МАГІСТЕРСЬКА РОБОТА

МР. ІПм - 55.00.00.000 ПЗ

Група ІПм-24-3

Шаран Ростислав

2025

Івано-Франківський національний технічний університет нафти і газу

Факультет інформаційних технологій

Кафедра інженерії програмного забезпечення

Шаран Ростислав Романович

(прізвище, ім'я, по батькові)

УДК 004.9

(індекс)

МАГІСТЕРСЬКА РОБОТА

Моделі та методи структуризації просторово-часових даних

(назва роботи)

Інженерія програмного забезпечення

(назва освітньої програми)

121 - Інженерія програмного забезпечення

(шифр і назва спеціальності)

Шаран Р.Р.

(підпис, ініціали та прізвище здобувача освітнього ступеня)

Науковий керівник Юрчишин Володимир Миколайович, д.т.н., професор

(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Допущено до захисту

Завідувач кафедри

доц. Бандура В.В.

(посада) (підпис) (дата) (ініціали та прізвище)

Нормоконтроль

доц. Вовк Р.Б.

(посада) (підпис) (дата) (ініціали та прізвище)

Робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

Івано-Франківськ – 2025

Івано-Франківський національний технічний університет нафти і газу

Факультет інформаційних технологій

Кафедра інженерії програмного забезпечення

Освітній рівень магістр

Спеціальність 121 – Інженерія програмного забезпечення

ЗАТВЕРДЖУЮ:
Зав. кафедрою **ІПЗ**
доц. В.В. Бандура
“ 04 ” вересня 2025 р.

ЗАВДАННЯ

НА МАГІСТЕРСЬКУ РОБОТУ СТУДЕНТУ

Шарану Ростиславу Романовичу

(прізвище, ім’я, по-батькові)

1. Тема магістерської роботи “ Моделі та методи структуризації просторово-часових даних ”

керівник проекту (роботи) Юрчишин В.М., д.т.н., професор

затверджені наказом закладу вищої освіти від “ 05 ” листопада 2025 р. № 695/7

2. Строк подання студентом проекту (роботи) 15 грудня 2025 р.

3. Вихідні дані до проекту (роботи) Формальні моделі і методи побудови технологій структуризації просторово-часових даних

4. Зміст розрахунково - пояснювальної записки(перелік питань, які потрібно розробити)

1. Аналіз предметної області використання методів ШІ для обробки просторово-часових даних
2. Теоретичне обґрунтування обробки та аналізу великомасштабних баз даних
3. Концептуальні особливості процесів видобування та структуризації просторово-часових даних
4. Реалізація моделі та підходу обробки та структуризації просторово-часових даних

5. Перелік графічного матеріалу (з точним зазначенням обов’язкових креслень)

1. Концептуальна архітектура системи (рис. 1.1)
2. Концептуальна схема видобутку просторово-часових даних (рис. 1.2)
3. Графічна інтерпретація процесу стиснення даних (рис. 1.3)
4. Процес дискретизації (рис. 1.4)
5. Процес зменшення розмірності (рис. 1.5)

6. Консультанти розділів проекту (роботи)

Розділ	Консультант	Підпис, дата
Перевірка на плагіат	доц., к.т.н. Вовк Р.Б.	

7. Дата видачі завдання 04 вересня 2025 р.

Керівник _____
(підпис)

Завдання прийняв до виконання_____
(підпис)

КАЛЕНДАРНИЙ ПЛАН

№ п/п	Назви етапів магістерської роботи	Строк виконання етапів роботи	Примітка
1	Підбір і вивчення літератури по темі магістерської роботи	15.09.2025	виконано
2	Аналіз предметної області використання методів ШІ для обробки просторово-часових даних	01.10.2025	виконано
3	Теоретичне обґрунтування обробки та аналізу великомасштабних баз даних	17.10.2025	виконано
4	Концептуальні особливості процесів видобування та структуризації просторово-часових даних	02.11.2025	виконано
5	Реалізація моделі та підходу обробки та структуризації просторово-часових даних	19.11.2025	виконано
6	Оцінка продуктивності алгоритмів зменшення даних	02.12.2025	виконано
7	Затвердження пояснювальної записки роботи завідувачем кафедри	15.12.2025	виконано

Студент – магістр _____
(підпис)

Керівник роботи _____
(підпис)

АНОТАЦІЯ

Магістерська робота: 79 с., 12 рис., 43 джерел.

Тема: Моделі та методи структуризації просторово-часових даних

Метою магістерської роботи є розроблення та дослідження моделей і методів структуризації просторово-часових даних, спрямованих на їх ефективне зменшення, оптимізацію обчислювальних процесів та збереження інформаційної цілісності для подальшого аналізу.

Об'єктом дослідження є процеси обробки, структуризації та зменшення великомасштабних просторово-часових даних у сучасних інформаційних системах.

Предметом дослідження є моделі, методи та алгоритми кластеризації й зменшення просторово-часових даних, а також їх вплив на якість і продуктивність аналітичної обробки.

Результати дослідження

У роботі розроблено концептуальну архітектуру системи видобування просторово-часових даних, яка охоплює модулі збору, попередньої обробки, зменшення даних, моделювання залежностей та інтелектуального аналізу.

Висновок

Створено фреймворк обробки просторово-часових масивів, який забезпечує поетапну інтеграцію попередньої обробки, зменшення та структуризації. Доведено, що запропонований підхід дозволяє оптимізувати обчислювальні процеси та забезпечити високу якість представлення даних для подальшої аналітики.

ПРОСТОРОВО-ЧАСОВІ ДАНІ, СТРУКТУРИЗАЦІЯ ДАНИХ, ЗМЕНШЕННЯ ДАНИХ, КЛАСТЕРИЗАЦІЯ, KDD, ВІЗУАЛІЗАЦІЯ ДАНИХ, ОБЧИСЛЮВАЛЬНА ГЕОМЕТРІЯ, АЛГОРИТМИ НА ОСНОВІ ЩІЛЬНОСТІ, СТИСНЕННЯ ЗНАНЬ.

ABSTRACT

Master Thesis: 79 pp., 12 fig., 43 sources.

Topic: Models and methods of spatiotemporal data structuring

The purpose of the master's thesis is to develop and study models and methods of spatiotemporal data structuring aimed at their effective reduction, optimization of computational processes and preservation of information integrity for further analysis.

The object of the study is the processes of processing, structuring and reduction of large-scale spatiotemporal data in modern information systems.

The subject of the study is models, methods and algorithms of clustering and reduction of spatiotemporal data, as well as their impact on the quality and productivity of analytical processing.

Research results

The work developed a conceptual architecture of a spatiotemporal data mining system, which includes modules for collection, preprocessing, data reduction, dependency modeling and intelligent analysis.

Conclusion

A framework for processing spatiotemporal arrays has been created, which provides a step-by-step integration of preprocessing, reduction and structuring. It is proven that the proposed approach allows to optimize computational processes and ensure high quality of data representation for further analytics.

SPOT-TEMPORARY DATA, DATA STRUCTURING, DATA REDUCTION, CLUSTERIZATION, KDD, DATA VISUALIZATION, COMPUTATIONAL GEOMETRY, DENSITY-BASED ALGORITHMS, KNOWLEDGE COMPRESSION.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ 9
ВСТУП 10

РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ВИКОРИСТАННЯ МЕТОДІВ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ОБРОБКИ ПРОСТОРОВО-ЧАСОВИХ ДАНИХ 14
1.1. Огляд та пропонований підхід аналізу і обробки просторово-часових даних 14
1.1.1. Проблема обробки великомасштабних даних 14
1.1.2. Пропонована методологія 15
1.2. Теоретичне обґрунтування обробки та аналізу великомасштабних баз даних 16
1.2.1. Стратегічна значущість даних у сучасних організаційних середовищах 16
1.2.2. Феномен експоненційного зростання даних 17
1.2.3. Критична необхідність аналізу даних 18
1.2.4. Необхідність попередньої обробки даних 19
1.3. Виклики аналізу та обробки просторово-часових даних великомасштабного масиву 20
1.3.1. Експоненційне зростання обсягів просторово-часових даних 20
1.3.2. Обмеження традиційного аналізу та складність просторово-часових даних 21
1.3.3. Концепція зменшення даних для збереження інформаційної цілісності 22
1.4. Задачі дослідження та мета магістерської роботи 23

РОЗДІЛ 2. КОНЦЕПТУАЛЬНІ ОСОБЛИВОСТІ ПРОЦЕСІВ ВИДОБУВАННЯ ТА СТРУКТУРИЗАЦІЇ ПРОСТОРОВО-ЧАСОВИХ ДАНИХ 27
2.1. Видобуток даних як ключовий етап відкриття знань 27

2.1.1. Процес відкриття знань у базах даних (KDD)..... 28
2.1.2. Проблеми та виклики у видобутку даних..... 30
2.2. Концептуальні основи та типологія завдань у просторово-часовому видобутку даних..... 31
2.3. Архітектура системи видобутку просторово-часових даних 34
2.4. Візуалізація та методології зменшення просторово-часових даних 37
2.4.1. Просторово-часова візуалізація 37
2.4.2. Методології зменшення даних..... 38
Висновки до розділу 44

РОЗДІЛ 3. РЕАЛІЗАЦІЯ МОДЕЛІ ТА ПІДХОДУ ОБРОБКИ ТА СТРУКТУРИЗАЦІЇ ПРОСТОРОВО-ЧАСОВИХ ДАНИХ 46
3.1. Модель стиснення знань для просторово-часових даних..... 46
3.2. Кластеризація як стратегія зменшення просторово-часових даних..... 49
3.3. Підхід до розробки фреймворку стиснення просторово-часових даних..... 53
3.4. Попередня обробка підготовка просторово-часових даних 55
3.5. Представлення фази зменшення даних та алгоритми кластеризації..... 61
3.5.1. Просторова кластеризація на основі щільності з використанням найближчих сусідів 61
3.5.2. Просторова кластеризація на основі метрики найближчих сусідів та щільності..... 63
3.5.3. Пошук так зменшення кількості основних представників 65
3.5.4. Оцінка результатів та візуалізація 66
3.6. Оцінка продуктивності алгоритмів зменшення даних..... 68
Висновки до розділу 70

ВИСНОВКИ 72
ПЕРЕЛІК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ..... 75

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

SNN – Shared Nearest Neighbor

SNNDBSC – Spatial Nearest Neighbor Density-Based Spatial Clustering

SNNMDBSC – Spatial Nearest Neighbor Metric-based Density-Based Spatial Clustering

DBSCAN – Density-Based Spatial Clustering of Applications with Noise

KNN – k-Nearest Neighbors

k-d tree – k-dimensional tree

NCAR – National Center for Atmospheric Research

UNDEF (UNDEFINED_CID) – Undefined Cluster ID

CORE – Core Point

NOISE – Noise Point

BORDER – Border Point

ВСТУП

Актуальність теми.

Стрімкий розвиток сучасних інформаційних технологій, зокрема засобів дистанційного моніторингу, супутникових систем спостереження, мобільних пристроїв та інтелектуальних сенсорних мереж, зумовлює формування безпрецедентних за масштабом масивів просторово-часових даних. Такі дані виступають основою для дослідження, моделювання та прогнозування складних динамічних процесів, що відбуваються у природних, соціальних та техногенних системах. Значне збільшення обсягів, різнорідність структури та складна просторово-часова взаємозалежність цих даних створюють нові виклики для аналітичних систем, які покликані забезпечити ефективне управління потоками інформації та отримання достовірних знань.

У сучасних умовах просторово-часові дані стають стратегічним ресурсом, що визначає якість прийняття рішень у таких галузях, як транспортне моделювання, моніторинг навколишнього середовища, урбаністика, логістика, геоінформаційні технології та інтелектуальні системи керування. Однак традиційні методи статистичного аналізу та класичні алгоритми обробки виявляються недостатньо ефективними через обмежену масштабованість, низьку стійкість до шумів і складність роботи з багатовимірними нелінійними структурами просторово-часових масивів. Саме тому актуальною постає потреба розроблення нових моделей і методів, здатних забезпечити якісну структуризацію, зменшення та інтелектуальний аналіз таких даних.

Представлена магістерська робота спрямована на формування комплексного підходу до обробки просторово-часових даних шляхом інтеграції методів штучного інтелекту, алгоритмів кластеризації та концепцій зменшення даних. Результати дослідження дозволяють не лише оптимізувати обчислювальні процеси, а й підвищити точність, інтерпретованість і

прикладну цінність отриманих знань. Робота формує теоретичні засади і пропонує практичні рішення, здатні бути інтегрованими у сучасні інформаційні системи та програмні комплекси.

Актуальність дослідження зумовлена необхідністю ефективної обробки просторово-часових даних, обсяги яких зростають експоненційно під впливом цифровізації та розширення мереж сенсорних і супутникових систем. У багатьох галузях дані такого типу становлять основу для аналізу, прогнозування та прийняття рішень, тому їх коректна обробка без втрати інформаційної цінності є критично важливою. Складність просторово-часових даних визначається не лише великим обсягом, а й високою динамічністю, багатовимірністю, нерегулярністю та залежністю між елементами у просторі та часі.

Відсутність достатньо ефективних моделей і методів структуризації негативно впливає на точність аналітичних висновків та продуктивність інформаційних систем. Наявні підходи до зберігання та аналізу часто не справляються з навантаженням або вимагають надмірних обчислювальних ресурсів, що унеможливлює їх використання в реальних умовах. Тому розробка методів зменшення та структурування даних, які зберігають ключові закономірності й забезпечують високу якість обробки, є науково та практично значущою задачею.

Таким чином, актуальність роботи визначається потребою у створенні нових моделей, здатних забезпечити масштабовану, адаптивну та інформаційно ефективну обробку просторово-часових даних, що відповідає сучасним вимогам наукових, промислових і управлінських систем.

Метою магістерської роботи є розроблення та дослідження моделей і методів структуризації просторово-часових даних, спрямованих на їх ефективне зменшення, оптимізацію обчислювальних процесів та збереження інформаційної цілісності для подальшого аналізу.

Об’єктом дослідження є процеси обробки, структуризації та зменшення великомасштабних просторово-часових даних у сучасних інформаційних системах.

Предметом дослідження є моделі, методи та алгоритми кластеризації й зменшення просторово-часових даних, а також їх вплив на якість і продуктивність аналітичної обробки.

Завдання дослідження

1. Проаналізувати особливості просторово-часових даних та визначити ключові виклики їх обробки.
2. Дослідити сучасні концепції структуризації та видобування знань зі складних даних.
3. Розробити концептуальну архітектуру системи для обробки та зменшення просторово-часових даних.
4. Створити модель стиснення знань для оптимізації структури просторово-часових масивів.
5. Реалізувати модифіковані алгоритми кластеризації, адаптовані до просторово-часової специфіки.
6. Розробити фреймворк зменшення даних і провести його експериментальну перевірку.

Методи дослідження

У роботі використано методи інтелектуального аналізу даних, алгоритмічні методи кластеризації та зменшення даних, математичне моделювання, методи багатовимірної статистики, просторового аналізу та обчислювальної геометрії. Також застосовано експериментальні методи оцінювання продуктивності алгоритмів, програмного моделювання та візуалізації просторово-часових структур.

Наукова новизна отриманих результатів

У роботі запропоновано інтегровану модель структуризації просторово-часових даних, яка поєднує концепції стиснення знань та адаптивної кластеризації. Розроблено модифіковані алгоритми просторової

кластеризації, що враховують щільність розподілу та метрику найближчих сусідів, що підвищує ефективність зменшення даних без втрати їх ключових характеристик.

Практичне значення отриманих результатів

Результати роботи можуть бути використані в системах моніторингу навколишнього середовища, транспортної аналітики, навігаційних сервісах, геоінформаційних системах, системах міського планування та прогнозування динамічних процесів. Розроблений фреймворк може бути інтегрований у програмні комплекси для оптимізації обробки великих масивів просторово-часових даних, забезпечуючи скорочення обсягу інформації при збереженні її аналітичної придатності.

Структура магістерської роботи. Представлена робота складається зі вступу, трьох розділів та висновків. Загальний обсяг роботи становить 79 сторінок, і містить 12 рисунків та перелік використаних джерел із 43 найменувань.

РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ВИКОРИСТАННЯ МЕТОДІВ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ОБРОБКИ ПРОСТОРОВО-ЧАСОВИХ ДАНИХ

1.1. Огляд та пропонований підхід аналізу і обробки просторово-часових даних

У сучасних умовах відбувається експоненційне зростання обсягів даних, які містять просторові та часові компоненти (просторово-часові дані). Ці дані генеруються з різноманітних джерел, включаючи, але не обмежуючись, метеорологічними сенсорами, системами глобального позиціонування (GPS), дистанційним зондуванням Землі (зокрема, супутниковими знімками) та сенсорними мережами Інтернету речей (IoT).

Ефективний аналіз цих просторово-часових наборів даних є фундаментально складним завданням, що обумовлено їх великим обсягом, високою розмірністю, неоднорідністю та складною кореляцією між просторовими та часовими атрибутами. Успішне вирішення цієї проблеми набуває критичного економічного та соціального значення у таких сферах, як прогнозування погоди, моніторинг навколишнього середовища, міське планування та управління транспортними потоками.

Саме тому виникла та активно розвивається галузь досліджень з видобутку просторово-часових даних (Spatio-Temporal Data Mining). У рамках цієї дисципліни розробляються та застосовуються інноваційні обчислювальні, статистичні та машинні техніки для виявлення прихованих закономірностей, тенденцій, аномалій та знань у цих надзвичайно великих просторово-часових базах даних.

1.1.1. Проблема обробки великомасштабних даних

Обсяг цих баз даних, а також швидкість їх безперервного генерації (потік даних) виступають основними обмежувальними факторами

(bottlenecks) для забезпечення своєчасного (реально-часового або близького до реально-часового) аналізу даних. Великі обсяги даних значно збільшують обчислювальну складність та час виконання алгоритмів видобутку знань.

Таким чином, існує нагальна потреба у розробці високоефективних методів попередньої обробки даних (Data Preprocessing), зокрема технік зменшення даних (Data Reduction), які дозволять підготувати просторово-часові дані до подальшого аналізу шляхом зниження їх розмірності або обсягу без суттєвої втрати інформаційної цінності та цілісності.

У рамках цього дослідження представлено та валідовано комплексну методологічну структуру зменшення даних, спеціально розроблену для обробки надзвичайно великих просторово-часових наборів даних (Very Large-Scale Spatio-Temporal Datasets).

1.1.2. Пропонована методологія

Пропонована структура включає модель стиснення даних (Data Compression Model), яка базується на застосуванні методів кластеризації за щільністю (Density-Based Clustering Techniques). Вибір методів, заснованих на щільності (наприклад, DBSCAN або його модифікації), обґрунтовується їхньою здатністю ефективно ідентифікувати кластери довільної форми та виявляти шуми/аномалії (які можуть бути важливими), що є критичним для просторово-часових даних, де розподіл об'єктів часто є нерівномірним.

В роботі буде надано детальне теоретичне обґрунтування кожної застосованої техніки (зокрема, механізмів кластеризації за щільністю та їхніх просторово-часових адаптацій).

Проведено аналітичне порівняння пропонованої моделі з існуючими state-of-the-art методами зменшення просторово-часових даних, акцентуючи увагу на часовій та просторовій складності та коефіцієнті стиснення. Здійснено комплексну оцінку ефективності та робастності розробленої моделі. Оцінка буде виконана на просторово-часових наборах даних великого обсягу (даних про рух транспорту, метеорологічних або

геофізичних даних). Ключові метрики оцінки включатимуть ступінь зменшення обсягу даних, збереження інформаційної ентропії та вплив стиснення на точність подальших аналітичних завдань (класифікації або прогнозування).

1.2. Теоретичне обґрунтування обробки та аналізу великомасштабних баз даних

1.2.1. Стратегічна значущість даних у сучасних організаційних середовищах

Протягом останнього десятиліття спостерігається фундаментальна зміна в інформаційній парадигмі функціонування різноманітних установ, включаючи державні інституції, наукові центри, комерційні підприємства та інші бізнес-структури.

Комп'ютерні системи перестали використовуватися виключно для обчислювальних операцій, перетворившись на ключові сховища для великомасштабних баз даних, які регулярно генеруються або акумулюються з зовнішніх джерел.

У контексті конкурентного середовища ці дані набувають статусу стратегічного ресурсу. Вони містять емпіричну інформацію, необхідну для підтримки прийняття рішень, а також для екстракції нових знань із результатів складних та трудомістких експериментів. Дані фіксуються та зберігаються, виходячи з апріорного припущення про їхню потенційну інформаційну цінність.

Це явище є універсальним і охоплює всі сфери людської діяльності: від рутинного збору операційних даних (наприклад, метадані телефонних дзвінків, деталі транзакцій з кредитними картками, урядова статистична звітність) до складних науково-дослідних комплексів (наприклад, астрономічні каталоги, геномні та протеомні послідовності, молекулярні бази даних, електронні медичні записи).

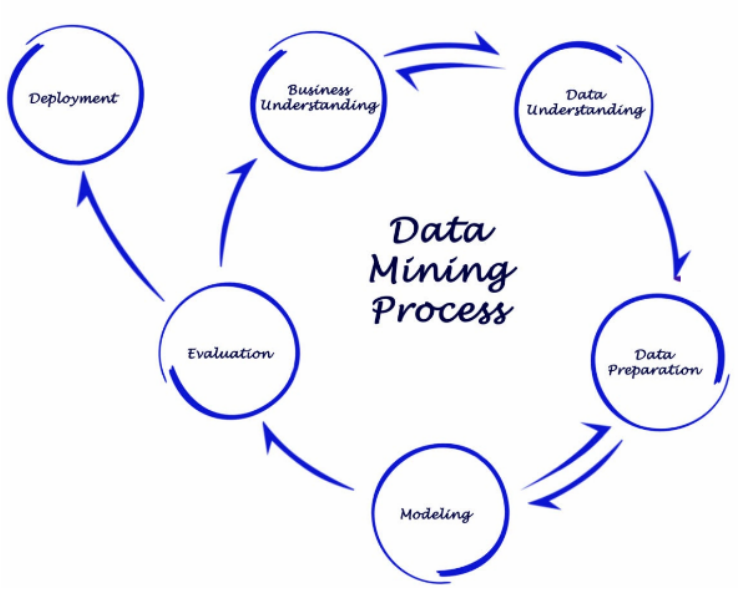

Рис. 1.1. Процеси Data Mining

1.2.2. Феномен експоненційного зростання даних

Технологічні можливості щодо генерації та акумуляції даних зросли на порядки, що призвело до створення баз даних терабайтного та петабайтного масштабу як у комерційному секторі, так і в різних наукових дисциплінах.

Ключові фактори, що сприяють цьому зростанню, включають:

- Тотальна комп'ютеризація, переведення бізнесових, наукових та державних транзакцій у цифрову форму.

- Широке застосування цифрових камер, інструментів публікації та систем штрих-кодів для комерційних продуктів.

- Прогрес у засобах збору даних від високошвидкісного сканування текстів та зображень до складних систем дистанційного зондування на основі супутникових платформ.

- Масове використання всесвітньої павутини (World Wide Web), яка генерує величезні обсяги неструктурованої та напівструктурованої інформації.

Для забезпечення ефективного управління інфраструктурою та валідації наукових гіпотез, аналітики, інженери та вчені повинні мати засоби для ефективного аналізу своїх даних. Це передбачає виявлення ключових кореляцій, закономірностей та причинно-наслідкових зв'язків, що дозволяє

підвищувати точність результатів, оптимізувати підходи до вирішення проблем та отримувати відповіді на дослідницькі запитання.

1.2.3. Критична необхідність аналізу даних

Ми перебуваємо у стані інформаційного вибуху, що обумовлює ургентну потребу у розробці та застосуванні методологій, здатних структурувати та впорядкувати цей феноменальний обсяг даних.

Експоненційне зростання обсягів постійно збережених та потокових (temporal) даних створило гостру необхідність у нових техніках та автоматизованих інструментах. Ці інструменти повинні інтелектуально сприяти трансформації сирих, великомасштабних даних у корисну інформацію та оперативні знання. Оскільки фізично неможливо для людини опрацювати та переглянути такі обсяги, людська увага стала дефіцитним ресурсом.

Отже, головною задачею є розробка методів для автоматизованого аналізу даних, включаючи:

- автоматичну класифікацію
- автоматичне резюмування
- автоматичне виявлення та характеризація тенденцій
- автоматичну ідентифікацію аномалій

У багатьох високотехнологічних галузях, таких як космічні дослідження (зображувальні дані високої роздільної здатності) та телекомунікації (моніторинг великих мережевих операцій), обсяг та швидкість генерації даних часто є обмежувальними чинниками (limiting factors) для своєчасного аналізу. Обсяг даних може перевищувати пропускну здатність доступного апаратного та програмного забезпечення. Традиційні статистичні методи підсумовування та аналізу баз даних виявляються неадекватними для обробки такого масштабу, а також для інтелектуального видобутку прихованої інформації або знань, які мають евристичну цінність

для дослідження предметної області та підтримки процесів прийняття рішень.

Дисципліна, що присвячена екстракції цих знань, відома як Data Mining. Data Mining перетворився на важливу дослідницьку галузь із значним потенціалом практичного застосування.

1.2.4. Необхідність попередньої обробки даних

Сучасні реальні бази даних є іманентно схильними до проблем, пов'язаних із зашумленістю, відсутніми значеннями та внутрішніми суперечностями (inconsistencies). Це обумовлено їх колосальним розміром (зазвичай, мульти-терабайтний або навіть петабайтний обсяг) та гетерогенним походженням із численних джерел. Проблеми з якістю даних є невід'ємною рисою реальних даних; їхня природа та серйозність залежать від низки факторів, що часто знаходяться поза контролем операторів.

Попередня обробка даних (Data Preprocessing) є критично важливим етапом у процесі аналізу. Її виконання спрямоване на:

- Усунення проблем з даними, які можуть перешкоджати якісному проведенню подальшого аналізу.

- Глибше розуміння внутрішньої природи даних.

- Отримання більш значущих знань із заданого набору даних.

Існує низка основних методів попередньої обробки:

1. Очищення даних - спрямоване на заповнення відсутніх значень, згладжування шуму шляхом ідентифікації та обробки викидів, а також корекцію суперечностей у даних.

2. Зменшення даних - отримати компактне представлення даних при мінімізації втрати інформаційного вмісту. Це досягається шляхом агрегування, усунення надлишкових ознак або кластеризації.

3. Інтеграція даних - процес об'єднання даних із декількох гетерогенних джерел для формування узгодженого сховища даних (наприклад, Data Warehouse).

4. Трансформація даних - зміна даних до відповідної форми для застосування аналітичних алгоритмів (наприклад, нормалізація).

Застосування методів попередньої обробки є необхідною умовою для всього процесу аналізу даних, оскільки якість рішень прямо залежить від якості вхідних даних. Своєчасне виявлення та виправлення аномалій у поєднанні зі зменшенням обсягу аналізованих даних призводить до суттєвих переваг у підтримці рішень, підвищуючи точність та обчислювальну ефективність наступних аналітичних процесів.

1.3. Виклики аналізу та обробки просторово-часових даних великомасштабного масиву

1.3.1. Експоненційне зростання обсягів просторово-часових даних

Сучасна здатність до акумуляції та зберігання даних значно перевершила спроможність до їхньої оперативної обробки, аналізу та ефективного використання. Ця невідповідність є особливо вираженою у сфері просторово-часових даних.

Багато організацій регулярно фіксують колосальні обсяги історичних (часових) даних, що деталізують їхню оперативну діяльність, взаємодію з клієнтами та характеристики продуктів. Паралельно, науковці та інженери у численних доменах збирають все більш складні експериментальні набори даних. Це включає терабайти даних, що надходять щоденно від космічних приладів, систем дистанційного зондування Землі (ДЗЗ) з високою просторовою, часовою та спектральною роздільною здатністю, а також від пристроїв моніторингу навколишнього середовища.

Значна частина природних явищ за своєю суттю володіє внутрішніми просторовими та часовими характеристиками. Крім традиційних застосувань, аналіз просторово-часових даних набуває критичної актуальності для вирішення сучасних глобальних проблем, таких як моделювання зміни клімату, прогнозування пандемічних загроз та моніторинг динаміки

переміщення об'єктів (наприклад, транспортних або безпекових). Завдяки інноваціям в апаратному забезпеченні, тепер можливо збирати та зберігати високороздільні просторово-часові набори даних для детального вивчення важливих змін у часі та просторових закономірностей конкретних подій.

Як приклад, охоплення та обсяг цифрових географічних наборів даних є значним і невпинно зростає. Ще у 2002 році було оцінено, що приблизно 80% даних, які зберігаються у корпоративних базах, інтегрують просторову інформацію. Це призводить до необхідності аналізу та обробки величезних масивів георефернованих даних. Ці набори даних є критично важливими для підтримки прийняття рішень, але їхня цінність прямо залежить від здатності екстрагувати корисну інформацію для дослідження та розуміння базових явищ, що генерують ці дані. Таким чином, пріоритетом досліджень стала розробка ефективних та дієвих методологій для аналізу та візуалізації просторово-часових наборів даних.

1.3.2. Обмеження традиційного аналізу та складність просторово-часових даних

Просторово-часові набори даних характеризуються великим розміром, високою багатовимірністю (multidimensionality) та гетерогенністю. Ці атрибути роблять традиційні методи аналізу даних неадекватними для цього класу інформації, оскільки такі методи були розроблені для малих, науково відібраних та однорідних наборів даних.

Ключовим обмеженням є припущення традиційного аналізу про те, що дані генеруються незалежно та однаково розподілені (Independently and Identically Distributed, I.I.D.). Проте, при аналізі просторово-часових даних, це припущення, як правило, не виконується, оскільки такі дані є сильно самокорельованими. У просторовій статистиці ця тенденція відома як автокореляція.

Ця просторова та/або часова автокореляція є надзвичайно поширеною, оскільки об'єкти, близькі один до одного у просторі та/або часі, зазвичай

демонструють подібні характеристики. Як наслідок, ігнорування автокореляції при аналізі даних із просторовими та часовими властивостями може призвести до формулювання некоректних гіпотез або побудови моделей, які є неточними чи неузгодженими з емпіричними даними. З огляду на фундаментальне значення просторово-часових наборів даних для підтримки рішень у багатьох прикладних контекстах, виник значний інтерес до застосування методів видобутку даних для ефективного вирішення цих складних питань.

1.3.3. Концепція зменшення даних для збереження інформаційної цілісності

Аналіз бази даних навіть обсягом у кілька гігабайт є обчислювально складним завданням, що часто вимагає використання витончених алгоритмів та паралельних апаратних архітектур. Великомасштабні набори даних створюють комбінаторно вибухові простори пошуку для алгоритмів аналізу, що може зробити процес екстракції корисної інформації нездійсненним через обмеження часу та пам'яті (простору).

Таким чином, було б бажано замінити великі бази даних невеликою репрезентативною підмножиною даних за умови, що точність оцінок (наприклад, щільності ймовірності, залежностей, меж класів), отриманих із цього зменшеного набору, буде зіставною з точністю, досягнутою при використанні повного набору даних.

Концепція зменшення даних традиційно відома під різними термінами, такими як вибірка, конденсація, фільтрація або проріджування, залежно від конкретної мети завдання.

Методи зменшення даних застосовуються для отримання скороченого представлення даних, яке є значно меншим за обсягом, але при цьому зберігає інформаційну цілісність оригінальних даних. Це означає, що видобуток даних, проведений на зменшеному наборі, повинен бути більш ефективним та давати зіставні аналітичні результати.

Дослідження у цій сфері призвели до формування двох основних підходів:

1. Зменшення кількості екземплярів (Instance Reduction) спрямоване на зменшення кількості записів або точок даних (наприклад, через вибірку або кластеризацію).

2. Зменшення розмірності (Dimensionality Reduction) спрямоване на вибір підмножини ознак або трансформацію наявних ознак (наприклад, через PCA або вибір ознак).

Існує низка обмежень щодо використання існуючих методів зменшення даних, які переважно стосуються балансу між зменшенням розміру та збереженням критичної інформації. Деякі методи ігнорують внутрішню природу даних, зокрема, у випадку просторово-часових даних, вони можуть не враховувати автокореляцію та просторову форму розподілу даних. Інші методи стикаються з труднощами щодо збереження цілісності даних після трансформації вимірів для цілей підсумовування або стиснення.

1.4. Задачі дослідження та мета магістерської роботи

Попередня обробка даних (Data Preprocessing) є критично важливим етапом у процесі аналізу надзвичайно великих реальних наборів даних. Видобуток даних (Data Mining) виступає як потужна технологія, застосування якої може бути використане для ефективного зменшення розміру цих масивів.

Основними цілями цього дослідження є:

- Надання концептуального огляду технологій, які використовуються для аналізу великомасштабних реальних наборів даних, та обґрунтування критичної важливості етапу попередньої обробки в аналітичному процесі.

- Демонстрація життєздатності застосування методів видобутку даних як надійного інструменту для зменшення даних. Зокрема, акцентується на використанні методів кластеризації для експлуатації фундаментальної

властивості просторово-часових даних — просторової та часової автокореляції (принцип, згідно з яким просторово та/або часово близькі об'єкти мають тенденцію до подібності).

- Використання результатів кластеризації для створення зменшеної версії даних. Це досягається шляхом екстракції корисних знань у формі репрезентантів кластерів. Таким чином, замість маніпулювання повним обсягом сирих даних, для візуалізації або аналізу можуть бути використані ці представники, забезпечуючи при цьому збереження критично важливої інформації.

- Розробка комплексної структури для стиснення просторово-часових даних, заснованої на принципах видобутку даних. Ця система повинна бути ефективною та дієвою, спеціально адаптованою для обробки дуже великих реальних наборів даних, особливо тих, що мають складні просторові та часові розширення. Очікується, що така система забезпечить скорочення часу аналізу при одночасному наданні надійних та відтворюваних результатів.

- Використання новостворених зменшених даних як вхідного масиву для системи видобутку просторово-часових даних. Ця система повинна надавати інтерактивне візуальне середовище для полегшення інтерпретації результатів процесу відкриття знань.

Основною метою цієї роботи є вирішення проблеми відкриття знань у дуже великих просторово-часових наборах даних. Зокрема, фокус зосереджено на застосуванні методів видобутку даних на етапі попередньої обробки для зменшення значного обсягу набору даних.

Для досягнення цієї мети, використано алгоритм кластеризації, що застосовує залежну від щільності метрику подібності для групування подібних об'єктів. Це є доцільною технікою для просторово-часових даних, оскільки просторова та часова близькість об'єктів корелює з їхньою внутрішньою подібністю, що дозволяє досягти ефективного стиснення даних.

В пропонованій роботі виконано ґрунтовний аналіз існуючих підходів до зменшення даних як рішення складної проблеми навчання на основі

великомасштабних баз даних. Розглянуті техніки намагаються вирішити обчислювальну складність аналізу шляхом отримання редукованого представлення даних, яке має значно менший обсяг, але при цьому зберігає високий ступінь цілісності оригінальних даних. Запропоновано стратегію зменшення даних для дуже великих просторово-часових наборів даних, засновану на алгоритмах видобутку даних. Ця стратегія базується на комбінації методів кластеризації на основі щільності та графових структур, що призводить до створення нової, зменшеної версії даних, представленої репрезентантами кластерів.

Створено надійну та масштабовану структуру стиснення просторово-часових даних, що базується на розробленій стратегії. Вона інтегрує техніку кластеризації для ефективного використання властивості просторової та часової автокореляції в даних.

Розроблена структура була інтегрована у комплексну систему видобутку та візуалізації просторово-часових даних. Ця система має двошарову архітектуру (видобуток та візуалізація), яка забезпечує підготовку, візуалізацію та інтерпретацію як вхідних даних, так і результатів видобутку.

Висновки до розділу

У першому розділі було здійснено системний аналіз предметної області обробки просторово-часових даних, який засвідчив стрімке зростання їх обсягів і складності в сучасних інформаційних системах. Було встановлено, що експоненційне збільшення даних зумовлене розвитком сенсорних мереж, мобільних пристроїв та супутникових технологій, що робить просторово-часові масиви критично важливими для прийняття рішень. З’ясовано, що традиційні методи аналізу не забезпечують належної ефективності й масштабованості при роботі з великими динамічними масивами даних. Розглянуто теоретичні засади значущості даних у сучасних організаційних

середовищах, які вимагають інтелектуальних підходів до їх обробки. Доведено, що попередня обробка даних є ключовим етапом, який визначає якість подальшого аналізу й зменшує ризики викривлення результатів. Особливу увагу приділено викликам, пов’язаним із багатовимірністю, гетерогенністю та складною структурою просторово-часових масивів. Обґрунтовано необхідність застосування методологій зменшення даних як засобу оптимізації обчислювальних процесів. Визначено, що якісне структурування таких даних базується на інтеграції формальних моделей та методів штучного інтелекту. Також сформульовано основні задачі дослідження, які охоплюють створення концептуальної моделі, розробку алгоритмів та оцінку їх ефективності.

РОЗДІЛ 2. КОНЦЕПТУАЛЬНІ ОСОБЛИВОСТІ ПРОЦЕСІВ ВИДОБУВАННЯ ТА СТРУКТУРИЗАЦІЇ ПРОСТОРОВО-ЧАСОВИХ ДАНИХ

2.1. Видобуток даних як ключовий етап відкриття знань

Досягнення в технологіях генерації та збору даних призвели до створення баз даних величезного розміру як у комерційному секторі, так і в різноманітних наукових дисциплінах. Реєстрація даних здійснюється, виходячи з апріорного припущення про їхню потенційну інформаційну цінність. Це явище є повсюдним і охоплює всі сфери людської діяльності: від збору рутинних операційних даних (наприклад, деталі телефонних дзвінків, транзакції кредитних карток, урядова статистика, взаємодії в соціальних мережах) до складних науково-дослідних масивів (наприклад, астрономічні, метеорологічні, агролісові, геномні та медичні дані).

Потенційно корисна інформація та знання інкапсульовані в цих великомасштабних базах даних. Дисципліна, що займається екстракцією цієї інформації, відома як видобуток даних (Data Mining, DM). Видобуток даних утвердився не лише як важлива галузь досліджень, але і як напрямок із значним прикладним потенціалом.

Галузь видобутку даних виникла у відповідь на обмеження традиційних методів аналізу даних щодо ефективної роботи з викликами, які ставлять нові типи та обсяги наборів даних. Об'єднані метою подолання цих викликів, дослідники з гетерогенних дисциплін сфокусувалися на розробці більш ефективних та масштабованих інструментів, здатних обробляти різноманітні типи даних.

Видобуток даних є мультидисциплінарною областю, яка інтегрує ідеї та методи з:

- Статистики - Вибірка (sampling), оцінювання (estimation) та перевірка гіпотез.

- Штучного інтелекту - алгоритми пошуку, методи моделювання та теорії навчання.

- Розпізнавання образів та машинного навчання.

Видобуток даних — це дослідження та аналіз, які здійснюються автоматичними або напівавтоматичними засобами, з метою виявлення або відкриття корисних, нових та досі невідомих знань із реальних даних.

У роботі [4] відкрите знання описується як цікаві закономірності (interesting patterns), що є невипадковими властивостями та відношеннями, які задовольняють наступним критеріям:

1. Дійсність - закономірність повинна бути достатньо загальною, щоб застосовуватися до нових (небачених) даних; це не повинна бути просто аномалія поточного набору даних.

2. Новизна - закономірність має бути нетрівіальною та несподіваною (тобто невідомою попередньо).

3. Корисність - стосується властивості, за якої закономірність може бути використана для підтримки прийняття рішень або для подальшого наукового дослідження.

4. Зрозумілість - закономірність повинна бути достатньо простою для інтерпретації людиною.

2.1.1. Процес відкриття знань у базах даних (KDD)

Видобуток даних часто розглядається в ширшому контексті процесу відкриття знань у базах даних (Knowledge Discovery in Databases, KDD). KDD — це ітеративний процес, що складається з багатьох етапів, спрямований на пошук корисної інформації та закономірностей у даних. Видобуток даних є одним із етапів KDD, а саме: застосування алгоритмів для екстракції інформації та закономірностей із попередньо оброблених даних.

Процес KDD зазвичай складається з наступних п'яти взаємопов'язаних етапів:

1. Вибірка - збір (отримання) даних із різних джерел (бази даних, файли, неелектронні джерела). Характерною особливістю є можливість походження даних із множини гетерогенних джерел.

2. Попередня обробка. На цьому етапі виконується низка операцій, спрямованих на покращення якості даних. Це включає виправлення або видалення помилок (шумів), а також заповнення або прогнозування відсутніх значень.

3. Трансформація. Дані з різних джерел повинні бути перетворені або трансформовані у більш зручні формати для аналізу. Зменшення даних (Data Reduction) може бути використано для скорочення кількості можливих значень або розмірності, які розглядаються. Метою цього етапу є підготовка даних для ефективного видобутку та отримання більш значущих результатів.

4. Видобуток даних - застосування цільових алгоритмів (вибір яких залежить від конкретної задачі видобутку) до трансформованих даних для генерації бажаних результатів (моделей, правил, кластерів).

5. Інтерпретація / Оцінка - спосіб представлення результатів видобутку даних користувачеві є надзвичайно важливим, оскільки корисність отриманих результатів залежить від їхньої зрозумілості. На цьому етапі застосовуються різноманітні стратегії візуалізації та графічні інтерфейси користувача (GUI).

Завдання видобутку даних традиційно поділяються на три основні групи:

- Прогностичний видобуток даних. Мета полягає у визначенні моделі або набору моделей у даних, які можуть бути використані для прогнозування певної цільової відповіді (тобто значення певного атрибуту). Типові методи включають статистичний аналіз, класифікацію та дерева рішень.

- Дослідний видобуток даних спрямований на виявлення прихованих закономірностей та структур або розпізнавання подібностей та відмінностей між об'єктами даних. Методи включають асоціативні правила (Association

Rules), кластеризацію (Clustering), нейронні мережі та візуальний видобуток даних.

- Редуктивний видобуток даних. Основна мета — зменшення даних. Завдання полягає в агрегуванні або консолідації даних у дуже великих наборах у менші, керовані підмножини.

2.1.2. Проблеми та виклики у видобутку даних

Незважаючи на значний потенціал, видобуток даних є складним процесом, що обумовлено низкою факторів, зокрема гетерогенністю даних та обчислювальною складністю алгоритмів при роботі з великими обсягами.

Основні виклики включають:

1. Масивні набори даних та висока розмірність.

Великі набори даних породжують комбінаторно вибухові простори пошуку, що може зробити процес видобутку закономірностей нездійсненним через обмеження простору (пам'яті) та часу. Тому алгоритми видобутку повинні бути ефективними та масштабованими.

2. Перенавчання та оцінка статистичної значущості.

Великі та розподілені набори даних можуть містити випадкові точки даних, що призводить до перенавчання моделі. У процесі видобутку можуть бути необхідні методи регуляризації та перехресної валідації (resampling).

3. Управління змінними даними та знаннями.

У базах даних, які швидко модифікуються, доповнюються або видаляються, раніше виявлені закономірності можуть втратити свою дійсність. Можливі рішення передбачають використання інкрементальних методів для оновлення закономірностей.

4. Взаємодія з користувачем та попередні знання.

Видобуток даних за своєю суттю є інтерактивним та ітеративним процесом. Користувачі повинні мати можливість взаємодіяти на різних етапах, а предметні знання можуть бути використані як для високорівневої

специфікації моделі, так і на більш детальному рівні. Також бажаною є візуалізація отриманої моделі.

5. Зрозумілість закономірностей.

Існує потреба зробити відкриті закономірності більш інтерпретованими для людини. Це може бути досягнуто шляхом використання правил, представлення природною мовою або візуалізації отриманих знань.

6. Нестандартні та неповні дані.

Дані часто бувають відсутніми та/або зашумленими. Якщо дані не оброблені належним чином, точність виявлених закономірностей буде низькою.

7. Обробка реляційних та складних типів даних.

Виникають складнощі при видобутку знань із складних типів даних, таких як мультимедійні об'єкти, просторові дані, часові ряди тощо.

8. Інтеграція.

Бажаною є безшовна інтеграція інструментів видобутку даних як із системою управління базою даних (СУБД), так і з кінцевою процедурою прийняття рішень.

2.2. Концептуальні основи та типологія завдань у просторово-часовому видобутку даних

Видобуток просторово-часових даних (STDM - Spatio-Temporal Data Mining) становить інтерсекцію кількох ключових наукових областей, включаючи машинне навчання, теорію інформації, статистику, системи баз даних та географічну візуалізацію. STDM охоплює набір дослідних, обчислювальних та інтерактивних підходів, спрямованих на аналіз надзвичайно великих просторових та просторово-часових наборів даних. Ефективна візуалізація та відкриття корисних знань із цих масивів є комплексною науковою проблемою і набуває критичного економічного

значення. Видобуток даних утвердився як ключова технологія для екстракції прихованих знань із цих значних обсягів інформації.

Області видобутку просторових даних та видобутку тимчасових даних досліджувалися незалежно протягом десятиліть у спільнотах Knowledge Discovery in Databases (KDD) та Data Mining. Видобуток просторових даних фокусується на відкритті цікавих відносин та характеристик, які можуть існувати імпліцитно в просторових даних. Видобуток тимчасових даних стосується аналізу подій, впорядкованих за одним або кількома вимірами часу.

Видобуток просторово-часових даних інтегрує ці напрями, включаючи адаптацію, модифікацію або вдосконалення "традиційних" технік, спираючись на солідний досвід KDD для розробки нових прикладних методологій видобутку даних, які враховують взаємозв'язок простору і часу.

У STDM просторовий та часовий виміри додають суттєвої складності до конвенційного процесу видобутку даних. Більшість досліджень у STDM зосереджено на адаптації, модифікації або розширенні "традиційних" технік. У результаті, форми просторово-часових правил (отриманих за допомогою алгоритмів видобутку) є розширеннями їхніх статичних аналогів, але водночас унікально відмінні від них.

Можна виділити п'ять основних класів завдань STDM:

1. Просторово-часові асоціації - це розширення асоціативних правил з інтеграцією просторово-часових предикатів. Ці правила вимагають використання як просторових, так і часових предикатів для визначення зв'язків між об'єктами або подіями.

2. Просторово-часові узагальнення - цей метод вимагає наявності фонових знань у формі концептуальних ієрархій. Концептуальні ієрархії використовуються для агрегування даних, дозволяючи виявляти сильніші правила за рахунок зниження специфічності.

3. Просторово-часові кластери. Кластеризація, визначена як групування подібних об'єктів, на відміну від узагальнення, не вимагає апріорних знань

(концептуальних ієрархій). У STDM вона спрямована на виявлення груп об'єктів, які є подібними у просторі та часі.

4. Правила еволюції. Просторово-часові правила еволюції виводяться з просторово-часових описів еволюції, які безпосередньо репрезентують зміни та рух просторових об'єктів і явищ у часі.

5. Метаправила - це правила, які отримуються шляхом порівняння існуючих наборів правил, згенерованих для одного й того ж предметного домену у різні моменти часу.

Значущість аналізу та видобутку просторово-часових даних зростає пропорційно до збільшення доступності великомасштабних наборів даних, зібраних у різних прикладних сферах, таких як метеорологія, геофізика, моніторинг навколишнього середовища, кримінологія, біологія, агролісівництво, охорона здоров'я, соціальні медіа та транспорт.

У реальних застосуваннях зустрічається кілька типів просторово-часових даних, які різняться за способом використання простору та часу у процесі збору та представлення:

- дані подій - складаються з дискретних подій, що відбуваються у певних просторових точках та часових моментах (наприклад, реєстрація випадків злочинних подій у межах міста).

- дані траєкторій - описують шлях рухомих тіл (наприклад, міграційні моделі тварин або відстеження транспортних засобів).

- дані точкових посилань. У цьому випадку безперервне просторово-часове поле вимірюється у рухомих просторово-часових точках відліку (наприклад, вимірювання температури, зібрані метеозондами, що рухаються).

- растрові дані. Спостереження просторово-часового поля збираються у фіксованих комірках просторово-часової сітки (наприклад, дані наземних сенсорів, що вимірюють якість повітря).

На основі моделей, розроблених для просторових та часових даних у реальних застосуваннях, ці застосування можна класифікувати на чотири категорії:

1. Застосування, де час не є частиною зареєстрованих даних (або його важливість є мінімальною).

2. Застосування, де дані реєструються як впорядковані послідовності подій відповідно до певного відношення (наприклад, "до/після", часова мітка).

3. Застосування, де дані реєструються з регулярними часовими інтервалами (наприклад, часові ряди).

4. Застосування, де вимір часу повністю інтегрований у зареєстровані дані (історичні просторові об'єкти, просторово-часові поля).

Розробка повноцінних просторово-часових систем стає реальністю завдяки розробці нових просторово-часових моделей даних та наявності значної обчислювальної потужності та доступних носіїв зберігання. Ці системи повинні забезпечувати інтеграцію кількох форматів даних, розширення аналітичних можливостей та покращену візуалізацію як запитів, так і вмісту бази даних.

2.3. Архітектура системи видобутку просторово-часових даних

Для вирішення методологічних та обчислювальних проблем, окреслених у попередньому підрозділі, було розроблено систему для видобутку просторово-часових даних.

Ключові функціональні цілі цієї розробки полягали у наступному:

- Просторово-часова локалізація - надати інструментам видобутку даних можливість забезпечувати певну форму локалізації в аналізованих масивах.

- Інтерактивна візуалізація - забезпечити інтерактивну 3D візуалізацію результатів процесу видобутку, що підвищує ефективність інтерпретації та значущість отриманих знань.

Для реалізації цих цілей була спроектована та імплементована система, що включає двигун видобутку даних (Data Mining Engine), здатний

інтегрувати різноманітні алгоритми видобутку (для роботи з конкретними типами наборів даних), а також інструменти 3D геовізуалізації.

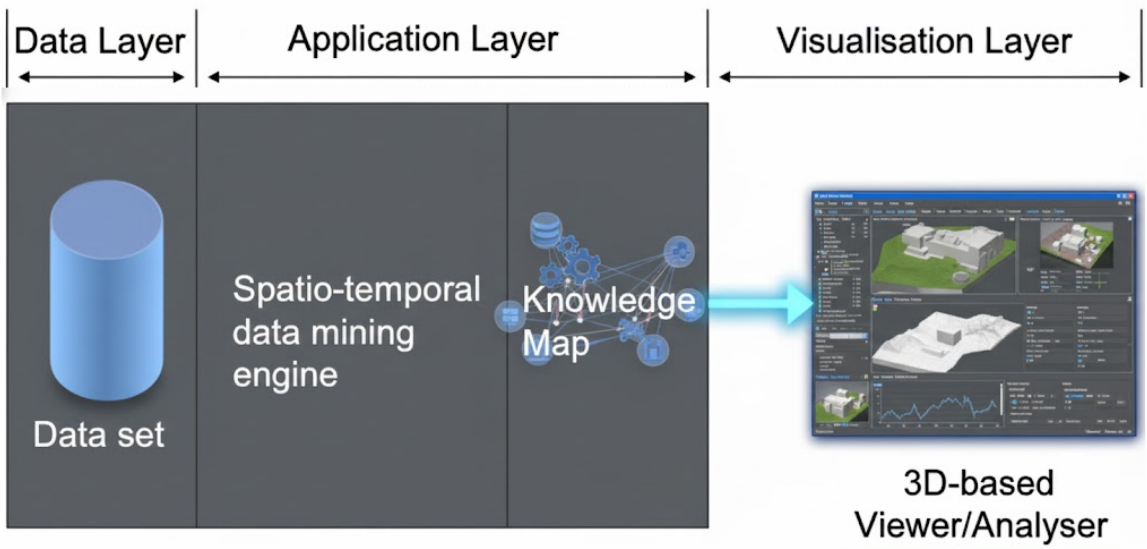

Рис. 2.1. Концептуальна архітектура системи

Концептуальна архітектура системи, ілюстрована на рис. 2.1, має трирівневу структуру:

1. Шар даних (Data Layer).

2. Шар застосування / процес видобутку даних (Application / Data Mining Process Layer).

3. Шар візуалізації (Visualization Layer).

Шар інтерфейсу даних (Data Interface Layer) виконує функцію забезпечення підключення як до внутрішніх, так і до зовнішніх джерел даних. Він відповідає за вирішення питань форматування та типізації даних, а також посередництво у відмінностях контексту джерела даних, таких як одиниці вимірювання. Крім того, шар даних здійснює обмін даними з джерелом для зберігання та доступу і взаємодіє з шаром видобутку даних для підготовки даних та побудови відповідних карт знань.

Процес видобутку для просторово-часових даних є суттєво складнішим, ніж для реляційних даних, як з точки зору обчислювальної ефективності, так і з огляду на складність можливих закономірностей.

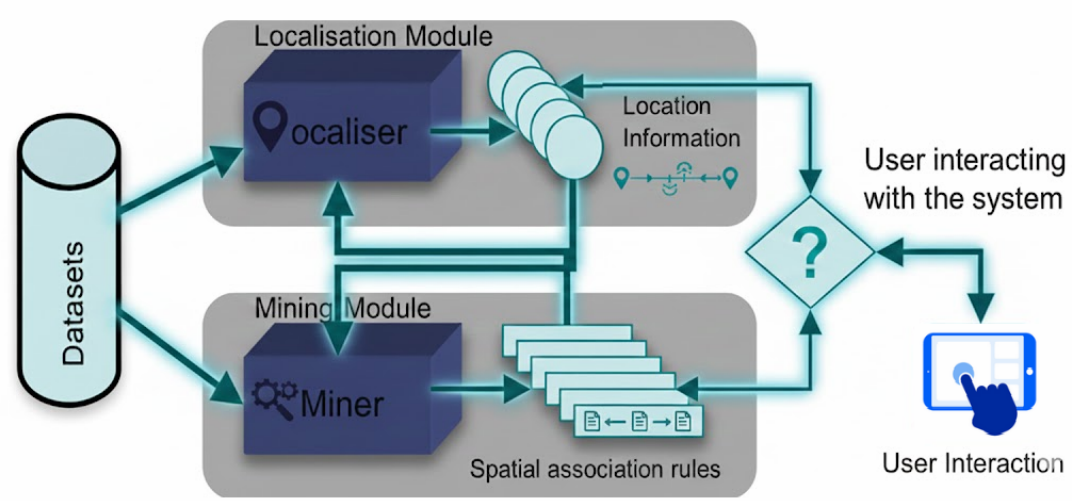

Рис. 2.2. Концептуальна схема видобутку просторово-часових даних

У цьому дослідженні пропонується новий підхід до видобутку просторово-часових даних, концептуальна схема якого представлена на рис. 2.2. Підхід структурований на два основні компоненти:

- Локалізатор (Localizer) - працює переважно з атрибутами даних, приділяючи особливу увагу просторовим та часовим вимірам. Його функція полягає у виявленні локальних залежностей.

- Видобувач (Miner) - обробляє дані, покладаючись на просторово-часові відношення, які були попередньо надані Локалізатором.

Розділення цих двох завдань дозволяє здійснювати обчислення локально, підвищуючи їхню простоту та швидкість виконання. Вихідні дані Видобувача (наприклад, карта знань, кластери, асоціативні правила) також можуть бути передані Локалізатору для ітеративного подальшого вдосконалення.

Для підтримки видобутку просторово-часових даних за допомогою геовізуалізації було застосовано дві альтернативні візуальні платформи:

- на основі Google Earth - використання відомого географічного інтерфейсу для накладання аналітичних результатів.

- на основі реалізації Java 3D.

Метою обох візуальних застосувань є надання розподіленого, колаборативного середовища для експертів з даних та видобутку знань. Це

середовище дозволяє працювати з даними, використовуючи тривимірну візуалізацію у геореферованому просторі, що значно полегшує інтерпретацію складних просторово-часових закономірностей.

2.4. Візуалізація та методології зменшення просторово-часових даних

2.4.1. Просторово-часова візуалізація

Методи візуалізації широко визнані як потужні інструменти для аналізу великомасштабних просторових та просторово-часових наборів даних, оскільки вони ефективно використовують здатність людського сприйняття до обробки та інтерпретації візуальних закономірностей (visual patterns). Основна передумова полягає в тому, що графічні представлення сприяють глибшому розумінню вмісту даних, оскільки людська візуальна система більш схильна до обробки візуальної інформації, ніж текстової.

Візуалізація передбачає застосування графічних/візуальних методів для представлення інформації, даних або знань з метою пояснення, аналізу та передачі концептуальної інформації про складні масиви. Вона вважається необхідною на етапі аналізу даних та відкриття знань (KDD) для отримання інсайтів про дані та явища, які вони репрезентують.

Однак існуючі методи просторової візуалізації виявляються недостатніми для систем підтримки прийняття рішень (DSS) при їх самостійному використанні. Ключові виклики включають:

1. Проблема візуалізації багатовимірних просторових наборів даних.

2. Складність визначення ефективних візуальних інтерфейсів для огляду та маніпулювання всіма геометричними компонентами просторових даних.

З огляду на це, виникає потреба в альтернативних рішеннях, які повинні інтегрувати не лише статичні графічні відображення результатів видобутку, але й надавати динамічну та інтерактивну можливість отримання та взаємодії з різними просторовими та часовими проекціями.

Перспективними напрямками є візуальний видобуток та іммерсивна аналітика. Ці підходи можуть використовувати недорогі високоякісні системи VR/AR для створення мультисенсорних інтерфейсів, що підтримують співпрацю. Наприклад, комбінування 2D та 3D візуалізацій для занурення користувачів у комп'ютерно-генеровані сцени може забезпечити ефективне середовище для підтримки аналізу та прийняття рішень.

Виклики візуального видобутку великомасштабних просторово-часових даних:

- необхідність враховувати специфічні характеристики багатовимірних просторово-часових даних для визначення найбільш ефективного способу візуалізації та передачі корисної та актуальної інформації, що посилить людські когнітивні можливості.

- методи та інструменти візуалізації повинні бути масштабованими щодо: обсягу даних, гетерогенності інформації, якості даних та характеристик різних дисплеїв і середовищ (розмір, роздільна здатність, можливості взаємодії).

- критично важливим є забезпечення ефективної інтеракції з візуальними інтерфейсами для огляду та маніпулювання геопросторовими та часовими атрибутами ST-даних.

- слід застосовувати орієнтований на користувача та завдання підхід (User-Centric and Task-Oriented) для визначення відповідних методів візуалізації та взаємодії. Це забезпечить ефективну підтримку цільових акторів та сприятиме продуктивній співпраці.

2.4.2. Методології зменшення даних

Аналіз колекції даних навіть обсягом у кілька гігабайт є обчислювально складним та тривалим завданням. Великі набори даних створюють комбінаторно вибухові простори пошуку для деяких алгоритмів видобутку даних, що робить процес аналізу нездійсненним через обмеження ресурсів (простору та часу).

Ключовий підхід до вирішення проблеми навчання на основі великих баз даних полягає у виборі меншої підмножини репрезентативних даних для видобутку. Мета полягає у зменшенні оригінальних даних до меншої підмножини представників таким чином, щоб точність оцінок (наприклад, щільності ймовірності, залежностей, меж класів), отриманих із цієї зменшеної підмножини, була зіставною з точністю, отриманою при використанні повного набору даних.

Концепція зменшення даних традиційно має низку назв (редагування, конденсація, фільтрація, проріджування), залежно від конкретної мети. Причина застосування цих методів полягає в отриманні зменшеного представлення даних, яке значно менше за обсягом, але при цьому зберігає цілісність оригінальних даних. Тобто, аналіз зменшених даних має бути більш ефективним, але продукувати ідентичні аналітичні результати.

Дослідження призвели до двох основних підходів:

- Зменшення кількості екземплярів (Instance Reduction) - зменшення кількості записів або точок даних.

- Зменшення розмірності (Dimensionality Reduction) - вибір підмножини ознак або перетворення наявних ознак.

Останній підхід може бути реалізований через вибір ознак (Feature Selection) та виділення ознак (Feature Extraction). Вибір ознак зменшує розмірність шляхом відкидання надлишкових, домінуючих або малоінформативних атрибутів. Методи виділення ознак використовують всю інформацію, щоб отримати новий, трансформований простір меншої розмірності.

Більшість існуючих методів зменшення даних зосереджуються на просторі зберігання або загальній розмірності, часто ігноруючи їхні просторові та часові властивості.

1. Вибірка (Sampling)

Найпростіший підхід полягає у виборі бажаної кількості випадкових зразків із усього набору. Існують випадкові, детерміністичні та засновані на

щільності стратегії вибірки. Хоча наївні методи вибірки прості, вони не підходять для реальних проблем із зашумленими даними, оскільки їхня продуктивність може непередбачувано змінитися. Метод випадкової вибірки ефективно ігнорує інформацію у невибраних зразках, що негативно впливає на точність та збіжність алгоритмів. Розвинений алгоритм зменшення даних повинен інтегрувати інформацію з усіх зразків.

2. Стиснення (Compression)

Стиснення даних — це набір методів для представлення інформації у компактній формі. Його метою є представлення джерела з якомога меншою кількістю бітів, дотримуючись мінімальних вимог для відновлення оригіналу.

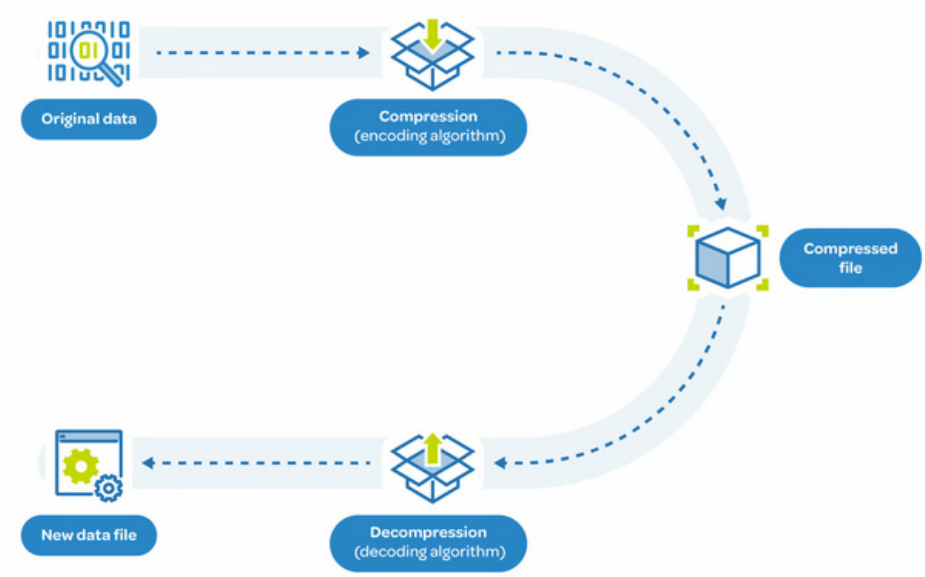

Рис. 2.3. Графічна інтерпретація процесу стиснення даних

Стиснення без втрат повинно точно відновлювати оригінальні дані зі стисненої версії. Стиснення з втратами допускає потенційну втрату важливої інформації.

Незважаючи на те, що стиснення даних можна розглядати як вид зменшення, воно фокусується виключно на просторі зберігання наборів даних, а не на екстракції прихованих у них знань.

3. Дискретизація (Discretization)

Методи дискретизації даних використовуються для зменшення кількості значень для заданого безперервного атрибуту шляхом поділу його діапазону на фіксовану кількість інтервалів. Мітки інтервалів замінюють фактичні значення, що спрощує оригінальні дані. Однак, неправильний вибір інтервалів може призвести до відкидання значного відсотка даних та втрати потенційно важливих знань.

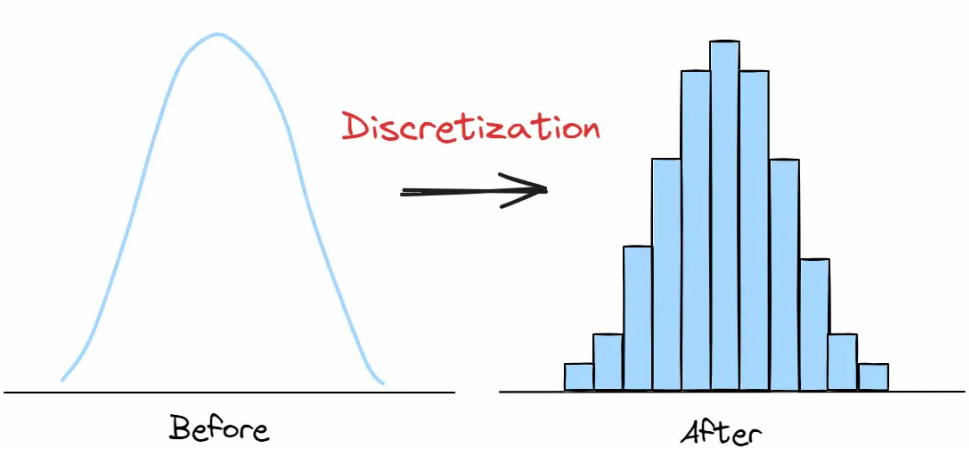

Рис. 2.4. Процес дискретизації

4. Генерація концептуальної ієрархії (Conceptual Hierarchy Generation)

Концептуальна ієрархія для числового атрибуту визначає його дискретизацію. Ці ієрархії використовуються для зменшення даних шляхом збору та заміни низькорівневих концепцій (наприклад, числових значень зросту) високорівневими концепціями (наприклад, низький, середній, високий). Хоча деталі втрачаються, узагальнені дані можуть бути більш значущими та легшими для інтерпретації. Аналіз зменшеного набору даних вимагає менше операцій I/O та є ефективнішим.

5. Зменшення розмірності (Dimensionality Reduction)

Зменшення розмірності — це процес скорочення кількості атрибутів з метою уникнення прокляття розмірності (curse of dimensionality).

Вибір ознак (Feature Selection), мета — знайти мінімальний набір атрибутів, який забезпечує точність оцінки, зіставну з використанням усіх атрибутів. Це підвищує інтерпретованість виявлених закономірностей та скорочує час обробки.

Виділення ознак (Feature Extraction) створює нові ознаки з функцій оригінальних ознак, трансформуючи дані у менший простір. Аналіз Головних Компонент (PCA) є прикладом, який використовує ортогональні перетворення для проектування даних у набір лінійно некорельованих змінних (головних компонент).

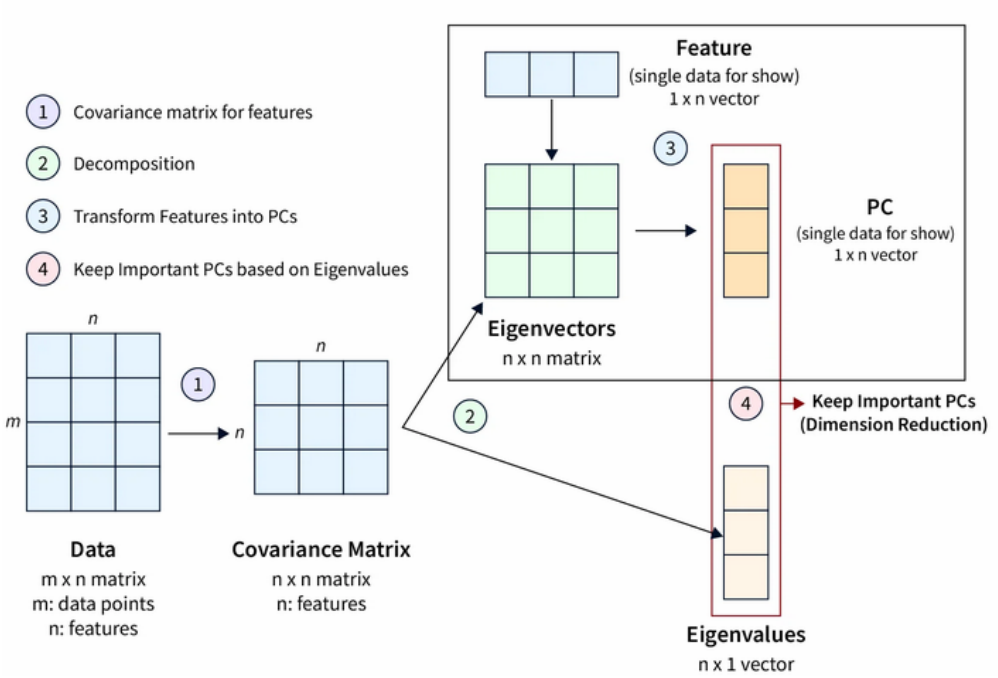

Рис. 2.5. Процес зменшення розмірності

6. K-Найближчих Сусідів (K-Nearest Neighbors, KNN)

Деякі схеми зменшення даних ґрунтуються на підходах класифікації, зокрема на правилі k-найближчих сусідів (KNN). Ефективність зменшеної множини вимірюється в термінах точності класифікації. Ці методи прагнуть отримати мінімальну узгоджену множину, яка правильно класифікує всі оригінальні зразки (наприклад, правило конденсованого найближчого сусіда,

CNN). Відстань до k-го найближчого сусіда також слугує локальною оцінкою щільності.

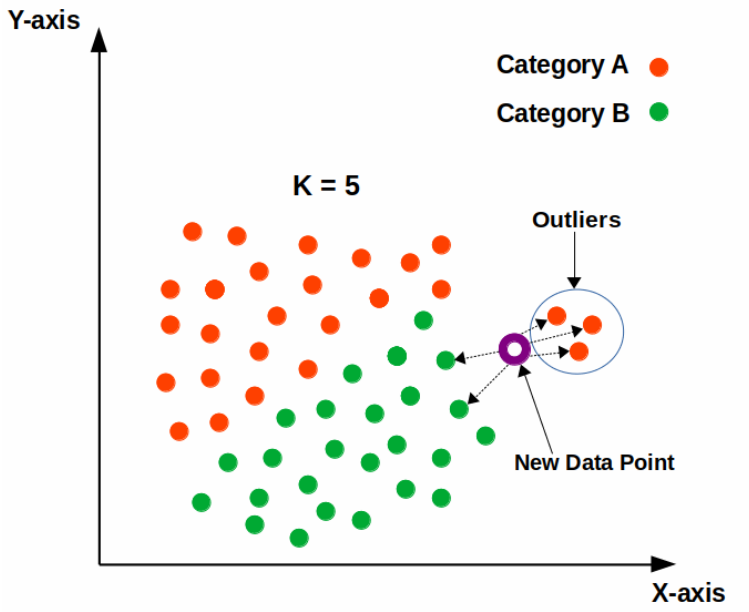

Рис. 2.6. Представлення правила k-найближчих сусідів

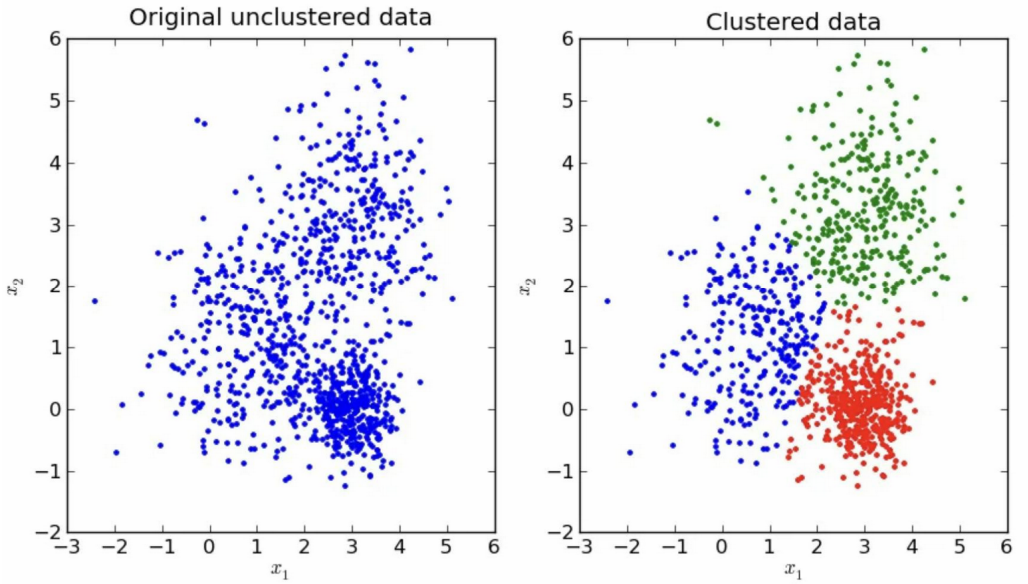

Рис. 2.7. Процес кластеризації даних

7. Кластеризація (Clustering)

Кластеризація — це метод неконтрольованого машинного навчання (unsupervised learning), спрямований на групування набору об'єктів таким

чином, що об'єкти в одній групі (кластері) є більш подібними один до одного, ніж до об'єктів в інших групах (кластерах). Кластеризація розділяє об'єкти даних на групи (кластери) так, що об'єкти в межах кластера подібні та відмінні від об'єктів в інших кластерах. Кластеризація може бути застосована для зменшення даних, коли представники кластерів (наприклад, центри кластерів, медоїди, або щільні точки) використовуються для заміни фактичних даних. Ефективне зменшення даних досягається при невеликому коефіцієнті зменшення (відношення кількості представників до кількості об'єктів) та високій здатності представників описувати свої кластери.

Висновки до розділу

Другий розділ присвячений детальному дослідженню концептуальних особливостей процесів видобування та структуризації просторово-часових даних, що дало змогу визначити ключові етапи відкриття знань у базах даних. У ході аналізу було показано, що процес KDD (Knowledge Discovery in Databases) є багатоступеневим і включає інтеграцію підготовки, трансформації та моделювання даних. Виявлено, що специфіка просторово-часового видобування даних пов’язана з необхідністю урахування як просторової взаємозалежності, так і часової динаміки.

Було узагальнено типологію задач, які варіюються від кластеризації й виявлення аномалій до прогнозування руху об’єктів. З’ясовано, що головні виклики у видобуванні просторово-часових даних пов’язані зі складною структурою, великою розмірністю та наявністю шумів. Обґрунтовано архітектуру системи видобування просторово-часових даних, яка включає модулі збору, обробки, зменшення, моделювання та візуалізації даних. Значну увагу приділено методологіям зменшення даних, що дозволяють зберегти їхню інформаційну цінність при значному скороченні обсягів. Проаналізовано сучасні підходи до візуалізації просторово-часових структур, які сприяють інтерпретуванню прихованих закономірностей.

Розділ сформував цілісне розуміння процесів структуризації даних на концептуальному та технологічному рівнях. У підсумку, отримані результати підтверджують необхідність комплексного, багаторівневого підходу до просторово-часового аналізу із застосуванням передових методів штучного інтелекту.

РОЗДІЛ 3. РЕАЛІЗАЦІЯ МОДЕЛІ ТА ПІДХОДУ ОБРОБКИ ТА СТРУКТУРИЗАЦІЇ ПРОСТОРОВО-ЧАСОВИХ ДАНИХ

3.1. Модель стиснення знань для просторово-часових даних

У цьому розділі пропонується модель стиснення даних, спеціально розроблена для просторово-часових наборів даних. Ця модель використовує гібридну стратегію зменшення даних, засновану на кластеризації, для ефективного аналізу дуже великих просторово-часових масивів.

Незважаючи на значні дослідження у сферах аналізу просторово-часових даних та зменшення даних [22, 11, 14], у науковій літературі існує обмежена кількість інформації щодо зменшення даних, яке базується на методах видобутку даних, особливо при застосуванні до просторово-часових масивів. Як зазначено раніше, більшість сучасних методів зменшення даних ігнорують просторові властивості наборів. Наскільки нам відомо, лише два підходи в цій парадигмі пропонують метод зменшення даних, орієнтований на знання, з використанням алгоритмів кластеризації. У цих роботах дослідники вивчали можливість застосування методів видобутку даних для зменшення обсягу просторово-часових наборів без втрати важливої просторової інформації шляхом екстракції суттєвих знань.

На рисунку 3.1 проілюстровано огляд запропонованої моделі стиснення даних.

Основна ідея полягає у зменшенні розміру даних шляхом створення меншого, орієнтованого на знання представлення набору даних, що контрастує з традиційним стисненням даних, яке вимагає подальшого розпакування для повторного використання.

Модель оперує чотирма послідовними станами даних:

1. Сирі дані (raw data) - містить вихідні реальні дані. Ці дані є великомасштабними та потребують очищення (cleaning). Вони типово неповні (відсутні значення атрибутів, агреговані дані), неузгоджені (різні

формати) та зашумлені (містять помилки, викиди). На рисунку 3.1 представлені порожніми колами.

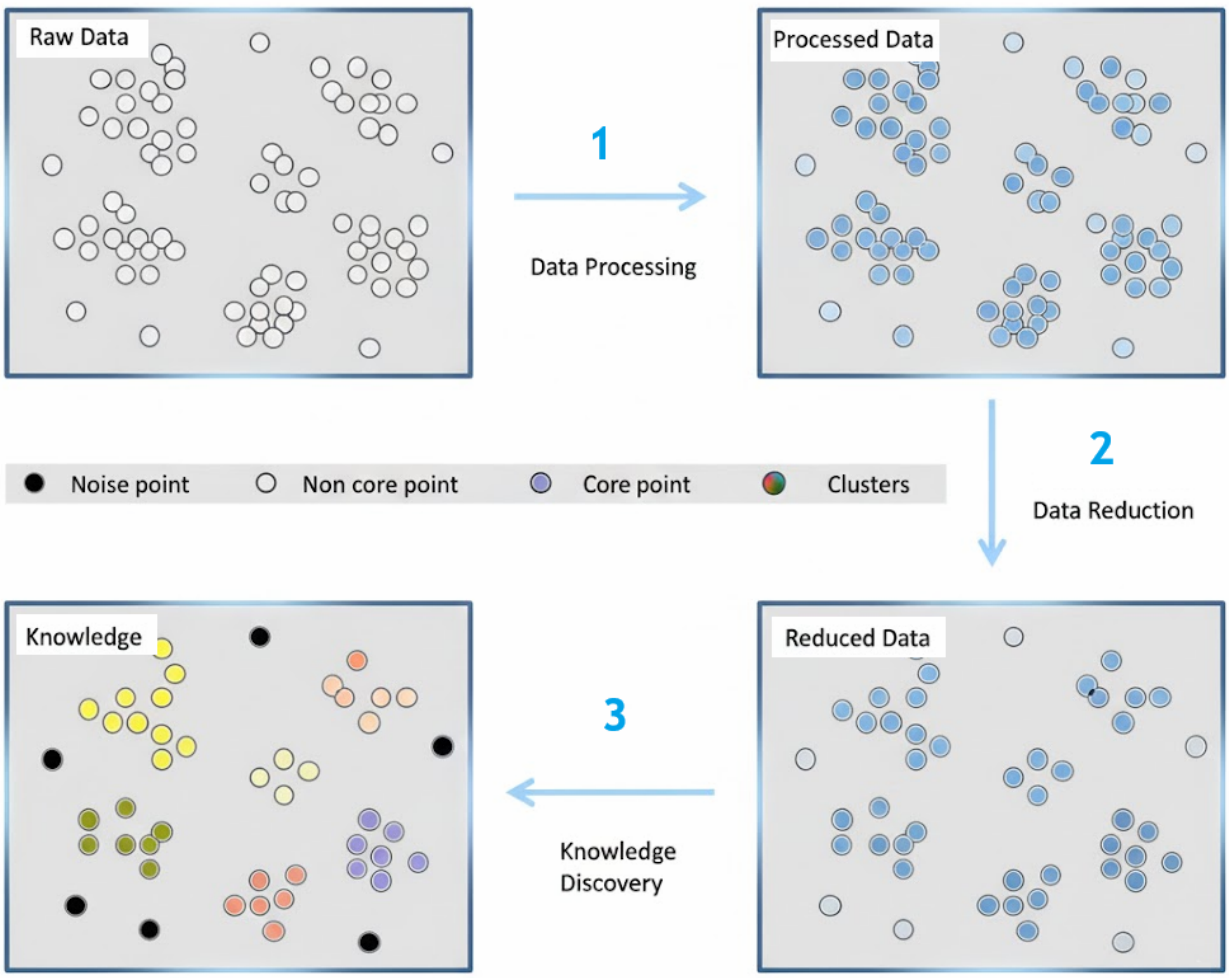

Рис. 3.1. Модель стиснення даних

2. Оброблені дані (processed data) - містить результати всієї необхідної попередньої обробки (очищення, трансформація, інтеграція), а також результати застосування нашої стратегії зменшення даних на основі кластеризації. Представлені синіми колами, що позначають вибраних представників кластерів.

3. Зменшені дані (reduced data) - містить зменшене представлення оригінальних даних, сформоване на основі результатів другого етапу. Врешті-решт, зберігаються лише найбільш корисні дані. На рисунку 3.1 кластерні дані, які не є представниками, видаляються, створюючи новий

зменшений набір, що включає представників кластерів (сині кола) та некластерні дані (порожні кола).

4. Знання (Knowledge). Містить знання, отримані в результаті візуалізації або застосування алгоритмів видобутку даних до зменшених даних третього етапу. Ці знання зазвичай мають форму цікавих закономірностей (правила класифікації, регресійні моделі, кластери, нейронні мережі). На рисунку 3.1 знання отримані з результатів алгоритму кластеризації, що чітко ідентифікує шість кластерів (кольорові кола) та викиди/шум (чорні кола).

Модель функціонує через три основні етапи:

1. Обробка даних (data processing).

Цей етап виконує основний обсяг роботи і є еквівалентним етапу попередньої обробки. Він включає очищення даних, трансформацію та всю підготовчу роботу для розробленої стратегії зменшення даних. На основі просторово-часової інформації тут розраховуються індексація даних, матриця подібності та граф подібності SNN (Shared Nearest Neighbor).

2. Зменшення даних (data reduction).

Після завершення обробки на попередньому етапі, знання, отримані в процесі видобутку, використовуються для вибору репрезентантів для формування зменшеного набору даних.

3. Відкриття знань (knowledge discovery).

На цьому етапі зменшені дані можуть бути проаналізовані (самостійно або у поєднанні з іншими зменшеними масивами) для генерації корисних знань (моделей, закономірностей, правил) шляхом застосування інших методів видобутку даних.

Оскільки дуже великі просторово-часові набори даних є обчислювально складними для будь-якого традиційного алгоритму видобутку, модель застосовує двоетапну стратегію для подолання цього виклику. Мета полягає у зменшенні розміру даних шляхом створення меншого, керованого представлення набору даних.

Етап 1 (групування та репрезентація). Метою є групування даних відповідно до їхньої подібності та репрезентація цих груп без втрати релевантної інформації (відповідає крокам 1 та 2).

Етап 2 (аналіз та інтерпретація). Метою є застосування методів видобутку (кластеризація, асоціативні правила) до нових представників даних для екстракції нових знань та підготовки до їхньої оцінки та інтерпретації (відповідає кроку 3).

3.2. Кластеризація як стратегія зменшення просторово-часових даних

Для сприяння зменшенню просторово-часових наборів даних ми пропонуємо застосування кластеризації як методу видобутку даних. Кластеризація є одним із фундаментальних методів у видобутку даних і володіє ключовими характеристиками, що роблять її корисною для обробки великомасштабних масивів:

1. Групування за властивостями здійснює групування об'єктів даних на основі їхніх внутрішніх характеристик та взаємовідносин.

2. Оптимізація подібності прагне максимізувати подібність об'єктів у межах кластера (внутрішньокластерна когезія) та максимізувати відмінність між кластерами (міжкластерна сепарація) з метою виявлення цікавих структур у базових даних.

3. Репрезентативність дозволяє представляти великі групи даних за допомогою різноманітних кластерних властивостей (наприклад, центр кластера, представники кластерів або основні точки). Це дозволяє використовувати ці репрезентанти для візуалізації або аналізу замість роботи з повним обсягом сирих даних, зберігаючи при цьому важливу інформацію.

Переваги кластеризації для просторово-часового аналізу:

- Візуалізація кластерів сприяє інтерпретації внутрішньої структури просторово-часових даних.

- Використовує спрощені метрики подібності для нівелювання складності даних, включаючи високу кількість атрибутів.

- Використовує представників кластерів для фільтрації (зменшення) даних без втрати критичної/цікавої інформації.

- Використовує важливу властивість просторово-часових даних (тобто об'єкти, які є фізично та часово близькими, мають тенденцію до подібності).

На рисунку 3.2 представлено розширений вигляд запропонованої моделі стиснення з інтегрованою стратегією на основі кластеризації.

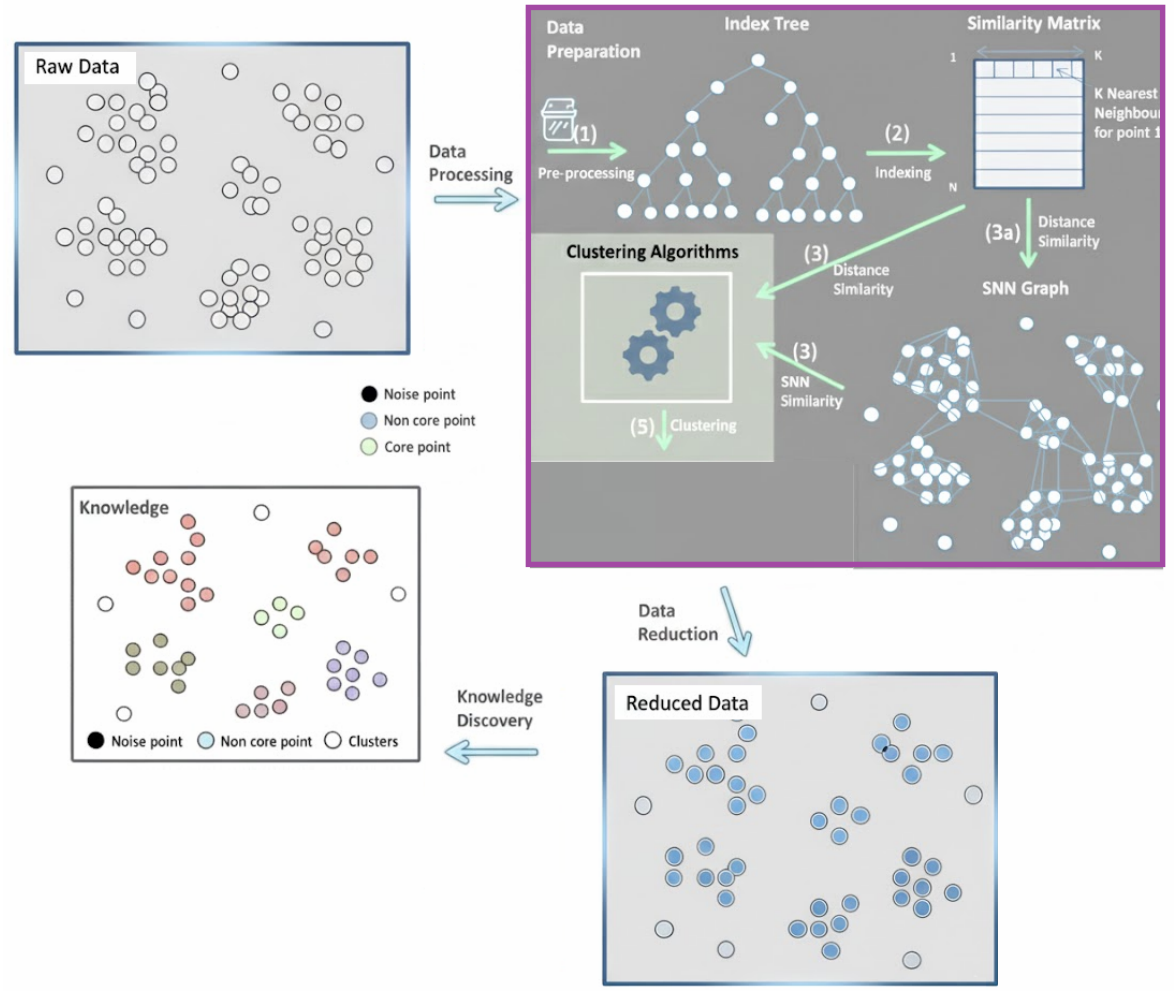

Рис. 3.2. Модель стиснення на основі стратегії кластеризації

Процес обробки сирих даних включає п'ять послідовних кроків:

1. Попередня обробка (Pre-processing).

2. Індексація (Indexing).

3. Подібність відстані (Distance Similarity).

4. Подібність найближчих сусідів (SNN) (Shared Nearest Neighbor Similarity).

5. Кластеризація (Clustering).

Ці кроки можна логічно об'єднати у три основні фази: попередня обробка даних, підготовка даних та зменшення даних.

Крок попередньої обробки є фундаментальним у будь-якому процесі видобутку даних, оскільки реальні дані зазвичай демонструють такі проблеми:

- Характеризується відсутніми значеннями атрибутів, відсутністю певних атрибутів інтересу або наявністю лише агрегованих даних.

- Містить помилки або викиди.

- Містить розбіжності у кодах або назвах.

Для вирішення цих проблем може бути виконана низка завдань (залежно від стану даних):

- Заповнення відсутніх значень, згладжування зашумлених даних, виявлення або видалення викидів та вирішення неузгодженостей.

- Об'єднання даних, отриманих із кількох баз даних, кубів даних або файлів.

- Трансформація даних, що включає нормалізацію та агрегацію даних.

- Скорочення обсягу даних при збереженні еквівалентних або зіставних аналітичних результатів.

- Перетворення числових атрибутів на номінальні (категоріальні).

На цьому етапі може бути виконана будь-яка необхідна кількість цих завдань.

Функціональне навантаження фази підготовки даних визначається потребами наступних фаз, зокрема, кластеризації просторово-часових даних. Ця фаза охоплює три ключові кроки:

1. Індексація (Indexing).

Оскільки робота ведеться з просторовими даними, необхідне застосування відповідного просторового індексу для локалізації об'єктів з метою прискорення часу відгуку. Метою індексації є підготовка та структурування об'єктів даних для ефективної обробки. Можливе попереднє визначення індексів, які є оптимальними для групи алгоритмів кластеризації. Для індексації точок даних було застосовано низку топологій на основі дерев (наприклад, R-дерево, R*-дерево, квадродерево, kd-дерево). У нашому підході для індексації точок даних використовується топологія kd-дерева.

2. Подібність відстані (Distance Similarity)

На цьому етапі обчислюється подібність відстані для кожного об'єкта даних. Як метрика використовується евклідова відстань. Ця подібність є вхідним параметром для алгоритмів кластеризації. Матриця подібності відстані розраховується ефективно, оскільки використовується індексне дерево для швидкого визначення k найближчих просторових сусідів для кожної точки, що вимагає обчислення лише k відстаней для кожної точки даних. Значення k є вхідним параметром для нашого підходу стиснення і повинно бути:

- Достатньо великим, щоб не виключати важливих відносин у даних.
- Достатньо малим, щоб не погіршувати продуктивність.

3. Подібність SNN (Shared Nearest Neighbor Similarity)

Граф подібності SNN будується для всіх даних з використанням матриці подібності відстані, розрахованої на попередньому кроці. Кожен вузол у графі представляє точку даних, а зв'язки від кожної точки даних вказують на найближчих сусідів точки відповідно до кількості спільних сусідів між ними. Інформація з цього графа буде використана алгоритмами кластеризації на наступному етапі.

У цьому підрозділі була представлена розроблена стратегія стиснення даних, яка базується на методі видобутку даних — кластеризації. Ми запропонували новий гібридний підхід для зменшення великомасштабних просторово-часових наборів даних.

Ключова інновація підходу полягає у поєднанні кластеризації на основі щільності та графів. Використовуючи концепцію спільних найближчих сусідів (Shared Nearest Neighbors, SNN), підхід додатково застосовує метрику евклідової відстані для уточнення подібності найближчих сусідів. Інтеграція метрики евклідової відстані була необхідна для посилення розрахунку подібності SNN для кожної основної точки, забезпечуючи більш надійну оцінку, ніж просте використання кількості спільних найближчих сусідів.

Основна перевага запропонованого гібридного підходу полягає в тому, що він ефективно вирішує проблеми низької подібності та варіативності щільності у даних. Завдяки цьому результати алгоритмів можуть бути використані як надійні представники великих наборів даних. Як наслідок, новий зменшений набір даних є значно меншим за обсягом порівняно з оригінальним масивом і може бути швидко та ефективно проаналізований для екстракції корисних знань (тобто моделей, закономірностей, правил тощо) шляхом застосування інших методів видобутку даних.

3.3. Підхід до розробки фреймворку стиснення просторово-часових даних

У цьому розділі детально описано розроблений підхід для фреймворку стиснення просторово-часових даних, що включає стратегію стиснення, представлену в попередньому підрозділі.

Основна мета фреймворку полягає у досягненні подвійної функціональності:

1. Забезпечення стиснення даних.

Дозволити інструментам видобутку даних застосовувати певну форму стиснення до аналізованих масивів.

2. Візуальна оцінка.

Надати можливість візуальної оцінки результатів процесу видобутку в інтерактивному середовищі.

Спочатку для валідації та тестування було використано великомасштабний набір даних у якості кейс-стаді. Дизайн структури був орієнтований на створення гнучкої платформи для кластеризації просторово-часових даних, здатної інтегрувати різні алгоритми кластеризації.

Одним із ключових викликів для цієї структури є обробка надзвичайно великого обсягу просторово-часових даних, оскільки їхня обчислювальна складність є непомірно високою для будь-якого традиційного алгоритму видобутку.

Для вирішення цієї проблеми у структурі застосовується двоетапна стратегія, спрямована на зменшення розміру даних шляхом створення меншого, репрезентативного представлення набору даних, що забезпечує ефективне керування та видобуток.

Етап 1 (індексація та подібність) - мета полягає в індексації даних відповідно до їхніх просторових атрибутів та розрахунку їхньої подібності найближчих сусідів (SNN).

Етап 2 (групування та репрезентація) - метою є групування даних відповідно до їхньої подібності та репрезентація цих груп без втрати релевантної інформації. Отримані репрезентанти цих груп потім готуються для оцінки та інтерпретації.

На даний час фреймворк включає дві адаптовані техніки кластеризації та один інструмент візуалізації. Ці техніки кластеризації були спеціально адаптовані до характеристик даних, використаних у кейс-стаді.

Архітектура фреймворку базується на чотиришаровій моделі (рис. 3.3):

- Шар попередньої обробки даних (data preprocessing): відповідає за очищення, інтеграцію та трансформацію сирих даних.

- Шар підготовки даних (data preparation): виконує індексацію даних та розрахунок метрик подібності (відстані, SNN).

- Шар кластеризації (clustering algorithms): запуск основних алгоритмів кластеризації для групування даних та вибору представників.

- Шар оцінки виходу (output evaluation): оцінка та візуалізація отриманих результатів кластеризації/стиснення.

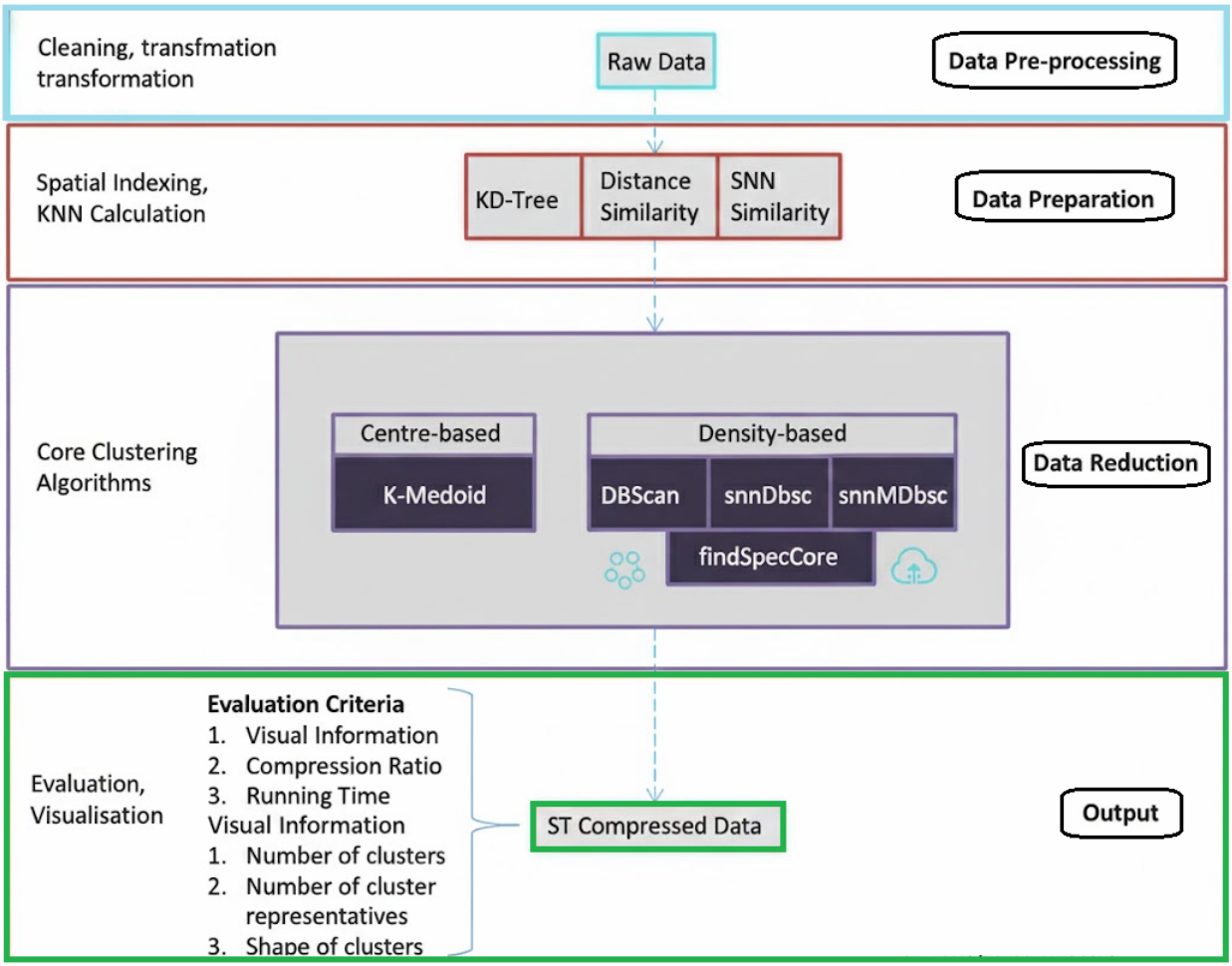

Рис. 3.3. Чотиришарова модель фреймворку стиснення просторово-часових даних

3.4. Попередня обробка підготовка просторово-часових даних

На початковому етапі обробки даних може виникнути потреба у фільтрації та трансформації даних для задоволення обмежень, накладених інструментами, алгоритмами або користувачами. Критично важливим є забезпечення низького рівня шуму в даних, а також необхідність деяких трансформацій для ефективної візуалізації великомасштабних наборів даних.

У рамках пропонованого фреймворку структури ми здійснюємо лише мінімальні, але необхідні операції попередньої обробки:

1. Фільтрація недійсних значень - видалення невизначених значень (таких як NULL або наземні значення), які можуть скомпрометувати кінцеві результати.

2. Нормалізація значень. Аналізуючи просторові та непросторові атрибути, ми стикаємося з різними діапазонами значень (наприклад, [0..500] для позиції та [0.001..0.032] для ваги пари). Отже, нормалізація є необхідною для уніфікації масштабу атрибутів.

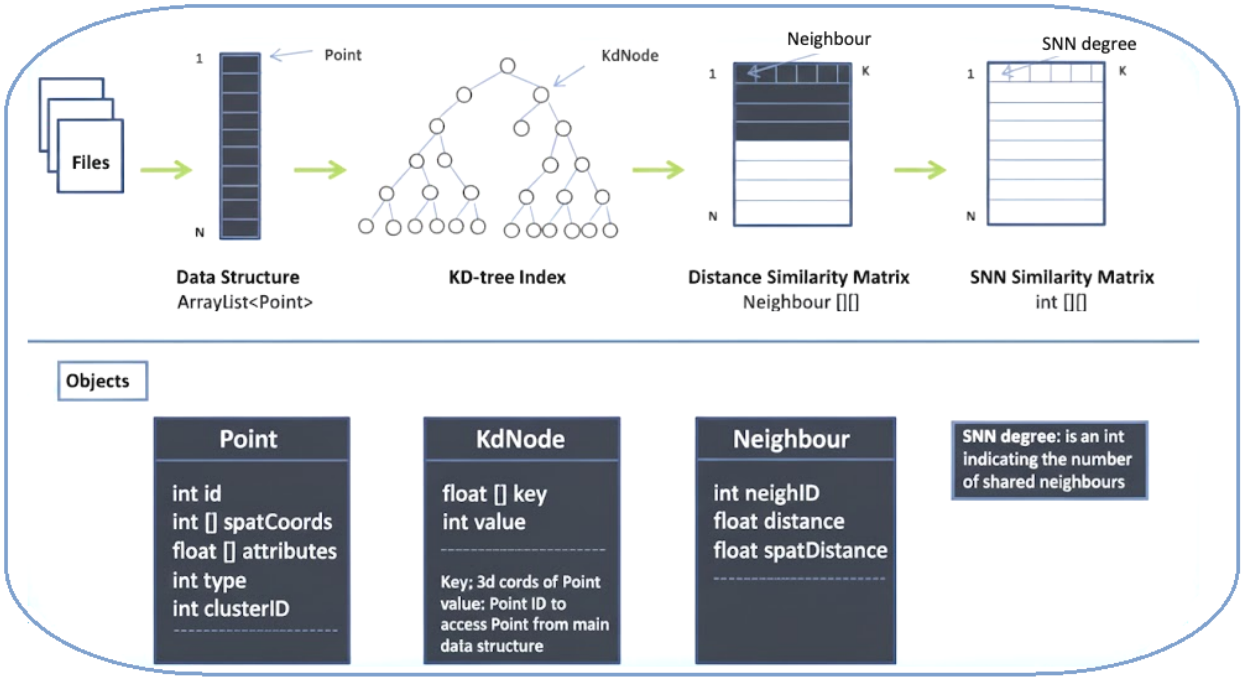

Рис. 3.4. Імплементація об’єктів для етапу підготовки даних

Ці дві операції можуть бути виконані сумісно за одне сканування набору даних.

Фаза підготовки даних безпосередньо залежить від техніки видобутку та аналізу. У нашому випадку, ця фаза включає два основні завдання:

- Визначення відповідного індексу для ефективної роботи з просторово-часовими даними.

- Розрахунок подібностей k-найближчих сусідів за евклідовою відстанню та матриць щільності SNN (Shared Nearest Neighbor) для підтримки кластеризації на основі щільності.

На рисунку 3.4 показано реалізовані об'єкти для етапу підготовки даних.

Для індексації просторово-часових даних використовується k-вимірне дерево (k-d дерево) [7]. k-d дерево є бінарним деревом просторового поділу, призначеним для організації точок у k-вимірному просторі.

Принцип побудови:

1. Кожен рівень дерева розділяє всіх нащадків уздовж певного виміру за допомогою гіперплощини, перпендикулярної відповідній осі.

2. На вершині (корені) поділ відбувається за першим виміром. На кожному наступному рівні використовується наступний вимір у циклічному порядку.

3. Для побудови збалансованого k-d дерева використовується метод поділу, подібний до Quick Sort, де медіанна точка поміщається в корінь. Це забезпечує приблизно однакову відстань кожного листового вузла від кореня, що критично важливо для ефективного пошуку найближчих сусідів.

Особливості реалізації:

- Для наших просторово-часових даних ми працюємо з тривимірними просторовими координатами, тому k-d дерево є тривимірним (k=3), що дозволяє уникнути прокляття розмірності у високовимірних просторах.

- Цикл площин поділу: X-вісь → Y-вісь → Z-вісь → X-вісь...

- Найгірша часова складність побудови такого збалансованого дерева становить O(nlogn).

Використання цієї індексації дозволяє швидко розрахувати k-найближчих сусідів для точки, що експлуатує властивість просторово-часових даних: просторово та тимчасово близькі об'єкти мають тенденцію до подібності.

Після побудови індексного дерева, формується матриця подібності відстані, яка містить k-найближчих сусідів для кожної точки.

Лістинг 3.1. Алгоритм пошуку k-найближчих сусідів

```
def searchKnn(cNode, pId, knnList, lev, K):

    if cNode is None:
        return knnList

    if knnList is not None and len(knnList) > 0:
        knnDist = max(knnList).getDistance()
    else:
        knnDist = float('inf')


    if cNode.val != pId:
        current_distance = calcDistance(cNode.val, pId)

        if current_distance < knnDist:
            if knnList.isFull():
                knnList.remove(max(knnList))

            knnList.add(Neighbour(cNode.val, current_distance))

            knnDist = max(knnList).getDistance()

    searchKey = getPoint(pId).getRealCoords()
    nextLevel = (lev + 1) % K

    if compare(searchKey, cNode.key, lev) > 0:
        child = cNode.right
        knnList = searchKnn(child, pId, knnList, nextLevel, K)
    else:
        child = cNode.left
        knnList = searchKnn(child, pId, knnList, nextLevel, K)

    if knnList is not None and len(knnList) > 0:
        knnDist = max(knnList).getDistance()

    if knnList is not None:
        if distanceAxis(cNode.key, searchKey, lev) < knnDist:

            if compare(searchKey, cNode.key, lev) > 0:
                child = cNode.left
                knnList = searchKnn(child, pId, knnList, nextLevel, K)
            else:
                child = cNode.right
                knnList = searchKnn(child, pId, knnList, nextLevel, K)

    return knnList
```

Алгоритм пошуку k-найближчих сусідів рекурсивним пошуком k-найближчих сусідів у k-d дереві, використовуючи відсікання пошукового простору для оптимізації (лістинг 3.1):

- Використовує k-d дерево для рекурсивного пошуку.

- Спочатку досліджує близьку гілку, а потім віддалену гілку, лише якщо відстань від поточної точки до площини розділу менша за відстань k-го сусіда (для оптимізації відсікання пошукового простору).

- Середній час пошуку k-найближчих сусідів для точки становить O(klogn).

Лістинг 3.2. Створення матриці відстані

```
def createDistMatrix(dataset, numOfNeighbours, root):
    # dataset is the set of points to create the distanceMatrix for
    # numOfNeighbours is the fixed value for the number of neighbours to calculate
    # distanceMatrix is the matrix of distances for the knn of each point to be re

    # Initialization
    # Assuming Neighbour is a custom data structure/class used for the matrix
    distanceMatrix = newNeighbour[dataset.size][numOfNeighbours]

    for i in range(dataset.size):
        # Assuming searchKnn takes root, point_id, initial_knnList (NULL), and ini
        # Assuming K is globally defined or accessible
        knnList = searchKnn(root, i, NULL, 0)

        # sort based on non-spatial dimensions
        sort(knnList)

        for j in range(numOfNeighbours):
            distanceMatrix[i][j] = knnList.get(j)

    return distanceMatrix
```

Створення матриці відстані (лістинг 3.2):

- Для кожної точки викликається searchKnn.

- Повернений список k-найближчих сусідів сортується відповідно до непросторових атрибутів перед додаванням до матриці подібності. Це забезпечує, що непросторові властивості враховуються при виборі найближчих сусідів.

- Матриця зберігає відстані до k найближчих просторово-часових сусідів.

Матриця подібності SNN будується з використанням інформації з матриці подібності відстані (лістинг 3.3).

Лістинг 3.3. Алгоритм створення матриці подібності спільних найближчих сусідів

```
def createSNNMatrix(dataset, numOfNeighbours, distanceMatrix):
    # dataset is the set of points to create the snnMatrix for
    # numOfNeighbours is the fixed value for the number of neighbours to calculate
    # distanceMatrix is the matrix of distances for the knn of each point (assumed
    # snnMatrix is the matrix of SNN degrees for the knn of each point to be retur

    # Initialization
    # Assuming int is the type for SNN degrees
    snnMatrix = int[dataset.size][numOfNeighbours]

    for pointA in range(dataset.size):
        for j in range(numOfNeighbours):

            # Get the ID of the j-th nearest neighbour of pointA
            pointB = distanceMatrix[pointA][j].getId()

            # Check if pointA is in the neighbourhood of pointB
            if completeNeighbours(pointA, pointB) == true:
                snnDegree = 0

                # Calculate the SNN degree
                for neighA in range(numOfNeighbours):
                    for neighB in range(numOfNeighbours):

                        # Compare the IDs of the neighbours of pointA and pointB
                        if distanceMatrix[pointA][neighA].getId() == distanceMatri
                            snnDegree = snnDegree + 1

                # Assign the calculated SNN degree
                snnMatrix[pointA][j] = snnDegree

            else:
                # If pointA is not in the neighbourhood of pointB, SNN degree is 0
                snnMatrix[pointA][j] = 0

    return snnMatrix
```

Має такий самий розмір, як і матриця відстані (n рядків на k стовпців). Кожне значення вказує ступінь SNN між точкою та кожним з її сусідів (тобто кількість спільних найближчих сусідів). Використання матриці SNN та подібності відстані будуть використані на наступній фазі для розрахунку щільності точки.

Використання підходу SNN значно зменшує вимоги до пам'яті порівняно з повною матрицею несумісності (несхідності).

Для набору даних, що містить N точок, повна матриця несхідності вимагає O(N2) пам'яті. Використовуючи алгоритм SNN, необхідно зберігати лише відстані до k-найближчих сусідів для кожної точки. Це вимагає $O(N \cdot k)$

пам'яті, що для того ж набору даних становить в сотні раз менше пам’яті, демонструючи високу масштабованість підходу.

3.5. Представлення фази зменшення даних та алгоритми кластеризації

Фаза зменшення даних є критичним етапом, на якому репрезентанти кластерів ідентифікуються з використанням інформації, підготовленої на попередньому етапі. У цій фазі було реалізовано чотири алгоритми кластеризації: K-Medoids (на основі центрів) та три алгоритми на основі щільності. Три алгоритми на основі щільності включають оригінальний DBSCAN та дві реалізовані версії, які поєднують концепцію SNN (Shared Nearest Neighbor) з кластеризацією за щільністю.

3.5.1. Просторова кластеризація на основі щільності з використанням найближчих сусідів

Першим реалізованим алгоритмом є SNNDBSC (Spatial Nearest Neighbor Density-Based Spatial Clustering). Код реалізації представлений в лістингу 3.4.

Лістинг 3.4. Алгоритм SNNDBSC

```
def SNNDBSC(dataset, ε, minPts):
    # dataset is the set of points to be clustered
    # ε is the minimum number of SNN between two points
    # minPts is minimum number of neighbours with high SNN.

    cId = 1
    for i in range(dataset.size()):
        p = dataset.getPoint(i)
        if p.getCId() == UNDEFINED_CID:
            if expandCluster(dataset, p, cId, ε, minPts) == true:
                cId = cId + 1
    return cId

def expandCluster(dataset, p, cId, ε, minPts):
    seeds = regionQuery(p.getId(), ε)
    seedsize = seeds.size()

    if seedsize < minPts:
        p.setType(NOISE)
```

```
            p.setCId(NOISE)
            return false
        else:
            p.setType(CORE)
            p.setCId(cId)

            for pt in seeds:
                dataset.getPoint(pt).setCId(cId)

            i = 0
            while i < seedsize:
                currentSeeds = regionQuery(seeds.get(i), ε)
                if currentSeeds.size() >= minPts:
                    dataset.getPoint(seeds.get(i)).setType(CORE)
                    for j in range(currentSeeds.size()):
                        resultP = currentSeeds.get(j)
                        if dataset.getPoint(resultP).getCId() == UNDEF:
                            seeds.add(resultP)
                            seedsize = seedsize + 1
                            dataset.getPoint(resultP).setCId(cId)
                        if dataset.getPoint(resultP).getCId() == NOISE:
                            dataset.getPoint(resultP).setCId(cId)
                            dataset.getPoint(resultP).setType(BORDER)
                else:
                    dataset.getPoint(seeds.get(i)).setType(BORDER)
                i = i + 1

            return true

    # NEW Region query for SNNDBSC
    def regionQuery(pid, minSNN):
        seeds = NEWList
        for i in range(numOfNeighbours):
            # Different to original DBSCAN
            if snnMatrix[pid][i] >= minSNN:
                seeds.add(distanceMatrix[pid][i].getId())
        return seeds
```

SNNDBSC тісно наслідує оригінальний DBSCAN, але ключова відмінність полягає у функції regionQuery:

- DBSCAN використовує distanceMatrix для визначення близькості сусідів за евклідовою відстанню відносно порогового значення ε.

- SNNDBSC звертається до snnMatrix для перевірки, скільки сусідів точки знаходяться поблизу, вимірюючи їхній ступінь SNN (Shared Nearest Neighbor) відносно параметра ε (що тут позначає мінімальну кількість SNN).

Вибір параметрів ε та minPts є важливим для всіх алгоритмів, що є похідними від DBSCAN. Для SNNDBSC критичним є значення k, яке використовується для розрахунку KNN у матрицях відстані та SNN.

Низьке k призводить до великої кількості малих, але щільних кластерів, особливо за наявності локальних варіацій у подібності.

Високе k згладжує локальні варіації, що призводить до меншої кількості великих, добре розділених кластерів.

Оскільки в SNNDBSC точка може бути подібною максимум до k інших точок, значення параметрів minPts та ε (мінімальний ступінь SNN) мають бути лише часткою розміру списку найближчих сусідів k.

3.5.2. Просторова кластеризація на основі метрики найближчих сусідів та щільності

Другий алгоритм, SNNMDBSC (Spatial Nearest Neighbor Metric-based Density-Based Spatial Clustering), є гібридним алгоритмом SNN. Код реалізації представлений в лістингу 3.5.

Лістинг 3.5. Алгоритм SNNMDBSC

```
def SNNMDBSC(dataset, ε, minPts):
    # dataset is the set of points to be clustered
    # ε is the metric radius to define the neighbourhood of a point
    # minPts is minimum value for the sum of SNN within ε neighbourhood.

    cId = 1
    for i in range(dataset.size()):
        p = dataset.getPoint(i)
        if p.getCId() == UNDEFINED_CID:
            if expandCluster(dataset, p, cId, ε, minPts) == true:
                cId = cId + 1
    return cId

def expandCluster(dataset, p, cId, ε, minPts):
    seeds = regionQuery(p.getId(), ε)
    seedsize = seeds.size()

    # New for SNNMDBSC: check if point is core using snnSum
    if snnSum(p.getId(), ε) < minPts:
        p.setType(NOISE)
        p.setCId(NOISE)
        return false
    else:
        p.setType(CORE)
        p.setCId(cId)

        for pt in seeds:
            dataset.getPoint(pt).setCId(cId)

        i = 0
        while i < seedsize:
            currentSeeds = regionQuery(seeds.get(i), ε)
            # New for SNNMDBSC: check if point is core using snnSum
```

```
            if snnSum(seeds.get(i), ε) >= minPts:
                dataset.getPoint(seeds.get(i)).setType(CORE)
                for j in range(currentSeeds.size()):
                    resultP = currentSeeds.get(j)
                    if dataset.getPoint(resultP).getCId() == UNDEF:
                        seeds.add(resultP)
                        seedsize = seedsize + 1
                        dataset.getPoint(resultP).setCId(cId)
                    if dataset.getPoint(resultP).getCId() == NOISE:
                        dataset.getPoint(resultP).setCId(cId)
                        dataset.getPoint(resultP).setType(BORDER)
            else:
                dataset.getPoint(seeds.get(i)).setType(BORDER)
            i = i + 1

        return true

# Region query for SNNMDBSC, same as original DBScan
def regionQuery(pid, ε):
    seeds = NEWList
    for i in range(numOfNeighbours):
        if distanceMatrix[pid][i].getDist() < ε:
            seeds.add(distanceMatrix[pid][i].getId())
    return seeds

# Calculation to check if point is core for SNNMDBSC
def snnSum(pid, ε):
    sum = 0
    # Assuming numOfNeighbours is known/accessible in this scope
    for i in range(numOfNeighbours):
        # Assuming snnMatrix and distanceMatrix are accessible
        if distanceMatrix[pid][i].getDist() < ε:
            sum = sum + snnMatrix[pid][i]
    return sum
```

Основна відмінність SNNMDBSC від SNNDBSC полягає у додаванні значення метричного радіусу (ε). Це значення використовується для посилення розрахунку щільності точки шляхом врахування суми ступенів SNN лише для тих сусідів, які знаходяться в межах метричного радіусу ε цієї точки.

Зміни в алгоритмі поданому в лістингу 3.5:

Функція regionQuery. Повертається до методу, який використовується в оригінальному DBSCAN (використовує евклідову відстань для визначення ε-околиці).

Перевірка основної точки (Core Point). Використовується новий метод snnSum. Точка є основною, якщо сума ступенів SNN її сусідів у межах метричного радіусу ε більша за задане значення minPts.

Функція snnSum розраховує суму ступенів SNN для сусідів точки, які знаходяться на відстані, меншій за ε.

Параметризація SNNMDBSC

- ε (метричний радіус) - визначається евклідовою відстанню та використовується для визначення околиці точки, подібно до DBSCAN.

- minPts вибрати складніше, оскільки це значення не обмежено розміром k, як у SNNDBSC. Воно визначає мінімальне значення для суми SNN у межах ε-околиці.

3.5.3. Пошук так зменшення кількості основних представників

Після ідентифікації основних точок (core points) наступним кроком є зменшення остаточної кількості представників шляхом знаходження специфічних основних точок (specific core points). Ця концепція базується на визначенні, що специфічна основна точка не повинна бути "покрита" іншою специфічною основною точкою.

Лістинг 3.6. Алгоритм findSpecCore

```
def findSpecCore(dataset, ε, numOfNeighbours, distanceMatrix, sCore):
    # dataset – набір точок для кластеризації
    # ε – метричний радіус для визначення околиці точки (евклідова відстань).
    # numOfNeighbours – фіксована кількість сусідів, використана в KNN
    # distanceMatrix – попередньо розрахована матриця відстаней
    # sCore – булевий список для вказівки, чи є точка вже специфічною основною точкою

    setOfSCore = NEWList # Список для зберігання ідентифікованих Специфічних Основних Т

    for i in range(dataset.size()):
        p = dataset.getPoint(i)

        # Обробляємо лише основні точки (CORE points)
        if p.getType() == CORE:
            j = 0
            foundSCore = false # вказує, чи покрита основна точка іншою специфічною ос
            distMax = distanceMatrix[i][j].getDist() # для SNNMDBSC: Ініціалізуємо м

            # Ітеруємося по k-найближчих сусідах
            while j < numOfNeighbours AND foundSCore == false:
                numNeigh = distanceMatrix[i][j].getId()
                neighbour = dataset.getPoint(numNeigh)
                dist = distanceMatrix[i][j].getDist()
```

```
            # Перевірка: відстань менша за ε АБО сусід є CORE точкою АБО належить до тог
            if dist < ε AND neighbour.getType() == CORE AND neighbour.getClusterId()

                # Оновлюємо максимальну відстань (використовується для об'єкта SCorePoin
                if dist > distMax:
                    distMax = dist

                # Перевіряємо, чи є сусід вже Специфічною Основною точкою
                if sCore[numNeigh]: # Булевий список для вказівки sCore точок
                    foundSCore = true

            j = j + 1

        # Якщо p не покрита жодною існуючою SCore точкою, тоді p є SCore точкою
        if foundSCore == false:
            # Зберігаємо точку разом із максимальною відстанню покриття
            setOfSCore.add(SCorePoint(p, distMax)) # для SNNMDBSC
            sCore[p.getId()] = true

    return setOfSCore
```

Цей алгоритм використовується для ідентифікації специфічних основних точок (Specific Core Points) після виконання кластеризації SNNMDBSC. Основна точка p вибирається як специфічна основна точка, якщо жодна інша специфічна основна точка у її ε-околиці не має меншого ступеня SNN (тобто ступінь SNN ≥ε між ними). Використовує ε як мінімальний ступінь SNN.

Для кращої візуалізації та оцінки результатів було реалізовано механізм об'єднання кластерів (cluster merging). Об'єднує два кластери, якщо вони мають прикордонні точки (BORDER), між якими ступінь SNN більший або дорівнює заданому пороговому значенню ε.

3.5.4. Оцінка результатів та візуалізація

Остання фаза структури спрямована на оцінку та візуалізацію результатів процесу зменшення даних. При цьому процес зменшення повинен гарантувати не лише ефективний коефіцієнт стиснення, але й збереження критично важливої інформації вихідного набору даних.

Для оцінки ефективності та якості зменшення даних пропонується використовувати такі критерії:

- Продуктивність усіх алгоритмів кластеризації на основі щільності, включаючи SNNDBSC та SNNMDBSC, критично залежить від вибору початкових параметрів. Оскільки обидва розроблені алгоритми є похідними від DBSCAN, основними розглянутими параметрами є (minPts, ε). Вибір ефективних параметрів залишається складним завданням.

- Коефіцієнт стиснення: кількісний показник, що визначається як співвідношення розміру даних до та після процедури зменшення.

- Кількість кластерів. Хоча алгоритми SNNDBSC та SNNMDBSC зосереджені на виборі репрезентативних об'єктів (специфічних основних точок), а не на кластерах як таких, кількість згенерованих кластерів впливає на вибір представників. Надмірно висока кількість кластерів може свідчити про наявність шумових об'єктів серед представників, оскільки розміри деяких кластерів можуть бути відносно малими. Кількість кластерів також тісно залежить від вибору початкових параметрів.

- Збереження форми кластерів. Цей критерій використовується для підтвердження того, що вибрані представники зберігають важливу інформацію про форму просторово-часових наборів даних.

У цьому підрозділі було детально описано реалізацію двох нових гібридних підходів до кластеризації, заснованих на комбінації кластеризації за щільністю та графів. Використання ідеї основної точки (core point) як репрезентанта даних забезпечує кращу відповідність формі даних порівняно з репрезентантами на основі центрів. Застосування концепції SNN (Shared Nearest Neighbor) дозволяє ефективно вирішувати проблему варіації щільності в даних, а відмінність між SNNDBSC та SNNMDBSC полягає у способі визначення щільності точки щодо її SNN. Обидва підходи інтегровані в запропоновану структуру зменшення просторово-часових даних.

Таким чином, запропонована структура долає обмеження більшості поточних технік зменшення даних, які часто ігнорують просторові властивості наборів даних.

3.6. Оцінка продуктивності алгоритмів зменшення даних

У цьому розділі представлена оцінка підходів до зменшення даних на основі кластеризації. Ми порівнюємо та оцінюємо продуктивність різних методів кластеризації: K-Medoids, SNNDBSC, та SNNMDBSC. Додатково проводиться порівняння з алгоритмом DBSCAN на тому ж тестовому наборі даних. Метою є демонстрація ефективності нашого рішення та порівняння різних підходів у створенні структури для зменшення даних на основі кластеризації.

Експерименти проводились на наборі візуально-просторових даних (НВПД), який має реальні характеристики. Дані мають сіткову специфікацію розміром 500×500×100.

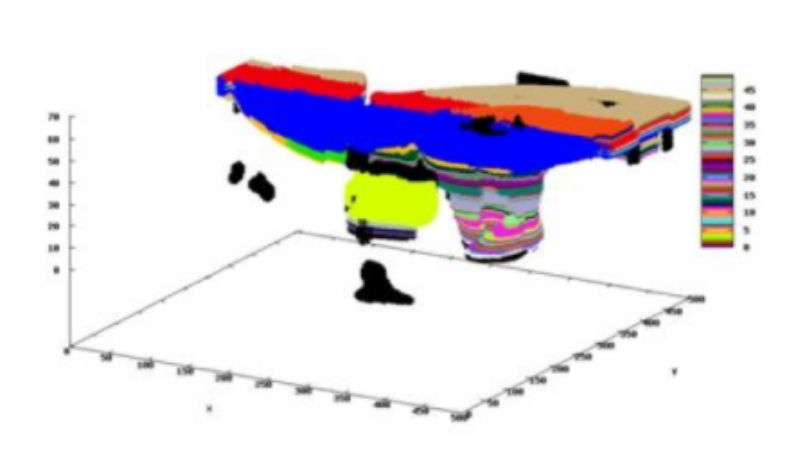

Рис. 3.5. Візуалізація початкових даних

Кожен часовий крок містить 13 непросторових атрибутів. Для оцінки було обрано один атрибут — Q, що представляє вагу хмарної води (наприклад, QCLOUD) з діапазоном значень [0..0.00332].

Тестовий набір містить точки даних із чотирма вимірами: X, Y, Z (просторові координати) та Q (непросторовий атрибут) для кожного часового кроку. Дані були очищені шляхом фільтрації значень NULL та наземних значень.

Ми починаємо з підходу масштабування як базової лінії для порівняння з кластеризацією. Як метод застосовується коефіцієнт масштабування 50×50×50 до просторових координат X,Y,Z. Це досягається вибором кожного десятого значення для X та Y, і кожного другого значення для Z (виходячи з повної сітки 500×500×100). Отримано представницькі точки (коефіцієнт зменшення 200:1). Підхід масштабування зберігає загальну форму оригінальних даних .

При застосуванні алгоритмів видобутку даних, таких як DBSCAN, до масштабованого набору, втрачається важлива географічна інформація . Кластери, утворені на масштабованому наборі, не порівнянні з кластерами, утвореними на повному наборі даних.

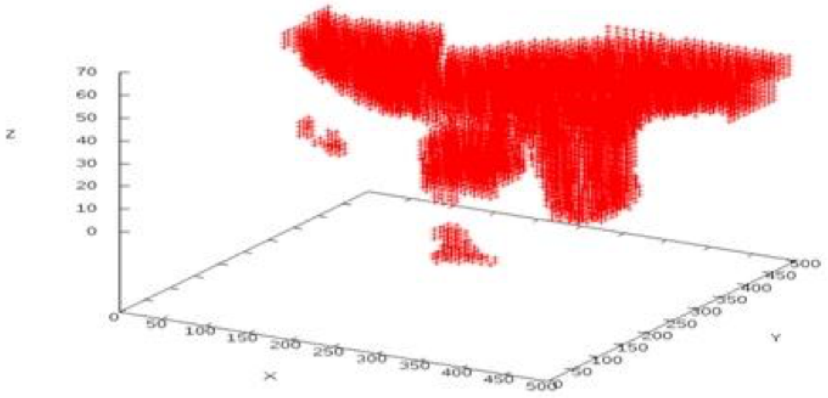

Рис. 3.6. Масштабовані дані 50 x 50 x 50

На рисунку 3.6 показано результати експериментів з масштабуванням на тестовому наборі даних для просторових координат X,Y,ZX, Y, ZX,Y,Z, атрибуту та вибраного часового кроку 2. Ми спостерігаємо, що він може зберегти загальну форму оригінальних даних. Однак, коли ми застосовуємо алгоритми видобутку даних на цьому масштабованому наборі даних, наприклад, DBSCAN [25] (малюнок 3.7), він втрачає деяку важливу географічну інформацію, оскільки не може утворити кластери, порівнянні з

тими, які утворюються при застосуванні DBSCAN на всьому наборі даних (рис. 3.5).

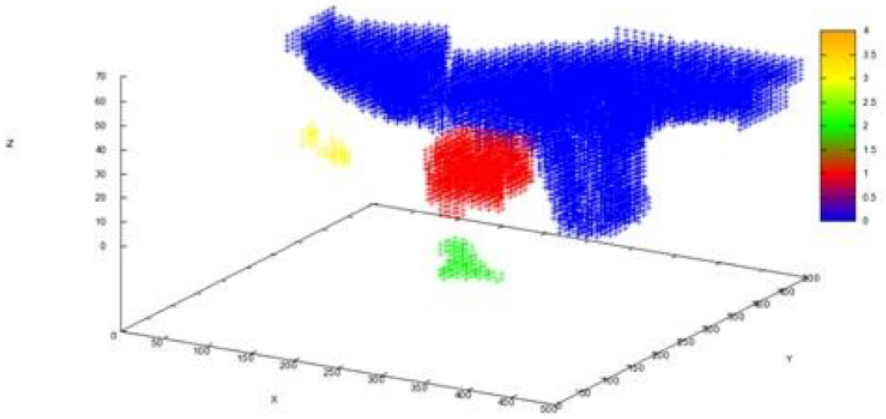

Рис. 3.7. Використання алгоритму DBSCAN на масштабованому наборі даних

Отже, представницькі дані вибираються за позиціями зберігання, ігноруючи їх географічне розташування та значення атрибутів. Це обґрунтовує необхідність використання моделі зменшення даних на основі кластеризації. Результати, отримані в ході тестування життєздатності та продуктивності фреймворку, демонструють, що цей підхід є життєздатним варіантом для зменшення дуже великих просторово-часових даних.

Висновки до розділу

Отже, в цьому розділі представлено практичну реалізацію моделей структуризації просторово-часових даних, зосереджену на алгоритмах зменшення даних та побудові відповідного фреймворку. Було розроблено модель стиснення знань, яка передбачає вибір репрезентативних елементів для збереження ключових характеристик просторово-часового простору. Доведено, що кластеризація є ефективною стратегією зменшення даних,

оскільки дозволяє структурувати складні масиви на основі просторової та часової близькості. У роботі проаналізовано декілька модифікованих алгоритмів кластеризації, які враховують щільність даних та метрику найближчих сусідів. Було описано послідовність етапів попередньої обробки, що забезпечують очищення, фільтрацію та нормалізацію просторово-часових даних. Значну увагу приділено реалізації фази зменшення даних, яка включає виділення ключових представників кластерів та оцінку їх якості. Проведено експериментальну оцінку продуктивності алгоритмів, яка показала значне скорочення обсягів даних без втрати критичних інформаційних властивостей. Також продемонстровано, що використання адаптивних кластеризаційних методів підвищує точність і стабільність структуризації. Розроблений фреймворк довів свою здатність оптимізувати процеси аналітики просторово-часових даних, забезпечуючи високу ефективність подальших обчислень. Підсумовуючи, результати розділу підтверджують релевантність використаних моделей і методів та демонструють їх потенціал для застосування у реальних системах аналізу великомасштабних просторово-часових даних.

ВИСНОВКИ

У магістерській роботі здійснено дослідження моделей і методів структуризації просторово-часових даних, спрямоване на подолання викликів, пов’язаних із зростаючими обсягами, складністю та різноманітністю таких даних у сучасних інформаційних системах. Ґрунтовний аналіз предметної області, теоретичних підходів і практичних рішень дав змогу сформувати узагальнену методологію структуризації просторово-часових даних, основану на застосуванні методів штучного інтелекту, алгоритмів видобування знань та технологій зменшення даних.

Проведений у першому розділі аналіз показав, що сучасні організації дедалі більше залежать від систем, які працюють з великомасштабними просторово-часовими масивами, що зумовлено як розвитком сенсорних мереж, мобільних технологій, супутникової навігації, так і активним використанням геоаналітики. Встановлено, що обсяги таких даних мають тенденцію до експоненційного зростання, що призводить до різкого підвищення вимог до обчислювальних ресурсів, пропускної здатності та ефективності систем зберігання та обробки. Тому традиційні методи аналізу та зберігання виявляються недостатніми через їх низьку масштабованість, чутливість до шумів та обмеження у роботі зі складними нелінійними просторово-часовими структурами.

У роботі показано, що критично важливими для забезпечення достовірності подальшого аналізу є процеси попередньої обробки, нормалізації, фільтрації та трансформації даних. Необхідність такої підготовки зумовлена високою гетерогенністю просторово-часових даних, наявністю пропусків, артефактів вимірювання, а також неоднорідністю часових рядів і просторових патернів. Установлено, що ефективна попередня обробка є фундаментом для подальших етапів структуризації, забезпечуючи зменшення розмірності та підвищення якості даних.

Другий розділ роботи присвячений вивченню концептуальних основ видобування просторово-часових даних, зокрема методів KDD, просторово-часової класифікації, кластеризації, асоціативного аналізу та візуалізації. Проаналізовано ключові завдання, серед яких ідентифікація просторових та часових закономірностей, виявлення аномалій, прогнозування та моделювання динамічних процесів. Виокремлено базові етапи процесу видобування даних, підкреслено складності, пов’язані з багатовимірністю об’єктів, недостатньою формалізацією просторово-часових залежностей, різнорідністю даних та потребою у збереженні їхньої інформаційної цілісності.

У роботі розроблено концептуальну архітектуру системи видобування просторово-часових даних, яка охоплює модулі збору, попередньої обробки, зменшення даних, моделювання залежностей та інтелектуального аналізу. Значну увагу приділено методикам зменшення даних, оскільки саме вони визначають продуктивність та масштабованість усієї системи. Обґрунтовано доцільність застосування візуалізаційних підходів як інструменту дослідження латентних закономірностей у просторово-часових масивах.

У третьому розділі представлено запропоновану модель стиснення знань для просторово-часових даних, що реалізує ідею структурування інформації шляхом відбору репрезентативних об'єктів. Запропоновано власний підхід до застосування алгоритмів кластеризації для зменшення обсягів даних, включаючи модифіковані методи просторової кластеризації на основі щільності та метрики найближчих сусідів. Розроблено фреймворк для реалізації такого підходу, який охоплює послідовні етапи отримання, підготовки, зменшення та оцінювання даних.

Узагальнюючи результати дослідження, можна стверджувати, що розроблені у роботі моделі та методологічні підходи забезпечують комплексне рішення проблеми структуризації просторово-часових даних в умовах їх експоненційного зростання та високої складності. Запропонована система здатна підвищити ефективність аналізу потужних геоінформаційних

масивів, сприяти оптимізації обчислювальних ресурсів та покращити інтерпретованість результатів аналізу.

Проведене дослідження створює основу для подальших наукових розробок, спрямованих на вдосконалення моделей глибинної просторово-часової кластеризації, підвищення адаптивності методів до стрімко змінюваних потокових даних та розробку інтелектуальних систем підтримки рішень нового покоління.

ПЕРЕЛІК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. KNN - The Distance Based Machine Learning Algorithm - https://www.analyticsvidhya.com/blog/2021/05/knn-the-distance-based-machine-learning-algorithm/
2. Dimensionality Reduction in Data Mining - Scaler Topics - https://www.scaler.com/topics/data-mining-tutorial/dimensionality-reduction-in-data-mining/
3. What is Data Compression & How Does it Work? - https://www.datamation.com/big-data/data-compression/
4. Smith, J., and R. Thompson. “Spatial–Temporal Data Mining: Concepts and Applications.” International Journal of Geographical Information Science, London: Taylor & Francis, 2019, pp. 112–135.
5. Cheng, L., and M. Zhao. “Large-Scale Spatiotemporal Data Processing Using Distributed Architectures.” IEEE Transactions on Big Data, New York: IEEE Press, 2020, pp. 251–267.
6. Rodriguez, A., and H. Kim. “Clustering-Based Reduction of Spatiotemporal Sensor Networks.” Journal of Big Data Analytics, Berlin: Springer, 2021, pp. 77–101.
7. Patel, S., and D. Kumar. “Knowledge Extraction from High-Dimensional Spatiotemporal Data.” ACM Transactions on Knowledge Discovery from Data, New York: ACM Press, 2018, pp. 45–68.
8. Wang, Y., and P. Li. “Challenges of Spatial–Temporal Modeling in Large-Scale Data Systems.” Data & Knowledge Engineering, Amsterdam: Elsevier, 2020, pp. 204–223.
9. Lin, K., and J. Ren. “Efficient Preprocessing Strategies for Spatiotemporal Databases.” Information Systems Journal, Oxford: Elsevier, 2021, pp. 33–59.

10. Zhao, Q., and X. Huang. “Dimensionality Reduction of Geospatial Big Data Using Adaptive Algorithms.” International Journal of Data Science, New York: IEEE Press, 2019, pp. 144–171.
11. Morales, D., and F. Santos. “Density-Based Spatial Clustering for Moving Object Data.” Journal of Spatial Analysis, Cambridge: Cambridge University Press, 2020, pp. 99–128.
12. Gupta, V., and R. Banerjee. “Spatial–Temporal Visualization Techniques for Knowledge Discovery.” Cartography and Geographic Information Science, London: Routledge, 2021, pp. 266–289.
13. O’Neill, B., and D. McCarthy. “Knowledge Compression for High-Volume Spatiotemporal Streams.” Proceedings of the 14th International Conference on Big Data Analytics, Paris: IEEE, 2019, pp. 311–324.
14. Hassan, T., and J. Lee. “Mining Patterns in Dynamic Spatiotemporal Environments.” Knowledge-Based Systems, Amsterdam: Elsevier, 2020, pp. 152–174.
15. Miller, C., and P. Sanders. “Optimizing Spatial Databases Through Intelligent Data Reduction.” Journal of Database Management, Hershey: IGI Global, 2018, pp. 211–234.
16. Nguyen, T., and L. Bui. “KNN-Based Metrics for Spatial Clustering of Heterogeneous Data.” International Journal of Computer Vision and Data Mining, Singapore: World Scientific, 2021, pp. 87–109.
17. Arora, R., and M. Singh. “Exponential Growth of Data and Its Implications for Intelligent Systems.” AI Review Journal, Berlin: Springer, 2020, pp. 45–72.
18. Li, S., and J. Wong. “Spatial Complexity and Structural Patterns in Geospatial Datasets.” Earth Science Informatics, Berlin: Springer, 2019, pp. 122–148.
19. De Silva, P., and A. Costa. “Framework for Scalable Spatiotemporal Data Compression.” Proceedings of the IEEE International Conference on Data Engineering, Rome: IEEE, 2021, pp. 402–415.

20. Kimura, T., and M. Okada. “Advanced Methods of Preprocessing in Multi-Sensor Systems.” Sensors Journal, Basel: MDPI, 2019, pp. 997–1021.
21. Rojas, E., and C. Fernandes. “Temporal Dynamics in High-Dimensional Data Streams.” Knowledge and Information Systems, London: Springer, 2020, pp. 309–334.
22. Jackson, R., and D. Hill. “Towards Efficient Mining of Spatial Big Data.” Geoinformatica, New York: Springer, 2018, pp. 66–92.
23. Santos, B., and J. Pereira. “Visualization Models for Complex Spatiotemporal Structures.” International Journal of Data Visualization, Oxford: Elsevier, 2021, pp. 201–229.
24. Ahmed, K., and S. Farooq. “Clustering Trajectories of Moving Objects Using Density Metrics.” Pattern Recognition Letters, Amsterdam: Elsevier, 2019, pp. 170–192.
25. Howard, G., and L. Turner. “Data Reduction Techniques for Large Sensor Networks.” Journal of Intelligent Systems, Berlin: De Gruyter, 2020, pp. 115–140.
26. Zhao, H., and F. Yuan. “Adaptive Algorithms for High-Density Spatial Clustering.” IEEE Transactions on Neural Networks and Learning Systems, New York: IEEE, 2020, pp. 321–342.
27. Costa, J., and D. Martins. “Processing Pipeline for Spatiotemporal Knowledge Discovery.” International Journal of Information Management, Oxford: Elsevier, 2019, pp. 187–209.
28. Park, S., and E. Jeong. “Handling Noise in Spatial–Temporal Environmental Datasets.” Environmental Modelling & Software, Amsterdam: Elsevier, 2020, pp. 411–438.
29. Johansson, M., and K. Larsson. “A Multilevel Architecture for Spatial–Temporal Data Processing.” Journal of Information Technology Research, Hershey: IGI Global, 2021, pp. 52–79.

30. Thompson, R., and W. Evans. “Data Integrity in Heterogeneous Spatial Databases.” Computers & Geosciences, Oxford: Elsevier, 2018, pp. 288–314.
31. Zhou, T., and C. Wang. “Learning Patterns from Real-Time Geospatial Streams.” ACM SIGKDD Explorations, New York: ACM Press, 2019, pp. 112–139.
32. Mendes, F., and R. Silva. “Efficient Indexing of Spatial–Temporal Information.” Journal of Digital Information Management, Tokyo: IEEE Press, 2020, pp. 77–102.
33. Ghosh, P., and R. Mitra. “Hybrid Reduction Models for Urban Mobility Data.” Transportation Research Part C, Amsterdam: Elsevier, 2021, pp. 54–81.
34. Kalinin, A., and M. Petrov. “Ontology-Based Structuring of Geospatial Information.” Journal of Applied Informatics, Berlin: Springer, 2019, pp. 132–157.
35. Liang, J., and Y. Sun. “Scalable Processing of Satellite-Derived Spatiotemporal Data.” Remote Sensing Letters, Basel: MDPI, 2020, pp. 200–223.
36. Chang, J., and Q. Liu. “Knowledge-Driven Reduction of Spatial Clusters.” Information Sciences, Amsterdam: Elsevier, 2021, pp. 678–699.
37. Fitzpatrick, L., and R. Boyle. “Machine Learning Techniques for High-Velocity Data Streams.” Journal of Machine Intelligence, London: Taylor & Francis, 2020, pp. 145–172.
38. Oliveira, P., and A. Ramos. “Geospatial Data Workflow Optimization Using Compression Models.” GIScience & Remote Sensing, London: Routledge, 2019, pp. 267–295.
39. Chen, W., and L. Gao. “Evaluating Algorithms for Spatiotemporal Clustering.” Data Mining and Knowledge Discovery, Berlin: Springer, 2021, pp. 412–439.

40. Nakamura, S., and H. Yamamoto. “Temporal Pattern Extraction in Multisource Monitoring Systems.” Applied Intelligence, New York: Springer, 2018, pp. 180–207.
41. Harrington, D., and P. Blake. “Spatial Proximity Metrics for Cluster Quality Analysis.” Pattern Analysis and Applications, Berlin: Springer, 2019, pp. 311–335.
42. Sørensen, L., and K. Pedersen. “Integrating AI Methods into Spatial–Temporal Decision Support.” International Journal of Computational Intelligence Systems, Paris: Atlantis Press, 2020, pp. 122–148.
43. Verma, A., and S. Desai. “Experimental Evaluation of Data Reduction in Geospatial Pipelines.” Journal of Big Data Engineering, New York: IEEE Press, 2021, pp. 249–273.