

АНОТАЦІЯ

Долгих А.О. Розробка програмного комплексу аналізу та ансамблевого прогнозування рядів динаміки. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 121 «Інженерія програмного забезпечення». – Дніпровський національний університет імені Олеся Гончара, Дніпро, 2021.

Актуальність розробок математичного та програмного забезпечення аналізу та ансамблевого прогнозування часових рядів викликана потребою побудови якомога більш точних прогнози щодо майбутніх значень ряду за допомогою аналізу минулих записів. Огляд останніх досліджень показав, що сьогодні ансамблі моделей прогнозування є одним з напрямків, що активно розвиваються, особливо в тих задачах, де прогностична точність більш важлива, ніж можливість інтерпретувати модель. Для побудови адекватної прогнозувальної моделі динамічного ряду необхідним є етап його попереднього аналізу. На цьому кроці важливим завданням є ідентифікація аномальних значень часового ряду, або викидів. Такі значення суттєво відрізняються від решти рівнів ряду та можуть негативно впливати на якість побудованої моделі прогнозування або свідчити про появу нового шаблону чи тренду у досліджуваному процесі. Тому розробка методів та програмних засобів, які дозволяють знаходити такі значення автоматично та виконувати прогнозування у режимі реального часу набуває особливої актуальності.

Дисертаційна робота спрямована на розв'язання науково-технічної задачі розробки програмного комплексу аналізу та ансамблевого прогнозування рядів динаміки, який може використовуватися для розв'язку широкого кола задач, пов'язаних з аналізом та обробкою даних у фінансовому секторі, області економіки, медицини та інших галузях соціальної інфраструктури.

В роботі розроблено модифікацію алгоритму побудови ансамблів моделей прогнозування часових рядів. При використанні ансамблів особливо

важливим є питання забезпечення достатнього різноманіття моделей прогнозування. Для розв'язку цієї задачі реалізовані моделі з різних класів, а саме модель нечіткої логіки, адаптивні моделі прогнозування, метод групового урахування аргументів, поінтегровану модель авторегресії-ковзкого середнього (ARIMA), модель сингулярного спектрального аналізу (SSA). Запропонований підхід дозволяє як знаходити оптимальні параметри вище перерахованих моделей автоматично, так і змінювати їх за допомогою інтерфейсу програми. Апробація методу проводилася на часових рядах, що представляють собою щоденні коливання цін на акції міжнародних компаній. Встановлено, що застосування ансамблевого методу дозволяє підвищити точність прогнозів у 77.5% часових рядів, які підлягали аналізу. Також було показано, що використання паралельних обчислень є ефективним рішенням, яке дозволяє скоротити час, необхідний для тренування моделей, в 4.5-8.0 разів.

Запропоновано нову, багатокритеріальну методику вибору моделі прогнозування для часового ряду. Ідея підходу полягає в оцінюванні не тільки якості результатів прогнозування та складності моделі, а й проведенні аналізу залишків моделі. Алгоритм відбору оптимальної моделі прогнозування передбачає виконання двох етапів: проведення статистичних тестів та надання рейтингів за результатами їх проходження. Перший етап включає розрахунок скорегованого інформаційного критерію якості Акаїке (AIC_c) та виконання аналізу залишків побудованої моделі прогнозування, а саме: перевірку рівності середнього залишків нулю, відсутності гетероскедантичності у ряді залишків, автокореляції та тенденції. Розроблена методика використовується у процесі побудові ансамблю прогнозування для побудови комбінованого прогнозу. Встановлено, що при її використанні для вибору моделі часового ряду вдається досягти більш точних прогнозів, ніж при застосуванні стандартних методів відбору, таких як інформаційні критерії якості Акаїке та Байєса, середньоквадратична та квадратична похибка у 67% часових рядах, що підлягали аналізу.

Розроблено новий метод ідентифікації викидів у часових рядах. Запропонований алгоритм надає можливість не тільки знаходити місця появи аномальних значень, а й визначати їх тип. Метод ґрунтується на використанні адаптивних моделей прогнозування, також відомих під назвою моделі експоненційного згладжування. Адаптивні моделі прогнозування мають ряд переваг при апроксимації значень ряду, необхідній для коректної роботи методу, а саме: не накладають обмежень на довжину часового ряду, швидко підлаштовуються під динаміку змін часового ряду, не потребують складних математичних обчислень. Для оцінки якості роботи запропонованого методу був проведений ROC-аналіз результатів поділу значень ряду на два класи: нормальні значення та викиди. Результати показали, що значення AUC-ROC на досліджуваних часових рядах змінюється від 0.848 до 1.0, що свідчить про точність роботи методу при розпізнаванні аномальних рівнів ряду. Також був проведений аналіз середніх часових затрат, необхідних для коректної роботи методу. На більшості досліджуваних часових рядів запропонованому методу знадобилося в середньому менше часу, ніж при використанні методів, що ґрунтуються на використанні поінтегрованих моделей авторегресії-ковзкого середнього. Зниження середніх часових затрат, потрібних для коректної роботи методу, свідчить про перспективність використання цього підходу в аналізі систем, які потребують швидкої обробки великих об'ємів вхідних даних типу big data.

Окрім цього, у ході виконання дослідження були реалізовані обчислювальні схеми сучасних методів прогнозування та аналізу часових рядів, а саме: статистичні тести для визначення присутності у часовому ряді трендової та сезонної компоненти, алгоритми розподілу ряду на складові та виділення форм тренду і періодичної компоненти, згладжування, прогнозування часового ряду з використанням сучасних моделей, таких як модель нечіткої логіки, адаптивні моделі, метод групового урахування аргументів, поінтегрована модель авторегресії-ковзкого середнього (ARIMA)

та модель сингулярного спектрального аналізу (SSA), оцінка їх за допомогою дисперсійного аналізу.

На основі представлених вище розробок на платформі .NET (мова програмування C#) був створений програмний комплекс аналізу та ансамблевого прогнозування часових рядів. Для розробки програмного забезпечення був застосований об'єктно-орієнтований підхід. Реалізовані моделі прогнозування відносяться до єдиної ієрархії та є наслідниками одного, який надає загальний інтерфейс для роботи з усіма моделями прогнозування та втілює загальні методи. Для організації процесу навчання та оцінки моделей ансамблю використовуються паралельні обчислення, а саме засоби бібліотеки розпаралелювання задач (TPL) платформи .NET. Використаний у програмному комплексі підхід дозволяє раціонально використовувати ресурси комп'ютері та слідкувати за тим, щоб уникнути створення більшої кількості потоків, ніж це потрібно.

Розроблений програмний комплекс містить достатньо широкий спектр функцій для аналізу та прогнозування часових рядів. Деякі з них, такі як побудова ансамблю моделей прогнозування та вибір оптимальної моделі досліджуваних даних, можуть потребувати істотних часових затрат. Для того, щоб інтерфейс користувача не був заблокований на цей період та дозволяв проводити аналіз часового ряду з використанням інших засобів, у програмі використовуються асинхронні виклики. Для роботи з асинхронними подіями була використана модель програмування TAP (Task-based asynchronous pattern), яка представляє собою зручну абстракцію для асинхронного коду.

У дисертаційній роботі представлено опис процесу проектування архітектури програмного забезпечення аналізу та ансамблевого прогнозування рядів динаміки за допомогою за допомогою уніфікованої мови моделювання (UML). Показані та охарактеризовані діаграми класів, варіантів використання та діяльності.

Засобами розробленого програмного комплексу проведено аналіз та прогнозування часових рядів фінансової та економічної природи, а саме,

часових рядів, які представляють собою щоденні коливання цін на акції міжнародних компаній, значення попиту на товар, вартість комплектуючих деталей підприємств сервісного обслуговування для українських та міжнародних підприємств.

Ключові слова: ряди динаміки, прогнозування, аналіз часових рядів, модель нечіткої логіки, модель нейронної мережі, адаптивні моделі прогнозування, метод групового урахування аргументів, поінтегрована модель авторегресії-ковзкого середнього (ARIMA), модель сингулярного спектрального аналізу (SSA), ансамблеве прогнозування, викиди у часових рядах, розробка програмного забезпечення, засоби організації паралельних обчислень, асинхронні події.

SUMMARY

Dolhikh A.O. Development of the software system for time series analysis and ensemble forecasting. – Qualifying scientific work as manuscript.

Dissertation for the degree of Ph. Dissertation on speciality 121 «Software Engineering». – Oles Honchar Dnipro National University, Dnipro, 2021.

The actuality of software development for time series analysis and ensemble forecasting is caused by the need to build as accurate predictions as possible regarding the future values of time series using the analysis of past recordings. The overview of recent studies has shown that today the ensemble of prognostication models is one of the directions which are actively developing, especially in those objectives, where prognostic accuracy is more important than the opportunity to interpret the model. To build an adequate predictive model of the time series, a stage of its preliminary analysis is necessary. At this step, an important task is to identify anomalous values of the time series, or outliers. Such values differ significantly from the other levels of the series and may negatively affect the quality of the constructed prognostic model or indicate the appearance of a new pattern or trend of the process being studied. Consequently, elaboration of methods and programme devices that

enable you to detect these values automatically and perform prognostication in real time is of particular relevance.

The thesis work is aimed at solving the scientific and technical problem of developing a software complex of time series analysis and ensemble forecasting, which can be used to solve a wide range of problems related to the analysis and processing of data in the financial sector, economics, medicine and other branches of social infrastructure.

The algorithm of constructing models ensembles of time series forecasting has also been modified in present research. Analysis of modern research has shown that it is especially important to ensure a sufficient variety of prognostic models, using ensembles. To solve this problem, models from different classes were implemented, namely, a fuzzy logic model, adaptive prognostic models, a group method of data handling, an autoregressive integrated moving average model (ARIMA) model and a singular spectrum analysis model (SSA). The proposed algorithm of constructing ensembles allows both to find the optimal parameters of the listed models automatically and to change them using the program interface. The method was tested on time series representing daily share price wave of international companies. It has been established that the use of the ensemble method makes it possible to improve the accuracy of forecasts in 77.5% of the time series to be analyzed. It was also determined, that the use of parallel computing is an efficient solution that allows you to reduce the time required for models training by 4.5-8.0 times.

A new method of prognostic model selection for time series was offered. The novelty of this approach lies in assessing not only the quality of prognostic results and the complexity of the model, but also in analyzing the model remains. The algorithm of the optimal selected prognostic model provides two stages: conducting statistical tests and assigning ratings based on their results. The first stage includes the calculation of corrected Akaike information criterion (AICc) and the analysis of the constructed prognostic model remains, namely: checking the equality of the main remains to zero, the absence of heteroscedantism in a number of residuals, autocorrelation and trends. The developed methods are used in the process of

prognostic ensemble constructing to build a combined forecast. It was determined, when it is used to select a time series model, it is possible to achieve more accurate predictions than when using standard selection methods, such as such as Akaike and Bayesian quality information criteria, root mean square error and square error, in 67% of time series, qualify for an analysis.

A new method of outliers identifying in time series has been developed. The proposed algorithm allows not only finding the places of occurrence of abnormal values, but also determining their type. The method is based on the use of adaptive forecasting models, also known as exponential smoothing models. Adaptive predictive models have a number of advantages while approximating range values necessary for proper operation of the method: do not impose restrictions on the time series length, quickly adapt to the dynamics of the time series of changes, do not require complex mathematical calculations. To assess the quality of the proposed method, an ROC-analysis of the results of dividing the range values into two classes was carried out: normal values and outliers. The results showed that the AUC-ROC values on the studied time series varies from 0.848 to 1.0, which indicates the accuracy of the method in recognizing the anomalous levels of the series. We also analyzed the average time required for the correct operation of the method. On most of the studied time series, the proposed method took on average less time than when using methods based on the use of autoregressive integrated moving average models. The decrease in the average time required for the correct operation of the method indicates the prospects of using this approach in the analysis of systems that require fast processing of large volumes of input data such as big data.

In addition, in the course of the study, calculating schemes of modern methods of forecasting and analyzing time series were implemented, namely: statistical tests to determine the presence of a trend and seasonal component in a time series, algorithms for distributing the series into components and identifying trend and periodic components, smoothing, the time series forecasting using modern models, namely, fuzzy logic model, adaptive models, the group method of data handling, the autoregressive integrated moving average model (ARIMA) and the singular

spectrum analysis model (SSA), their assessment using analysis of variance (ANOVA).

Based on the developments described above on the platform .NET (programming language C#) a software package for analysis and ensemble forecasting of time series was created. An object-oriented approach was used for programme development. The implemented forecasting models belong to a single hierarchy and are inheritors of the one, which provides a common interface for working with all forecasting models and implements common methods. To organize the learning process and to evaluate the ensemble models, parallel computations are used, namely the task parallel library (TPL) of the .NET platform. The approach used in the software package makes it possible to use rationally the computer resources and to ensure that the creation of a larger stream number than necessary is avoided.

The developed software package contains a fairly wide range of functions for analyzing and forecasting time series. Some of them, such as building an ensemble of forecasting models and choosing the optimal model of the data under study, can be time-consuming. In order, that the user interface has not been blocked for that period and allows the analysis of time series using other means, asynchronous calls are used in the program. To work with asynchronous events the task-based asynchronous pattern (TAP) was used, which is a convenient abstraction for an asynchronous code.

A description of the design process of software architecture for analysis and ensemble forecasting of dynamics series using the unified modeling language (UML) is presented. Class diagrams, use cases and activity diagrams are shown and characterized.

By means of the developed software complex, the analysis and forecasting of time series of financial and economic nature, namely, time series, which represent daily fluctuations in the international companies share prices, the value of demand for goods, the cost of component parts of service enterprises for Ukrainian and international companies were carried out.

Keywords: dynamic series, forecasting, time series analysis, fuzzy logic model, neural network model, adaptive forecasting models, group method of data handling, autoregressive integrated moving average model (ARIMA), singular spectrum analysis model (SSA), ensemble forecasting, time series outliers, software development, means of organizing parallel calculations, asynchronous events.

Список опублікованих праць за темою дисертації

Статті у наукових виданнях, включених до Переліку наукових фахових видань України:

1. A.Dolhikh., O.Baibuz. The software development for time series forecasting with using adaptive methods and analysis of their efficiency. Mathematical modeling. 2019. Vol. 41, No 2, PP. 7-16.

2. А.О. Долгіх, О.Г. Байбуз. Аналіз методів, моделей та програмних засобів прогнозування часових рядів. Збірник наукових праць Національного аерокосмічного університету ім. Жуковського «ХАІ» «Відкриті інформаційні та комп'ютерні інтегровані технології». 2018. Т.79, С. 74-87.

3. Долгіх А.О., Байбуз О.Г. Порівняльний аналіз ефективності моделей прогнозування часових рядів за допомогою багатокритеріальної процедури на прикладі фінансових показників. Вісник Житомирського державного технологічного університету. Серія: Технічні науки. 2019. Т.83(1), С. 131-141.

4. Долгіх А.О., Байбуз О.Г. Огляд сучасних розробок прогнозування часових рядів з використанням прихованих марківських моделей. Актуальні проблеми автоматизації та інформаційних технологій. 2017. Т.21, С. 60-73.

Статті у наукових періодичних виданнях Європейського Союзу з наукового напрямку, з якого підготовлено дисертацію:

5. Dolhikh A.O., Baybuz O.G. The software technology of time series outliers' identification. Austrian Journal of Technical and Natural Sciences. 2019. Vol. 11-12, PP. 7-14.

Наукові праці, які засвідчують апробацію матеріалів дисертації:

6. О.Г. Байбуз, А.О. Долгіх. Використання паралельних обчислень при прогнозуванні фінансових показників. Математичне та програмне

забезпечення інтелектуальних систем (МПЗІС-2019): тези доповідей XVII міжнародної науково-практичної конференції (Дніпро, Україна, 20-22 листопада 2019). Дніпро: ДНУ, 2019 р., С. 12.

7. А.О. Долгіх, О.Г. Байбуз. Розробка технології пошуку викидів у часових рядах фінансових показників. Сучасні інформаційні і комунікаційні технології на транспорті, в промисловості та освіті: тези доповідей XIII міжнародної науково-практичної конференції (Дніпро, Україна, 11-12 грудня 2019). Дніпро: ДНУЗТ , 2019 р., С. 69.

8. Долгіх А.О., Байбуз О.Г., Білобородько О.І. Застосування адаптивних моделей для прогнозування економічних часових рядів. Математичне та програмне забезпечення інтелектуальних систем (MPZIS-2016): тези доповідей XIV міжнародної науково-практичної конференції (Дніпро, Україна, 16-18 листопада 2016). Дніпропетровськ: ДНУ, 2016 р., С. 64-66.

9. Долгіх А.О., Байбуз О.Г. Використання марківських моделей у прогнозуванні часових ряді. Інформаційні технології в металургії та машинобудуванні : матеріали міжнародної науково-технічної конференції (Дніпро, Україна, 28-30 березня 2017). Дніпро: Національна металургійна академія України, 2017 р., С.63.

10. Долгіх А.О. Огляд сучасних методів прогнозування часових рядів з використанням марківських моделей. Людина і космос: тези доповідей XIX міжнародної молодіжної науково-практичної конференції (Дніпро, Україна, 12-14 квітня 2017). Дніпро: Національний центр аерокосмічної освіти молоді ім. Макарова, 2017 р., С.167.

11. Долгіх А.О., Байбуз О.Г. Аналіз моделей та програмних засобів прогнозування фінансових часових рядів. Математичне та програмне забезпечення інтелектуальних систем (MPZIS-2018): тези доповідей XVI міжнародної науково-практичної конференції (Дніпро, Україна, 21-23 листопада 2018). Дніпро: ДНУ, 2018 р., С 57-59.

12. Долгіх А.О., Байбуз О.Г. Аналіз сучасних розробок в області прогнозування часових рядів з використанням прихованих моделей Маркова.

Сучасні інформаційні та комунікаційні технології на транспорті, в промисловості та освіті: тези доповідей XII міжнародної науково-практичної конференції (Дніпро, Україна, 12-13 грудня 2018). Дніпро: Дніпропетровський національний університет залізничного транспорту ім. В. Лазаряна, 2018 р, С.88.

13. Долгіх А.О., Байбуз О.Г. Використання функцій пакету R у програмах, розроблених на платформі .NET у Microsoft Visual Studio. Science progress in European countries: new concepts and modern solutions: матеріали IV міжнародної наукової конференції (Stuttgart, Germany, 28 грудня 2018). Stuttgart, Germany: ORT Publishing, 2018 р., С. 227-233.

14. А.О. Долгіх, О.Г. Байбуз. Розробка програмного забезпечення попереднього аналізу часових рядів. Science progress in European countries: new concepts and modern solutions: матеріали V міжнародної наукової конференції (Stuttgart, Germany, 28 лютого 2019 р.). Stuttgart : ORT Publishing, 2019 р., С. 204-214.

15. А.О. Долгіх, О.Г. Байбуз. Прогнозування фінансових часових рядів з використанням нейронної мережі. Молоді вчені 2019 – від теорії до практики: матеріали X міжнародної конференції молодих вчених (Дніпро, Україна, 07 березня 2019 р.). Дніпро: ТОВ «Дніпровський освітній центр», 2019 р., С. 174-179.

16. А.О. Долгіх, О.Г. Байбуз. Розробка програмного забезпечення ідентифікації викидів часового ряду. Інформаційні технології в металургії та машинобудуванні: тези доповідей міжнародної науково-технічної конференції (Дніпро, Україна, 26-28 березня 2019 р.). Дніпро: НМетАУ, 2019 р., С. 53.

17. А.О. Долгіх, О.Г. Байбуз. Розробка програмного забезпечення часових рядів з використанням інтегрованої моделі авторегресії-ковзкого середнього. Perspectives of science and education: матеріали VIII міжнародної молодіжної конференції (New York, USA, 29 березня 2019). New York: СЛОВО/WORD, 2019 р., С.83-88.

18. А.О. Долгіх, О.Г. Байбуз. Розробка програмного забезпечення прогнозування часових рядів з використанням методу групового урахування аргументів. Сучасні тенденції у розвитку науки та освіти: матеріали V Всеукраїнської науково-практичної конференції молодих вчених (Кам'янець-Подільський, Україна, 04 квітня 2019 р.). Кам'янець-Подільський: ТОВ «Друкарня Рута», 2019 р., С. 108-114.

19. А.О. Долгіх. Прогнозування фінансових часових рядів з використанням нейронної мережі типу МГУА. Радіоелектроніка і молодь у ХХІ столітті: тези доповідей 23-го міжнародного молодіжного форуму (Харків, Україна, 16-18 квітня 2019 р.). Харків: ХНУРЕ, 2019 р., С. 146-148.

20. А.О. Долгіх, О.Г. Байбуз. Розробка програмного забезпечення прогнозування часових рядів з використанням процедури нечіткого виводу. V Всеукраїнський форум студентів, аспірантів і молодих учених: тези доповідей (Дніпро, Україна, 25-26 квітня 2019 р.). Дніпро: ДНУ, 2019 р., С. 95-98.

21. А.О. Долгіх. Прогнозування фінансових часових рядів з використанням моделей Маркова. Проблеми прикладної математики та інформаційних технологій: тези доповідей за підсумками науково-дослідної роботи Дніпровського національного університету імені Олеся Гончара за 2018 рік (Дніпро, Україна, 17 червня 2019 р.). Дніпро: ДНУ, 2019 р., С. 30-32.

22. O. Baibuz, A. Dolhikh. Development of the software system for ensemble time series forecasting. Математичне та програмне забезпечення інтелектуальних систем (МПЗІС-2020): тези доповідей XVIII міжнародної науково-практичної конференції (Дніпро, Україна, 18-20 листопада 2020). Дніпро: ДНУ, 2020 р., С. 103-105.