ВІДГУК

офіційного опонента на дисертаційну роботу
Павлишенка Богдана Михайловича
на тему: «Методи інтелектуального аналізу консолідованих даних
для підтримки прийняття рішень»,
поданої на здобуття наукового ступеня доктора технічних наук
за спеціальністю 05.13.23 – системи та засоби штучного інтелекту

Актуальність теми досліджень

Задача обробки та аналізу даних при побудові моделей складних об'єктів та процесів на цей час є дуже актуальною. У сучасній інформаційній епосі існує багато різноманітних джерел даних різної структури, які містять кількісні та якісні величини різноманітних ознак. Актуальною проблемою обробки та аналізу є об'єднання усіх даних отриманих з різних джерел, дотичних до аналізованої задачі, в єдиній аналітичній моделі. Складність полягає у тому, що різні процеси характеризуються даними з різною структурою. Виявлення та формування ефективних аналітичних ознак цих процесів є різним у різних предметних областях часто базується на експертному досвіді. Важливим етапом в аналізі даних є їхня консолідація, під якою розуміють об'єднання масивів даних із різних джерел та з різною структурою для вирішення певної аналітичної проблеми. Актуальним є створення узагальнених моделей та методів у аналізі даних, виявлення та створення аналітичних ознак даних та їхнього узагальнення для підтримки прийняття рішень у заданому класі проблем. На різних етапах такого аналізу стає очевидною доцільність розроблення нових методів та підходів, шляхом поєднання наявних методів та алгоритмів. Аналізованим процесам властива деяка міра невизначеності, тому важливо враховувати та аналізувати невизначеність факторів впливу та цільової змінної для того, щоб оцінити ризики, пов'язані з неточністю прогнозування. Прикладом слабоструктурованих типів даних можуть бути текстові масиви. Сучасний аналіз текстової інформації вимагає розвитку нових ефективних методів семантичного аналізу із заглибленням у зміст інформації, використовуючи методи машинного навчання. Підбір, об'єднання прогнозних моделей та формування аналітичних ознак є об'єднаною комплексною проблемою інтелектуального аналізу, розв'язок якої базується як на сучасних методах аналізу даних, так і на знаннях у предметній області, до якої належать аналізовані процеси. Виникає потреба в удосконаленні наявних та розробці нових методів та підходів

інтелектуального аналізу для підтримки прийняття рішень з урахуванням особливостей структури даних та предметної області.

У дисертаційній роботі розглянута актуальна науково-прикладна проблема розроблення, вибору, поєднання та оптимізації моделей та методів інтелектуального аналізу різнотипних консолідованих даних з метою підвищення інформативності, точності та достовірності результатів для підтримки прийняття рішень в інформаційно-аналітичних системах.

Достовірність отриманих результатів. Достовірність викладених в дисертаційній роботі наукових положень, результатів і висновків, зроблених здобувачем, підтверджується даними, що були отримані при розв'язанні практичних завдань, впровадженням результатів роботи, а також апробацією на міжнародних наукових конференціях.

Наукова новизна результатів дисертації. Аналіз дисертаційної роботи дозволяє зробити висновок, що автором у процесі досліджень отримані такі наукові результати:

Вперше:

- Розроблено метод оптимізації прогнозної аналітики часових рядів з використанням стекінгового об'єднання та відбору різнотипних моделей на основі лінійної регресії LASSO та байєсівської регресії, що забезпечує підвищення точності прогнозування та формування оптимального прогнозного ансамблю моделей.
- Розроблено метод виявлення технічних відмов, який, за рахунок поєднання байєсівської, лінійної та машино-навчальної логістичних регресій, забезпечує підвищення точності та достовірності результатів, що дозволяє побудувати ефективні диверсифіковані процеси прийняття рішень.
- Розроблено метод векторного представлення текстових даних, який, за рахунок використання теорії семантичних та тематичних полів, дозволяє представляти текстові документи у низькорозмірному просторі семантичних ознак та забезпечує зменшення складності розрахунків і підвищення достовірності результатів в аналізі текстових даних.
- Розроблено метод аналізу текстових даних на основі алгоритмів машинного навчання з використанням кількісних ознак семантичних і тематичних полів,

а також метод генетичної оптимізації набору цих ознак, що забезпечує підвищення достовірності результатів інтелектуального аналізу текстових масивів.

- Розроблено метод виявлення додаткових аналітичних ознак на основі лексемних поєднань у семантичних структурах текстових масивів, який, за рахунок використання теорії частих множин та асоціативних правил, розширює інформаційну основу для підтримки прийняття рішень в аналітиці консолідованих даних.
- Розроблено модель семантичних концептів текстових масивів на основі теорії формальних концептів, що дозволяє виявляти ефективні аналітичні ознаки з урахуванням семантичної структури текстових масивів.

Отримали подальший розвиток :

- Методи оптимізації послідовності дій інтелектуального агента в задачах аналітики попиту з використанням глибокого Q-навчання та імітаційного моделювання середовища взаємодії на основі параметричної моделі та з використанням історичних даних, що забезпечує підвищення ефективності прийняття бізнес рішень.

Удосконалено:

- Метод класифікаційного та регресійного аналізу різнотипних консолідованих даних на основі поєднання LSTM нейромережі з вхідними текстовими даними та нейромережі з повністю з'єднаними шарами з вхідними кількісними ознаками, що забезпечує підвищення точності та достовірності результатів.

Практичне значення отриманих результатів полягає у тому, що запропоновані методи методи є складовою технологією для підтримки прийняття рішень у комплексних інформаційних системах і забезпечують підвищення інформативності та надійності інтелектуального аналізу даних у прогнозній аналітиці різнотипних консолідованих даних.

Одержані результати дають можливість: підвищити точність прогнозування та зменшити кількість моделей у стекінговому ансамблі на 30% для певного класу задач за рахунок розроблених методів стекінгового об'єднання різнотипних моделей у прогнозні ансамблі; оцінити невизначеність та прогнозні ризики складових моделей при прийнятті експертних рішень щодо формування

прогнозного ансамблю моделей за рахунок розробленого методу використання байєсівської регресії для стекінгу прогнозних моделей; підвищити точність та інформативність результатів у задачах аналізу динаміки попиту та в аналітиці фінансових часових рядів за рахунок розроблених методів застосування лінійних, ймовірнісних та машинно-навчальних прогнозних моделей з урахуванням аналітичних ознак консолідованих даних заданої предметної області інтелектуального аналізу; оптимізувати набір прогнозних ознак та підвищити точність прогнозування за рахунок розроблених методів у прогнозуванні технічних відмов на лініях збірки на виробництві з використанням стекінгового об'єднання моделей; зменшити кількість аналітичних семантичних ознак текстових даних у 3-10 разів у порівнянні з набором лексемних частотних ознак для заданих характеристик інтелектуального аналізу текстових даних за рахунок розроблених методів використання теорії семантичних та тематичних полів; Отримані у роботі результати використовуються у компанії **8ой8єгує** Іпс. для розробки програмного забезпечення у задачах аналізу даних, а також впроваджені у відповідні навчальні курси у Львівському національному університеті імені Івана Франка.

Практичне значення одержаних результатів. Отримані результати були досліджені експериментально на тестових і реальних даних, де довели свою перевагу над відомими методами. Запропоновані в роботі методи високопродуктивного оброблення даних можуть бути використані в різних областях таких як біоінформатика, імуноінформатика та інші, де дані представлені у чисельному вигляді. Запропоновані методи довели свою ефективність при розв'язанні практичних задач. Усі впровадження підтверджено відповідними актами. Одержані у дисертаційному дослідженні результати та розроблені методи є складовою технологією для підтримки прийняття рішень у комплексних інформаційних системах і забезпечують підвищення інформативності та надійності інтелектуального аналізу даних у прогнозній аналітиці різнотипних консолідованих даних.

Структура та обсяг дисертаційної роботи. Дисертаційна робота складається зі вступу, шести розділів, висновків, списку літератури з 361 джерела та додатків, загальним обсягом 407 сторінки друкованого тексту, з яких 314 . сторінок основного тексту.

Зміст роботи

У першому розділі наведено літературний огляд основних моделей, методів та підходів, які використовуються в інтелектуальному аналізі даних. Розглянуто методи аналізу даних табличного та текстового типів. Наведено основні положення лексемної семантики та теорії семантичних полів. На основі наведених літературних даних зроблено висновки та сформульовано невирішені питання.

У другому розділі розглянуто моделювання, формування ознак та інтелектуальний аналіз даних табличного типу. Розроблено комплексний підхід у прогнозній аналітиці табличних даних на основі параметричних та машинно-навчальних моделей, який дає змогу утворювати оптимальний набір аналітичних ознак та формувати ефективний підхід у побудові прогнозних моделей. Розглянуто об'єднання моделей різних типів у прогнозний ансамбль на основі стекінгового підходу. Розглянуто використанням LASSO регресії як стекінгової моделі другого рівня. У роботі проведено дослідження застосування байєсівської регресії. Досліджено реалізацію стекінгу різних прогнозних моделей за допомогою байєсівської регресії, що дозволяє отримати розподіли для регресійних коефіцієнтів моделей першого рівня прогнозного ансамблю і оцінити невизначеність, внесену кожною моделлю в результат стекінгу. Розглянуто використання моделей глибокого Q-навчання у задачах часових рядів продажів.

У третьому розділі розглянуто концепції семантичних та тематичних лексикографічних полів із точки зору їхнього використання в алгоритмах інтелектуального аналізу текстових масивів. На основі концепцій семантичних полів створено теоретико-множинну модель, яка об'єднує поняття семантичного та тематичного лексемних. Розглянуто векторну модель текстових документів у семантичному просторі, базис якого утворено частотно-дистрибутивними характеристиками семантичних та тематичних полів.

У четвертому розділі проаналізовано використання концепції семантичних полів в аналітиці текстових даних на основі методів машинного навчання. Розглянуто текстові вибірки різних типів, зокрема, масив авторських текстів англомовної художньої прози, повідомлення груп новин та текстові повідомлення соціальної мережі Твіттер. Як семантичні ознаки, розглянуто частотні характеристики семантичних та тематичних полів, а також компонент тематик латентного розміщення Діріхле. Розроблено метод кластеризації текстових документів у семантичному просторі. Розроблено метод класифікації

текстових даних за експертно сформованими семантичними ознаками, які відображають семантичну сторону аналізу предметної області. Розроблено метод використання семантичних ознак у комбінованих нейронних мережах із використанням рекурентних підмереж для текстових даних та підмереж. Розглянуто використання генетичних алгоритмів для оптимізації набору семантичних полів, що дозволяє формувати ефективні низькорозмірні простори семантичних ознак у задачах інтелектуального аналізу текстових даних. Розглянуто квантовий алгоритм пошуку ключових семантичних образів у масивах текстових об'єктів.

У п'ятому розділі розглянуто використання теорії частих множин та асоціативних правил в аналітиці текстових повідомлень соціальних мереж, зокрема Твіттера, що дає можливість сформувати тематичне семантичне поле, яке в подальшому можна використовувати для пошуку асоціативних правил.

Розглянуто використання теорії графів для аналізу повідомлень мережі Твіттер, зокрема, для аналізу зв'язків між користувачами та виявлення різних спільнот. Використовуючи алгоритми аналізу графів, а також теорію частих множин та асоціативних правил, проведено інтелектуальний аналіз повідомлень мережі Твіттер, пов'язаних з пандемією COVID-19.

У шостому розділі на основі теорії аналізу формальних концептів запропоновано модель семантичного контексту, яка відображає структурну семантичну організацію текстових масивів. Розроблено метод використання моделі семантичного контексту в аналітиці текстових повідомлень соціальних мереж. Запропоновано застосування теорії аналізу формальних концептів в інтелектуальній обробці повідомлень Твіттера.

Повнота відображення результатів дисертації у публікаціях. За результатами досліджень опубліковано 52 наукові праці, серед яких 30 статей у наукових фахових журналах і 22 публікації у матеріалах конференцій. Серед публікацій 7 статей опубліковано у наукових журналах зі списку Scopus, а також 5 статей опубліковано у матеріалах конференцій, які реферуються у Scopus. Усі наукові результати, які виносяться на захист дисертаційної роботи, отримані автором самостійно. Усі наукові праці опубліковано одноосібно.

Автореферат та дисертація за змістом, структурою та обсягом відповідають чинним вимогам, викладені логічно послідовно, коректно та грамотно. Зміст автореферату відповідає змісту дисертації.

Відповідність змісту автореферату основним положенням дисертації. Оформлення автореферату за своїм обсягом, структурою та змістом відповідає чинним вимогам. Зміст автореферату ідентичний змісту основних положень дисертації, автореферат адекватно відображає результати дисертації.

Відповідність дисертації встановленим вимогам. Дисертаційна робота є завершеним і цілісним дослідженням, її матеріал є досить добре структурованим і логічно викладеним. Роботу написано коректно з використанням сучасної науково-технічної термінології.

Оформлення дисертації відповідає встановленим вимогам до докторських дисертацій згідно «Порядку присудження наукових ступенів» (Постанова КМУ №567 від 24 липня 2013 р.), а також вимогам МОН України до дисертацій на здобуття наукового ступеня доктора технічних наук.

Зауваження по дисертаційній роботі. Серед недоліків дисертаційної роботи слід зазначити такі:

1. У першому розділі автор наводить сім "основних" на його погляд методів інтелектуального аналізу табличних даних. Однак він не аргументує чому саме ці та на підставі яких критеріїв він вибрав саме ці "основні" методи.
2. При прогнозуванні часових рядів автором проведено аналіз лише точності отриманої моделі і оцінка параметрів моделі. Однак було б доцільно хоча б щонайменше додатково оцінити адекватність отриманої моделі хоча б за допомогою перевірки незалежності значень ряду залишків за допомогою тесту Дарбіна-Уотсона або аналіз залишків за допомогою критерію Колмогорова-Смирнова, а також F-критерію.
3. У розділі 1.2.7. дається узагальнений опис генетичного алгоритму. Проте здобувач дуже поверхневій формі описує даний алгоритм. З тексту не ясно переваги і недоліки обраного алгоритму.
4. У розділах 2.5. опис методів інтелектуального агентів даних з використанням Q-навчання, однак здобувач не описує запропонованого методу.
5. В розділі 2.4. дається опис роботи методів X_Boost, байєсовської та лінійної регресії, для побудови ансамблю, було б доцільно обґрунтувати чому саме ці лінійні моделі обрані для розв'язання даної задачі при використанні ансамблевого підходу.
6. При кластерному аналізі текстових документів, було б доцільним провести оцінку якості отриманих результатів кластеризації.

7. Доцільно було б провести більш детальніший аналіз різних методів класифікації текстових документів, зокрема методів на основі метода опорних векторів та випадкового лісу.

8. У роботі запропоновано метод формування та використання ознак для прогнозування. Доцільно було би дослідити та проаналізувати використання даних ознак у алгоритмах машинного навчання.

9. При використанні ансамблевих моделей, було би доцільним провести порівняльні дослідження результатів використання беггінгу, бустінгу та стекінгу

10. Цікавими є запропоновані методи прогнозування подій з використанням аналітичних ознак на основі частих множин та асоціативних правил у повідомленнях Твіттера. Доцільним було б більш детальніше розглянути використання таких ознак у алгоритмах машинного навчання.

Вказані зауваження та наведені недоліки не впливають на загальну високу оцінку одержаних науково-прикладних результатів.

Загальний висновок

Дисертаційна робота Павлишенка Богдана Михайловича, представлена на здобуття наукового ступеня доктора технічних наук за спеціальністю 05.13.23 – системи та засоби штучного інтелекту, є завершеним науковим дослідженням, у якому вирішено актуальну науково-прикладну проблему вибору, поєднання та оптимізації методів інтелектуального аналізу консолідованих даних шляхом розроблення методів моделювання, формування інформативних аналітичних ознак та інтелектуального аналізу табличних та текстових даних з урахуванням предметної області аналізу, що дозволило створювати ефективні прогнозні багаторівневі моделі, розширити інформативність інтелектуального аналізу різнотипних даних та вдосконалити підтримку прийняття рішень у комплексних інформаційно-аналітичних системах. Дисертаційна робота відповідає паспорту спеціальності 05.13.23 – системи та засоби штучного.

Вважаю, що за новизною, актуальністтю, обсягом, науковим рівнем та практичним значенням отриманих результатів дисертація відповідає вимогам пунктів 9, 10, 12-14 «Порядку присудження наукових ступенів», затвердженого

постановою Кабінету Міністрів України № 567 від 24.07.2013, а її автор, Павлишенко Богдан Михайлович, заслуговує присудження йому наукового ступеня доктора технічних наук за спеціальністю 05.13.23 – системи та засоби штучного інтелекту.

Офіційний опонент:
завідувач кафедри інформатики та
комп'ютерних наук
Херсонського національного
технічного університету, доктор
технічних наук, професор В. І. Литвиненко

Підпис Литвиненка
Володимира Івановича
засвідчую.
Начальник відділу кадрів
ХНТУ М.В. Танська

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХЕРСОНСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
05480298