Outlier detection and classification in sensor data streams for proactive decision support systems

M V Shcherbakov[1], **A Brebels**[2], **N.L. Shcherbakova**[1], **V.A. Kamaev**[1], **O.M. Gerget**[3] , **D. Devyatykh**[3]

[1] Volgograd State Technical University, 28, Lenina Ave., Volgograd, 400066, Russia
[2] KULeuven University, Oude Markt 13, 3000 Leuven, Belgium
[3] Tomsk Polytechnic University, 30, Lenina Ave., Tomsk, 634050, Russia

E-mail: maxim.shcherbakov@vstu.ru

Abstract. A paper has a deal with the problem of quality assessment in sensor data streams accumulated by proactive decision support systems. The new problem is stated where outliers need to be detected and to be classified according to their nature of origin. There are two types of outliers defined; the first type is about misoperations of a system and the second type is caused by changes in the observed system behavior due to inner and external influences. The proposed method is based on the data-driven forecast approach to predict the values in the incoming data stream at the expected time. This method includes the forecasting model and the clustering model. The forecasting model predicts a value in the incoming data stream at the expected time to find the deviation between a real observed value and a predicted one. The clustering method is used for taxonomic classification of outliers. Constructive neural networks models (CoNNS) and evolving connectionists systems (ECS) are used for prediction of sensors data. There are two real world tasks are used as case studies. The maximal values of accuracy are 0.992 and 0.974, and F1 scores are 0.967 and 0.938, respectively, for the first and the second tasks. The conclusion contains findings how to apply the proposed method in proactive decision support systems.

1. Introduction

Increasing the volume and intensity of information and data flows is one of the most essential challenge in information technologies development. This expansion of data leads to allocation of the separate research domain known as data stream mining [1, 2] and knowledge database discovery (KDD) [3, 4]. Data quality assessing is the crucial process in engineering systems where sensor data gathering and transmited into a centralized repository [5]. Data may be distorted or lost due to different reasons and unpredicted circumstances. If corrupted data is used for management and decision-making, then decisions might be inaccurate and inadequate. As a result, actions performed based on wrong data can be incorrect and can lead to undesirable consequences. The situation becomes especially critical in the case of automatic control.

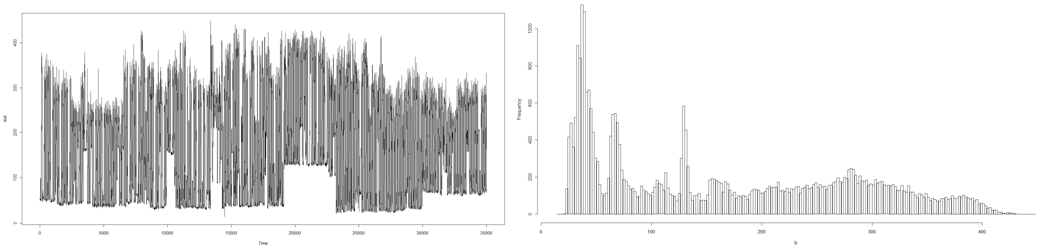
As an example, we consider an energy management process for the network including various buildings. Generally, sensors or meters with digital outputs are polled by data acquisitions devices to get data about energy consumption. After, data acquisitions devices transmit data to the server for storing information in a database. As sensors and meters with digital outputs are installed in residential or non-residential buildings, therefore, they receive various negative impacts. This leads to misoperation of meters and data acquisition systems, thereby data may be corrupted or lost. These situations may be interpreted as existence of outliers in data streams. In such cases, human or machine feedback and actions should be performed as soon as possible, so the response time needs to be minimal. Needless to say, that if manual operations are replaced by algorithms based on machine learning and computational intelligence then the training sets is need to be prepared [5, 6]. Despite the fact that preprocessing of the data is included in the Cross Industry Standard Process for Data Mining (CRISP-DM), preprocessing is most likely art than the science due to an unstructured kind of task statements [7, 8].

Another aspect is associated with different types of corrupted data: gaps, outliers and anomalies. Let us note that literature reviews show the difference between the terms of 'an outlier' and 'anomaly'. Outliers are values that differ from other values significantly [9]. Generally, these situations are connected with failures of observed devices or data transmitters [10, 11]. In the presented study, this type of outliers belongs to type I of an outlier group. If the observed system or the process have many stages or operation modes [12], anomalies indicate the shift to the unproper operation mode or a transition to the negative stage [13]. In the presented study, these anomalies belong to type II of an outlier group. We shall note that the undertaken actions differ from each other and the choice depends on types of outliers. Finally, the detector must find out outliers of different types in data streams and classify them properly.

So the method for detecting and classifying the Ist and IInd types of the outlier must meet the following requirements. Firstly, it must handle data streams in the real time mode. Secondly, the method should contain the algorithm for splitting detected wrong data into the classes. This study provides the data-driven solution for the new task of identification and determination of outliers of different types in sensor data streams. This solution is based on forecasting and clustering machine learning techniques.

2. Task statement and background

A stream of sensory data is a continuous stream of data generated by heterogeneous sources, and this stream is considered as input data for monitoring systems [14]. Due to the fact, that the investigated sensor data are generated by systems with several operation modes (multi-instance), we can provide task statement for identifying outliers for a one-dimensional case.



a b

Figure 1. A 15-minute time series of energy consumption for a year period (a) and its histogram (b).

The outlier detecting in sensor data streams is a part of the research field called sensor data stream minig incuding patterns recognition in continuous and constantly incoming data streams [6]. The literature review shows the three main groups of outliers detecting methods: based on machine learning algorithms with a supervised scenario of training, an unsupervised scenario and a semi-supervised scenario [13]. This research was based on the third group of methods, as an expert should be involved in the process as less as possible. The latter class contains the following methods [15, 16]: distribution-based [17], depth-based [19], density-based [21,22], distance-based [15, 23, 24], and clustering-based methods [25, 26]. All those methods have sortcomings in case they are applied to sensor stream data analysis. The new promising methods were based on a data-driven approach [5, 28].

The literature review shows that the data-driven approach might be used as a basis for the new method which allows one to split outliers into two main groups: the Ist and IInd types of outliers.

3. A method

The main purpose of the proposed method is detecting the Ist and IInd types of the outlier in the real time mode to be implemened in proactive decision support systems. The architecture of the system for the Ist and IInd type outlier detection is based on the architectures of sensor data stream processing [6]. The architecture includes an online component for predicting data stream values (marked as *DSP*), a component - for the Ist type outlier identification (marked as *OD*), a component - for the IInd type outlier identification (marked as *AD*) and an offline component (marked as *OFLF*) - for configuring components *DPS*, *OD* and *AD*.

The method uses short-term forecasting models which are fitted based on historical data. When the newly observed value is captured by the system, simultaneously, component *DSP* generates a predicted value for the expected time stamp. So, the system has a real (observed) value and a predicted value based on previous observation. Both of these values are passed into outlier identification component *OD* as inputs values. Then the following check is made. If the observed value differs from the predicted one, i.e. the difference between these values exceeds the calculated threshold, then this value is marked as an outlier. When an outlier is detected, the next step is to clarify a type of an outlier. In this case, component *OA* checks whether the detected outlier belongs to one of the previously defined clusters containing values marked as the Ist and IInd types of outliers. When an outlier is detected and the type of the outlier is defined, the system generates an appropriate alarm for users. Finally, the sequence of action regarding the proposed method is the following: (i) initialization of parameters of components *DSP*, *OD* and *OA*; (ii) identification of the Ist and IInd types of outliers; (iii) offline adjustment of parameters of components *DSP*, *OD* and *OA*. As the task of the outlier classification falls into the definition of classification and clustering tasks, the following quality criteria are used: accuracy, precision, recall and F1 score. Also, the three new criteria have been added. *TPO* is a number of cases where the Ist type of the outlier identified as the Ist type of the outlier correctly. TPA is a number of cases where the IInd type outlier is identified as the IInd type outlier correctly. And FAOR is a number of cases where the IInd type outlier is wrongly detected as the Ist type outlier.

Component *DSP* uses time series forecasting models to obtain the predicted value of the data stream for the expected time. This research is built on two machine learning techniques for time series forecasing: constructive neural networks (CoNN) [32] and evolving connectionists systems (ECS) [29]. This choice was made based on well-proven results of neural network techniques for energy consumption forecasting. The main benefit of chosen approaches is the ability to change the network structure during the training process. This allows one to (i) discover the proper network structure during the training and exclude human intuition from the structure initialization process and (ii) minimise the total cost function [29; 31]. Let us note that for the training, the assessor creates a labelled training set which is used for training CoNN and ECS. After training, mean squared error MSE is calculated over the time interval with size *D*. We assume, that absolute errors of forecasting of the training set are normally distributed. Many experiments proved this intuition [30]. This assumption will be used for the outlier detection algorithm. After a series of experiments, we found that the ensemble of CoNN performed better than single CoNN. So the ensemble of CoNN is used for calculations.

Component *OD* using MSE was calculated by the *PSD* component for interval *D*. We shall note that *D* is a number of discrete time values, so the MSE is a calculation for interval [t - D + 1, t - 1]. Threshold *Q* for outliers detection is calculated as the product of MSE and defined confidence level a. The next step is checking for the condition: if an error (the difference between a single real obtained value at time t and a predicted value by PSD for time *t*) falls into interval [m - Q, m + Q], then the obtained value is normal. m is a mean error in interval [t - D +1, t -1]. Otherwise, it belongs to a group of the Ist type outliers. Figure 2 explains how the Ist type outlier identification works.

Let us describe the functionality of component *OA*. As the ensemble of *n* CoNN or ESC models is used, the training data set has been split into *n* training sets. For instance, if a time series contains values obtained for every 15-minutes, then there are *96* observations during a day. We assume that the time series has a very strong daily-based pattern, so the initial time series is split into *96* time series. So the first 'artificial' time series contains values only for *0:15* time stamps for every day in the initial time series, the second one contains values for *0:30* and so on. For every artificial time series, the clustering algorithm is applied to separate values in a time series into *k* clusters. The the *k*-mean algorithm was used with a number of clusters equal to *3*. It means all values are separated into the clusters with "normal", the Ist type outliers, the IInd type outliers. As we declared early, the IInd type outliers identify changes of the operation mode or object's states. If there is no additional information about the process, we assume that the observed value is the IInd type outlier if the following conditions are true. So a value is the IInd type outlier if the (i) *i*-th class contains values marked as the IInd type outliers and/or the (ii) *i*-th cluster contains the continuous sequence of the Ist type outliers with length *Z*, and $Z \geq 3$. Let *i* be the label of the nearest cluster's centre relatively a value according to the selected distance.

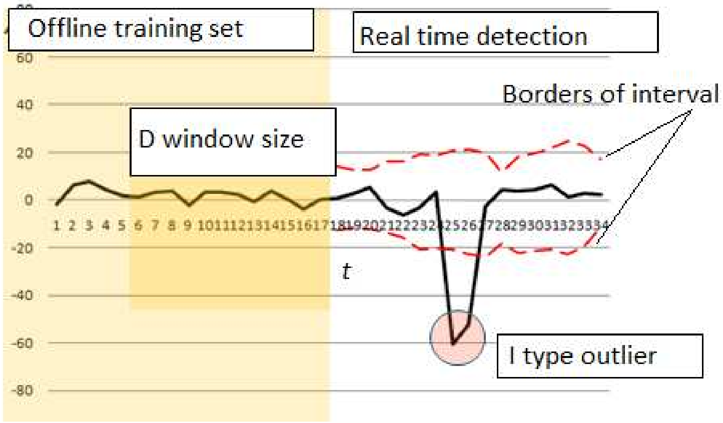


Figure 2. The scheme for explanation of *OD* component functionality.

The *OFLF* component adjusts parameters of *OA*, *OD* and *PDF* components in the batch (offline) mode. The adjustment is performed according to the schedule or when a new chunk of data of the preset volume arrives. If the time series have daily patterns, it may be reasonable to make recalculation once a *24* hour period.

4. Results and discussion

The proposed method was implemented in the energy management system with functions of proactive actions development and proactive decision support (ProEMS) [31]. ProEMS collects data about energy consumption for the network of buildings. Figure 1a shows the time series of the energy consumption data during a year. Let us note that the collected time series have daily, weekly and annual patterns. There are two main modes of building usage: night-time and day-time [32]. Based on the properties of the seasonality, *96* predictive models have been included in the ensemble forecasting model. The same idea has been applied for clustering data for each time series. Two real samples of the labelled time series were prepared for evaluating the efficiency of the proposed method. This data were extracted from the real database of energy management system EcoScada [32]. These time series contained outliers of the both types: I and II types. The first sample consists of 1,728 records with 250 outliers of the both types (14.47% of total). There are 4 records of the Ist type outliers and 246 - of the second one. The training set contains 672 values. The second sample consists of 2,688 values with 120 outliers (4.46% of total). There are 2 records of the Ist type outliers and 118 - of the second one. The training set contains 1,632 values. In the experiments, the following parameters were varied: significance level (*a*), window parameter (*D*), type of neural network (CoNN or ECS) and number of clusters (*K*). In total, 44 experiments were carried out. Table 1 contains selected results of experiments.

Firstly, the proposed method has high accuracy: the minimal value for all experiments is 0.933 and the maximal - 0.998. Also, the precision is high as well: the minimal precision is 0.458 (deviant value) and the maximal precision is 0.992. The minimal value of recall is 0.625 and the maximal is 0.992. The F1 score has the minimal value of 0.628 and the maximal one is 0.967. The experiment labelled as 1.5 is the best one among the experiments for the first case study. But in this best case, 14 outliers of the IInd type were recognised as normal. For the second use cases, the best experiment is 2.11. But the system found 3 more outliers in comparison to the expert defined. Rows, indicating the better result, are highlighted in bold in Table 1.

Table 1. Selective results of experiments

Experiment number and neural network type	Settings (*a*, *D*, *K*)	True positive rates TPO/TPA	Accuracy	Precision	Recall	F1-score
1.5/ CoNNS	**7/7/3**	**3/230**	**0.979**	**0.992**	**0.944**	**0.967**
1.7/CoNNS	7/3/3	4/212	0.973	0.829	0.932	0.878
1.14/ESC	3/7/3	4/238	0.987	0.82	0.968	0.888
1.15/ESC	4/7/3	4/237	0.986	0.906	0.964	0.934
2.6/ CoNNS	8/7/3	0/59	0.940	0.964	0.667	0.788
2.11/ CoNNS	**5/9/3**	**1/112**	**0.991**	**0.927**	**0.95**	**0.938**
2.16/ ESC	5/7/3	1/99	0.980	0.930	0.892	0.911
2.18/ ESC	7/7/3	1/62	0.946	0.966	0.7	0.812

Based on results analysis, we can see that accuracy depends on significance level *a*. For instance, changing values of the significance level in experiments *1.1 - 1.6*, the accuracy was decreasing. Simultaneously, precision was increasing, for values $a > 7$ precision was decreasing. The F1 score had the maximal value for $a = 7$. F1 might be chosen as a compromise measurement for quality evaluation. We shall note that ESC is more sensitive to parameters changing then CoNSS. Experiments showed that the approach based on ensemble is more preferable in term of quality.

5. Conclusion

The article presents the authors' findings regarding the new task statement of the Ist and IInd type outlier detection in sensor data streams. The proposed detection method is based on the forecasting approach and the clustering technique. The novel aspects of the method are: (i) the Ist type outlier detection is based on forecasting of expectation values in the sensor data stream and (ii) the IInd type outlier detection is obtained by splitting values into clusters using clustering techniques. CoNNS and ESC structures of neural networks were used for forecasting as these paradigms create the neural network structure during the training procedure, and the training process takes less time in comparison with other neural network approaches. Splitting outliers into classes allows understanding changes in operation modes of observed systems.

The method was implemented and tested on the couple of cases of application. Results show the effective detection of the Ist and IInd types of outliers, the F1 score has the maximal value of 0.967. On average, F1 is 0.938 for the first experiment and 0.861 for the second one. This method can be applied in decision support systems where proactive decisions need to be made.

6. Acknowledgments

The reported study was partially supported by RFBR research projects 16-37-60066_mol_a_dk, Project MD-6964.2016.9 and the project of the Russian Ministry of Education (VSTU State task project # 2.1917.2014K_2014).

References

[1] Gaber M.M. 2012 Advances in data stream mining *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2(1)** 79-85

[2] Jiawei H.J.G., Han J., Gao J 2009 Research Challenges for Data Mining in Science and Engineering *Challenges* 1-18

[3] Gama J., Gaber M 2007 Learning from Data Streams: Processing Techniques in Sensor Networks *Springer*

[4] Omitaomu O. A. et al 2010 Knowledge discovery from sensor data (SensorKDD) *SIGKDD* **11(2)** 84-87

[5] Hill D.J., Minsker B.S 2010 nomaly detection in streaming environmental sensor data: A data-driven modeling approach *Environmental Modelling & Software* **25(9)** 1014-1022

[6] Alzghoul A., Lofstrand M. 2011 Increasing availability of industrial systems through data stream mining *Comput. Ind. Eng.* **60(2)** 195-205

[7] Larose D.T 2005 Discovering knowledge in data: an introduction to data mining *John Wiley & Sons, Inc.*

[8] Nisbet R., Elder J., Miner G 2009 Handbook of Statistical Analysis and Data Mining Applications *Academic Press*

[9] Hawkins D.M 1980 *Identification of outliers* (Chapman and Hall)

[10] Barnett V., Lewis T 1994 Outliers in statistical data *John Wiley & Sons, Inc.*

[11] Applied multivariate statistical analysis 1998 *Prentice-Hall, Inc.*

[12] Wu, O., Gao, J., Hu, W., Li, B., Zhu, M. 2010 Identifying Multi-instance Outliers *SIAM* 430-44

[13] Kriegel H.-P., Kroger P., Zimek A. 2010 Outlier Detection Techniques *Tutorial*

[14] Abadi, D., et al. 2013 Aurora: A Data Stream Management System

[15] Breunig, M.M et al 2000 LOF:Identifying Density-Based Local Outliers *SIGMOD* **29(2)** 93-104

[16] Liua M. A 2011 Novel Approach to Mining Local Outliers *Energy Procedia* **13** 6332-6339

[17] Grubbs F.E., Beck G 1972 Extension of Sample Sizes and Percentage Points for Significance Tests of Outlying Observations *Technometrics* **14(4)** 847-854

[18] Arning A., Agrawal R., Raghavan P 1996 Method for Deviation in Large Databases *KDD* 146-169

[19] Tukey J 1977 Exploratory Data Analysis. *Addison-Wesley*

[20] Ruts I., Rousseeuw P.J 1996 Computing depth contours of bivariate point clouds *Computational Statistics & Data Analysis* **23** 153-168

[21] Knorr E.M., Ng R.T 1998 Algorithms for Mining Distance-Based Outliers in Large Datasets *Proc. of the 24th International Conference on Very Large DataBases*

[22] Wei X., Huang H., Tian S. A 2007 Grid-Based Clustering Algorithm for Network Anomaly Detection *First Int. Symp on Data, Privacy, and E-Commerce* 104-106

[23] Jin W., Tung A. K. H., Han J., & Wang W 2006 Ranking Outliers Using Symmetric Neighborhood Relationship *PAKDD*

[24] Papadimitriou S., Kitawaga H., Gibbons P.B 2002 LOCI: Fast Outlier Detection Using the Local Correlation Integral *ICDE*

[25] Loureiro A., Torgo L., Soares C. 2004 Outlier Detection Using Clustering Methods: a data cleaning application *KDNet*

[26] Zoubi B.A 20009 An Effective Clustering-Based Approach for Outlier Detection *European Journal of Scientific Research* **28(2)** 310-316

[27] Gu X., Papadimitriou S., Yu P. S., and Chang S.P 2008 Online Failure Forecast for Fault-Tolerant Data Stream Processing *ICDE* 1388-1390

[28] Tian L., Li A., Zou P 2006 Research on prediction models over distributed data streams *Int.*

Conf. on Web Information Systems 25-36

[29] Kasabov N. 2007 Evolving Connectionist Systems *Perspectives in Neural Computing*

[30] De Veaux, R. D., Schweinsberg, J., Schumi, J., & Ungar, L. H. 1998 Prediction intervals for neural networks via nonlinear regression *Technometrics* **40(4)** 273-282

[31] Tyukov A., Brebels A., Shcherbakov M. 2011 Automatic two way synchronization between server and multiple clients for HVAC system *iiWAS '11* 467-470

[32] Kamaev V., Shcherbakov M., Panchenko D., Shcherbakova N., Brebels A. 2012 Using connectionist systems for electric energy consumption forecasting in shopping centers *Automation and Remote Control* **73(6)** 1075-1084