Stream Data Preprocessing: Outlier Detection Based on the Chebyshev Inequality with Applications

Georgy Shevlyakov
Peter the Great St. Petersburg Polytechnic University
Institute of Applied Astronomy of Russian Academy of Sciences
Institute of Problems of Mechanical Engineering
of Russian Academy of Sciences
St. Petersburg, Russia
Georgy.Shevlyakov@phmf.spbstu.ru

Margarita Kan
Institute of Applied Astronomy
of Russian Academy of Sciences
Peter the Great St. Petersburg Polytechnic University
St. Petersburg, Russia
mo.kan@iaaras.ru

***Abstract*—A novel method of outlier detection for preprocessing stream data in the conditions of uncertainty when the data means and standard errors as the only available summary information about the initial data statistics is proposed. As neither the initial data samples nor their sample sizes are known, the classical methods of outlier detection, including nonparametric methods of statistics, cannot be applied in this case. The principal idea of the proposed approach to outlier detection is based on the use of the classical Gauss-Chebyshev type probability inequalities—the corresponding confidence intervals constructed on these inequalities allow to set up the problems of hypotheses testing similar to the classical settings as the problems of minimizing the upper bound of the Bayesian risk and maximizing the lower bound of the test power in the Neyman-Pearson sense. The results of the processing of the real-life data (Lunar Laser Ranging data) and the model data manifest unexpectedly good outlier detection performance.**

I. Introduction

In the framework of data mining and machine learning, the problem of data preprocessing is vitally important for the well-known saying "garbage in, garbage out". Data preprocessing includes a number of stages: cleaning, instance selection, normalization, transformation, feature extraction and selection, etc. with its product as the final training set [1], [2].

At the cleaning stage, outlier detection is primary. However, there is no a satisfactory definition of an outlier or an anomaly in the data. One of the common definitions is the following [3], [4]: "An outlying observation, or outlier, is one that deviates markedly from other members of the sample in which it occurs." Generally, this definition is neither mathematically or statistically correct (for details, see [5]).

Outliers in the data can be explained by data input errors, experiment conducting errors, measurement errors, mixing data from various sources, and unaccounted features of sample. For univariate data, outliers usually have high magnitudes.

In present, there does not exist a general method of outlier detection: as a rule, outlier detection methods and algorithms depend on the various purposes of studying, distribution models and data types [4], [6], [7], [8].

Existing methods of outlier detection may be classified by two groups: methods using only initial data (data-based) and methods using the information about data distribution laws (model-based). For example, the outlier detection method based on the "three sigma"-rule is oriented on the normal data distribution since only the 0.0027 fraction of data values is observed out of the "three sigma" boundaries in this case. Also, Grubb's method [3] is also related to the model-based group. On the other hand, Tukey's boxplot [9] and the method of k nearest neighbors are oriented only on the data.

The method of outlier detection proposed in this work, generally, belongs to the group of model-based methods, in which the lengths of confidence intervals used in this method are constructed basing on the underlying data distribution.

The main idea of this method can be described as follows. First, the available data for processing are presented as the pairs of means and their standard errors, so, they are the summaries of the initial unavailable data samples including their sizes— to the best of our knowledge, neither of the existing methods of outlier detection can be applicable in this case. It is worth noting that this form of data representation is common for physics and for other natural and technical sciences.

Second, the opportunity of using the whole information contained in data pairs (mean value and its standard error) is given by the classical tools in the form of Gauss-Chebyshev type probability inequalities [10], [11]. The corresponding confidence intervals constructed on these inequalities allow to set up the problems of hypotheses testing as the problems of minimizing the upper bound of the Bayesian risk and maximizing the lower bound of the test power in the Neyman-Pearson sense.

An outline of the remainder of the paper is as follows. In Section II, the Gauss-Chebyshev type probability inequalities are briefly reviewed. In Section III, problem settings for outlier detection are given. In Section IV, main results are formulated. In Section V, real-life and simulated data are processed by the proposed methods. In Section VI, some conclusions are drawn.

II. Gauss – Chebyshev type inequalities

The first in this list of results is the Gauss inequality [10]. Let X be a unimodal random variable with mode m and $\tau^2 = E(X-m)^2$. Next, $\tau^2 = (\mu-m)^2+\sigma^2$, where $\mu = E(X)$ and

$\sigma = \sqrt{D(X)}$. Then for any positive k the following inequality holds:

$$P[|X - m| \geq k] \leq \begin{cases} \left(\frac{2\tau}{3k}\right)^2 & \text{for} \quad k \geq \frac{2\tau}{\sqrt{3}}, \\ 1 - \frac{k}{\tau\sqrt{3}} & \text{for} \quad 0 \leq k \leq \frac{2\tau}{\sqrt{3}}. \end{cases} \tag{1}$$

Example 1 Consider the standard normal distribution with $\mu = m = 0$, $\sigma = 1$ and $k = 3$. Then inequality (1) yields

$$P[|X - m| \geq k] \leq (2/9)^2 = 4/81 \approx 0.05.$$

Secondly, we consider the well-known Chebyshev inequality [11]. Let X be a random variable with mean μ and variance σ^2. Then for any $k \geq 1$ the we have:

$$P[|X - \mu| \geq k\sigma] \leq 1/k^2. \tag{2}$$

For the Chebyshev inequality, the interval is constructed with the center at the mean value and without the condition of unimodality.

Example 2 Consider again the standard normal distribution $X \sim N(0, 1)$ with $k = 3$. In this case inequality (2) yields the "three sigma" rule

$$P[|X - \mu| \geq 3\sigma] \leq 1/9 \approx 0.11.$$

Finally, we consider the Vysochanskij – Petunin inequality [12]. Let X be a unimodal random variable with mean μ and variance σ^2. Then for any $k > \sqrt{8/3}$ we get:

$$P[|X - \mu| \geq k\sigma] \leq 4/9k^2. \tag{3}$$

The Vysochanskij – Petunin inequality is similar to the Gauss inequality with the only difference that the center of the interval is at the mean value.

Example 3 Consider the standard normal distribution $X \sim N(0, 1)$ with different values of k; inequality (3) yields the following bounds:

1) for $k = \sqrt{8/3}$, it is $1/2$ — the limit case,

2) for $k = 2$, it is $1/9$ like in the Chebyshev case,

3) for $k = 3$, it is $4/81$ like in the Gauss case.

III. Problem setting

A. Data presentation form

The data for processing are given in the form of stream data as mean values with their standard errors at time instants $t_0, t_1, \ldots, t_n, \ldots$

$$(\bar{x}_0, s_0), (\bar{x}_1, s_1), \ldots, (\bar{x}_n, s_n), \tag{4}$$

where

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \quad i = 0, 1, \ldots, n,$$

$$s_i = \frac{1}{\sqrt{n_i}} \left(\frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \right)^{1/2}.$$

Here we underline that the initial data samples $\{x_{0j}\}_1^{n_0}, \{x_{1j}\}_1^{n_1}, \ldots, \{x_{ij}\}_1^{n_i}$ together with their sample sizes are unavailable, only summary data in the form (4) is available; that is why, the classical methods of outlier detection cannot be used in this case. Moreover, we repeat that this form of data presentation is common for natural and technical sciences.

In what follows, we apply the classical inequalities of the Gauss – Chebyshev type for revealing the statistically significant outliers (shifts) in the data. Recall that the aforementioned inequalities estimate the probabilities of the deviations of a random variable X from its mean values (in different senses) via its distribution moment values.

B. Confidence intervals based on the Chebyshev inequality

The available information on the real-life statistical data as the mean and its standard error values $(\bar{x}_0, s_0), (\bar{x}_1, s_1), \ldots, (\bar{x}_n, s_n)$ can be represented by applying the Chebyshev inequality as the sequence of confidence intervals $\Delta_0(k_0), \Delta_1(k_1), \ldots, \Delta_n(k_n)$ for random variables $\bar{X}_0, \bar{X}_1, \ldots, \bar{X}_n$:

$$\begin{aligned} \Delta_0(k_0) &= (\bar{x}_0 - k_0 s_0, \bar{x}_0 + k_0 s_0), \\ \Delta_1(k_1) &= (\bar{x}_1 - k_1 s_1, \bar{x}_1 + k_1 s_1), \\ &\ldots\ldots\ldots\ldots\ldots, \\ \Delta_n(k_n) &= (\bar{x}_n - k_n s_n, \bar{x}_n + k_n s_n). \end{aligned} \tag{5}$$

From (5) it follows that the probabilities of belonging of random variables $\bar{X}_0, \bar{X}_1, \ldots, \bar{X}_n$ to the intervals $\Delta_0(k_0), \Delta_1(k_1), \ldots, \Delta_n(k_n)$ have the form:

$$\begin{aligned} P[\bar{X}_0 \in \Delta_0(k_0)] &= P[\bar{x}_0 - k_0 s_0 \leq \bar{X}_0 \leq \bar{x}_0 + k_0 s_0] \\ &= P[|\bar{X}_0 - \bar{x}_0| \leq k_0 s_0] \geq 1 - \frac{s_0^2}{k_0^2 s_0^2} = 1 - \frac{1}{k_0^2}, \\ P[|\bar{X}_1 - \bar{x}_1| \leq k_1 s_1] &\geq 1 - \frac{s_1^2}{k_1^2 s_1^2} = 1 - \frac{1}{k_1^2}, \\ &\ldots\ldots\ldots\ldots\ldots, \\ P[|\bar{X}_n - \bar{x}_n| \leq k_n s_n] &\geq 1 - \frac{s_n^2}{k_n^2 s_n^2} = 1 - \frac{1}{k_n^2}, \end{aligned} \tag{6}$$

where $k_0, k_1, \ldots, k_n$ are some positive values.

Equations (6) yield the minimum values for the confidence probabilities of the localization of random variables $\bar{X}_0, \bar{X}_1, \ldots, \bar{X}_n$ in the confidence intervals $\Delta_0(k_0), \Delta_1(k_1), \ldots, \Delta_n(k_n)$. This situation is illustrated by Fig. 1 for confidence intervals $\Delta_0(k_0)$, ..., $\Delta_($(k_4)$ for different variants of their relative location.

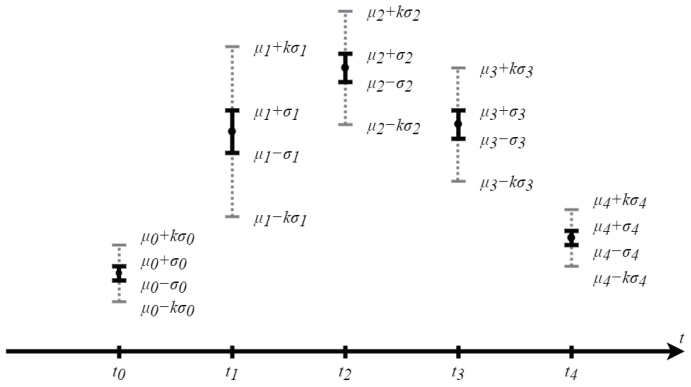

Fig. 1. Observations and their confidence intervals

C. Minimizing the upper bound of Bayesian risk

Consider the problem of outlier detection as the problem of hypotheses testing

$$H_0 : X = \bar{X}_0 \in \Delta_0(k_0) \text{ versus } H_1 : X = \bar{X}_1 \in \Delta_1(k_1). \quad (7)$$

Denote the prior probabilities of hypotheses as $P[H_0] = p_0$, $P[H_1] = p_1$, $p_0 + p_1 = 1$. Next, we introduce the function of Bayesian risk in the form

$$R(k_0, k_1) = c_0 p_0 P[H_1|H_0] + c_1 p_1 P[H_0|H_1], \quad (8)$$

where $P[H_1|H_0]$ is the probability of type I error (the alternative H_1 is accepted when the null hypothesis H_0) is true), $P[H_0|H_1]$ is the probability of type II error (the the null hypothesis H_0 is accepted when the alternative H_1 is true), $c_0, c_1 > 0$ are given error costs.

Now we rewrite Equation (8) in the following form

$$R(k_0, k_1) = c_0 p_0 \alpha + c_1 p_1 \beta, \quad (9)$$

where

$$\alpha = P[H_1|H_0] = P[\bar{X}_0 \in \Delta_1(k_1)]$$
$$= P[\bar{x}_1 - k_1 s_1 \le \bar{X}_0 \le \bar{x}_1 + k_1 s_1],$$
$$\beta = P[H_0|H_1] = P[\bar{X}_1 \in \Delta_0(k_0)]$$
$$= P[\bar{x}_0 - k_0 s_0 \le \bar{X}_1 \le \bar{x}_0 + k_0 s_0].$$

The error probabilities α and β can be expressed via their complementary probabilities:

$$\alpha = P[H_1|H_0] = 1 - P[H_1|H_1]$$
$$= 1 - P[\bar{x}_1 - k_1 s_1 \le \bar{X}_1 \le \bar{x}_1 + k_1 s_1],$$
$$\beta = P[H_0|H_1] = 1 - P[H_0|H_0]$$
$$= 1 - P[\bar{x}_0 - k_0 s_0 \le \bar{X}_0 \le \bar{x}_0 + k_0 s_0].$$

Now we apply the Chebyshev inequality for estimating the bounds of the corresponding probabilities:

$$P[H_0|H_0] = P[\bar{X}_0 \in \Delta_0(k_0)] \ge 1 - \frac{1}{k_0^2},$$

$$P[H_1|H_1] = P[\bar{X}_1 \in \Delta_1(k_1)] \ge 1 - \frac{1}{k_1^2}.$$

Further, we get

$$\alpha \le \frac{1}{k_0^2}, \quad \beta \le \frac{1}{k_1^2},$$

and the corresponding upper bound for the Bayesian risk

$$R(k_0, k_1) = c_0 p_0 \alpha + c_1 p_1 \beta$$

$$\le \bar{R}(k_0, k_1) = \frac{c_0 p_0}{k_0^2} + \frac{c_1 p_1}{k_1^2}.$$

The unconstrained minimization of the upper bound $\bar{R}(k_0, k_1)$ is obviously senseless as its minimum is attained in the limit case when $k_0, k_1 \to \infty$, hence some upper constraints upon the values of the parameters k_0, k_1 (explicit or implicit), that is, upon the lengths of the corresponding confidence intervals, should be imposed. It is natural to consider the problem of minimization of $\bar{R}(k_0, k_1)$ under the side condition of the nonintersecting confidence intervals $\Delta_0(k_0)$ and $\Delta_1(k_1)$ (see Fig. 1)

$$\bar{R}(k_0, k_1) \to \min_{k_0, k_1}, \quad (10)$$

$$\bar{x}_1 - k_1 s_1 \ge \bar{x}_0 + k_0 s_0. \quad (11)$$

The solution of problem (10) has sense only with the sufficiently small probabilities of type I and II errors, that is, with sufficiently large values of the parameters k_0, k_1. These conditions are also taken into account in the Neyman-Pearson problem setting.

D. Neyman-Pearson hypotheses testing problem setting

Consider the hypotheses testing problem (7) with the corresponding Neyman-Pearson test

$$P_D = P[H_1|H_1] = P[\bar{X}_1 \in \Delta_1(k_1)]$$

$$\ge 1 - 1/k_1^2 \to \max_{k_1}, \quad (12)$$

$$\alpha = P[H_1|H_0] \le \bar{\alpha} = 1/k_0^2, \quad (13)$$

$$\bar{x}_1 - k_1 s_1 \ge \bar{x}_0 + k_0 s_0$$

as the problem of maximizing the lower bound of the power of test (the probability of true detection) under the bounded probability of type I error and under the side condition of the nonintersecting confidence intervals $\Delta_0(k_0)$ and $\Delta_1(k_1)$ (11) (see Fig. 1).

IV. Main results

The problem settings considered in Section III imply the following results.

A. Precise results

Theorem 1 The optimal test for outlier detection minimizing the upper bound of the Bayesian risk is given by the solution of the optimization problem (10) under the side condition (11).

In this case, the optimal test is represented by the optimal lengths of the non-intersecting confidence intervals $\Delta_0(k_0^*)$ and $\Delta_1(k_1^*)$ with the corresponding probabilities of type I and II errors satisfying the following inequalities

$$\alpha \le \bar{\alpha} = \frac{1}{k_0^{*2}}, \quad \beta \le \bar{\beta} = \frac{1}{k_1^{*2}}. \tag{14}$$

Theorem 2 The optimal test for outlier detection maximizing the lower bound of of the power of test is given by the solution of the optimization problem (12) under the side conditions (11) and (13).

These results hold under the conditions of the validness of the Chebyshev inequality, namely, for data distributions with bounded first two moments.

B. A low-complexity algorithm for the approximate solution of the constrained optimization problem

Consider the simplification of optimization problem (10) with setting $k_0 = k_1 = k$:

$$\bar{R}(k) = \frac{c_0 p_0 + c_1 p_1}{k^2} \to \min \tag{15}$$

under the side condition

$$\bar{x}_1 - k s_1 \ge \bar{x}_0 + k s_0.$$

The optimal value of k is given by

$$k^* = \frac{|\bar{x}_1 - \bar{x}_0|}{s_0 + s_1}. \tag{16}$$

Similarly to problem (10), the solution (16) has sense only with sufficiently small values of the probabilities of type I and II errors

$$\alpha \le \frac{1}{k^{*2}}, \quad \beta \le \frac{1}{k^{*2}}, \tag{17}$$

that is, with sufficiently large values of the parameter k^*, say, $k^* > \bar{k}$. Thus, if $k^* > \bar{k}$ then the hypotheses H_0 and H_1 significantly differ from each other: an outlier (or an anomaly) is observed with the transition from the point $(\bar{x}_0, s_0)$ to the point $(\bar{x}_1, s_1)$.

In our study, the threshold $\bar{k}$ is set equal to 2 with the corresponding upper bound upon the probabilities of type I and II errors equal to $1/4$: $\alpha, \beta \le 1/4$ as it follows from the Chebyshev inequality (17). Further in Section V, while processing the real-life and model data, we show that real probabilities of type I and II errors are far smaller than their rather pessimistic upper bounds.

Now we describe our algorithm of outlier detection in a stepwise way.

1) Consider two adjacent observations: i-th and $(i+1)$-th, where $i = \overline{1,\ N-1}$.
2) For this pair of observations, we find the corresponding $k^\star$-value according to (16):
$$k_i^\star = \frac{|\mu_1 - \mu_0|}{\sigma_0 + \sigma_1},$$
where $\mu_0 = \overline{x}_i,\ \mu_1 = \overline{x}_{i+1},\ \sigma_0 = s_i,\ \sigma_1 = s_{i+1}$.
3) If $k_i^\star \ge 2$, we decide that an anomaly has occurred between i-th and $(i+1)$-th observations.

Formally, this algorithm is given as follows.

Algorithm 1 Outlier detection

Input: integer *N* – number of observations, double *m*[*N*], double *s*[*N*] – arrays of means and standard errors of observations, integer *threshold* – value regulating maximum amount of outliers following each other
Output: bool *ifOutlier*[*N*] – array of boolean flags showing if an observation is an outlier or not

```
0: integer anomaly[] := createEmptyIntegerArray();
1: push(anomaly, 1);
2: integer n := 1;
3: for i := 1, 2, ..., N do
       ifOutlier[i] := False;
   end for
4: for i := 1, 2, ..., (N-1) do
       k := abs(m[i+1] - m[i]) ÷ (s[i] + s[i+1]);
       if k ≥ 2 then
         push(anomaly, i+1);
         n := n+1;
       end if
   end for
5: push(anomaly, N+1);
6: n := n+1;
7: for i := 1, 2, ..., (n-1) do
       if anomaly[i+1] - anomaly[i] ≤ threshold then
         for j := anomaly[i], ..., (anomaly[i+1] - 1) do
             ifOutlier[j] := True;
         end for
       end if
   end for
```

V. Performance evaluation

A. Modeling the real-life data

In order to test the proposed outlier detection algorithm, we model the real-life data (Lunar Laser Ranging data): only in the case of model data when the outliers in the data are exactly known, it is possible to do this correctly.

Lunar Laser Ranging data are generated by the following procedure: (i) there are five retroreflector arrays installed on Moon; (ii) each of them represents a reflective surface; (iii) hitting such surfaces, the light ray reflects back; (iv) in the earth observatories, lasers are installed aiming at the retroreflector arrays on Moon; (v) for each pulse transmitted, the response time is measured and this response time is proportional to the distance between the observatory and the retroreflector array.

Lunar Laser Ranging data is presented in special databases. Actually, the rows of database files correspond not to individual observations but to the groups of observations. Each

group consists of the observations obtained at approximately one point in time which have been considered to belong to the same distribution (unknown). Thus, the data for processing consists of pairs: (mean value $\bar{x}$, its standard error s).

Real-life data turns out to be presented as a sequence of segments of observations. These segments obey different distributions and are separated from each other with a pronounced shift (however, in this work we are not going to deal with the shifts). Real-life data is contaminated by outliers.

The real-life data is manually divided into segments. The division is based on visual assessment. After that, also manually, the outlier-like observations are removed from data. As a result, we expect to get data segments distributed unimodally or close to.

After that using the *R* packages we attempt to match the distributions and their parameters with the segments so that the real-life data is described in the best way. For the means the following distributions are used: normal, Cauchy, logistic, skewed normal and skewed Student. For the standard errors we use the following distributions: exponential, gamma, log-normal and Weibull. The choice of distributions for matching is determined, first of all, by the quality of their fitting to the real-life data; secondly, it is due to their availability in the *R* statistical packages.

For each real-life data segment the best fitting pair of distributions is chosen by applying the χ^2 fitting test. These results are represented in Table I.

TABLE I. THE FRACTIONS OF SEGMENTS WITH DIFFERENT DISTRIBUTIONS PAIRS

	Weibull	Log-normal	Gamma	Exponential
Normal	0	0.023	0	0
Cauchy	0	0.023	0.047	0
Logistic	0.023	0.093	0	0.047
Skewed normal	0	0.070	0	0.047
Skewed Student	0.070	0.255	0.047	0.255

Based on the information obtained as a result of real-life data analysis, model data is generated as follows:

1) The amount of observations in the generated segment is chosen randomly using the inverse distribution function of the number of observations in segments.
2) The pair of distributions for the generated segment is selected randomly in accordance with the fractions of different pairs. The parameter values set is chosen from the real-life sets of parameter values.
3) The generated segment is randomly contaminated by outliers so that they account to $2-3\%$ of the total amount of observations. The basic condition for the outliers to satisfy is that their means and standard errors should significantly differ from the means and the standard errors of the observations considered to be true.
4) All the generated segments are concatenated.

B. Processing model data

We now conduct the experiment under the same condition (i.e., with the same number of segments) N times. Each time we calculate the probabilities of interest:

1) outlier detection probability
$$P_D = \frac{N_D}{N_{out}},$$
where N_D stands for the number of correctly detected outliers, N_{out} stands for the total number of outliers;
2) outlier false alarm probability (probability of taking a true observation as an outlier)
$$P_F = \frac{N_F}{N - N_{out}},$$
where N_F stands for the number of incorrectly detected outliers, N stands for the total amount of observations, N_{out} stands for the total amount of outliers.

Here we give the illustration of results got on model data.

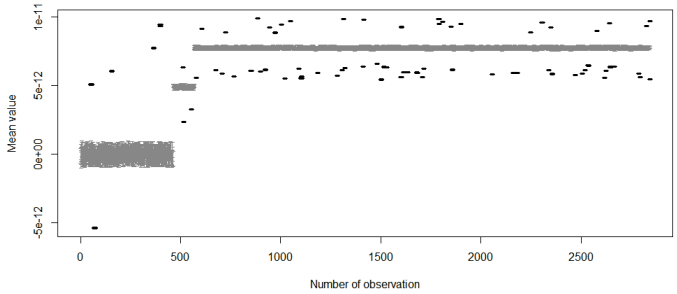

Fig. 2. Outlier detection with 3-segment model data

In Fig. 2 observations considered to be true are grey-colored and outliers are black.

For the experiment conducted with the number of Monte Carlo cycles equal to 1000, the probability of outlier detection P_D is approximately 0.99, whereas the false alarm probability $P_F \approx 0$ within the margin of error - the obtained results are unexpectedly good, especially if to recall that the Chebyshev upper bound upon the probability of errors is 0.25.

C. Processing real-life data

Here we give the illustrations of the method application to real-life data. Observations which were marked as outliers are black-colored.

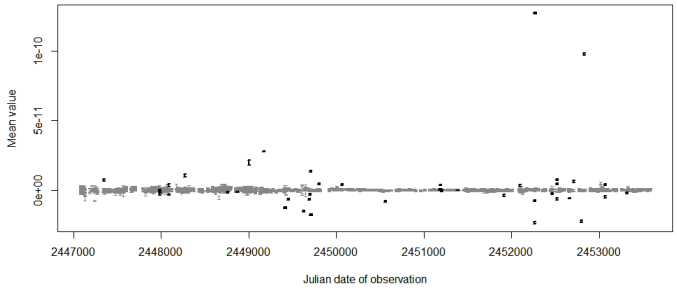

Fig. 3. Outlier detection with real-life data (Cerga2)

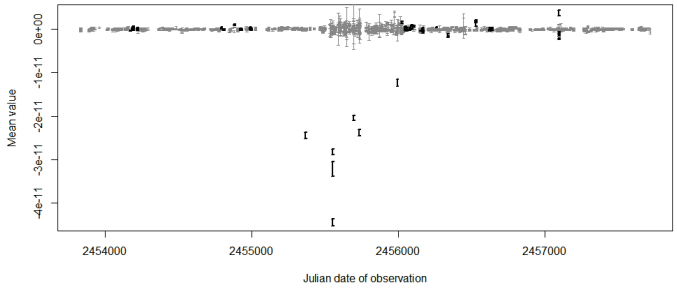

Fig. 4. Outlier detection with real-life data (Apache)

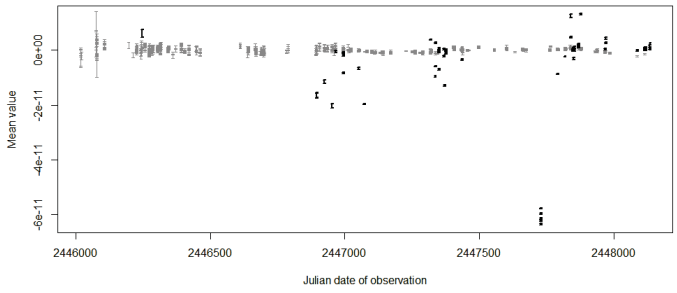

Fig. 5. Outlier detection with real-life data (Haleakala)

The feature of real-life data is that we never know if an observation is an outlier or not as we have no "correct answer" to compare the results with. Thus, we have two main ways to analyze the results: visual analysis and expert reviewing. Analyzing the results visually, we see that some of the observations which could be expected to be filtered are marked as outliers—though not each of them, as well as we can not claim that each observation marked is an *obvious* outlier. Anyway, most of the observations which have significantly different mean values and/or large standard values have been detected – that is exactly what we expected. From the expert reviews we managed to understand that the results obtained *resemble* what we supposed to get.

VI. Conclusions

1) Classical probability tools, namely, the Gauss-Chebyshev inequalities are applied to outlier detection in stream data described as the set of pairs of observation means and their standard errors as the only available summary information. This approach allows to get the corresponding confidence intervals and on their basis to formulate the problem of outlier detection as the problems of hypotheses testing both in the Bayesian and Neyman-Pearson senses: minimizing the upper bound of the Bayesian risk and maximizing the lower bound of the test power, respectively.
2) These optimization problems are considered under the natural side conditions of non-intersecting confidence intervals—after simplifying the optimization setting, a low-complexity effective algorithm of outlier detection is proposed.
3) The proposed algorithm has been applied to the model data—it exhibits unexpectedly splendid performance.
4) A rather fair performance of this algorithm is observed while processing the real-life data.
5) In case of the availability of the additional information about data distributions, the quality of outlier detection can considerably be improved with the use of the Gauss and Vysochanskij-Petunin inequalities.
6) The successful experience of application of the proposed algorithm to the real-life data lets us expect its advantageous applicability to other problems of anomaly detection, for example, shift detection.

Acknowledgment 1

The reported study was partially funded by RFBR according to the research project 18-29-03250.

Acknowledgment 2

The authors are very much grateful to our colleague, Dmitry A. Pavlov, for his stimulating help in the performing of this work.

References

[1] S.A. Alasadi and W.S. Bhaya, "Review of data preprocessing techniques in data mining". *Journal of Engineering and Applied Sciences,* vol. 12, 2017, pp. 4102-4107.

[2] S. Srivastava, "Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining". *International Journal of Computer Applications. Foundation of Computer Science,* vol. 88, 2014, pp. 26-29.

[3] F.E. Grubbs, "Procedures for Detecting Outlying Observations in Sample". *Technometrics,* vol. 11, 1969, pp. 1-21.

[4] V.J. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies". *Artificial Intelligence Review*, vol. 22, 2004, pp. 85-126.

[5] L.B. Klebanov, J. Antoch, A. Karlova and A.V. Kakosyan, "Outliers and related problems". *arXiv:1701.06642 [math.PR]*, 2017.

[6] V. Barnett and T. Lewis, *Outliers in Statistical Data*. Wiley, 1994.

[7] I. Ben-Gal, Outlier detection, In: Maimon O. and Rockach L. (Eds.) "*Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers,*" Kluwer Academic Publishers, 2005.

[8] A.F. Pimentel Marco, D.A. Clifton, L. Clifton and L. Tarassenko, "A review of novelty detection". *Signal Processing*, vol. 99, 2014, pp. 215-249.

[9] J.W. Tukey, *Exploratory Data Analysis.* Addison-Wesley, 1977.

[10] C.F. Gauss, "Gauss's work (1803-1826) on the Theory of Least Squares" / English translation by H.F. Trotter. *Princeton, Princeton University Press*, 1957, pp. 10-13.

[11] P. Tchebichef, "Des valeurs moyennes". *Journal de Mathematiques Pures et Appliquees*, vol. 2. 1867, pp. 177-184.

[12] D.F. Vysochanskij and Y.I. Petunin, "Justification of the 3-σ rule for unimodal distributions". *Theory of Probability and Mathematical Statistics*, vol. 21, 1980, pp. 25-36.