ВІДГУК

офіційного опонента – доктора фізико-математичних наук, професора, завідувача відділу математичних методів дослідження операцій № 130 інституту кібернетики ім. В.М. Глушкова НАН України Кнопова Павла Соломоновича на кваліфікаційне дослідження «Оцінки параметрів авторегресійних моделей» Книгніцької Тетяни Василівни, яка здобуває науковий ступінь доктора філософії у галузі знань 11 «Математика та статистика» за спеціальністю 113 «Прикладна математика».

Актуальність теми виконаної роботи.

Основна увага дисертаційного дослідження Книгніцької Т.В. присвячена дослідженню часових рядів, встановленню подібності між часовими рядами та визначенню оптимальної кількості кластерів для даних, які представлені у вигляді графів. Часові ряди та графи є представниками неструктурованих типів даних. Будь-який процес життєдіяльності людини у наш час характеризується акумулюванням великих обсягів даних, правильне опрацювання яких допомагає знаходити відповіді на багато питань з управління бізнесом, прогнозування економічних показників, поширення вірусів, діагностики захворювань, визначення шахрайських банківських операцій тощо. Усе це стало можливим завдяки впровадження машинного навчання у технології, якими користується людина щодня. Одним із напрямків машинного навчання без учителя, який активно обговорюється та розвивається у науковій літературі є кластеризація часових рядів, яка і є основним об'єктом дисертаційного дослідження.

Мета кластеризації полягає в тому, щоб визначити структуру в немаркованому наборі даних шляхом об'єктивної організації даних в однорідні групи, де подібність між об'єктами всередині групи зведена до мінімуму, а відмінність між об'єктами різних груп – максимальна. Кластеризація необхідна, коли недоступні дані з мітками, незалежно від того, чи є дані двійковими, категоріальними, числовими, інтервальними, порядковими, реляційними, текстовими, просторовими, часовими, просторово-часовими, зображеннями, мультимедійними або сумішами вищевказаних типів даних. Для здійснення кластеризації часових рядів необхідно, щоб вхідні дані були стаціонарними. Тобто, щоб математичне сподівання та автокореляційна функція часового ряду не залежали від часу. На практиці, не всі вхідні дані задовольняють умову стаціонарності. Тому існують підходи до приведення часового ряду до стаціонарного типу. У наш час усі існуючі підходи до кластеризації даних можна умовно поділити на п'ять основних категорій: методи поділу, ієрархічні методи, методи на основі щільності, методи на основі сітки та методи на основі моделей. У кваліфікаційному дослідженні Книгніцької Тетяни Василівни розглянуто метод кластеризації даних, який базується на моделях часових рядів. У такому випадку для кластеризації даних використовуються параметри моделей часових рядів, а не самі дані. Даний напрямок наукового дослідження є дуже актуальним зараз, так як дані з будь-якої сфери життєдіяльності людини зберігаються у формі часових рядів. Наприклад, метеорологічні дані, дані про продажі товарів та послуг, біомедичні вимірювання (артеріальний тиск, електрокардіограма, тощо), ціни на акції, біометричні дані тощо. Відповідно, наукові дослідження та застосування кластеризації даних, які представлені часовими рядами або графами, можна знайти в різних сферах, таких як енергетика, фінанси, біоінформатика чи біологія. Багато досліджень, пов'язаних з аналізом часових рядів, використовується в різних областях для різних цілей, таких як: зіставлення підпослідовностей, виявлення аномалій, індексування, візуалізація, сегментація, ідентифікація закономірностей, аналіз тенденцій, узагальнення та прогнозування. Тому актуальність тематики дисертаційного дослідження Книгніцької Т.В. важко переоцінити.

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

Напрям дослідження відповідає програмі наукової тематики кафедри прикладної математики та інформаційних технологій Чернівецького національного університету імені Юрія Федьковича «Математичне моделювання і числово-аналітичні методи дослідження динамічних та інформаційних процесів».

Нові факти, отримані здобувачем та їх наукова новизна полягають в тому, що:

- У дисертаційному дослідженні Книгніцької Т.В. вперше запропоновано нову метрику для встановлення міри подібності між стаціонарними часовими рядами, які задано моделями $ARMA(p,q)$.
- Здійснено порівняння запропонованої метрики з класичними моделями визначення подібності між часовими рядами.
- Показано, що описана дисертанткою метрика є більш стійкою до викидів і дає більш точні результати для часових рядів з великою кількістю вимірювань.
- Встановлено, що складність алгоритму обчислення з використанням запропонованої метрики для N часових рядів складає $O(T*N^2)$, в той же час аналогічна складність алгоритмів DTW, ERP становить $O(T^2N^2)$. За рахунок стійкості до викидів дана метрика дозволяє отримувати більш стійкі до шумів кластери.
- Доведено, що відносна похибка вимірювань для запропонованого дисертанткою методу зростає логарифмічно, у той час як відносна похибка для класичних методів зростає лінійно. Відносна похибка вимірювань для запропонованого методу спадає з ростом кількості вимірювань часового ряду, у той же час відносна похибка вимірювань для класичних методів не змінюється з ростом кількості вимірювань у часовому ряді.
- Запропоновано новий метод визначення оптимальної кількості кластерівпри розгляді задач кластеризації об'єктів, що задаються неструктурованими даними (графами, часовими рядами тощо) на основі спектрального аналізу стохастичної матриці графу.
- Досліджено спектр стохастичної матриці графу, на основі якої вдається оцінювати оптимальну кількість кластерів для вхідних даних.
- Використовуючи симуляцію методом Монте-Карло, показано, що запропонований метод дає кращі результати для визначення оптимальної кількості кластерів у порівнянні з рядом класичних методів (марковський алгоритм з двома типами параметрів та метод ліктя). Симуляція Монте-Карло використана для утворення багатовимірних даних – графу з фіксованою кількістю сукупностей (кластерів). Таким чином, для порівняння запропонованого методу вибору оптимальної кількості кластерів з марковським алгоритмом та методом ліктя дисертантка наперед володіла інформацією про оптимальну кількість кластерів.
- Встановлено, що розроблений алгоритм знаходження оптимальної кількості кластерів є менш чутливим до наявності кластерів різного розміру.

Обґрунтованість і достовірність наукових положень, висновків і рекомендацій забезпечена тим, що матеріали дисертації опубліковані у рейтингових міжнародних журналах, які рецензуються фаховими науковцями. Публікації (3 статті) у наукових журналах з теми дисертації повністю висвітлюють проблематику та основні положення наукової роботи. Наукові праці (4), які додатково відображають наукові результати дисертації, містять приклади застосування запропонованих підходів до аналізу реальних даних. Апробація основних результатів дисертації відбулася у формі доповідей на 8 наукових конференціях. Частина розглянутих задач у дисертаційному дослідженні розв'язана точно із застосуванням добре апробованих теоретичних методів – розрахунок власних значень стохастичної матриці

графу, вибір $ARMA(p,q)$моделі для стаціонарного вхідного часового ряду, нормалізація та оцінка параметрів часового ряду, а ті, які розв'язані наближено – перевірені на збіжність та не суперечать загальноприйнятим міркуванням і принципам. Професійний огляд наукової літератури з напрямку дослідження та рівень володіння сучасними підходами до розв'язання поставлених задач свідчать про те, що дисертантка повністю володіє необхідною методологією наукового дослідження.

Наукове і практичне значення роботи.

Двома основними проблемами при кластеризації даних є визначення оптимальної кількості кластерів та визначення метрики подібності в даних. Кваліфікаційна наукова праця Книгніцької Т.В. описує нові підходи до розв'язання вище вказаних задач. Авторка вміло використала існуючі наукові дослідження у даному напрямку та розвинула їх, запропонувавши власні ідеї. Порівняння результатів, отриманих у дисертаційному дослідженні, з класичними підходами показали, що описані дисертантом алгоритми дають кращі результати для визначення подібності між часовими рядами за певних умов та встановлення оптимальної кількості кластерів для даних, які представлені графами. Варто підкреслити, що досі не існує універсального алгоритму для визначення оптимальної кількості кластерів. Даний напрямок досліджень активно розвивається. Тому наукове значення роботи є вагомим у напрямку науки про дані.

Наукове значення роботи полягає у тому, що:

- Описано алгоритм для визначення подібності між даними, які представлені часовими рядами, який базується на моделях часових рядів. У такому випадку відстань знаходиться не між самими вимірюваннями, а між параметрами моделей часових рядів.
- Показано, що описаний алгоритм дає кращі результати при розгляді довгих часових рядів та у випадку великої кількості викидів у даних у порівнянні з існуючими алгоритмами.
- Розроблено методику визначення оптимальної кількості кластерів для даних, які задаються неструктурованими типами даних. Дана методика базується на знаходженні власних значень стохастичної матриці графу.
- Встановлено, що на основі власних значень стохастичної матриці графу можна оцінювати оптимальну кількість кластерів.

Практичне значення роботи полягає у тому, що розроблена теорія дозволяє точніше аналізувати будь-які дані. У наш час наука про дані проникла у майже всі сфери життєдіяльності людини. Результати, отримані у дисертаційному дослідженні, можуть бути використані при

- Кластеризації медичних даних (підбирати індивідуальні підходи до лікування на основі схожості пацієнтів і їх реакції на терапію, розробляти програми попередження захворювань і проводити обов'язкові медичні обстеження);
- дослідженні економічних процесів (вивчення конкурентної ситуації та сегментації ринку, що дозволяють компаніям розробляти ефективні стратегії маркетингу та розвитку, оцінки ризику та портфельного управління; прогнозування економічних трендів та розвитку стратегії під них);
- групуванні користувачів у рекламній галузі (рекламодавці можуть створювати кластери споживачів на основі їхніх інтересів, демографічних характеристик і поведінки, щоб розробляти більш ефективні рекламні кампанії, персоналізовану рекламу для кожного сегмента аудиторії);
- визначенні шахрайських операцій у банківській сфері.

Усі ці приклади підкреслюють важливість кластеризації даних у великій кількості галузей, де вона допомагає у зрозумінні і використанні складних наборів даних для прийняття рішень, підвищення ефективності та досягненнябільшого розуміння ключових питань.

Оцінка змісту дисертації, її завершеність. Кваліфікаційна робота Кнігніцької Т.В. має загальний обсяг 150 сторінок машинописного тексту та структурно складається з анотації (українською та англійською мовами), списку опублікованих праць автора, переліку умовних позначень, вступу, трьох розділів, списку використаних джерел (122 позиції) та додатку (список публікацій здобувачки за темою дисертації).

У вступі обґрунтовано актуальність теми дослідження, сформульовано мету, завдання, предмет, об'єкт та методи дослідження, вказано наукову новизну, практичне значення отриманих результатів, зв'язок роботи з науковими дослідженнями та особистий внесок здобувачки, а також наведено дані про те, де доповідались, обговорювались та були опубліковані основні результати дисертації.

У першому розділі здійснено огляд наукової літератури, присвяченої дослідженню часових рядів, зокрема, визначенню метрик подібності між часовими рядами та підходи до кластеризації даних, які представлені у вигляді неструктурованих типів даних. Детально проаналізовано хронологію розвитку наукових підходів до задач кластеризації, класифікації, зменшення розмірності часових рядів. Перший пункт розділу 1 відображає загальний огляд розвитку наукових досліджень при дослідженні часових рядів та існуючі метрики для встановлення подібності між часовими рядами. У другому пункті наведено методи дослідження структурних стрибків у часових рядах. У третьому пункті зроблено огляд наукових досліджень, які стосуються неперервних часових рядів. Вибір оптимальної кількості кластерів при поділі даних на групи представлено у пункті четвертому.

У другому розділі запропоновано визначати подібність або відстань між часовими рядами за допомогою моделей часових рядів. Запропонований алгоритм для встановлення подібності двох наборів даних використовує параметри моделі, а не самі вимірювання. У якості моделей часових рядів розглянуто стаціонарні $ARMA$ моделі. Отриманий алгоритм порівнюється з уже існуючими метриками знаходження відстаней у випадку збільшення вимірювань часового ряду та у випадку зростання кількості викидів у вхідному часовому ряді. Отриманий алгоритм має меншу обчислювальну складність, ніж алгоритми Евкліда, DTW та ERP. Запропоновану відстань можна використовувати для кластеризації сильно зашумлених даних.

Наукову новизну висновків, зроблених на основі отриманих у другому розділі результатів, розкривають такі положення:

- Описано алгоритм для знаходження відстані між часовими рядами на основі моделей часових рядів. Отримана відстань є більш стійкою до викидів у часових рядах. У випадку збільшення кількості викидів запропонований у дисертаційному дослідженні алгоритм дає кращі результати (відносна похибка зростає логарифмічно), ніж аналогічні алгоритми (Евклідова відстань, ERP, DTW) для знаходження відстані між часовими рядами (відносна похибка зростає лінійно).
- Запропонований метод знаходження відстані між вимірюваннями часового ряду дає кращі результати для великих часових рядів, коли кількість вимірювань $T > 1000$. До того ж обчислювальна складність отриманого алгоритму є меншою за обчислювальну складність уже існуючих алгоритмів.

У третьому розділі розглянуто проблему кластеризації на графах на основі власних значень стохастичної матриці графа. Доведено, що власні значення стохастичної матриці для

великих графів $(N \geq 100)$ поділяються на три групи, одна із яких є визначальною для числа кластерів у графі. Використовуючи теорію випадкових матриць, вдалося показати, що асимптотичний розподіл підгрупи дійсних частин власних значень стохастичної матриці графу описується напівколовим розподілом Вігнера. Використання стохастичних матриць дало змогу точно локалізувати власні значення, що відповідають за кількість кластерів, чого не вдавалося зробити для матриць суміжності. Основні припущення моделі пов'язані з властивостями дискретних ланцюгів Маркова, що дозволяє розширити область застосування отриманих результатів на більш широкий клас об'єктів. Теоретичні результати перевірені на кластеризації часових рядів, що описують вартості $N = 470$ акцій $S\&P500$ в період з 2013 до 2018 року.

Наукову новизну висновків, зроблених на основі отриманих у третьому розділі результатів, розкривають такі положення:

- У роботі запропоновано новий метод визначення оптимальної кількості кластерів k_{opt} при кластеризації об'єктів, що задаються неструктурованими даними (графами та часовими рядами) на основі спектрального аналізу стохастичної матриці даного графу.
- Використовуючи метод Монте-Карло, вдалося показати, що запропонований метод дає кращі результати для визначення оптимальної кількості кластерів k_{opt} у порівнянні із деякими класичними методами.
- Оскільки запропонований алгоритм є спектральним, то його складність збігається зі складністю знаходження власних значень для стохастичної матриці P.
- Описаний алгоритм не є чутливим до кластерів різного розміру, тобто співвідношення між розмірами кластерів практично не впливають на точність алгоритму.
- Теоретичні результати роботи перевірено на реальних даних $N = 470$ акцій $S\&P500$, розглянутих в період з 2013 до 2018 року. Результати оцінки оптимального значення k_{opt} збіглися із відповідними оцінками для даних компаній в інший період часу.

Дисертаційні положення та побажання щодо вдосконалення змісту дисертації. Позитивно оцінюючи отримані результати дослідження, їх наукову новизну та практичну значущість, водночас, вважаю за доцільне звернути увагу на деякі аспекти.

1) У роботі наявні описки;
2) У роботі розглянуто випадок гомоскедастичних моделей $ARIMA$ та не враховується зміна дисперсії часових рядів. Варто було б перевірити актуальність застосованої теорії на гетероскедастичних моделях, наприклад $GARCH$ моделях.
3) Для порівняння оцінки кількості кластерів варто було б використати більше методів, не лише метод ліктя та марковський алгоритм.

Відсутність порушень академічної доброчесності. Кваліфікаційне дослідження є самостійною науковою працею автора. Висновки, рекомендації та пропозиції, що характеризують наукову новизну кваліфікаційного дослідження, одержані автором особисто. При використанні праць інших вчених для аргументації актуальних положень дослідження обов'язково вказано посилання на відповідні праці.

Загальний висновок. Кваліфікаційна наукова робота Книгніцької Тетяни Василівни «Оцінки параметрів авто регресійних моделей» за актуальністю, науковою новизною, загальним переліком отриманих результатів, а також їх взаємозв'язком та повнотою їх

викладу в журнальних публікаціях та апробацією цілком відповідає вимогам «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 №44 від 12 січня 2022 року зі змінами, внесеними згідно з Постановою Кабінету Міністрів України №341 від 21 березня 2022 року, а також «Вимогам до оформлення дисертації», затверджених Наказом Міністерства освіти і науки України №40 від 12 січня 2017 року, а авторка кваліфікаційної наукової роботи Книгніцька Тетяна Василівна заслуговує присудження їй наукового ступеня доктора філософії з галузі знань 11 Математика та статистика за спеціальністю 113 Прикладна математика.

Офіційний опонент:
Член-кор. НАН України
доктор фізико-математичних наук, професор,
завідувач відділу математичних методів
дослідження операцій
Інституту кібернетики ім. В.М. Глушкова
НАН України Павло КНОПОВ

Академія наук України * Національна * Інститут кібернетики імені В.М.Глушкова ідентифікаційний код 05417176 * Україна * м.Київ *

Підпис П. Кнопова
ЗАСВІДЧУЮ
Зав. канц.
ІК НАН України