МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Західноукраїнський національний університет
Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно-обчислювальних систем і управління

ТРАЧ Юлія Іванівна

Нейромережевий метод оцінювання якості великих даних / A Neural Network Method for Assessing the Quality of Big Data

спеціальність: 122 – Комп'ютерні науки
освітньо-професійна програма – Комп'ютерні науки

Кваліфікаційна робота

Виконала студентка групи
КНм-21
Ю.І. Трач
_____

Науковий керівник:
д.т.н., доцент М.П. Комар
_____

Кваліфікаційну роботу
допущено до захисту:
«___» _____ 2022 р.
Завідувач кафедри
_____ М. П. Комар

ТЕРНОПІЛЬ – 2022

Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно-обчислювальних систем і управління
Освітній ступінь «магістр»
спеціальність: 122 – Комп'ютерні науки
освітньо-професійна програма – Комп'ютерні науки

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ М.П. Комар
«____» _____20__р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТЦІ
ТРАЧ Юлія Іванівна
(прізвище, ім'я, по батькові)

1. Тема кваліфікаційної роботи
Нейромережевий метод оцінювання якості великих даних / A Neural Network Method for Assessing the Quality of Big Data
керівник роботи д.т.н., доцент М.П. Комар
затверджені наказом по університету від 31 грудня 2021 року № 606.

2. Строк подання студентом закінченої кваліфікаційної роботи 16 листопада 2022 року.

3. Вихідні дані до кваліфікаційної роботи: завдання на кваліфікаційну роботу студента, наукові статті, технічна література.

4. Основні питання, які потрібно розробити
- дослідження актуальності задачі оцінювання якості великих даних;
- аналіз відомих інструментів оцінювання якості великих даних;
- аналіз відомих методів оцінювання якості великих даних;
- постановка задачі дослідження;
- визначення параметрів якості великих даних;
- визначення дослідницьких наборів для оцінки параметрів якості великих даних;
- дослідження моделей оцінювання якості великих даних;
- визначення узагальненого результату параметрів якості даних;
- проведення експериментальних досліджень запропонованих рішень.

5. Перелік графічного матеріалу у роботі
- структура нейронної мережі, що відображає всі параметри якості даних;
- діаграма розсіювання;
- параметри якості даних: прогнозовані та фактичні.

6. Консультанти розділів кваліфікаційної роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		Завдання видав	Завдання прийняв

7. Дата видачі завдання 11 жовтня 2021 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів кваліфікаційної роботи	Примітка
1	Аналіз відомих методів та засобів оцінювання якості великих даних	12.2021 р. – 03.2022 р.	
2	Моделі оцінювання якості великих даних	03.2022 р. – 05.2022 р.	
3	Метод оцінювання якості великих даних та експериментальні дослідження	05.2022 р. – 11.2022 р.	
4	Повне завершення та представлення кваліфікаційної роботи на кафедру	16.11.2022 р.	

Студентка _____ Ю.І. Трач
підпис

Керівник роботи _____ д.т.н., доцент М.П. Комар
підпис

РЕЗЮМЕ

Кваліфікаційна робота на тему «Нейромережевий метод оцінювання якості великих даних» на здобуття освітнього ступеня «Магістр» зі спеціальності 122 «Комп'ютерні науки» освітньої програми «Комп'ютерні науки» написана обсягом в 75 сторінок і містить 7 ілюстрацій, 3 таблиці, 1 додаток та 26 використаних джерел.

Метою кваліфікаційної роботи є дослідження моделей оцінювання якості великих даних та розробка методу оцінювання якості великих даних.

Методи досліджень: узагальнення інформації з літературних джерел; методи обробки та аналізу великих даних; метод Монте-Карло; нейромережеві методи обробки інформації; метод лінійного регресійного аналізу; методи проведення експериментальних досліджень; методи візуалізації та інтерпретації отриманих результатів.

Результати дослідження: запропоновано метод оцінювання якості великих даних, який базується на восьми параметрах якості даних, який на відміну від відомих дозволить обробляти дані, які генеруються програмним забезпеченням або датчиками у великих обсягах, на високих швидкостях і у різних форматах. Досліджено моделі оцінювання якості великих даних. Для розрахунку якості сукупності даних використано метод Монте-Карло та автоенкодерну нейронну мережу. Вивчення конкретного прикладу проводиться за допомогою лінійного регресійного аналізу, і всі параметри якості великих даних підтверджуються позитивними результатами.

Результати роботи можна використовувати для створення контрольних показників та протоколів для оцінки та оптимізації якості даних.

Ключові слова: ВЕЛИКІ ДАНІ, ЯКІСТЬ ДАНИХ, АВТОЕНКОДЕРНА НЕЙРОННА МЕРЕЖА, МЕТОД МОНТЕ-КАРЛО, РЕГРЕСІЙНИЙ АНАЛІЗ.

ABSTRACT

Qualification work on the topic «A Neural Network Method for Assessing the Quality of Big Data» for Master's degree on speciality 122 «Computer Science» educational and professional program «Computer Science» is written on 75 pages and it contains 7 figures, 3 tables, 1 annex and 26 sources.

The purpose of the qualification work is to study the models for assessing the quality of big data and to develop a method for assessing the quality of big data.

Research methods: generalization of information from literary sources; methods of processing and analyzing big data; the Monte Carlo method; neural network methods of information processing; method of linear regression analysis; methods of conducting experimental research; methods of visualization and interpretation of the obtained results.

Research results: a method for evaluating the quality of big data is proposed, which is based on eight parameters of data quality, which, unlike the known ones, will allow processing data generated by software or sensors in large volumes, at high speeds and in various formats. The models for assessing the quality of big data were studied. The Monte Carlo method and an auto-encoder neural network were used to calculate the quality of the data set. A case study is conducted using linear regression analysis, and all parameters of the quality of big data are confirmed by positive results.

The results of the work can be used to create benchmarks and protocols for evaluating and optimizing data quality.

Keywords: BIG DATA, DATA QUALITY, AUTOENCODERN NEURAL NETWORK, MONTE CARLO METHOD, REGRESSION ANALYSIS.

ЗМІСТ

ВСТУП

Актуальність теми. Завдяки прогресу таких технологій, як хмарні обчислення, Інтернет речей, пристрої соціальних мереж тощо, використання мобільних додатків на сьогодні генерує велику кількість даних, ніж будь-коли раніше. За даними технологічної дослідницької компанії Gartner, станом на 2020 рік 25 мільярдів пристроїв підключених до мережі [17].

Однак через величезний обсяг даних, які генеруються, високу швидкість, з якою надходять нові дані, і велику різноманітність різнорідних даних, поточна якість даних далека від ідеальної [8].

За оцінками, помилкові дані обходяться підприємствам США приблизно в 600 мільярдів доларів на рік [5]. Наразі не існує стандартного методу вимірювання якості даних, тому все ще потрібно встановити надійні контрольні показники.

Таким чином, існує велика потреба в забезпеченні якості великих даних, яке можна визначити як «вивчення та застосування різноманітних процесів забезпечення, методів, стандартів, критеріїв і систем для забезпечення якості великих даних з точки зору набору параметрів якості» [6].

На сьогодні актуальною задачею є забезпечення якості та перевірки великих даних, а саме:

1. Необхідно підвищити усвідомлення важливості забезпечення якості великих даних.

2. Існує потреба в чітко визначених стандартах забезпечення якості.

3. Необхідно провести дослідження моделей якості великих даних [6].

Щоб задовольнити ці потреби, необхідно розробити чітко визначені стандарти забезпечення якості та перевірки великих даних. З цією метою необхідно структурувати відповідні програми забезпечення якості великих даних, а також визначити та дослідити моделі якості великих даних.

Мета і завдання дослідження. Метою кваліфікаційної роботи є дослідження моделей оцінювання якості великих даних та розробка методу

оцінювання якості великих даних.

Для досягнення поставленої мети визначено ряд завдань:

– дослідження актуальності задачі оцінювання якості великих даних;

– аналіз відомих інструментів оцінювання якості великих даних;

– аналіз відомих методів оцінювання якості великих даних;

– постановка задачі дослідження;

– визначення параметрів якості великих даних;

– визначення дослідницьких наборів для оцінки параметрів якості великих даних;

– дослідження моделей оцінювання якості великих даних;

– визначення узагальненого результату параметрів якості даних;

– проведення експериментальних досліджень запропонованих рішень.

Об’єкт дослідження – процеси збору, попередньої обробки та аналізу великих даних.

Предмет дослідження – моделі та методи оцінювання якості великих даних.

Методи дослідження. Для вирішення поставлених задач в роботі використовувалися наступні методи:

– узагальнення інформації з літературних джерел;

– методи обробки та аналізу великих даних;

– метод Монте-Карло;

– нейромережеві методи обробки інформації;

– метод лінійного регресійного аналізу;

– методи проведення експериментальних досліджень;

– методи візуалізації та інтерпретації отриманих результатів.

Наукова новизна одержаних результатів. Запропоновано метод оцінювання якості великих даних, який базується на восьми параметрах якості даних, який на відміну від відомих дозволить обробляти дані, які генеруються

програмним забезпеченням або датчиками у великих обсягах, на високих швидкостях і у різних форматах.

Практичне значення отриманих результатів. Досліджено моделі оцінювання якості великих даних. Для розрахунку якості сукупності даних використано метод Монте-Карло та автоенкодерну нейронну мережу. Вивчення конкретного прикладу проводиться за допомогою лінійного регресійного аналізу, і всі параметри якості великих даних підтверджуються позитивними результатами.

Публікації та апробація. Результати кваліфікаційної роботи апробовані та опубліковані у матеріалах:

– міжнародної наукової Інтернет-конференції «Інформаційне суспільство: технологічні, економічні та технічні аспекти становлення», 2022 р.

–- школи-семінару молодих вчених і студентів «Комп’ютерні інформаційні технології» (CIT’2022).

Кваліфікаційна робота складається із вступу, трьох розділів, висновків, списку використаних джерел та додатків.

1 АНАЛІЗ ВІДОМИХ МЕТОДІВ ТА ЗАСОБІВ ОЦІНЮВАННЯ ЯКОСТІ ВЕЛИКИХ ДАНИХ

1.1 Дослідження актуальності задачі оцінювання якості великих даних

Одним із наслідків неякісних даних є втрата бізнесу. Як зазначено в [2], низька якість даних може спричинити багато матеріальних і нематеріальних втрат для підприємств. Орієнтовні витрати можуть становити від 8% до 12% доходів типової організації та можуть генерувати приблизно 40%-60% витрат обслуговуючої організації [19].

Очевидно, що погані дані можуть перешкоджати досягненню цілей щодо прибутків компаній. Вони також можуть викликати помилки в спілкуванні, що може призвести до незадоволених клієнтів [6].

Іншим негативним ефектом низької якості даних є більша витрата ресурсів. Однак, оскільки організації часто не знають, чому якість даних важлива, 65% підприємств чекають, доки виникнуть проблеми з даними, перш ніж шукати рішення. Таким чином вони витрачають значну кількість праці та часу [6].

Нарешті, погане обслуговування на основі неправильних даних призводить до неправильного прийняття рішень і, отже, до низької якості продукції. В результаті послуги не будуть відповідати очікуваним стандартам якості, тому вся важка робота, час та вкладена праця можуть бути практично непотрібними [6].

Розглянемо проблеми з якістю великих даних.

П’ять основних характеристик великих даних (різноманітність, обсяг, цінність, швидкість і достовірність), хоч і важливі, також призводять до проблем у вимірюванні якості великих даних. Оскільки обсяг даних великий, складно підтримувати якість даних протягом заданого часу. Також важко інтегрувати дані через кілька форматів даних.

Управління підприємством для великих даних. Різні організації мають різні потреби у даних, тому всім їм потрібні свої методи обробки даних. Їм

також необхідно мати власні методи управління великими даними та забезпечення їх якості. Погане управління в будь-якій із цих областей призведе до поганої якості даних.

Обробка та обслуговування великих даних. Сюди входять такі фактори, як збір даних, перетворення даних, масштабованість служби даних та перетворення даних [6]. Через великий обсяг великі дані створюють проблеми з точки зору збору і перетворення даних. Зрештою, це призводить до неякісної організації даних.

1.2 Аналіз відомих інструментів оцінювання якості великих даних

З появою великих даних і сенсорних мереж велика увага приділяється якості даних датчиків. На сьогоднішній день проведено багато досліджень загальних параметрів якості великих даних [1, 6, 20,].

Laranjeiro, Soydemir і Bernardino [11]), а також Clarke [3] і Loshin [13] відзначили різні параметри та визначення якості великих даних.

Cai і Zhu [2] описує підходи до системи показників, які можна використовувати для вимірювання якості великих даних. Більш того, організації придумали їх визначення, моделі або методи вимірювання або прогнозування якості.

Дослідження щодо якості даних (наприклад, Cai і Zhu) проводяться з 1950-х років. Галузеві експерти запропонували багато визначень і параметрів якості даних.

Група з Массачусетського технологічного інституту, Total Data Quality Management, провела масштабне дослідження в цій галузі. Вони дослідили та визначили чотири основні категорії, які містять близько п’ятнадцяти параметрів якості даних.

Стаття [6] представляє корисні ідеї щодо забезпечення якості великих даних, включаючи відповідні проблеми та потреби. Він розглядає, наскільки якість великих даних є такою ж, як і звичайні дані, способи перевірки якості

великих даних та інші ключові фактори. Він визначає параметри якості, такі як точність, актуальність, своєчасність, правильність, послідовність, зручність використання, повнота, доступність, звітність і масштабованість. Він також описує процес перевірки якості великих даних і пропонує всебічне дослідження факторів, які викликають проблеми з якістю великих даних.

Процес перевірки великих даних.

Розглянемо п’ять основних сервісів великих даних, які застосовуються у процесі перевірки великих даних:

– збір даних;
– очищення даних;
– перетворення даних;
– завантаження даних;
– аналіз даних.

Збір даних – це процес накопичення даних і обчислення різноманітної інформації щодо важливих змінних, що покращує розуміння даних, що призводить до кращого прийняття рішень.

Очищення даних – це, як випливає з назви, процес пошуку пошкоджених або неточних даних і їх виправлення.

Перетворення даних – перетворює формат даних із вихідної системи даних у формат цільової системи даних..

Завантаження даних – це процес, у якому дані завантажуються у великі сховища даних. Залежно від вимог організації цей процес значно варіюється.

Аналіз даних стосується процесу виконання всіх раніше обговорених послуг великих даних, таких як збір, очищення, перетворення та завантаження з головною метою прийняття кращих рішень і отримання більшої інформації про самі дані. Агрегація даних стосується збору інформації з баз даних з метою підготовки комбінованих наборів даних для обробки.

Інструменти перевірки якості великих даних.

Програмне забезпечення MS-Excel, що є частиною офісного пакету Microsoft, є інструментом очищення та перевірки даних. Його можна

використовувати для перевпорядкування та форматування даних для аналізу. Можна також використовувати його для створення діаграм і графіків, які можуть добре ілюструвати дані. MS-Excel підтримує CSV, XLSX та інші формати даних. Однак, незважаючи на хорошу роботу з невеликими обсягами даних, Excel не може працювати з великими даними.

Zoho Reports – це служба онлайн-звітів і бізнес-аналітики. Це рішення для обробки великих даних і аналітики, яке дозволяє користувачам створювати глибокі звіти та інформаційні панелі. Це інструмент платформи SaaS, який дуже простий у використанні.

DataCleaner – це інструмент із відкритим вихідним кодом для якості даних, зберігання даних, профілювання даних, керування основними даними, бізнес-аналітики та корпоративного управління ефективністю. Він сумісний з кількома платформами, такими як Windows, Linux і IOS. Основна увага приділяється підключенню Apache Hive та Apache HBase. Він може підтримувати дані з файлів TXT, CSV і TSV, а також таблиць реляційних баз даних, листів MS Excel, MongoDB і Couch DB.

Основні функції DataCleaner:

– виявлення дублікатів на основі принципів машинного навчання;

– перевірка цілісності між кількома таблицями за один крок;

– профілювання та аналіз бази даних за лічені хвилини, але він повільніший порівняно з іншими інструментами перевірки великих даних;

– служить ефективним і плановим монітором справності даних.

QuerySurge – це інструмент для тестування великих даних, ETL і сховищ даних. Він знаходить пошкоджені дані та дає уявлення про стан даних.

Splunk – це провідний інструмент оперативної аналітики. Клієнти використовують цей інструмент для моніторингу, пошуку, аналізу та візуалізації даних. Він може генерувати графіки, візуалізації, звіти та створювати інформаційні панелі. Splunk простий у використанні та працює як з неструктурованими, так і зі структурованими даними. Він доступний як у вигляді програмного забезпечення, так і у вигляді хмарного сервісу.

Talend – це основний інструмент перевірки даних з відкритим кодом. Він складається з різних модулів, таких як інтеграція великих даних, хмарна інтеграція та інтеграція додатків. Він працює в Hadoop і Spark. Він підтримує кілька операційних систем, включаючи Windows, Linux і Mac OS. Він імпортує дані з реляційних баз даних, NoSQL і з файлів CSV. Він також виконує численні перевірки якості даних і створює графіки, аналізуючи певні критерії.

Tableau є провідним інструментом бізнес-аналітики та аналізу даних. Він може підключатися до різних джерел даних, таких як файли CSV, Cloudera Hadoop, MySQL і аналітика Google. Він має функції перевірки типу даних, відповідності та діапазону. Фільтри даних можна застосовувати, а клієнти також можуть створювати власні фільтри. Він простий у використанні, а наявність діаграм і графіків дозволяє чітко аналізувати дані.

Pentaho – це платформа для інтеграції великих даних і бізнес-аналітики. Вона складається з багатьох інструментів, таких як інтеграція даних, вбудована аналітика, бізнес-аналітика, хмарна бізнес-аналітика, аналітика Інтернету речей тощо. Цей продукт інтеграції даних надає клієнтам точні дані з будь-якого джерела даних. Pentaho має механізм паралельної обробки, який забезпечує високу продуктивність і масштабованість. Він надає вбудовані відладчики для тестування та виконання завдань. Він має вбудовану бібліотеку з компонентами, які використовуються для перетворення та перевірки даних.

1.3 Аналіз відомих методів оцінювання якості великих даних

Для проведення якісної оцінки великих даних слід дотримуватися належної методології. Cai і Zhu [2] пропонують один із таких механізмів. Ця модель (показана на рисунку 1.1) визначає мету збору даних і визначає параметри. На основі цих параметрів останнім кроком є вибір різних індикаторів оцінки, кожен з яких вимагатиме власних інструментів і методів.

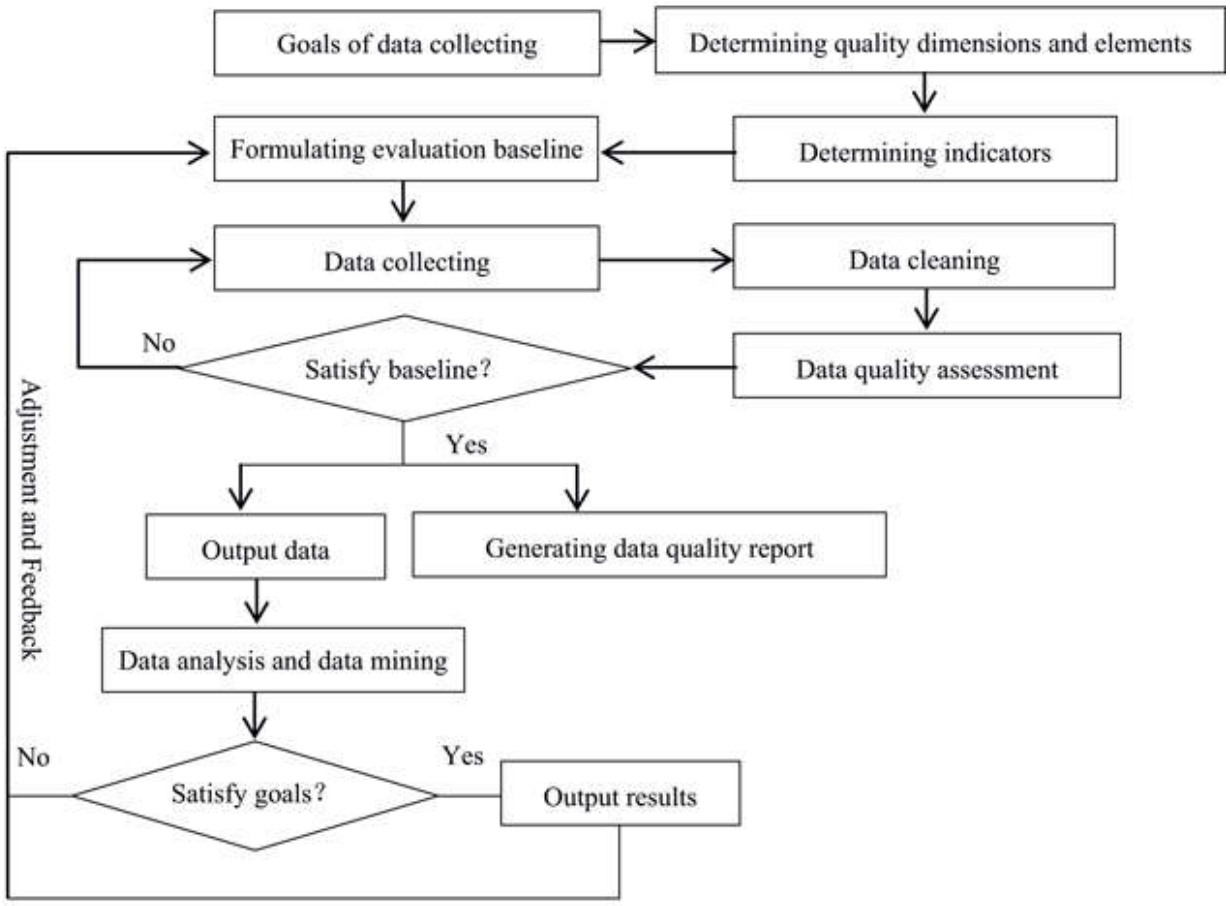


Рисунок 1.1 – Процес оцінки якості великих даних [2]

Після збору всієї необхідної інформації для оцінки даних дані збираються та очищаються. Потім проводиться оцінка якості даних шляхом порівняння результатів з вихідним рівнем початкових цілей. На основі результатів або створюється звіт про якість, або весь процес від «формулювання базової лінії оцінки» повторюється.

Структура якості даних. Gudivada та ін. [7] пропонують структуру якості даних (DQF), показану на рисунку 1.2.

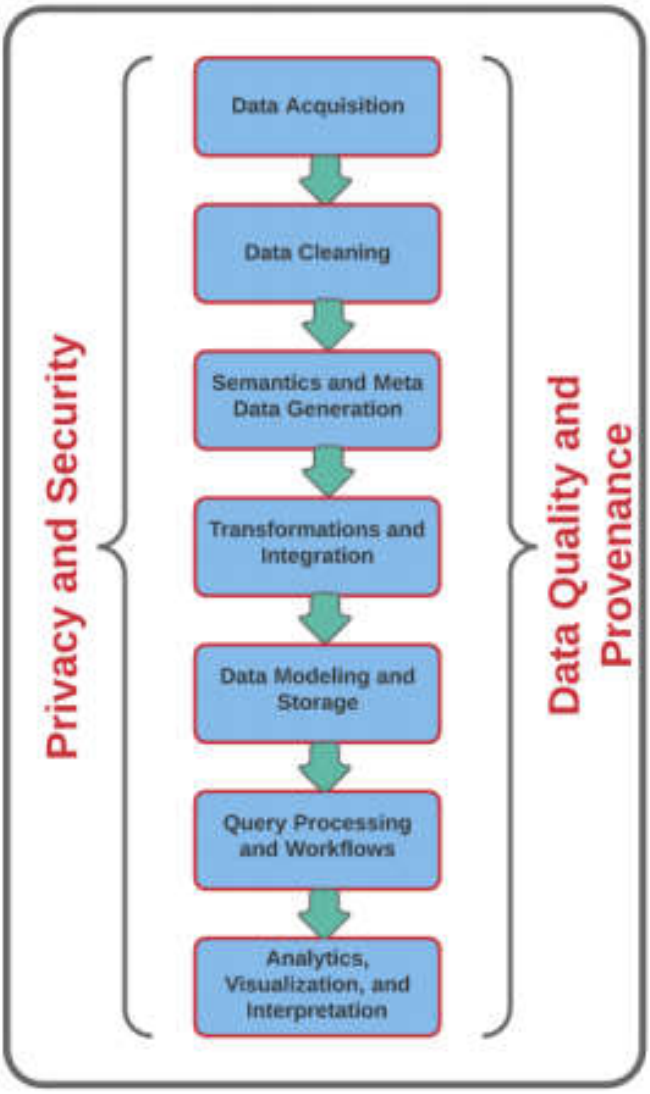


Рисунок 1.2 – Структура якості даних [7]

У представленому рисунку процес починається зі збору даних і супроводжується очищенням даних. На третьому етапі генеруються семантика та метадані. Тут неструктуровані дані, такі як зображення, графіка, аудіо, відео та твіти, перетворюються на напівструктуровані дані. На наступних етапах трансформації та інтеграції даних відбувається моделювання даних, обробка запитів, аналітика та візуалізація.

Після порівняння моделей Cai and Zhu (2015) і Gudivada та ін. [7] можна побачити, що більшість фаз однакові. Що відрізняється, так це часові рамки. У Cai і Zhu [2] збір даних відбувається на набагато пізнішому етапі. Тоді як у

Gudivada та ін. [7], першим кроком є збір даних. Cai і Zhu [2] підкреслюють важливість прийняття корисних рішень для підтримки якості на ранній стадії.

На поточному рівні запропонованого рішення бракує моделей якості великих даних, які можна було б застосовувати на основі параметрів.

Перевірка якості великих даних здійснюється для оцінки якості даних, щоб переконатися, що вони мають високу якість. Згідно Ludo [14], дані мають високу якість, якщо вони придатні для використання за призначенням у роботі, прийнятті рішень і плануванні. Високоякісні дані є точними, доступними, повними, послідовними, достовірними, такими, що піддаються обробці, актуальними та своєчасними. Виходячи з наведеного вище визначення високоякісних даних, ця теза спирається на вісім параметрів якості (рисунок 1.3), які використовуватимуться для перевірки стандартів якості для великих даних:

– повнота: чи всі необхідні значення доступні в наборі даних?
– точність: чи дані точно описують події чи об’єкти?
– своєчасність: чи надходять дані в очікуваний час?
– унікальність: чи є надмірність у наборі даних?
– валідність: чи відповідають дані певним правилам?
– узгодженість: чи є якісь протиріччя в даних?
– надійність вимірювального приладу/датчика: чи надійний стан машини, що збирає дані?
– зручність використання: чи відповідають дані заданим потребам?

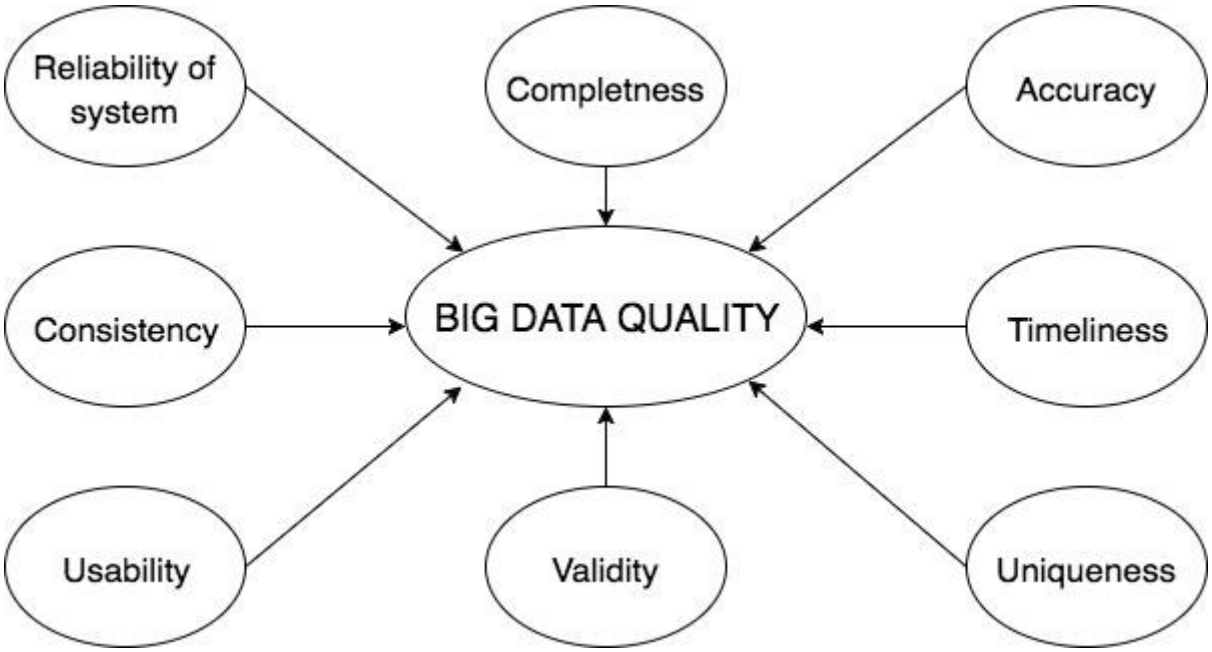


Рисунок 1.3 – Параметри якості великих

Повнота великих даних – це міра кількості доступних даних у порівнянні з бажаною кількістю для їх цільового призначення. Повнота використовується для перевірки, чи вплинуть недоліки в даних на зручність їх використання. Повноту великих даних можна визначити як частку збережених даних проти 100% повних даних [1]. Для вимірювання повноти використовується кількість доступних значень у заданому наборі даних і обчислюється співвідношення із загальною очікуваною кількістю значень. Одиницею вимірювання є відсоток.

Точність великих даних можна визначити як ступінь, до якого дані правильно описують об’єкт або подію «реального світу», які беруться до уваги [1]. Щоб виміряти точність набору даних або елемента даних, дані порівнюються з даними «реального світу». Зазвичай використовують довідкові дані третьої сторони, які, як правило, вважаються надійними та мають такий же тип [1]. Одиницею вимірювання є відсоток записів даних, які відповідають вимогам до точності даних. У деяких випадках точність легко виміряти, наприклад, розрізняючи стать (тобто чоловічу чи жіночу). Інші випадки можуть бути не так чітко диференційовані, що ускладнює вимірювання точності.

Точність допомагає відповісти на запитання, наприклад, чи є надані дані точними, чи викликають вони двозначність і чи відображають вони реальний стан джерела даних.

Своєчасність великих даних є важливим фактором для оцінки якості великих даних, оскільки дані змінюються щосекунди. Своєчасність великих даних вимірюється ступенем даних, які представляють реальність у потрібний момент часу [1]. Щоб виміряти своєчасність, позначають різницю в часі між моментом, коли відбувається подія, та моментом її реєстрації. Іншими словами, це різниця між тим, коли очікуються дані про час, і коли вони готові для використання. Одиницею вимірювання є відсоток різниці в часі. Своєчасність допомагає визначити, чи дані надійшли вчасно та чи регулярно оновлюються дані.

Унікальність великих даних визначається як вимірювання елемента даних відносно самого себе або його аналога в іншому наборі даних або базі даних [1]. Одиницею вимірювання є відсоток. Цей параметр використовується для підтвердження того, що набір даних не має повторюваних значень. У великих даних перевірка цього фактора допомагає усунути надмірності.

Достовірність великих даних або правильність даних. Дані дійсні, якщо вони відповідають синтаксису (формат, тип і діапазон) своїх визначень [1]. Щоб виміряти валідність, дані порівнюються з визначеними для них дійсними правилами. Одиницею вимірювання є відсоток. Це допомагає дізнатися, чи дійсні дані для використання за призначенням чи ні. Ця теза моделює дійсність на рівнях транзакції та параметрів.

Послідовність великих даних означає, наскільки логічний зв’язок між корельованими даними є правильним і повним [2]. Askham та ін. [1] визначають узгодженість як відсутність різниці при порівнянні двох або більше представлень однієї речі. Щоб виміряти узгодженість, один вимірює елемент даних проти самого себе або його аналога в іншому наборі даних [1]. Припустімо, що однакові дані надходять на дві різні станції, надходячи з кількох шляхів і накопичуючись на базовій станції. Для узгодженості, обидва

набори даних повинні мати однакове значення. З цієї причини необхідно перевірити узгодженість між ними. Ця теза моделює цінність і узгодженість даних у часі.

Надійність системи великих даних визначається як здатність мережі забезпечувати надійну передачу даних у стані постійної зміни структури мережі [12]. Щоб виміряти надійність системи, потрібно визначити, чи компонент або система належним чином працює відповідно до своїх специфікацій протягом певного часу. Датчики перевіряють на надійність.

Зручність використання великих даних можна визначити корисність даних та їх відповідність потребам користувачів [1]. Щоб виміряти зручність використання, обчислюють своєчасність, точність і повноту, оскільки значення цих трьох параметрів якості визначає, придатні дані для використання чи ні. Одиницею вимірювання є відсоток.

1.4 Постановка задачі дослідження

Важливим для отримання достовірних та якісних результатів при використанні інформаційних технологій є не тільки методи, способи та засоби їх отримання, але і якість початкових даних. Вони мають ряд характеристик, від яких суттєво залежить результат. До таких характеристик, зокрема, можуть належати повнота, точність, змістовність тощо. Певні з цих характеристик можуть бути більш вагомими, інші ні. Але в сукупності вони дають змогу оцінити отриманий результат, як такий, що базується на якісних даних. В процесі застосування інформаційних технологій можуть виникати проблеми через наявність неповних або надлишкових даних. Тоді, потрібна оцінка якості цих даних, бо вони впливатимуть на отриманий результат.

Сучасні інформаційні технології обробки даних отримують на вході, як правило багато різнотипних даних і їх великі обсяги. Кріт того, ІТ побудовані з використанням методів, які спрямовані на обробку таких великих даних, а також при їх імплементації використовуються складні алгоритми. Зокрема,

наприклад для глибокого навчання з використанням штучних нейронних мереж потрібні досить якісні дані, бо як правило, нейронним мережам потрібно більше даних для створення більш потужних абстракцій. Хоча в сценаріях з великими даними міститься багато даних, вони не завжди правильні або не завжди мають досить високу якість для навчання. Невеликі варіації або несподівані особливості початкових даних, а особливо неповнота даних можуть повністю розбалансувати налаштування в моделях нейронних мереж.

Враховуючи такі проблеми з початковими даними великих обсягів, які використовуються в ІТ, що побудовані на основі сучасних методів, зокрема і інтелектуальних та еволюційних, необхідним є початкова оцінка якості великих даних.

Тому, метою кваліфікаційної роботи є дослідження моделей оцінювання якості великих даних та розробка методу оцінювання якості великих даних.

Завдання кваліфікаційної роботи:

– дослідження актуальності задачі оцінювання якості великих даних;

– аналіз відомих інструментів оцінювання якості великих даних;

– аналіз відомих методів оцінювання якості великих даних;

– постановка задачі дослідження;

– визначення параметрів якості великих даних;

– визначення дослідницьких наборів для оцінки параметрів якості великих даних;

– дослідження моделей оцінювання якості великих даних;

– визначення узагальненого результату параметрів якості даних;

– проведення експериментальних досліджень запропонованих рішень.

Висновки до розділу 1

1. Важливим для отримання достовірних та якісних результатів при використанні інформаційних технологій є не тільки методи, способи та засоби їх отримання, але і якість початкових даних. Вони мають ряд характеристик, від яких суттєво залежить результат. До таких характеристик, зокрема, можуть належати повнота, точність, змістовність тощо. Певні з цих характеристик можуть бути більш вагомими, інші ні. Але в сукупності вони дають змогу оцінити отриманий результат, як такий, що базується на якісних даних. В процесі застосування інформаційних технологій можуть виникати проблеми через наявність неповних або надлишкових даних. Тоді, потрібна оцінка якості цих даних, бо вони впливатимуть на отриманий результат.

2. Сучасні інформаційні технології обробки даних отримують на вході, як правило багато різнотипних даних і їх великі обсяги. Кріт того, ІТ побудовані з використанням методів, які спрямовані на обробку таких великих даних, а також при їх імплементації використовуються складні алгоритми. Зокрема, наприклад для глибокого навчання з використанням штучних нейронних мереж потрібні досить якісні дані, бо як правило, нейронним мережам потрібно більше даних для створення більш потужних абстракцій. Хоча в сценаріях з великими даними міститься багато даних, вони не завжди правильні або не завжди мають досить високу якість для навчання. Невеликі варіації або несподівані особливості початкових даних, а особливо неповнота даних можуть повністю розбалансувати налаштування в моделях нейронних мереж.

3. Враховуючи проблеми з початковими даними великих обсягів, які використовуються в ІТ, що побудовані на основі сучасних методів, зокрема і інтелектуальних та еволюційних, необхідним є початкова оцінка якості великих даних.

2 МОДЕЛІ ОЦІНЮВАННЯ ЯКОСТІ ВЕЛИКИХ ДАНИХ

2.1 Параметри якості великих даних

Розглянемо особливості якості великих даних за їх властивостями та характеристиками. Це необхідно для забезпечення якості великих даних на основі оцінки їх якості для того, щоб переконатися, що вони можуть бути використаними як початкові дані. Придатність даних для використання за призначенням в процесі експлуатації, прийняття рішень та планування можливе за умови їх високої якості. Якісні дані характеризуються такими параметрами: точність; доступність; повнота; послідовність; достовірність; обробність; актуальність; своєчасність.

Введемо для відображення характеристик якості даних множину $\mathfrak{W}$, елементами якої будуть параметри якості даних. Тоді, множина $\mathfrak{W} = \{p_1, p_2, p_3, p_4, p_5, p_6, p_7\}$, в якій задано сім параметрів якості для перевірки стандартів якості великих даних. Кількість параметрів може змінюватись. Задамо значення восьми параметрів якості. Вагу значення кожного з них на цьому етапі не розглядатимемо.

Нехай p_1 відповідає повноті даних, p_2 – точності, p_3 – своєчасності, p_4 – унікальності, p_5 – дійсності, p_6 – послідовності, p_7 – надійності, тоді таку формалізацію початкових даних великого обсягу можна представити в моделі оцінки великих даних.

Зміст введених параметрів якості великих даних є важливим і не повинен відноситись до декількох параметрів одночасно, а відповідно за відношенням «один до одного». Таким чином, зміст повноти означає, всі потрібні значення є доступними в множині значень. Зміст точності пов’язаний з відображенням точності опису даними події чи об’єкту. Надходження даних у передбачуваний час визначає зміст своєчасності. Наявність надмірності у множині значень даних відповідає за зміст унікальності, тобто відсутності повторюваних значень або корельованих між собою груп значень. Зміст дійсності вказує на необхідність відповідності даним певним правилам.

Суперечність в даних відповідає змісту послідовності. Надійність стану машинного збору даних відповідає змісту надійності. Відповідність даних певним потребам відповідає змісту потреб. Розглянемо суть введених параметрів детальніше.

Параметр p_1, що відповідає повноті даних - це показник кількості доступних даних по відношенню до бажаної кількості. Він використовується для підтвердження того, як недоліки даних вплинуть на їхню корисність. Повноту великих даних можна визначити як частку збережених даних проти потенціалу 100% повних даних. Для вимірювання повноти даних використаємо кількість доступних значень у даному наборі даних та обчислимо її відношення до загальної передбачуваної кількості значень у відсотках.

Параметр p_2, що відповідає точності великих даних можна визначити як ступінь, в якому дані правильно описують об’єкт чи реальний світ, що враховується. Для вимірювання точності набору даних або елементу даних дані порівнюються з еталонами або даними, що є загально використовуваними чи прийнятими. Одиницею виміру є відсоток записів даних, які відповідають вимогам точності даних. У деяких випадках точність можна виміряти, якщо є повні відомості про вичерпну множину об’єктів певної предметної області, данні про яких становлять повну групу подій. Але не завжди є можливість здійснити встановлення точності даних, особливо великих і різнорідних. Точність великих даних показує у відсотках відображення реального стану джерела даних.

Своєчасність великих даних, задана параметром p_3, відноситься до критичних параметрів оцінки якості великих даних, бо вони можуть змінюватись душе швидко і при нехтуванні важливістю цього параметру втрачається актуальність результату. Своєчасність великих даних вимірюється ступенем даних, які представляють реальність у необхідний момент часу. Щоб виміряти своєчасність, позначають різницю в часі між тим, коли подія відбувається та коли вона записана. Іншими словами, це різниця

між часом, коли очікуються дані часу, і часом, коли вони легко доступні для використання. Одиницею виміру є відсоток різниці в часі. Своєчасність допомагає визначити, чи дані надходять вчасно і чи регулярно оновлюються дані.

Унікальність великих даних, яку задано параметром p_4, визначається як вимірювання елемента даних щодо самого себе або його аналогу в іншому наборі даних або базі даних. Одиницею виміру є відсоток, який використовується для підтвердження того, що набір даних не має повторюваних значень. У великих даних перевірка цього показника допомагає усунути надмірності.

Дійсність великих даних, задана параметром p_5, вказує на відповідність синтаксису, тобто коректності даних, їх формату, типу та діапазону. Для вимірювання достовірності порівнюють дані з дійсними правилами, визначеними для них. Одиницею виміру є відсоток, що показує відповідність дійсності даних цільовому використанню.

Параметр p_6, що відповідає за узгодженість великих даних стосується того, наскільки логічний зв'язок між корельованими даними є правильним і повним, тобто ним визначають відсутність різниці при порівнянні двох або більше даних з події чи об'єкту. Для виміру параметру узгодженості даних здійснюють вимірювання елементу даних стосовно об'єкту чи події та його аналогу в іншому наборі даних. Припустимо, одні й ті самі дані надходять на дві різні станції, надходячи з декількох шляхів і накопичуючись на базовій станції. Для узгодженості обидва набори даних повинні мати однакове значення та однакове значення. З цієї причини необхідно перевірити узгодженість між ними.

Параметр p_7, що відповідає за надійність системи великих даних, визначається як здатність мережі забезпечувати надійну передачу даних у стані постійної зміни структури мережі або здатність пристрою здійснювати надійну видачу даних. Для вимірювання надійності даних системи характеризують на предмет правильної роботи її компонент або відповідно до

її специфікацій протягом певного часу. Також, здійснюється перевірка датчиків, щоб визначити їх надійність.

2.2 Дослідницькі набори для оцінки параметрів якості великих даних

З метою визначення моделей якості великих даних можна досліджувати два набори даних. До цих наборів даних віднесемо очікувані та отримані набори даних. Їх можна використати в якості дослідницьких наборів, щоб оцінити параметри якості великих даних.

Розглянемо n станцій в комп'ютерній мережі. Представимо їх множиною $M = \{m_1, m_2, \dots, m_n\}$. Задамо кожним m_i, де $i = 1, 2, \dots, n$, джерело даних у комп'ютерній станції. Припустимо, від джерела даних m_i очікується, що множина даних надходить із кількістю транзакцій n, і кожна транзакція складається з кількості параметрів r. Крім того, джерело даних m_i отримує набір даних із числом транзакцій, і кожна транзакція має кількість параметрів. Нехай Q є очікуваною множиною даних. Представимо цю множину даних матрицею таким чином:

$$Q = \begin{pmatrix} Q_{1,1} & Q_{1,2} & \cdots & Q_{1,n} \\ Q_{2,1} & Q_{2,2} & \cdots & Q_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ Q_{r,1} & Q_{r,2} & \cdots & Q_{r,n} \end{pmatrix}, \quad (2.1)$$

де $Q_{i,j}$ є значенням для *i*-ї транзакції та *j*-го параметра.

Введемо для результуючого (отриманого) набору даних на певний момент часу матрицю W, таким чином, щоб на перетині для *i*-ї транзакції та *j*-го параметра знаходилось результуюче значення. Для кількості результуючих значень введемо такі позначення: n_W- кількість отриманих транзакцій, r_W - загальна кількість отриманих параметрів на транзакцію. Відмінності між матрицями Q та W полягають в тому, що матриця Q сформована з очікуючих значень, тобто властивість повноти для неї виконується, а матриця W містить

таку саму кількість елементів або меншу, що пов'язано з реальним отриманням даних від джерел даних. Враховуючи такі початкові умови визначення матриці W отримуємо співвідношення: $n_W \leq n, r_W \leq r$. В такому випадку матрицю W задамо так:

$$W = \begin{pmatrix} W_{1,1} & W_{1,2} & \dots & W_{1,n_W} \\ W_{2,1} & W_{2,2} & \dots & W_{2,n_W} \\ \dots & \dots & \dots & \dots \\ W_{r_W,1} & W_{r_W,2} & \dots & W_{r_W,n_W} \end{pmatrix}, \quad (2.2)$$

де $W_{i,j}$ представляє значення, які вже отримані на певний момент часу, для *i*-ї транзакції та *j*-го параметра.

Для вимірювання параметрів якості даних необхідно розрахувати загальну кількість значень для очікуваних та отриманих наборів даних. Введемо позначення s_Q для визначення загальної кількості значень для очікуваних та отриманих наборів даних. Тобто s_Q – це загальна кількість очікуваних елементів з n транзакціями, і кожна транзакція має r параметрів. Тоді, s_Q визначається так:

$$s_Q = n \times r. \quad (2.3)$$

Аналогічно задамо загальну кількість s_W отриманих елементів із n_W транзакціями, де кожна транзакція має r_W параметрів. Таким чином, s_W визначається так:

$$s_W = n_W \times r_W. \quad (2.4)$$

В результаті такого представлення значень початкових і результуючих даних отримано моделі, які використовуватимуться в подальшому для визначення якості великих даних. Для використання цих моделей спочатку потрібно визначення кожного параметра якості. При цьому можуть бути

враховані різні вимірювання за часом, мітки часу транзакцій, кількість транзакцій на день та інтервали між транзакціями. Такі вимірювання дозволять розрахувати значення за день, місяць та рік для різних параметрів.

Для розгляду моделей якості великих даних потрібно враховувати відсоток завершених транзакцій, тобто за розглядуваний час частина транзакцій завершиться, а частина не завершиться. Результат отриманий за цим параметром суттєво впливає на повноту даних. Тому, проаналізуємо параметр повноти великих даних з точки зору транзакції. Щоб визначити повноту, необхідно знати, скільки даних слід вважати набором даних як завершений. Можна встановити загальну кількість очікуваних даних, використовуючи формулу (2.3). Крім того, доповнююча кількість даних до повної визначається за загальною кількістю відсутніх значень у великих даних. Набір даних Q - початковий набір даних. Введемо загальну кількість s_V відсутніх даних у результуючому наборі даних W із n_W - транзакціями та параметрами r_W. Також, у результуючому наборі даних W можуть бути нульові значення. Нехай n_Z - це загальна кількість нульових значень у отриманому наборі даних. Отже, s_V для отриманого набору даних W може бути заданий наступним чином:

$$s_V = s_Q - s_W + n_Z, \tag{2.5}$$

де s_Q, s_W отримані з формул (2.3) і (2.4) відповідно;

n_Z - кількість нульових значень у наборі даних W.

Згідно формули (2.5) для отримання загальної кількості відсутніх значень загальна кількість отриманих значень віднімається від кількості очікуваних значень. Нульові значення додаються до загальної кількості відсутніх значень, інакше не будуть враховані не значущі значення, які отримані на певний момент часу.

2.3 Моделі оцінювання якості великих даних

Для того, щоб виміряти повноту кожної транзакції, яка є частиною в загальній оцінці якості даних, здійснимо обчислення при $r = 1$ для набору даних Q формули (2.2) та $r_W = 1$ для набору даних W формули (2.3). Введемо величину s_{P_i}, як відповідатиме за повноту транзакції для набору даних матриці W та номера транзакції i. Індекс нижнього індексу означає, що повнота вимірюється з точки зору транзакції. Значення s_{P_i} визначимо так:

$$s_{P_i} = \frac{s_Q - s_V}{s_Q}, \tag{2.6}$$

де s_Q і s_V отримані з формул (2.3) і (2.5) відповідно.

У формулі (2.6) загальна кількість відсутніх даних s_V у наборі даних віднімається із загальної очікуваної кількості даних s_Q. Це ціле значення дає фактичну кількість елементів, доступних у наборі даних W. Ділення цього віднімання на s_Q дає коефіцієнт повноти.

Для визначення відсоткового значення повноти в формулі (2.7) введемо його позначення $s_{P_i}\%$ для номера транзакції i. Визначення відсоткового значення повноти здійнюватимемо за формулою:·

$$s_{P_i}\% = s_{P_i} \cdot 100\ \%, \tag{2.7}$$

де значення s_{P_i}, яке отримане за формулою (2.6).

Щоб визначити вимірювання параметрів якості даних за день, необхідно визначити загальну кількість транзакцій за день. Введемо величину $s_{P_{D,i}}$, яка відповідатиме за загальну кількість транзакцій за день i. Величина $s_{P_{D,i}}$ обчислюється з використанням різниці в часі між двома транзакціями та загальною кількістю годин транзакції за формулою:

$$s_{P_{D,i}} = \frac{t_S}{t_{S_D}}, \tag{2.8}$$

де t_S - різниця в часі між двома транзакціями;

t_{S_D} - загальна кількість годин, за які трансакції відбувались протягом дня.

Якщо ця величина $s_{P_{D,j}}$ представляє середню повноту дня j з точки зору транзакції, тоді визначаємо її за формулою:

$$s_{P_{D,j}} = \frac{\sum_{i=1}^{s_{P_{D,i}}} s_{P_i}\%}{s_{P_{D,i}}}, \tag{2.9}$$

де $s_{P_{D,i}}$ - транзакції за день j, отримані за формулою (2.8);

$s_{P_i}\%$ - відсоток повноти транзакції i, що обчислюється за формулою (2.7).

Сумування застосовується за відсотком повноти транзакції і для всіх значень i, які дорівнюють від 1 до кількості транзакцій за день $s_{P_{D,i}}$. Для обчислення повноти всіх транзакцій, що відбулися протягом дня j, значення підсумовування ділиться на $s_{P_{D,i}}$, отримуючи середню повноту транзакції за день j.

Введемо величину $N_{D,j}$ для фіксування кількості днів, коли транзакції відбувались у місяці j. А також, величину $N_{D,s,j}$, яка відповідатиме за середню повноту за місяць j з точки зору транзакції. Тоді, обчислення середньої повноти за місяць j з точки зору транзакції здійснимо за формулою:

$$N_{D,s,j} = \frac{\sum_{i=1}^{N_{D,j}} s_{P_{D,i}}}{N_{D,j}}, \tag{2.10}$$

де чисельник величини $N_{D,s,j}$ обчислюється сумуванням над $s_{P_{D,i}}$ для всіх значень i, які дорівнюють від 1 до $N_{D,j}$.

Для обчислення повноти усіх транзакцій, що відбулись протягом $N_{D,s,j}$ днів у місяці j, це підсумоване значення ділиться на $N_{D,i}$, щоб визначити середню повноту транзакції за місяць.

Нехай величина $N_{M,j}$ представляє кількість місяців, протягом яких відбувались операції в році j. Нехай величина $N_{M,s,j}$ представляє середню повноту за рік j з точки зору транзакції. Тоді, середню повноту за рік j з точки зору транзакції визначимо так:

$$N_{M,s,j} = \frac{\sum_{i=1}^{N_{M,j}} N_{D,s,i}}{N_{M,j}} , \tag{2.11}$$

де $N_{D,s,i}$ обчислюється за формулою (2.10).

Підсумовування застосовується для величини $N_{D,s,i}$ для всіх значень i, які дорівнюють від 1 до $N_{M,j}$ для обчислення повноти всіх операцій, що відбулись за $N_{M,s,j}$ місяців у році j. Це значення підсумовування, поділене на величину $N_{M,j}$, дає змогу обчислити середню завершеність операцій за рік.

Розглянемо детальніше величину p_1 - параметр, для якого обчислюється повнота. Отриманий набір даних Q представляє собою всі значення у параметрі p_1. За допомогою значення величини s_V з формули (2.5) обчислюємо загальну кількість відсутніх даних у отриманому наборі даних W із числом s_W транзакцій. Відповідно, для обчислення повноти необхідно знати обсяг даних, який, як очікується, вважатиме отриманий набір даних повним. За допомогою значення величини s_Q з формули (2.3) встановлюємо загальну кількість очікуваних значень у наборі даних.

Нехай $p_{1.i}$ - повнота кожного параметра для набору даних матриці W та параметра i. Індекс означає, що повнота вимірюється через параметр. Значення $p_{1.i}$ визначаємо за формулою:

$$p_{1.i} = \frac{s_Q - s_V}{s_Q}, \tag{2.12}$$

де величини s_Q та s_V обчислюються за формулами (2.3) і (2.5) відповідно.

Відсоткове значення величини повноти $p_{1.i}$ для параметра і визначаємо за формулою:

$$p_{1.i}\% = p_{1.i} \cdot 100\%, \tag{2.13}$$

де значення $p_{1.i}$ отримується за формулою (2.12).

Нехай $p_{1.D,j}$ представляє середню повноту дня j через параметр $p_{1.i}$, яку визначимо за формулою:

$$p_{1.D,j} = \frac{\sum_{i=1}^{s_{P_{D,j}}} p_{1.i}\%}{s_{P_{D,j}}}, \tag{2.14}$$

де $s_{P_{D,j}}$ - транзакції за день j, отримані за формулою (2.9);

$p_{1.i}\%$ - відсоткова повнота для параметра i, отриманого за формулою (2.13).

Підсумовування застосовується для величини $p_{1.i}\%$ для всіх значень i, які дорівнюють від 1 до $s_{P_{D,j}}$, для обчислення повноти всіх транзакцій, що відбулися протягом дня j. Це значення підсумовування ділиться на $s_{P_{D,j}}$, що дає змогу отримати середню повноту параметрів для дня j.

Нехай $p_{1.M,s,j}$ – середня повнота місяця j з точки зору параметра, тоді його обчислення здійснимо за формулою:

$$p_{1.M,s,j} = \frac{\sum_{i=1}^{N_{p_{1.D,j}}} p_{1.D,i}}{N_{p_{1.D,j}}}, \tag{2.15}$$

де $N_{p_{1.D,j}}$ – кількість днів, протягом яких відбулися транзакції в місяці j, а величина $p_{1.D,i}$ обчислюється за формулою (2.14).

Сумування застосовується до величин $p_{1.D,i}$ для всіх значень i, які дорівнюють від 1 до $N_{p_{1.D,j}}$, щоб обчислити повноту всіх транзакцій, що відбулися протягом $N_{p_{1.D,j}}$ днів у місяці j. Це значення підсумовування, поділене на $N_{p_{1.D,j}}$, дає середню повноту параметрів на місяць.

Нехай $p_{1.Y,s,j}$ представляє середню повноту за рік j через параметр. Тоді величина $p_{1.Y,s,j}$ обчислюється за формулою:

$$p_{1.Y,s,j} = \frac{\sum_{i=1}^{N_{p_{1.M,j}}} p_{1.M,s,i}}{N_{p_{1.M,j}}}, \tag{2.16}$$

де величина $N_{p_{1.M,j}}$ є кількістю місяців, протягом яких відбувалися транзакції в році j, а величина $p_{1.M,s,i}$ обчислюється за формулою (2.15), підсумовування застосовується по $Cp_{1.M,s,i}$ для всіх значень i, які дорівнюють від 1 до $N_{p_{1.M,j}}$, щоб обчислити повноту всіх операцій, що відбулися за $N_{p_{1.M,j}}$ місяців у році j. Поділивши на $N_{p_{1.M,j}}$, частка дає змогу отримати повноту параметрів на рік.

Розглянемо детальніше параметр точності з врахуванням його опису моделями в залежності від розглядуваних властивостей і ознак. Точність для кожної моделі транзакції перевіряє точність кожного елемента в одній транзакціях. Модель параметра точності p_2 перевіряє кожен елемент у параметрі під час усіх транзакцій протягом заданого часу. Обидва використовують відсоток як одиницю виміру.

Для розрахунку точності необхідний набір довідкових даних. Очікуваний набір даних, заданий матрицею Q, є еталонним набором даних для всіх розрахунків. Для розрахунку точності за транзакцією приймемо n = 1 у матриці Q, яку задано формулою (2.2), та n_W= 1 у матриці W, яку задано формулою (2.3). Отриманий набір даних дорівнює W. Відстань між обома вибраними наборами даних надає їх точність. Введемо величину N_{p_T}, яка

відповідатиме за максимальну кількість параметрів на транзакцію між посиланням та отриманими наборами даних.

Точність для транзакції визначимо величиною $p_{2,T}$ для транзакції k, де $p_{2,T,i,j}$ – різниця між еталонним та отриманим набором даних для транзакції i та параметром j. Обчислення точності здійснимо так:

$$p_{2,T,i,j} = 1, \text{якщо } Q_{i,j} - W_{i,j} = 0, \tag{2.17}$$

де i – кількість транзакцій;

j – кількість параметрів.

Введемо величину $p_{2,T,k}$, що відповідатиме за точність для транзакції k. Її обчислення здійснимо за формулою:

$$p_{2,T,k} = \frac{\sum_{j=1}^{N_{p_T}} p_{2,T,k,j}}{N_{p_T}}, \tag{2.18}$$

де $p_{2,T,k,j}$ обчислюються за формулою (2.17).

Підсумовування застосовується за $p_{2,T,k,j}$ для всіх значень j, що дорівнюють від 1 до кількості параметрів N_{p_T}.

Нехай $p_{2,T,i}\%$ – величина, що відображає відсоткову точність транзакції i. Обчислення здійснимо за формулою так:

$$p_{2,T,k}\% = p_{2,T,k} \cdot 100\%, \tag{2.19}$$

де значення $p_{2,T,k}$ обчислюються за формулою (2.18).

Нехай $p_{2,T,D,j}$ представляє середню точність дня j з точки зору транзакцій, що відбулися в день j. Визначимо величину $p_{2,T,D,j}$ за формулою:

$$p_{2,T,D,j} = \frac{\sum_{i=1}^{s_{P_{D,j}}} p_{2,T,i}\%}{s_{P_{D,j}}}, \quad (2.20)$$

де $s_{P_{D,j}}$ – транзакції за день j, отримані за формулою (2.8);

$p_{2,T,i}\%$ – відсоткова точність транзакції i, отримана за формулою (2.19).

Сумування застосовується за $p_{2,T,i}\%$ для всіх значень i, які дорівнюють від 1 до $s_{P_{D,j}}$, для обчислення точності для всіх транзакцій, що відбулися протягом дня j. Це сумування ділиться на $s_{P_{D,j}}$ і отримується середня точність параметрів для дня j.

Нехай величина $p_{2,T,M,j}$ – середня точність місяця j з точки зору транзакції. Тоді, величина $p_{2,T,M,j}$ обчислюється за формулою:

$$p_{2,T,M,j} = \frac{\sum_{i=1}^{N_{p_{2.D,j}}} p_{2,T,D,i}}{N_{p_{2.D,j}}}, \quad (2.21)$$

де $N_{p_{2.D,j}}$ – кількість днів, протягом яких відбулися транзакції в місяці j, а $p_{2,T,D,i}$ обчислюється за формулою (2.20).

Сумування застосовується для всіх значень $p_{2,T,D,i}$ для всіх значень i, які дорівнюють від 1 до $N_{p_{2.D,j}}$, щоб обчислити точність для всіх транзакцій, що відбулися протягом $N_{p_{2.D,j}}$ днів у місяці j. Це підсумоване значення, поділене на $N_{p_{2.D,j}}$, дає змогу обчислити середню точність транзакції на місяць.

Нехай величина $p_{2,T,Y,j}$ представляє середню точність за рік j з точки зору кожної транзакції. Тоді, величина $p_{2,T,Y,j}$ обчислюється за формулою:

$$p_{2,T,Y,j} = \frac{\sum_{i=1}^{N_{p_{2.M,j}}} p_{2,T,M,i}}{N_{p_{2.M,j}}}, \quad (2.22)$$

де значення $N_{p_{2.M,j}}$ – кількість місяців, протягом яких відбувались транзакції в році j; значення $p_{2,T,M,i}$ обчислюється за формулою (2.20).

Сумування застосовується по значеннях $p_{2,T,M,i}$ для всіх значень i, які дорівнюють від 1 до $N_{p_{2.M,j}}$, для обчислення точності для всіх операцій, що відбулися за $N_{p_{2.M,j}}$ місяців у році j. Це підсумоване значення, поділене на $N_{p_{2.M,j}}$, дає змогу обчислити середню точність транзакції на рік.

Для точності кожного параметра підставляємо m= 1 у формулу (2.2) та r_W= у формулу (2.3). Отриманим набором даних є W. Нехай набір даних очікуваного набору даних з матриці Q буде набором еталонних даних. Введемо для позначення кількості транзакцій величину $N_{p_2,T}$, що позначає максимальну кількість транзакцій на параметр між опорними та отриманими наборами великих даних. Визначаємо точність кожного параметра для транзакції $p_{2,T,k}$ за формулою:

$$p_{2,T,k} = \frac{\sum_{i=1}^{N_{p_{2.T,j}}} p_{2,T,i,k}}{N_{p_{2.T,j}}}, \tag{2.23}$$

де $p_{2,T,i,k}$ обчислюється за формулою (2.17);

$N_{p_{2.T,j}}$ - кількість параметрів на транзакцію.

Введемо величину $p_{2,T,k}\%$ для задання відсоткової повноти параметра i та визначимо її так:

$$p_{2,T,k}\% = p_{2,T,k} \cdot 100\%, \tag{2.24}$$

де значення $p_{2,T,k}$ обчислюється за формулою (2.23).

Введемо величину $p_{2,T,D,j}$ для представлення середньої точності дня j з точки зору параметра та обчислюватимемо його за формулою:

$$p_{2,T,D,j} = \frac{\sum_{i=1}^{sP_{D,j}} p_{2,T,i}\%}{sP_{D,j}}, \tag{2.25}$$

де значення $s_{P_{D,j}}$ – транзакції за день j обчислюються за формулою (2.8);

$p_{2,T,i}\%$ – відсоткова точність параметра i обчислюється за формулою (2.24).

Сумування застосовується за $p_{2,T,i}\%$ для всіх значень i, які дорівнюють від 1 до $s_{P_{D,j}}$, для обчислення точності для всіх транзакцій, що відбулися протягом дня j. Це значення підсумовування ділиться на $s_{P_{D,j}}$, що дає змогу отримати середню точність параметрів для дня j.

Введемо величину $p_{2,T,M,j}$ для представлення середньої точності за місяць j за параметром. Обчислюватимемо значення величини $p_{2,T,D,j}$ за формулою:

$$p_{2,T,M,j} = \frac{\sum_{i=1}^{N_{p_{2.D,j}}} p_{2,T,D,i}}{N_{p_{2.D,j}}}, \tag{2.26}$$

де значення $N_{p_{2.D,j}}$ – кількість днів, протягом яких відбулися транзакції в місяці j; значення $p_{2,T,D,i}$ обчислюється за формулою (2.25).

Це сумування застосовується над значеннями $p_{2,T,D,i}$ для всіх значень i, які дорівнюють від 1 до $N_{p_{2.D,j}}$, для обчислення точності $p_{2,T,M,j}$ для всіх операцій, що відбулися протягом всіх днів у місяці j. Це значення підсумовування, поділене на $N_{p_{2.D,j}}$, дає змогу обчислити середню точність параметрів на місяць.

Введемо величину $p_{2,T,Y,j}$ для представлення середньої точність за рік j через параметр i здійснимо її обчислення так:

$$p_{2,T,Y,j} = \frac{\sum_{i=1}^{N_{p_{2.M,j}}} p_{2,T,M,i}}{N_{p_{2.M,j}}}, \tag{2.27}$$

де значення величини $N_{p_{2.M,j}}$ представляє собою кількість місяців, протягом яких відбувались транзакції в році j; значення $p_{2,T,M,i}$ обчислюється за формулою (2.26).

Сумування застосовується над значеннями $p_{2,T,M,i}$ для всіх значень i, які дорівнюють від 1 до $N_{p_{2.M,j}}$, для обчислення точності для всіх операцій, що відбулися за всі місяці у році j. Це значення сумування, поділене на $N_{p_{2.M,j}}$, дає змогу обчислити середню точність параметрів на рік.

Згідно з введеним параметром своєчасності, потрібно виміряти різницю в часі між часом прибуття та отриманим часом. Щоб виміряти своєчасність, потрібно зберігати позначку часу для кожної транзакції. Нехай структура E представляє масив міток часу для часу початку та закінчення кожного запису. Задамо її так: $E = \{t_{1,Q}, t_{1,W}, t_{2,Q}, t_{2,W}, r_{2,r}, t_{m,Q}, t_{m,W}\}$, де зв'язок відображує собою очікуваний час прибуття транзакції i, а $t_{i,W}$ вказує фактично отриманий час транзакції i.

Введемо величину $p_{3,l,T,i}$, яка відповідатиме за параметр своєчасність здійснення транзакції i та обчислюватимемо її так:

$$p_{3,l,T,i} = 1, \text{якщо } t_{1,Q} - t_{1,W} = 0. \tag{2.28}$$

Нехай величина $p_{3,l,T,i}\%$ - відсоток своєчасності транзакції i та визначимо її так:

$$p_{3,l,T,i}\% = p_{3,l,T,i} \cdot 100\%, \tag{2.29}$$

де значення $p_{3,l,T,i}$ обчислюється за формулою (2.28).

Нехай величина $p_{3,D,T,j}$ представляє середню своєчасність для дня j з точки зору транзакції та обчислюватимемо її так:

$$p_{3,D,T,j} = \frac{\sum_{i=1}^{s_{P_{D,j}}} p_{3,l,T,i}\%}{s_{P_{D,j}}}, \quad (2.30)$$

де значення $s_{P_{D,j}}$ - транзакції за день j обчислюються за формулою (2.8);

$p_{3,l,T,i}\%$ - це відсоток своєчасності транзакції i та обчислюється за формулою (2.29).

Сумування застосовується до значень $p_{3,l,T,i}\%$ для всіх значень i, які дорівнюють від 1 до $s_{P_{D,j}}$, для обчислення своєчасності всіх транзакцій, що відбулися протягом дня j. Це значення сумування ділиться на $s_{P_{D,j}}$, даючи змогу отримати середню своєчасність параметра для дня j.

Введемо величину $p_{3,M,T,j}$, що визначатиме середню своєчасність місяця j з точки зору транзакції та обчислюватимемо її так:

$$p_{3,M,T,j} = \frac{\sum_{i=1}^{N_{p_{3.D,j}}} p_{3,D,T,i}}{N_{p_{3.D,j}}}, \quad (2.31)$$

де значення $N_{p_{3.D,j}}$ - кількість днів, протягом яких відбувались транзакції в місяці j; значення $p_{3,D,T,i}$ обчислюються за формулою (2.30).

Сума застосовується до усіх значень $p_{3,D,T,i}$ для всіх значень i, які дорівнюють від 1 до $N_{p_{3.D,j}}$, щоб обчислити своєчасність усіх транзакцій, що відбулися протягом $N_{p_{3.D,j}}$ днів у місяці j. Коли це підсумоване значення ділиться на $N_{p_{3.D,j}}$, то дає змогу визначити середню своєчасність транзакції на місяць.

Введемо величину $p_{3,Y,T,j}$ для представлення середньої своєчасності за рік j з точки зору транзакції та обчислюватимемо її так:

$$p_{3,Y,T,j} = \frac{\sum_{i=1}^{N_{p_{3.M,j}}} p_{3,M,T,i}}{N_{p_{3.M,j}}}, \quad (2.32)$$

де значення $N_{p_{з.M,j}}$ – кількість місяців, коли відбулися транзакції в році j; значення $p_{3,M,T,i}$ обчислюється за формулою (2.31).

Сумування застосовується до всіх значень $p_{3,M,T,i}$ для всіх значень i, які дорівнюють від 1 до $N_{p_{з.M,j}}$, для обчислення точності для всіх операцій, що відбулися за $N_{p_{з.M,j}}$ місяців у році j. Це значення підсумовування, поділене на $N_{p_{з.M,j}}$, дає змогу обчислити середню точність параметрів на рік.

Унікальність великих даних вимірюється порівнянням даних із їх аналогом у тому самому наборі даних для перевірки надмірності. Розглянемо параметр p_4 унікальності кожної транзакції, здійсненої за один день. Припустимо, що наявна одна транзакція, тоді щоб обчислити її унікальність, порівняємо її з іншими транзакціями.

Введемо параметр $p_{4,i}$ - унікальність транзакції i. Щоб визначити унікальність транзакції, порівняємо цю транзакцію з рештою транзакцій в наборі даних і визначатимемо її так:

$$p_{4,i} = 1, \text{якщо збіг не знайдено в наборі, інакше } p_{4,i} = 0. \quad (2.33)$$

Формула (2.34) визначає відсоткове значення унікальність транзакції i. Введемо величину $p_{4,i}\%$, яка відповідатиме відсотковій унікальності транзакції i та визначатиметься так:

$$p_{4,i}\% = p_{4,i} \cdot 100\%, \quad (2.34)$$

де значення $p_{4,i}$ – обчислюється за формулою (2.33).

Введемо параметр унікальності $p_{4,D,i}$, який представляє середню унікальність дня j з точки зору транзакції і його обчислення здійснимо так:

$$p_{4,D,i} = \frac{\sum_{i=1}^{sP_{D,j}} p_{4,i}\%}{sP_{D,j}}, \quad (2.35)$$

де $s_{P_{D,j}}$ – це транзакції за день j;

$p_{4,i}\%$ – це відсоткова унікальність транзакції i та обчислюються за формулою (2.34).

Сумування застосовується для всіх величин $p_{4,i}\%$ для всіх значень і, які дорівнюють від 1 до $s_{P_{D,j}}$, для обчислення унікальності для всіх транзакцій, що відбулися протягом дня j. Це значення підсумовування ділиться на $s_{P_{D,j}}$, що дає змогу отримати середню унікальність параметра для дня j.

Введемо величину $p_{4,M,i}$ – середня унікальність місяця j з точки зору транзакції та обчислюватимемо її так:

$$p_{4,M,T,j} = \frac{\sum_{i=1}^{N_{p_{4.D,j}}} p_{4,D,i}}{N_{p_{4.D,j}}}, \tag{2.36}$$

де $N_{p_{4.D,j}}$ - кількість днів, протягом яких відбулися транзакції в місяці j; значення величини $p_{4,D,i}$ обчислюємо за формулою (2.35).

Знаходження суми усіх значень $p_{4,D,i}$ здійснюється для всіх значень i, які дорівнюють від 1 до $N_{p_{4.D,j}}$, для обчислення унікальності для всіх транзакцій, що відбулися протягом $N_{p_{4.D,j}}$ днів у місяці j. Це значення підсумовування, поділене на $N_{p_{4.D,j}}$, дає змогу обчислювати середню унікальність транзакції на місяць.

Введемо величину $p_{4,Y,i}$, що представляє середню унікальність за рік j з точки зору унікальності та обчислюватимемо її так:

$$p_{4,Y,T,j} = \frac{\sum_{i=1}^{N_{p_{4.M,j}}} p_{4,M,T,i}}{N_{p_{4.M,j}}}, \tag{2.37}$$

де значення величини $N_{p_{4.M,j}}$ представляє собою кількість місяців, протягом яких відбувались транзакції в році j; значення величини $p_{4,M,T,i}$ обчислюється за формулою (2.36).

Сумування застосовується над всіма значеннями $p_{4,M,T,i}$ для всіх значень i, які дорівнюють від 1 до $N_{p_{4.M,j}}$, щоб обчислити унікальність для всіх операцій, що відбулись за for місяців у році j. Це значення підсумовування, поділене на $N_{p_{4.M,j}}$, дає змогу отримати середню унікальність транзакції на рік.

Визначення дійсності великих даних правильно припускає, що воно передбачає міру валідності. Важливим є наявність правила або синтаксису, за допомогою якого можна оцінювати точність. Розглянемо параметр дійсність p_5 на рівні транзакцій та параметрів.

Для перевірки даних мають бути визначені певні правила, які дозволяють вважати ці дані дійсними. Припустимо, отриманим набором даних є W із n_W транзакціями та загальною кількістю отриманих параметрів на транзакцію r_W. Для кожного параметра, присутнього в наборі даних W, введемо, критерії перевірки $B = \{b_1, b_2, \dots, b_n,\}$. Щоб визначити валідність за параметром, приймемо $n_W = 1$ згідно формули (2.3). Здійснимо перевірку валідності кожного елемента значення у параметрі, щоб визначити, для якого саме елементу потрібно обчислювати. Щоб перевірити параметр, кожне значення параметра вимірюється відповідно до його правил для перевірки достовірності.

Дійсність для кожного значення в наборі даних визначимо величиною $p_{5,i}$ для значення і та обчислимо так:

$$p_{5,i} = 1, \text{ якщо всі правила дійсності передані, інакше } p_{5,i} = 0, \quad (2.38)$$

Для обчислення всіх присутніх у параметрі складових здійснимо їх сумування для цього параметру так:

$$p_{5,i} = \frac{\sum_{j=1}^{n_W} p_{5,j}}{n_W}, \tag{2.39}$$

де значення n_W - загальна кількість транзакцій; значення величини $p_{5,j}$ обчислюється за формулою (2.38).

Введемо величину $p_{5,i}\%$, що визначатиме відсоток повноти для параметра *i* та обчислюватимемо її так:

$$p_{5,i}\% = p_{5,i} \cdot 100\%, \tag{2.40}$$

де значення $p_{5,i}$ обчислюється за формулою (2.39).

Введемо величину $p_{5,D,i}$ для представлення середньої достовірності дня *j* з точки зору параметра та обчислюватимемо її так:

$$p_{5,D,j} = \frac{\sum_{i=1}^{s_{P_{D,j}}} p_{5,i}\%}{s_{P_{D,j}}}, \tag{2.41}$$

де значення $s_{P_{D,j}}$ – це транзакції за день *j*;

значення $p_{5,i}\%$ – це валідність параметра і та обчислюються за формулою (2.40).

Сумування застосовується над всіма значеннями $p_{5,i}\%$ для всіх значень *i*, які дорівнюють від 1 до $s_{P_{D,j}}$, для обчислення дійсності для всіх транзакцій, що відбулися протягом дня *j*. Це значення підсумовування ділиться на $s_{P_{D,j}}$, отримуючи середню достовірність параметра для дня *j*.

Введемо величину $p_{5,M,j}$ для представлення середньої дійсності місяця *j* з точки зору параметра та обчислюватимемо її так:

$$p_{5,M,j} = \frac{\sum_{i=1}^{N_{p_{5.D,j}}} p_{5,D,i}}{N_{p_{5.D,j}}}, \tag{2.42}$$

де $N_{p_{5.D,j}}$ – кількість днів, протягом яких відбулися транзакції в місяці j; значення величини $p_{5,D,i}$ обчислюємо за формулою (2.41).

Сумування застосовується для всіх значень $p_{5,D,i}$ для всіх значень i, які дорівнюють від 1 до $N_{p_{5.D,j}}$, для обчислення дійсності $p_{5,M,j}$ для всіх транзакцій, що відбулись протягом $N_{p_{5.D,j}}$ днів у місяці j. Це значення підсумовування, поділене на $N_{p_{5.D,j}}$, дає змогу обчислювати середню дійсність параметра на місяць.

Введемо величину $p_{5,Y,j}$ для представлення середнього терміну дії року j через параметр та обчислюватимемо його так:

$$p_{5,Y,j} = \frac{\sum_{i=1}^{N_{p_{5.M,j}}} p_{5,M,,i}}{N_{p_{5.M,j}}}, \tag{2.43}$$

де значення величини $N_{p_{5.M,j}}$ представляє собою кількість місяців, протягом яких відбувались транзакції в році j; значення величини $p_{5,M,i}$ обчислюється за формулою (2.42).

Сумування здійснюється над всіма значеннями $p_{5,M,i}$ для всіх значень i, які дорівнюють від 1 до $N_{p_{5.M,j}}$, для обчислення дійсності для всіх операцій, що відбулись за $N_{p_{5.M,j}}$ місяців у році j. Це значення підсумовування, поділене на $N_{p_{5.M,j}}$, дає змогу обчислити середню дійсність параметра за рік.

Для вимірювання дійсності кожної транзакції необхідно мати правила дійсності або синтаксис для кожного значення транзакції. Це означає, що кожну величину потрібно порівнювати з її правилами. Припустимо, отриманим набором даних є W із n_W транзакціями, і кожна транзакція має параметр r_W. Для кожного значення в операції здійснюватимемо перевірку її дійсності відповідно до формули (2.38). А далі, здійснюватимемо сумування всіх значень, присутніх у транзакції. Тоді, величина $p_{5,T,i}$ для транзакції і обчислюватиметься так:

$$p_{5,T,i} = \frac{\sum_{j=1}^{n_W} p_{5,j}}{n_W}, \quad (2.44)$$

де значення n_W – загальна кількість параметрів на транзакцію; значення величини $p_{5,j}$ обчислюється за формулою (2.38).

Визначимо відсоткове значення величини $p_{5,T,i}$. Припустимо, що величина $p_{5,T,i}\%$ - відсотковий термін дії транзакції і та обчислюватимемо його так:

$$p_{5,T,i}\% = p_{5,T,i} \cdot 100\%, \quad (2.45)$$

де значення $p_{5,T,i}$ обчислюється за формулою (2.44).

Введемо величину $p_{5,D,T,i}$ для представлення середньої дійсності дня j з точки зору транзакції та обчислюватимемо її так:

$$p_{5,D,T,j} = \frac{\sum_{i=1}^{s_{P_{D,j}}} p_{5,T,i}\%}{s_{P_{D,j}}}, \quad (2.46)$$

де значення $s_{P_{D,j}}$ – це транзакції за день j, які визначаються за формулою (2.8);

значення $p_{5,T,i}\%$ - відсотковий термін дії транзакції і та обчислюються за формулою (2.45).

Сумування застосовується до всіх значень $p_{5,T,i}\%$ для всіх значень i, які дорівнюють від 1 до $s_{P_{D,j}}$, для обчислення дійсності для всіх транзакцій, що відбулися протягом дня j. Це значення підсумовування ділиться на $s_{P_{D,j}}$, дає змогу обчислити середню достовірність параметра для дня j.

Якщо величина $p_{5,M,T,j}$ представляє середній термін дії місяця j з точки зору транзакції, то обчислюватимемо її так:

$$p_{5,M,T,j} = \frac{\sum_{i=1}^{N_{p_{5.D,j}}} p_{5,D,T,i}}{N_{p_{5.D,j}}}, \quad (2.47)$$

де $N_{p_{5.D,T,j}}$ – кількість днів, протягом яких відбулися транзакції в місяці j; значення величини $p_{5,D,T,i}$ обчислюємо за формулою (2.46).

Сумування застосовується над всіма значеннями $p_{5,D,T,i}$ для всіх значень i, які дорівнюють від 1 до $N_{p_{5.D,j}}$, для обчислення дійсності для всіх транзакцій, що відбулися протягом $N_{p_{5.D,j}}$ днів у місяці j. Це підсумоване значення, поділене на $N_{p_{5.D,j}}$, дає змогу обчислити середню термін дії транзакції на місяць.

Введемо величину $p_{5,Y,T,j}$, що представляє середню термін дії року j з точки зору транзакції та обчислюватимемо її так:

$$p_{5,Y,T,j} = \frac{\sum_{i=1}^{N_{p_{5.M,j}}} p_{5,M„,T,i}}{N_{p_{5.M,j}}}, \quad (2.48)$$

де значення величини $N_{p_{5.M,j}}$ представляє собою кількість місяців, протягом яких відбувались транзакції в році j; значення величини $p_{5,M,T,i}$ обчислюється за формулою (2.42).

Сумування застосовується до всіх значень $p_{5,M„,T,i}$ для всіх значень i, які дорівнюють від 1 до $N_{p_{5.M,j}}$, для обчислення дійсності для всіх операцій, що відбулись за $N_{p_{5.M,j}}$ місяців у році j. Це підсумоване значення, поділене на $N_{p_{5.M,j}}$, дає змогу обчислити середню термін дії транзакції на рік.

Розглянемо параметр p_6, що відповідає за послідовність великих даних за двома варіантами: за параметрами; за часом. Таким чином, буде два типи послідовності. За послідовності параметрів кожне значення порівнюється зі значенням з іншого набору даних. Тоді, як на основі часу позначки часу порівнюються з обома наборами даних.

Спочатку проведемо дослідження за послідовністю параметрів для параметра і набору даних джерела даних $Q_{n,1,1}$ і приймемо набір даних джерела даних $Q_{n,1,2}$ як еталонний набір даних. Представимо введені розглядувані структури матрицями:

$$Q_{n,1,1} = \begin{pmatrix} Q_{1_{1,1}} \\ Q_{1_{2,1}} \\ \cdots \\ Q_{1_{n,1}} \end{pmatrix}, \qquad Q_{n,1,2} = \begin{pmatrix} Q_{2_{1,1}} \\ Q_{2_{2,1}} \\ \cdots \\ Q_{2_{n,1}} \end{pmatrix}. \tag{2.49}$$

Розмір набору даних $Q_{n,1,1}$ повинен бути рівним набору даних розміру $Q_{n,1,2}$. Якщо ж розмірність різна, то для вирівнювання розмірності підставимо нуль у відсутньому вимірі, щоб зробити його збіжним, так що n_W - це загальна кількість транзакції та дорівнює максимальному номеру транзакції обох наборів даних. Узгодженість на станції щодо джерел даних $Q_{n,1,1}$ щодо $Q_{n,1,2}$ може бути представлена за формулою (2.50) для параметра *i*. Далі, здійснимо порівняння кожного елементу даних, присутнього у параметрі *i* так:

$$p_6 = 1, \text{ якщо різниці в обох наборах даних не виявлено,}$$
$$\text{інакше } p_6 = 0. \tag{2.50}$$

Введемо величину $p_{6,i}$ для представлення узгодженості кожного параметра для параметра *i* та обчислюватимемо її так:

$$p_{6,P,i} = \frac{\sum_{j=1}^{n_W} p_{6,j}}{n_W}, \tag{2.51}$$

де значення величини $p_{6,j}$ обчислюється за формулою (2.50);

n_W – загальна кількість транзакцій.

Введемо величину $p_{6,P,i}\%$, що представлятиме відсоткову повноту для параметра *i* та обчислюватимемо її так:

$$p_{6,P,i}\% = p_{6,P,i} \cdot 100\%, \tag{2.52}$$

де значення $p_{6,P,i}$ обчислюватимемо за формулою (2.51).

Введемо величину $p_{6,D,P,i}$, що представлятиме середнє значення дня *j* з точки зору параметра та обчислюватимемо її так:

$$p_{6,D,P,j} = \frac{\sum_{i=1}^{s_{P_{D,j}}} p_{6,P,i}\%}{s_{P_{D,j}}}, \tag{2.53}$$

де значення $s_{P_{D,j}}$ – це транзакції за день *j*, які визначаються за формулою (2.8);

значення $p_{6,P,i}\%$ – відсоток узгодженості параметра *i* та обчислюються за формулою (2.52).

Сумування застосовується до всіх значень $p_{6,P,i}\%$ для всіх значень *i*, які дорівнюють від 1 до $s_{P_{D,j}}$, для обчислення узгодженості всіх транзакцій, що відбулися протягом дня *j*. Це значення підсумовування ділиться на $s_{P_{D,j}}$ та дає змогу отримати середню узгодженість параметрів для дня *j*.

Введемо величину $p_{6,M,P,i}$, що представлятиме середнє значення місяця *j* з точки зору параметра та обчислюватимемо її так:

$$p_{6,M,P,j} = \frac{\sum_{i=1}^{N_{p_{6.D,j}}} p_{6,D,P,i}}{N_{p_{6.D,j}}}, \tag{2.54}$$

де значення $N_{p_{6.D,T,j}}$ – кількість днів, протягом яких відбулися транзакції в місяці *j*; значення величини $p_{6,D,P,i}$ обчислюємо за формулою (2.52).

Сумування застосовується до всіх значень $p_{6,D,P,i}$ для всіх значень i, які дорівнюють від 1 до $N_{p_{6.D,T,j}}$, щоб обчислити узгодженість для всіх транзакцій, що відбулися протягом $N_{p_{6.D,T,j}}$ днів у місяці j. Це значення підсумовування, поділене на $N_{p_{6.D,T,j}}$, дає змогу обчислити середню узгодженість параметрів за місяць.

Введемо величину $p_{6,Y,P,j}$ представляє середнє значення за рік j з точки зору параметра та обчислюватимемо так:

$$p_{6,Y,P,j} = \frac{\sum_{i=1}^{N_{p_{6.M,j}}} p_{6,M_{,,}P,i}}{N_{p_{6.M,j}}}, \tag{2.55}$$

де значення величини $N_{p_{6.M,j}}$ представляє собою кількість місяців, протягом яких відбувались транзакції в році j; значення величини $p_{6,M_{,,}P,i}$ обчислюється за формулою (2.49).

Сумування застосовується до всіх значень $p_{6,M_{,,}P,i}$ для всіх значень i, які дорівнюють від 1 до $N_{p_{6.M,j}}$, щоб обчислити узгодженість усіх операцій, що відбулись за $N_{p_{6.M,j}}$ місяців у році j. Це значення підсумовування, поділене на $N_{p_{6.M,j}}$, дає змогу обчислити середню узгодженість параметрів на рік.

Послідовність часу вимірюється для того, щоб показати узгодженість часу між двома наборами даних. Для обох джерел даних $Q_{n,1,1}$ та $Q_{n,1,2}$ вимірюють отримані транзакції часу, щоб перевірити, чи підтримують вони узгодженість часу між тими самими транзакціями. Нехай структура $Q_{n,1,3}$ визначається як масив отриманих міток часу для джерела даних $Q_{n,1,1}$. Для джерела даних $Q_{n,1,1}$, $Q_{n,1,3}$ може бути заданий так:

$$Q_{n,1,3} = \begin{pmatrix} Q_{3_{1,1}} \\ Q_{3_{2,1}} \\ \cdots \\ Q_{3_{n,1}} \end{pmatrix}, \tag{2.56}$$

де значення величини $Q_{3_{i,1}}$ представляє собою мітку часу для i-ї транзакції; n_W представляє загальну кількість транзакцій.

Введемо величину $p_{6,T,P,i}$, яка відображатиме узгодженість часу для транзакції i. Щоб визначити узгодженість часу транзакції, порівняємо часову мітку цієї транзакції з міткою часу набору довідкових даних. Розглянемо джерела даних $Q_{n,1,1}$ та $Q_{n,1,2}$, щоб перевірити відповідність часу між ними. Тоді, $p_{6,T,P,i}$ для джерел даних $Q_{n,1,1}$ та $Q_{n,1,2}$ обчислюватимемо так:

$$p_{6,T,P,i} = 1, \text{ якщо різниці між } Q_{i,1,1} \text{ та } Q_{i,1,2} \text{ не знайдено,}$$
$$\text{інакше } p_{6,T,P,i} = 0, \tag{2.57}$$

де $(i, 1, j)$ представляє мітку часу для i-ї транзакції джерела даних j.

Введемо величину $p_{6,T,P,i}\%$ для визначення відсоткового значення з формули (2.57) та обчислюватимемо відсоткову узгодженість часу для транзакції i так:

$$p_{6,T,P,i}\% = p_{6,T,P,i} \cdot 100\%. \tag{2.58}$$

Введемо величину $p_{6,T,T,P,j}$ представляє середню узгодженість часу для дня j з точки зору транзакції та обчислюватимемо її так:

$$p_{6,D,T,P,j} = \frac{\sum_{i=1}^{s_{P_{D,j}}} p_{6,T,P,i}\%}{s_{P_{D,j}}}, \tag{2.59}$$

де значення $s_{P_{D,j}}$ – це транзакції за день j, які визначаються за формулою (2.8);

значення $p_{6,T,P,i}\%$ – відсоток узгодженості часу для транзакції i та обчислюються за формулою (2.58).

Підсумовування застосовується до всіх значень $p_{6,T,P,i}\%$ для всіх значень i, які дорівнюють від 1 до $s_{P_{D,j}}$, для обчислення узгодженості часу для всіх транзакцій, що відбулися протягом дня j. Це підсумоване значення ділиться на $s_{P_{D,j}}$ та дає змогу обчислити середню узгодженість часу параметра для дня j.

Введемо величину $p_{6,M,T,P,j}$ - середня узгодженість часу за місяць j з точки зору транзакції та обчислюватимемо її так:

$$p_{6,M,T,P,j} = \frac{\sum_{i=1}^{N_{p_{6.D,j}}} p_{6,D,T,P,i}}{N_{p_{6.D,j}}}, \tag{2.60}$$

де значення $N_{p_{6.D,T,j}}$ – кількість днів, протягом яких відбулися транзакції в місяці j;

значення величини $p_{6,D,T,P,i}$ обчислюємо за формулою (2.59).

Сумування застосовується до всіх значень $p_{6,D,T,P,i}$ для всіх значень i, які дорівнюють від 1 до $N_{p_{6.D,j}}$, для обчислення узгодженості часу для всіх транзакцій, що відбулися протягом $N_{p_{6.D,j}}$ днів у місяці j. Це підсумоване значення, поділене на $N_{p_{6.D,j}}$, дає змогу обчислювати середню узгодженість часу транзакції за місяць.

Введемо величину $p_{6,Y,T,P,j}$, яка представляє середню узгодженість часу за рік j з точки зору транзакції та обчислюватимемо її так:

$$p_{6,Y,T,P,j} = \frac{\sum_{i=1}^{N_{p_{6.M,j}}} p_{6,M,T,P,i}}{N_{p_{6.M,j}}}, \tag{2.61}$$

де значення величини $N_{p_{6.M,j}}$ представляє собою кількість місяців, протягом яких відбувались транзакції в році j; значення величини $p_{6,M,T,P,i}$ обчислюється за формулою (2.60).

Сумування застосовується до всіх значень $p_{6,M,T,P,i}$ для всіх значень i, які дорівнюють від 1 до $N_{p_{6.M,j}}$, для обчислення узгодженості часу для всіх операцій, що відбулись за $N_{p_{6.M,j}}$ місяців у році j. Це підсумовування, значення, поділене на $N_{p_{6.M,j}}$, дає змогу обчислити середню постійність часу транзакції на рік.

Розглянемо параметр надійність p_7, який опосередковано пов'язаний з якістю великих даних. Він є важливим, бо якість даних може погіршитися, якщо система, яка отримує дані, несправна. Нехай i-та станція має джерела даних $M = \{m_1, m_2, \dots, m_n\}$ протягом певного часового інтервалу за допомогою пошуку надійності джерел даних. Надійність великих даних визначається на рівні станції.

Введемо величину $p_{7,i}$ – надійність для станції i. Визначення надійності станції i з l_1 ненадійним джерелом даних і l_2 як загальну кількість джерел здійснимо за формулою:

$$p_{7,i} = \frac{l_2 - l_1}{l_2}, \tag{2.62}$$

де l_1 – кількість ненадійних джерел даних;

l_2 – загальна кількість джерел даних.

Відсоткове значення величини $p_{7,i}$ відповідатиме за відсоток надійності станції i та обчислюватиметься так:

$$p_{7,i}\% = p_{7,i} \cdot 100\%. \tag{2.63}$$

де значення $p_{7,i}$ обчислюється за формулою (2.62).

Зручність використання великих даних можна змоделювати, просто вимірявши три різні параметри якості, такі як повнота, точність і своєчасність.

$$p_{8,i} = \frac{p_{1,i}+p_{2,i}+p_{3,i}}{3}, \quad (2.64)$$

Отже, досліджено моделі якості даних на основі їх параметрів, що реалізуються в проекті: p_1відповідає повноті даних, p_2 – точності, p_3 – своєчасності, p_4 – унікальності, p_5 – достовірність, p_6 – послідовності, p_7 – надійності, p_8 – зручності використання.

Висновки до розділу 2

1. Досліджено моделі якості великих даних, які будуть основою для їх подальшої оцінки та застосування в розроблюваних методах з використанням великих даних. На відміну від відомих, запропоновані моделі також включають такий параметр як зручність використання.

2. Досліджені моделі якості великих даних враховували вісім параметрів якості для перевірки стандартів. Така деталізація параметрів якості великих даних дозволила визначити характеристики цих параметрів за рахунок відсутності кореляційного зв’язку між ними. Отримані моделі в сукупності дозволять здійснити оцінку якості великих даних.

3 МЕТОД ОЦІНЮВАННЯ ЯКОСТІ ВЕЛИКИХ ДАНИХ ТА ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ

3.1 Визначення узагальненого результату параметрів якості даних

У цьому розділі показано, як розрахувати узагальнений результат вимірювань, зроблених для кожного параметра якості даних за допомогою двох, добре відомих, методів, таких як Монте-Карло та нейронних мереж.

У методі Монте-Карло до даних, розрахованих на основі розглянутих вище моделей, застосовується певний ваговий коефіцієнт (%) для отримання узагальненого результату для оцінки якості даних. Оцінка вагового коефіцієнту вимагає особливої уваги і ґрунтуватиметься на взаємозв'язку між даними та результатами. Іноді результати можуть відрізнятися, якщо вагові коефіцієнти визначені неправильно.

Регресійний аналіз може бути використаний для визначення наступного набору вагових коефіцієнтів. Підхід змінної ваги дуже ефективний і більш практичний у використанні.

З іншого боку, нейронні мережі включають багаторівневий підхід. Вимога деяких рівнів (шарів) та їх нейронів потребує ретельного відбору. Навчання мережі також дуже важливим і потребує багато даних та часу. Неправильне навчання та методи, які використовуються для оцінки ваги, можуть призвести до помилок до 40%. Для даних із високими внутрішніми зв'язками методи нейронних мереж дуже ефективні. Однак, якщо дані дискретні та мають мінімальні зв'язки з іншими даними, методи нейронних мереж можуть стати «дорогими». Без будь-якого внутрішнього взаємозв'язку між даними метод на основі нейронних мереж дає результат дуже близький до результату, отриманого за допомогою методу Монте-Карло. Для обох методів використовуються вісім параметрів визначення узагальненого результату якості даних на рівні датчиків: повнота, точність, своєчасність, унікальність, достовірність, несуперечність, надійність і зручність використання.

Використання методу Монте Карло. В таблиці 3.1 показаний базовий розрахунок з використанням методу Монте-Карло з метою оцінки даних на основі моделей, визначених у попередньому розділі. Тут вага може бути вказана відповідно до вимог даних. Припустимо, що повнота не є важливою характеристикою даних для оцінки якості, тоді приймемо W1 за 0. Зазвичай сума вагового коефіцієнту дорівнює 100. Якщо бажаний результат відомий і це R, можна провести регресійний аналіз з використанням методу найменших квадратів для переоцінки вагових коефіцієнтів (w_1, w_2 і т. д.).

Таблиця 3.1 – Узагальнення параметрів

№ п/п	Параметри якості даних	Вагові коефіцієнти	Обчисленні дані	Результат
1	Повнота	W_1	D_1	$W_1 \times D_1$
2	Точність	W_2	D_2	$W_2 \times D_2$
3	Своєчасність	W_3	D_3	$W_3 \times D_3$
4	Унікальність	W_4	D_4	$W_4 \times D_4$
5	Достовірність	W_5	D_5	$W_5 \times D_5$
6	Послідовність	W_6	D_6	$W_6 \times D_7$
7	Надійність	W_7	D_7	$W_7 \times D_7$
8	Зручність	W_8	D_8	$W_8 \times D_8$
	Сума	$\sum_{i=1}^{8} w_i = 100$	-	$W = \sum_{i=1}^{8} w_i \times Di$

На рисунку 3.1 нижче показано графічне представлення нейронної мережі для даної системи моделювання даних з урахуванням лише одного прихованого шару з трьома нейронами, тобто структура нейронної мережі 8-3-1.

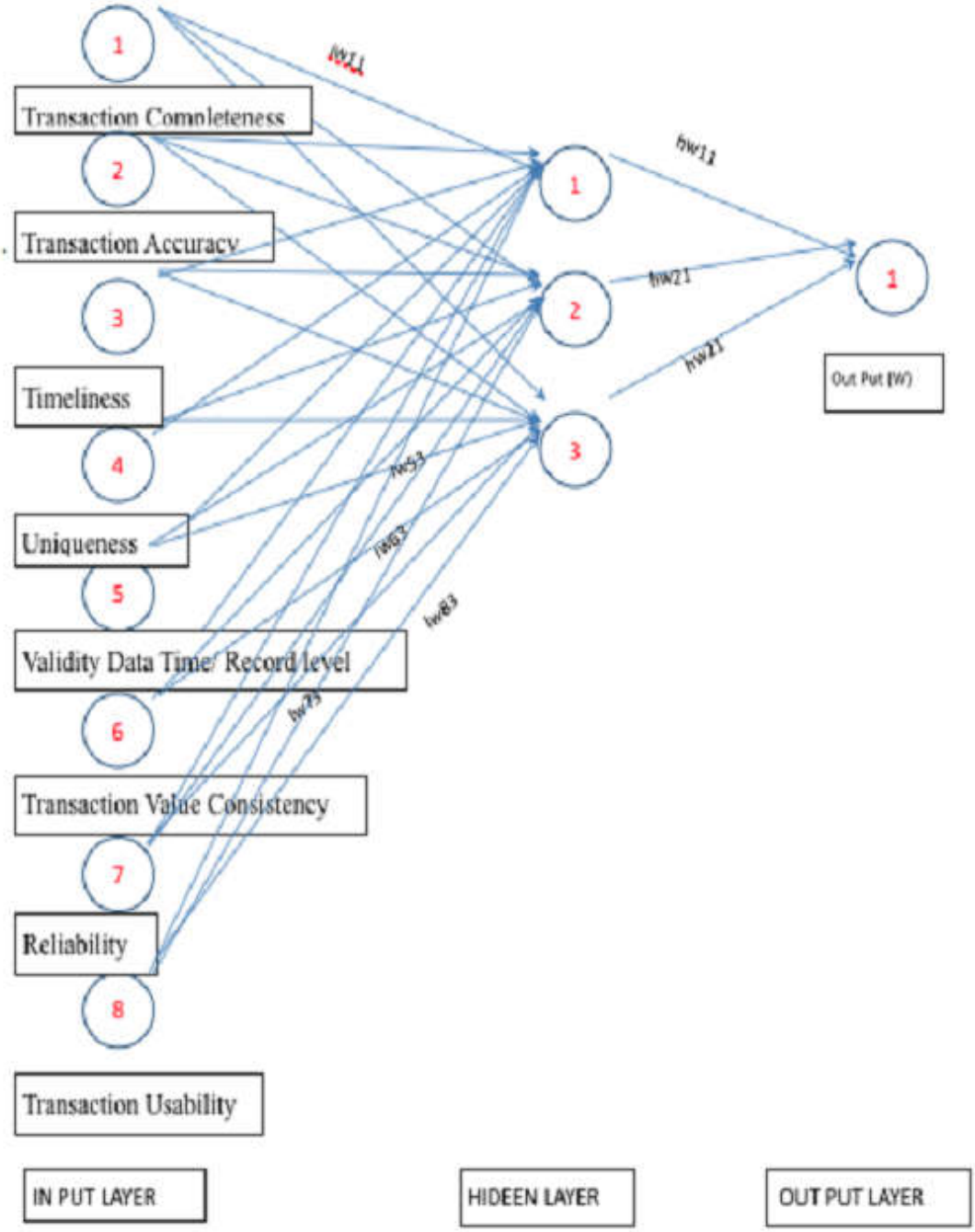


Рисунок 3.1 – Структура нейронної мережі, що відображає всі параметри якості даних

Слід зауважити, для того щоб зробити мережу читабельною, не всі ваги відображаються на рисунку. В даній структурі передбачено лише один прихований шар із трьома нейронами, але його можна змінити, тобто зміни кількість нейронних елементів.

Зазвичай для менш складної системи один прихований шар дає хороші результати. Система навчається, використовуючи вихідне значення порівняно з бажаним результатом R. Один прихований шар із трьома нейронами оцінює 8*3+3*1 = 27 ваг проти восьми в попередньому методі (Монте-Карло). Якщо до прихованого шару додати більше нейронів, буде більше ваг, які потрібно оцінити та оптимізувати. Отже, більше навчання означає більше даних. Після видалення прихованого шару система стає схожою на попередній метод (Монте-Карло).

3.2 Експериментальні дослідження

Експериментальні дослідження спрямовані на перевірку правильності параметрів якості, розглянутих в цій роботі, шляхом застосування прогнозного аналізу до параметрів якості даних про якість води, зібраних з Estuarine Research Reserve System (NERRS) [15]. Коректність моделей можна перевірити, розрахувавши значення параметрів з допомогою моделей, розглянутих в цій роботі, порівняно з прогнозованими значеннями з допомогою регресійного аналізу.

Параметри якості, що реалізуються в проекті: p_1відповідає повноті даних, p_2 – точності, p_3 – своєчасності, p_4 – унікальності, p_5 – достовірність, p_6 – послідовності, p_7 – надійності, p_8 – зручності використання.

Значення розраховуються на 2001–2014 роки, і виконуються прогнози на 2015 та 2016 роки.

3.2.1 Аналіз даних

Аналіз даних передбачає процес перевірки, очищення, перетворення та моделювання даних з метою виявлення корисної інформації, яку можна використовувати для підтримки процесу прийняття рішень [9]. Набір даних для прикладу складається з щоденних транзакцій даних одного датчика протягом багатьох років. Набір даних складається зі структурованих даних і

завантажується у форматі CSV. Після застосування процесу Extract Transformation Load (ETL) дані були збережені в MongoDB у структурованому форматі. Потім до даних було застосовано моделі параметрів якості та розраховано значення для кожного параметра якості. Розраховані значення були збережені у файлах CSV і далі використовувалися для прикладу.

3.2.2 Моделі прогнозування

Перед проведенням експериментальних досліджень та обговорення їх результатів, розглянемо моделі прогнозування та різні алгоритми. Прогнозне моделювання – це процес створення та перевірки моделі для найкращого визначення ймовірності результату [9]. У прогнозній аналітиці є кілька методів моделювання з машинного навчання, штучного інтелекту та статистики. Кожен з них має свої слабкі та сильні сторони, тому кожен найкраще підходить для певних задач. Ці моделі поділяються на три категорії, які подані у таблиці 3.2.

Таблиця 3.2 – Категорії моделей

Категорія	Визначення
Прогнозні моделі	Аналізують минулі показники для прогнозування майбутніх.
Описові моделі	Кількісно визначають зв’язки в даних, щоб класифікувати набори даних на групи.
Моделі рішень	Зображують зв’язки між усіма змінними рішення, щоб переглянути результати рішень, що включають багато змінних.

У таблиці 3.3 наведено приклади різних алгоритмів, які виконують статистичний аналіз та інтелектуальний аналіз даних для прогнозування закономірностей та тенденцій у даних.

Таблиця 3.3 – Прогнозні моделі

Моделі	Результат роботи	Приклади
Кластеризація	Групування результатів по категоріях	Kohonen, K-means, TwoStep
Регресія	Прогнозування зв’язків між змінними	Лінійна, експоненціальна, логарифмічна, геометрична і багатолінійна.
Часові ряди	Прогнозування по часу	Одинарне, подвійне та потрійне експоненціальне згладжування.
Асоціативні правила	Щоб визначити правила асоціації, цей алгоритм знаходить шаблони у великих наборах транзакційних даних	Apriori
Дерева рішень	Класифікація та визначення однієї або декількох дискретних змінних на основі інших змінних.	C 4.5, CNR Tree
Нейронні мережі	Прогнозування, класифікація та статистичне розпізнавання образів.	NNet Neural Network, MONMLP Neural Network

3.2.3 Регресійний аналіз

Регресійний аналіз допомагає оцінити зв'язок між залежними та незалежними (пояснювальними) змінними. Якщо є лише одна пояснювальна змінна, то це називається простою лінійною регресією, тоді як якщо присутні

декілька пояснювальних змінних, це називається множинною лінійною регресією.

Для проведення лінійного регресійного аналізу кожного параметра якості було отримано спостереження вимірювань одного параметра якості з 2001 по 2014 роки та нанесено на графік у Excel.

На рисунку 3.2 представлено діаграму розсіювання для параметра повноти; воно дає рівняння, за допомогою якого було прогнозовано значення для 2015 та 2016 років. Значення R^2 становить 0,822, що вказує на те, що рівняння регресії може пояснити 80% мінливості даних.

Years	Calculated	Predicted
2001	97.6	
2002	98.6	
2003	98.7	
2004	97.3	
2005	97.6	
2006	97.6	
2007	98.3	
2008	98.2	
2009	97.5	
2010	99.4	
2011	98.6	
2012	98.5	
2013	97.9	
2014	98.3	
2015	97.5	98.39
2016	98.7	98.42

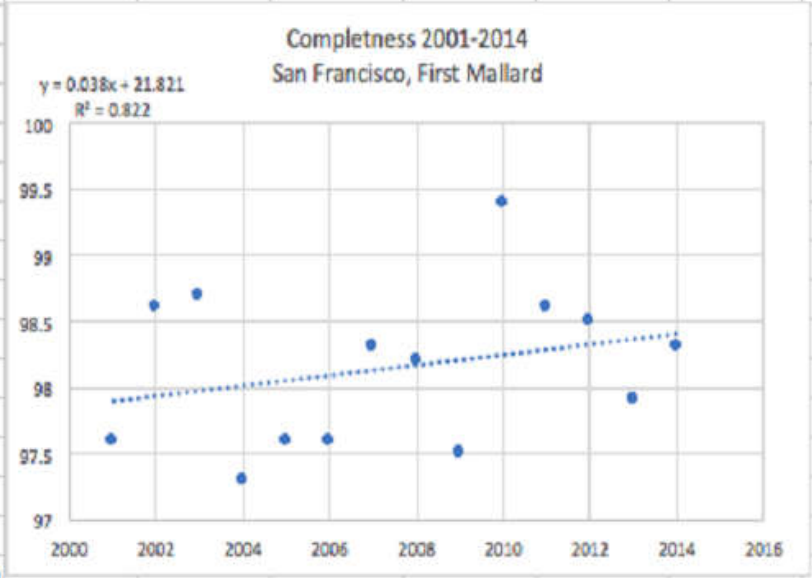


Рисунок 3.2 – Діаграма розсіювання для повноти параметрів (2001-2014 роки)

3.2.4 Результати експериментальних досліджень

На рисунках 3.3 та 3.4 показана діаграма, на якій показано розрахункові та прогнозні значення для 2015 та 2016 років відповідно.

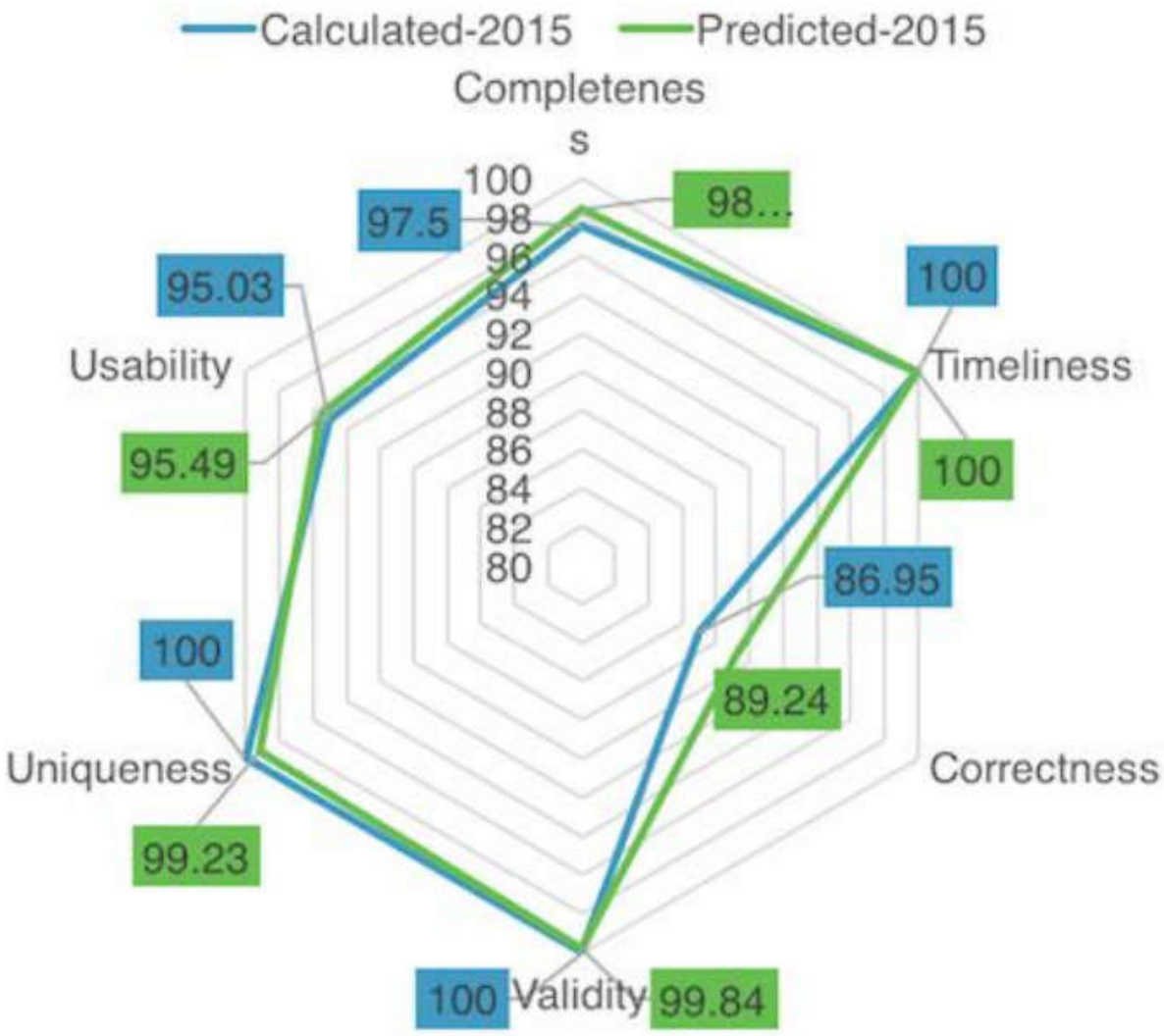


Рисунок 3.3 – Параметри якості даних: прогнозовані та фактичні значення за 2015 рік

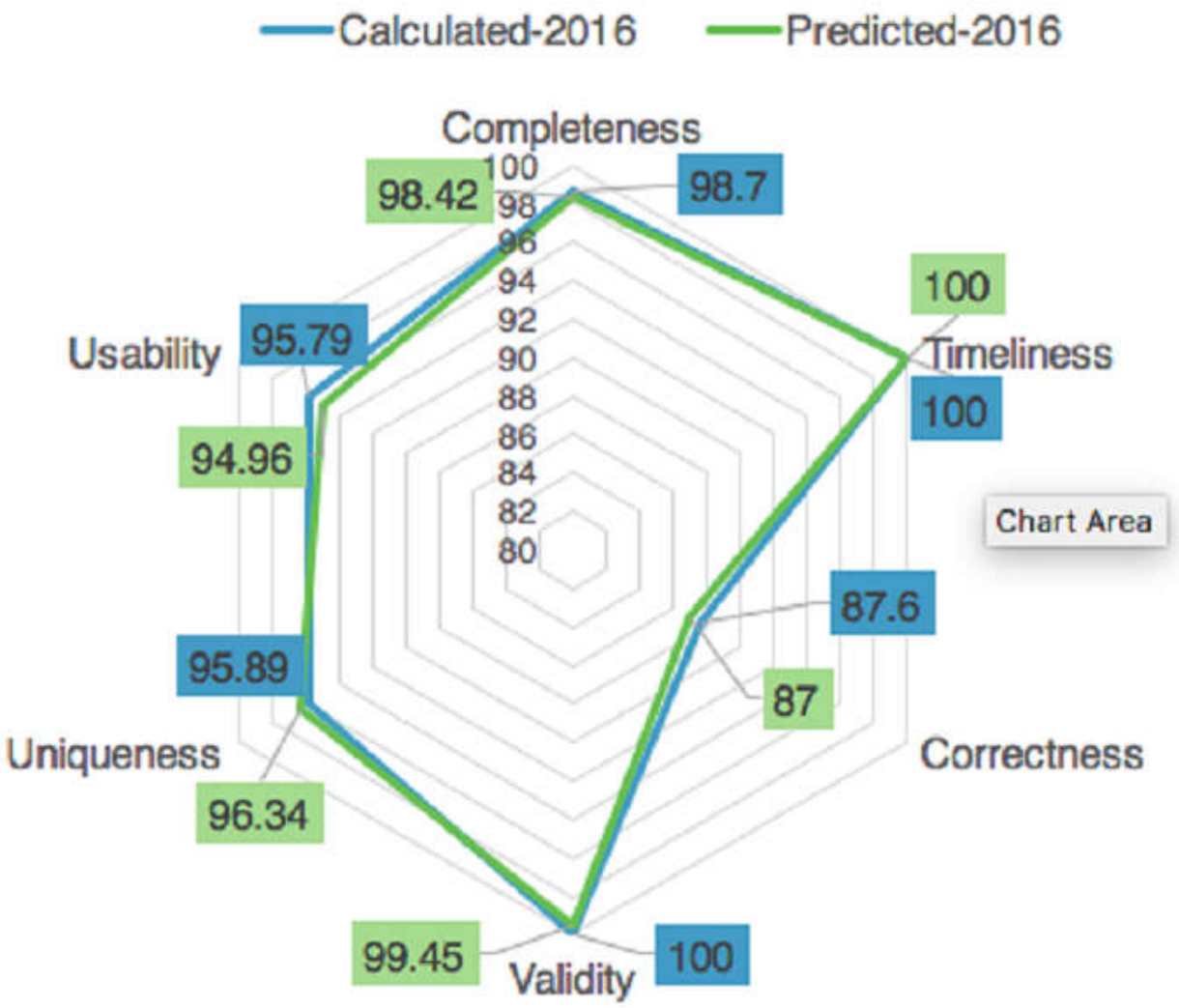


Рисунок 3.4 – Параметри якості даних: прогнозовані та фактичні значення за 2016 рік

Діаграми використовуються для відображення значень усіх розрахованих та прогнозованих параметрів. Нанесені лінії майже перекриваються, а довірчий інтервал всіх параметрів якості становить близько 95%. Це хороші показники, моделі, які запропоновані в цій роботі, є прийнятними.

Висновки до розділу 3

1. Показано, як розрахувати узагальнений результат вимірювань, зроблених для кожного параметра якості даних за допомогою методів Монте-Карло та нейронних мереж. Для обох методів використовуються вісім параметрів визначення узагальненого результату якості даних на рівні датчиків: повнота, точність, своєчасність, унікальність, достовірність, несуперечність, надійність і зручність використання.

2. Проведено експериментальні дослідження, які спрямовані на перевірку правильності параметрів якості, розглянутих в цій роботі, шляхом застосування прогнозного аналізу до параметрів якості даних. Коректність моделей перевіряється з допомогою регресійного аналізу, розрахувавши значення параметрів з допомогою моделей, розглянутих в цій роботі, порівняно з прогнозованими значеннями.

3. Запропоновані моделі оцінки якості даних можна використовувати для створення контрольних показників та протоколів для оцінки та оптимізації якості даних у більших масштабах. Ці інструменти вимірювання та прогнозування корисні при порівнянні різних даних, оскільки ці дані можуть використовуватися для надійного прийняття рішень. Отримані результати та запропоновані моделі можуть бути розширені в майбутньому, якщо вони будуть вивчені та доопрацьовані.

ВИСНОВКИ

1. Важливим для отримання достовірних та якісних результатів при використанні інформаційних технологій є не тільки методи, способи та засоби їх отримання, але і якість початкових даних. Вони мають ряд характеристик, від яких суттєво залежить результат. До таких характеристик, зокрема, можуть належати повнота, точність, змістовність тощо. Певні з цих характеристик можуть бути більш вагомими, інші ні. Але в сукупності вони дають змогу оцінити отриманий результат, як такий, що базується на якісних даних. В процесі застосування інформаційних технологій можуть виникати проблеми через наявність неповних або надлишкових даних. Тоді, потрібна оцінка якості цих даних, бо вони впливатимуть на отриманий результат.

2. Сучасні інформаційні технології обробки даних отримують на вході, як правило багато різнотипних даних і їх великі обсяги. Кріт того, ІТ побудовані з використанням методів, які спрямовані на обробку таких великих даних, а також при їх імплементації використовуються складні алгоритми. Зокрема, наприклад для глибокого навчання з використанням штучних нейронних мереж потрібні досить якісні дані, бо як правило, нейронним мережам потрібно більше даних для створення більш потужних абстракцій. Хоча в сценаріях з великими даними міститься багато даних, вони не завжди правильні або не завжди мають досить високу якість для навчання. Невеликі варіації або несподівані особливості початкових даних, а особливо неповнота даних можуть повністю розбалансувати налаштування в моделях нейронних мереж.

3. Враховуючи проблеми з початковими даними великих обсягів, які використовуються в ІТ, що побудовані на основі сучасних методів, зокрема і інтелектуальних та еволюційних, необхідним є початкова оцінка якості великих даних.

4. Досліджено моделі якості великих даних, які будуть основою для їх подальшої оцінки та застосування в розроблюваних методах з використанням великих даних.

5. Досліджені моделі якості великих даних враховували вісім параметрів якості для перевірки стандартів. Така деталізація параметрів якості великих даних дозволила визначити характеристики цих параметрів за рахунок відсутності кореляційного зв'язку між ними. Отримані моделі в сукупності дозволять здійснити оцінку якості великих даних.

6. Показано, як розрахувати узагальнений результат вимірювань, зроблених для кожного параметра якості даних за допомогою методів Монте-Карло та нейронних мереж. Для обох методів використовуються вісім параметрів визначення узагальненого результату якості даних на рівні датчиків: повнота, точність, своєчасність, унікальність, достовірність, несуперечність, надійність і зручність використання.

7. Проведено експериментальні дослідження, які спрямовані на перевірку правильності параметрів якості, розглянутих в цій роботі, шляхом застосування прогнозного аналізу до параметрів якості даних. Коректність моделей перевіряється з допомогою регресійного аналізу, розрахувавши значення параметрів з допомогою моделей, розглянутих в цій роботі, порівняно з прогнозованими значеннями.

8. Запропоновані моделі оцінки якості даних можна використовувати для створення контрольних показників та протоколів для оцінки та оптимізації якості даних у більших масштабах. Ці інструменти вимірювання та прогнозування корисні при порівнянні різних даних, оскільки ці дані можуть використовуватися для надійного прийняття рішень. Отримані результати та запропоновані моделі можуть бути розширені в майбутньому, якщо вони будуть вивчені та доопрацьовані.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Askham, N., Denise, C., Martin, D., Helen, F., Mike, G., Ulrich, L., Rob, L., Chris, M., Gary, P., and Julian, S. (2013). The six primary parameters for data quality assessment. Technical report, *DAMA UK Working Group*.

2. Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal, 14*, 2.

3. Clarke, R. (2014). Quality factors in big data and big data analytics. Xamax Consultancy Pry Ltd. Retrieved from http://www.rogerclarke.com/EC/BDQF.html.

4. Cloern, J.E., & Schraga, T. S. (2016). USGS Measurements of water quality in San Francisco Bay (CA). 1969-2015, version 2: U. S. Geological Survey data release, doi:10.5066/F7TQ5ZPR.

5. Eckerson, W. W. (2002). Data quality and the bottom line: Achieving business success through a commitment to high quality data. *The Data Warehousing Institute,* 1-36.

6. Gao, J., Xie, C., & Tao, C. (2016, March). Big data validation and quality assurance-Issues, challenges, and needs. In *Service-Oriented System Engineering (SOSE), 2016 IEEE Symposium* (433-441). IEEE.

7. Gudivada, V. N., Rao, D., & Grosky, W. I. (2016). Data quality centric application framework for big data. *ALLDATA 2016*, 33.

8. IDC Forecast: Big data technology and services to hit $32.4 billion in 2017. (2013, December 18). Retrieved from http://www.hostingjournalist.com/cloud-hosting/idcforecast-big-data-technology-and-services-to-hit-32-4-billion-in-2017/.

9. IDC Forecast: Big data technology and services to hit $32.4 billion in 2017. Retrieved from http://www.hostingjournalist.com/cloud-hosting/idcforecast-big-data-technology-and-services-to-hit-32-4-billion-in-2017.

10. Kang, G., Gao J. Z., & Xie, G. (2017, April). Data-Driven water quality analysis and prediction: A survey. *2017 IEEE Third International Conference on*

Big Data Computing Service and Applications (BigDataService), San Francisco, CA, 2017, 224-232.

11. Laranjeiro, N., Soydemir, S. N., & Bernardino, J. (2015, November). A survey on data quality: Classifying poor data. In *Dependable Computing (PRDC),* 2015 IEEE 21st Pacific Rim International Symposium on (179-188). IEEE.

12. Lavanya, S., & Prakasm, S. (2014). Reliable techniques for data transfer in wireless sensor networks. *International Journal of Engineering and Computer Science, 3 (12)*.

13. Loshin, D. (2010). The practitioner's guide to data quality improvement. Amsterdam, Netherlands: Elsevier.

14. Ludo, H., & Beek, J. (2013, December 16). Open data quality. Presentation published in *Technology.* Retrieved from https://www.slideshare.net/OpenDataSupport/open-dataquality-29248578.

15. National Estuarine Research Reserve System, https://coast.noaa.gov/nerrs/

16. Sharma, A. B., Golubchik, L., & Govindan, R. (2010). Sensor faults: Detection methods and prevalence in real-world data sets. *ACM Transactions on Sensor Networks (TOSN), 6(3),* 23.

17. Vass, L. (2016, 28 June). Terabyte terror: It takes special databases to lasso the Internet of Things. *Ars Technica*. Retrieved from https://arstechnica.com/informationtechnology/2016/06/building-databases-for-the-internet-of-data-spewing-things/.

18. What is predictive modeling? (2017) *Predictive Analytics Today*. Retrieved from http://www.predictiveanalyticstoday.com/predictive-modeling/.

19. Wigan, M. R., & Clarke, R. (2013). Big data's big unintended consequences. *Computer, 46(6),* 46-53.

20. Woodall, P., Gao, J., Parlikad, A., & Koronios, A. (2015). Classifying data quality problems in asset management. *In Engineering Asset Management-Systems, Professional Practices and Certification* (321-334). New York, NY: Springer, Cham.

21. Zhu, X., Lu, Y., Han, J., & Shi, L. (2016). Transmission reliability evaluation for wireless sensor networks. *International Journal of Distributed Sensor Networks, 12(2)*.

22. Комар М. П. Методологічні основи інформаційної технології інтелектуального аналізу та обробки великих даних: дис. д-ра техн. наук: 05.13.06 / Українська академія друкарства. Львів, 2021. 363 с.

23. Трач Ю.І. та ін. Особливості навчання глибоких нейронних мереж для обробки та аналізу великих даних. Збірник тез доповідей міжнародної науково-практичної інтернет-конференції «Інформаційне суспільство: технологічні, економічні та технічні аспекти становлення». – 2022. – Вип. 72, http://www.konferenciaonline.org.ua/ua/article/id-800/

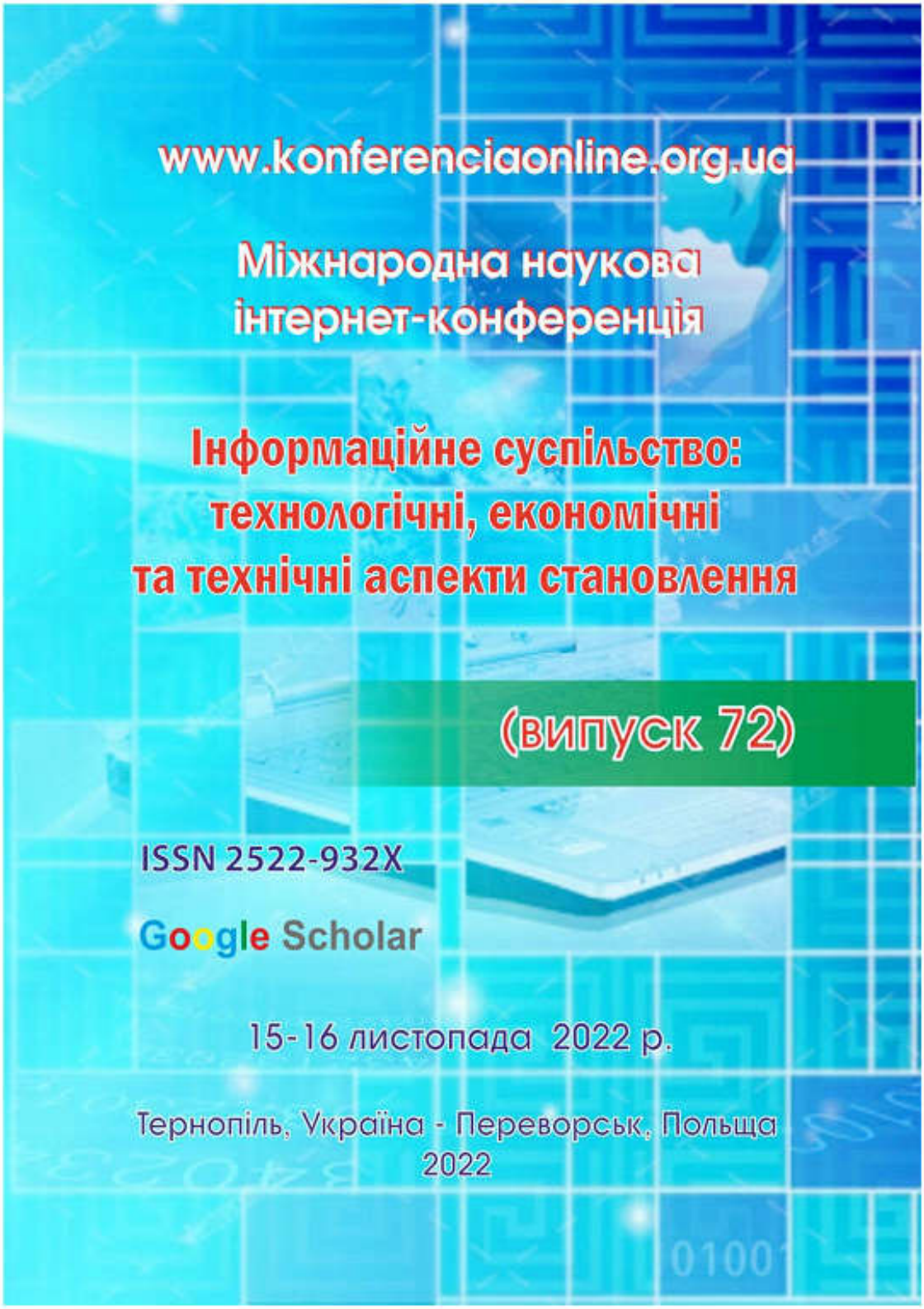
24. Трач Ю.І. та ін. Підвищення ефективності побудови систем обробки та аналізу великих даних на основі глибоких нейронних мереж. Збірник матеріалів школи-семінару молодих вчених і студентів «Комп'ютерні інформаційні технології» (CIT'2022). С. 00-00.

25. Загальні рекомендації з підготовки, оформлення, захисту та оцінювання випускних кваліфікаційних робіт здобувачів вищої освіти першого «бакалаврського» і другого «магістерського» рівнів / За ред. доц. М.І. Шинкарика. Тернопіль: ТНЕУ, 2018. 67 с.

26. Комар М.П., Саченко А.О., Васильків Н.М. Методичні рекомендації до виконання кваліфікаційної роботи з освітньо-професійної програми «Комп'ютерні науки» спеціальності 122 «Комп'ютерні науки» за другим (магістерським) рівнем вищої освіти. Тернопіль: ЗУНУ, 2021. 32 с.

Додаток А

Копії публікацій



УДК 001 (063)

Інформаційне суспільство: технологічні, економічні та технічні аспекти становлення (випуск 72)*:* матеріали Міжнародної наукової інтернет-конференції, (м. Тернопіль, Україна – м. Переворськ, Польща, 15-16 листопада 2022 р.) / [редкол. : О. Патряк та ін.] ; ГО "Наукова спільнота"; WSSG w Przeworsku. – Тернопіль : ФО-П Шпак В.Б. – 223 с. – ISSN 2522-932X

Збірник тез доповідей підготовлено за матеріалами Міжнародної наукової інтернет-конференції (випуск 72) 15-16 листопада 2022 р. на сайті www.konferenciaonline.org.ua

Наша адреса: Оргкомітет МНІК "Конференція онлайн"
а/с 797, м. Тернопіль 46005
тел. моб. 068 366 0 525
e-mail: inetkonf@ukr.net

URL Інтернет-конференції: http://www.konferenciaonline.org.ua/
ISSN 2522-932X

Зміст

Секція 1. Інформаційні системи і технології

***Біблий Олег Сергійович**, студент, Західноукраїнський національний університет, м. Тернопіль;*
***Трач Юлія Іванівна**, студентка, Західноукраїнський національний університет, м. Тернопіль;*
***Хлібойко Михайло Ярославович**, студент, Західноукраїнський національний університет, м. Тернопіль;*
***Цвик Роман Богданович**, студент, Західноукраїнський національний університет, м. Тернопіль*

***Науковий керівник: Комар Мирослав Петрович**, доктор технічних наук, доцент, Західноукраїнський національний університет, м. Тернопіль*

ОСОБЛИВОСТІ НАВЧАННЯ ГЛИБОКИХ НЕРОННИХ МЕРЕЖ ДЛЯ ОБРОБКИ ТА АНАЛІЗУ ВЕЛИКИХ ДАНИХ

Інтернет-адреса публікації на сайті:
http://www.konferenciaonline.org.ua/ua/article/id-800/

Великі дані стають все більш важливими, оскільки багатьом організаціям і компаніям необхідно збирати корисну інформацію з величезних обсягів даних. Традиційні алгоритми машинного навчання були розроблені, щоб змусити машини пізнавати і розуміти реальний світ, а це означає, що комп'ютери можуть самостійно вивчати нові знання та досвід в обмеженому наборі даних за допомогою деяких спеціальних методів машинного навчання.

Однак у середовищі великих даних складно вивчати та аналізувати традиційні алгоритми машинного навчання, тому що великі дані мають велику кількість вибірок даних, складну структуру та широку різноманітність. На щастя, глибоке навчання – дуже перспективний спосіб вирішення аналітичних задач у великих даних. Важливою особливістю глибокого навчання, яке також є ядром аналітики великих даних, є автоматичне вивчення представлень високого рівня та складних структур із величезних обсягів необроблених вхідних даних для отримання значущої інформації. Навчання у великомасштабних мережах глибокого навчання з мільярдами чи навіть більше параметрами може значно підвищити точність глибоких мереж. Але навчання в цих великих глибоких мережах забирає багато часу і потребує величезної кількості обчислювальних ресурсів.

Отже, необхідно прискорити ці великі глибокі мережі за допомогою високопродуктивних обчислювальних ресурсів (наприклад, графічних процесорів, суперкомп'ютерів та розподілених кластерів).

Глибоке навчання здатне виявляти складні структури в багатовимірних даних, що зрештою приносить користь у багатьох сферах життя суспільства. Так, наприклад, рекорд класифікації зображень був побитий в ImageNet Challenge 2012 з використанням глибокої нейронної згорткової мережі (CNN) [1].

Крім того, глибоке навчання значно впливає на вирішення інших проблем комп'ютерного зору, такі як виявлення осіб, сегментація зображень, загальне виявлення об'єктів і оптичне розпізнавання символів. Глибоке навчання також можна використовувати для розпізнавання мови, розуміння природної мови та багатьох інших галузей, таких як системи рекомендацій, фільтрація веб-контенту, прогнозування захворювань та ін. [2].

З покращенням архітектури глибоких мереж, навчальних вибірок та високопродуктивних обчислень глибоке навчання успішно застосовуватиметься у більшій кількості додатків у найближчому майбутньому.

Мережі глибокого навчання хороші для виявлення складних структур багатовимірного набору навчальних даних і добре підходять для вирішення великомасштабних задач. Навчання на великому наборі даних та великомасштабних глибоких мережах, які мають велику кількість шарів т а кількість параметрів, може показати хороші результати. Але це також означає, що навчання на таких великих моделях займає набагато більше часу, доводиться чекати дуже довго (кілька місяців або навіть років), щоб отримати добре навчену модель.

Зі швидким розвитком сучасних обчислювальних пристроїв і паралельних методів стало можливим навчати ці великомасштабні моделі за допомогою високопродуктивних обчислювальних методів, таких як розподілені системи з тисячами ядер центрального процесора (CPU), графічних процесорів (GPU) з тисячами обчислювальних потоків та інших паралельних обчислювальних пристроїв.

Література:

1. Krizhevsky A., Sutskever I., Hinton G. ImageNet classification with deep convolutional neural networks. In: Proc. Advances in Neural Information Processing Systems. 2012. 25. Pp. 1090-1098.
2. LeCun Y., Bengio Y., Hinton G. Review: Deep learning. Nature. 2015. 521. Pp. 436-444.

УДК 004.6+004.9

ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ ПОБУДОВИ СИСТЕМ ОБРОБКИ ТА АНАЛІЗУ ВЕЛИКИХ ДАНИХ НА ОСНОВІ ГЛИБОКИХ НЕЙРОННИХ МЕРЕЖ

Трач Ю.І., Хлібойко М.Я., Біблий О.С., Цвик Р.Б. [1)]
Західноукраїнський національний університет
[1)] магістранти

І. Постановка проблеми

На сьогодні великі дані стали визначальною і постійно зростаючою характеристикою сучасної економіки, вектором глобальних перетворень майже у всіх областях. Аналіз великих даних дозволяє побачити приховані закономірності, непомітні обмеженому людському сприйняттю. Це дає безпрецедентні можливості оптимізації всіх сфер нашого життя: національна розвідка, кібербезпека, виявлення шахрайства, маркетинг і медична інформатика, державне управління, телекомунікації, фінанси, транспорт, виробництво і т. д. [1].

Але необроблені дані, самі по собі не приносять користі. Протягом останніх два десятиліття розроблено технологію Big Data [2], яка може швидко отримувати, зберігати, обробляти, аналізувати великі дані і, таким чином, отримувати з них корисну інформацію. Технологія Big Data добре справляється з структурованими, трохи гірше з напівструктурованими даними. Але складні, неструктуровані дані важко зрозуміти за допомогою традиційних аналітичних методів.

Для того, щоб ефективно обробляти великі обсяги даних при прийнятних часових затратах, необхідні особливі технології. Сьогодні методи машинного навчання, зокрема глибокого навчання [3] разом з досягненнями в області обчислювальної потужності, відіграють важливу роль у аналітиці великих даних. Глибокі нейронні мережі мають велику ефективність нелінійного перетворення і представлення даних в порівнянні з традиційними нейронними мережами [4].

Глибокі нейронні мережі також успішно застосовуються в промислових продуктах, які використовують переваги великого обсягу цифрових даних. Такі компанії, як Apple, Google, Microsoft, IBM Facebook та ін., які щоденно збирають і аналізують величезні обсяги даних, наполегливо просувають проекти, пов'язані з глибоким навчанням. За версією вчених Массачусетського технологічного інституту глибокі нейронні мережі входять в список 10 найбільш перспективних високих технологій, здатних в недалекому майбутньому в значній мірі перетворити повсякденне життя більшості людей на нашій планеті.

На сьогодні глибоке навчання (Deep Learning) і великі дані (Big Data) є одними з найбільш гарячими тенденціями в швидко зростаючому цифровому світі. Хоча глибоке навчання має величезний потенціал для отримання більшої користі з великих даних порівняно з традиційними аналітичними методами, воно все ще знаходиться початковому етапі розвитку, і перед дослідниками і практиками стоїть ряд серйозних проблем [5].

1. Для глибокого навчання потрібно досить якісні дані - як правило, нейронним мережам потрібно більше даних для створення більш потужних абстракцій. Хоча в сценаріях з великими даними міститься багато даних, вони не завжди вірні або не завжди мають досить високу якість для навчання. Невеликі варіації або несподівані особливості вхідних даних, а особливо наявність відсутніх даних можуть повністю зіпсувати моделі нейронних мереж.

2. Проблеми безпеки – глибоке навчанню потребує для навчання і перенавчання на масивних, реалістичних наборах даних, і в процесі розробки алгоритму ці дані необхідно безпечно передавати, зберігати і обробляти. Коли алгоритм глибокого навчання розгортається в критично важливому середовищі, зловмисники можуть вплинути на вихідні дані нейронної мережі, вносячи невеликі шкідливі зміни у вхідні дані. Це може повністю змінити отримані результати, привести до неправильного діагнозу пацієнта або до аварії безпілотного автомобіля.

3. Глибоке навчання має труднощі зі зміною контексту – моделі нейронної мережі, навченої на певній проблемі, буде важко вирішити дуже схожі задачі, представлені в іншому контексті. Наприклад, системи глибокого навчання, які можуть ефективно виявляти набір зображень, можуть не коректно працювати, коли представлені одними і тими ж зображеннями, але поверненими під різними кутами або з різними характеристиками (відтінки сірого в порівнянні з кольором, різне розширення і т. д.).

4. Рішення в режимі реального часу – більшість великих даних в світі передається в режимі реального часу, і важливість аналізу даних в режимі реального часу стає все більш актуальним. Глибоке навчання складно використовувати для аналізу даних в реальному часі, оскільки воно вимагає великих обчислювальних та часових ресурсів для навчання глибокої нейронної мережі.

II. Мета роботи

Метою роботи є розробка рішень, які дозволять підвищити ефективність побудови систем обробки та аналізу великих даних на основі глибоких нейронних мереж.

III. Напрямки підвищення ефективності побудови систем обробки та аналізу великих даних на основі глибоких нейронних мереж

Для вирішення цих та інших проблем великих даних авторами запропоновано ряд рішень.

По-перше, важливим для отримання достовірних та якісних результатів при використанні інформаційних технологій є не тільки методи, способи та засоби їх отримання, але і якість початкових даних. Вони мають ряд характеристик, від яких суттєво залежить результат. До таких характеристик, зокрема, можуть належати повнота, точність, змістовність тощо. Певні з цих характеристик можуть бути більш вагомими, інші ні. Але в сукупності вони дають змогу оцінити отриманий результат, як такий, що базується на якісних даних.

Враховуючи проблеми з початковими даними великих обсягів, які використовуються в ІТ, що побудовані на основі сучасних методів, зокрема і інтелектуальних та еволюційних, необхідним є початкова оцінка якості великих даних. Для цього запропоновано використати автоенкодерну нейронну мережу для дослідження моделей якості великих даних.

По-друге, ще одна проблема полягає в тому, що стандартні методи обчислювального інтелекту неспроможні обробляти вхідні дані з пропущеними значеннями і, отже, неспроможні виконувати задачі класифікації, регресії та ін. На сьогодні, існує велика кількість методів відновлення відсутніх даних (імпутації даних). Проте, такі методи мають ряд обмежень, зокрема щодо структурованості даних. В багатьох випадках отримані результати є суперечливими. Жоден з проаналізованих методів для відновлення відсутніх даних (імпутації даних) не має переваг над іншими і тому доцільно провести додаткові дослідження щодо можливості аналізу великих даних без відновлення відсутніх значень.

Для цього запропоновано метод роботи з відсутніми даними для оперативного моніторингу стану складних об’єктів. Спочатку, було вирішено задачу класифікації за наявності пропущених даних, коли робилися спроби відновити пропущені значення. Потім проблемну область було розширено до задачі регресії.

По-третє, навчання глибоких нейронних мереж вимагає значних обчислювальних ресурсів. Обчислення на центральному процесорі є досить повільними для великих об’ємів даних, тому паралельні обчислення активно застосовуються для інтелектуального аналізу даних. Використання розподілених систем для паралельних обчислень є досить складним і вимагає додаткових затрат на устаткування. Через це було вирішено дослідити методи та засоби паралельного навчання нейронних мереж на графічних процесорах Nvidia з підтримкою технології CUDA, яка має високу продуктивність паралельних обчислень.

По-четверте, щоб реалізувати ідею структурного синтезу для створення глибоких нейронних мереж, використано наступні обчислювальні конструкції, що імітують такі механізми біологічної еволюції як: спадковість, природний відбір, випадкові мутації. Використано механізм спадковості для представлення архітектурних характеристик глибокої нейронної мережі. Ідея спадковості моделюється шляхом кодування архітектурних характеристик глибоких нейронних мереж в формі синаптичних ймовірнісних моделей, які використовуються для передачі ознак від покоління до покоління.

Висновок

Запропоновані рішення дозволять підвищити ефективність побудови систем обробки та аналізу великих даних на основі глибоких нейронних мереж.

Список використаних джерел

1. Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute [Електронний ресурс] – Режим доступу: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

2. Lynch C. Big data: science in the petabyte era / C. Lynch // Nature. – 2008. – Vol. 455. – P. 1-50.

3. LeCun Y. Deep learning / Y. LeCun, Y. Bengio, G. Hinton // Nature. – 2015. – Vol. 521 (7553). – pp. 436–444.

4. Hinton G.E. A fast learning algorithm for deep belief nets / G. E. Hinton, E.S. Osindero, Y. Teh // Neural Computation. – 2006. – Vo1. 18. – pp. 1527–1554.

5. Neural Network Concepts: Deep Learning for Big Data. https://missinglink.ai/guides/neural-network-concepts/deep-learning-for-big-data.