АНОТАЦІЯ

Книгніцька Т.В. Оцінки параметрів авторегресійних моделей. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 113 – Прикладна математика. – Чернівецький національний університет імені Юрія Федьковича МОН України, Чернівці, 2023.

Дисертаційна робота присвячена знаходженню відстаней між вимірюваннями даних, які представлені часовими рядами, та визначенню оптимальної кількості кластерів на основі власних значень стохастичної матриці графа. Дисертація складається із вступу, трьох розділів, висновків та переліку використаних джерел.

У вступі обгрунтовано 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 здобувача, 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 пункту 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:

1. 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).

2. 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 ($N > 100$) 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:

1. 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.

2. 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.

3. Оскільки запропонований алгоритм є спектральним, то його складність збігається зі складністю знаходження власних значень для стохастичної матриці P.

4. Описаний алгоритм не є чутливим до кластерів різного розміру,тобто співвідношення між розмірами кластерів практично не впливають на точність алгоритму.

5. 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.

Практичне значення отриманих результатів

Питання про визначення відстані між вимірюваннями часових рядів (даних) та знаходження оптимальної кількості кластерів залишається відкритим. Досі не існує універсального підходу для визначення метрики подібності між часовими рядами та встановлення оптимальної кількості кластерів для даних, який однаково добре працює для наборів даних з різних сфер життєдіяльності людини. У даному дисертаційному дослідженні описано нові ідеї та підходи до розв'язання вище згаданих проблем. Результати дисертації можуть бути використані для поділу на групи (кластеризації) даних, які представлені графами або часовими рядами. Кластеризація дозволяє групувати подібні дані в категорії або кластери, що спрощує їхнє вивчення і використання у майбутньому.

Результати, отримані у даному дисертаційному дослідженні, можуть бути використані при кластеризації медичних даних: за допомогою аналізу симптомів і медичних даних можна класифікувати пацієнтів за різними хворобами або ступенями важкості захворювань; підбирати індивідуальні підходи до лікування на основі схожості пацієнтів і їх реакції на терапію; розробляти програми попередження захворювань і проводити цільові медичні обстеження.

Використання запропонованих у дисертаційному дослідженні підходів до рекламної галузі: рекламодавці можуть створювати кластери споживачів на основі їхніх інтересів, демографічних характеристик і поведінки, щоб розробляти більш ефективні рекламні кампанії; кластеризація даних допомагає рекламодавцям створювати персоналізовану рекламу для кожного сегмента аудиторії; аналіз кластерів споживачів допомагає передбачати попит на продукти і послуги в майбутньому.

В економіці кластеризація даних корисна для: дослідження конкурентної ситуації та сегментації ринку, що дозволяють компаніям розробляти ефективні стратегії маркетингу та розвитку; для оцінки ризику та портфельного управління; прогнозування економічних трендів та розвитку стратегії під них.

У наш час науковці вивчають генетичні схожості і родові зв'язки саме за допомогою кластеризації. Кластеризація може допомогти у виділенні регіонів зі схожим кліматом для дослідження змін клімату. За допомогою кластеризації у соціологічних та психологічних дослідженнях виділяють групи осіб зі схожими характеристиками для причинно-наслідкового аналізу поведінки.

Усі ці приклади підкреслюють важливість кластеризації даних у великій кількості галузей життєдіяльності людини. Завдяки кластеризації даних можна приймати правильні рішення в бізнесі, підвищувати ефективність виробництва у промисловості, оптимізувати роботу розумних мереж тощо.

Ключові слова: марковське перемикання, параметри регресії, динаміка, модель, моделювання, часові ряди, машинне навчання, нейронні мережі, збурене випадкове блукання, стохастичні диференціальні рівняння, слабка збіжність, стійкість, стохастична модель оптимізації, ройовий алгоритм, подібність кластерів.

ABSTRACT

Knignitska T. V. Estimates of parameters of autoregressive models. – Qualifying scientific project on manuscript rights.

Thesis for obtaining the scientific degree of Doctor of Philosophy in specialty 113 – Applied mathematics. – Yury Fedkovich Chernivtsi National University named after, Ministry of Education and Science of Ukraine, Chernivtsi, 2023.

The dissertation paper is devoted to finding the distances between data measurements, which are represented by time series, and determining the optimal number of clusters based on the eigenvalues of the stochastic matrix of the graph. The dissertation consists of an introduction, three sections, conclusions, and a list of used sources.

The introduction substantiates the relevance of the research topic, formulates the goal, task, subject, object, and methods of the research, indicates the scientific novelty, the practical significance of the results obtained, the connection of the work with scientific research and the personal contribution of the recipient, and also provides data about where the main results of the dissertation were reported, discussed and published.

In the first chapter a review of the scientific literature devoted to the study of time series, in particular, the definition of similarity metrics between time series and approaches to clustering data, which are presented in the form of unstructured data types, is carried out. The chronology of the development of scientific approaches to the problems of clustering, classification, and dimensionality reduction of time series is analyzed in detail. The first paragraph of Section 1 reflects a general overview of the development of scientific research in the study of time series and existing metrics for establishing similarity between time series. The second point describes the methods of researching structural

jumps in time series. In the third point, an overview of scientific research related to continuous time series is made. The selection of the optimal number of clusters when dividing the data into groups is presented in the fourth point.

The second Chapter suggests determining the similarity or distance between time series using time series models. Stationary $ARMA$ models are considered, as time series models. The resulting algorithm is compared with already existing metrics for finding distances in the case of an increase in time series measurements and in the case of an increase in the number of outliers in the input time series. The resulting algorithm has lower computational complexity than the DTW and ERP algorithms. The proposed distance can be used for clustering highly noisy data.

The scientific novelty of the conclusions drawn on the basis of the results obtained in the second chapter is revealed by the following provisions:

1. An algorithm for finding the distance between time series based on time series models is described. The resulting distance is more robust to outliers in the time series. In the case of an increase in the number of emissions, the algorithm proposed in the dissertation research gives better results (the relative error increases logarithmically) than similar algorithms (Euclidean distance, ERP, DTW) for finding the distance between time series (the relative error increases linearly).

2. The proposed method of finding the distance between time series measurements gives better results for large time series when the number of measurements $T > 1000$. In addition, the computational complexity of the obtained algorithm is lower than the computational complexity of already existing algorithms.

The third Chapter deals with the problem of clustering on graphs based on the eigenvalues of the stochastic matrix of the graph. It is proved that the

eigenvalues of the stochastic matrix for large graphs ($N > 100$) are divided into three groups, one of which is decisive for the number of clusters in the graph. Using the theory of random matrices, it was possible to show that the asymptotic distribution of the subgroup of the real parts of the eigenvalues of the stochastic matrix of the graph is described by the semicircular Wigner distribution. The use of stochastic matrices made it possible to precisely localize the eigenvalues responsible for the number of clusters, which could not be done for adjacency matrices. The main assumptions of the model are related to the properties of discrete Markov chains, which makes it possible to expand the scope of the obtained results to a wider class of objects. The theoretical results were tested on the clustering of time series describing the values of $N = 470$ shares of $S\&P500$ in the period from 2013 to 2018.

The scientific novelty of the conclusions drawn on the basis of the results obtained in the third chapter is revealed by the following provisions:

1. The paper proposes a new method for determining the optimal number of clusters k_{opt} when clustering objects given by unstructured data (graphs and time series) based on the spectral analysis of the stochastic matrix of the given graph.

2. Using the Monte Carlo method, it was possible to show that the proposed method gives better results for determining the optimal number of clusters k_{opt} in comparison with some classical methods.

3. Since the proposed algorithm is spectral, its complexity coincides with the complexity of finding eigenvalues for the stochastic matrix P.

4. The described algorithm is not sensitive to clusters of different sizes, that is, the ratio between the sizes of clusters practically does not affect the accuracy of the algorithm.

5. The theoretical results of the work were verified on real data ($N = 470$ shares of $S\&P500$, considered in the period from 2013 to 2018. The results of estimating the optimal value of k_{opt} coincided with the corresponding estimates for these companies in another time period.

Practical significance of the obtained results

The question of determining the distance between measurements of time series (data) and finding the optimal number of clusters remains open. There is still no universal approach for determining the similarity metric between time series and establishing the optimal number of clusters for data, which works equally well for datasets from different areas of human activity. This dissertation study describes new ideas and approaches to solving the above-mentioned problems. The results of the thesis can be used to divide the data into groups (clustering), which are represented by graphs or time series. Clustering allows you to group similar data into categories or clusters, which simplifies their further study and use.

The results obtained in this dissertation research can be used in the clustering of medical data: by means of the analysis of symptoms and medical data, it is possible to classify patients according to different diseases or degrees of severity of diseases; select individual approaches to treatment based on the similarity of patients and their response to therapy; develop disease prevention programs and conduct targeted medical examinations.

Applying the approaches proposed in the dissertation research to the advertising industry: advertisers can create clusters of consumers based on their interests, demographics, and behaviors to design more effective advertising campaigns; data clustering helps advertisers create personalized ads for each audience segment; analysis of consumer clusters helps predict future demand for products and services.

In economics, data clustering is useful for: competitive situation research and market segmentation, allowing companies to develop effective marketing and development strategies; for risk assessment and portfolio management; forecasting economic trends and developing strategies for them.

Nowadays, scientists study genetic similarities and ancestral relationships precisely with the help of clustering. Clustering can help in identifying regions with similar climates for climate change research. With the help of clustering in sociological and psychological research, groups of individuals with similar characteristics are distinguished for further analysis of behavior.

All these examples highlight the importance of data clustering in a large number of industries, where it helps in understanding and using complex data sets to make decisions, improve efficiency and achieve greater understanding of key issues.

Key words: Markov switching, regression parameters, dynamics, model, simulation, time series, machine learning, neural networks, perturbed random walk, stochastic differential equations, weak convergence, robustness, stochastic optimization model, swarm algorithm, cluster similarity.