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

ХЛІБОЙКО Михайло Ярославович

Метод аналізу великих даних з відсутніми значеннями на основі нейронних мереж / Method for Analyzing Big Data with Missing Values Based on Neural Networks

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

Кваліфікаційна робота

Виконав студент групи
КНм-21
М.Я. Хлібойко
_____

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 СТУДЕНТУ
Хлібойко Михайло Ярославович
(прізвище, ім'я, по батькові)

1. Тема кваліфікаційної роботи
Метод аналізу великих даних з відсутніми значеннями на основі нейронних мереж / Method for Analyzing Big Data with Missing Values Based on Neural Networks
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
- дослідити основні визначення та характеристики великих даних;
- аналіз підходів для обробки та аналізу великих даних;
- провести аналіз методів відновлення відсутніх даних;
- зробити постановку задачі дослідження;
- дослідити теоретичні основи нечіткої нейронної мережі ARTMAP;
- дослідити теоретичні основи багатошарових персептронів;
- дослідити алгоритм зворотного поширення помилки для навчання багатошарових персептронів;
- запропонувати вдосконалений метод на основі нейромережевих технологій для аналізу великих даних з відсутніми значеннями;
- розробити нейромережеві архітектури для вирішення задачі аналізу великих даних з відсутніми значеннями;
- провести експериментальні дослідження вдосконаленого методу на основі нейромережевих технологій для аналізу великих даних з відсутніми значеннями;
- провести порівняльний аналіз запропонованого методу на основі нейромережевих технологій для аналізу великих даних з відсутніми значеннями.

5. Перелік графічного матеріалу у роботі

– схема запропонованого підходу на основі ансамблю нейронних мереж для відсутніх даних;

– графік продуктивності відповідно до кількості класифікаторів.

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	Огляд основних визначень та аналіз методів обробки і аналізу великих даних	12.2021 р. – 03.2022 р.	
2	Теоретичні основи аналізу великих даних з відсутніми значеннями на основі нейронних мереж	03.2022 р. – 05.2022 р.	
3	Реалізація та експериментальні дослідження методу аналізу великих даних з відсутніми значеннями на основі нейронних мереж	05.2022 р. – 11.2022 р.	
4	Повне завершення та представлення кваліфікаційної роботи на кафедру	16.11.2022 р.	

Студент _____ М.Я. Хлібойко
підпис

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

РЕЗЮМЕ

Кваліфікаційна робота на тему «Метод аналізу великих даних з відсутніми значеннями на основі нейронних мереж» на здобуття освітнього ступеня «Магістр» зі спеціальності 122 «Комп'ютерні науки» освітньої програми «Комп'ютерні науки» написана обсягом в 69 сторінок і містить 7 ілюстрацій, 3 таблиці, 1 додаток та 77 використаних джерел.

Метою кваліфікаційної роботи є вдосконалення методу аналізу великих даних з відсутніми значеннями на основі нейронних мереж.

Методи досліджень: аналіз літературних джерел та узагальнення інформації; методи нейронних мереж; методи машинного навчання; методи проведення експериментів і порівняння результатів; розрахунок, візуалізація та інтерпретація порівняльних даних про результати експериментальних досліджень.

Результати дослідження: вдосконалено метод аналізу великих даних з пропущеними значеннями на основі нейромережевих технологій. На відміну від відомих, запропонований метод усуває необхідність пошуку найкращої оцінки даних, а отже, економить час, що дозволяє підвищити точність і ефективність моніторингу стану складних об'єктів. Розроблено нейромережеві архітектури для дослідження задачі аналізу великих даних з пропущеними значеннями. Проведено експериментальні дослідження оцінки ефективності запропонованих рішень.

Результати роботи дозволять розробити ефективне програмне забезпечення для аналізу великих даних з пропущеними значеннями.

Ключові слова: АНАЛІЗ ВЕЛИКИХ ДАНИХ, ІМПУТАЦІЯ ДАНИХ, НЕЙРОННА МЕРЕЖА, МАШИННЕ НАВЧАННЯ.

ABSTRACT

Qualification work on the topic «Method for Analyzing Big Data with Missing Values Based on Neural Networks» for Master's degree on speciality 122 «Computer Science» educational and professional program «Computer Science» is written on 69 pages and it contains 7 figures, 3 tables, 1 annex and 77 sources.

The purpose of the qualification work is to improve the method of analyzing big data with missing values based on neural networks.

Research methods: analysis of literary sources and generalization of information; methods of neural networks; machine learning methods; methods of conducting experiments and comparing results; calculation, visualization and interpretation of comparative data on the results of experimental studies.

Research results: the method of analyzing big data with missing values based on neural network technologies has been improved. Unlike known methods, the proposed method eliminates the need to search for the best estimate of the data, and therefore saves time, which allows to increase the accuracy and efficiency of monitoring the condition of complex objects. Neural network architectures have been developed to investigate the problem of big data analysis with missing values. Experimental studies were conducted to evaluate the effectiveness of the proposed solutions.

The results of the work will allow to develop effective software for analyzing big data with missing values.

Keywords: BIG DATA ANALYSIS, DATA IMPUTATION, NEURAL NETWORK, MACHINE LEARNING.

ЗМІСТ

ВСТУП

Актуальність теми. Термін «Великі дані (Big Data)» останнім часом, а саме в 2020-ті зустрічається все частіше і частіше. Однак його поширення не говорить про те, що фахівці у різних сферах використовують його за призначенням. Повсюдне його вживання говорить про підвищену до нього увагу. Ці два слова використовують і в промисловості, і в програмуванні, і в маркетингу. Також це можна почути в бізнесі практично в будь-яких сферах через те, що у багатьох організацій накопичуються дані про клієнтів, статистичні показники власної роботи і т. д.

Крім того, великі дані це не тільки дані всередині самої компанії, вони можуть бути зовнішніми. Так, наприклад, вивчення виборців усередині конкретного штату США стає дедалі точнішим при використанні технологій Big Data.

Однак, зі зростанням використання даного терміну, зростає необхідність у таких професіях, як data scientist і data analyst. Дані професії передбачають роботу з даними, найчастіше з сирими. Варто зазначити, що багато хто, хто використовує вищезгаданий термін, не уявляє, що з так званими даними робити, тому виникла необхідність нових професій.

Найчастіше під великими даними мають на увазі великий масив даних, яку структурований чи структурований не точно або зовсім не структурований. Такий масив необхідно обробляти та робити висновки з отриманих результатів. Зі зростанням продуктивності комп'ютерів збільшуються і варіації аналізу, і швидкість його проведення, проте це потребує певних навичок. Також технічні можливості дозволяють зберігати дані, які обчислюються десятками та сотнями терабайтів (1 терабайт = 1024 гігабайт). Такі дані дозволяють виявляти не тільки статистичні закономірності та прогнозувати плани бюджетів та збільшувати прибуток, але й працювати з відео, аудіо, текстом та фото файлами.

Проблема, пов'язана з відсутністю даних для процесу прийняття рішень, більш очевидна в онлайн-додатках, де дані повинні використовуватися відразу після отримання. Методи обчислювального інтелекту, такі як нейронні мережі та інші методи розпізнавання образів, останнім часом стали дуже поширеними інструментами в процесах прийняття рішень. Коли деякі змінні не вимірюються, стає важко продовжувати прийняття рішень. Найбільша проблема у тому, що стандартні методи обчислювального інтелекту, що неспроможні обробляти вхідні дані з пропущеними значеннями і, отже, неспроможні виконувати задачі класифікації чи регресії.

Тому доцільно провести додаткові дослідження щодо можливості аналізу великих даних без відновлення відсутніх значень.

Мета і завдання дослідження. Метою кваліфікаційної роботи є вдосконалення методу аналізу великих даних з відсутніми значеннями на основі нейронних мереж.

Для досягнення поставленої мети необхідно виконати наступні завдання:

– дослідити основні визначення та характеристики великих даних;
– аналіз підходів для обробки та аналізу великих даних;
– провести аналіз методів відновлення відсутніх даних;
– зробити постановку задачі дослідження;
– дослідити теоретичні основи нечіткої нейронної мережі ARTMAP;
– дослідити теоретичні основи багатошарових персептронів;
– дослідити алгоритм зворотного поширення помилки для навчання багатошарових персептронів;
– запропонувати вдосконалений метод на основі нейромережевих технологій для аналізу великих даних з відсутніми значеннями;
– розробити нейромережеві архітектури для вирішення задачі аналізу великих даних з відсутніми значеннями;
– провести експериментальні дослідження вдосконаленого методу на основі нейромережевих технологій для аналізу великих даних з відсутніми

значеннями;

– провести порівняльний аналіз запропонованого методу на основі нейромережевих технологій для аналізу великих даних з відсутніми значеннями.

Об’єкт дослідження – процеси обробки великих даних.

Предмет дослідження – метод аналізу великих даних з відсутніми значеннями на основі нейронних мереж.

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

– аналіз літературних джерел та узагальнення інформації;

– методи нейронних мереж;

– методи машинного навчання;

– методи проведення експериментів і порівняння результатів;

– розрахунок, візуалізація та інтерпретація порівняльних даних про результати експериментальних досліджень.

Наукова новизна одержаних результатів. Вдосконалено метод аналізу великих даних з пропущеними значеннями на основі нейромережевих технологій. На відміну від відомих, запропонований метод усуває необхідність пошуку найкращої оцінки даних, а отже, економить час, що дозволяє підвищити точність і ефективність моніторингу стану складних об’єктів.

Практичне значення отриманих результатів. Розроблено нейромережеві архітектури для дослідження задачі аналізу великих даних з пропущеними значеннями. Проведено експериментальні дослідження оцінки ефективності запропонованих рішень. Отримані результати дозволять розробити ефективне програмне забезпечення для аналізу великих даних з пропущеними значеннями.

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. (випуск 72);

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

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

1 ОГЛЯД ОСНОВНИХ ВИЗНАЧЕНЬ ТА АНАЛІЗ МЕТОДІВ ОБРОБКИ І АНАЛІЗУ ВЕЛИКИХ ДАНИХ

1.1 Огляд основних визначень великих даних

Для того, щоб вивчити термін Big Data, необхідно звернутися до історії та вияснити, де виник даний термін і ким він був започаткований. Створення цієї фрази відносять до Lynch C. Ця людина є редактором журналу «Nature», який випустив особливий випуск у вересні 2008 року на тему «Як можуть вплинути на майбутнє науки технології, що відкривають можливості роботи з великими обсягами даних?» Цю дату пов'язують не з прямим створенням даного терміну, а швидше з його закріпленням та початком широкого використання. У цій статті обговорювалося величезне зростання обсягів даних, пов'язане зі щорічним збільшенням можливості зберігання все більшої та більшої кількості даних. Крім того, далі порушувалося питання про обробку цих даних та нові можливості використання даних, що не могло бути реальним 5 років тому тощо. Крім того, зростає можливість більш якісної обробки даних, навіть штучного прогнозування деяких об'єктів або ситуацій [40].

Спочатку цей термін застосовувався в академічному середовищі, проте з урахуванням можливостей, що надаються даною областю, не дивно, що його використання поширилося і в діловому середовищі в різних сферах, а також протягом декількох років з'явилися перші продукти, створені для роботи з великими даними та обробляти їх.

Не дивно, що консалтингова фірма Gartner відзначила великі дані другим за значимістю трендом у сфері інформаційних технологій.

Звернемося до термінології та з'ясуємо, що має на увазі під собою даний термін:

«…великі дані характеризуються по 3V теорії, а саме обсяг, різноманітність і швидкість, зокрема із генерацією та збиранням розмір даних, масштаб даних стають дедалі більшими, своєчасність великих даних, зокрема, збір та аналіз даних повинен проводитися швидко і своєчасно, а також різні

типи даних, які включають напівструктуровані та неструктуровані дані, а також традиційні структуровані дані» писав Laney D. у [32; 73].

«…нове покоління технологій та архітектур, розроблених для економічного вилучення цінності з дуже великих обсягів найрізноманітніших даних, забезпечуючи можливість високошвидкісного захоплення, відкриття та/або аналізу» описують у своїх роботах Gantz J. і Reinsel D. [18; 73].

«…дані, розмір яких перевищує можливості типових програмних засобів баз даних фіксувати, зберігати, керувати та аналізувати їх» так під великими наборами даних розуміють Manyika J. та ін. [41; 73].

«…набори, розміри яких перевищують можливості загальновживаних програмних засобів для збору, управління та обробки даних протягом проміжку часу» відносять до великих даних Mashingaidze K. та Backhouse J. [42; 73].

«…дані, для яких їх обсяг, швидкість збору або подання обмежує можливості використання традиційних реляційних методів для проведення ефективного аналізу або дані, які можуть бути ефективно оброблені за допомогою важливих технологій горизонтального масштабування» розуміють під великими даними у [50; 73].

«…великі дані передбачають не лише здатність обробляти великі обсяги даних, натомість вони представляють широкий спектр нових аналітичних технологій та бізнес-можливостей. Ці нові системи обробляють широкий спектр даних, починаючи від даних датчиків, закінчуючи даними Інтернету та соціальних мереж, покращеними аналітичними можливостями, оперативною бізнес-аналітикою, що покращує спритність бізнесу, дозволяючи автоматизовані дії в режимі реального часу та прийняття рішень протягом дня, швидше включаючи попит на програмне забезпечення як послугу. Підтримка великих даних передбачає поєднання апаратних та хмарних обчислень для створення нових рішень, які можуть принести значну користь бізнесу» White C. вказує у [66; 73].

«…великі дані: об’ємні, високошвидкісні та/або різноманітні

інформаційні ресурси, які потребують нових форм обробки, щоб забезпечити розширене прийняття рішень, виявлення розуміння та оптимізацію процесів» Beyer M. та Laney D. пишуть у [8; 73].

«…великі дані, як і аналітика до них, прагнуть отримати дані з даних і перетворити це на бізнес-перевагу, однак є три ключові відмінності: швидкість, різноманітність, об’єм» Gantz J. та Reinsel D. зазначають у [19; 73].

«…культурний, технологічний та науковий феномен, що спирається на взаємодію: технологій – максимізація обчислювальної потужності та алгоритмічної точності для збору, аналізу, зв’язку та порівняння великих наборів даних; аналізу – спираючись на великі набори даних для визначити закономірності для пред'явлення економічних, соціальних, технічних та юридичних вимог та міфологія –поширена думка, що великі масиви даних пропонують вищу форму інтелекту та знань, які можуть генерувати представлення, які раніше були неможливими, з аурою істини, об’єктивності та точності» Boyd D. та Crawford K. описують у [10].

«…великі дані – це поєднання обсягу, різноманітності, швидкості та достовірності, що створює можливість організаціям отримати конкурентні переваги на сучасному оцифрованому ринку» Schroeck M. та ін. пишуть у [60].

«…великі дані відносяться до наборів даних, розміри яких перевищують можливості звичайних програмних засобів збирати, управляти та обробляти дані протягом визначеного часу» визначають Bharadwaj A. та ін. [9; 73].

«…великі дані – це великомасштабні дані з різними джерелами та структурами, які неможливо обробити звичайними методами і призначені для вирішення організаційних або соціальних проблем», так пишуть Kamioka T. і Tapanainen T. у [29].

«…великі дані включають зберігання даних, управління, аналіз та візуалізацію дуже великих і складних наборів даних, основна увага приділяється новим методам управління даними, які перевершують традиційні реляційні системи, і краще підходять для управління великими обсягами даних соціальних медіа», у свою чергу, зазначають Bekmamedova N. та Shanks G. у

[7; 73].

«…великі дані складаються з великих наборів даних (великих обсягів), які оновлюються швидко і часто (з високою швидкістю) і містять величезний спектр різних форматів та вмісту (широкий вибір)» пише Davis C. [15].

«…набори даних з неоднорідних та автономних ресурсів, із різноманітністю вимірів, складними та динамічними взаємозв'язками, за розміром, що перевищує можливості звичайних процесів чи інструментів для їх ефективного захоплення, зберігання, управління, аналізу та використання», розуміють під великими даними Sun E. та ін. [64; 73].

«…великі дані, як правило, стосуються таких типів даних:

- традиційні корпоративні дані;

- машинно згенеровані дані / дані датчиків (наприклад, веб-журнали, розумні лічильники, виробничі датчики, журнали обладнання);

- соціальні дані» зазначають Opresnik D. і Taisch M. у [51; 73].

«…великі дані часто представляють різні записи про місцеперебування великої кількості людей, які базується на різних форматах та режимах спілкування (наприклад, текст, зображення та звук), що порушує серйозні проблеми перекладу та сумісності значень, великі дані зазвичай використовуються для посилання на великі обсяги даних, що генеруються та надаються в Інтернеті та поточних екосистемах цифрових медіа» стверджують Constantiou I. та Kallinikos J. у своїй роботі [13].

«…великі дані відрізняються від «звичайних» даних у чотирьох вимірах, або 4V – об'єм, швидкість, різноманітність та достовірність», Abbasi A. та ін. кажуть у [1].

«…великі дані визначаються з точки зору 5V: обсягу, швидкості, різноманітності, достовірності та цінності, де «обсяг» відноситься до обсягів великих даних, які збільшуються в геометричній прогресії; «швидкість» – це швидкість збору, обробки та аналізу даних у реальному часі; «різноманітність» відноситься до різних типів даних, зібраних у середовищах великих даних; «достовірність» означає надійність джерел даних; «цінність» відображає

трансакційні, стратегічні та інформаційні переваги великих даних», говорять Akter S. та Wamba S. [3].

Крім того, різні дослідники зосереджуються на різних аспектах великих даних, за даними джерела [45] (таблиця 1.1).

Таблиця 1.1 – Визначення характеристик великих даних [45]

Характеристика	Визначення
Обсяг	Обсяг являє собою просторий розмір набору даних через агрегування великої кількості змінних та ще більший набір спостережень для кожної змінної.
Швидкість	Швидкість відображає швидкість збору та аналізу даних у реальному часі чи майже в реальному часі від сенсорів, транзакцій продажів, публікацій у соціальних мережах та даних настроїв для останніх новин та соціальних тенденцій.
Різноманітність	Різноманітність великих даних походить від безлічі структурованих та неструктурованих джерел даних, таких як текст, відео, мережі та графіки, серед інших.
Достовірність	Достовірність гарантує, що використані дані є надійними, достовірними та захищені від несанкціонованого доступу та модифікації.
Цінність	Цінність представляє ступінь, в якій великі дані генерують економічно варті ідеї та / або вигоди завдяки вилученню та перетворенню.

Варіативність	Варіативність стосується того, як інформація постійно змінюється, оскільки одна і та ж інформація може трактуватися по-різному, або нові потоки з інших джерел допомагають сформувати інший результат.
Візуалізація	Візуалізація може бути описана як інтерпретація закономірностей та тенденцій, які присутні в даних.
3Vs: обсяг, швидкість, різноманітність. 4Vs: обсяг, швидкість, різноманітність, достовірність. 5Vs: обсяг, швидкість, різноманітність, достовірність, цінність. 7Vs: обсяг, швидкість, різноманітність, достовірність, цінність, варіативність, візуалізація.	

1.2 Аналіз підходів для обробки та аналізу великих даних

Якщо дані відповідають вищевказаним ознакам, їх можна вважати великими. Незважаючи на те, що сфера Big Data досить молода, вона через свою популярність вже стала, тому що в багатьох сферах різних компаній її вже використовують не на тестовому рівні, навіть у деяких державних сферах.

Необхідно вивчити велику кількість технічних аспектів у цій темі для того, щоб при вивченні специфіки застосування Big Data не виникло питань та труднощів.

Для збору даних існують різні технології та різне обладнання. Для початку варто розібрати основні підходи для обробки даних:

1. MapReduce – певна модель обробки даних. Дана модель розподіляє навантаження при дуже велику кількість даних і проводить обчислення паралельно. На стадії Map дані передобробляються та фільтруються. Користувач сам задає необхідні значення, якими буде проводитися фільтр даних. На стадії Reduce вибрані дані вже агреговані.

2. SQL – певна мова програмування, яка часто застосовується для створення та зміни даних під час роботи з реляційною (пов'язані значення) базою даних.

3. NoSQL – цей термін визначається як SQL. Техніки при цьому підході використовуються при базах даних, які мають відмінності від моделей, що використовуються в традиційних базах даних зі зв'язаними значеннями. Їх використовують, коли структура даних мінлива, наприклад, при збиранні інформації у соціальних мережах.

4. Hadoop – цей підхід використовується для високонавантажених сайтів, часто для пошукових механізмів. Наприклад, eBay, який є міжнародним майданчиком продажу та купівлі товарів. Ця система якраз стійка до відмов. Якщо порушиться робота деяких вузлів кластера, це не завадить роботі, оскільки дані копіюється іншому вузлі.

5. R – мова програмування, яка дуже часто використовується для роботи з Big Data, оскільки вона призначена для статистичної обробки даних.

Крім того, для роботи з даними потрібні не лише мови, а й певні сервери та обладнання. Деякі компанії вже пропонують комплексні послуги для компаній роботи з великими даними.

Після аналізу популярних підходів для обробки даних, слід звернутися до методик аналізу великих даних. Одним із найзвичніших є спосіб, пов'язаний з обробкою даних статистично, також там використовуються способи, взяті з інформатики. Аналіз статистично дуже затребуваний у різних галузях, і він був найдоступнішим, коли великі дані не були такі популярні і нові способи ще не створили. Варто враховувати, що чим більше даних і чим вони різноманітні, то більша ймовірність отримати релевантні дані після аналізу. За методиками аналізу великих даних можна назвати такі:

1. A/B тестування. Ця методика передбачає послідовне порівняння певної кількості елементів з іншим кількістю елементів. Цей тест показує, за яких параметрів змінюється важливий для користувача факт, наприклад,

збільшується кількість покупок або замовлень. Також цей метод застосовується не лише серед Big Data.

2. Data mining – комплекс методик, що дозволяє вивчати дані та виявляти раніше приховані знання, які можуть бути корисними для бізнесу тощо. Ці знання можуть бути нетривіальними або прихованими і їх необхідно вичленувати серед інших. Дані методики складають методи класифікації, моделювання та прогнозування. Наприклад, кластерний аналіз дозволяє класифікувати об'єкти за кластерами, групами, де у кожної групи буде ознака або ознаки. Також є класифікація, це набір методик, які дозволяють виявити певні знання, наприклад, оцінювання кредитоспроможності позичальників. Для цієї програми, клієнт – це ознак, наявність деяких із них веде до схвалення видачі кредиту, а наявність інших до заборони видачі. Також є методика асоціації, яка виявляє взаємозв'язки.

3. Genetic algorithms – за допомогою даної методики рішення розглядаються так само, як і за природного відбору. Деякі рішення найвиживаніші і найуспішніші, деякі отпадают.

4. Machine learning – великий комплекс методик, пов'язаний із аналізом даних через самонавчання на основі існуючих даних. Навчання з учителем передбачає, що з конкретної ситуації задається необхідне рішення. Алгоритм з'ясовує залежності між даними і видає вже те знання, яке можна надалі використовувати в іншому аналізі або його вже досить. Сюди входять досить популярні нейронні мережі. Навчання без вчителя означає, що для даних задається лише ситуація без рішення та алгоритму необхідно створити кластери з об'єктами, використовуючи наявні дані про взаємозв'язки об'єктів. Навчання без учителя та з учителем це найбільш використовувані методики у Machine Learning. Також є навчання із підкріпленням, випадок навчання з учителем.

1.3 Аналіз методів відновлення відсутніх даних

В якості фундаментальної технології аналізу великих даних використовується кластеризація, яка поділяє об'єкти на різні кластери на основі подібності [24, 70, 71]. Традиційні алгоритми кластеризації даних зосереджені на повній обробці даних, таких як кластеризація зображень [33], кластеризація звуку [20] та кластеризація тексту [2]. Гетерогенні методи кластеризації даних, останнім часом, цікавлять науковців [55, 58, 72].

Крім того, було запропоновано багато алгоритмів - наприклад, в [43] оптимізовано уніфіковану цільову функцію за допомогою ітераційного процесу, і розроблено алгоритм спектральної кластеризації для кластеризації різнорідних даних на основі теорії графів. У [34] запропоновано алгоритм нечітких c-середніх значень високого порядку для розширення звичайного алгоритму нечітких c-середніх з векторного простору на тензорний простір. Для кластеризації даних в системах Інтернету речей (IoT) запропоновано алгоритм c-середніх значень високого порядку, заснований на тензорах [69]. Ці алгоритми ефективні для покращення продуктивності кластеризації неоднорідних даних. Однак вони можуть отримати лише результати кластеризації та позбавлені подальшого аналізу неповних даних з низькими розмірами. Тому їх продуктивність обмежена неоднорідними даними у середовищі великих даних. Що ще важливіше, інші існуючі алгоритми кластеризації об’єктів не враховують відновлення даних та відсутність даних.

Метод середнього заміщення (MS) дозволяє вирішити проблему неповноти даних, замінюючи кожну відсутність середньою змінною. Типи MS - це підміна медіани, моди, середнього значення підгрупи даних [31], значення з найвищою частотою (найбільш загальне значення), в деяких випадках замінює мінімальне/максимальне значення. Цей метод може призвести до багатьох небажаних результатів, таких як заниження реальної дисперсії, негативне упередження кореляцій та неправильне представлення загальної сукупності [37].

Проста hot-deck імпутація замінює кожне відсутнє значення випадковим значенням, взятим із існуючого набору даних [47]. Його суттєвим недоліком є спотворення кореляцій.

Імпутація Cold-deck (CD) реалізує заміну кожного проходу на деяке постійне значення із зовнішнього джерела [62]. Особливим випадком CD є нульова підміна. Як і послідовна та випадкова гаряча заміна, вона має ті ж недоліки, що і проста гаряча заміна.

Перевага підходів на основі заміщення полягає в тому, що вони дають внутрішньо узгоджений набір значень, але не можуть визначити взаємозв'язки між змінними. Методи заповнення прогалин на основі моделі оцінюють значення параметрів моделі, які послідовно використовуються для обчислення прогалин. Ці методи враховують взаємозв'язки між змінними, але вони досить складні, оскільки спричиняють помилки, коли всі параметри оцінюються одночасно.

Оцінка регресії (RE) передбачає заміщення прогалин даних прогнозованими значеннями, отриманими з рівняння регресії, побудованого з повного набору даних. Недоліки заповнення регресією включають наступне: необхідність точно ідентифікувати моделі регресії, перебільшення кореляції та коваріації, ймовірність виходу прогнозованих значень за межі логічного ряду та необхідність великих обсягів даних для отримання послідовних оцінок.

Метод максимальної ймовірності (ML) виводить оцінені параметри таким чином, що ймовірність відтворення даних на основі відомих значень максимізується. Метод ML розглядається як один із підходів до обробки неповних даних [52], і його широко рекомендують використовувати, оскільки він не призводить до упередженості при наявності відсутніх схем значень MCAR та MAR [49]. Однак метод ML не вирішує цього питання за наявності MNAR, хоча величина такого упередження, як правило, набагато менша, ніж традиційні підходи.

Метод максимізації очікувань (EM) заснований на загальному ітераційному алгоритмі, який реалізує послідовне заповнення відсутніх значень за їхніми оцінками та максимізує умовне сподівання ймовірності отримання повних даних до процесу конвергенції [49]. Основним недоліком методів ML та EM є їх ітераційна властивість і висока часова складність в подальшому, особливо для обробки великих даних.

Метод k-найближчого сусіда (kNN) заснований на визначенні найближчого сусіда для кожного пропущеного значення за допомогою певної метрики [36]. Після пошуку k-найближчих сусідів за кожний прохід воно замінюється середнім значенням характеристик для сусідів, що містять повні дані. Окремим випадком методу kNN є зважений метод kNN [6]. Найбільша складність у його застосуванні полягає у визначенні адекватного ступеня близькості.

Алгоритм опорних векторних машин (SVM) [4] дозволяє уникнути оцінки ймовірності даних, які є стабільними. Вибір ядра впливає на якість обчислення відсутніх даних. Ось чому реалізація цього алгоритму залежить від набору даних та його параметрів.

Метод випадкових лісів (RF) [63] заснований на побудові класифікатора, сформованого групою дерев рішень, навчальним набором повних даних та прогнозуванням відсутніх значень у наборі тестів. На першому кроці всі прогалини в наборі тестів заповнюються середнім значенням на основі наявних даних. Далі розраховується матриця близькості для першого набору даних. Середні ваги, розраховані на матриці близькості для першої ітерації, використовуються для заміщення пропусків на наступній ітерації і т. д. до досягнення критерію зупинки [53]. Хоча цей метод є ітеративним, він дозволяє розпаралелювати обробку за рахунок зменшення часової складності.

Метод асоціативних правил (AR) передбачає побудову асоціативних правил щодо надмірності наявних даних [54, 61]. По-перше, усі отримані правила сортуються за якимись параметрами (наприклад, автентичністю). Потім здійснюється пошук набору даних за кожним пропуском та відповідним

правилом, яке не суперечить іншим значенням, що знаходяться у відновлюваному рядку. Алгоритм AR має кілька модифікацій. Наприклад, один з них, FP-алгоритм може ефективно паралельно працювати, саме тому його можна використовувати для попередньої обробки великих даних. Однак часову складність цього методу доводиться вдосконалювати.

Метод багатошарового персептрону (MLP) використовує навчену мережу для оцінки відсутніх значень. Оскільки вхідними даними є непусті набори даних, виходами є неповні значення змінних, які потрібно визначити [27, 49, 59]. Заповнення відсутніх значень за допомогою MLP складається з трьох етапів:

– формування повного (навчального) та неповного (тестового) набору;

– побудова MLP на навчальному наборі, де значення змінної є значеннями прогалин, а вхідними значеннями, відповідно, є заповнені значення змінної;

– прогнозування невідомих значень для кожної неповної структури даних за допомогою навченої мережі.

Недоліком цього підходу є прив'язка до набору змінних, що містять відсутні значення, тому потрібно будувати окрему модель MLP для кожної комбінації.

Метод самоорганізованої карти (SOM) [28] дозволяє навчати мережу на основі неповного набору даних. Він запускається завдяки здатності алгоритму ігнорувати відсутні значення та обчислювати відстань між поточним спостереженням і вузлом та використовувати навчену карту для оцінки значень. Недолік SOM такий же, як і MLP.

Глибоке навчання (DL) активно застосовується у багатьох додатках завдяки його потужній здатності до імпутації даних [46]. Глибоко вбудована кластеризація (DEC) вчиться перетворювати з простору даних у простір об’єктів низьких розмірів [67].

У [30] показано здатність представлення ознак автокодера (VAE). VAE вивчає багатогранну структуру даних і досягає високих показників кластеризації [35]. Крім того, VAE має потужну здатність до вилучення та реконструкції функцій, і це може бути хорошим інструментом для обробки неповних даних.

У роботі [68] запропоновано алгоритм VAE, який може покращити ефективність кластеризації мультимедійних даних. На відміну від багатьох існуючих технологій, цей алгоритм вивчає функції даних, проектуючи вдосконалену мережу VAE, і використовує fuzzy c-means алгоритм на основі тензора для кластеризації даних у просторі функцій.

У роботі [56] пропонується розширення варіаційного автокодера (VAE), який називається варіаційним автокодером із зваженими втратами (VAE-WL). Він використовує спеціальну функцію втрат, яка більше підходить для обчислення відсутніх значень.

На основі проведеного дослідження можна зробити висновки, що, 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.

Таким чином, на сьогодні, 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.

1.4 Постановка задачі дослідження

На основі проведеного дослідження можна зробити висновки, що, 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.

Таким чином, на сьогодні, 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 того, щоб ефективно обробляти великі обсяги даних при прийнятних часових затратах, необхідні особливі технології. Такими технологіями можна вважати нейронні мережі, які мають велику ефективність нелінійного перетворення і представлення даних в порівнянні з традиційними методами.

Отже, метою кваліфікаційної роботи є вдосконалення методу аналізу великих даних з відсутніми значеннями на основі нейронних мереж.

Для досягнення поставленої мети необхідно виконати наступні завдання:

– дослідити основні визначення та характеристики великих даних;
– аналіз підходів для обробки та аналізу великих даних;
– провести аналіз методів відновлення відсутніх даних;
– зробити постановку задачі дослідження;
– дослідити теоретичні основи нечіткої нейронної мережі ARTMAP;
– дослідити теоретичні основи багатошарових персептронів;
– дослідити алгоритм зворотного поширення помилки для навчання багатошарових персептронів;
– запропонувати вдосконалений метод на основі нейромережевих технологій для аналізу великих даних з відсутніми значеннями;

– розробити нейромережеві архітектури для вирішення задачі аналізу великих даних з відсутніми значеннями;

– провести експериментальні дослідження вдосконаленого методу на основі нейромережевих технологій для аналізу великих даних з відсутніми значеннями;

– провести порівняльний аналіз запропонованого методу на основі нейромережевих технологій для аналізу великих даних з відсутніми значеннями.

Висновки до розділу 1

1. Проведено огляд основних визначень великих даних та зроблено деякі висновки. Великі дані – серія підходів, інструментів і методів обробки структурованих і неструктурованих даних величезних обсягів, і значного різноманіття для отримання результатів, що сприймаються людиною, ефективних в умовах безперервного приросту, розподілу по численних вузлах обчислювальної мережі. Основними характерними ознаками є великий обсяг, вкрай висока швидкість накопичення, різноманіття даних, достовірність та висока цінність інформації.

2. Для обробки та аналізу великих даних часто використовуються наступні технології: MapReduce, Hadoop, мова R та інше. Крім того, Аналізують Big Data за допомогою Machine Learning, Data Mining та інші. Для того, щоб ефективно обробляти великі обсяги даних при прийнятних часових затратах, необхідні особливі технології. Такими технологіями можна вважати нейронні мережі, які мають велику ефективність нелінійного перетворення і представлення даних в порівнянні з традиційними методами.

3. 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.

2 ТЕОРЕТИЧНІ ОСНОВИ АНАЛІЗУ ВЕЛИКИХ ДАНИХ З ВІДСУТНІМИ ЗНАЧЕННЯМИ НА ОСНОВІ НЕЙРОННИХ МЕРЕЖ

2.1 Теоретичні основи нечіткої нейронної мережі ARTMAP

Fuzzy ARTMAP – це архітектура нейронної мережі, розроблена Carpernter et al. у 1992 році та заснована на теорії адаптивного резонансу (Adaptive Resonance Theory – ART). Fuzzy ARTMAP використовувалася для моніторингу стану складних об’єктів Javadpour та Knapp [26], але їх застосування не було онлайн.

Архітектура Fuzzy ARTMAP здатна до швидкого онлайнового контрольованого поетапного навчання, класифікації та прогнозування [11]. Нечіткий ARTMAP працює шляхом поділу вхідного простору на кілька гіпербоксів, які відображаються у вихідний простір. Використовується навчання на основі екземплярів, де кожен окремий вхід зіставляється з міткою класу.

Для управління процесом навчання використовується три параметри, а саме пильність $\rho \in [0, 1]$, швидкість навчання $\beta \in [0, 1]$ та параметр вибору α. Параметр вибору зазвичай роблять невеликим, наприклад 0,01. Параметр β управляє швидкістю адаптації, де 0 означає повільну швидкість, а 1 найвищу. Якщо $\beta = 1$, гіпербокси збільшуються, щоб увімкнути точку, представлену вхідним вектором. Пильність є ступенем приналежності і контролює, наскільки великим може стати будь-який гіпербокс, що призводить до формування нових гіпербоксів. Більші значення ρ можуть призвести до формування менших гіпербоксів і можуть призвести до «проліферації категорій», що можна розглядати як надмірне навчання.

Перевага надається нечіткому ARTMAP через його здатність до поступового навчання. У міру вибірки нових даних не буде потреби перенавчати мережу, як у випадку з MLP.

Для великих наборів даних час обробки та обсяг пам'яті створюють декілька проблем. Поступове навчання стає привабливою рисою, оскільки

навчання не обов'язково має проводитися одразу [5]. Таким чином, нейронна мережа, яка використовується у великих наборах даних, повинна [5, 73]:

– вміти отримувати додаткову інформацію з нових даних;

– зберегти раніше набуті знання;

– бути в змозі врахувати нові категорії даних, які можуть бути введені з новими даними;

– не вимагає доступу до вихідних даних, які використовуються для початкового навчання системи.

Нечітка ARTMAP пропонує всі ці характеристики і, як наслідок, їй буде віддана перевага. Fuzzy ARTMAP також вважається однією з провідних архітектур нейронних мереж, як обговорювалося [73].

Fuzzy ARTMAP використовувалася при розробці реактивної моделі автомобіля на основі агентів і застосовувалася так само, як і методи зворотного поширення, тоді як Fuzzy ARTMAP пропонує можливість поступового навчання.

Taylor and Wolf [73] також використовували Fuzzy ARTMAP для оптимізації на основі агентів, і результати продемонстрували хороше узагальнення та швидке навчання, ніж інші класифікації. Завдяки успіху в різних програмах, нечітка ARTMAP буде використовуватися в даній роботі.

ARTMAP має мережу, яка здатна засвоювати нові знання, зберігаючи при цьому знання або інформацію, які вже були надані раніше.

Структура Fuzzy ARTMAP показана на рисунку 2.1.

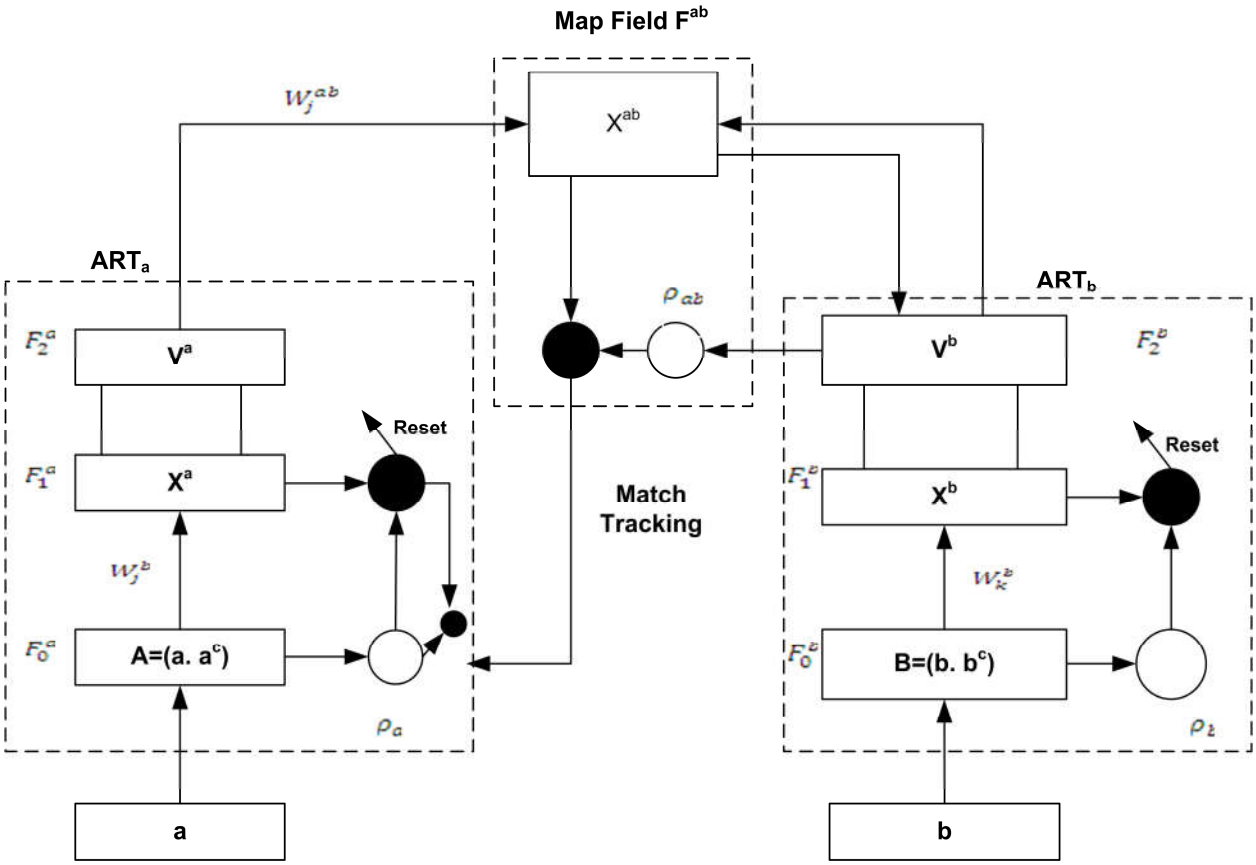


Рисунок 2.1 – Структура нечіткої нейронної мережі ARTMAP

Параметри ART мають функціональні можливості, визначені таким чином [39]:

– F_0 представляє вхідний рівень, який отримує та зберігає нові вхідні шаблони;

– F_1 представляє рівень порівняння, головною функціональністю якого є усунення шуму на вході;

– F_2 – це рівень розпізнавання, який зберігає категорії, до яких потрапляють вхідні дані.

ART має дві вимоги до пам’яті, які допомагають у його функціональності. Ці пам’яті називаються довгостроковою пам’яттю (LTM) і короткочасною пам’яттю (STM). STM відповідає за моделі активності, розроблені в шарах F_1 і F_2. Продуктивність мережі ART пояснюється наступним чином:

1. Вхідний шаблон подається на мережу.

2. Необхідно перевірити подібність вхідного шаблону до шаблону, який уже зберігається в LTM мережі.

(a) ящо така подібність існує, то шаблон уже відомий;

(b) в іншому випадку вхідний шаблон не відноситься до жодної вже сформованої категорії, і мережа сформує нову категорію, яка зберігатиме новий вхідний шаблон.

Модуль Fuzzy ART класифікує вхідні вектори за категоріями, що складаються з аналогових даних, які перетворюються на двійкові значення активним конвертером коду. Вхідний набір нечіткого модуля ARTMAP складається з двійкових даних.

Вхід.

Вхідним шаблоном ART_a є вектор $a = [a_1, a_2, ..., a_{Ma}]$, де M_a є розмірністю, а вхідні дані для другого ART, позначеного *як* ART_b, є $b = [b_1, b_2, ..., B_M]$.

Параметри.

Є три основні параметри, які важливі для продуктивності та навчання нечіткої мережі ARTMAP. Ці параметри наведено нижче:

– обраний параметр α, який діє на вибір категорії та задовольняє умову $\alpha > 0$;

– швидкість навчання $\beta \in [0, 1]$, яка контролює швидкість, з якою мережі адаптуються;

– параметр пильності ($\rho \in [0, 1]$] і керує резонансом мережі. Він також відповідає за кількість категорій, сформованих у F_2.

Якщо параметр пильності встановлено дуже великим, це дає хорошу класифікацію з багатьма категоріями. Навпаки, якщо цей параметр встановлений дуже низьким, мережа має гарне узагальнення [39].

Алгоритм.

Нехай J буде активною категорією для модуля ART_a, а K буде активною категорією для ART_b. Використовуючи процес, який називається відстеженням відповідності, активна категорія ART_a перевіряється на відповідність

бажаному результату в модулі ART_b. Якщо вони збігаються, наступним кроком є перевірка умови пильності, що задається виразом

$$\frac{|y^b \wedge w_{JK}^{ab}|}{|y^b|} \geq \rho_{ab}$$

де y^b – вихідний вектор модуля ART_b. Коли резонанс не досягається, параметр пильності трохи збільшується, щоб виключити поточну категорію та вибрати іншу категорію, яка буде використана. Цю процедуру повторюють, доки не буде виконано попереднє рівняння.

Оператор ∧ визначається як

$$(p \wedge q) \equiv min(p_i, q_i)$$

Норма |.| визначається

$$|p| \equiv \sum_{i=1}^{M} |p_i|$$

для *M*-вимірного вектору [11].

Навчання.

Усі ваги спочатку мають значення 1, що вказує на відсутність активної категорії. Навчання відбувається в процесі адаптації ваг. Адаптація ART_a та ART_b представлена

$$w_J^{new} = \beta(I \wedge w_J^{old}) + (1 - \beta)w_J^{old}$$

де J представляє активну категорію. Параметр β контролює швидкість навчання, де $\beta = 1$ дає найшвидшу швидкість адаптації, а $0 < \beta < 1$ дозволяє вагам адаптуватися повільно.

Це впливає на адаптацію модуля inter-ART

$$w_{JK}^{ab} = 1$$

$$w_{JK}^{ab} = 1 \quad for \quad k \neq K$$

Припустимо, що вхід і вихід позначено I і O відповідно. Першим кроком у навчанні є обчислення висхідних вхідних даних для кожного вузла j у F_2^a наступним чином [12]:

$$T_j^a = \frac{|\mathbf{I}^r \wedge \mathbf{w}_j^a|}{|\mathbf{w}_j^a| + \beta_a}$$

Потім вузол j_{max} перевіряється, чи відповідає він критерію пильності. Наступний крок виконується, лише якщо критерій пильності пройдено. Цей вузол додатково перевіряється для тестування прогнозування, де перевіряється, чи вузол відповідає точному вихідному вектору O. Усі ці операції повторюються, доки тестування прогнозування не буде пройдено.

2.2 Теоретичні основи багатошарових персептронів

В даному розділі розглядаються багатошарові персептрони (multilayer per-ceptrons), які є багатошаровими нейронними мережами з прямим поширенням сигналів (multilayer feedforward neural networks).

Багатошаровий персептрон здатний здійснити будь-яке відображення вхідних векторів у вихідні. Цю властивість багатошарових нейронних мереж відзначали ще у 60-х роках Мінський та Пайперт. Однак на той момент часу не було ефективного алгоритму навчання таких мереж. Тому їх висновки щодо перспектив розвитку багатошарових персептронів були дуже песимістичні.

У 1986 р. ряд авторів (Rumelhart, Hinton, Williams) незалежно один від одного запропонували алгоритм зворотного поширення помилки (backpropagation algorithm), який став ефективним засобом навчання багатошарових персептронів [57].

До 2006 року в науковому середовищі була пріоритетною парадигма, що персептрон з одним або максимум двома прихованими шарами є достатнім для вирішення різних задач. Використання персептрона з більш ніж двома прихованими шарами немає великого сенсу. В даний час пріоритетною є парадигма, що персептрон з більш ніж двома прихованими шарами є більш ефективним. У результаті персептрони відродилися під новою назвою глибокі нейронні мережі (deep neural networks), які в загальному випадку є персептроном з більш ніж двома прихованими шарами. Завдяки багатошаровій архітектурі вони дозволяють обробляти та аналізувати великі обсяги даних, а також моделювати різноманітні когнітивні процеси [57].

Архітектура багатошарового персептрона складається з багатьох шарів нейронних елементів (рисунок 2.2).

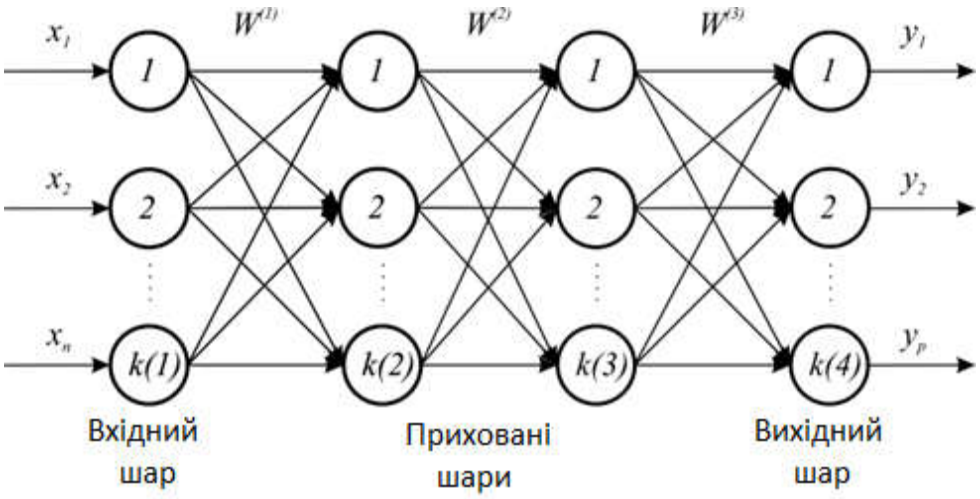


Рисунок 2.2 – Багатошаровий персептрон

Вхідний шар виконує розподільні функції. Вихідний шар (output layer) служить для обробки інформації, отриманої від попередніх шарів, і видачі остаточних результатів. Шари нейронних елементів, розташовані між вхідним та вихідним шарами, називаються проміжними або прихованими (hidden layers). Як і вихідний, приховані шари обробляють інформацію. Вихід кожного нейронного елементу попереднього шару нейронної мережі з'єднаний синаптичними зв'язками з усіма входами нейронних елементів наступного шару. Таким чином, топологія багатошарової нейронної мережі є однорідною та регулярною [21–23, 57].

Нехай $W^{(i)}$ – матриця вагових коефіцієнтів i-го шару багатошарового персептрону. Тоді вихідні значення Y для нейронної мережі з двома прихованими шарами можна визначити так:

$$Y = F(F(F(XW^{(1)})W^{(2)})W^{(3)}),$$

де $X = (x_1, \ x_2, \ \ldots, \ x_n)$ – вектор вхідних сигналів;

F – оператор нелінійного перетворення.

Загальна кількість синаптичних зв'язків багатошарового персептрону дорівнює:

$$V = \sum_{i=1}^{p-1} k(i) \cdot k(i+1) + \sum_{i=1}^{p-1} k(i+1),$$

де p − загальна кількість шарів нейронної мережі;

$k(i)$ − кількість нейронних елементів у і-му шарі.

Кількість шарів характеризує здатність багатошарової нейронної мережі до розбиття вхідного простору образів на підпростори меншої розмірності. Персептрон з одним прихованим шаром і нелінійною функцією активації нейронних елементів дозволяє формувати в просторі рішень будь-які розділяючі поверхні, як опуклі, так і неопуклі. За допомогою персептрона з двома та більш прихованими шарами також можна отримувати області рішень будь-якої форми та складності.

Розглянемо можливості багатошарового персептрону залежно кількості в ньому прихованих шарів. З цього питання у літературі існує безліч нюансів. Так, у роботі [38] стверджується, що персептрон з одним прихованим шаром може формувати тільки опуклу розділяючу поверхню з точки зору класифікації образів. Хоча пізніше було показано, що такий персептрон може формувати неопуклу поверхню. Перш за все, слід зазначити, що можливості персептрону з одним прихованим шаром різняться в залежності від функції активації, що використовується в ньому. Розглянемо персептрони з різною кількістю шарів.

У 1957 Колмогоров доказав, що будь-яку безперервну функцію *n* змінних на одиничному відрізку [0,1] можна представити у вигляді суми кінцевого числа одновимірних функцій:

$$f(x_1, x_2, \dots, x_n) = \sum_{p=1}^{2n+1} g(\sum_{i=1}^{n} \lambda_i \phi_p(x_i)),$$

де g і ϕ_p – є одновимірними та безперервними функціями;

$\lambda_i = const$ для всіх i.

Дане твердження лягло основою побудови багатошарових нейронних мереж для апроксимації функцій. З нього випливає, що будь-яку безперервну функцію $f\colon [0,1]^n \to [0,1]$ можна апроксимувати нейронною мережею з одним прихованим шаром, якщо вона має n вхідних нейронів, $(2n+1)$ прихованих і один вихідний. Основна проблема в даному випадку пов'язана з вибором відповідних функцій g і ϕ_p. У 1989 р. ряд авторів [14, 25] узагальнили наведені вище результати на багатошаровий персептрон з одним прихованим шаром. Отримані ними результати можна подати у вигляді наступної теореми.

Персептрон з одним прихованим шаром є універсальним апроксиматором, тобто він здатний з будь-яким ступенем точності апроксимувати будь-яку безперервну функцію, якщо в якості функції активації нейронних елементів прихованого шару використовується безперервна, монотонно зростаюча, обмежена функція. При цьому точність апроксимації функції залежить від кількості нейронів у прихованому шарі. Чим більша кількість нейронів, тим більша точність апроксимації. Однак, при занадто великій розмірності прихованого шару може наступити явище, яке називається перетренуванням мережі, коли мережа має погану узагальнюючу здатність [14, 25].

Наведена вище теорема є основою апроксимації і класифікації функцій з допомогою багатошарових нейронних мереж. З теореми випливає, що в якості функцій активації нейронних елементів можуть використовуватися сигмоїдні функції активації: сигмоїдна, біполярна сигмоїдна та гіперболічний тангенс.

Персептрон з одним прихованим шаром та безперервною, монотонно зростаючою, обмеженою функцією активації нейронних елементів є універсальним класифікатором.

Проведемо аналіз персептрону з одним прихованим шаром в якості класифікатора при використанні порогової функції активації нейронних елементів (рисунок 2.3).

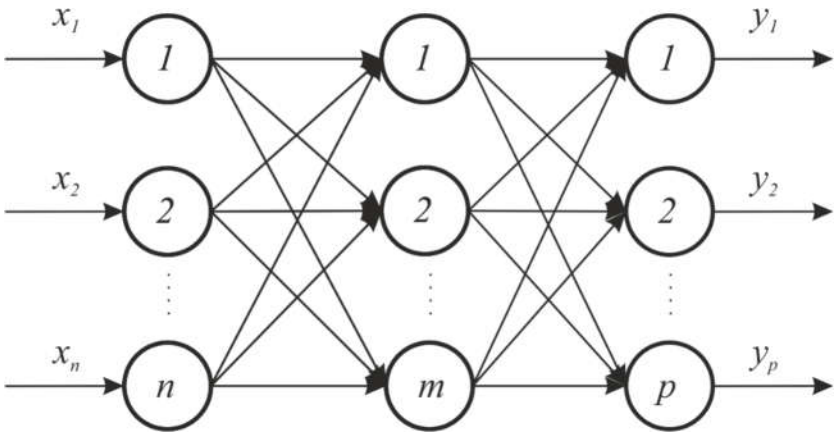


Рисунок 2.3 – Персептрон з одним прихованим шаром

Нейрони прихованого шару є локальними детекторами ознак. Вони розбивають вхідний простір образів на області за допомогою гіперплощин (прямих у двовимірному випадку). Число нейронів у скритому шарі характеризує кількість таких областей:

$$c \leq 2^m.$$

Вихідний шар нейронних елементів здійснює об'єднання одержаних областей у відповідні класи.

Отже, персептрон з одним прихованим шаром при використанні безперервної, монотонної, зростаючої обмеженої функції активації є універсальним апроксиматором. Однак при занадто великій розмірності прихованого шару може наступити перетренування мережі (overfitting), коли вона має погану узагальнюючу здатність. Тому, аж до 2006 року вважалося, що якщо на навчальній вибірці за допомогою одного прихованого шару не вдається отримати необхідну точність апроксимації функції та прийнятну узагальнюючу здатність, то необхідно переходити до персептрону з двома прихованими шарами. Використання персептрону з більш ніж двома прихованими шарами вважалося неефективним.

Розглянемо функції різних шарів нейронних елементів. Нейрони першого прихованого шару є локальними детекторами ознак. Вони виділяють

примітивні ознаки із вхідного простору образів. Другий прихований шар нейронних елементів детектує простір ознак вищого рівня абстракції на основі композиції ознак першого прихованого шару. Вихідний шар нейронних елементів відображає простір ознак прихованого шару у відповідні класи.

2.3 Алгоритм зворотного поширення помилки для навчання багатошарових персептронів

Алгоритм зворотного поширення помилки був запропонований в [57] і є ефективним засобом навчання багатошарових нейронних мереж.

Розглянемо нейронну мережу, що складається із чотирьох шарів (рисунок 2.4). Позначимо шари нейронних елементів від входу до виходу відповідно.

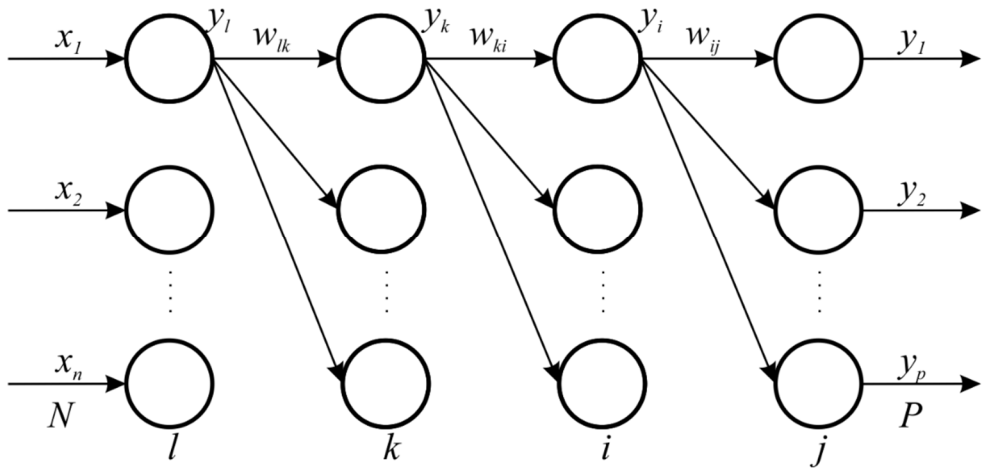


Рисунок 2.4 – Нейронна мережа з двома прихованими шарами

Вихідне значення *j*-го нейрону останнього шару дорівнює:

$$y_j = F(S_j),$$

$$S_j = \sum_i \omega_{ij} y_i - T_j,$$

де S_j– зважена сума *j*-го нейрона вихідного шару;

y_i – вихідне значення *i*-го нейрона передостаннього шару;

ω_{ij} і T_j – відповідно ваговий коефіцієнт і поріг *j*-го нейрону вихідного шару.

Аналогічно, вихідне значення *i*-го нейрону передостаннього шару визначається, як:

$$y_i = F(S_i),$$

$$S_i = \sum_k \omega_{ki} y_k - T_i.$$

Відповідно для *k*-го шару:

$$y_k = F(S_k),$$

$$S_k = \sum_l \omega_{lk} x_l - T_k.$$

Алгоритм зворотного поширення помилки мінімізує сумарну квадратичну помилку нейронної мережі, яка характеризує різницю між реальними і еталонними вихідними значеннями для всіх образів навчальної вибірки:

$$E_s = \frac{1}{2} \sum_{k=1}^{L} \sum_{j=1}^{p} ({y_j}^k - e_j^k)^2 ,$$

де ${y_j}^k$ і e_j^k – відповідно вихідне та еталонне значення нейронної мережі для *k*-го образу;

L – розмірність навчальної вибірки.

Для мінімізації сумарної квадратичної помилки мережі будемо використовувати метод градієнтного спуску в просторі вагових коефіцієнтів і

порогів нейронної мережі. Існує послідовне (online learning) і групове (batch learning) навчання [57].

Висновки до розділу 2

1. Архітектура Fuzzy ARTMAP здатна до швидкого онлайнового контрольованого поетапного навчання, класифікації та прогнозування. Нечіткий ARTMAP працює шляхом поділу вхідного простору на кілька гіпербоксів, які відображаються у вихідний простір. Використовується навчання на основі екземплярів, де кожен окремий вхід зіставляється з міткою класу.

2. Глибокі нейронні мережі (deep neural networks) в загальному випадку є персептроном з більш ніж двома прихованими шарами. Завдяки багатошаровій архітектурі вони дозволяють обробляти та аналізувати великі обсяги даних, а також моделювати різноманітні когнітивні процеси.

3. Алгоритм зворотного поширення помилки є ефективним засобом навчання багатошарових нейронних мереж. Для мінімізації сумарної квадратичної помилки мережі використовується метод градієнтного спуску в просторі вагових коефіцієнтів і порогів нейронної мережі. Існує послідовне (online learning) і групове (batch learning) навчання.

3 РЕАЛІЗАЦІЯ ТА ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ МЕТОДУ АНАЛІЗУ ВЕЛИКИХ ДАНИХ З ВІДСУТНІМИ ЗНАЧЕННЯМИ НА ОСНОВІ НЕЙРОННИХ МЕРЕЖ

В даному розділі досліджується задача моніторингу стану, коли методи обчислювального інтелекту використовуються для класифікації та регресії за наявності відсутніх даних без фактичного прогнозування відсутніх значень. Представлено новий підхід, коли не робляться спроби відновити відсутні значення як для регресії, так і для задач класифікації. Запропоновано ансамбль класифікаторів fuzzy-ARTMAP для класифікації за наявності відсутніх даних. Запропонований підхід розширено до задачі регресії, де MLP використовуються для отримання правильного результату з обмеженими вхідними змінними. Запропонований метод порівнюється з підходом, який поєднує нейронні мережі з генетичним алгоритмом для апроксимації відсутніх даних.

3.1 Алгоритм обробки вхідних даних із відсутніми значеннями на основі нечіткої нейронної мережі ARTMAP

Запропонований метод використовує ансамбль нейронних мереж для вирішення задач класифікації та регресії за наявності відсутніх даних. Підходи, засновані на ансамблі, були добре досліджені, і було виявлено, що вони покращують ефективність класифікації в різних програмах [17]. Потенціал використання підходу на основі ансамблю для вирішення задачі відсутніх даних залишається невивченим як у задачах класифікації, так і в задачах регресії. У запропонованому методі виконується пакетне навчання, тоді як тестування проводиться в режимі онлайн. Навчання здійснюється за допомогою кількох нейронних мереж, кожна з яких навчається з різною комбінацією функцій. Для системи моніторингу стану, яка містить n датчиків, користувач повинен вказати значення n_{avail}, яке є кількістю функцій, які,

швидше за все, будуть доступні в будь-який момент часу. Таку інформацію можна отримати з датчиків, як зазначено виробниками. Такі специфікації, як середній час між відмовами (Mean-time-between-failures - MTBF) і середній час до відмови (Mean-time-to-failure - MTTF), стають більш корисними для виявлення того, які датчики найімовірніше вийдуть з ладу.

Коли визначено кількість датчиків, які, найімовірніше, будуть доступні, можна розрахувати кількість всіх можливих мереж, використовуючи:

$$\mathbf{N} = \begin{pmatrix} n \\ \\ n_{avail} \end{pmatrix} = \frac{n!}{n(n - n_{avail})!} \tag{3.1}$$

де N – загальна кількість усіх можливих мереж, n – загальна кількість функцій, а n_{avail} – кількість функцій, які, швидше за все, будуть доступні в будь-який час.

Хоча n_{avail} можна обчислити статистично, воно впливає на кількість доступних мереж. Розглянемо простий приклад, коли вхідний простір має п’ять функцій, позначених як: *a*, *b*, *c*, *d* і *e*, і є 3 функції, які, швидше за все, будуть доступні в будь-який час. Використовуючи рівняння (3.1), змінна N дорівнює 10. Ці класифікатори будуть навчені ознаками [*abc, abd, abe, acd, ace, ade, bcd, bce, bde, cde*].

У випадку, коли одна змінна відсутня, скажімо, для тестування можна використовувати лише чотири мережі, і це класифікатори, які не використовують у своїй навчальній вхідній послідовності. Якщо існує ситуація, коли відсутні дві змінні, скажімо, *a* і *b*, можна використовувати лише один класифікатор. У результаті кількість класифікаторів зменшується зі збільшенням кількості відсутніх входів.

Кожна нейронна мережа навчається за допомогою n_{avail} функцій. Потім виконується процес перевірки, а результат використовується для прийняття

рішення щодо схеми комбінування. Процес навчання вимагає наявності повних даних, оскільки навчання проводиться в автономному режимі. Доступний набір даних поділяється на «навчальний набір» і «тестовий набір». Кожна створена мережа перевіряється на тестовому наборі та отримує вагу відповідно до її продуктивності на тестовому наборі.

Схематична ілюстрація запропонованого ансамблевого підходу представлена на рисунку 3.1.

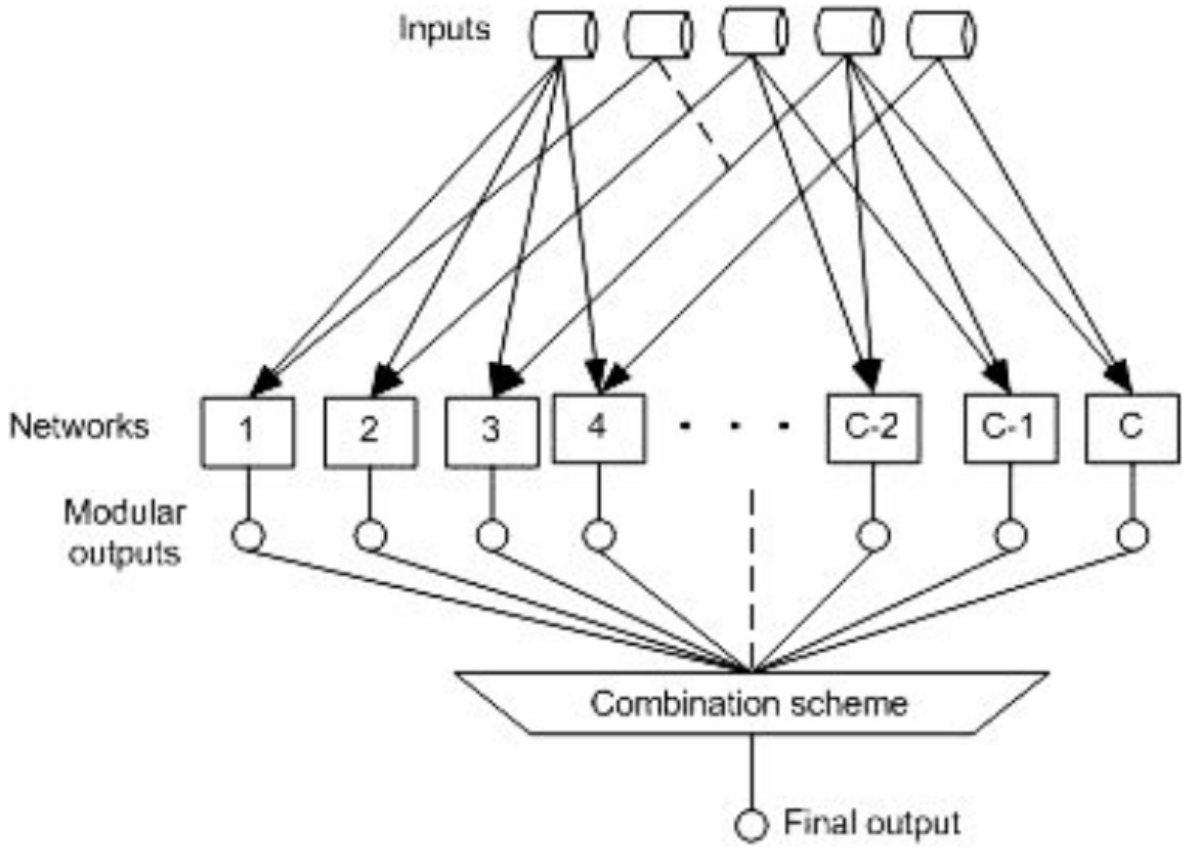


Рисунок 3.1 – Схема запропонованого підходу на основі ансамблю нейронних мереж для відсутніх даних

Для задачі класифікації вага присвоюється з використанням схеми зваженої більшості, наведеної у [44]:

$$\alpha_i = \frac{1 - E_i}{\sum_{j=1}^{N}(1 - E_i)}$$

де E_i – оцінка помилки моделі i на перевірочному наборі. Такий вид розподілу ваги бере свій початок у тому, що називається бустингом, і базується на тому факті, що набір мереж, які дають різні результати, можна об'єднати для отримання кращих результатів, ніж кожна окрема мережа в ансамблі [44].

Алгоритм навчання представлено на рисунку 3.2, а параметр $ntwk_i$ представляє i-ту нейронну мережу в ансамблі.

Процедура тестування відрізняється для класифікації та регресії. У класифікації тестування починається з вибору елітного класифікатора. Це класифікатор з найкращим коефіцієнтом класифікації в наборі перевірки. До цього елітного класифікатора поступово додаються ще два класифікатори, забезпечуючи збереження непарної кількості. Голосування зваженою більшістю використовується в кожному випадку, поки продуктивність не покращиться або поки не будуть використані всі класифікатори.

У випадку регресії всі мережі використовуються одночасно, а їх прогноз разом із ваговими коефіцієнтами використовується для обчислення остаточного значення.

Кінцеве прогнозоване значення обчислюється таким чином:

$$f(x) = y \equiv \sum_{i=1}^{i=N} \alpha_i f_i(x)$$

де α – вага, присвоєна на етапі перевірки, коли не було пропущених даних;

N – загальна кількість регресорів.

Параметр α присвоюється таким чином, щоб

$$\sum_{i=1}^{i=N} \alpha_i = 1$$

Враховуючи, що не всі мережі будуть доступні під час тестування, існує необхідність визначити N_{usable} як кількість регресорів, які можна використовувати для отримання значення регресії екземпляра j.

У результаті

$$\sum_{i=1}^{i=N_{usable}} \alpha_i \neq 1$$

Ваги перераховуються таким чином щоб

$$\sum_{n=1}^{N_{usable}} \alpha_n = 1$$

input : all variable ∈ InputSpace & n_{avail} obtained from the user

output: Decision

Calculate number of maximum Networks N using equation (8.1)

forall (*variables* $1 \rightarrow X_n$) **do**
| Create all possible networks, $ntwk_1 \rightarrow ntwk_C$, each with n_{avail} inputs

end

while *Training* **do**
| ← Train $ntwk_i$ with a different combination of n_{avl} inputs
| **forall** $i \rightarrow C$ **do**
| | ← Subject $ntwk_i$ to a validation set as follows:
| | ⟶ Select the corresponding features used;
| | ⟶ Obtain network performance;
| | ⟶ Assign weights, α according to equation (6.2) and store for future use
| **end**

end

while *Testing* **do**
| ← Load parameters from trainning;
| **if** *A Classification problem* **then**
| | **foreach** *instance with missing values* **do**
| | | ← Select networks, starting with those with bigger α;
| | | ← Bring 2 more networks, using their αas the selection criteria;
| | | ← Use majority voting to obtain the final classification
| | **end**
| **end**
| **if** *A Regression Problem* **then**
| | **foreach** *instance with missing values* **do**
| | | ← Get regression estimates from all networks trained without the current missing variable
| | | ← Use their weights to compute the final value.
| | **end**
| **end**

end

Рисунок 3.2 – Алгоритм прийняття рішень без прогнозування відсутніх даних

3.2 Експериментальні дослідження методу відновлення відсутніх даних на основі нейронної мережі

У даному параграфі представлені результати, отримані під час експериментів, проведених із застосуванням описаного вище алгоритму. Спочатку будуть представлені результати запропонованого методу для задачі класифікації, а пізніше метод буде перевірено для задачі регресії. В обох випадках результати порівнюються з результатами, отриманими після відновлення відсутніх значень за допомогою комбінації нейронної мережі та генетичного алгоритму.

3.2.1 Результати запропонованого методу для задачі класифікації

3.2.1.1 Набір даних

Експеримент проводився з використанням даних аналізу розчинених газів, одержаних від трансформатора. Дані складаються з 10 ознак, які є газами, розчиненими в маслі. Гіпотеза в цьому експерименті полягає в тому, щоб дослідити, чи можна визначити стан прохідного ізолятора (несправний або справний) за відсутності деяких даних. Дані були поділені на навчальну вибірку та тестову вибірку, кожна з яких містила 2000 екземплярів.

3.2.1.2 Постановка експериментів

Класифікаційний тест було реалізовано з використанням ансамблю мереж Fuzzy-ARTMAP. Два входи вважалися відсутніми з більшою ймовірністю, і в результаті 8 вважалися найімовірніше доступними. Було змодельовано оперативний процес, при якому дані відбираються по одному примірнику за раз для тестування. Параметри мережі були визначені емпірично, а параметр пильності 0,75 використовувався для Fuzzy-ARTMAP. Отримані результати порівнювали з результатами, отриманими з використанням підходу NN-GA, де для GA використовувалася швидкість

кросовера 0,1 для 25 поколінь, кожне з розміром популяції 20. Ці параметри визначені емпірично.

3.2.1.3 Результати експериментальних досліджень

За допомогою рівняння (3.1) було знайдено, що максимально можливими є 45 мереж. Продуктивність була розрахована лише після оцінки 4000 випадків та показана на рисунку 3.3. Класифікація збільшується зі збільшенням кількості використовуваних класифікаторів. Хоча всі ці класифікатори не були навчені всім вхідним даним, їхня комбінація працює краще, ніж одна мережа. Точність класифікації, отримана за відсутності даних, досягає 98,2%, що дуже близько до 100.

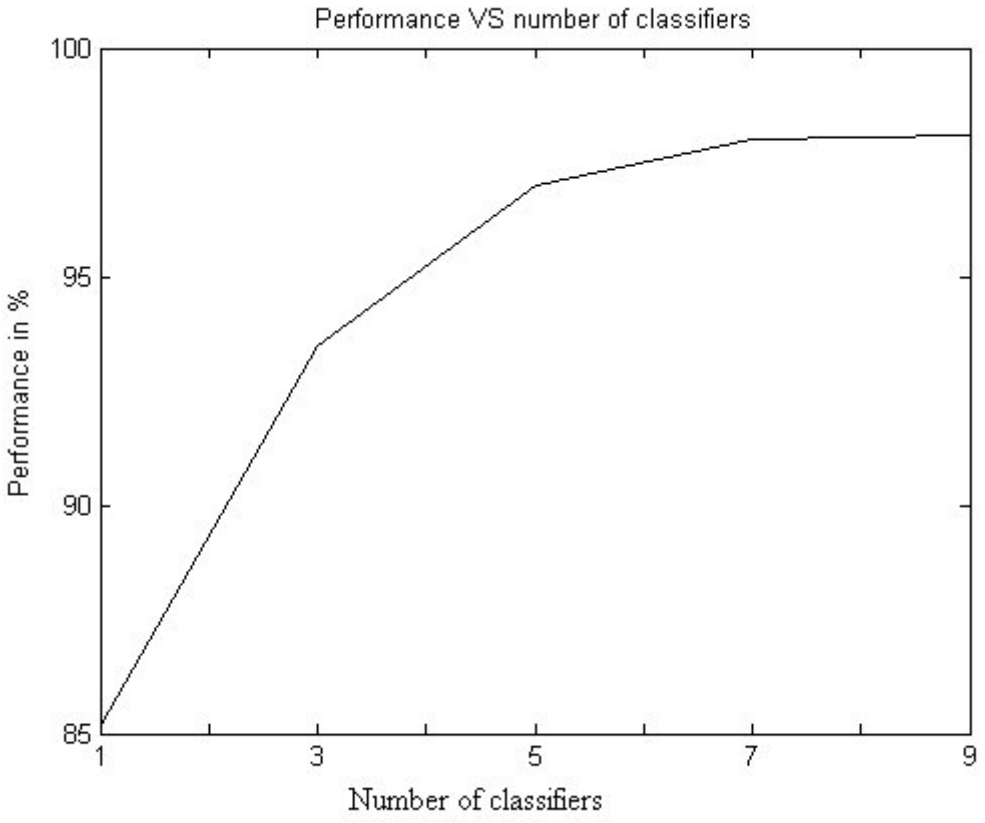


Рисунок 3.3 – Продуктивність відповідно до кількості класифікаторів

Результати, отримані за допомогою цього методу, порівнюються з результатами, отриманими з використанням підходу NN-GA. Результати порівняння представлені у таблиці 3.1.

Таблиця 3.1 – Порівняння запропонованого методу та підходу NN-GA для задачі класифікації

	Пропонований підхід		NN-GA	
Кількість пропущених значень	1	2	1	2
Точність, %	98,2	97,2	99	89,1
Час роботи, с	0,86	0,77	0,67	1,33

Результати, представлені в таблиці 3.1, ясно показують, що запропоновані алгоритми можна використовувати як засіб вирішення задач відсутніх даних. Пропонований алгоритм дуже добре порівняти з добре відомим підходом NN-GA. Час виконання для перевірки продуктивності методу значно варіюється. З таблиці видно, що з методу NN-GA час виконання збільшується зі збільшенням кількості відсутніх змінних. На відміну від підходу NN-GA, пропонований метод пропонує час виконання, який зменшується зі збільшенням кількості вхідних даних. Причиною цього є те, що кількість доступних мереж Fuzzy-ARTMAP зменшується зі збільшенням кількості вхідних даних, як згадувалося раніше. Однак це покращення швидкості досягається за рахунок різноманітності. Крім того, цей метод повністю зазнає невдачі у випадку, коли одночасно буде відсутнє більше, ніж n_{avl} входів.

3.2.2 Результати запропонованого методу для задачі регресії

У цьому параграфі алгоритм, реалізований вище, використовується для задачі регресії. Пропонований алгоритм реалізований з використанням MLP в основному для того, щоб показати, що він може працювати незалежно від типу нейронної мережі, що використовується.

3.2.2.1 База даних

Для цієї задачі використовувалися дані моделі парогенератора на електростанції Abbott [16]. Ці дані мають чотири входи: паливо, повітря, контрольний рівень та збурення. Є два виходи, які мають бути спрогнозовані з використанням пропонованого підходу за наявності даних, що відсутні. Цими вихідними даними є тиск у барабані та витрата пари.

3.2.2.2 Постановка експериментів

MLP тут регресує, щоб отримати два виходи: тиск у барабані та витрата пари. Припустимо, що $n_{avl} = 2$, і в результаті можна використовувати лише два входи. Було створено ансамбль мереж MLP, кожна з яких має п'ять прихованих вузлів та навчена з використанням лише двох вхідних даних для отримання вихідних даних. Через обмежені характеристики в наборі даних, ця робота повинна враховувати не більше однієї відмови датчика в кожному випадку. Кожна мережа була навчена 1200 навчальним циклам з використанням алгоритму масштабованого спряжного градієнту та функції активації гіперболічного тангенсу. Усі ці параметри тренування знову було визначено емпірично.

Оскільки тестування виконується в режимі он-лайн, коли за раз надходить один вхідний сигнал, оцінка продуктивності в кожному екземплярі не дасть загального уявлення про те, як працює алгоритм. Таким чином, робота оцінює загальну ефективність за допомогою наступної формули лише після того, як було передбачено N випадків.

$$Error = \frac{n_\tau}{N} \times 100\%$$

де n_τ – кількість прогнозів у певному допуску. Тут використовується допуск у 20%, який було обрано довільно. Результати зведені в таблицю 3.2.

Таблиця 3.2 – Точність регресії, отримана без оцінки відсутніх значень порівняно з підходом на основі NN-GA

	Пропонований підхід		NN-GA	
Кількість пропущених значень	Точність, %	Час роботи, с	Точність, %	Час роботи, с
Тиск в барабані	86	0,71	68	126
Потік пари	0,86	0,77	84	98

Результати показують, що запропонований метод добре підходить для досліджуваної проблеми.

Запропонований метод працює краще, ніж комбінація генетичних алгоритмів та автокодерних нейронних мереж в досліджуваній задачі регресії. Причина полягає в тому, що помилки, які виникають під час оцінювання відсутніх даних у підході NN-GA, далі поширюються на етап прогнозування вихідних даних.

Підхід, заснований на ансамблі, запропонований тут, не страждає від цієї проблеми, оскільки немає спроби наближення відсутніх змінних. Також можна помітити, що підхід на основі ансамблю займає менше часу, ніж метод NN-GA. Причиною цього є те, що GA може зайняти більше часу, щоб зблизитися до надійних оцінок відсутніх значень залежно від цільової функції, яку потрібно оптимізувати. Хоча час прогнозування незначно менший, метод, заснований на ансамблі, потребує більше часу для навчання, оскільки навчання включає багато мереж.

Висновки до розділу 3

1. Запропоновано метод роботи з відсутніми даними для оперативного моніторингу стану складних об’єктів. По-перше, було вирішено задачу класифікації за наявності пропущених даних, коли робилися спроби

відновити пропущені значення. Потім проблемну область було розширено до задачі регресії.

2. Проведено порівняння запропонованого методу з відомими, який показав, що запропонований метод працює краще, ніж підхід NN-GA, як щодо точності, так і ефективності часу під час тестування. Основний внесок у полягав у тому, щоб запропонувати нову архітектуру, а й у тому, щоб показати, що ця архітектура також надійна. Пропонована архітектура здатна працювати з будь-яким типом нейронної мережі. Це було показано при використанні Fuzzy ARTMAP, так і MLP в запропонованій конфігурації. Перевага запропонованого методу у тому, що він усуває необхідність пошуку найкращої оцінки даних, отже, економить час.

ВИСНОВКИ

1. Проведено огляд основних визначень великих даних та зроблено деякі висновки. Великі дані – серія підходів, інструментів і методів обробки структурованих і неструктурованих даних величезних обсягів, і значного різноманіття для отримання результатів, що сприймаються людиною, ефективних в умовах безперервного приросту, розподілу по численних вузлах обчислювальної мережі. Основними характерними ознаками є великий обсяг, вкрай висока швидкість накопичення, різноманіття даних, достовірність та висока цінність інформації.

2. Для обробки та аналізу великих даних часто використовуються наступні технології: MapReduce, Hadoop, мова R та інше. Крім того, Аналізують Big Data за допомогою Machine Learning, Data Mining та інші. Для того, щоб ефективно обробляти великі обсяги даних при прийнятних часових затратах, необхідні особливі технології. Такими технологіями можна вважати нейронні мережі, які мають велику ефективність нелінійного перетворення і представлення даних в порівнянні з традиційними методами.

3. 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.

4. Архітектура Fuzzy ARTMAP здатна до швидкого онлайнового контрольованого поетапного навчання, класифікації та прогнозування. Нечіткий ARTMAP працює шляхом поділу вхідного простору на кілька гіпербоксів, які відображаються у вихідний простір. Використовується навчання на основі екземплярів, де кожен окремий вхід зіставляється з міткою класу.

5. Глибокі нейронні мережі (deep neural networks) в загальному випадку є персептроном з більш ніж двома прихованими шарами. Завдяки багатошаровій архітектурі вони дозволяють обробляти та аналізувати великі обсяги даних, а також моделювати різноманітні когнітивні процеси.

6. Алгоритм зворотного поширення помилки є ефективним засобом навчання багатошарових нейронних мереж. Для мінімізації сумарної квадратичної помилки мережі використовується метод градієнтного спуску в просторі вагових коефіцієнтів і порогів нейронної мережі. Існує послідовне (online learning) і групове (batch learning) навчання.

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

8. Проведено порівняння запропонованого методу з відомими, який показав, що запропонований метод працює краще, ніж підхід NN-GA, як щодо точності, так і ефективності часу під час тестування. Основний внесок у полягав у тому, щоб запропонувати нову архітектуру, а й у тому, щоб показати, що ця архітектура також надійна. Пропонована архітектура здатна працювати з будь-яким типом нейронної мережі. Це було показано при використанні Fuzzy ARTMAP, так і MLP в запропонованій конфігурації. Перевага запропонованого методу у тому, що він усуває необхідність пошуку найкращої оцінки даних, отже, економить час.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Abbasi A., Sarker S., Chiang R.H. Big data research in information systems: toward an inclusive research agenda. *Journal of the Association for Information Systems*. 2016. Vol. 17. Issue 2. Pp. 1–32.

2. Abualigah, L.M.; Khader, A.T.; Al-Betar, M.A. Unsupervised feature selection technique based on genetic algorithm for improving the Text Clustering. In Proceedings of the 2016 7th International Conference on Computer Science and Information Technology (CSIT), Amman, Jordan, 13–14 July 2016; pp. 1–6.

3. Akter S., Wamba S.F. Big data analytics in E-commerce: a systematic review and agenda for future research. *Electronic Markets.* 2016. Vol. 26. Pp. 173–194.

4. Alhroob, A., Alzyadat, W., Almukahel, I., Altarawneh, H. Missing Data Prediction Using Correlation Genetic Algorithm and SVM Approach. Population, 2020, 11(2). https://doi.org/10.14569/IJACSA.2020.0110288 [Access 18.07.2021].

5. Andonie, R., Sasu, L. and Beiu, V. Fuzzy ARTMAP with relevence factor, *IEEE International Joint Conference on Neural Networks*, Portland, USA, 2003, pp. 1975–1980.

6. Aydilek, I. B., Arslan, A. A Hybrid Method for Imputation of Missing Values Using Optimized Fuzzy C-Means with Support Vector Regression and a Genetic Algorithm. Information Sciences, 2013, 233, 25-35.

7. Bekmamedova N., Shanks G. Social media analytics and business value: A theoretical framework and case study. *47th Hawaii International Conference on System Sciences* : Proceedings (Waikoloa, HI, USA, 2014). Waikoloa, 2014. Pp. 3728-3737.

8. Beyer M.A., Laney D. The importance of big data: a definition. *Gartner, Stamford*. 2012. Pp 2014–2018.

9. Bharadwaj A., El Sawy O.A., Pavlou P.A., Venkatraman N.V. Digital business strategy: toward a next generation of insights. *MISQ*. 2013. Vol. 37(2). Pp. 471–482.

10. Boyd D., Crawford K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*. 2012. Vol. 15(5). Pp. 662–679.

11. Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H. and Rosen, D. B. Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks.* 1992. 3, 698–713.

12. Castro, J., Georgiopoulos, M., Secretan, J., DeMara, R. F., Anagnostopoulos, G. and Gonzaleza, A. Parallelization of fuzzy ARTMAP to improve its convergence speed: The network partitioning approach and the data partitioning approach, *Nonlinear Analysis.* 2005. 63, 877–889.

13. Constantiou I.D., Kallinikos J. New games, new rules: big data and the changing context of strategy. *Journal of Information Technology.* 2015. Vol. 30(1). Pp. 44–57.

14. Cybenko, G. Approximations by Superpositions of a Sigmoidal Function. Math.Contrl., Signals, Syst. 1989. Vol. 2. P. 303–314.

15. Davis C.K. Beyond data and analysis. *Commun ACM*. 2014. Vol. 57(6). Pp. 39–41.

16. De Moor, B. L. R. Database for the identification of systems, department of electrical engineering, ESAT/SISTA, Internet Listing. URL: http://http://www.esat.kuleuven.ac.be/sista/daisy.

17. Freund, Y. and Schapire, R. E. A decision theoretic generalization of on-line learning and an application to boosting, *Proceedings of the Second European Conference on Computational Learning Theory pages*, Springer Verlag, 1995, pp. 23–37.

18. Gantz J., Reinsel D. Extracting value from chaos. *IDC iView*. 2011. Pp. 1–12.

19. Gantz J., Reinsel D. The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east [Електронний ресурс]. – Режим доступу : https://www.emc.com/leadership/digital-universe/2012iview/index.htm.

20. Gebru, I.D.; Alameda-Pineda, X.; Forbes, F.; Horaud, R. EM Algorithms for Weighted-Data Clustering with Application to Audio-Visual Scene Analysis. IEEE Trans. Pattern Anal. Mach. Intell. 2016, 38, 2402–2415.

21. Hinton, G. E. A practical guide to training restricted Boltzmann machines. Toronto : Machine Learning Group, University of Toronto, 2010.

22. Hinton, G. E., Osindero S. , Teh Y. A fast learning algorithm for deep belief nets. Neural Computation. 2006. № 18. P. 1527–1554.

23. Hinton, G., Salakhutdinov R. Reducing the dimensionality of data with neural networks. Science. 2006. –313 (5786). P. 504–507.

24. Hoecker, M.; Polsterer, K.L.; Kugler, S.D.; Heuveline, V. Clustering of Complex Data-Sets Using Fractal Similarity Measures and Uncertainties. In Proceedings of the 2015 IEEE 18th International Conference on Computational Science and Engineering, Porto, Portugal, 21–23 October 2015; pp. 82–91.

25. Hornik, K., Stinchcombe, M., White H. Multilayer feedforward networks are universal approximators. Neural Networks. 1989. № 2. P. 359–366.

26. Javadpour, R., Knapp, G. M. A fuzzy neural network approach to condition monitoring. *Computers and Industrial Engineering*. 2003. 45, 323 –330.

27. Jung, S., Moon, J., Park, S., Rho, S., Baik, S. W., Hwang, E. Bagging Ensemble of Multilayer Perceptrons for Missing Electricity Consumption Data Imputation. Sensors, 2020, 20(6). https://doi.org/10.3390/ s20061772 [Access 18.07.2021].

28. Jurgelevičius, A., Sakalauskas, L. Big Data Mining Using Public Distributed Computing. Information Technology and Control, 2018, 47(2), 236-248.

29. Kamioka T., Tapanainen T. Organizational use of big data and competitive advantage–exploration of antecedents. In: Pacific Asia conference on information systems (PACIS). 2014. p 372.

30. Kingma, D.P.;Welling, M. Auto-Encoding Variational Bayes. arXiv 2013, arXiv:1312.6114 [Access 18.07.2021].

31. Kowarik, A., Templ, M. Imputation with the R Package. VIM, 2016. https://doi.org/10.18637/jss.v074.i07 [Access 18.07.2021].

32. Laney D. 3D data management: Controlling data volume, velocity and variety. *META group research note*. 2001. Vol. 6. p. 1..

33. Li, F.; Qiao, H.; Zhang, B. Discriminatively boosted image clustering with fully convolutional auto-encoders. Pattern Recognit. 2018, 83, 161–173.

34. Li, P.; Chen, Z.; Yang, L.T.; Zhao, L.; Zhang, Q. A privacy-preserving high-order neuro-fuzzy C-means algorithm with cloud computing. Neurocomputing 2017, 256, 82–89.

35. Li, X.; Chen, Z.; Poon, L.K.M.; Zhang, N.L. Learning Latent Superstructures in Variational Autoencoders for Deep Multidimensional Clustering. arXiv 2018, arXiv:1803.05206 [Access 18.07.2021].

36. Lim, S. Y., Mohamad, M. S., Chai, L. E., Deris, S., Chan, W. H., Omatu, S., Ibrahim, Z. Investigation of the Effects of Imputation Methods for Gene Regulatory Networks Modelling Using Dynamic Bayesian Networks. Distributed Computing and Artificial Intelligence, 13th International Conference, 2016, 413-421.

37. Lin, T. H. A Comparison of Multiple Imputation with EM Algorithm and MCMC Method for Quality of Life Missing Data. Quality and Quantity, 2010, 44, 277-287.

38. Lipmann, R. An introduction to computing with neural nets. IEEE Acoustic, Speech and Signal Processing Magazine. 1987. № 2. P. 4–22.

39. Lopes, M. L. M., Minussi, C. R. and Lofuto, A. D. P. Electric load forecasting using a fuzzy ART & ARTMAP neural network, *Applied Soft Computing.* 2005, 5, 235–244.

40. Lynch C. Big data: science in the petabyte era. *Nature*. 2008. Vol. 455. Pp. 1-50.

41. Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh C., Byers, A.H. Big data: The next frontier for innovation, competition, and productivity. 2011.

42. Mashingaidze K., Backhouse J. The relationships between definitions of big data, business intelligence and business analytics: a literature review.

International Journal of Business Information Systems, 2017. Vol.26. No.4, Pp. 488 – 505.

43. Meng, L.; Tan, A.H.; Xu, D. Semi-Supervised Heterogeneous Fusion for Multimedia Data Co-Clustering. IEEE Trans. Knowl. Data Eng. 2014, 26, 2293–2306.

44. Merz, C. J. Using correspondence analysis to combine classifiers, *Machine Learning.* 1997, pp. 1–26.

45. Mikalef P., Pappas I.O., Krogstie J., Giannakos M. Big data analytics capabilities: a systematic literature review and research agenda. *Journal of Information Systems and e-Business Management*. 2018. Vol. 3. Pp. 547–578.

46. Mohammadi, M.; Al-Fuqaha, A.; Sorour, S.; Guizani, M. Deep Learning for IoT Big Data and Streaming Analytics: A Survey. IEEE Commun. Surv. Tutor. 2018, 20, 2923–2960.

47. Myers, T. A. Goodbye, Listwise Deletion: Presenting Hot Deck Imputation as an Easy and Effective Tool for Handling Missing Data. Communication Methods and Measures, 2011, 5, 297-310.

48. Nelwamondo, F. V. and Marwala, T. Fuzzy artmap and neural network approach to online processing of inputs with missing values, *SAIEE Africa Research Journal.* 2007. 98(2), 45–51.

49. Nelwamondo, F. V., Mohamed, S., Marwala, T. Missing Data: A Comparison of Neural Network and Expectation Maximisation Techniques, 2007, arXiv:0704.3474 [Access 18.07.2021].

50. NIST Big Data [Електронний ресурс]. – Режим доступу : https://bigdata.nist.gov/home.php.

51. Opresnik D., Taisch M. The value of big data in servitization. *International Journal of Production Economics*. 2015. Vol. 165. Pp. 174–184.

52. Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., Petersen, I. Missing Data and Multiple Imputation in Clinical Epidemiological Research. Clinical Epidemiology, 2017, 9, 157-166.

53. Rahman, Md. G., Islam, M. Z. Missing Value Imputation Using Decision Trees and Decision Forests by Splitting and Merging Records: Two Novel Techniques. Knowledge-Based Systems, 2013, 53, 51-65.

54. Raković, L., Sakal, M., Matković, P., Marić, M. Shadow IT-Systematic Literature Review. Information Technology and Control, 2020, 49(1), 144-160.

55. Ramachandran, N.; Perumal, V. Delay-aware heterogeneous cluster-based data acquisition in Internet of Things. Comput. Electr. Eng. 2018, 65, 44–58.

56. Ricardo Cardoso Pereira , Joana Cristo Santos , Jos´e Pereira Amorim , Pedro Pereira Rodrigues and Pedro Henriques Abreu. Missing Image Data Imputation using Variational Autoencoders with Weighted Loss // In Proceedings European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2020), 2-4 October 2020. http://www.i6doc.com/en/ [Access 18.07.2021].

57. Rumelhart, D., Hinton G. , Williams R. Learning representation by backpropagation errors. Nature. 1986. № 323. P. 533–536.

58. Saadaoui, F.; Bertrand, P.R.; Boudet, G.; Rouffiac, K.; Dutheil, F.; Chamoux, A. A Dimensionally Reduced Clustering Methodology for Heterogeneous Occupational Medicine Data Mining. IEEE Trans. Nanobiosci. 2015, 14, 707–715.

59. Sachenko A., Kochan V., Turchenko V. Instrumentation for Gathering Data. IEEE Instrumentation and Measurement Magazine, 2003, 6 (3), 34-40.

60. Schroeck M., Shockley R., Smart J., Romero-Morales D., Tufano P. Analytics: The real-world use of big data. *IBM Global Business Services*. 2012. Pp. 1–20.

61. Shakhovska, N., Kaminskyy, R., Zasoba, E., Tsiutsiura, M. Association Rules Mining in Big Data. International Journal of Computing, 2018, 17(1), 25-32.

62. Shakhovska, N., Strubytskyi, R. Model of Data Warehouse with Uncertain Consolidated Data. Applied Mathematics & Information Sciences, 2015, 9(4). http://dx.doi.org/10.12785/amis/090412 [Access 18.07.2021].

63. Siroky, D. S. Navigating Random Forests and Related Advances in Algorithmic Modeling. Statistics Surveys, 2009, 3, 147-163.

64. Sun E.W., Chen Y.T., Yu M.T. Generalized optimal wavelet decomposing algorithm for big financial data. *International Journal of Production Economics*. 2015. Vol. 165. Pp. 194–214.

65. Taylor, G. W. and Wolf, C. Reinforcement learning for parameter control of text detection in images from video sequences, *Proceedings of the 1st IEEE International Conference on Information and Communication Technologies*, 2004, Damascus, Syria, pp. 1–6.

66. White C. Using big data for smarter decision making IBM. Yorktown Heights, New York. 2011.

67. Xie, J.; Girshick, R.; Farhadi, A. Unsupervised Deep Embedding for Clustering Analysis. arXiv 2015, arXiv:1511.06335 [Access 18.07.2021].

68. Yu X, Li H, Zhang Z, Gan C. The Optimally Designed Variational Autoencoder Networks for Clustering and Recovery of Incomplete Multimedia Data. Sensors. 2019; 19(4):809.

69. Zhang, Q.; Yang, L.T.; Chen, Z.; Li, P. High-order possibilistic c-means algorithms based on tensor decompositions for big data in IoT. Inf. Fusion 2018, 39, 72–80.

70. Zhang, S.; Yang, Z.; Xing, X.; Gao, Y.; Xie, D.; Wong, H.S. Generalized Pair-Counting Similarity Measures for Clustering and Cluster Ensembles. IEEE Access. 2017, 5, 16904–16918.

71. Zhou, L.;Wu, D.; Zheng, B.; Guizani, M. Joint physical-application layer security for wireless multimedia delivery. IEEE Commun. Mag. 2014, 52, 66–72.

72. Zhou, Q. Research on heterogeneous data integration model of group enterprise based on cluster computing. Clust. Comput. 2016, 19, 1275–1282.

73. 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.

74. Хлібойко М. Я. 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/

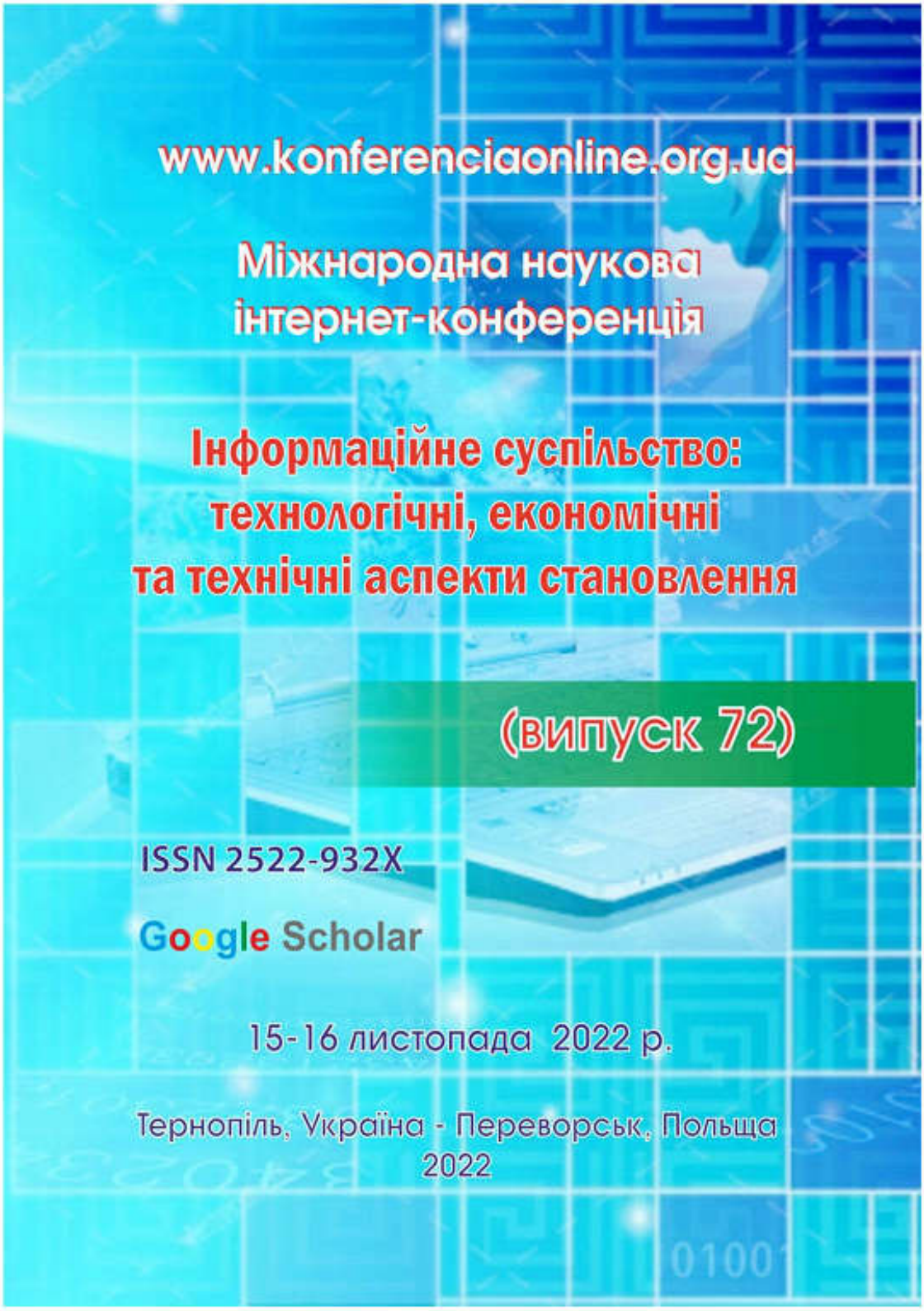
75. Хлібойко М. Я. та ін. Підвищення ефективності побудови систем обробки та аналізу великих даних на основі глибоких нейронних мереж. Збірник матеріалів школи-семінару молодих вчених і студентів «Комп'ютерні інформаційні технології» (CIT'2022), 29 листопада 2022 р. С. 00-00.

76. 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.

77. 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 IT, 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.