ВІДГУК

офіційного опонента – доктора технічних наук, професора,

професора кафедри систем штучного інтелекту

Національного університету «Львівська політехніка»

Виклюка Ярослава Ігоровича

на дисертаційну роботу **Кириченка Євгена Олександровича**

на тему «***Оптимізація структури гетерогенних даних в Big Data***»,

подану на здобуття наукового ступеня доктора філософії

за спеціальністю 121 – «Інженерія програмного забезпечення»

галузі знань 12 – «Інформаційні технології»

Актуальність теми дисертаційного дослідження.

Складність і великі обсяги сучасних гетерогенних даних (табличних даних) призводять до невисокої ефективності класичних методів при оперуванні повною інформацією. Актуальним рішенням є перехід до стислих інтерпретованих форм представлення даних. Це дозволяє проводити порівняльний аналіз та групування об'єктів без надмірних витрат ресурсів, не втрачаючи при цьому значущих властивостей вибірки. Тут слід зауважити, що основна увага також повинна звертатися і на моделі, для яких будуть використовуватися компактні копії даних, що дозволяє зберігати лише необхідні метрики даних.

Незважаючи на розвиток методів аналізу табличних даних, актуальними залишаються проблеми автоматичного розпізнавання типів змінних у масивах невідомого або незадокументованого походження та розробки компактних моделей представлення гетерогенних даних. Окремої уваги потребує створення універсальних метрик для порівняння різнорідних структур, які будуть чутливі до різних типів та враховувати особливості гетерогенних даних. Особливий інтерес становить моделювання близькості даних у вигляді графів, що дозволяє залучити спектральний аналіз для пошуку кластерів та неявних закономірностей у великих корпусах даних.

Сучасний науковий дискурс навколо аналізу гетерогенних даних охоплює широкий спектр методологічних рішень. Значна увага приділяється інтеграції структурованої та неструктурованої інформації за допомогою глибинного навчання для створення узгоджених представлень, які включають в себе узгодження як і з повним набором даних, тка і з моделями, за допомогою яких ці дані будуть аналізуватися. Важливим етапом обробки є формування компактних моделей, які зменшують розмірність даних, зберігаючи їхні основні описові властивості. Також розроблено ефективні методи класифікації змішаних наборів (числових, категоріальних, часових і рядкових), адаптовані для конкретних прикладних задач. Паралельно розвиваються напрями інтелектуального керування обчислювальними ресурсами та моделювання багатовимірних процесів на основі статистичних характеристик. Для порівняння складних структур активно залучаються ансамблеві методи машинного навчання та спеціалізовані статистичні моделі.

Враховуючи зазначені виклики, актуальним науково-прикладним завданням є створення теоретичної бази та відповідної інформаційної технології для комплексного аналізу даних. Усе це зумовлює значну актуальність теми дисертаційної роботи Кириченка Є.О. Це передбачає розробку методів класифікації змінних, генерацію компактних дескрипторів (Compact Data Representation, CDR), впровадження універсальної метрики подібності (Data Information Structure Similarity, DISS) та візуалізацію взаємозв'язків через графи й дерева подібності з подальшою реалізацією системи у масштабованих хмарних інфраструктурах.

Аналіз змісту дисертації та основні результати роботи.

Дисертаційна робота складається з переліку умовних позначень, вступу, чотирьох розділів із підрозділами, висновків, списку використаних джерел (232 найменування) та додатків.

У вступі автором дисертаційного дослідження окреслено основну тематику дисертаційного дослідження, обґрунтовано актуальність дисертаційного дослідження, розглянуто мету, задачі та методи дослідження,

теоретичне та практичне значення одержаних результатів, апробацію одержаних теоретичних результатів.

У **першому розділі** розкрито теоретичний фундамент аналізу великих масивів гетерогенних даних. Автор обґрунтовує доцільність використання спектральних методів та теорії випадкових матриць для ідентифікації їхньої внутрішньої структури. Описано математичний інструментарій стохастичних матриць і спектральні властивості систем великої розмірності, що дозволяє моделювати складні об'єкти без обмежень щодо параметрів розподілу (включаючи дані, взаємозв'язок між якими моделюється на основі розподілів з важкими хвостами). Запропоновано критерій визначення оптимальної кількості кластерів, що ґрунтується на аналізі власних значень стохастичної матриці. Також проаналізовано потенціал спектральних підходів у соціальних, біологічних та технологічних доменах Big Data, де дані мають виражену стохастичну природу.

Другий розділ присвячений розробці математичного та алгоритмічного забезпечення для компактного подання, класифікації та порівняльного аналізу великих гетерогенних масивів даних. На основі сучасних концепцій Big Data запропоновано модель CDR, яка дозволяє оперувати стислими представленнями замість повних копій даних, зберігаючи їхні ключові статистичні та структурні властивості. Важливим елементом є розроблений алгоритм автоматичної типізації, що ідентифікує числові, категоріальні, часові та текстові атрибути без апріорних відомостей про схему даних. У розділі також систематизовано та обґрунтовано вибір метрик для різних типів змінних.

У **третьому розділі** проаналізовано принципи проєктування масштабованих систем у хмарних середовищах та специфіку розподіленої обробки даних. Обґрунтовано перехід до багаторівневої модульної архітектури, яка базується на принципах низької зв'язності та чіткого розподілу відповідальності між компонентами. Технологічну незалежність системи від конкретних хмарних провайдерів забезпечено використанням Python, Apache Spark для паралельних обчислень та Apache Airflow як інструменту оркестрації.

Розділ також містить детальний опис конвеєра трансформації неструктурованих масивів у валідні представлення.

Четвертий розділ присвячено програмній реалізації та апробації розробленої інформаційної системи, в основу якої покладено метод компактного представлення CDR та метрики DISS. Експериментальна перевірка на реальних фінансових часових рядах підтвердила стійкість і універсальність запропонованого підходу до варіацій ринкових режимів та джерел даних. Результати тестування демонструють високу точність виявлення структурних змін і збалансованість класифікації. Архітектурні особливості системи забезпечують її легку інтеграцію в хмарні конвеєри обробки Big Data та платформи машинного навчання.

Роботу завершують **висновки**, що підтверджують наукову новизну та практичну цінність отриманих результатів.

Додатки містять наукові праці з ключовими результатами дослідження, документальні підтвердження їх практичної апробації (акти та довідки про впровадження), а також фрагменти програмного коду розробленого програмного забезпечення.

Наукова новизна, оцінка обґрунтованості наукових положень дисертаційного дослідження та їх достовірності.

Дисертаційна робота містить ряд нових цікавих теоретичних результатів, розробку інформаційної системи та аналіз реальних даних на основі розроблених методів та підходів. Зокрема, **вперше** ***розроблено інтегровану модель CDR+DISS***, яка завдяки поєднанню класифікації змінних, методів компактного; ***запропоновано метод генерації дерев подібностей***, що базується на алгоритмах мінімального остовного дерева та метриці DISS; ***створено багатоетапну хмарно-орієнтовану інформаційну технологію*** для аналізу табличних даних. Рішення реалізовано на базі AWS-сервісів, що забезпечує масштабованість процесів класифікації та кластеризації. Також, **дістали подальшого розвитку** ***теорія структурної подібності*** табличних даних, що базується на узагальненні статистичних відстаней для різнотипних змінних;

графово-спектральні підходи до кластеризації, де застосування компактних дескрипторів для опису подібності та аналіз спектра відповідних матриць дозволяють ідентифікувати кластерну структуру даних.

Дисертаційна робота Кириченка Є.О. має чітку та логічну структуру і є цілісним та завершеним науковим дослідженням.

Зв'язок роботи з науковими програмами, планами та темами.

Дисертаційну роботу виконано в межах науково-дослідних робіт Чернівецького національного університету імені Юрія Федьковича, зокрема згідно з планами ***кафедри програмного забезпечення комп'ютерних систем*** за держбюджетною темою «Дослідження, моделювання та розробка програмного забезпечення складних динамічних систем» (номер державної реєстрації 0121U109232); ***кафедри математичних проблем управління і кібернетики*** за держбюджетною темою «Інформаційні технології в аспекті сучасних задач прийняття рішень» (номер державної реєстрації 0121U109159).

Теоретичне та практичне значення результатів. Робота має виражений практичний характер, хоча містить ряд теоретичних тверджень. Основні положення та наукові (теоретичні) результати дисертаційного дослідження впроваджені в навчальний процес Чернівецького національного університету імені Юрія Федьковича, а практичні результати – в діяльність компаній ТОВ «Кодерс ПРО» та ТОВ «Палетний сервіс», про що свідчать акти впрвадження.

Повнота викладу результатів дисертації в опублікованих працях.

Основні положення та результати дисертаційної роботи висвітлено у **13 наукових працях**. У публікаціях розкрито теоретичні аспекти, запропоновані алгоритмічні рішення та особливості технологічної реалізації системи класифікації й аналізу подібності табличних даних, серед яких **5 статей** у рецензованих фахових виданнях (зокрема **2** – у журналах, що індексуються наукометричною базою **Scopus**, та **3** – у провідних фахових виданнях України), **1** – праця, яка додатково відображає наукові результати дисертації, а також **6 тез** доповідей у матеріалах міжнародних науково-практичних конференцій.

Дискусійні положення та зауваження до змісту дисертаційного дослідження.

1. Поряд із добре розробленим математичним апаратом, математична модель містить ряд невідомих параметрів (ваг w) та неоднозначність вибору метрик при використанні DISS. Слід більш детально звернути увагу саме на підбір даних параметрів.

2. У роботі зазначається про широкий спектр проведених експериментальних досліджень, але основна увага приділена опису експериментів над фінансовими даними.

3. Теоретичні результати зосереджені на розробці методів кластеризації графів за допомогою спектральних методів для випадку важких хвостів елементів матриці суміжності графу для $\alpha \in (2,3]$, тобто за умови існування скінченного математичного сподівання. Оскільки відповідна стохастична матриця будується на основі нормування рядків, то варта провести моделювання відповідних графів та дослідити валідність одержаних оцінок для $\alpha \in (1,2]$.

Однак, зазначені недоліки не є принциповими, не впливають на загальну позитивну оцінку дисертаційної роботи, не змінюють її наукової новизни та практичної цінності.

Загальний висновок.

Оцінюючи дисертаційну роботу в цілому, є всі підстави стверджувати, що за актуальністю теми, обсягом виконаних досліджень і науковою новизною та цінністю одержаних в ній результатів і науково-теоретичним рівнем їх обґрунтованості дисертаційна робота Євгена Олександровича Кириченка на тему «Оптимізація структури гетерогенних даних в Big Data» є завершеним науковим дослідженням, що вносить значний внесок у розвиток методів кластеризації графів та побудови репрезентативних копій табличних (гетерогенних) даних.

Дисертація є завершеною науковою працею, яка цілком відповідає вимогам пунктів 6, 7, 8, 9 «Порядку присудження ступеня доктора філософії та

скасування рішення разової спеціалізованої вченої ради закладу вищої освіти, наукової установи про присудження ступеня доктора філософії», затвердженого Постановою Кабінету Міністрів України від 12 січня 2022 р. №44 (зі змінами, внесеними згідно з Постановами Кабінету Міністрів України № 341 від 21.03.2022, № 502 від 19.05.2023, № 507 від 03.05.2024), а її автор Кириченко Євген Олександрович заслуговує на присудження ступеня доктора філософії за спеціальністю 121 – «Інженерія програмного забезпечення» в галузі знань 12 – «Інформаційні технології».

Офіційний опонент:

доктор технічних наук, професор,

професор кафедри системи штучного інтелекту Ярослав ВИКЛЮК

Національного університету «Львівська політехніка»

Підпис засвідчую.

Вчений секретар Р. Брилинський