Agnieszka DURAJ[1], Łukasz CHOMĄTEK[1]

Politechnika Łódzka, Instytut Informatyki, Wydział Fizyki Technicznej, Informatyki i Matematyki Stosowanej (1)

Detection of outliers in data streams using grouping methods

Abstract. *Efficient processing of data streams usually requires their initial processing, including on the removal of anomalies caused by, for example, measuring errors. Such errors may result in misinterpretation of the phenomena being analyzed .The literature describes several methods for detecting exceptions in data streams. Each of them requires proper selection of operating parameters. In addition, the effectiveness of methods may vary depending on the data set being analyzed. The article describes current methods for detecting exceptions in data streams and analyzed their operation on gas consumption data.*

Streszczenie. *Przedmiotem niniejszej pracy jest wykrywanie wyjątków w strumieniach danych przy użyciu metod grupowania. Przetwarzanie strumieni danych wymaga wstępnej analizy a przede wszystkim usuwania wszelkiego rodzaju anomalii spowodowanych błędami pomiarowymi. Błędy te prowadzą do niewłaściwej interpretacji analizowanych zjawisk. W literaturze można odnaleźć metody wykrywania wyjątków w strumieniach danych oparte na metodach statystycznych, grupowaniu danych. Każda metoda wymaga odpowiedniego doboru parametrów operacyjnych. Skuteczność jest uzależniona od analizowanego zestawu strumienia. W pracy podano kilka metod grupowania używanych do detekcji wyjątków w strumieniach danych. Metody te sprawdzono dla strumieni dotyczących zużycia gazu.* ***(Wykrywanie wyjątków w strumieniach danych przy użyciu metod grupowania)***

Keywords: outliers, data stream analysis
Słowa kluczowe: wykrywanie wyjątków, strumienie danych.

Introduction

A data stream is a set of observations recorded in time intervals, i.e. containing a unit of time. They do not have to be data recorded at regular intervals, but usually these are data which have an equal time interval. An example of a data stream can be:

- Air temperature measurement in the room every one second, which gives us a data stream consisting of the time value and the measurement result assigned to it in the form of degrees Celsius.
- 24-hour registration of the electrocardiographic signal using the Holter method, recording heart activity;
- Electricityconsumption;
- Record of monitoring at the airport;
- Monitoring the workload of networks and websites;
- Monitoring and recording of work;
- Banking, telemetric and surveillance systems;
- Tracking and analysis of biological and medical data;
- Stock market data;
- Data on all types of physical devices

The data stream is defined as an ordered set of values of the analyzed feature or a specific phenomenon at different times (intervals) of time. It is also a series of observations recorded in a strictly defined time.

Collection and storage of data in the form of streams increases the development of methods of processing them. There is also an increased problem of detection of outliers in data streams.

The problem of detecting outliers in large data collections is, according to the authors, an important research problem. Among others, deterministic statistical methods, as well as methods based on distance and density, are used. However, none of the methods proposed in the literature is universal. A rich overview of this field is given in [1,2]. In addition, their effectiveness depends on the data set as well as on the parameters required for their operation. The authors dealt with the detection of outliers in their earlier works. Among other things, a method of dete-cting outliers using linguistic summaries was developed [3,4,7]. It was also proposed to solve the problem of dete-cting outliers through multicriteria optimization. See e.g. [5,6,8].

Outliers in data streams

Outliers are occasional deviations, random variations are often undesirable. Random fluctuations occurring in the exiting streams usually have a negative effect on the further prototyping process (prediction, forecasting)

Data stream is often considered as an ordered sequence of values of the analyzed feature or the phenomenon in different time. The other works, defines a data stream as series of observations in a specified time.

Outliers in the streams can lead to a change in the trend which is very unfavorable for the analysis of a given stream. The data stream decomposes by determining seasonal fluctuations, trend and random fluctuations, which in this work are called exceptions. The decomposition of the example stream is shown in Fig. 1 with the trend, seasonal fluctuations and random variations that constitute exceptions in the data stream.

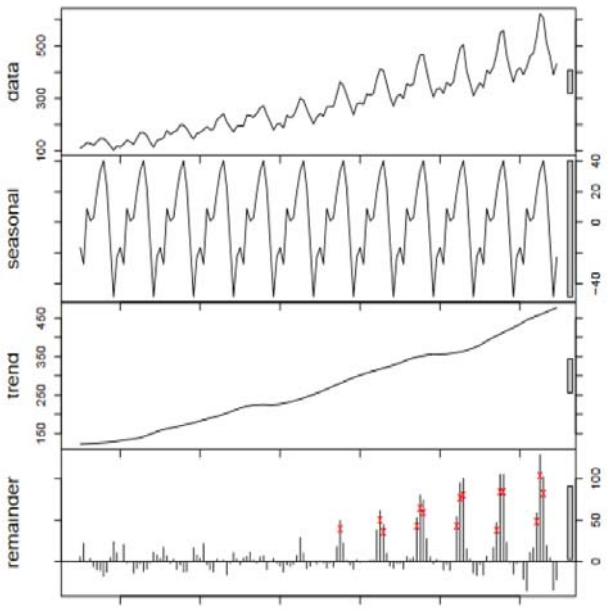

Fig.1 Decomposition of the data stream (successively from the top): the original data, seasonal fluctuations, trend and random fluctuations.

The similarity of data streams was determined using the Dynamic Time Warping (DTW) measure introduced by Sakoe and Chiba in 1978. The DTW measure matches the two series very flexibly so that the distance function is the smallest. The two time series Q and C are compared

respectively with distances of *n* and *m*, $n \times m$. A matching path is created as a vector between two points q_i and c_j. in the form $W = (w_1, w_2, \dots, w_k)$ such that$w_k = (i, j)$, which represents the adjustment of the series Q and C. For this vector the distance DTW is calculated according to

$$DTW(Q,C) = min\sqrt{\sum_{k=1}^{K}\frac{w_k}{K}}$$

where *K* is used to eliminate warping results due to path length.

It should be noted that there may be multiple paths in the data stream. One can, therefore, introduce a limitation of monotonicity, which forces the monotonic distribution of points in time. The second limitation concerns boundary conditions. The matching path must always start and end in the opposite corners of the matrix, which comes down to fixing $w_i = (1,1)$ and $w_k = (m,n)$. Continuity should also be maintained, what means that all points must be adjacent to predecessors, $i_k = i_{k-1} \leq 1$ and $j_k = j_{k-1} \leq 1$.
The optimal path is determined on the basis of DTW.

The article focuses on the detection of exceptions in data streams using data mining methods. The work is organized as follows: chapter II acquaints the reader with the current state of knowledge in the field of detection of outliers in data streams. Section III is a practical example of the methods used for the gas consumption data set. The work is finished with a summary.

Related works

The collection of data in the form of streams has become very common. Created systems allows to gather the data about User navigation on the Web, time of telephone conversations, banking operations or operations with the credit cards, states of devides, etc. [9, 10, 11].

When searching for similarities between time series or its fragments, one can usedifferent methods originatedfrom the analysis of the signals. Often, in this case the data transformation in the area of frequency. For digital data the discrete Fourier transform (DFT) or discrete Wavelet transform (DWT) are used. More information on the classification and ways of time series analysis can be found in the work [12]. There is also a proposal to replace the values of the time series from real to symbolic, which can facilitate data analysis using knowledge discovery algorithms

During our research, the RStudio environment and its packages adapted to streams were utilized. The studies took into account hierarchical methods and algorithms dedicated to data streams. Research related to the classification of data streams was given in the authors' earlier work.

The Anomalus algorithm generates a feature vector of a given data stream fragment and applies a strong decomposition of features by making the two first components of the anomaly detection vector decomposition.

The Wsher method or the so-called Venturini[21] method is most often used for streams in which there are no negative values. For each observation, the previous and subsequent observation is selected. The observations are connected by a line and in this way you can easily see the exceptions. The process of combining observations in the stream is shown in Fig. 2.

Another method that marksoutliers in the data streams is the Chen and Liu procedure[23], called TsOutliers. Outliers are obtained on the basis of determined linear regression. The masking effect is limited. Maximum credibility of the regression model parameters is determined on the basis of the original or corrected data series. In this way deviations are estimated - outliers appearing in the data stream. Of course, the original series is used only for the first iteration.

For calculated deviations for each $t = 1,2,\dots,n\$$ we calculate$\tau_{IO}(t)$, $\tau_{AO}(t)$, $\tau_{LS}(t)$.- maximum values of standardized deviation statistics. We get values approximately in the normal distribution in the order from the left for innovative, additive deviations and level changes.

In the next stage, the maximum values of *η* are determined for the types of deviations *IO, AO, LS* according to

$$\eta_t = \max(\tau_{IO}(t), \tau_{AO}(t), \tau_{LS}(t))$$

If for *n* this iteration the maximum critical value is greater than the determined critical value *WK*,then there is a probability of an outlier (anomaly) of one or more types for $t_1 = t_p$ wheret_p can be one of the types of deviations *IO, AO, LS*.

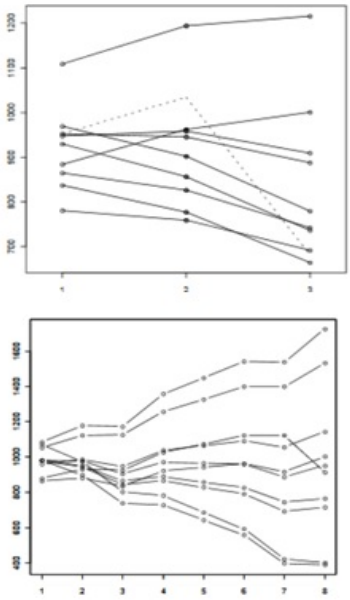

Fig. 2 Vintirini's method which connects observations in the data stream.

Practical example of outlier's detection

In the research we used gas consumption collection. The collection was downloaded from [http://datahub.io/core/natural-gas/r/natural-gas-monthly.csv]. Original collection was extended and modified by adding outliers. Then the analysis of the detection of outliers with the methods described in Section II was started. Outliers were analyzed using the hierarchical method and anomalous algorithms, TsOutliers. Graphic illustrations are presented in sequence in Fig. 3, Fig. 4 and Fig. 5.

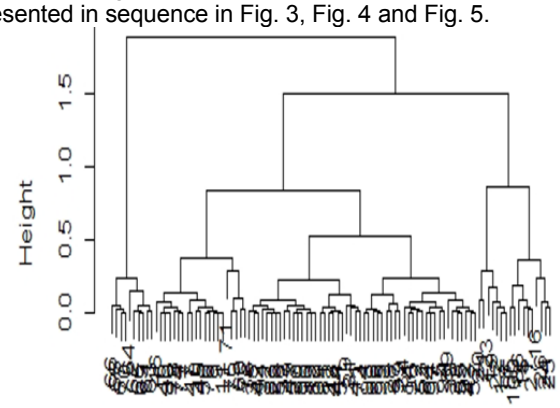

Fig.3 Grouping a stream using a hierarchical method

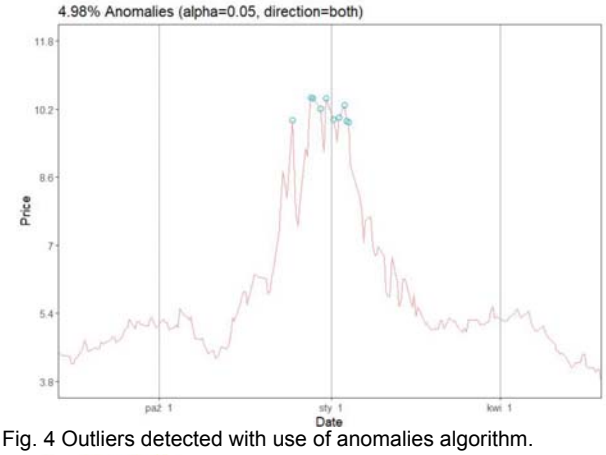

Fig. 4 Outliers detected with use of anomalies algorithm.

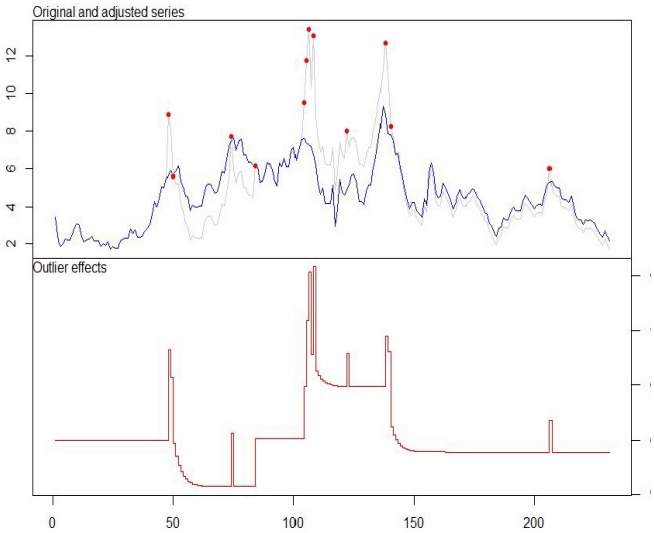

Fig.5 Outliers detected by TsOutliers algorithm

The best results for the anomalous algorithm were obtained for Ralpha = 0.05. The algorithm detected 86% of outliers in the analyzed dataset.

Fig.5 shows the detected anomalies on the gas consumption, using TsOutliers. The gray line shows the original values along with the points included in the anomalies. The outliers are marked in red. The blue line shows the matched values of the time series. It should be noted that the graph obtained as a result of the algorithm is more static, has fewer violent deviations in relation to the original one. 12 anomalies of various types have been detected. The most frequently detected anomalies were a sharp jump in the value up, change in signal level and time changes.

The best results were obtained for the TsOutliers algorithm. For the same fragment of the analysed stream, 97% of the detected anomalies were obtained.

Conlusions

As demonstrated in the research, the methods used in the work detect deviations in the streams. The hierarchy methods have the least accuracy.

Research work on the detection of outliers in data streams will be continued. Particular attention will be devoted in the future to the rapid parametric transformations described in [16, 17]. We will also check outlier detection methods for data obtained in [14,15].

REFERENCES

[1] Aggarwal, Charu C., Outlier Analysis, Springer, 2013.
[2] Barnett, V., Lewis, T., Outliers in statistical data, Wiley, 1994.
[3] A. Duraj A., P.S. Szczepaniak, Information Outliers and Their Detection in: M. Burgin and W. Hofkirchner (Eds.): Information Studies and the Quest for Transdisciplinarity World Scientific Publishing Company, Vol.9, Chapter 15, pp. 413—436
[4] A. Duraj,Outlier Detection in Medical Data Using Linguistic Summaries, 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications INISTA 2017, Gdynia, Poland, 3-5 July 2017.
[5] Ł. Chomątek, A. Duraj,Multiobjective Genetic Algorithm for Outliers Detection, 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications INISTA 2017, Gdynia, Poland, 3-5 July 2017, pp. 379-384
[6] A.Duraj., Ł.Chomątek, Outlier Detection Using the Multiobjective Genetic Algorithm, Journal of applied computer science, Institute of Information Technology, pp. 29-42, 10p.
[7] A. Duraj, A.Niewiadomski, P.S. Szczepaniak, Outlier detection using linguistically quantified statements, International Journal of Intelligent Systems, Vol.33, No 8, pp.1590-1601
[8] A.Duraj, Ł.Chomątek,Outlier Detection Using the Multiobjective Genetic Algorithm, Journal of Applied Computer Science, Vol.25, No 1, pp.29-42
[9] Berndt D.J, Clifford J. (1994) Using dynamic time warping to find patterns in time series.KDD Workshop.
[10] Suchwałko, A. i Zagdański, A. Analiza i prognozowanie szeregów czasowych. Praktyczne wprowadzenie na podstawie środowiska R. Wydawnictwo Naukowe PWN. 2016 Warszawa
[11] Vallis, O., Hochenbaum, J., & Kejariwal, A. A Novel Technique for Long-Term Anomaly Detection in the Cloud. 2014.
[12] López-de-Lacalle, J. Tsoutliers. 27 Maj 2017. https://cran.r-project.org/web/packages/tsoutliers/tsoutliers.pdf
[13] Yanchang, Z. R and Data Mining. Examples and Case Studies Academic Press. 2013
[14] Lebioda, M.; Rymaszewski, J.; Korzeniewska, E., Simulation of Thermal Processes in Superconducting Pancake Coils Cooled by GM Cryocooler, MICROTHERM' 2013 - MICROTECHNOLOGY AND THERMAL PROBLEMS IN ELECTRONICS Book Series:Journal of Physics Conference Series Volume:494Article number:012018 Published:2014
[15] Rymaszewski, Jacek; Lebioda, Marcin; Korzeniewska, Ewa, Simulation of the loss of superconductivity in a three-dimensional model of the metal-superconductor connection, PRZEGLAD ELEKTROTECHNICZNY, Volume:88, Issue:12B, Pages:183-186, Published:2012.
[16] Puchala, D.: Approximating the klt by maximizing the sum of fourthordermoments. IEEE Signal Processing Letters 20(3) (2013) 193–196
[17] Puchala, D., Yatsymirskyy, M.: Joint compression and encryption ofvisual data using orthogonal parametric transforms. Bulletin of thePolish Academy of Sciences Technical Sciences 64(2) (2016) 373–382
[18] Kejariwal, A. *Introducing practical and robust anomaly detection in a time series*. Access: Nov, 22, 2018, blog.twitter.com: https://blog.twitter.com/engineering/en_us/a/2015/introducing-practical-and-robust-anomaly-detection-in-a-time-series.html
[19] Shadan, M. *Manipulating Time Series Data with xts*. Access: Nov, 22, 2018, RPubs: https://rpubs.com/mohammadshadan/288218
[20] Hyndman, Rob J. *Highest Density Regions and Conditional Density Estimation,*Access: Nov, 22, 2018, z cran.r-project.org: https://cran.r-project.org/web/packages/hdrcde/hdrcde.pdf
[21] Venturini, Andrea *Time Series Outlier Detection: A New Non-Parametric Methodology (Washer)*. 2011. Access: Nov, 22, 2018, https://rivista-statistica.unibo.it/article/viewFile/3617/2968
[22] Rosner, Bernard *On the detection of many outliers*. Technometrics 1975, Issue 2, Volume 17, Pages 221-227
[23] Chan, Liu *Journal of the American Statistical Association*, 1993, No. 421, Volume 88, Pages 284-297, Access: Nov, 22, 2018, istat.it: https://www.istat.it/it/files/2014/06/Joint-Estimation-of-Model-Parameters-and-Outlier-Effects-in-Time-Series.pdf
[24] Hyndman, Rob J. *Computing and Graphing Highest Density Regions.* The American Statistician, May 1996, No. 2, Vol. 50, Pages 120-126,Access: Nov, 22, 2018, webpages.uidaho.edu: http://www.webpages.uidaho.edu/~stevel/517/Computing%20and%20Graphing%20HDR.pdf