концепції Big Data.
3. Експериментальне дослідження модуля виявлення аномалій розробленої системи.

Об'єктом дослідження є процес виявлення аномалій в інформаційно-телекомунікаційних системах.

Предметом дослідження є методи, моделі, системи виявлення аномалій для операторів стільникового зв'язку.

Наукова новизна: вперше запропоновано структуру системи 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.

Практична цінність:

- Можливість інтеграції розробленої системи виявлення аномалій у мережу операторів стільникового зв’язку.
- Підвищення відсотку виявлення загроз за рахунок використання сигнатурного модуля, що дозволяє виявляти відомі атаки.
- Експериментальне доведення високої точності модуля виявлення аномалій.

Розділ 1. АНАЛІЗ СУЧАСНИХ МЕТОДІВ ВИЯВЛЕННЯ АНОМАЛІЙ

1.1. Підходи і методи до виявлення аномалій

Виявлення аномалій є однією з найважливіших концепцій аналізу даних. Інформаційний об’єкт вважається аномалією, якщо він суттєво відрізняється від звичайної поведінки даних у певній галузі. В загальному, це означає, що об’єкт є не таким як інші в конкретному масиві даних. Важливо виявляти такі об’єкти під час аналізу даних, щоб розглядати їх під іншим кутом та використовуючи інші методи.

Приклад аномалій зображено на рис. 1.1, де області N1 і N2 є нормальними, а точки O1 і O2, а також область O3 – аномальними.

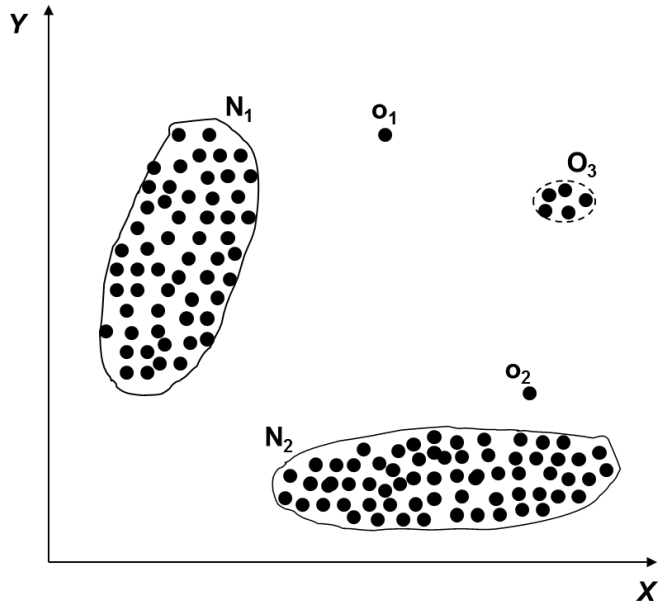

Рис. 1.1. Приклад аномалій

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 й постійно еволюціонує.

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.

Методи виявлення аномалій відрізняються в залежності від того, в якій формі представлені вихідні дані. Зазвичай, представлення даних буває наступним: метричні дані; дані, що еволюціонують; мультиструктуровані дані.

Метричні дані є найбільш поширеною формою представлення даних. У цьому випадку кожний об’єкт у наборі даних має певний набір властивостей, що дозволяє взаємодіяти з такими поняттями як «несхожість» і «наближеність».

Дані, що еволюціонують – це добре вивчені об’єкти: дискретні послідовності, часові серії і багаторозмірні потоки даних.

Мультиструктуровані дані – це дані, які представлені у неструктурованому, напів-структурованому або структурованому вигляді. Такі дані можуть не мати чіткої структури, але між ними можуть бути певні залежності.

Методи, що оперують метричними даними. Дані в цьому випадку представлені об’єктами в просторі, так званими точками. Тому завданням таких методів є визначення звичайних і незвичайних точок, спираючись на такі атрибути даних, як відстань між об’єктами, взаємодія між ними та розподіл даних.

Ці дані є структурованими і не залежать від часу. Метрична форма даних є найбільш розповсюдженою через те, що майже всі об’єкти можуть бути представлені через набір характеристик, а відтак – як точка в конкретному просторі.

1. Методи, засновані на відстані між об’єктами.

Базовий набір таких методів включає в себе *кластеризацію*, *метод K-nearest neighbors* та їхні похідні.

Методи кластеризації використовують відстань, визначену в просторі, щоб відокремити дані в однорідні щільні групи (кластери). Якщо точка не

знаходиться у великих кластерах, вона класифікується як аномалія. Також можна припустити, що невеликі кластери можуть бути аномальними, тому що аномалії можуть також мати подібну структуру, а значить, кластеризуватися.

Метод K-nearest neighbors заснований на концепції наближення. Розглядаються k найближчих точок на основі певних правил, які вирішують чи є об’єкт аномалією. Найпростішим прикладом такого правила є відстань між об’єктами, тобто найвіддаленіша точка від своїх «сусідів» швидше за все є аномалією.

Існує багато видів правил, починаючи від заснованих на відстані і до заснованих на розподілі «сусідів». Наприклад, LOF (local outlier factor) аналізує щільність розміщення «сусідів» в «районі».

2. Методи, засновані на взаємодії між об’єктами.

Головна ідея таких методів спирається на концепцію взаємодії між характеристиками об’єктів. Така ситуація часто зустрічається в реальних даних, тому що різні характеристики можуть бути породжені одним і тим самим процесом.

Прикладом таких методів є лінійна регресія. Використовуючи її ми будуємо певну площину, яка описує наявні дані, а тоді в якості аномальних об’єктів ми використовуємо, як знаходяться далеко від цієї площини. Також часто використовується PCA (principal component analysis) як інструмент для зменшення розмірності даних, тому він іноді застосовується в попередній обробці даних. Проте його можна використати напряму для відокремлення аномалій. У такому випадку основою є той факт, що в нових розмірах легше відокремити аномальні об’єкти.

3. Методи, засновані на ймовірнісному розподілі.

У даних методах основним підходом є припущення, що дані відповідають певному закону розподілу. Тому, аномальні об’єкти можуть бути визначені, як такі, що не відповідають такому правилу.

Прикладом такого методу є EM (expectation-maximization), ітеративний алгоритм, заснований на знаходженні максимальної схожості. Кожна ітерація складається з «очікування» та «максимізації».

На етапі очікування відбувається розрахунок функції схожості, а на етапі максимізації – відшукується параметр, який максимізує функцію схожості.

Іншими прикладами є методи, що використовують статистичні закономірності та включають спостережний аналіз розподілу і використовують, наприклад, нерівність Маркова.

4. Методи, що застосовуються на категоризованих даних.

Дані методи оперують з безперервними даними, тому одним з підходів є перетворення інформації, що входить до певної категорії, в безперервні характеристики. Як приклад, категоризовані дані можуть бути представлені у вигляді бінарних характеристик. Зазвичай, такого роду трансформація може викликати збільшення розмірності даних, але ця проблема може бути вирішена із застосуванням методів зниження розмірності.

Також можуть бути застосовані інші методи, але в такому випадку їх доведеться модифікувати для використання категоризованих даних.

5. Методи, що застосовуються на велико-розмірних даних.

У різних застосунках завжди виникає проблема великої кількості характеристик. Ця проблема означає додаткові характеристики, неправильне визначення відстані між об’єктами та викривлення результатів функціонування методів.

Так, наприклад, методи, засновані на взаємодії між об’єктами, працюватимуть набагато гірше з великою кількістю характеристик. Можливим виходом з даної ситуації можна вважати PCA, але тоді виникає інша проблема. При зменшенні кількості характеристик, втрачається інформація. Через невелику кількість прикладів аномалій, або появу нових типів аномалій, попередні «ненормальні» характеристики можуть бути втрачені.

Більш витонченим підходом є Sparse Cube Method (метод розсіяного куба). Він заснований на аналізі щільності розподілів проекцій даних, після чого

відбувається дискретизація сітки (дані формуються в розсіяний куб на цьому етапі) й застосовується еволюційний алгоритм, щоб знайти підходящий підпростір для зручнішого виявлення аномалій.

Методи, що оперують даними, які еволюціонують. Реальні дані дуже часто представлені у тимчасовому вигляді. Зазвичай це залежить від походження даних. Тимчасова характеристика може дискретною чи нескінченною, тому дані можуть бути представлені як послідовності чи часові серії.

1. Методи, які застосовуються на дискретних послідовностях.

Існує декілька шляхів визначити аномалію в даних, представлених у вигляді дискретних послідовностей: аналіз значень на певних позиціях; аналіз усієї послідовності.

Незалежно від цього, методи поділяються на три групи: засновані на послідовностях (Kernel Based Techniques, Window Based Techniques); засновані на суміжних підпослідовностях (Window Scoring Techniques, Segmentation Techniques); засновані на шаблонах (Substring Matching, Subsequence Matching, Permutation Matching).

2. Методи, які застосовуються на часових серіях.

У випадку, якщо дані залежать від часу, тоді постає проблема передбачити майбутні дані і проаналізувати нинішні тенденції.

Такі методи засновані на досконало розвиненому механізмі аналізу часових серій, включаючи наступні: фільтр Калмана; авторегресивне моделювання; виявлення незвичних форм перетворенням Хаара; численні статистичні методи.

Методи, що оперують мультиструктурованими даними. У деяких випадках дані представлені в більш складному вигляді, ніж числова таблиця «характеристика – значення». Тоді важливо зрозуміти, який саме об’єкт є аномалією, використавши відповідний метод аналізу.

1. Методи, які застосовуються на текстових даних.

У зв’язку з розвитком комунікацій, всесвітньої мережі та соціальних мереж, збільшився інтерес і до аналізу текстів в Інтернеті. Враховуючи

взаємодію між виявленням аномалій та пошуком текстів, можна виділити такі основні задачі: пошук «ненормальних» текстів – виявлення спаму; пошук нестандартних текстів – виявлення новизни.

Вирішуючи ці проблеми, головною з них є представлення тестів у вигляді метричних даних для використання вищеописаних методів. Використовуючи техніки обробки природньої мови, такі як LSA (latent semantic analysis), виникає можливість згрупувати текст, інтегруючи його зі стандартними методами виявлення аномалій. Виходячи з великої кількості текстів, частіше використовується машинне «навчання з учителем».

Використання tf-idf вимірювань (статистичний показник, що дозволяє оцінити важливість слів у контексті документа) дало можливість розробити алгоритм виявлення спаму, який базується на обчисленні відстаней між повідомленнями. Після цього він створює «нормальну» зону використовуючи тренувальний набір даних. Наприкінці межа зони визначає чи є лист спамом.

2. Методи, які застосовуються на даних графів.

Основною проблемою є той факт, що граф може бути великим і складним, або ж навпаки, складатися з великої кількості менших і простіших графів. Наступним завданням є витяг потрібних характеристик з вузлів, граней та підграфів, що дозволить використання методів, які оперують з метричними даними.

У випадку великих графів, аномалії можуть визначатись у вузлах, з’єднаннях, та підграфах. Методи, які аналізують вузлові аномалії, зазвичай витягують характеристики з даного вузла та його сусідів, але у випадку аномалій з’єднань, концепція аномалії стає дуже складною. Тоді грань розглядається в якості аномалії, якщо вона з’єднує вузли з кластерів вузлів різної щільності.

Найбільш популярними методами є такі, що засновані на: теорії випадкових графів; факторизації матриць; техніках спектрального аналізу.

Іншою проблемою може стати виявлення підграфів з девіантною поведінкою та визначення їх структури і витяг потрібних характеристик.

Для визначення аномальних підграфів зазвичай використовують принцип Minimal Description Length (мінімальної довжини опису). В одному з найбільш важливих застосувань сьогодні – даних соціальних мереж – використовуються такі сучасні методи: Байесові моделі; випадкові поля Маркова; модель Ізинга; EM-алгоритм; LOF.

1.2. Методи виявлення аномалій в стільникових мережах

Постійне зростання 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 і т.д.

Основними вимогами до поступово і неминуче зростаючих мобільних стільникових мереж є: висока пропускна спроможність; низькі витрати капіталу; низькі операційні витрати.

Ці аспекти продиктовані вимогами високошвидкісного доступу до послуг зв'язку за невеликі гроші. Тому технології радіодоступу і стільникові мережі постійно розвиваються і намагаються досягти більш ефективного використання радіоресурсів.

Проте, незважаючи на зростаючу кількість подібних загроз, більшість мобільних операторів реагують на загрози вже після їх реалізації, а не діючи на випередження. Тим не менш, останнім часом стає все більш популярною розробка систем виявлення аномалій, які б пропонували значну кількість переваг над існуючими рішеннями:

1. Проекція: аналіз на мережевому рівні не включає в себе додатковий моніторинг, обробку, зберігання чи комунікації з мобільними пристроями, які незважаючи на те, що мають досить велику обчислювальну потужність, все ж обмежені у плані акумуляторів та пропускної спроможності.

2. Дієвість: більшість користувачів не мають поняття про темпи зростання ризиків атак на мобільні пристрої, у той час як інші можуть неохоче використовувати традиційні антивірусні рішення, через їх властивість відбирати для свого функціонування вільні ресурси, що спричиняє зниження продуктивності, а також велику кількість помилкових спрацьовувань, що спричинені невідомою антивірусам поведінкою.

3. Швидкість і зручність розгортання: такі системи можуть бути легко модифіковані без потреби користувачу встановлювати виправлення та оновлення, в той час як існуючі системи повинні підтримувати значну кількість мобільних операційних систем та апаратних платформ.

4. Надійність: виявлення на мережевому рівні не є вразливим до атак, які дозволяють злоякісному програмному забезпеченню обійти захист на мобільних пристроях, але його можна обійти кваліфікованим зловмисникам, які приховуватимуть свою діяльність як звичайні користувачі.

5. Всеохоплення: системи виявлення аномалій забезпечує широкий огляд злоякісної активності всередині мережі стільникового оператора, включаючи атаки нульового дня, які звичайні технології виявлення швидше не розпізнають. Справді, в той час, як деякі атаки можуть бути виявлені вивчаючи поведінку користувача, інші стають більш видимими, коли певні характеристики розглядаються в сукупності від багатьох користувачів. Більше того, доступ оператора до даних, що підлягають оподаткуванню може полегшити виявлення атак, які впливають безпосередньо на платіжні рахунки.

Виявлення аномалій у стільникових мережах має свої особливості, через те, що дані з’являються постійно, а тому аналізувати їх потрібно у реальному часу. Також, нормальна поведінка даних змінюється в залежності від часу, а отже профіль нормальної поведінки потрібно оновлювати динамічно. Ключова ідея полягає в тому, щоб оновлювати профіль такими даними, які скоріше за все є нормальними, тобто мають відносно низький аномальний рівень.

Процес надходження даних в реальному часі зображений на рис. 1.2:

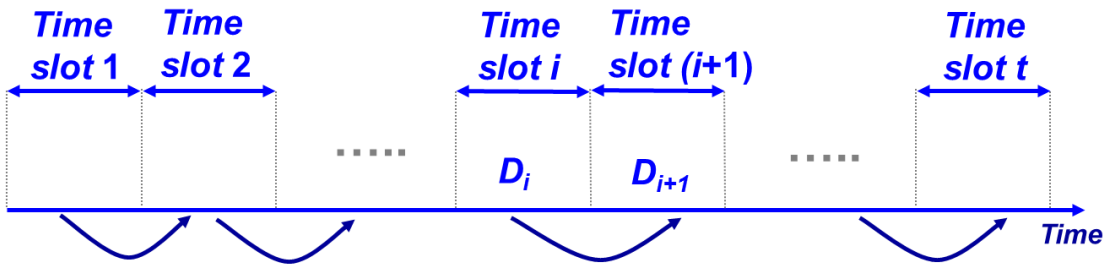

Рис. 1.2. Надходження даних в реальному часі

Тут *Time slot i* та набір даних D_i - модель нормальної поведінки. Алгоритм виявлення аномалій в *Time slot (i+1)* заснований на профілі, який було вирахувано в попередньому слоті.

Серед існуючих методів виявлення аномалій в системах стільникового зв'язку можна виділити наступні:

1. Використання графів.
2. Використання випадкових нейронних мереж.
3. Виявлення знань в базах даних.
4. Використання модифікацій алгоритму SVM.
5. Використання алгоритму SHESD.

Метод з використанням графів передбачає представлення даних, які циркулюють в стільникових мережах у вигляді графів.

При використанні методу графів основними аномаліями можуть бути: вставки, модифікації та видалення.

Вставки можуть визначені, якщо у графі з'являється неочікуваний вузол чи грань. Модифікації виявляються у випадку зміненої назви вузла чи грані. Видалення – у випадку відсутності вузла чи грані.

Метод виявлення аномалій на основі графів (ОГ) базується на використанні структури в даних, що представленні графом. Даний метод виявляє аномальні зразки структурних шаблонів в даних, які представлені сутностями, відносинами та діями.

ОГ викриває реляційну сторону проблеми, замість того, щоб працювати виключно із традиційним статистичним відхиленням індивідуальних

характеристик даних. Відхилення характеристик оцінюються у контексті відносин структурно подібних сутностей. До того ж більшість методів виявлення аномалій використовують підхід «навчання з учителем», які на початку потребують чітко розмежованих даних (тобто, що є аномалією, а що ні) для того, щоб тренувати систему.

ОГ є прикладом «самостійного навчання», що не потребує жодної базової інформації про значущі чи відомі аномалії. Даний метод шукає таку діяльність, яка намагається відповідати нормальній поведінці, але насправді суттєво відрізняється.

Для оцінки ОГ використовує два показники: мінімальну довжину опису та розмір. Для прикладу, використовуючи перший показник, можна виявити аномальну модифікацію (рис. 1.3), а другий – аномальну вставку (рис. 1.4).

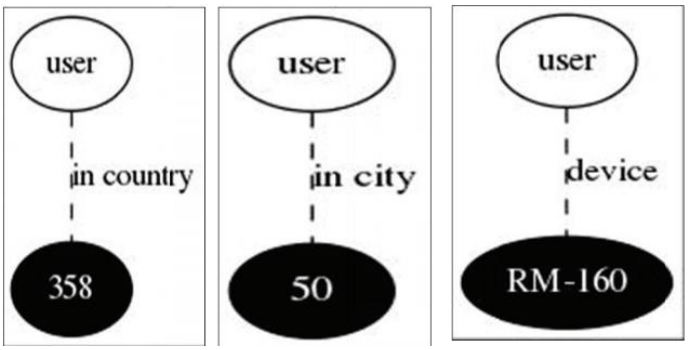

Рис. 1.3. Аномальна модифікація

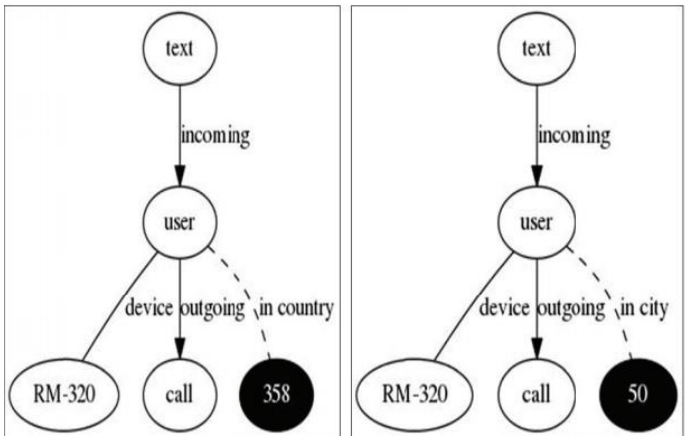

Рис. 1.4. Аномальна вставка

Метод виявлення аномалій на основі випадкових нейронних мереж (ВНМ) являє собою засновану на біології обчислювальну модель, у якій нейрони

обмінюються сигналами у формі стрибків одиниць амплітуди. У ВНМ позитивні та негативні сигнали представляють збудження та гальмування відповідно і накопичуються у нейронах. Позитивні сигнали скасовуються негативними і нейрон може вигоріти, якщо його потенціал буде зі знаком «+».

Сигнал може перейти від нейрона i до нейтрона j, як позитивний сигнал з ймовірністю p^+_{ij}; як негативний, з ймовірністю p^-_{ij} або може залишити мережу з ймовірністю d_i, де $\sum_j p^+_{ij} + p^-_{ij} + d_i = 1$.

ВНМ зображена на рис. 1.5, де k – потенціал нейрона, w – сигнали, які передаються нейронами, $q(t)$ – ймовірність того, що нейрон збуджений в даний момент часу.

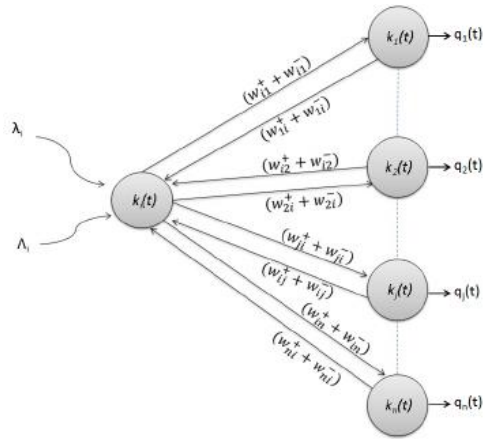

Рис. 1.5. Випадкова нейронна мережа

ВНМ відстежує активність об’єктів у мережі (користувачів, пристроїв) і вимірює набір виразних ознак, який описує різні характеристики тієї чи іншої поведінки. У кінці часового слоту (серії) даний метод надає ймовірность того, чи вхідні дані є прикладом нормальної поведінки, чи аномалії. Остаточне рішення приймається кінцевими вузлами нейронної мережі, в залежності від значення ймовірності.

ВНМ є гнучкою, та включає достатню кількість параметрів, щоб забезпечити компроміс між швидкістю виявлення, точністю і т.д. Даний підхід також є масштабованим, тобто може бути застосований для виявлення атак, як проти простих користувачів, так і проти мережі стільникового провайдера.

Змінюватиметься тільки набір відповідних ознак та певна модифікація параметрів алгоритму.

Метод виявлення аномалій на основі використання знань у базах даних (ЗБД). В загальному, такий метод використовується для звичайного аналізу даних, проте його можна застосувати також і до виявлення аномалій.

Процес добування необхідної інформації зображений на рис. 1.6.

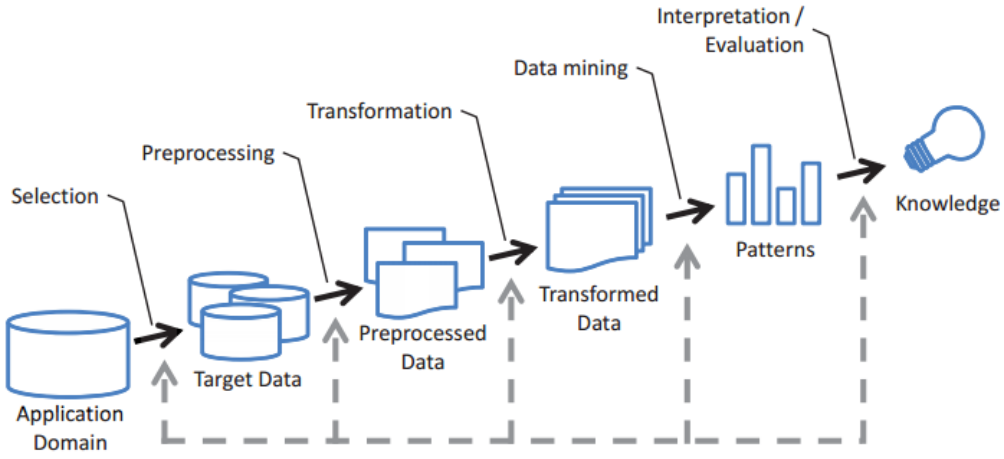

Рис. 1.6. Добування знань

ЗБД – послідовний процес, проте кожний з його етапів може включати ітерації, тому що в будь-який момент часу дані можуть змінитися, тому деякі кроки доведеться повторювати декілька разів. Детальніше розглянемо кожний з етапів ЗБД:

- Вибірка – починає вибирати дані з цільового набору. Цільовий набір – набір даних чи їх джерел, чия взаємодія та спільні шаблони поведінки повинні бути розглянуті. Часто набір наявних джерел даних чи записів є надлишковим чи, навпаки, неінформативним. Тому потрібно налаштовувати параметри алгоритму і обирати такі характеристики, які будуть корисні для аналізу.
- Підготовка – якість обраних даних не завжди є достатньою для подальшого аналізу (через пропущені чи втрачені записи, великий рівень шуму). Підготовка включає в себе злиття даних з деяких джерел і конвертацію у відповідні формати.

- Трансформація – підготовлені дані описані певною кількістю ознак. Це створює суттєві перешкоди у знаходженні їхньої ідеальної комбінації для вирішення початкової проблеми. Трансформація проектує вихідні дані у простір нижчого виміру – вбудований простір – і включає лінійні та нелінійні методи. Зменшений набір вбудованих ознак дозволяє провести візуальне дослідження та полегшує наступні етапи аналізу.
- Видобуток – є ключовим процесом ЗБД. Тут починає працювати обраний алгоритм для пошуку відповідних шаблонів.
- Інтерпретація/оцінювання – фінальний крок ЗБД, який включає в себе візуальний аналіз і набір параметрів продуктивності роботи алгоритму. Правильна інтерпретація є важливою, тому що вона дозволяє перевірити чи є правильними налаштування попередніх кроків ЗБД.

Один з можливих варіантів розгортання ЗБД зображений на рис. 1.7.

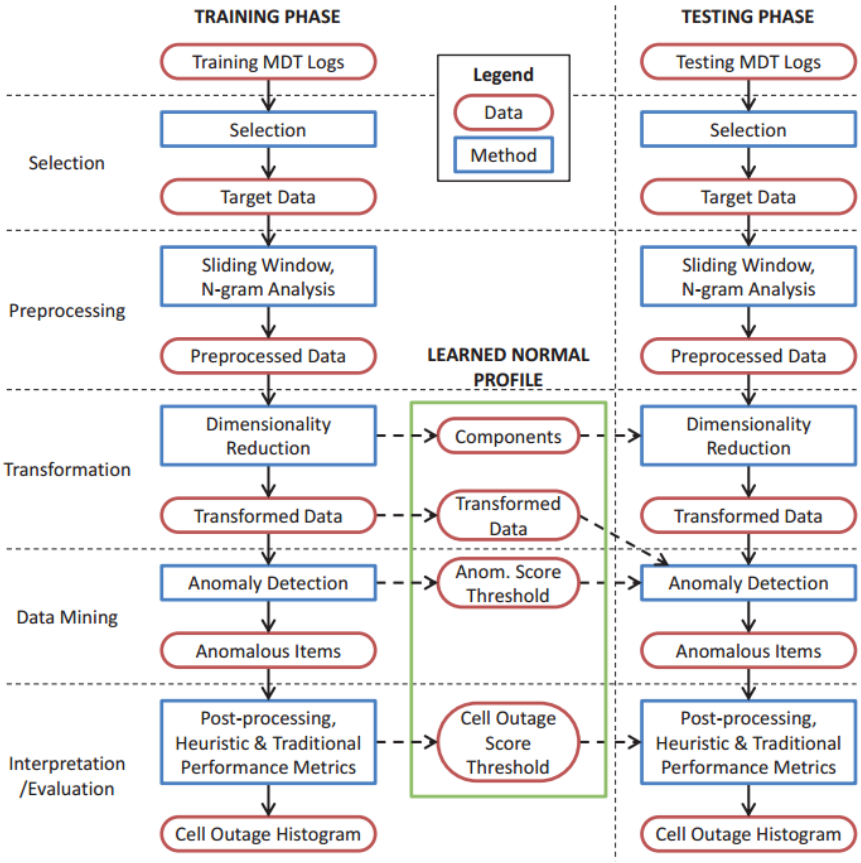

Рис. 1.7. Можливий варіант ЗБД

Модифікація алгоритму SVM (MSVM) знадобилася для більшої стійкості алгоритму проти різноманітних завад і зменшення впливу на результат його роботи надлишкових даних. Серед таких модифікацій виділяють:

- Робастний однокласовий SVM (P-SVM);
- η-однокласовий SVM (η-SVM).

Перший підхід використовується з метою підвищення робастності – стійкості системи до зовнішніх та внутрішніх хибних факторів. Ключовою ідеєю є мінімізації середньої квадратичної помилки для знаходження аномалій, використовуючи центр класу в якості усередненої інформації. Це дозволило зменшити кількість опорних векторів і підвищити продуктивність алгоритму.

Основна модифікація в P-SVM відбувається з урахуванням нечітких змінних. Тобто, використання звичайного алгоритму із нечіткими змінними при наявності аномалій може призвести до такої ситуації, що опорні вектори будуть вважатися аномаліями. Приклад такого випадку зображений на рис. 1.8.

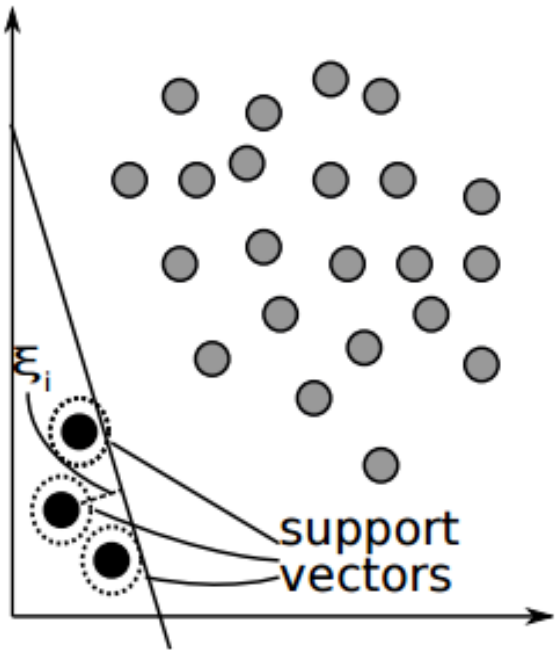

Рис. 1.8. Вплив аномалій на опорні вектори

Як бачимо, межа, що відокремлює нормальні дані від аномалій виділяє опорні вектори як аномалії, а це є неприпустимим.

У P-SVM нечіткі змінні є пропорційними до відстані до центроїда. Це дозволяє точкам, які знаходяться на великій відстані від центру мати великі нечіткі змінні. А так як вони є фіксованими, то їх можна виключити із задачі мінімізації.

З іншої сторони, в такому випадку межа зміщується у напрямку нормальних даних, але залишає значний простір для розміщення опорних векторів, щоб вони не вважалися аномаліями (рис. 1.9)

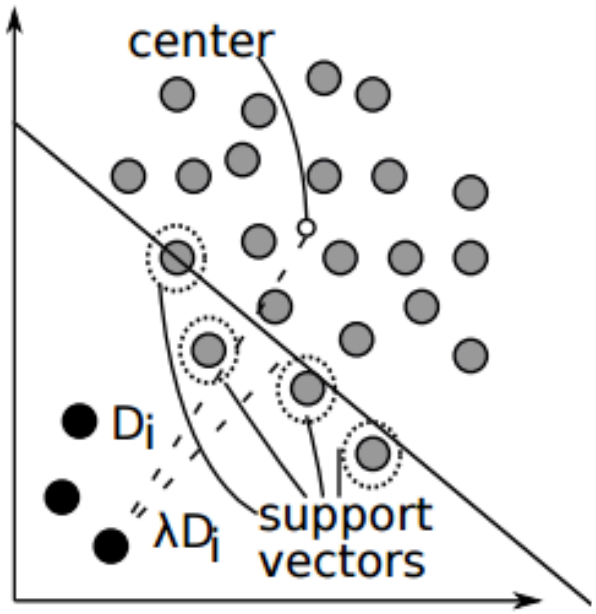

Рис. 1.9. Результат P-SVM

Використання η-SVM має на меті використання механізму явного придушення аномалій та передбачає введення додаткової змінної η, яка представляє оцінку, що дані є нормальними.

В аномальних даних змінна η приймає значення, близьке до 0, тому такі дані не використовуватимуться при проведенні межі, а отже на межу будуть впливати лише нормальні дані. Приклад η-SVM зображено на рис. 1.10.

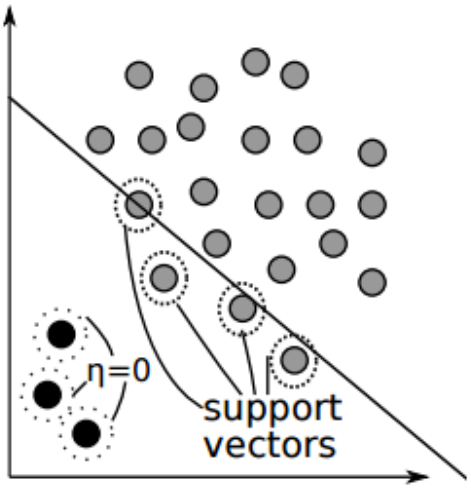

Рис. 1.10. Приклад η-SVM

Алгоритм SHESD, вдосконалений та впроваджений компанією Twitter, включає в себе наступні кроки: student-t розподіл; ESD-тест; узагальнений ESD; лінійна поліноміальна регресія; сезонний та тенденційний розподіл.

Алгоритм дозволяє виявити та ілюструвати такі типи аномалій, виділені компанією Twitter:

1. Глобальні/локальні.
2. Позитивні/негативні.

На рис. 1.11 зображений приклад роботи алгоритму і виділені наведені вище типи аномалій.

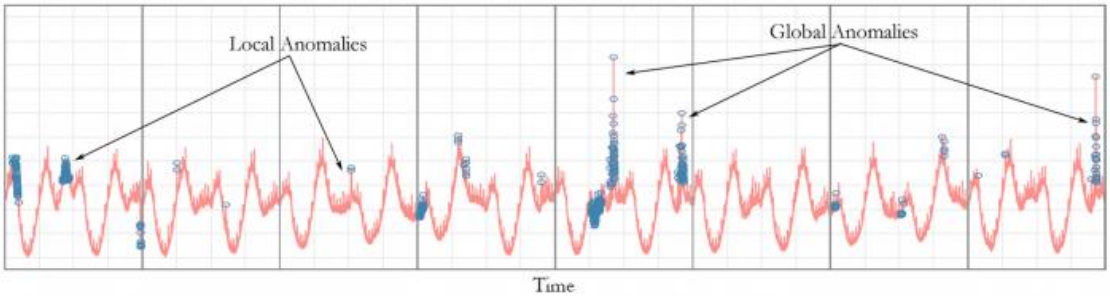

Рис. 1.11. Результат роботи алгоритму SHESD

Аналізуючи наведені дані, можна помітити, що глобальні аномалії значно відокремлюються від нормальних і тому їх доволі легко виявити. Локальні ж знаходяться на межі нормальних і тому вимагають більш уважного підходу.

1.3. Переваги та недоліки існуючих підходів до виявлення аномалій та шляхи їх удосконалення.

В загальному, усі методи виявлення аномалій можна виділити в декілька категорій і розглядати саме ці групи, тому що алгоритми в конкретних групах мають багато спільного, як переваги, так і недоліки. Розподіл методів виявлення аномалій по категоріях викладений у табл. 1.1-1.6.

Таблиця 1.1

Класифікаційні методи

Приклади методів	Переваги	Недоліки
1. Нейронні мережі; 2. Байесові мережі; 3. Метод опорних векторів; 4. Decision Tree.	• Методи, що базуються на класифікації, особливо багатокласовій, використовують потужні алгоритми, що дозволяє провести чітку межу між сутностями, які відносяться до різних класів. • Тестувальна фаза таких методів є швидкою, так як кожна тестова сутність повинна порівнюватися з попередньо обчисленою моделлю.	• Методи, які базуються на багатокласовій класифікації, вимагають наявності точних міток для різноманітних нормальних класів. • Наявність таких міток у кожного класу може стати недоліком, коли потрібно дати оцінку аномаліям у тестових випадках.

Таблиця 1.2

Методи «найближчого сусіда»

Приклади методів	Переваги	Недоліки
1	2	3
1. Використання відстані до k -го «найближчого сусіда». 2. Використання відносної щільності.	• Ключовою перевагою таких методів є «навчання без вчителя» і тому не мають жодного впливу на розподіл даних. «Напів-навчені» методи мають вищу продуктивність, ніж «ненавчені» відносно пропущених аномалій, так як ймовірність того, що аномалії створять близьке сусідство у тренувальному наборі даних дуже низька.	• У випадку, якщо нормальні дані не мають достатньої кількості близьких «сусідів» або, навпаки, аномалії їх мають, для «ненавчених» методів існує загроза неправильного визначення типу даних (нормальні чи аномальні), що стає причиною пропущених аномалій.

Продовження табл. 1.2

1	2	3
1. Використання відстані до k-го «найближчого сусіда». 2. Використання відносної щільності.	• Адаптування таких методів до різних типів даних є прямолінійним і головним чином потребує визначення підходящої • техніки вимірювання відстані для вихідних даних.	• Якщо нормальним даним у тестовому наборі не відповідає достатня кількість схожих нормальних даних у тренувальному, то для • «напів-навчених» методів існує загроза частих хибно-позитивних спрацювань. • Вимогливість до обчислювальних ресурсів. • Продуктивність таких методів залежить від вимірювання відстані, що визначена між нормальними та аномальними даними, особливо коли дані є графами чи послідовностями.

Таблиця 1.3

Кластеризаційні методи

Приклади методів	Переваги	Недоліки
Кластеризація	• Може діяти без попереднього навчання. • Методи можуть застосовані для складних даних просто задіявши кластеризаційний алгоритм, який працює з відповідним типами даних. • Тестувальна фаза для таких методів є швидкою, оскільки кількість кластерів, з якими повинні порівнюватися тестові дані є малою.	• Продуктивність таких методів залежить від ефективності кластеризаційного алгоритму у створенні структури кластеру нормальних даних. • Більшість методів виявляють аномалії, як побічний продукт кластеризаці, а відтак є неоптимізованими. • Деякі кластеризаційні алгоритми відносять абсолютно всі дані до певного кластеру, що стає причиною можливого віднесення аномалії до великого кластеру. • Деякі алгоритми є неефективними, коли аномалії формують власні суттєві кластери.

Таблиця 1.4

Статистичні методи

Приклади методів	Переваги	Недоліки
1. Параметричні методи 2. Непараметричні методи	• Якщо припущення щодо базового розподілу даних вірні, статистичні методи забезпечують виправдане рішення для виявлення аномалій. • Оцінка аномалії, надана такими методами, пов'язана з довірчим інтервалом, що може бути використаний в якості додаткової інформації при прийнятті рішення щодо будь-яких тестових даних. • Якщо оцінка розподілу є стійкою до аномалій в даних, такі методи можуть працювати «без навчання» без необхідності маркування даних.	• Ключовим недоліком таких методів є припущення, що дані формуються з певного розподілу. Таке припущення не завжди є правдивим, особливо для велико-просторових наборів реальних даних. • Навіть якщо статистичне допущення є виправданим, існує кілька гіпотетичних тестових статистик, а обрати кращу з них не завжди легко. Зокрема, побудова гіпотетичних тестів для складних розподілів, які мають відповідати велико-просторовим даним є нетривіальною задачею. • Деякі алгоритми можуть бути легкими у впровадженні, проте вони не мають змоги виявити взаємодію між різними характеристиками. Деякі аномалії можуть мати такі значення характеристик, які окремо зустрічаються дуже часто, а в сукупності – рідко, тому дані методи не виявлять таких аномалій.

Таблиця 1.5

Інформаційно-теоретичні методи

Приклади методів	Переваги	Недоліки
1. Складність Коломогорова. 2. Ентропія.	• Можуть діяти «без навчання». • Не роблять ніяких припущень щодо базового статистичного розподілу даних.	• Продуктивність залежить від обраного методу. Інколи метод може виявити аномалії лише за високої концентрації їх присутності в даних. • При застосуванні таких методів до послідовностей та просторових даних може виникнути проблема розміру структур. • Проблемне асоціювання оцінки аномалій з тестовими наборами даних.

Таблиця 1.6

Спектрально-аналітичні методи

Приклади методів	Переваги	Недоліки
Засновані на алгоритмі PCA	• Автоматично знижують розмірність даних, тому можуть використовуватись для велико-просторових наборів даних. • Можуть діяти «без навчання»	• Корисні лише у випадку, якщо нормальні та аномальні дані легко відділити у нижчому просторовому вкладенні даних. • Вимогливість до обчислювальних ресурсів.

Тим не менш, недоліками, що об’єднують усі наведені вище методи є наступні:

- В 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.

1.4. Висновки до першого розділу

Для вирішення даних 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.

Розділ 2. РОЗРОБКА МОДЕЛІ СИСТЕМИ ВИЯВЛЕННЯ АНОМАЛІЙ

2.1. Застосування систем виявлення аномалій в мережах операторів

Однією з галузей, де виявлення аномалій є справді актуальною проблемою є телекомунікаційна сфера, а саме – мережі операторів стільникового зв'язку.

Згідно зі звітом Українського державного центру радіочастот тільки у період з лютого по березень 2016 було зафіксовано близько 5000 порушень маршрутизації міжнародного трафіку (рефайлу) у вітчизняних операторів, що фактично є основним джерелом прибутку. Сигнатурні методи виявлення загроз не дозволяють вчасно виявити витік трафіку, оскільки це не є прямою атакою на мережу оператора. Факти такої крадіжки трафіку можна виявити вже постфактум, що приносить значні збитки компаніям.

Іншими видами шахрайства в мобільних мережах є:

- Клонування SIM-карт з метою організації міжміських переговорних пунктів, одночасного використання необмежених тарифних планів, протиправних дій під виглядом справжнього власника SIM-картки, у випадку її клонування без відома власника.
- Незаконне завершення з'єднання.
- Створення «чужих» SIM-центрів (в результаті «свій» SIM-центр не отримує інформацію про відправлене повідомлення і воно не тарифікується).
- Роумінгове шахрайство (генерація трафіку на дорогі направлення з розрахунком на затримку своєчасної тарифікації роумера в білінгу домашньої мережі).
- Шахрайство із службами замовлення білетів, сплати за послуги, переказ грошей з банківських рахунків з допомогою SMS-транзакцій тощо.

Види внутрішнього шахрайства – співробітників стільникового оператора:

- Умисна активація неоплачених сервісів.
- Шахрайство при використанні передплачених послуг.
- Помилкова тарифікація.

Одним із варіантів запобігання цьому – дотримання рекомендацій стандартів QoS у сфері телекомунікацій, які безпосередньо впливають на безпеку даних клієнтів.

Насамперед, потрібно зрозуміти визначити схему взаємодії оператора і користувача у розумінні QoS. Приклад такої схеми зображений на рис. 2.1.

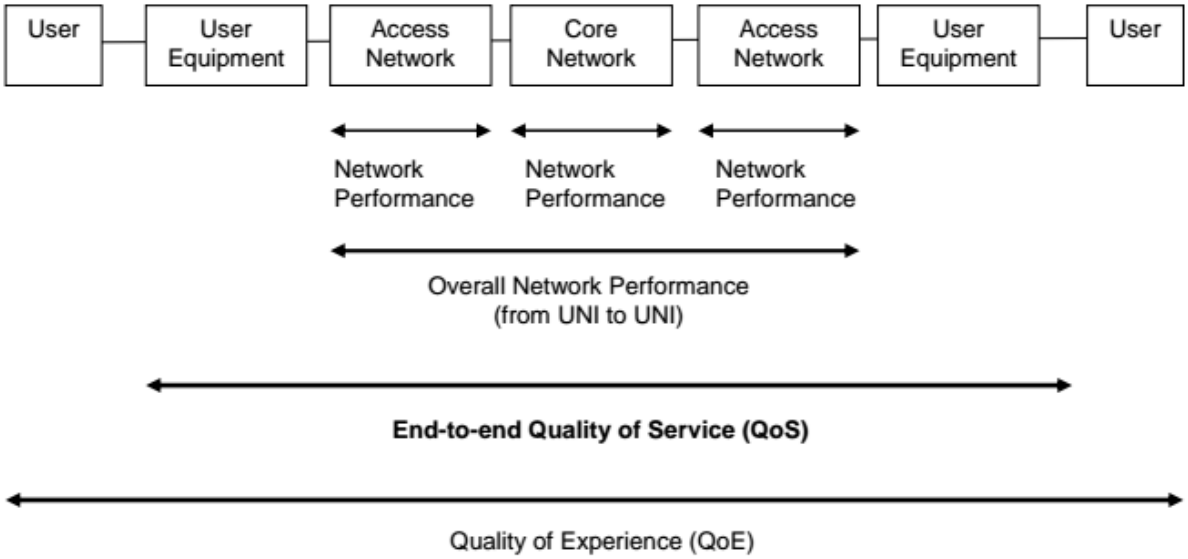

Рис. 2.1. Схема QoS

Тобто, QoS для оператора визначається наступними складовими:

1. Продуктивність мережі – оцінюється як частина мережі або підмережі. Зазвичай, продуктивність мережі в технічному плані залежить від оцінки технічних параметрів, які описують продуктивність конкретної частини мережі. Прикладами таких параметрів є коефіцієнт бітових помилок, потужність відправлення і прийому, затримка передачі тощо.

2. Загальна продуктивність мережі – якщо декілька частин мережі розглядаються як одна цілісна частина мережі («чорний ящик»), тоді потрібно оцінювати загальну продуктивність. Таким чином може бути оцінений зв’язок між двома мережевими інтерфейсами користувача.

3. End-to-End QoS – оцінка загального ланцюга передачі з точки зору користувача проводиться об’єктивним чином. Тобто, мається на увазі, що

потрібно оцінити найбільш повний відрізок передачі даних, не беручи до уваги власне користувача.

4. Якість досвіду (QoE) – включення користувача до загальної якості у телекомунікаціях розширює більш об’єктивну QoS до надзвичайно суб’єктивної QoE, яка відрізняється від користувача до користувача, оскільки на неї впливає особистий досвід та очікування кожного конкретного користувача.

Також, QoS може розглядатися з так званої «Four-viewpoint» моделі, що зображена на рис. 2.2.

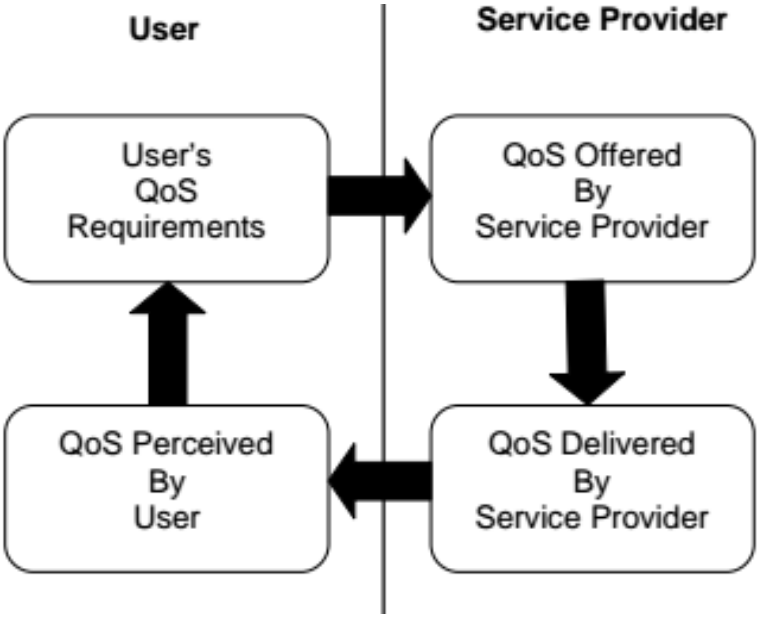

Рис. 2.2. Модель «Four-viewpoint»

Отже, QoS може розглядатися з чотирьох перспектив:

- QoS, яку потребує користувач – такі вимоги описані в загальному, не технічному плані. Вони є відображанням очікувань клієнта з кінцевої точки зору. Тобто клієнт виражає очікування від послуг, які надаються за допомогою мережі. Він не повинен думати про технічні недоліки чи обмеження, які можуть виникати. Для оператора такі вимоги є базовим рівнем, від якого він повинен відштовхуватися, щоб задовольнити потреби клієнтів.
- QoS, яку пропонує оператор – бажаний рівень для постачальника зв’язку. Може бути досягнута кількома шляхами: полегшення доставки

інформації до користувача; запросити експертів для встановлення угод рівня послуг чи спрощення цілей технічного планування.

- QoS, яку надає оператор – реальний рівень QoS. Порівняння пропонованої і реальної якості послуг дозволяє оцінити можливості оператора і визначити слабкі ланки.
- QoS, яку отримує користувач – відображує його суб'єктивний погляд на конкретну ситуацію. Одним із засобів її оцінки є проведення анкетування серед користувачів чи інший зворотний зв'язок. На цьому рівні користувач порівнює індивідуальний досвід із більш технічно-орієнтованою якістю послуг, що надаються. Загалом, так визначається персональна QoE кожного клієнта.

Сам процес надання послуг з боку оператора включає наступні аспекти:

1. Наявність мережі – вірогідність того, що послуги пропонуються користувачу через мережеву інфраструктуру.
2. Доступність мережі – вірогідність того, що користувач проходить успішну реєстрацію в мережі, яка надає послугу. Мережа може бути доступною лише за умови наявності.
3. Доступність послуги – вірогідність того, що користувач отримає доступ до послуги, якою хоче скористатися.
4. Цілісність послуги – описується якістю протягом часу, якого послуга надається і включає в себе такі елементи: якість мови, відео чи кількістю бітових помилок у файлі, що передається. Цілісність послуги може бути визначена, лише за умови успішного доступу.
5. Стійкість послуги – описується можливістю завершити послугу (згідно чи супроти бажання користувача). Прикладами можуть бути будь які неочікувані завершення дзвінків чи переривання передачі даних.

Для підтримання даних аспектів у належному стані, оператору потрібно звертати увагу на наступні параметри:

- Швидкість – швидкість, з якою послуга надається користувачеві.

- Точність – достовірність та повнота реалізації комунікаційної функції враховуючи еталонний рівень.
- Доступність – ймовірність, з якою користувач може отримати доступ до важливих компонент послуги у зручний для нього момент.
- Надійність – ймовірність того, що послуга буде надаватися без будь-яких обмежень швидкості, точності чи доступності в заданий період.
- Безпека – конфіденційність, з якою надається послуга, тобто неможливість отримати доступ до інформації третім особам без дозволу конкретного користувача.
- Простота – легкість надання певної послуги.
- Гнучкість – варіанти, яких потребує користувач і які пропонує оператор для задоволення спеціальних вимог.

Проте лише дотримання рекомендацій стандартів QoS недостатньо. Саме тому, проаналізувавши недоліки методів, розглянутих у попередньому розділі, було прийнято рішення розробити модель гібридної системи виявлення аномалій, що б поєднувала в собі як виявлення відомих атак, так і нових, ще не внесених у бази даних.

Вибір методу реалізації такої системи залежить від типу даних, які будуть відстежуватися і стануть вхідними для алгоритму виявлення, не беручи до уваги власне тип наявного методу виявлення.

Базуючись на типі даних, що відстежуються, виявлення аномалій може бути класифіковане як:

1. Host-based (моніторинг вузла мережі).
2. Network-based (моніторинг власне мережі).

Network-based моніторинг зазвичай виконується з даними типу мережевого трафіку, пакетів даних і т.д. У свою чергу, host-based моніторинг виконується на таких даних мобільного пристрою, як:

- події системного рівня;
- активність процесора;

- споживання пам’яті;
- активність мережевого та файлового вводу/виводу;
- події рівня операційної системи;
- події рівня застосунків.

Обидва підходи мають свої переваги та недоліки. Їх порівняння наведене у табл. 2.1.

Таблиця 2.1

Порівняння Network-based та Host-based моніторингу

Метод	Переваги	Недоліки
Network-based	• Зменшення навантаження на обчислювальні потужності конкретного телефону; • Виявлення зовнішніх вторгнень.	• Відсутність доступу до даних на мобільному пристрої, які можуть бути корисними для вторгнень; • Комунікаційне середовище є надзвичайно фрагментованим, так як мобільний пристрій може бути під’єднаний до різних джерел через різні інтерфейси у той самий час; • Неможливість виявлення вторгнень конкретно на пристрої (злоякісне ПЗ, віруси тощо) • Збирання всіх даних для всіх мереж чи комунікаційних інтерфейсів, з якими взаємодіє мобільний пристрій може бути надзвичайно складним.
Host-based	• Доступ до інформації на мобільному пристрої, яка може бути корисною для виявлення вторгнень, так як дані зібрані з конкретного пристрою точно відображують його поведінку. • Більш надійний результат збору даних та спостереження.	• Складність впровадження через обмежені обчислювальні потужності мобільних пристроїв; • Складність забезпечення безпеки даних, які збираються напряму з мобільного пристрою.

2.2. Структурна схема системи виявлення аномалій

Згідно з нещодавніми дослідженнями у сфері стільникового зв'язку, у мережах 3 найбільших операторів досі знаходиться суттєва частка мобільних апаратів, які або зовсім не мають операційних систем нового покоління (Android, iOS, Windows Phone), або ці системи вже є застарілими і не мають перспектив розвитку (Symbian, Bada OS). Проблема таких апаратів полягає в тому, що їхньої обчислювальної потужності не вистачить для участі у Host-based системі. Тому логічним рішенням є розробка системи з використанням Network-based підходу.

Загальна структура системи зображена на рис. 2.3.

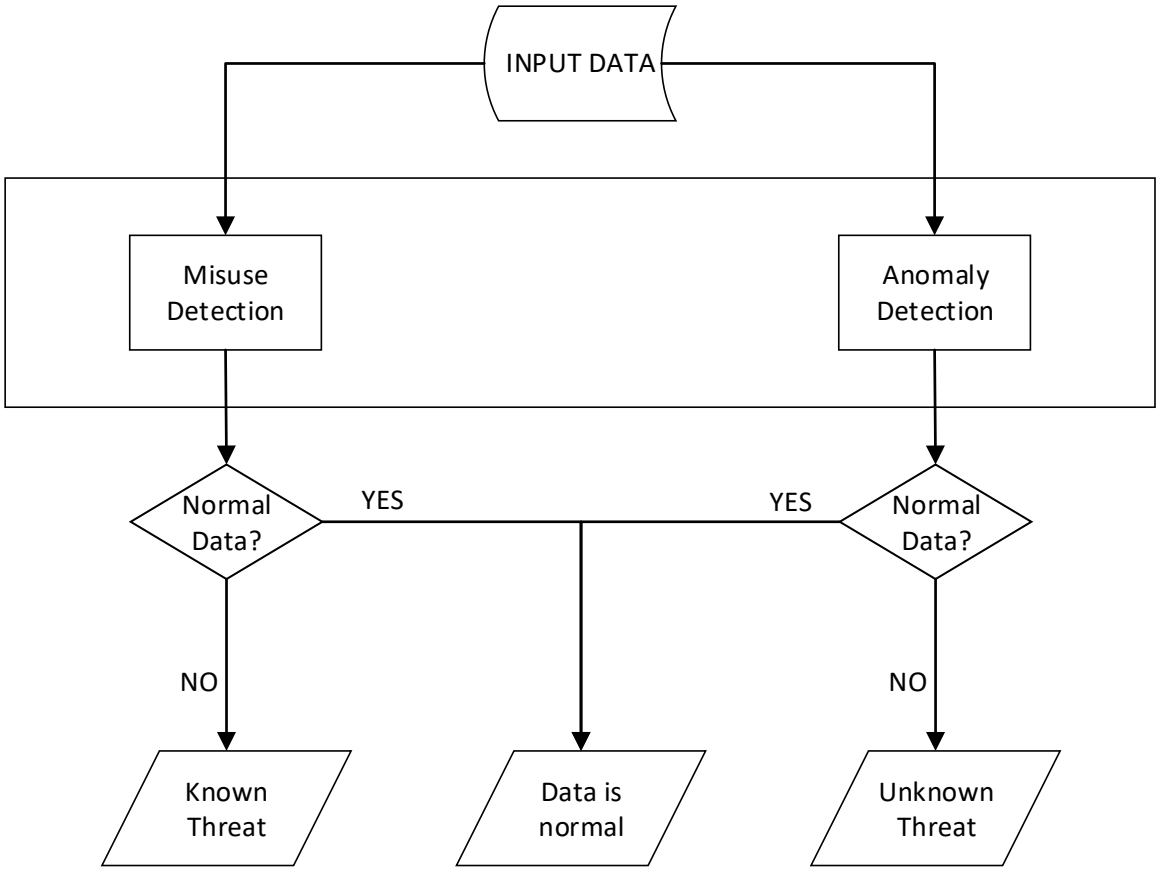

Рис. 2.3. Загальна структура гібридної системи

На вхід до системи подається трафік, який паралельно аналізується двома модулями – Misuse Detection (сигнатурне виявлення) і Anomaly Detection (виявлення аномалій). У випадку, якщо загроза виявляється тільки модулем сигнатурного виявлення – на вихід системи подається сам факт наявності загрози і її опис згідно базам даних.

Якщо загроза виявляється тільки модулем виявлення аномалій – на виході буде лише сам факт загрози.

Якщо загрози не виявлено – дані відмічається як нормальні.

Для зручності логіка системи наведена у таблиці 2.2, де 1 – означає спрацювання модуля, а 0 – відповідно, ні.

Таблиця 2.2

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*

У якості модуля Misuse Detection доцільно використати open-source утиліту Snort, яка працює як на операційних системах Windows, так 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.

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 файрволу, порівнює їх з сигнатурними правилами і ставить мітку «Відкинуто» у випадку якщо пакети відповідають правилу.

Схема взаємодії Snort із файрволом зображена на рис. 2.4.

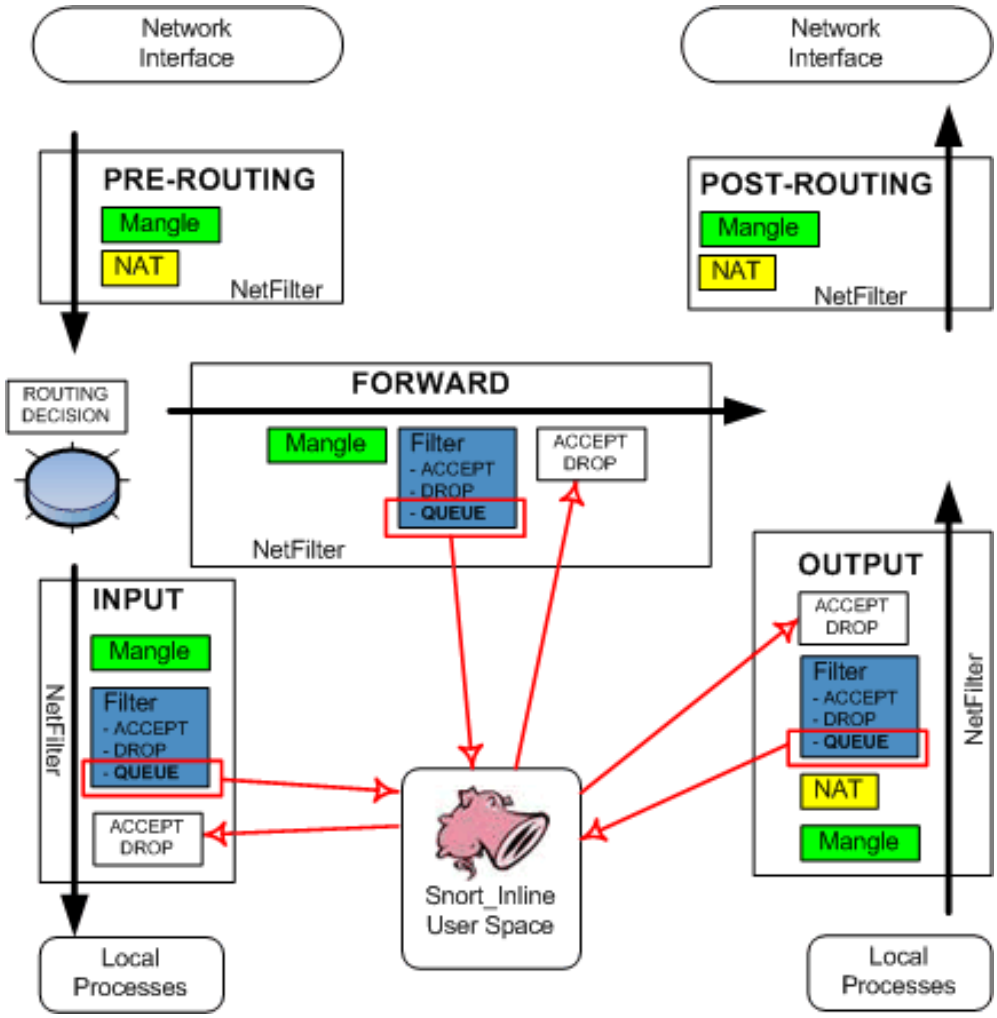

Рис. 2.4. Схема взаємодії Snort і файрвола

IPS mode у свою чергу має ще 2 режими роботи:

- Drop mode (режим відкидання) – пакет відкидається, якщо відповідає сигнатурному правилу. У режимі наявні три опції використання: drop (відкинути пакет, відправити команду reset до вузла мережі, записати подію у журнал); sdrop (відкинути пакет без посилання команди reset); ignore (відкинути пакет, відправити команду reset, не записувати подію в журнал).
- Replace mode (режим заміни) – пакет модифікується, якщо відповідає сигнатурному правилу.

Налаштування Snort відбуваються у конфігураційному файлі snort.conf (рис. 2.5).

```
# Setup the network addresses you are protecting
ipvar HOME_NET any

# Set up the external network addresses. Leave as "any" in most situations
ipvar EXTERNAL_NET any

# List of DNS servers on your network
ipvar DNS_SERVERS $HOME_NET

# List of SMTP servers on your network
ipvar SMTP_SERVERS $HOME_NET

# List of web servers on your network
ipvar HTTP_SERVERS $HOME_NET

# List of sql servers on your network
ipvar SQL_SERVERS $HOME_NET

# List of telnet servers on your network
ipvar TELNET_SERVERS $HOME_NET

# List of ssh servers on your network
ipvar SSH_SERVERS $HOME_NET

# List of ftp servers on your network
ipvar FTP_SERVERS $HOME_NET

# List of sip servers on your network
ipvar SIP_SERVERS $HOME_NET
```

Рис. 2.5. Вікно налаштувань Snort

Для зручності роботи окрім Snort у систему потрібно встановити систему відображення попереджень у графічному вигляді. Прикладом такої системи може бути Basic Analysis and Security Engine (BASE), яка зображена на рис. 2.6, проте це просто веб-сторінка, підключена до бази даних, тому за необхідності можна створити власний інтерфейс.

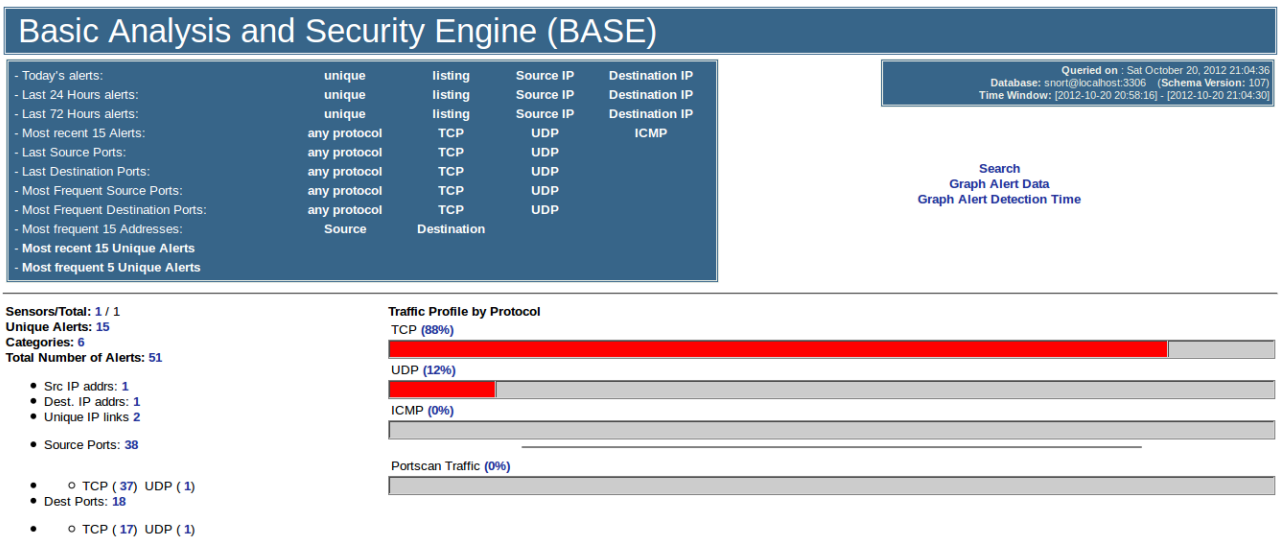

Рис. 2.6. Сторінка BASE

Провівши аналіз методів, розглянутих в попередньому розділі, була створена таблиця (див. табл. 2.3) їх порівняння за такими критеріями:

- 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 2.3

Порівняння 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	–	+/–	+/–	+/–	+

Згідно з проведеним аналізом, кращим методом для даної системи виявлення аномалій є метод Decision Tree.

Decision Tree – метод, що широко використовується в аналізі даних. Decision Tree є потужним інструментом для класифікації та передбачення. Вони

відображають правила, які зрозумілі користувачам і можуть бути використані в таких системах знань, бази даних.

Метою є створення моделі, яка передбачує значення цільової характеристики базуючись на декількох вхідних характеристиках. На рис. 2.7 зображена загальна схема Decision Tree.

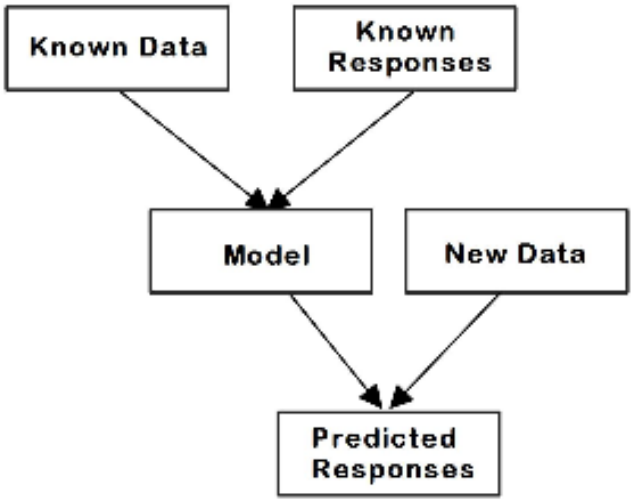

Рис. 2.7. Загальна схема Decision Tree

Схема відображає багатокроковий процес, на кожному етапі якого виконується бінарне рішення. Дерево складається з вузлів, які поділяються на внутрішні та термінальні та гілок.

Внутрішні вузли розділяються на двоє 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, у якості 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 в якості головного вузла.
- 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. Обчислити ентропію для вузла A (рис. 2.8):

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

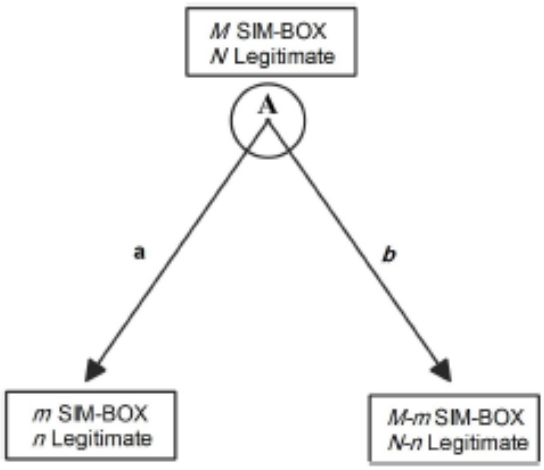

Рис. 2.8. Приклад моделі

2. Після того, як набір даних розподіляється на дві гілки за різними характеристиками, обрахувати ентропію для кожної гілки:

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

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

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

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

3. Ентропія для кожної 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 ж 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 ж поки не будуть використані всі характеристики.

Алгоритм Decision Tree:

1. Ініціалізація дерева з параметрами (приклади даних, цільова характеристика, список характеристик). Приклади даних є тренувальними

елементами. Цільова характеристика є тією, величину якої потрібно прогнозувати, використовуючи дерево. Список характеристик – перелік тих атрибутів, які можуть також бути класифіковані деревом.

2. Створення головного вузла для дерева.

2.1. Якщо всі приклади є аномаліями, повернути результат у вигляді дерева з єдиним вузлом із міткою – «Аномалія».

2.2. Якщо всі приклади є нормальними, повернути результат - дерево з єдиним вузлом із міткою – «Нормально».

2.3. Якщо список характеристик є порожнім, повернути дерево з єдиним вузлом із міткою, що відповідає найбільш поширеному значенню цільової характеристики у прикладах.

2.4. В іншому випадку:

2.4.1. Обираємо A із списку характеристик з умовою, що A дає найбільше інформаційне підсилення.

2.4.2. Вирішальна характеристика для головного вузла – A.

2.4.3. Для кожного можливого значення v_i для A:

2.4.3.1. Додати нову гілку дерева нижче головного вузла, відповідно до тесту A $= v_i$.

2.4.3.2. Підмножина $samples_{v_i}$ є такою, що містить значення v_i для A.

2.4.3.3. Якщо приклади v_i є порожніми:

2.4.3.3.1. Нижче даної нової гілки додати термінальний вузол з міткою, що відповідає найбільш поширеному значенню цільової характеристики у прикладах.

2.4.3.4. В іншому випадку нижче даної нової гілки додати «піддерево» з параметрами ($samples_{v_i}$, цільова характеристика, список характеристик без A).

3. Кінець та повернення головного вузла.

2.3. Концепція Big Data

Big Data – масивні набори даних, що мають велику, різноманітну і більш складну структуру, що викликає труднощі зберігання, аналізу та візуалізації для подальших процесів чи результатів.

Проте розвиток даного напрямку зумовив розширення його визначення:

Big Data – сукупність інструментів, методів, моделей, принципів для обробки, зберігання, аналізу, та виділення корисної інформації із неструктурованих масивів даних

Головними проблемами, для вирішення яких використовують методи Big Data (рис. 2.9):

- Volume (ємність).
- Variety (різноманітність).
- Velocity (швидкість).

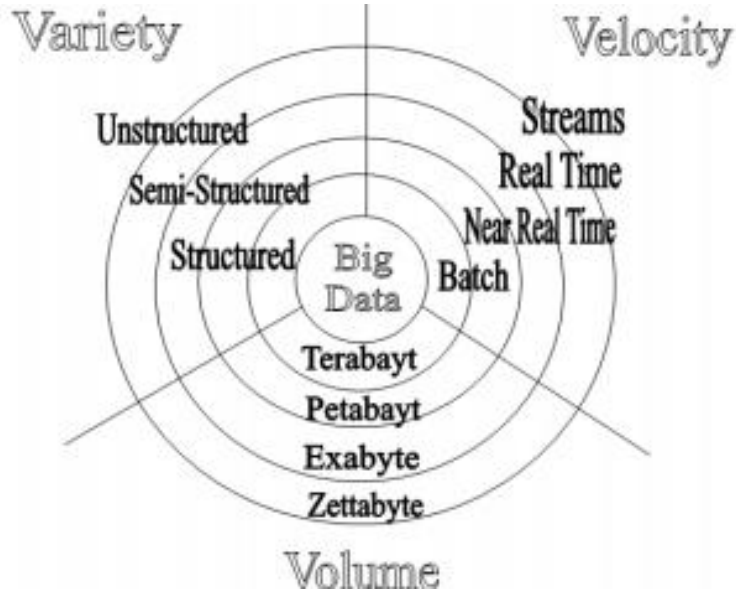

Рис. 2.9. Проблеми великих даних

Ємність означає величину даних. Дані починаються вважатися великими з декількох терабайтів і петабайтів. Визначення ємності великих даних є відносним і відрізняються за такими характеристиками як час та тип даних. Визначення великих даних може змінюватися з часом, оскільки місткість сховищ даних зростатиме, дозволяючи зберігати навіть більші об’єми інформації.

До того ж, тип даних також визначає поняття «великих». Два набори даних однакового розміру, можуть потребувати різних технологій управління даними, залежно від їх типу, наприклад, табличні і відеодані. Також, визначення великих даних залежить від галузі, що розглядається. Тому встановлення певної межі, після досягнення якої дані стають Big Data, є непрактичним.

Різноманітність означає структурну гетерогенність в масиві даних. Технологічний прогрес дозволяє компаніям використовувати різні типи структурованих, напів-структурованих та неструктурованих даних.

Структуровані дані, які складають лише 5% від усіх існуючих, представлені у вигляді табличних даних чи реляційних базах даних. Текст, зображення, відео є прикладами неструктурованими даними, яким бракує структурної організованості, що заважає проводити машинний аналіз.

Займаючи позицію між повністю структурованими та неструктурованими даними, формат напів-структурованих даних не відповідає суворим стандартам. Наприклад, Extensible Markup Language (XML), текстова мова для обміну даними в Інтернеті є типовим прикладом напів-структурованих даних. XML-документи містять визначені користувачами мітки даних, що робить їх можливими для читання машинами.

Високий рівень різноманітності, характеристики, що визначає великі дані, не є новим. Організації накопичували неструктуровані дані з внутрішніх джерел (дані сенсорів) та зовнішніх (соціальні медіа). Тим не менш, поява нових технологій управління та аналізу даних, дозволяє організаціям максимально ефективно використовувати їх у бізнес-процесах. Наприклад, технології розпізнавання обличчя дозволяють продавцям покращити збір інформації про трафік магазину, розподіл віку чи статі їхніх клієнтів та шаблони їх пересування у приміщенні магазину. Така безцінна інформація є ефективною у прийнятті рішень про просування продуктів, розміщення та стаффінгу.

Clickstream-дані надають багато інформації про поведінку користувача та його зразки його серфінгу в Інтернеті онлайн-магазинам. Таким чином можна отримати дані про час та послідовність перегляду сторінок користувачем.

Використовуючи аналітику великих даних, навіть невеликі та середні підприємства можуть добувати великі об’єми напів-структурованих даних для того, щоб покращити дизайн веб-сайтів та впровадити ефективні перехресні продажі та персональні рекомендації для кожного користувача.

Швидкість визначає темп, з яким дані генеруються і повинні бути проаналізовані і застосовані у процесі. Поширення цифрових пристроїв, таких як смартфони та сенсорні датчики призвело до неймовірної швидкості створення даних, що зумовлює необхідність аналізу у реальному часі та планування, заснованого на фактичних даних.

Дані, добуті з мобільних пристроїв та мобільних застосунків, продукують величезні масиви інформації, яка може бути використана для генерування персоналізованих пропозицій для кожного клієнта у реальному часі. Такі дані забезпечують наступну інформацію про клієнтів: геолокацію, демографічні дані та інформацію про минулі покупки, яка може бути проаналізована у реальному часі для створення фактичної цінності клієнта.

Враховуючи постійно зростаючу популярність смартфонів організації скоро зіштовхнуться з потребою аналізу в реальному часі сотень тисяч потокових джерел даних. Традиційні системи управління даними не здатні миттєво впоратися з великими масивами даних. У такому випадку доцільним є використання технологій Big Data.

На додачу до основних «трьох V», Big Data розглядається у контексті інших понять:

- Veracity (точність). Деяким джерелам даних властива проблема правдивості. Наприклад, настрій клієнтів у соціальних медіа є неявними по своїй природі, оскільки на нього впливає людське судження. Тим не менш, вони містять цінну інформацію, тому потреба обробляти неточну та неявну інформацію є ще одним викликом для методів Big Data.
- Variability (and complexity) – мінливість (і складність). Мінливість представляє собою зміну швидкості потоків даних. Часто швидкість великих

даних є непостійною і має свої вершини і западини. Складність представляє собою той факт, що великі дані генеруються через міріади різних джерел. Тут постає критичне завдання: потреба з'єднати, знайти відповідність, очистити і трансформувати дані, що отримуються від різних джерел.

- Value (цінність). Компанія Oracle запропонувала даний аспект Big Data, посилаючись на той факт, що величезні оригінальні масиви даних мають дуже низьку вартість. Але, з іншої сторони, аналізуючи такі дані, можна отримати солідну вигоду.

Процес роботи з великими даними зображений на рис. 2.10 і складається з таких етапів:

1. Управління даними:
   1.1. Накопичування та запис.
   1.2. Видобуток, очистка та тлумачення.
   1.3. Інтеграція, об'єднання та представлення.
2. Аналітика даних:
   2.1. Моделювання та аналіз.
   2.2. Тлумачення.

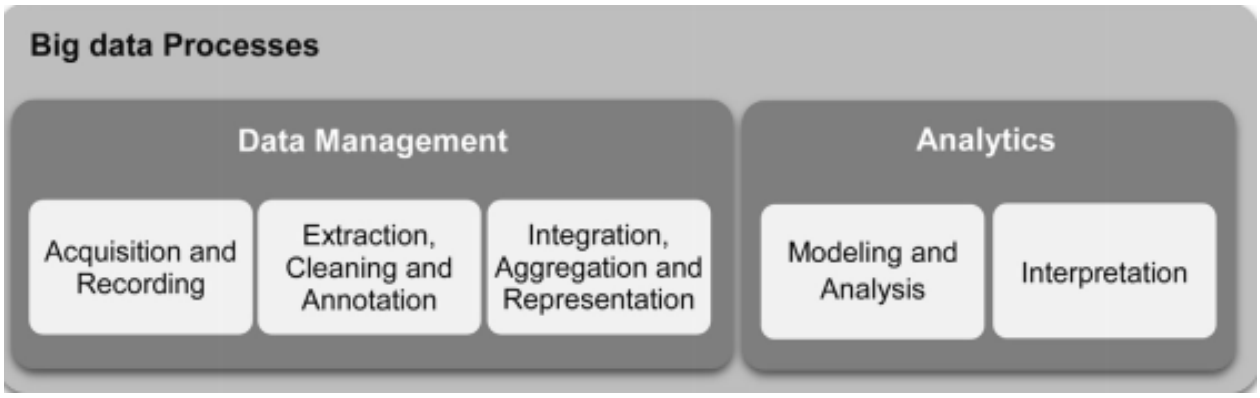

Рис. 2.10. Процес роботи з великими даними

Для роботи з великими даними використовують цілу низку спеціально призначених методів. Одним із прикладів є MapReduce.

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. Він не є заміною для баз даних чи стратегії «Видобувай, трансформуй, завантажуй».

У складі Hadoop наявні:

- розподілена файлова система;
- платформи для аналізу та зберігання даних;
- рівень, що управляє паралельними обчисленнями;
- адміністрування конфігураціями.

Компоненти Hadoop:

1. HDFS – надзвичайно відмовостійка розподілена файлова система, яка відповідає за зберігання даних на кластерах.
2. MapReduce – програмна техніка паралельних обчислень на кластерах.
3. HBase – масштабована, розподілена база даних для операцій випадкового читання/запису.
4. Pig – система обробки даних, призначена для аналізу масивів даних, що використовують мови високого рівня.
5. Hive – застосунок зберігання даних, який надає інтерфейс, схожий на SQL та реляційну модель.

6. Sqoop – проект, призначений для передачі даних між реляційними базами даних та Hadoop.

7. Avro – система серіалізації даних.

8. Oozie – робочий процес для залежних робіт Hadoop.

9. Chukwa – підпроцес Hadoop, визначений як система накопичення даних для моніторингу розподілених систем.

10. Flume – надійна та розподілена колекція потокових журналів подій.

11. ZooKeeper – централізована служба, призначена для забезпечення розподіленої синхронізації та групових процесів.

Загальний процес проходження даних через Hadoop зображений на рис. 2.11.

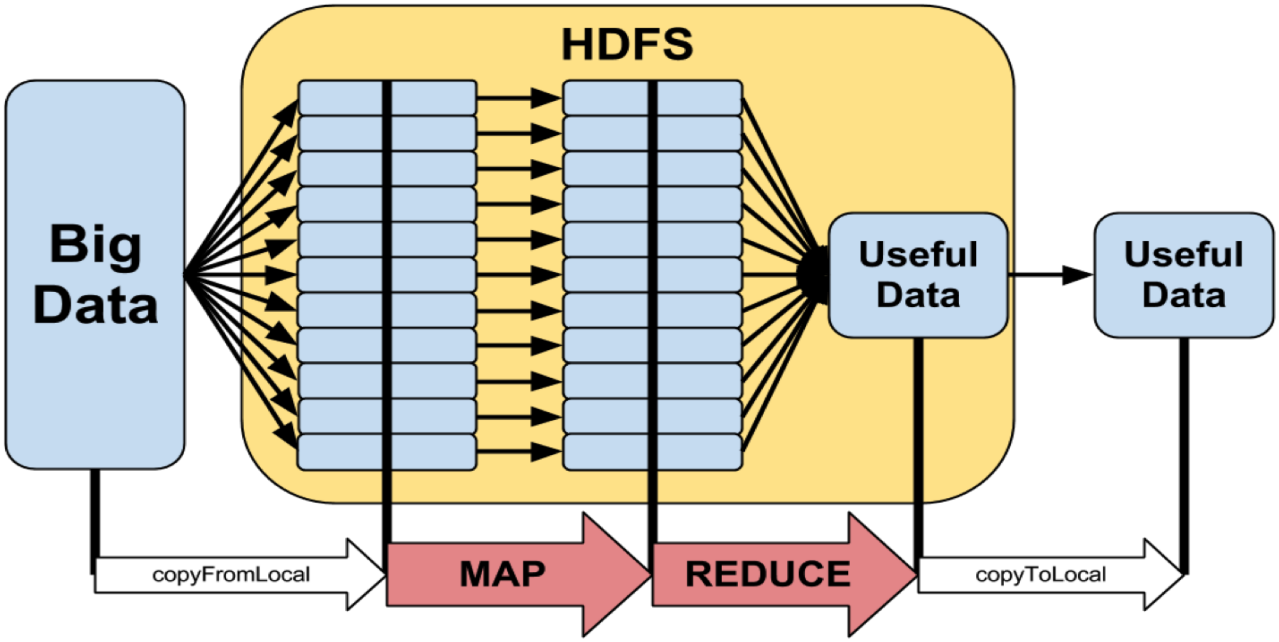

Рис. 2.11. Загальний процес проходження даних через Hadoop

У якості альтернативи Hadoop можна використовувати HPCC (high performance computing cluster). На відміну від Hadoop модель даних в HPCC визначається користувачем. Ключ до вирішення складних проблем може бути легко знайдений використовуючи стек ECL – мови програмування, спеціально орієнтованої на роботу з даними. HPCC контролює, щоб програми ECL виконувались у максимально короткий термін та вузли працювали у паралельному режимі.

Основними трьома компонентами HPCC є:

- Thor Data Refinery (обробка даних) – масивний паралельний ETL-двигунець, який робить можливим інтеграцію даних у масштабі і забезпечує управління згрупованими даними.
- Roxie Data Delivery (видача даних) – розпаралелений модуль, що забезпечує високу пропускну здатність і швидкість, низьку затримку, дозволяє ефективний багатокористувацький видобуток даних і можливість структурованих запитів.
- ECL – забезпечує автоматичне розподілення робочого навантаження між вузлами, має автоматичну синхронізацію алгоритмів та постійно зростаючу бібліотеку методів машинного навчання. Є досить простою мовою програмування, оптимізованою для операцій над великими даними та операціями запитів.

На рис. 2.12 та 2.13 зображені графічні моделі роботи Hadoop та HPCC.

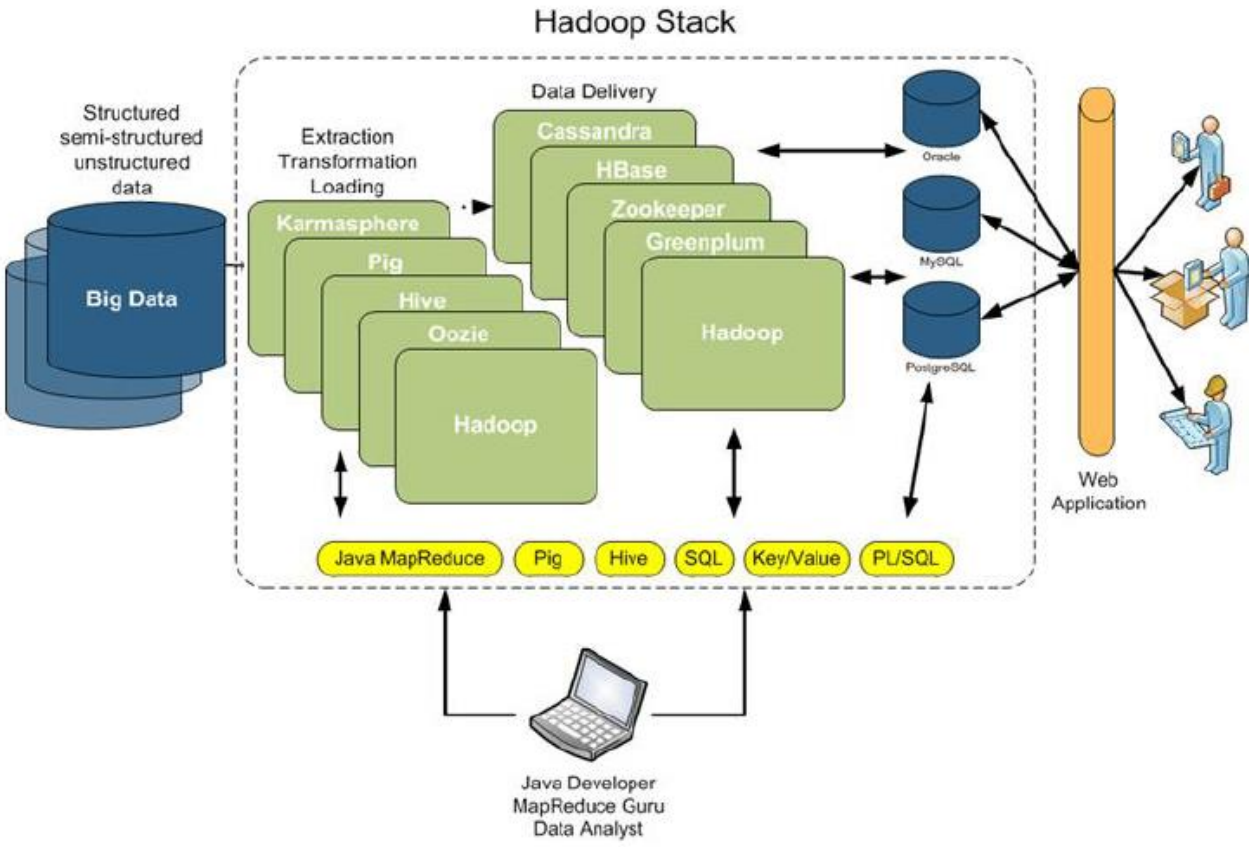

Рис. 2.12. Модель роботи Hadoop

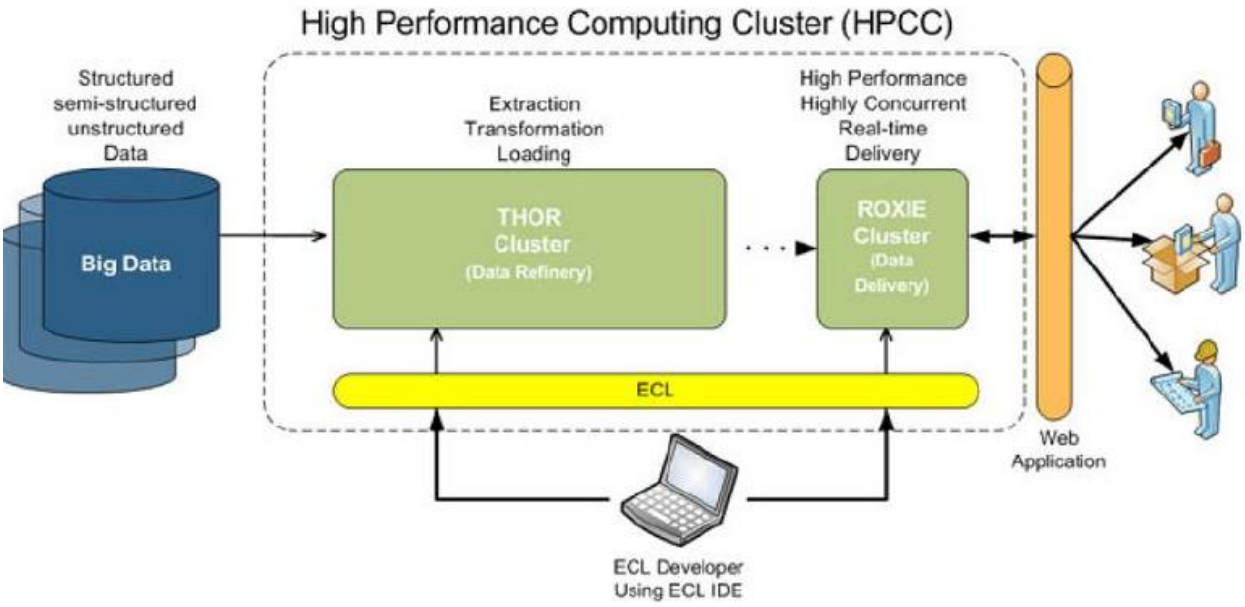

Рис. 2.13. Модель роботи HPCC

Проаналізувавши наведені вище моделі, можна виділити наступні відмінності:

- Кластери HPCC виконуються у модулях Thor і Roxie. Hadoop використовує MapReduce.
- У середовищі HPCC головною мовою програмування є ECL, у той час як процеси Hadoop MapReduce засновані на мові Java.
- Платформа HPCC створює багатоключові та багатоваріантні індекси у розподіленій файловій системі. Hadoop Hbase працює із column-орієнтованою базою даних.
- Зберігання даних використовується в Roxie для структурних запитів та аналітичних застосунків. Натомість, Hadoop Hive забезпечує можливість завантаження даних у HDFS.

Для управління вузлами обробки даних використовується спеціальний модуль YARN, структура якого зображена на рис. 2.14.

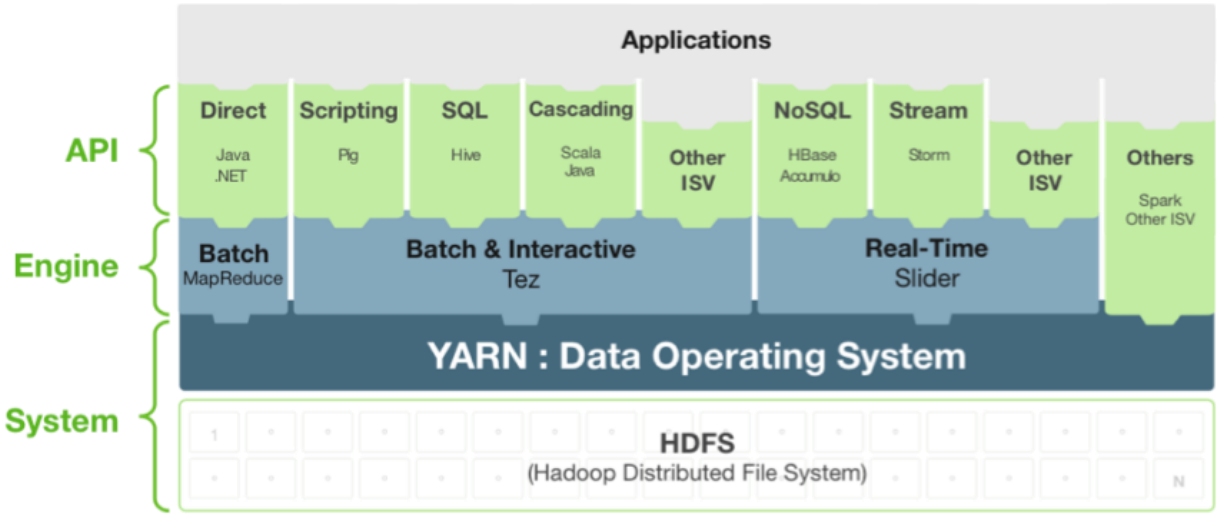

Рис. 2.14. Загальна структура YARN

YARN відповідає за розподіл навантаження між наявними ресурсами системи.

Проте обидві технології здебільшого використовуються для аналізу зібраних даних та тих, що вже пройшли певну обробку. Тим не менш, дедалі частіше аналізувати вхідні масиви інформації потрібно у режимі реального часу. Hadoop та HPCC не є основними інструментами для таких задач, оскільки мають порівняно низьку швидкість обробки великих даних. У таких випадках доцільно використовувати ще одну утиліту – Apache Spark.

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

Spark є обчислювальною системою від компанії Apache із статусом «інкубатора», цей інструмент призначений для прискорення аналізу даних, він зберігає швидкість як під час запуску певних процесів, так і в момент запису різноманітних даних.

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, а також пропоную загальну модель виконання, що оптимізує довільний граф оператора.

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:

- Найшвидший двигунець для обробки 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.

2.4. Проблеми інформаційної безпеки Big Data в мережах оператора стільникового зв’язку

Забезпечення інформаційної безпеки при роботі з великими даними не суттєво відрізняється від звичайного захисту даних. Але відмінності наявні і вони залежать від наступних чинників:

- Дані, які збираються, об’єднуються та аналізуються засобами та інструментами Big Data.
- Інфраструктура, що забезпечує зберігання та розміщення великих даних.

- Технології, які застосовуються для аналізу структурованих, напівструктурованих та неструктурованих великих даних.

Так як зазвичай основним пріоритетом є швидкість обробки великого об’єму інформації, то деякими аспектами безпеки часто нехтують, в основному через те, що немає чіткої класифікації даних, які будуть зберігатися та передаватися.

У випадку оператора стільникового зв’язку Big Data може використовуватися для наступних задач:

- Збільшення об’єму сховищ даних.
- Оптимізація послуг.
- Адаптивне впровадження електронних послуг (e-service).
- Надання послуг у реальному часі.
- Аналітика даних.
- Створення моделей прогнозування для систем підтримки прийняття рішень.
- Оцінка ризиків.

Оператори стільникового зв’язку постійно шукають шляхи розвитку інноваційних послуг, заснованих на даних, які можуть стати джерелом прибутку у випадку якісного аналізу величезної кількості даних, які генеруються їхньою інфраструктурою. Для операторів стільникового зв’язку Big Data – це бізнес-рішення, яке стосується не лише оптимізації внутрішніх процесів. У них є численні можливості розробки застосунків Big Data використовуючи інфраструктуру, пропонуючи ці послуги іншим індустріям, таким як: продаж, фінансові послуги, охорона здоров’я, маркетинг тощо. Тобто оператори стільникового зв’язку виступають не тільки користувачами, але й постачальниками.

У Big Data є багато застосувань у сфері телекомунікацій. Одним з таких є аналітика даних для управління телекомунікаційної інфраструктури. Інформація збирається для визначення трендів та аналізу поведінки мережі. Аналіз даних

також є корисним для внутрішніх сервісів (управління ресурсами, управління доходами, поставками тощо), які обробляють дані, що постійно генеруються та інтегрують техніки Big Data для оптимізації внутрішніх процесів, операцій чи для більш ефективного продажу продуктів.

Також, з використанням Big Data, інформація, яка виникає всередині стільникової мережі, особливо даних про геолокацію, результат аналізу дозволяє виявити тренди та шаблони з метою кращого розуміння поведінки користувачів, засновуючись на подіях, які виникають безперервно протягом року. Інформація збирається та об'єднується через виділені інтерфейси мережевого обладнання (базові станції, антени, мережеві контролери).

Основні проблеми інформаційної безпеки:

1. Перевірка джерела та фільтрація інформації.

Об'єднання персональних даних, які циркулюють у стільникових мережах, зумовлює виникнення важливих проблем безпеки, а конкретно контроль походження даних. Однією проблемою є те, що фреймворк застосунків Big Data збирає інформацію, яка генерується в обладнанні стільникової мережі, у якій використовуються технології різних виробників. Оператор стільникового зв'язку мусить перевіряти і довіряти компонентам інфраструктури – як апаратному, так і програмному забезпеченню – які створюють дані та події, що збираються фреймворками Big Data та бути переконаними у коректності захищеного походження даних.

2. Захист програмного забезпечення.

Необхідним є використання захищених версій програмного забезпечення. Іноді, використання open-source технологій, які можуть бути економічно вигідним рішенням, може стати причиною виникнення вразливостей.

3. Контроль доступу та автентифікація.

Повинні бути впроваджені механізми контролю та автентифікації для різних ролей (користувач/клієнт), які будуть користуватися послугою. Тим не менш, найбільшою проблемою може стати той факт, що автентифікація користувачів та доступ до даних може здійснюватися з різних локацій. Даний

процес не завжди може ефективно контролюватися, через використання децентралізованої моделі та численних взаємодій між системами.

Зокрема, у регульованих галузях захист доступу привілейованого користувача має бути головним пріоритетом. Певним користувачам має бути наданий доступ до високочутливих даних у конкретних бізнес-процесах, але запобігання потенціальному шахрайству з даними може бути проблематичним.

Захист доступу привілейованого користувача потребує чітко визначених політик безпеки й контролю, які впроваджують доступ, заснований на ролях.

4. Захист ланцюжка постачання.

Використання у стільникових мережах різних пристроїв (як смартфонів, так і звичайних мобільних) зумовлює проблему безпеки ланцюжка постачання, оскільки контроль за пристроями покладається на організації, чия діяльність пов’язана з наданням послуг чи доставкою товарів.

5. Захищене управління даними.

У випадку надання послуг Big Data, безпека має розглядатися в якості частини системи. Захист сховищ даних має бути першочерговою задачею, яка, втім, залежить від обраної бізнес-моделі: послуги можуть надаватися лише власними силами чи деякі елементи будуть придбані (орендовані) у третьої сторони.

6. Захист інфраструктури.

Захист інфраструктури має охоплювати не тільки кібер-фізичні системи (наприклад, сенсорні мережі), але й термінали кінцевого користувача, незалежно, чи це буде портативним пристроєм, чи ні. Велика кількість ненадійних пристроїв може бути підключена до мережі компанії у будь-який час, на яких циркулюватимуть критичні дані, що зумовлює необхідність «кінцевого» шифрування.

7. Використання захищеної «хмари».

З використанням Big Data у «хмарі», потрібно брати до уваги специфічні категорії ризиків відносно хмарних технологій: прив’язка до постачальника, порушення цілісності та управління даними.

2.5. Удосконалення початкової моделі за допомогою концепції Big Data

На рис. 2.15 зображена модель, удосконалена методами Big Data.

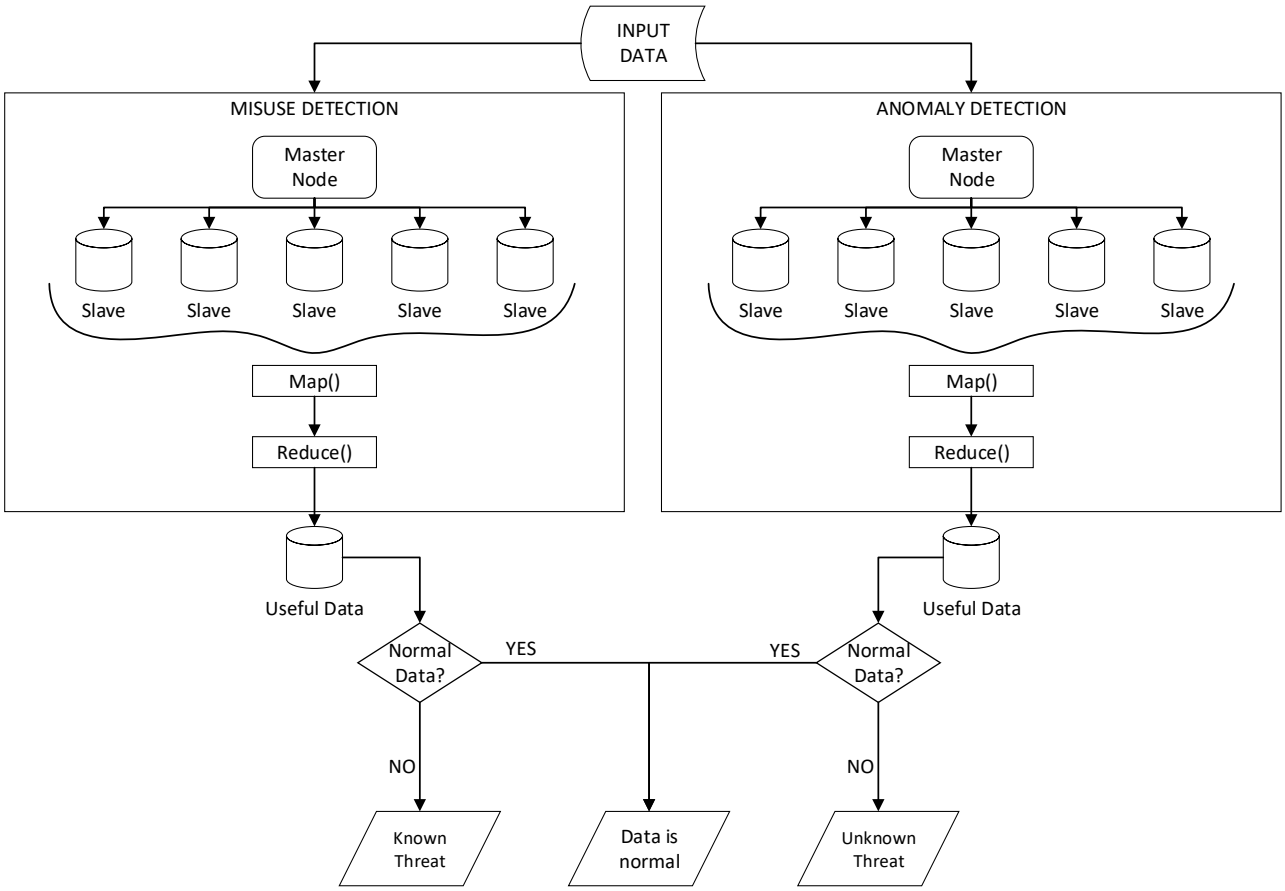

Рис. 2.15. Удосконалена методами Big Data модель

Вхідні дані у паралельному режимі подаються на два модуля, 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.

2.6. Висновки до другого розділу

Отже, проаналізувавши основні загрози та види шахрайства у мережах стільникових операторів, було розроблено гібридну систему виявлення аномалій, з використанням сигнатурного модуля на базі концепції Big Data. Комбінація даних компонентів дозволяє реалізувати виявлення відомих загроз та можливість виявлення нових, ще не внесених до баз даних, а також швидко обробляти вхідні потоки даних.

Розділ 3. ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ МОДУЛЯ ВИЯВЛЕННЯ АНОМАЛІЙ РОЗРОБЛЕНОЇ СИСТЕМИ

1.1. Методика експерименту

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).

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

1.2. Середовище Weka v.3.8

Weka являє собою зручну платформу, яка дозволяє використовувати наявні алгоритми машинного навчання для задач аналізу даних. Алгоритми можуть застосовані напряму до набору даних або викликані від власноруч написаного коду на мові програмування Java.

Weka має у своєму арсеналі інструменти для попередньої обробки даних, класифікації, регресії, кластеризації та візуалізації. Також у Weka наявна можливість розробки нових схем машинного навчання.

На рис. 3.1 зображено стартове вікно програми, у якому можна обрати режим роботи.

Рис. 3.1. Головне вікно Weka

Weka пропонує наступні режими роботи:

1. Explorer.

Розглянемо основні вкладки Explorer. На першій вкладці Preprocess (рис. 3.2) можна обрати початковий масив даних (з файлу, з сайту, з бази даних, згенерувати власноруч), відредагувати його, зберегти в іншому файлі; відфільтрувати масив даних, подивитися характеристики, переглянути результат початкового аналізу (рис. 3.3) та перевірити журнал подій.

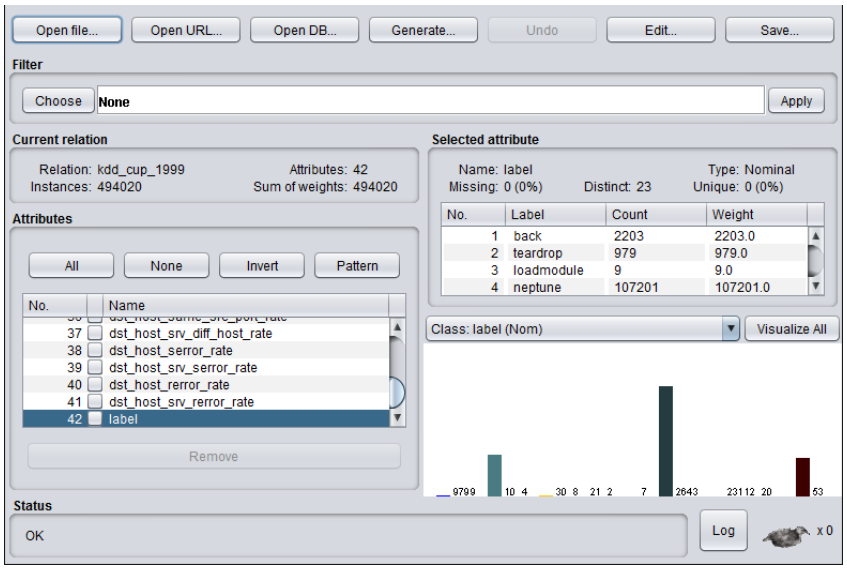

Рис. 3.2. Вікно Explorer

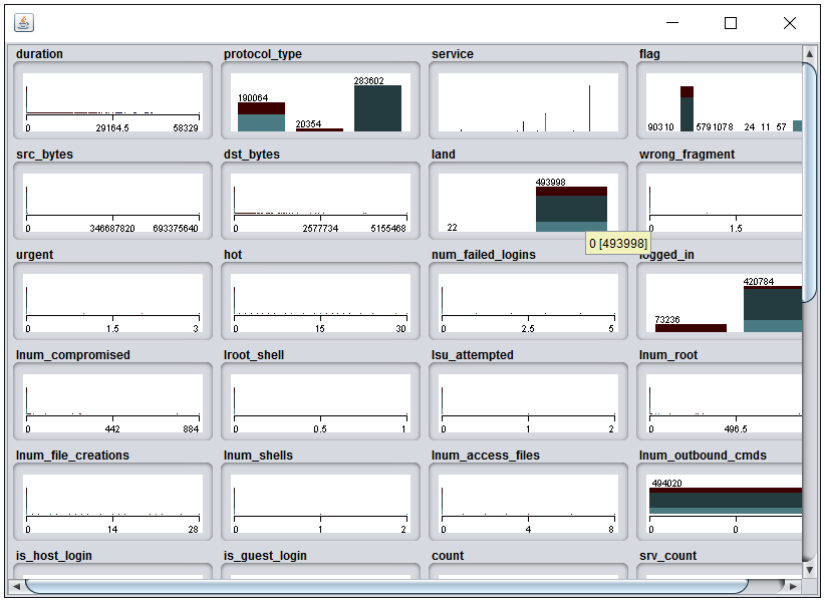

Рис. 3.3. Результат початкового аналізу

На наступній вкладці Classify (рис. 3.4) можна обрати класифікаційний алгоритм (рис. 3.5), опції до нього та режими визначення точності алгоритмів (рис. 3.6):

- Використання тренувальних даних;
- Використання тестових даних;
- Застосування крос-валідації: задається ціле число *n* частин, на які розбивається вихідний набір даних. Для тренування моделі використовується *n–1* частин набору, для тестування – одна, що залишилися;
- Застосування користувацького розподілу даних у відсотках.

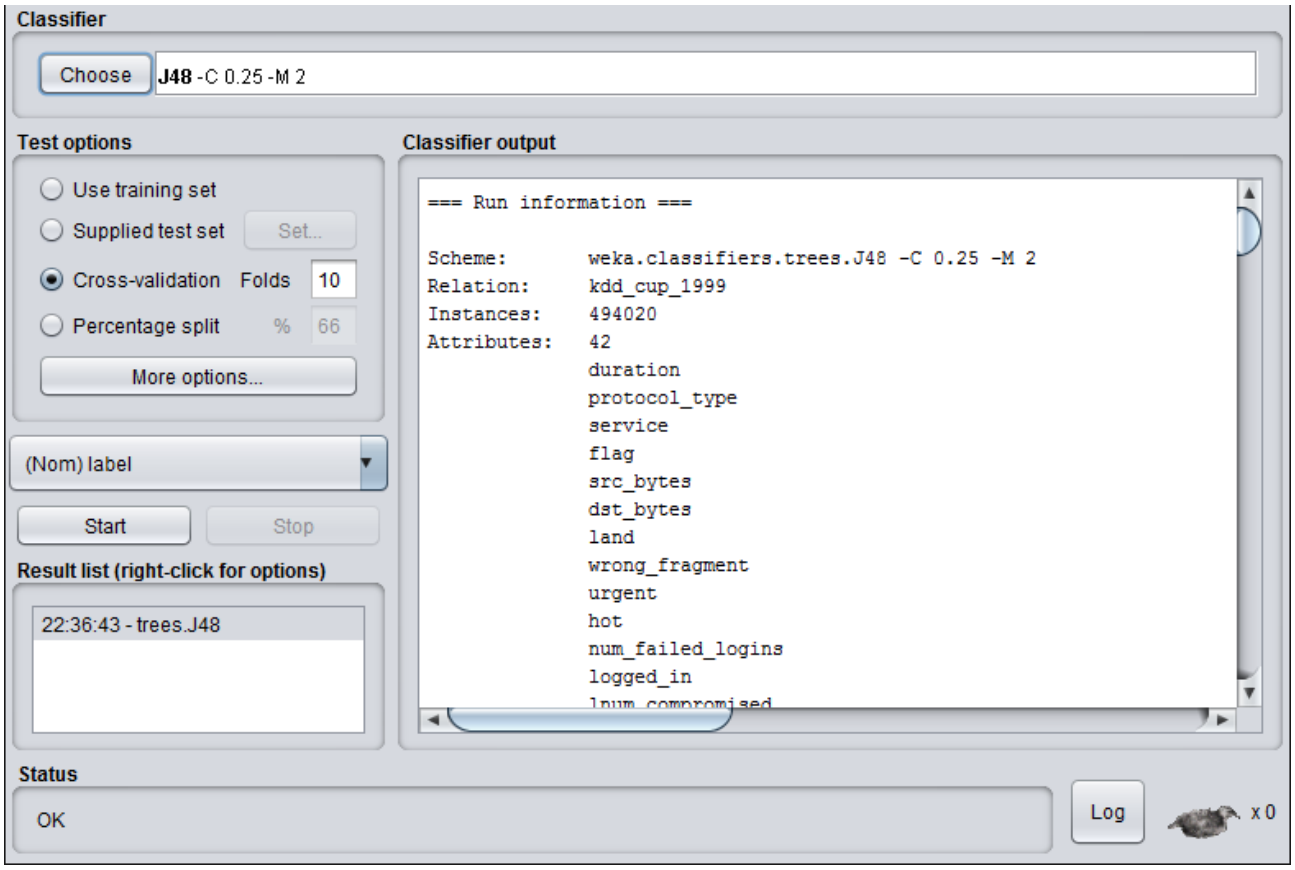

Рис. 3.4. Вкладка Classify

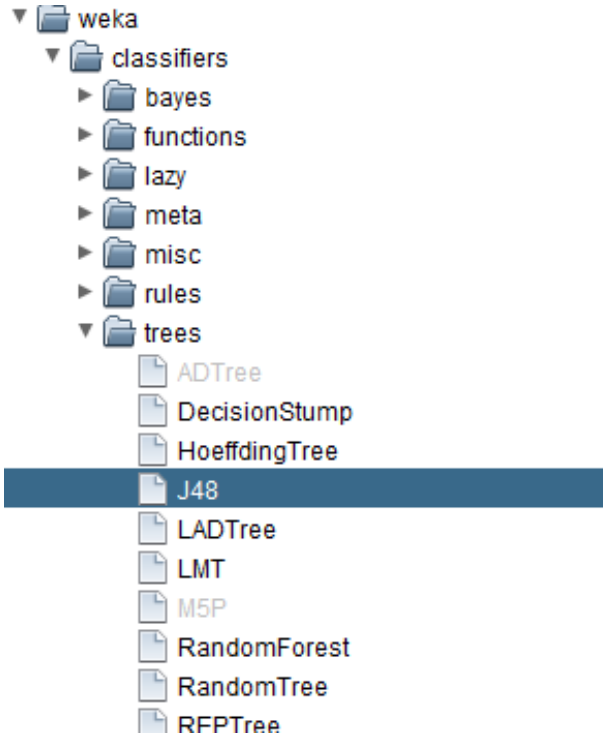

Рис. 3.5. Вікно вибору класифікаційного алгоритму

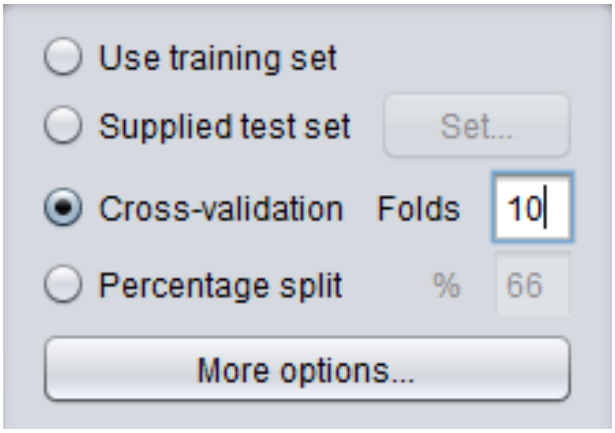

Рис. 3.6. Вікно режимів перевірки точності алгоритмів

На наступній вкладці Cluster можна обрати кластеризаційний алгоритм та задати додаткові опції (рис. 3.7). Структура вікна є схожою до попередньої вкладки.

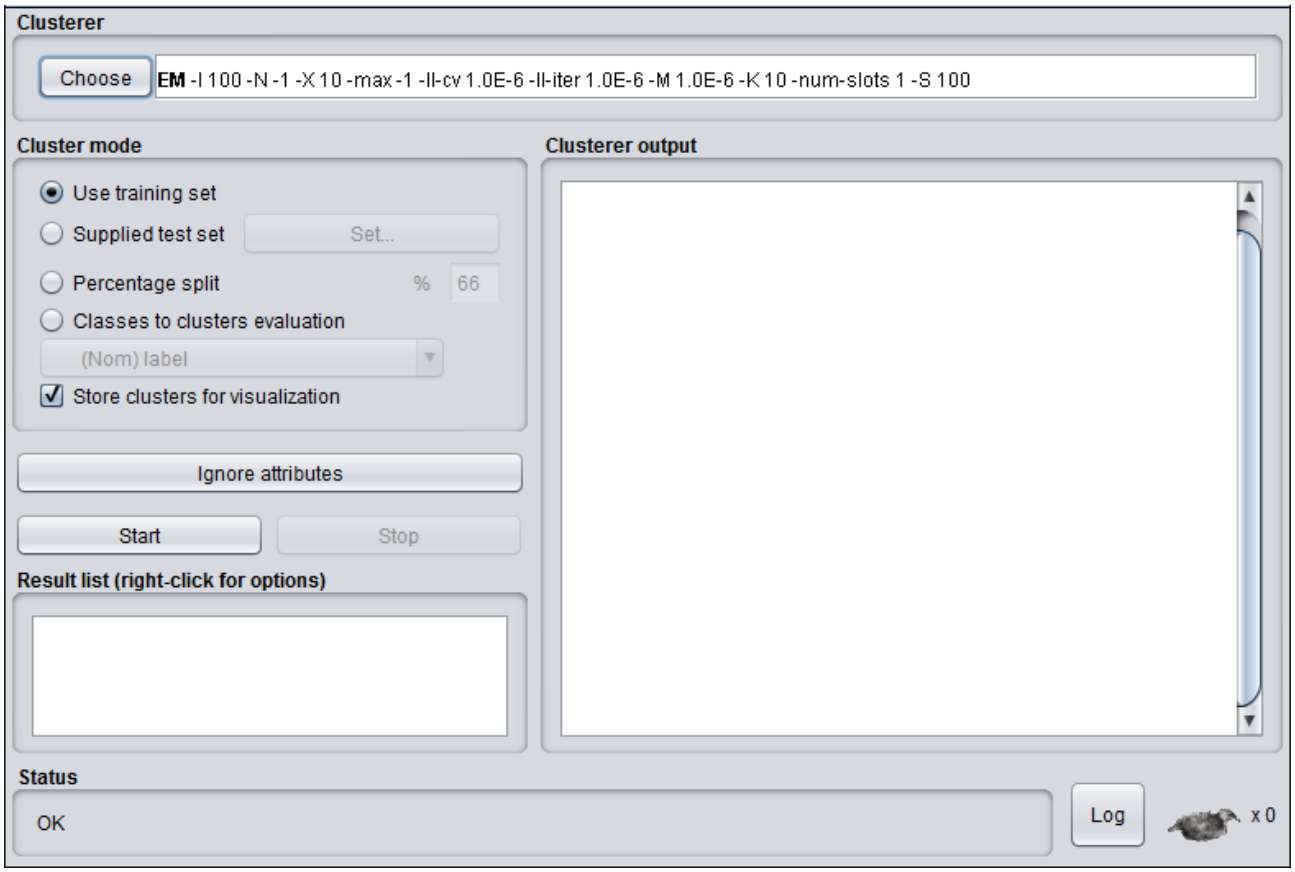

Рис. 3.7. Вкладка Cluster

2. Experiment (рис. 3.8). Даний режим призначений для більш простого запуску алгоритмів та можливості зберігання результатів експерименту у файл. Проте у цьому режимі неможливо переглянути проміжні результати (час побудови моделі, графічний вигляд, відсоток правильно визначених аномалій), тому він є менш придатним для проведення експерименту.

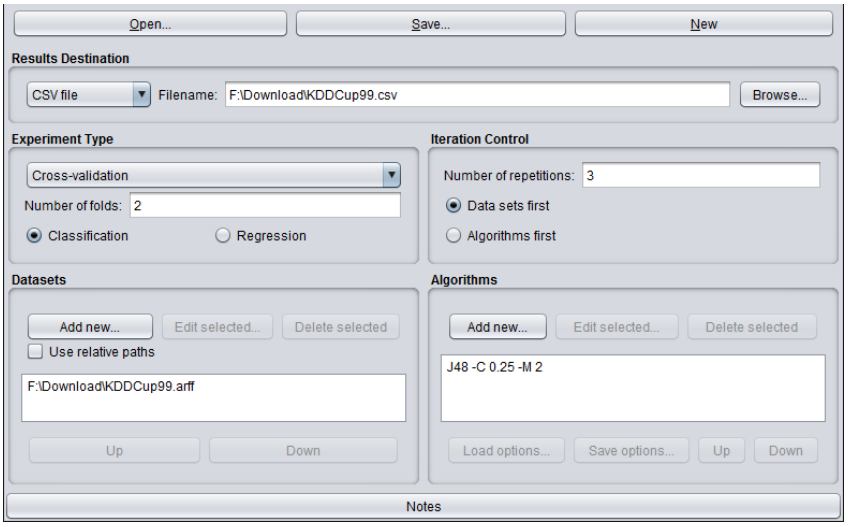

Рис. 3.8. Вікно Experiment

3. Knowledge Flow (рис. 3.9).
4. Workbench (рис. 3.10).
5. Simple CLI (рис. 3.11).

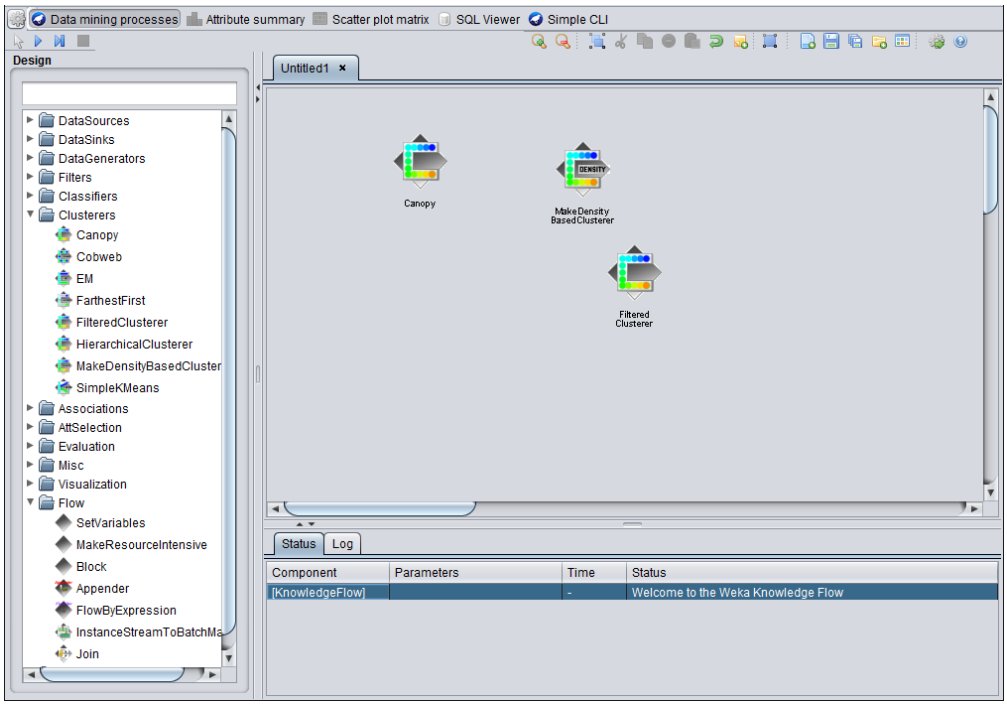

Рис. 3.9. Вікно KnowledgeFlow

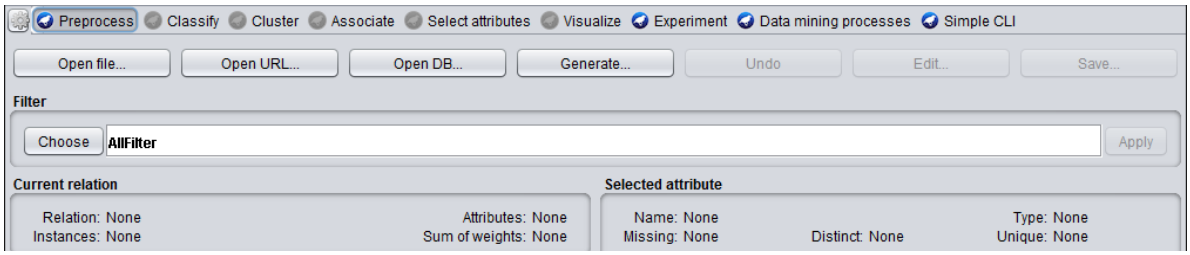

Рис. 3.10 Вікно Workbench

```
Command must be one of:
        java <classname> <args> [ > file]
        kill
        capabilities <classname> <args>
        cls
        history
        exit
        help <command>
```

Рис. 3.11. Вікно Simple CLI

Дані режими є більш розширеними інструментами для побудови моделей потоків даних, створення власної «майстерні» з можливістю гнучких налаштувань для кожного етапу, та можливості використання командного рядка для виконання певних операцій.

1.3. Проведення експерименту

Етап 1: завантаження вихідних даних у середовище (рис. 3.12).

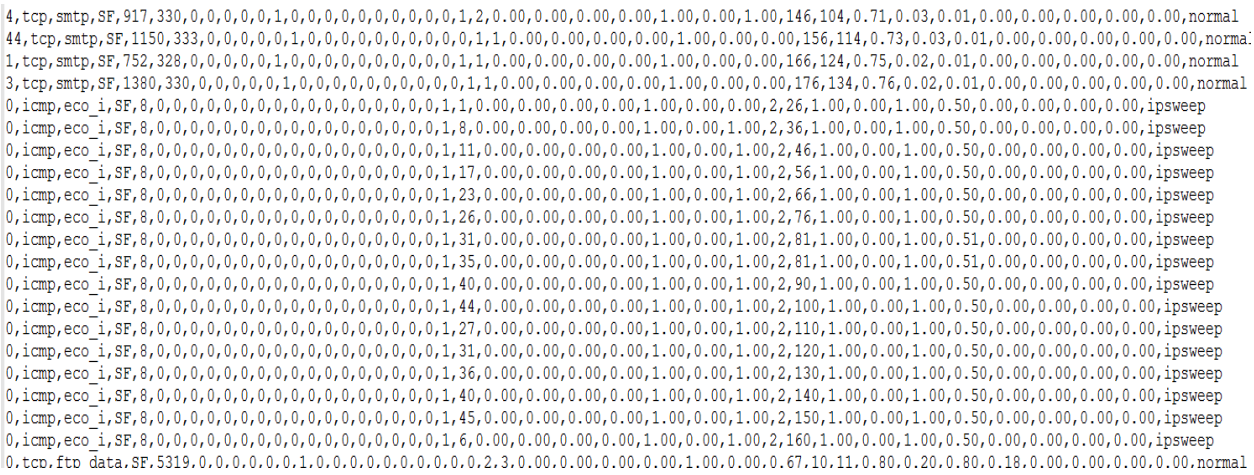

Рис. 3.12. Вихідні дані масиву даних KDDCup

Етап 2: обирання класифікаційного алгоритму, у даному випадку – це J48 – реалізація на мові програмування Java одного з найпопулярніших алгоритмів Decision Tree (рис. 3.13)

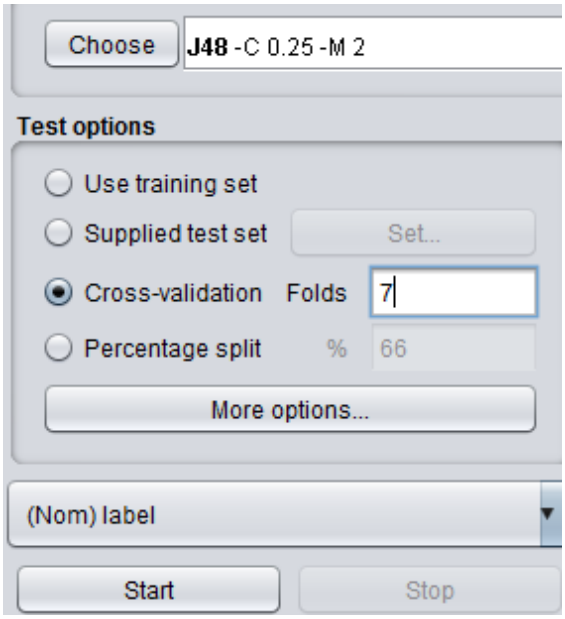

Рис. 3.13. Опції для алгоритму J48

Етап 3: побудова моделі Decision Tree.

На цьому етапі можна відстежити час побудови моделі (рис. 3.14), текстовий вигляд дерева (рис. 3.15) та розмір дерева (рис. 3.16).

```
Time taken to build model: 61.06 seconds
```

Рис. 3.14. Час побудови моделі

```
dst_host_same_src_port_rate <= 0.99
|   srv_serror_rate <= 0.58
|   |   src_bytes <= 6
|   |   |   protocol_type = tcp
|   |   |   |   dst_host_diff_srv_rate <= 0.13: normal (2270.0/1.0)
|   |   |   |   dst_host_diff_srv_rate > 0.13
|   |   |   |   |   dst_host_diff_srv_rate <= 0.88
|   |   |   |   |   |   dst_host_rerror_rate <= 0.16
|   |   |   |   |   |   |   duration <= 0: normal (14.0)
|   |   |   |   |   |   |   duration > 0: satan (3.0)
|   |   |   |   |   |   dst_host_rerror_rate > 0.16: portsweep (10.0/1.0)
|   |   |   |   |   dst_host_diff_srv_rate > 0.88: ipsweep (4.0)
|   |   |   protocol_type = udp: satan (44.0/2.0)
|   |   |   protocol_type = icmp: normal (0.0)
|   |   src_bytes > 6
|   |   |   lnum_file_creations <= 0
|   |   |   |   count <= 46
|   |   |   |   |   dst_bytes <= 0
|   |   |   |   |   |   dst_host_srv_diff_host_rate <= 0.11
|   |   |   |   |   |   |   dst_host_same_src_port_rate <= 0.8
|   |   |   |   |   |   |   |   src_bytes <= 353
|   |   |   |   |   |   |   |   |   src_bytes <= 326: normal (3782.0/12.0)
|   |   |   |   |   |   |   |   |   src_bytes > 326: warezclient (36.0)
|   |   |   |   |   |   |   |   src_bytes > 353: normal (2810.0/13.0)
```

Рис. 3.15. Текстовий вигляд дерева

```
Size of the tree :      833
```

Рис. 3.16. Розмір дерева

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.

Етап 4: результати експерименту.

У ході експерименту були отримані наступні результати:

- Кількість та відсоток правильно та помилково визначених даних (рис. 3.17).
- Графічний вигляд побудованого дерева (рис. 3.18).

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	

Рис. 3.17. Відсоток правильно та помилково визначених даних

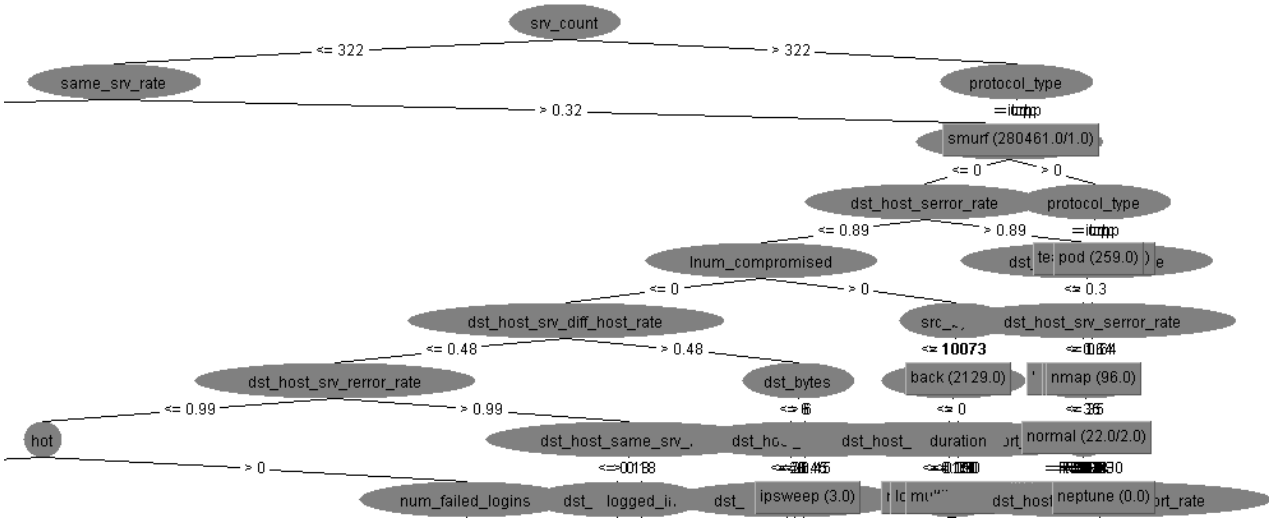

Рис. 3.18. Графічний вигляд частини побудованого дерева

1.4. Висновки до третього розділу

Отже, у ході експерименту було виявлено, що обраний алгоритм виявлення аномалій Decision Tree, а конкретно – одна з його реалізацій – J48, здатний забезпечити високу точність виявлення аномалій – 99,96%, що в сукупності з використанням сигнатурного модуля та концепції Big Data, дає можливість припустити, що обрана система функціонуватиме коректно і не допускатиме велику кількість помилок, навіть при роботі у режимі реального часу з даними набагато більшого розміру та з менш чіткою класифікацією.

ВИСНОВКИ

Результатом виконаної роботи є розроблена модель система виявлення аномалій аномалій для оператора стільникового зв’язку на основі концепції Big Data. Під час виконання роботи було отримано наступні висновки:

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.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Бондаровець С.С. Сучасні методи виявлення аномалій / С.С. Бондаровець // «ITSEC»: VI міжнар. наук.-техн. конф.: тези доп. – К.: НАУ, 2016.– С. 76-77.

2. Abe N. Outlier detection by active learning / N. Abe. – New York: ACM Press, 2006. – P. 504-509.

3. Abraham B. Bayesian analysis of some outlier problems in time series / B. Abraham. – New York: Biometrika, 1979. – P. 229-236.

4. Abraham B. Outlier detection and time series modeling / Abraham B. – New York: Technometrics, 1989. – P. 241-248.

5. Addison J. Effectiveness of feature extraction in neural network architectures for novelty detection / J. Addison // 9th International Conference on Artificial Neural Networksm, P. 976-981.

6. Agarwal D. An empirical bayes approach to detect anomalies in dynamic multidimensional arrays / D. Agarwal // 5th IEEE International Conference on Data Mining. – Washington: IEEE Computer Society, 2005. – P. 26-33.

7. Agarwal D. Detecting anomalies in cross-classified streams: a bayesian approach / D. Agarwal. – Washington: Knowledge and Information Systems, 2006. – P. 29-44.

8. Bakar Z. A comparative study for outlier detection techniques in data mining / Z. Bakar. – USA: IEEE, 2006. – P. 1-6.

9. Baker D. A hierarchical probabilistic model for novelty detection in text / D. Baker. // International Conference on Machine Learning, 1999. – P. 25-36.

10. Barbara D. Bootstrapping a data mining intrusion detection system / D. Barbara. – USA: ACM Press, 2003. – P. 421-425.

11. Barbara D. Detecting novel network intrusions using bayes estimators / D. Barbara. – San Francisco: First SIAM, 2001. – P. 69-98.

12. Chen H. Business intelligence and analytics: From big data to big impact / H. Chen // MIS Quarterly. – 2012. – No.36(4). – P 1165-1188.

13. Chung W. BizPro: Extracting and categorizing business intelligence factors from textual news articles / W. Chung // International Journal of Information Management. – 2014. – No.34(2). – P. 272-284.

14. Crook P. A robot implementation of a biologically inspired method for novelty detection / P. Crook. – Manchester: Intelligent Mobile Robots Manchester, 2013. – P. 268-289.

15. Crook P. A tale of two filters - on-line novelty detection / P. Crook // International Conference on Robotics and Automation, 2011. –P. 3894–3899.

16. Cukier K. Data, data everywhere: A special report on managing information / K. Cukier // The Economist. – 2010. – No. 2016. – P. 7-27.

17. Cun Y. Handwritten digit recognition with a back-propagation network. Advances in neural information processing systems / Y. Cun. – Cambridge: MIT Press, 1990. – P. 396-404.

18. Das K. Detecting anomalous records in categorical datasets / K. Das. – USA: ACM Press, 2007. – P. 1039-1044.

19. Diebold F. A personal perspective on the origin(s) and development of “big data”: The phenomenon, the term, and the discipline / F. Diebold // Social Science Research Network. – 2012. – No.25. – P. 113-139.

20. Fan J. Challenges of big data analysis. National Science Review / J. Fan // Journal of the Royal Statistical Society: Series B (Statistical Methodology). – 2014. – No.70(5). – P. 849-911.

21. Feldman R. Techniques and applications for sentiment analysis / R. Feldman. – USA: Communications of the ACM, 2013. – P. 82-89.

22. Gwadera R. Markov models for identification of significant episode / R. Gwadera // 5th SIAM International Conference on Data Mining, 2005. – P. 126-136.

23. Hakeem A. Video analytics for business intelligence / A. Hakeem // Business Intelligenceю – 2012. – No.9. – P. 309-354.

24. Hasan M. A survey of link prediction in social networks / M. Hasan // Social network data analytics. – 2011. – No.17. – P. 243–275.

25. He W. / Social media competitive analysis and text mining: A case study in the pizza industry // International Journal of Information Management. – 2015. – No.33(3). – P. 464-472.

26. Hollmen J. Call-based fraud detection in mobile communication networks using a hierarchical regime-switching model / J. Hollmen. – Cambridge: MIT Press, 1999. – P. 889-895.

27. Jakubek S. Fault-diagnosis using neural networks with ellipsoidal basis functions / S. Jakubek // American Control Conference, 2002. – P. 3846-3851.

28. Kosoresow A. Intrusion detection via system call traces / A. Kosoresow. – USA: IEEE Software, 1997. – P. 35-42.

29. Kou Y. Spatial weighted outlier detection / Y. Kou // SIAM Conference on Data Mining, 2006. – P. 126-159.

30. Kruegel C. Bayesian event classification for intrusion detection / C. Kruegel. – Portland: IEEE Computer Society, 2003. – 129 p.

31. Mahoney M. A machine learning approach to anomaly detection / M. Mahoney V. – CS: Tech. Rep., 2003. – 203 p.

32. Mahoney M. Learning nonstationary models of normal network traffic for detecting novel attacks / M. Mahoney. –USA: ACM Press, 2002. – 385 p.

33. Mahoney M. Learning rules for anomaly detection of hostile network traffic / M. Mahoney // The 3rd IEEE International Conference on Data Mining. IEEE Computer Society, 2003. – P. 590-601.

34. Moya M. One-class classifier networks for target recognition applications / M. Moya // In Proceedings on World Congress on Neural Networks, International Neural Network Society, Portland, 1993. – P. 797-801.

35. Murray A. Novelty detection using products of simple experts - a potential architecture for embedded systems / A. Murray. – New York: Neural Networks, 2001. – P. 1257-1264.

36. Nairac A. A system for the analysis of jet engine vibration data / A. Nairac. – San Franc.: Integrated Computer-Aided Engineering, 1999. – 56 p.

37. Nairac A. Choosing an appropriate model for novelty detection // In Proceedings of the 5th IEEE International Conference on Artificial Neural Networks, 1997, P. 227-232.

38. Ng R. Efficient and effective clustering methods for spatial data mining // R. Ng // In Proceedings of the 20th International Conference on Very Large Data Bases. Morgan Kaufmann (Publishers Inc., San Francisco), 1994, P. 144-155.

39. Odin T. . Novelty detection using neural network technology / T. Odin // In Proceedings of the COMADEN Conference (Houston), 2000, P. 12-19.

40. Otey M. Fast distributed outlier detection in mixed-attribute data sets / M. Otey. – USA:Data Mining and Knowledge Discovery, 2006. – p. 228.

41. Otey M. Towards nic-based intrusion detection / M. Otey // In Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining, 2003, P. 723-728.

42. Papadimitriou S. Fast outlier detection using the local correlation integral / S. Papadimitriou. – Pittsburgh: Intel Research Laboratory, 2002. – 16 p.

43. Parra L. Statistical independence and novelty detection with information preserving nonlinear maps / L. Parra. – New York: Neural Computing, 1996. – P. 260- 269.

44. Parzen E. On the estimation of a probability density function and mode. Annals of Mathematical Statistics / E. Parzen. – San Franc.:MS, 1962. – 1076 p.

45. Qin M. Frequent episode rules for internet anomaly detection / M. Qin // In Proceedings of the 3rd IEEE International Symposium on Network Computing and Applications, IEEE Computer Society, 2004, P. 12-26.

46. Ramadas M. Detecting anomalous network traffic with self-organizing maps / M. Ramadas. – USA: RAID print, 2003. – P. 36-54.

47. Ramaswamy S. Efficient algorithms for mining outliers from large data sets / S. Ramaswamy. – USA: ACM Press, 2000. – P. 427-438.

48. Ratsch G. Constructing boosting algorithms from svms: An application to one-class classification / G. Ratsch // IEEE conference Transactions on Pattern Analysis and Machine Intelligence, 2002, P. 1184-1199.

49. Roberts S. Extreme value statistics for novelty detection in biomedical signal processing / S. Roberts // 1st International Conference on Advances in Medical Signal and Information Processing, 2002, P. 166-172.

50. Roberts S. Novelty detection using extreme value statistics / S. Roberts // Vision, Image and Signal processing (IEEE). – 1999. – Vol. 146. – P. 124-129.

51. Scott S. Detecting network intrusion using a markov modulated nonhomogeneous poisson process / S. Scott // Journal of the American Statistical Association. – 2001. – No.12. – P. 125-139.

52. Sebyala A. Active platform security through intrusion detection using naive bayesian network for anomaly detection // The London Communications Symposium, 2002, P. 12-25.

53. Sekar R. A fast automaton-based method for detecting anomalous program behaviors / R. Sekar. – Washington: IEEE Computer Society, 2001. – p.144.

54. Sekar R. A high-performance network intrusion detection system / R. Sekar // The 6th ACM conference on Computer and communications security (ACM Press), 2015, P. 8-17.

55. Suzuki E. Detecting interesting exceptions from medical test data with visual summarization // The 3rd IEEE International Conference on Data Mining, 2003, P. 315-322.

56. Ye N. An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems / N. Ye // Journal of Quality and Reliability Engineering Internationaand Chen. – 2001. – No.3. – P. 105-112.

57. Yi B. Online data mining for co-evolving time sequences / B. Yi // The 16th International Conference on Data Engineering, IEEE Computer Society, 2000, P. 36-49.

58. Ypma A. Novelty detection using self-organizing maps / A. Ypma // Connectionist Based Information Systems. – Vol. 2. – P. 1322-1325.

59. Yu J. Finding centric local outliers in categorical/numerical spaces / J. Yu // Knowledge and Information Systems. – 2006. – No.9. – P. 309-338.

60. Zeevi A. Time series prediction using mixtures of experts / A. Zeevi // Advances in Neural Information Processing. – 1997. – Vol. 9. – P. 136-159.

61. Zhang J. Detecting outlying subspaces for high-dimensional data: the new task, algorithms, and performance / J. Zhang // Knowledge and Information Systems. – 2006. – No.12. – P. 333-355.

62. Zhang Z. Hide: a hierarchical network intrusion detection system using statistical preprocessing and neural network classification / Z. Zhang // IEEE Workshop on Information Assurance and Security. – 2001. – No.16. – P. 85-90.

63. Zimmermann J. Distributed intrusion detection in clusters based on non-interference / J. Zimmermann. – Darlinghurst, Australia: Australian Computer Society, Inc, 2006. – 96 p.