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

Інтелектуальний аналіз даних широко використовується у сучасних інформаційних системах. Дані відображають різноманітні явища та процеси у бізнесі, суспільстві, соціальних мережах, технічних пристроях тощо. У сучасній інформаційній епосі існує багато різноманітних джерел даних різної структури, які містять кількісні та якісні величини різноманітних ознак. Актуальним є об'єднання усіх даних, дотичних до аналізованої задачі, в єдиній аналітичній моделі. Складність полягає у тому, що різні процеси характеризуються даними з різною структурою, наприклад, частина даних може мати табличну структуру, а частина – текстову. Виявлення та формування ефективних аналітичних ознак цих процесів є різним у різних предметних областях і в основному 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. Це узагальнений етап аналізу даних, який може відрізнятись у різних предметних областях. Актуальним є створення узагальнених моделей та методів у аналізі даних, консолідації досліджуваних даних різних типів із різних джерел та різних предметних областей, виявлення та створення аналітичних ознак даних та їхнього узагальнення для підтримки прийняття рішень у заданому класі проблем. Розроблення ефективних методів інтелектуального аналізу різноструктурованих консолідованих даних із використанням

алгоритмічних моделей можливе шляхом аналізу різнотипних задач із різних предметних областей. На різних етапах такого аналізу стає очевидною доцільність розроблення нових методів та підходів, зокрема, 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. Стимулом розвитку методів інтелектуального аналізу текстів є значний ріст слабоструктурованої інформації текстового типу, зокрема, у мережі Інтернет. Сучасний аналіз текстової інформації поряд із традиційними статистичними методами вимагає розвитку нових ефективних методів семантичного аналізу із заглибленням у зміст інформації, використовуючи методи машинного навчання. Практика інтелектуального аналізу показує, що сучасні бізнес процеси настільки складні, що важко виробити єдиний для всіх задач підхід у прогнозній аналітиці. 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. Актуальним є розгляд типових задач такого аналізу з різних предметних областей та узагальнення методів і алгоритмів розв'язку прикладних задач, беручи до уваги особливості заданої предметної області знань.

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 наукових результатів забезпечується використанням сучасного математичного апарату, а саме: теорії та алгоритмів машинного та глибокого навчання для створення прогнозних моделей та їх ансамблів; теорії машинного навчання з підкріпленням для побудови моделей інтелектуальних агентів в алгоритмах оптимізації послідовності прийняття рішень; теорії ймовірності та математична статистика для формування частотних семантичних характеристик текстових лексем та для створення ймовірнісних прогнозних моделей інтелектуального аналізу даних; теорії множин для створення теоретико-множинних моделей семантичних та тематичних полів; теорії частих множин та асоціативних правил і теорія аналізу формальних концептів для розробки підходів в аналітиці текстових потоків даних. Достовірність наукових результатів та висновків дисертаційної роботи підтверджується отриманими експериментальними результатами та апробацією результатів роботи на різних науковоих конференціях.

Наукова новизна одержаних результатів

Унаслідок проведених теоретичних та експериментальних досліджень отримано такі нові результати:

1. 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.

2. Вперше 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.

3. 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.

4. 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.

5. 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.

6. розроблено модель семантичних концептів текстових масивів на основі теорії формальних концептів, що дозволяє виявляти ефективні аналітичні ознаки з урахуванням семантичної структури текстових масивів.

7. 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.

8. 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 одержаних результатів

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. 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; кількісно аналізувати семантичну складову авторського ідіолекта в текстових масивах за рахунок розробленого методу аналізу текстів із використанням теорії семантичних та тематичних полів; сформувати додаткові семантичні ознаки для прогнозних моделей та підвищити якість інформаційно-аналітичних систем за рахунок розроблених методів інтелектуального аналізу текстових потоків соціальної мережі Твіттер з використанням теорії частих множин і асоціативних правил та теорії формальних концептів. Отримані у роботі результати використовуються у компанії SoftServe Inc. для розробки програмного забезпечення у задачах аналізу даних, а також впроваджені у відповідні навчальні курси у Львівському національному університеті імені Івана Франка.

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, яка дає можливість оцінити невизначеність складових факторів аналізу і прогнозні ризики. Досліджено реалізацію стекінгу різних прогнозних моделей за допомогою байєсівської регресії, яку було використано на другому стекінговому рівні, 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 полів і дає можливість представляти текстові дані у просторі семантичних ознак з метою інтелектуального аналізу заданого семантичного спектру текстових даних. 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, який дає можливість отримувати новий структурний поділ документів за семантичними ознаками. Розроблено метод класифікації текстових даних за експертно сформованими семантичними ознаками, зокрема, квантитативними ознаками семантичних та тематичних полів, що дозволяє проводити інтелектуальний аналіз текстових масивів із відповідними семантичними акцентами, які відображають семантичну сторону предметної області аналізу. Розроблено метод використання семантичних ознак у комбінованих нейромережах із використанням рекурентних підмереж для текстових даних та підмереж із повністю з'єднаними шарами для кількісних ознак, що диверсифікує простір прогнозних ознак в алгоритмах глибокого навчання та покращує якість інтелектуального аналізу консолідованих даних. Розглянуто використання генетичних алгоритмів для оптимізації набору семантичних полів, які утворюють векторний простір документів в алгоритмах інтелектуального аналізу текстових даних, 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. Показано, що реалізація квантових алгоритмів аналізу семантичних образів текстових об’єктів для певного класу задач дає можливість поліноміально зменшити час виконання алгоритму у порівнянні з класичними алгоритмами внаслідок реалізації квантового паралелізму.

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. На основі відібраних частих множин семантичних ознак можна побудувати асоціативні правила, які будуть відображати семантичні зв’язки змісту повідомлень мікроблогів.

Розглянуто використання теорії графів для аналізу повідомлень мережі Твіттер, зокрема, для аналізу зв'язків між користувачами та виявлення різних спільнот.

Показано, що у потоках твітів, у яких обговорюються очікувані події, можна виявити ознаки на основі частих множин, які мають прогнозний отенціал стосовно цих подій.

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. У семантичному контексті формується частково впорядкована множина семантичних концептів, формальний зміст яких визначається семантичними полями, а формальний об’єм - текстовими документами. Розроблено метод використання моделі семантичного контексту в аналітиці текстових повідомлень соціальних мереж. Побудова ґратки семантичних концептів дає можливість описувати ієрархічну семантичну структуру в масиві документів та виявляти групи текстових документів, які об’єднані спільною групою семантичних ознак. 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.

Зауваження

Як зауваження та недоліки роботи можна зазначити такі:

1. Опис формування ознак для деяких наведених прикладів інтелектуального аналізу табличних даних наведено дуже стисло, варто було б розглянути формування прогнозних ознак більш детальніше.
2. У роботі не достатньо повно розкрито виникнення ефекту перенавчання в алгоритмах машинного навчання та методах уникнення такого ефекту.
3. LSTM нейромережі часто використовують в аналізі часових рядів, однак цей напрям не розглянуто у роботі.
4. Розгляд впливу пандемії COVID-19 розглянуто на короткому періоді часу, варто було б розширити дослідження із врахуванням механізмів поширення COVID-19 та зробити порівняльний аналіз для різних проміжків часу.
5. У роботі досліджено невизначеність прогнозування у моделях машинного навчання за допомогою байєсівської регресії, доцільно

було б розглянути методи оцінки невизначеності результатів прогнозування для нейромереж.

6. Опис використання ознак, які базуються на текстових повідомленнях Твіттера для аналізу оптимальної трейдингової стратегії на основі Q-навчання із підкріпленням бажано було б зробити більш детальнішим.
7. У методах кластеризації текстових документів розглянуто два методи кластеризації – агломеративна та к-середніх. Доцільно було б розглянути більшу кількість методів та порівняти їхні результати.
8. у роботі розглянуто теоретичні підходи у використанні квантових алгоритмів в аналізі слабоструктурованих даних, однак мало уваги приділено практичному використанню таких алгоритмів.

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 інтелекту.

Вважаю, що за новизною, актуальністю, 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 спеціальністтю 05.13.23 – системи та засоби штучного інтелекту.

Офіційний опонент:

професор кафедри кібербезпеки
Університету банківської справи
доктор технічних наук, професор **Д.Д. Пелешко**

Підпис Пелешка Д.Д. засвідчую
/
Ректор
Університету банківської справи
Доктор економічних наук, професор **А.Я.Кузнєцова**