UDC 004.5
DOI: 10.20535/SRIT.2308-8893.2023.1.06

ALGORITHMS OF STATISTICAL ANOMALIES CLEARING FOR DATA SCIENCE APPLICATIONS

O. PYSARCHUK, D. BARAN, Yu. MIRONOV, I. PYSARCHUK

Abstract. The paper considers the nature of input data used by Data Science algorithms of modern-day application domains. It then proposes three algorithms designed to remove statistical anomalies from datasets as a part of the Data Science pipeline. The main advantages of given algorithms are their relative simplicity and a small number of configurable parameters. Parameters are determined by machine learning with respect to the properties of input data. These algorithms are flexible and have no strict dependency on the nature and origin of data. The efficiency of the proposed approaches is verified with a modeling experiment conducted using algorithms implemented in Python. The results are illustrated with plots built using raw and processed datasets. The algorithms application is analyzed, and results are compared.

Keywords: anomaly removal, anomaly detection, noise removal, statistical techniques, data analysis, big data, data cleaning.

INTRODUCTION

The Data Science techniques and approaches are widely used to solve modern day technical problems. One of frequent use-cases is implementing Back-End software components for distributed CRM and ERP with intellectual features. A possible format of input data for such systems is Big Data arrays that could be interpreted as numeric statistical samples. This kind of data is used in a variety of application domains, such as Commodity Trade and Risk Management systems, automotive applications like automated driver assistants, Unmanned Aerial Vehicle software, Computer Vision software responsible for raster-to-vector conversion et cetera [1–6].

In solutions from the aforementioned application domains, the underlying mathematical models are a priory known. Therefore, dataset processing is reduced to mapping data features to model properties. This allows to determine the trend line of a considered process both within observation range and beyond it (by means of interpolation and extrapolations – retrospective and perspective prognoses) [2–4]. This paper suggests treating the process of analytical model configuring as an unsupervised learning activity, since any solution of Artificial Intelligence problem, in its first iteration, is implemented as a process of parameters configuration with respect to input dataset [2–4]. Moreover, such models could be associated with artificial intelligence technologies because of the aforementioned and because of their prognosis features.

Statistical dataset processing is based on the hypothesis of accidental factors model that result in input dataset bias. Statistical data are usually obtained through experiments and sampling.

The first stage of step-by-step statistical analysis includes a preparatory stage — data research [2–4, 7]. Tasks of this stage include checking input Big Data array for anomalies. Any value that has a significant difference from the other dataset values could be treated as anomaly. Abnormal data is a result of various miscalculations and malfunctions during the data sampling. Two major types of anomalies are value skips and crude measurements. Both of these anomaly types, if not handled during next stages, will result in distorted results of statistical analysis. The distortions are represented with increased dispersion of scoring and/or with bias of scoring result [2–4, 7]. Possible ways to address the problem of distorted data are smoothing or evaluation.

In order to reduce influence of anomalies on further data processing, it is necessary to detect abnormalities and to restore or remove abnormal values. These actions could be translated to a tuple of stages that will represent the entire process of clearing dataset from abnormal values [3, 4, 7–10].

Anomaly detection is based on their analyzing properties with respect to absolute values, trend dynamics and statistical properties change [3, 4, 7, 9–10].

Related works. A significant amount of known approaches to anomaly clearing has been considered [7, 8]. They possess a common flaw – it is necessary to manually tune parameters with respect to input dataset and anomaly properties. Moreover, the majority of them will result in NP problems when applied to Big Data array. This makes their application impossible.

Therefore, the task of designing simple and efficient data clearing algorithms for Data Science purposes remains relevant.

Goal. The goal of the paper is proposing precise, performant, efficient and relatively simple algorithms for clearing datasets from anomalies.

Proposal. Experience of multiple IT proejcts related to Data Science allows to formulate practical requirements to algorithms responsible for anomaly detection and clearing. They include:

1. High efficiency of detection, integrity and adequacy of output, precise values of dispersion and standard deviation.

2. Relatively low computing complexity, therefore high performance with large datasets.

3. Automatic adjustment of parameters with respect to the input data properties, or minimal amount of manual settings;

Three algorithms of anomaly detection and removal have been developed to address these requirements.

Data preprocessing algorithm – the “sliding_wind” algorithm

The main idea of the algorithm is application of smoothing within a trivial sliding window. Within the window, an arithmetic mean value is calculated. The size of the sliding window is calculated in accordance to the demand of quasi-linear trend that describes the change considered process withing the window.

Sliding_wind algorithm stages:

1. Considering statistical sample $X = \{x_i\}$, $i = 1...n$ as an input.

2. Formulating *Nwin*-dimensional sliding window.

3. Calculating an arithmetic mean:

$$j = 1,...,N_{win}.$$

4. Formulating a clean dataset $\hat{X} = \{\hat{x}_i\}$, $i = 1...n$ by replacing elements of X sample with arithmetic mean $\hat{x}_j$: $x_{i=Nwin} = \hat{x}_j$, starting from the last entry within sliding window.

5. Shifting the sliding window to the next dataset dimension to the right – to $i = i + 1$.

6. Repeating steps 2–5 within input statistical sample $i = 1...n$.

7. After processing the entire dataset, in order to adjust data within the first sliding window, a subset *N=2Nwin* is created. After that, steps 2–5 are applied to this subset, traversing it in reverse.

The advantages of sliding_wind algorithm are its simplicity, minimum number of manual settings (only window size), potential for effective usage on datasets with fuzzy trend properties and anomalies and equal sizes of input and output.

It could be possible to include nonlinear estimation model for $\hat{x}_j$ inside sliding_wind algorithm, but it is inadvisable due to major increase in complexity. This would especially affect Big Data inputs.

Results of modeling and efficiency estimation of sliding_wind algorithm. For the modeling, a *n* = 10000 dataset has been considered. Quadratic trend and normal distribution are present in the dataset. Normal distribution has expected value of 0 and standard deviation $\sigma_X = 5.0$. Abnormal entries are evenly distributed within selection and constitute 10% of values. The model of trend and stochastic components is additive. Computations are conducted using Python implementation of algorithm [5, 6] using features of numpy and matplotlib libraries.

Results of sliding_wind execution are provided in Fig. 1.

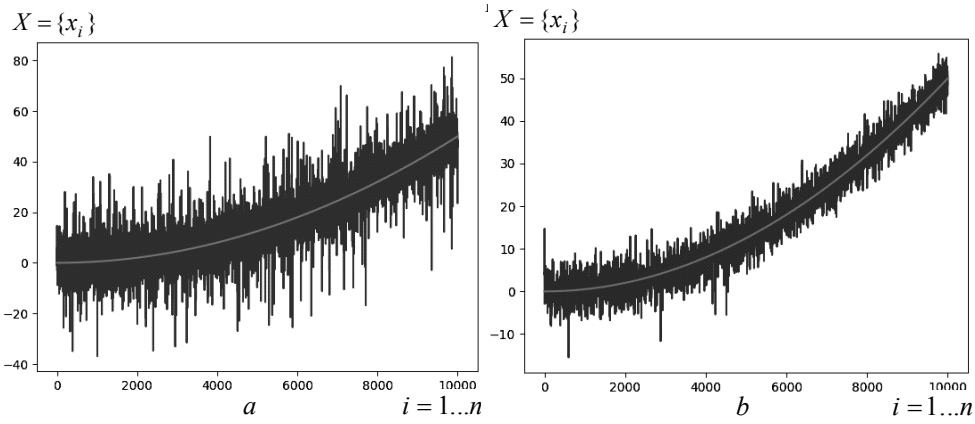

Fig. 1. Results of sliding_wind execution: *a* — input dataset, $\sigma_X =$ 6.64; *b* — processed dataset, $\sigma_X = 2.95$

Fig. 1, *a* represents a plot built from input dataset $X = \{x_i\}$, $i = 1,...,n$ with normal noise and anomalies, as well as the trend line obtained via least squares method [3, 4]. Standard deviation of input dataset with anomalies is $\sigma_X = 6.64$. Fig. 1, *b* represents a plot built from dataset cleared from anomalies by means of sliding_wind algorithm. Standard deviation of such a dataset is $\sigma_X = 2.95$. Comparing plots 1, *a* and 1, *b*, as well as reduced value of standard deviation indicates efficiency of the sliding_wind algorithm.

Algorithm of dataset statistic properties control – the "medium" algorithm

The main idea of the algorithm is detecting abnormal elements by determining etalon parameters of dataset (expected value, standard deviation) without anomalies and comparing them with initial statistical features. Detected anomalies are replaced with arithmetic mean. Etalon and initial statistical features are derived from datasets that formulate initial initial and current sliding windows respectively.

Stages of medium algorithm:

1. Considering statistical sample $X = \{x_i\},\ i = 1,...n$.

2. Within the first sliding window with size of *Nwin* dimensions, estimation of etalon expected value, standard deviation and dispersion is conducted:

$$\hat{x}_j^{\text{etalon}} = \frac{1}{Nwin}\sum_{j=1}^{Nwin} x_j\ ,\ \hat{D}_{\text{j}}^{\text{etalon}} = \frac{1}{Nwin-1}\sum_{j=1}^{Nwin}(\hat{x}_j - x_j)^2\ ,$$

$$\hat{\sigma}_{\text{j}}^{\text{etalon}} = \sqrt{\hat{D}_j}\ ,\ j = 1...Nwin.$$

3. Formulating the next sliding window with size of *Nwin* dimensions is conducted.

4. Determining expected value, dispersion and standard deviation for sliding windows:

$$\hat{x}_j = \frac{1}{Nwin}\sum_{j=1}^{Nwin} x_j\ ,\ \hat{D}_j = \frac{1}{Nwin-1}\sum_{j=1}^{Nwin}(\hat{x}_j - x_j)^2\ ,\ \hat{\sigma}_j = \sqrt{\hat{D}_j}\ ,\ j = 1...Nwin.$$

5. If the following condition is true

$$\hat{\sigma}_j > Q\hat{\sigma}_{\text{j}}^{\text{etalon}},$$

(Q — weighting factor), then current j-th dimension, added to sliding window is considered to be abnormal.

Abnormal j-th dimension is replaced with estimated expected value of current sliding window $x_i = \hat{x}_j$ inside $X = \{x_i\}$ dataset.

If condition $\hat{\sigma}_j > Q\hat{\sigma}_j^{\text{etalon}}$ is false, then the $X = \{x_i\}$ dataset is not modified.

6. Repeating steps 3–5 within statistical sample of $i = 1...n$.

7. The result of algorithm is dataset $\hat{X} = \{\hat{x}_i\},\ i = 1...n$ clear of anomalies.

The main advantages of suggested algorithm are simplicity of implementation, statistical approach to determining abnormality of values, immutability of dataset when no anomalies detected, keeping the same size of dataset after processing.

One peculiarity of the algorithm is that etalon dataset should contain no anomalies. Also, Q weighting factor is determined proportionally to the size of abnormality bias.

Results of modeling and efficiency estimation of medium algorithm. Modeling conditions are the same as for sliding_wind algorithm. Research results are provided on Fig. 2, in equivalence to Fig. 1.

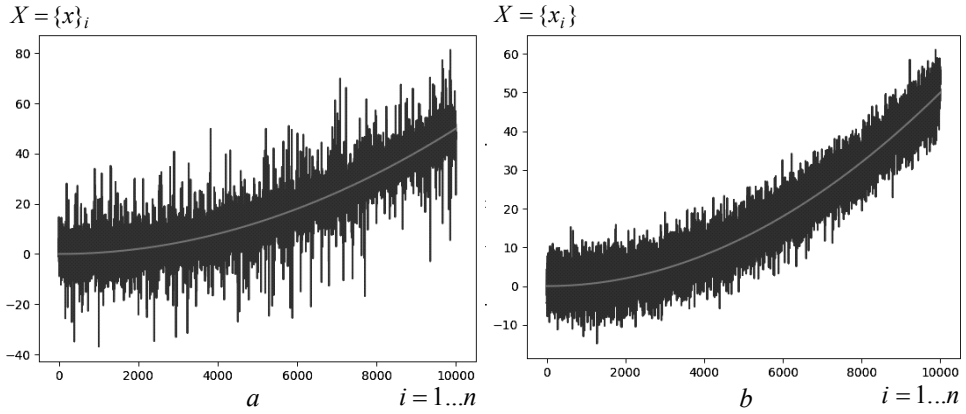

Fig. 2. Results of medium algorithm execution: *a* — input dataset $\sigma_X = 6.64$; *b* — rocessed dataset $\sigma_X = 4.45$

Comparing plots depicted at Fig. 2, *a* and 2, *b*, as well as standard deviation reduction ($\sigma_X = 4.45$) tells of medium algorithm efficiency.

Algorithm of control over dataset dynamic properties change — the LSM algorithm

The idea of the algorithm is in detecting abnormal values by determining etalon parameters of trend dynamic change speed, and comparing them to current properties using Least Squares Method. Etalon values of speed are calculated according to input dataset, and current values – according to sliding window selections. During comparison, the statistical estimation errors are respected.

LSM algorithm steps:

1. Considering input dataset $X = \{x_i\}$, $i = 1...n$.

2. Using polynomial model LSM, the etalon speed is calculated within dataset:

$Speed^{\text{etalon}} = c_2$, with LSM polynomial $T(x) = c_0 + c_1 x + c_2 x^2 + ...$

3. Formulating sliding window with $Nwin$ dimensions.

4. Determining current estimation of expected value, dispersion and standard error for the sliding windows:

$$\hat{x}_j = \frac{1}{Nwin}\sum_{j=1}^{Nwin} x_j\ ,\ \hat{D}_j = \frac{1}{Nwin-1}\sum_{j=1}^{Nwin}(\hat{x}_j - x_j)^2\ ,\ \hat{\sigma}_j = \sqrt{\hat{D}_j}\ ,\ j = 1...Nwin.$$

5. Determining controlled parameters for abnormality that are scaled up to size of the datasets with scores and respecting dimension errors:

$$\text{Ind\_1} = \left|Speed^{\text{etalon}} \cdot \sqrt{n}\right|,\ \text{Ind\_2} = \left|Q \cdot \hat{\sigma}_j \cdot Speed^{\text{etalon}} \cdot \sqrt{Nwin}\right|,$$

(Q — setting weighting factor).

6. If condition is true

$$\text{Ind\_2} > \text{Ind\_1}\ ,$$

then current j-th dimension added to sliding window is treated as anomaly.

Abnormal j-th dimension is replaced with LSM score $x_i = T(x = \hat{x}_j)$ inside dataset $X = \{x_i\}$.

If condition $\text{Ind\_2} > \text{Ind\_1}$ is false, $X = \{x_i\}$ dataset remains immutable.

6. Repeating steps 3–5 within input dataset $i = 1...n$.

7. The result of algorithm is dataset $\hat{X} = \{\hat{x}_i\},\ i = 1...n$ clear of anomalies.

The main advantages of suggested algorithm are simplicity of implementation, dynamic criteria of abnormality, modification of abnormal elements only, same size of input and result datasets.

One peculiarity of algorithm is that it does not require parameter control when determining etalon values; the weighting factor Q is determined proportionally to the size of abnormality bias.

Results of modeling and efficiency estimation of LSM algorithm. Modeling conditions are the same as for sliding_wind algorithm. Results are depicted at Fig. 3. Equivalent to Fig. 1.

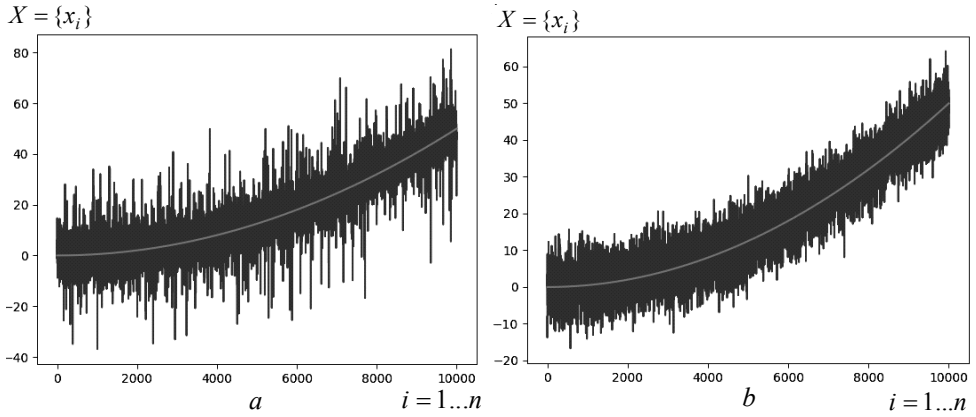

Fig. 3. Results on MNK algorithm execution: *a* — raw dataset $\sigma_X =$ 6.64; *b* — processed dataset $\sigma_X = 4.71$

Comparison of plots depicted at Fig. 3, *a* and 3, *b*, as well decreasing standard deviation ($\sigma_X = 4.7$) illustrates algorithm efficienccy.

Generalized statistical properties and reduction of bias with three proposed algorithms are depicted at table.

Generalized statistical properties of three proposed algorithms

Raw data with abnormalities	**sliding_wind**	**medium**	**MNK**
$\sigma_X = 6.64$	$\sigma_X = 2.95$	$\sigma_X = 4.45$	$\sigma_X = 4.71$

Data from table demonstrate that sliding_wind algorithm is the most successful precision-wise — it has the least standard deviation. This algorithm is also the most productive and has no need for configuration. However, it may be unapplicable to dynamic strongly nonlinear processes — its usage may result in biases. Medium and MNK algorithms are the opposites to sliding_wind. In order to choose specific algorithm, dataset properties and anomaly features (such as margin of error or the nature of the trend model) have to be considered. Comparing proposed methods with more sophisticated alternatives gave no results, since using data of equivalent complexity (*N*=10000) yielded no results from these alternatives in sensible time.

Conclusions. The proposed algorithms are not expensive to implement, they do not require major manual tuning and show acceptable precision and performance for Big Data arrays processing.

REFERENCES

1. F. Provost and T. Fawcett, *Data Science for Business*. USA: O'Reilly Media, Inc, 2013, 409 p.
2. D. Dietrich, *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Indianapolis, Indiana, USA: John Wiley & Sons, Inc, 2015, 420 p. doi: 10.1002/9781119183686.ch1.
3. O. Pysarchuk and V. Kharchenko, *Nonlinear multi-criteria process modeling in traffic management systems*, (in Ukrainian). Kyiv: Institute of Gifted Child, 2015, 248 p.
4. S. Kovbasiuk, O. Pysarchuk, and M. Rakushev, *Least Squares Method and its practical applications*, (in Ukrainian). Zhytomyr: Zhytomyr Military Institute, 2008, 228 p.
5. S. Raschka, Y. Liu, and V. Mirjalili, *Machine Learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python*. Birmingham: Packt, 2022.
6. P. Joshi, *Artificial Intelligence with Python*. Birmingham: Packt, 2017.
7. G. Kishan, K. Chilukuri, and H. HuaMing, *Anomaly Detection Principles and Algorithms*. Switzerland, Springer, 2017, 229 p. doi: 10.1007/978-3-319-67526-8.
8. O. Pysarchuk and Y. Mironov, *Chromosome Feature Extraction and Ideogram-Powered Chromosome Categorization*. Switzerland, Springer, 2022. doi: 10.1007/978-3-031-04812-8_36.
9. H. Blomquist and J. Möller, *Anomaly detection with Machine learning. Quality assurance of statistical data in the Aid community*. Uppsala: Uppsala University, 2015, 60 p.
10. S. Thudumu, P. Branch, J. Jin, and J. Singh, *A comprehensive survey of anomaly detection techniques for high dimensional big data*. Switzerland, Springer, 2017, 30 p. doi: 10.1186/s40537-020-00320-x.

Received 10.08.2022

INFORMATION THE ARTICLE

Oleksii O. Pysarchuk, ORCID: 0000-0001-5271-0248, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, e-mail: PlatinumPA2212@gmail.com

Danylo R. Baran, ORCID: 0000-0002-3251-8897, "Codeimpact B.V", Ukraine, e-mail: danil.baran15@gmail.com

Yurii G. Mironov, ORCID: 0000-0002-2291-5864, National Aviation University, Ukraine, e-mail: yuriymironov96@gmail.com

Illya O. Pysarchuk, ORCID: 0000-0003-4343-0142, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, e-mail: flimka134@gmail.com

АЛГОРИТМИ ОЧИЩЕННЯ СТАТИСТИЧНОЇ ВИБІРКИ ВІД АНОМАЛІЙ ДЛЯ ЗАДАЧ DATA SCIENCE / О.О. Писарчук, Д.Р. Баран, Ю.Г. Міронов, І.О. Писарчук

Анотація. Розглянуто природу даних, що використовуються в задачах сучасних прикладних областей. Запропоновано декілька алгоритмів очищення статистичної вибірки від аномалій в конвеєрі задач Data Science. Відзнакою та перевагою запропонованих алгоритмів є їх відносна простота та обмежена кількість параметрів налаштувань, що визначаються за технологіями навчання відповідно до властивостей вхідних статистичних даних. Запропоновані алгоритми є достатньо гнучкими у використанні і не залежать від природи та походження даних. Результати модельного експерименту запропонованих підходів у вигляді скриптів мовою Python та базових бібліотек довели їх ефективність. Результати проілюстровано графіками, побудованими з використанням початкових даних та даних, що змінені за допомогою запропонованих алгоритмів. Застосування алгоритмів проаналізовано та порівняно результати виконання алгоритмів.

Ключові слова: очищення від аномалій, виявлення аномалій, видалення шуму, статистичні методи, аналіз даних, великі дані, очищення даних.