UDC 519.2+004.8

THE METHOD OF ADAPTATION OF THE PARAMETERS OF ALGORITHMS FOR THE DETECTION AND CLEANING OF A STATISTICAL SAMPLE FROM ANOMALIES FOR DATA SCIENCE PROBLEMS

Pysarchuk O. O. – Dr. Sc., Professor, Professor of the Department of Computer Engineering, Faculty of Informatics and Computing, National Technical University of Ukraine "Ihor Sikorskyi Kyiv Polytechnic Institute".

Pavlova S. O. – Student of the Faculty of Informatics and Computing, National Technical University of Ukraine "Ihor Sikorskyi Kyiv Polytechnic Institute".

Baran D. R. – Assistant of the Department of Computer Engineering, Faculty of Informatics and Computing, National Technical University of Ukraine "Ihor Sikorskyi Kyiv Polytechnic Institute".

ABSTRACT

Context. Popularization of the Data Science for the tasks of e-commerce, the banking sector of the economy, for the tasks of managing dynamic objects – all this actualizes the requirements for indicators of the efficiency of data processing in the Time Series format. This also applies to the preparatory stage of data analysis at the level of detection and cleaning of statistical samples from anomalies such as rough measurements and omissions.

Objective. The development of the method for adapting the parameters of the algorithms for detecting and cleaning the statistical sample of the Time Series format from anomalies for Data Science problems.

Method. The article proposes a method for adapting the parameters of algorithms for detecting and cleaning a statistical sample from anomalies for data science problems. The proposed approach is based on and differs from similar practices by the introduction of an optimization approach in minimizing the dynamic and statistical error of the model, which determines the parameters of settings of popular algorithms for cleaning the statistical sample from anomalies using the Moving Window Method.

Result. The introduction of the proposed approach into the practice of Data Science allows the development of software components for cleaning data from anomalies, which are trained by parameters purely according to the structure and dynamics of the Time Series.

Conclusions. The key advantage of the proposed method is its simple implementation into existing algorithms for clearing the sample from anomalies and the absence of the need for the developer to select parameters for the settings of the cleaning algorithms manually, which saves time during development. The effectiveness of the proposed method is confirmed by the results of calculations.

KEYWORDS: anomaly detection, dynamic error, statistical error, model optimization, Moving Window, Data Science, Big Data, time series.

ABBREVIATIONS

AM is an Abnormal Measurements;

ARIMA is an Autoregressive Integrated Moving Average;

MAE is a mean absolute error.

NOMENCLATURE

θ is a dynamic error;

σ is a standard deviation (square root of variance);

σ^2 is a variance of the error sample ($D[y \mid x]$);

σ_y^2 is a variance of the dependent variable sample ($D[y]$);

$D[y]$ is a variance of the sample;

$D[y \mid x]$ is a conditional variance of the dependent variable given factors x (variance of the model error);

e_i is a model error;

n is a number of observations in the sample;

R^2 is a coefficient of determination;

$threshold$ is a threshold for anomaly detection;

$window\_size$ is a size of the sliding window;

x_i is an predicted values of the variable;

y_i is an actual values of the variable.

INTRODUCTION

At present, Data Science tasks have gained immense popularity, as they allow the use of large amounts of data (Big Data) to obtain valuable information and make informed decisions [1–14]. It should be noted that this trend has been maintained for many years, which is due to the development of information technologies and their implementation in many areas.

One of the areas of Data Science is the processing of data in the format of a time series, which characterizes the studied processes with a discrete series of values that change in time or depending on another argument (variable). Examples of time series processing tasks are the analysis of changes in time and forecasting: indicators of economic efficiency of trading companies; weather indicators; changes in exchange rate fluctuations; global statistical indicators of the state's economy – production of agricultural products, population growth, subsistence minimum, morbidity of the population, etc.; navigation parameters of the movement of dynamic objects – airplanes, cars, robotic/unmanned aerial vehicles and many other industries.

DOI 10.15588/1607-3274-2025-3-4

These examples are focused on high accuracy of time series processing. This is achieved by considering the heterogeneity of the input data due to the presence of abnormal measurements (AM). The problem of clearing the time series from AM is quite common [1–3, 6–10]. Since, depending on the ratio of the number and magnitude of AM to the number of measurements in the time series, anomalies can significantly distort the processing results. However, this is an additional stage, which on Big Data is critical to the conflict of attracting resource space and the efficiency of obtaining the result. Therefore, more often they prefer simple but effective algorithms built on the principles of a sliding window [3, 4, 8–14]. Here, simplicity is a positive and negative property at the same time. The negative is manifested in fixing the parameters of such algorithms. But this does not allow you to adapt to dynamic data properties. This phenomenon is significantly manifested by data with significant nonlinearities, seasonal variations, etc. That is, the algorithms for clearing the time series from AM with fixed parameters are fast, but "blind" to the dynamics of data changes. This leads to the need to support program implementations of such approaches, which is not always justified and possible.

One of the basic approaches to time series processing is statistical training methods. But they apply pre-prepared data through AM cleanup.

In connection with the above, the task of developing effective (in terms of speed and accuracy) approaches to adapting the parameters of algorithms for detecting and cleaning the statistical sample from anomalies for data science problems **is relevant**.

The object of study is the process of purifying the statistical sample from anomalous measurements

The subject of study is methods for cleaning the statistical sample from anomalous measurements.

The purpose of the work is to develop a method for adapting the parameters of algorithms for detecting and cleaning the statistical sample of the Time Series format from anomalies for Data Science problems.

1 PROBLEM STATEMENT

Time series processing methods are quite common and are represented by algorithms such as ARIMA, regression analysis and statistical training (such as the method of least squares (LSM) and others), deep learning using artificial neural networks [2–4]. The quality of application of all these approaches is largely determined by the quality of data preparation for processing. One of the stages of data preparation is to clean them from anomalous measurements – those that differ significantly in their values from other measurements and disrupt the dynamics of the time series, as well as data omissions. Depending on the absolute values of AM and the ratio of the number of AMs to the sample size of the time series, anomalies can distort the processing results quite strongly [6–9]. Therefore, in the process of data preparation, it is necessary to provide for the stage of clearing the sample of measurements from AM. In turn, the process of clearing time series from AM is and remains one of the most difficult and time-consuming tasks in the field of Data Science. This is due to the complex nature of the reasons for the appearance of AM and their negative impact on the results of processing. At the same time, quite high requirements for performance are put forward to the algorithms for clearing time series from AM (especially on Big Data arrays) and to autonomous adaptation (adaptation of parameters by "self-learning" depending on the properties of the Time Series – the nature of the trend, statistical characteristics, etc.).

Let us assume that a set of measurements y_i ,that form the Time Series. It is known that the measurements are distributed with normal law and a contain certain percentage of anomalous measurements. Detection of anomalous measurements is carried out using a sliding window algorithm, the efficiency of which is determined by the parameters $threshold$ – the anomaly detection threshold and $window\_size$ – the size of the sliding window.

It is necessary to develop a method for adapting the parameters of the algorithms for detecting and cleaning the statistical sample from anomalies – $threshold$, $window\_size$ to the properties of a specific sample of measurements.

Criteria and limitations. The method under development should ensure the minimization of dynamic θ and stochastic σ estimation of errors on a limited set of time series measurements n .

2 REVIEW OF THE LITERATURE

In the problems of clearing the sample from anomalies, there are quite a lot of varieties of methods and algorithms based on different approaches and principles [6–14]. All known approaches are based on unitary and/or combinatorial analysis of AM features. In general, there are AM of the rough dimensions and AM of the omission type. In both cases, the signs of AM are a change in the dynamics of the time series (dynamic properties); a difference in the value of a single dimension compared to other dimensions (properties of absolute values measurements); changes in the statistical properties of the sample in the presence of AM (statistical properties).

Depending on the signs used to detect AM and the principles of their detection, the following classes should be distinguished:

– methods of clustering according to the principles of machine learning [5, 6];

– methods for analyzing the dynamic properties of the time series [8];

– methods for analyzing the statistical properties of the time series [8, 9].

Despite the versatility and wide representation of these approaches, their key drawback is the empirical (research) adjustment of their parameters, depending on the nature of the properties of the time series. This may not be acceptable,

DOI 10.15588/1607-3274-2025-3-4

as finding the best solutions can take a significant amount of time during the development phase. It also complicates their practical implementation and scalability for time series with a wide range of properties that sometimes change during the operation of the software system. The disadvantages of known approaches to training according to the parameters of algorithms for clearing the sample from AM are also in the complexity of their implementation on Big Data arrays with significant nonlinearities and seasonalities.

3 MATERIALS AND METHODS

The method under development is aimed at supplementing the known time series cleaning algorithms based on the principle of a sliding window, for example: Moving Window Method, Median Filtering algorithm or Least Squares Method [9].

The main idea of the proposed method is as follows.

The parameters to be determined are the size of the sliding window and the threshold value (sensitivity) of the algorithms for detecting and cleaning the time series from the anomalies. These parameters are determined from the list of discrete values that ensure a minimum of dynamic and statistical error in the model of the results of statistical selection after cleaning the time series from the anomalies. The method of statistical learning is used as the Least Squares Method [9].

It is advisable to put forward the following requirements for the method of adaptation of the parameters of the algorithms for detecting and cleaning the statistical sample from anomalies:

1) The use of the method of parameter adaptation should lead to an improvement in the results of cleaning the sample from anomalies in accordance with the quality metrics of the statistical learning model given below.

2) The method of parameter adaptation should be based on the choice of a statistical learning model with the minimum combination of dynamic and statistical error.

3) Sample cleaning by the developed method should not remove structurally important properties of the sample.

We will introduce model quality indicators to understand how successful data cleaning from anomalies was. We will take the mean absolute error (MAE) and the coefficient of determination (R^2) as such metrics.

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n}, \quad (1)$$

$$R^2 = 1 - \frac{D[y \mid x]}{D[y]} = 1 - \frac{\sigma^2}{\sigma_y^2}, \quad (2)$$

where $D[y] = e_y^2$ is the variance of the random error of the measured sample y, and $D[y \mid x] = \sigma^2$ is the conditional (by factors x) variance of the dependent variable (variance of the model error).

Considering the above requirements and model quality indicators, a method for adapting the parameters of algorithms for detecting and cleaning a statistical sample from anomalies has been developed, based on minimizing the dynamic and statistical error of the statistical learning model.

A method for adapting the parameters of the algorithms for detecting and cleaning up the sample anomalies.

The size of the sliding window and the threshold for detecting the anomalies are subject to adaptation based on the characteristics of the input sample. This is done by finding a balance between the dynamic and statistical errors of the statistical learning model [3–5]. The dynamic error is the previously defined metric of mean absolute error (MAE), and the statistical error is the coefficient of determination (R^2).

For a representative sample, the mean absolute error is minimal, and the coefficient of determination is close to one. A small MAE value guarantees minimal discrepancy between data without anomalies and the results of anomaly removal algorithms. An R^2 value close to one means that the model reproduces the data well and considers all its variability.

If the parameters of the algorithms for detecting and cleaning the sample from the outliers are incorrectly defined, this will result in the outliers remaining in a posteriori sample. The presence of anomalies in the sample will lead to an increase in the mean absolute error (MAE) and a decrease in the coefficient of determination (R^2), which can be used as feedback for evaluating the next combination of parameters of the algorithms for detecting and cleaning the sample from the outliers.

Thus, the problem of parameter adaptation is reduced to minimizing the result of the expression calculation: $MAE + (1 - R^2)$ over the course of the values of the parameters of the algorithms for detecting and cleaning the sample from the anomalies. Reducing the result of calculating this expression means that the model has smaller errors (low MAE) and at the same time explains the data well (high R^2). This is a consequence of the quality of the algorithm for cleaning the sample from outliers.

The stages of the method of adapting the parameters of the algorithms for detecting and cleaning the sample from the anomalies include the following.

1. Determine the range for optimizing the $window\_size$ and $threshold$ parameters within the specified limits. The boundaries, i.e. the minimum and maximum values of $window\_size$ and $threshold$, are determined using the statistical parameters of the input sample – sample size, standard deviation, etc.

a) The $window\_size$ parameter affects how many neighboring values will be considered during data cleaning. Determining the optimal window size allows you to balance data smoothing and detail preservation.

DOI 10.15588/1607-3274-2025-3-4

b) The *threshold* parameter defines the acceptable level of deviation for anomaly detection. Values above this threshold are considered anomalies. Determining the optimal threshold allows you to effectively detect and eliminate anomalies without unnecessarily deleting correct data.

2. A nested loop is executed for *window_size* and *threshold*. At each iteration of the loop, one of all possible combinations of *window_size* and *threshold* within the previously defined limits is considered.

3. For each combination of *window_size* and *threshold*, one of the following data cleaning algorithms is used: Moving Window Method, Median Filtering, or Least Squares Method.

4. For each combination of *window_size* and *threshold*, the $MAE + (1 - R^2)$ values are calculated for the original and cleaned data.

5. For each combination of *window_size* and *threshold*, the $MAE + (1 - R^2)$ values of the current combination are compared with the best values of the previous combinations. If the current values are better (less MAE and more R^2), they become the best values.

6. The result is a combination of *window_size* and *threshold* parameters with the lowest $MAE + (1 - R^2)$ value.

To implement these stages of the method of adapting the parameters of the algorithms for detecting and cleaning the sample from anomalies, a software script was developed in the Python programming language with the numpy [11], pandas [10], and matplotlib libraries.

To evaluate the effectiveness of the proposed solutions, several computational experiments were conducted. The essence of the experiments is to process a stochastic sample with anomalies by a known algorithm and an algorithm using the developed method. The analysis of the results was carried out by comparing the initial and final characteristics of the sample obtained using the traditional and the proposed approaches.

4 EXPERIMENTS

We will conduct a series of experiments to evaluate the effectiveness of the method of adapting the parameters of the Moving Window Method [9], Least Squares Method, and Median Filtering algorithms for the task of cleaning the sample from anomalies.

A statistical sample of $n = 21$ measurements was subject to modeling. The basis was real data: statistics on Russian army losses for 1–21 September 2023. The data is presented by category: personnel, armored combat vehicles, tanks, artillery, aircraft, helicopters, ships.

For the Mowing Window Method, the standard deviation in the input sample is $\sigma = 0.2800$, and the dynamic error is $\theta = 0.2287$. After modeling the addition of 10% of anomalies, which are uniformly distributed over the sample, we have the following characteristics of the statistical sample: $\sigma = 0.3763$, $\theta = 0.2863$.

For the Least Squares Method, the input sample contained: $\sigma = 0.2669$, $\theta = 0.2033$. The sample with anomalies: $\sigma = 0.4115$, $\theta = 0.3209$.

For the Median Filtering algorithm, the input sample contained: $\sigma = 0.2710$, $\theta = 0.2035$. Sample with anomalies: $\sigma = 0.3463$, $\theta = 0.2705$.

5 RESULTS

The results of the study of the method of parameter adaptation based on the Mowing Window Method are shown in Fig. 1.

Fig. 1a shows the sample plot (dependence of the value of the controlled parameter “Values” on time “Time”) after using the well-known Moving Window Method: $\sigma = 0.1840$, $\theta = 0.1398$. The model quality indicators mean absolute error $MAE = 0.2876$, coefficient of determination $R^2 = 0.7649$.

Instead, Fig. 1b shows the sample plot after using the developed method of parameter adaptation, which has error values: $\sigma = 0.2535$, $\theta = 0.1933$. The following model quality indicators were obtained: $MAE = 0.2569$, $R^2 = 0.7761$.

The statistical characteristics show that the algorithm without a method of parameter adaptation also removes structurally important data. While the proposed approach allows preserving the structure of the input sample and provides better model accuracy.

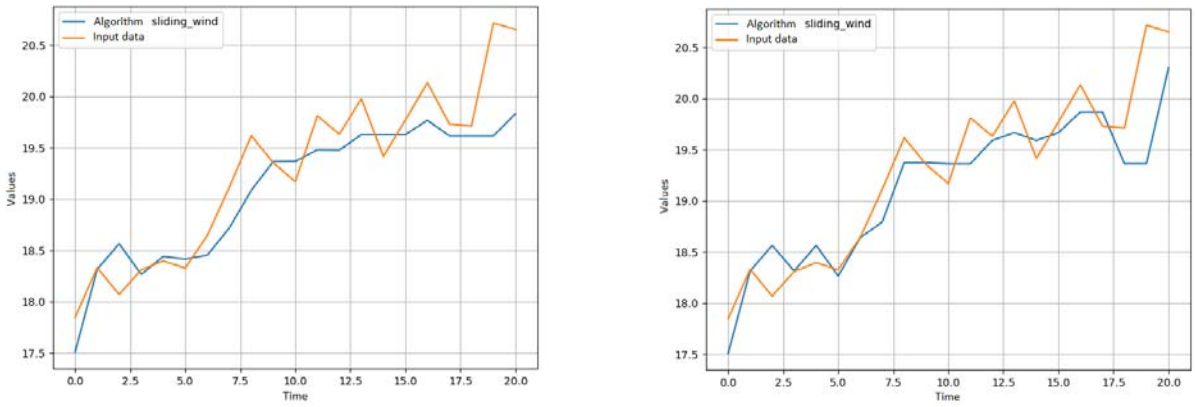
The results of the study of the method of parameter adaptation based on Least Squares Method are shown in Fig. 2, where the notation is like that of Fig. 1.

The use of the well-known Least Squares Method algorithm gave the following results: $\sigma = 0.1220$, $\theta = 0.0694$. When applying the developed method, we have: $\sigma = 0.0845$, $\theta = 0.0304$.

Comparison of the graphs of Fig. 2, and Fig. 2 b and the model quality criteria indicate that the sample is still representative when using the developed adaptation method. While the well-known Least Squares Method algorithm focuses more on the initial values of the sample.

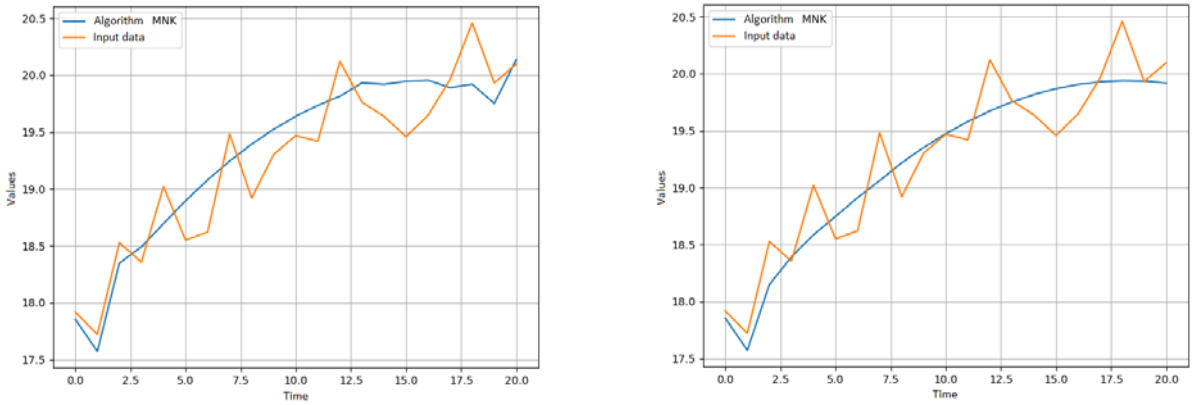
The results of the study of the method of parameter adaptation based on the Median Filtering are presented in Fig. 3, where the notation is like that of Fig. 1. The well-known Median Filtering showed the following results: $\sigma = 0.1653$, $\theta = 0.1229$. Whereas the optimized algorithm is: $\sigma = 0.1910$, $\theta = 0.1532$. Comparison of the graphs of Fig. 3a and Fig. 3b also demonstrates a decrease in the average absolute error and increase in the coefficient of determination when using the proposed approach, which indicates its effectiveness.

DOI 10.15588/1607-3274-2025-3-4



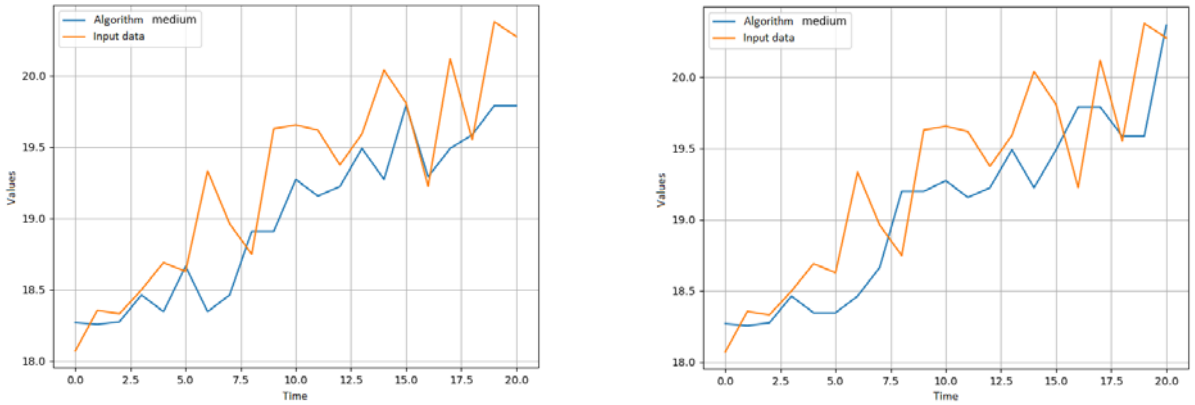
a b

Figure 1 – Results of the study of the method of parameter adaptation based on the Moving Window Method through comparison of the input and cleaned samples: a –using a well-known Moving Window Method ($MAE = 0.2876$, $R^2 = 0.7649$), b – using adaptation of Moving Window Method ($MAE = 0.2569$, $R^2 = 0.7761$)



a b

Figure 2 – The results of the study of the method of parameter adaptation based on the Least Squares Method: a – using well-known Least Squares Method ($MAE = 0.2598$, $R^2 = 0.8353$), b – using adaptation of Least Squares Method parameters ($MAE = 0.2154$, $R^2 = 0.8609$)



a b

Figure 3 – Results of the study of the method of parameter adaptation based on the Median Filtering: a – using well-known Median Filtering ($MAE = 0.3243$, $R^2 = 0.5809$), b – using adapting the parameters of the Median Filtering ($MAE = 0.2630$, $R^2 = 0.7685$)

DOI 10.15588/1607-3274-2025-3-4

Table 1 – Generalized statistical characteristics of the approaches under study

Algorithm	Input sample	A sample with anomalies	Standard algorithm	Minimization method
Moving Window Method	$\sigma = 0.2800$, $\theta = 0.2287$	$\sigma = 0.3763$, $\theta = 0.2863$	$\sigma = 0.1840$, $\theta = 0.1398$	$\sigma = 0.2535$, $\theta = 0.1933$
Least Squares Method	$\sigma = 0.2669$, $\theta = 0.2033$	$\sigma = 0.4115$, $\theta = 0.3209$	$\sigma = 0.1220$, $\theta = 0.0694$	$\sigma = 0.0845$, $\theta = 0.0304$
Median Filtering	$\sigma = 0.2710$, $\theta = 0.2035$	$\sigma = 0.3463$, $\theta = 0.2705$	$\sigma = 0.1653$, $\theta = 0.1229$	$\sigma = 0.1910$, $\theta = 0.1532$

Table 2 – Summary of model quality indicators

Algorithm	Standard algorithm	Minimization method
Moving Window Method	$MAE = 0.2876$, $R^2 = 0.7649$	$MAE = 0.2569$, $R^2 = 0.7761$
Least Squares Method	$MAE = 0.2598$, $R^2 = 0.8353$	$MAE = 0.2154$, $R^2 = 0.8609$
Median Filtering	$MAE = 0.3243$, $R^2 = 0.5809$	$MAE = 0.2630$, $R^2 = 0.7685$

6 DISCUSSION

A summary of the statistical characteristics of the approaches studied is presented in Table 1. The generalized quality indicators of the models are presented in Table 2. The analysis of the data in Tables 1 and 2 allows us to conclude that the application of the developed method of parameter adaptation for the Least Squares Method algorithm is not the best choice to preserve the statistical properties of the sample. Figures 2.a and 2.b show excessive smoothing of the data and, accordingly, the loss of their features, which is also demonstrated by the results presented in Table 1. However, even with such a loss of features, the use of the developed approach demonstrates the best quality of the model among the three algorithms considered the properties of time series [9]. The calculation results have proved the effectiveness of the proposed approach.

It is worth noting that the values of the dynamic and random component errors in the estimates of the controlled parameters after the applied solutions are sufficient to be no worse than the known analogues. This statement is true because the main advantage of the proposed approach is the adaptation of the parameters of the anomaly detection methods to the properties of the input sample. Therefore, the fact of preserving accuracy along with the adaptive properties of the proposed approach is evidence of achieving the goal and conditions and limitations of the pre-face part of the work.

CONCLUSIONS

The work solves the problem of developing a method for adapting the parameters of algorithms for detecting and cleaning statistical samples from anomalies for data science tasks.

The scientific novelty of the obtained results lies in the implementation of an optimization approach in minimizing the dynamic and statistical error of the model, which determines the parameters (sliding window size and sensitivity coefficient) of known algorithms for cleaning statistical samples from anomalies according to the principles of the sliding window.

The practical value of the proposed solution for Data Science tasks lies in the possibility of developing software components for cleaning data from anomalies, which are trained according to the parameters taking into account the structure and dynamics of changes in the time series. At the same time, high accuracy rates of estimation for the dynamic and stochastic components of errors are maintained. The advantage of the proposed method is also its simplicity and implementation into existing algorithms.

Prospects for further research lie in expanding the list of anomaly indicators (for example, to dynamic and influential) for multifactorial optimization of the parameters of detection algorithms.

ACKNOWLEDGEMENTS

These studies were conducted for educational purposes at the Department of Computer Engineering at the Faculty of Informatics and Computer Engineering of the National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

The results of these studies were used in scientific research project “Nonlinear and multicriterial mathematical models for Data Science and Embedded Systems technologie” (state registration number 0124U003323).

REFERENCES

1. Kumar J., Kumar A., Kumar R. Big Data and Analytics: The key concepts and practical applications of big data analytics. BPB Publications, 2024, 232 p.
2. Dietrich D., Heller B., Yang B. Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Present-

DOI 10.15588/1607-3274-2025-3-4

ing Data. Indianapolis, Indiana, John Wiley & Sons, 2015, 420 p.
3. Provost F., Fawcett T. Data Science for Business. New York: O'Reilly Media, Inc, 2013, 409 p.
4. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning Data Mining, Inference, and Prediction Second Edition. New York, Springer, 2017, 763 p.
5. Brockwell P. J., Davis R. A. Introduction to Time Series and Forecasting. New York, Springer, 2016, 439 p.
6. Tuhanskykh O., Baran D., Pysarchuk O. Method for Statistical Evaluation of Nonlinear Model Parameters in Statistical Learning Algorithms, *Proceedings of Ninth International Congress on Information and Communication Technology*. – Springer, Singapore, 2024, No. 1013, pp. 265–274. (Series "Lecture Notes in Networks and Systems"). DOI: 10.1007/978-981-97-3559-4_21
7. Nassif A. B., Talib M. A., Nasir Q., Dakalbab F. M. Machine Learning for Anomaly Detection: A Systematic Review, *IEEE Access*, 2021, No. 9, pp. 78658–78700. DOI: 10.1145/3439950.
8. Pang G., Shen C., Cao L., Hengel A. Deep Learning for Anomaly Detection, *ACM Computing Surveys*, 2021, Vol. 54(2), pp. 1–38. DOI: 10.1145/3439950.
9. Song X., Wu M., Jermaine C., Ranka S. Conditional Anomaly Detection, *IEEE Transactions on Knowledge and Data Engineering*, 2007, No. 19, pp. 631–645. DOI: 10.1109/TKDE.2007.1009.
10. Pysarchuk O., Baran D., Mironov Y., Pysarchuk I. Algorithms of statistical anomalies clearing for data science applications, *System research and information technologies.* – 2023, No. 1, pp. 78–84. DOI: 10.20535/SRIT.2308-8893.2023.1.06.
11. Mehrotra K. G., Mohan C. K., Huang H. Anomaly Detection Principles and Algorithms. Switzerland, Springer, 2017, 229 p.
12. McKinney W. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media, 2017, 550 p.
13. Nelli F. Python Data Analytics: With Pandas, NumPy, and Matplotlib, 2nd ed. Edition. Apress, 2018, 588 p.
14. Raschka S., Mirjalili V. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, Second Edition. Packt Publishing, 2017, 622 p.
15. Joshi P. Artificial Intelligence with Python: A Comprehensive Guide to Building Intelligent Apps for Python Beginners and Developers. Packt Publishing, 2017, 466 p.

Received 04.04.2025.
Accepted 29.06.2025.

УДК 004.5

СПОСІБ АДАПТАЦІЇ ПАРАМЕТРІВ АЛГОРИТМІВ ВИЯВЛЕННЯ ТА ОЧИЩЕННЯ СТАТИСТИЧНОЇ ВИБІРКИ ВІД АНОМАЛІЙ ДЛЯ ЗАДАЧ DATA SCIENCE

Писарчук О. О. – д-р техн. наук, професор, професор кафедри обчислювальної техніки, факультету інформатики та обчислювальної техніки, Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського».

Павлова С. О. – студентка факультету інформатики та обчислювальної техніки, Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського».

Баран Д. Р. – асистент кафедри обчислювальної техніки факультету інформатики та обчислювальної техніки, Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського».

АНОТАЦІЯ

Актуальність. Популяризація задачі Data Science для завдань електронної комерції, банківського сектору економіки, для задач управління динамічними об'єктами – актуалізує вимоги до показників ефективності обробки даних формату Time Series.Зазначене стосується і підготовчого етапу аналізу даних на рівні виявлення та очищення статистичних вибірок від аномалій типу грубі виміри та пропуски.

Метою роботи є розробка способу адаптації параметрів алгоритмів виявлення та очищення статистичної вибірки формату Time Series від аномалій для задач Data Science.

Метод. У статті запропоновано спосіб адаптації параметрів алгоритмів виявлення та очищення статистичної вибірки від аномалій для задач data science. Запропонований підхід базується та відрізняється від аналогічних практик запровадженням оптимізаційного підходу в мінімізації динамічної та статистичної похибки моделі, що визначає параметри налаштувань популярних алгоритмів очищення статистичної вибірки від аномалій з використанням ковзного вікна (Moving Window Method).

Результат. Запровадження запропонованого підходу в практику Data Science дозволяє розробляти програмні компонентидля очищення даних від аномалій, що навчаються за параметрами суто за структурою та динамікою Time Series. Це забезпечує підтримку широкого кола задач з нелінійними властивостями та сезонними закономірностями у даних. Отже спрощується процес супроводження подібних продуктів після впровадження їх в практику застосування.

Висновки. Ключовою перевагою запропонованого методу є його проста імплементації в існуючі алгоритми очищення вибірки від аномалій та відсутність необхідності розробнику підбирати параметри налаштувань алгоритмів очищення вручну, що економить час при розробці. Ефективність запропонованого способу підтверджується результатами розрахунків.

КЛЮЧОВІ СЛОВА: аномальні виміри, динамічна похибка, статистична похибка, оптимізація моделі, Moving Window, Data Science, Big Data, Time Series.

DOI 10.15588/1607-3274-2025-3-4

ЛІТЕРАТУРА

1. Kumar J. Big Data and Analytics: The key concepts and practical applications of big data analytics / J. Kumar, A. Kumar, R. Kumar. – BPB Publications, 2024. – 232 p.
2. Dietrich D. Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data / D. Dietrich, B. Heller, B. Yang. – Indianapolis, Indiana : John Wiley & Sons, 2015. – 420 p.
3. Provost F. Data Science for Business / F. Provost, T. Fawcett. – New York : O'Reilly Media, Inc, 2013. – 409 p.
4. Hastie T. The Elements of Statistical Learning Data Mining, Inference, and Prediction Second Edition / T. Hastie, R. Tibshirani, J. Friedman. – New York : Springer, 2017. – 763 p.
5. Brockwell P. J. Introduction to Time Series and Forecasting / P. J. Brockwell, R. A. Davis. – New York : Springer, 2016. – 439 p.
6. Tuhanskykh O. Method for Statistical Evaluation of Nonlinear Model Parameters in Statistical Learning Algorithms / O. Tuhanskykh, D. Baran, O. Pysarchuk // Proceedings of Ninth International Congress on Information and Communication Technology. – Springer, Singapore. – 2024. – 1013. – P. 265–274. (Series "Lecture Notes in Networks and Systems"). DOI: 10.1007/978-981-97-3559-4_21
7. Machine Learning for Anomaly Detection: A Systematic Review / [A. B. Nassif, M. A. Talib, Q. Nasir, F. M. Dakalbab] // IEEE Access. – 2021. – No. 9. – P: 78658–78700. DOI: 10.1145/3439950.
8. Deep Learning for Anomaly Detection: / [G. Pang, C. Shen, L. Cao, A. Hengel] // ACM Computing Surveys. – 2021. – No. 54(2). – P. 1–38. DOI: 10.1145/3439950.
9. Conditional Anomaly Detection [X. Song, M. Wu, C. Jermaine, S Ranka] // IEEE Transactions on Knowledge and Data Engineering. – 2007. – No. 19. – P. 631–645. DOI: 10.1109/TKDE.2007.1009.
10. Algorithms of statistical anomalies clearing for data science applications / [O. Pysarchuk, D. Baran, Y. Mironov, I. Pysarchuk] // System research and information technologies. – 2023. – No. 1. – P. 78–84. DOI: 10.20535/SRIT.2308-8893.2023.1.06.
11. Mehrotra K. G. Anomaly Detection Principles and Algorithms / K. G. Mehrotra, C. K. Mohan, H. Huang. – Switzerland : Springer, 2017. – 229 p.
12. McKinney W. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython / W. McKinney. – O'Reilly Media, 2017. – 550 p.
13. Nelli F. Python Data Analytics: With Pandas, NumPy, and Matplotlib, 2nd ed. Edition / F. Nelli. – Apress, 2018. – 588 p.
14. Raschka S. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, Second Edition / S. Raschka, V. Mirjalili. – Packt Publishing, 2017. – 622 p.
15. Joshi P. Artificial Intelligence with Python: A Comprehensive Guide to Building Intelligent Apps for Python Beginners and Developers / P. Joshi. – Packt Publishing, 2017. – 466 p.

DOI 10.15588/1607-3274-2025-3-4