Computational Intelligence, Information Systems and Data Mining

edited by
Małgorzata Charytanowicz
Paweł Karczmarek
Adam Kiersztyn

Wydawnictwo
Politechniki Lubelskiej

Lublin 2021

Reviewers:
Piotr A. Kowalski
Grzegorz Kozieł
Edyta Łukasik
Dorota Pylak

Publication approved by the Rector of Lublin University of Technology

ISBN: 978-83-7947-492-9

Publisher: Wydawnictwo Politechniki Lubelskiej
www.biblioteka.pollub.pl/wydawnictwa
ul. Nadbystrzycka 36C, 20-618 Lublin
tel. (81) 538-46-59

Printed by: Soft Vision Mariusz Rajski
www.printone.pl

Adam Kiersztyn[1], Pavel Urbanovich[2], Nadzeya Shutko[3]

The concept of random cluster based outlier detection

Abstract: Detection of outliers is one of the most common and important problems in modern data analysis. Sources of outliers are different. These could be the result of a database malfunction or user errors. The problem is very important due to the dynamic development of large data sets. Therefore, in this paper we present detailed results of work on the concept of using distribution properties to detect outliers. The aim of the study is to introduce an innovative solution that enables the use of statistical semantics of identification and classification of outliers. The undoubted advantages of the novel approach for outlier detection are the simplicity of interpretation and the possibility of its modification. The effectiveness of the proposed method was compared with other recognized techniques to detecting outliers on both artificially generated and empirical data sets.

Keywords: outlier detection, statistical semantic

1. Introduction

In the realities of the world around us, in every field, especially in biological sciences, we deal with the processing of large amounts of data. Data is growing at an alarming rate, but unfortunately data sets contain outliers. As a result of system malfunction or human error, there are numerous anomalies and outliers in data that can have a dramatic effect on the results of queries, reports, and analyzes performed on such data. Therefore, the process of data integration, in particular the process of detecting and classifying anomalies and outliers, is still under development, and the design of effective anomaly detection methods continues to be mainstreamed in data analysis research.

Researchers have proposed several main directions for working with outliers. They are related to the main areas of work in general machine learning approaches. One of the most important models is based on distance [40], in particular the k-nearest neighbour [7], [8], [26], [37] or the density of a dataset, for example Isolation Forest [31], [32]. Also, tests based on support vector machines [30], [42], hidden Markov models [28], [29] or Gaussian processes [48] have been widely discussed in the literature. Recently, with the increasing popularity of deep neural network applications, such approaches have also emerged carefully analyzed. Models such as self-organizing maps, long-term memory, or convolutional neural networks [12], [13], [33], [50] have been

[1] Department of Computer Science, Lublin University of Technology, Lublin, Poland; e-mail: adam.kiersztyn.pl@gamil.com, a.kiersztyn@pollub.pl
[2] Department of Information System and Technology, Belarusian State Technological University, Minsk, Belarus; e-mail: p.urbanovich@belstu.by
[3] Department of Informatics and Web-design, Belarusian State Technological University, Minsk, Belarus; e-mail: shutko_bstu@mail.ru

widely discussed. Several authors have considered the DBSCAN algorithm, see [43]. Finally, interesting techniques related to fuzzy sets were proposed [11], [16], [18], [21], [22], [36], [47], including fuzzy C-means [19], fuzzy rules [35] or linguistic prototypes [49]. The interested reader can find extensive discussions and method reviews in articles [4], [9], [15], [17]. Recently, there has been extensive research into the use of information granules to detect outliers [5], [10], [14], [17], [20], [24], [25], [51].

The concept described in this paper is based on the use of the distribution properties of the analyzed data. Clusters are randomly generated and the affiliation of individual points to the closest clusters is analyzed. It is reasonable to assume that outliers will not be located near other points.

The article is organized as follows. Section II provides a theoretical description of the proposed innovation. Section III includes detailed numerical experiments on an artificial dataset and two large publicly available databases. Particular attention in this section is devoted to the issue of contextual anomaly detection. Finally, Section IV contains conclusions and further research directions related to the development of the proposed approach.

2. Theoretical description

The starting point of the proposed solution (RCOD) is the use of statistical properties of the analyzed data. In the case of multivariate data, the key element of the analysis is to examine the distribution of the analyzed data.

Suppose we have a set D consisting of N records with K numeric fields each. Such a set can be identified with a matrix with N rows and K columns. For such a data set, we randomly select an *n*-element sample S. We will identify the elements of this sample with the centers of the clusters. The size of the sample should depend on the number of analyzed N elements. In the experimental section, the transformation given by the formula

$$n = \lceil lnN + 1 \rceil,$$

was applied. Where ⌈x⌉is rounding up the value of x. In the next step, the distances between all elements of the sample S are determined. In this way a square table with dimensions n × n is obtained, where the elements on the main diagonal are obviously equal to zero. Then, the distribution of distances between the individual elements of the S sample is analyzed and basic position measures are determined, such as minimum (min S), maximum (max S), quartile1 (Q1 S), quartile 2 (Me S) and quartile 3 (Q3 S) Of course, when determining the minimum value, elements from the main diagonal of the distance matrix are not taken into account. Then, for each element from the input set D, the distances from the centers of the clusters are determined. Information is stored whether the distance to any of the clusters is smaller than the analyzed

statistics. In other words, for each element $x \in D$ from the input data set, vector values are calculated, for which individual components are calculated using the formulas

$$x_{min} = \begin{cases} 1, if \ \ min_{y \in S} d(x,y) < min\, S \\ 0, if \ \ min_{y \in S} d(x,y) \geq min\, S, \end{cases} \quad (1)$$

$$x_{Q1} = \begin{cases} 1, if \ \ min_{y \in S} d(x,y) < Q1 \ S \\ 0, if \ \ min_{y \in S} d(x,y) \geq Q1\, S, \end{cases} \quad (2)$$

$$x_{Me} = \begin{cases} 1, if \ \ min_{y \in S} d(x,y) < Me\, S \\ 0, if \ \ min_{y \in S} d(x,y) \geq Me\, S, \end{cases} \quad (3)$$

$$x_{Q3} = \begin{cases} 1, if \ \ min_{y \in S} d(x,y) < Q3\, S \\ 0, if \ \ min_{y \in S} d(x,y) \geq Q3\, S, \end{cases} \quad (4)$$

$$x_{max} = \begin{cases} 1, if \ \ min_{y \in S} d(x,y) < max\, S \\ 0, if \ \ min_{y \in S} d(x,y) \geq max\, S. \end{cases} \quad (5)$$

The procedure described above is repeated predetermined number of times M. When subsequent repetitions of the values obtained by the formulas (1–5) are summed. Action proposed solution outlier detection data can be expressed by the following algorithm.

```
For j=1 to M do
  Random n-element sample S.
  Determine the distance between the elements of the set S.
  Determine min(S), Q1(S), Me(S), Q3(S), max(S).
  For i=1 to N do
      Calculate x_min, x_Q1, x_Me, x_Q3, x_max using formulas (1-5).
      Aggregate values x_min, x_Q1, x_Me, x_Q3, x_max
```

The values obtained in this way, describing in how many cases the examined element is located at a certain distance from random cluster centers, allow for the construction of a classifier determining whether a given point can be considered an outlier. Due to the specificity of the analyzed data sets and their significant diversity, it is necessary to develop a dedicated classifier for each set separately.

3. Numerical experiments

The effectiveness of the proposed solution was tested on 4 specially generated two-dimensional data sets and on 26 publicly available empirical data sets: Annthyroid, Arrhythmia, BreastW, Cardio, ForestCover, Glass, Ionosphere, Letter Recognition, Lympho, Mammography, Musk, Optdigits, Pendigits, Pima,

Satellite, Satimage-2, Shuttle, Speech, Thyroid, Vertebral, Vowels, Wbc, Wine [3], [27], [31], [34], [38], [41], [44], [46], [52] coming from the Outlier Detection DataSets (ODDS), and Nad, Unsw0 coming from Kaggle. As part of a series of experiments, the described algorithm was carried out for each data set with the parameter M = 10000. Then, half of the points were randomly selected from the data generated in this way and two classifiers were built on their basis. The first classifier uses Fuzzy Rule (FR) and the second uses Decision Trees (DT). Two well-known measures were used to compare the effectiveness of the proposed method, namely accuracy and precision given by formulas:

$$ACC = \frac{TP+TN}{TP+FP+TN+FN}, \tag{6}$$

$$PREC = \frac{TP}{TP+FP}. \tag{7}$$

The effectiveness of the proposed solution was compared with the classic Isolation Forest method (IF) [31], Gaussian Mixture (GM) [1], Support Vector Machine (SVM) [6], Elliptical Envelope (EE) [40], Local Outlier Factor (LOF) [8]. The characteristics of the analyzed data sets are presented in Table 1.

Table 1. Characteristics of the analyzed data sets

Dataset	**The number of records**	**The number of attributes**
Artificial 1	5090	2
Artificail 2	10400	2
Artificial 3	10600	2
Artificial 4	20400	2
Annthyroid	7200	6
Arrhythmia	452	274
Breastw	683	9
Cardio	1831	21
Cover	286048	10
Glass	214	9
Ionosphere	351	33
Letter	1600	32
Lympho	148	18
Mammography	11183	6
Musk	3062	166
Optdigits	5216	64
Pendigits	6870	16
Pima	768	8
Satellite	6435	36
Satimage-2	5803	36

Dataset	The number of records	The number of attributes
Shuttle	49097	9
Speech	3686	400
Thyroid	3772	6
Vertebral	340	6
Vowels	1456	12
Wbc	378	30
Wine	129	13
Nad	148517	42
Unsw0	257673	43

Source: own study

In the case of generated data sets, an easy-to-interpret graphic visualization of the obtained results is possible. The results for two different aggregation methods are summarized in Figure 1 and Figure 2.

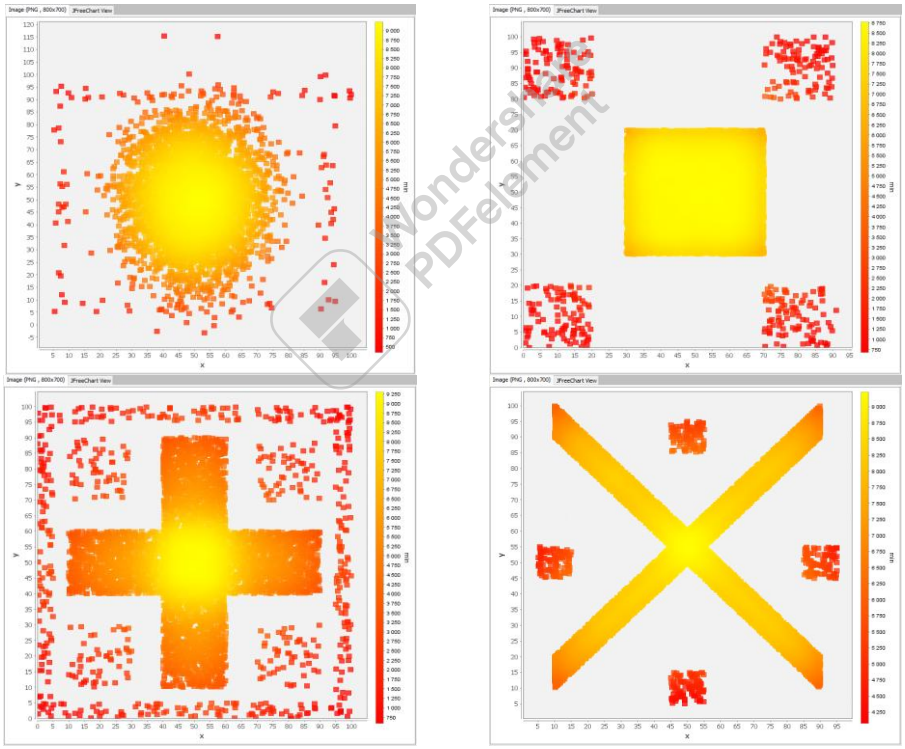

Fig. 1. Values determining the degree of anomaly when applying the minimum function to aggregation

Source: own study

By analyzing the results presented in Figure 1 and Figure 2, it can be stated that the use of the maximum function indicates outliers much more clearly. The differences obtained with this approach are more pronounced. It can be seen that the proper selection of the aggregating function is essential. Moreover, when selecting the aggregating function, one should be guided by the shape of the set and its distribution. When the maximum function is used, fewer elements are designated as outliers. On the other hand, the use of the minimum function determines the elements distant from the center as outliers. It is enough if only one coordinate is sufficiently far from the mean.

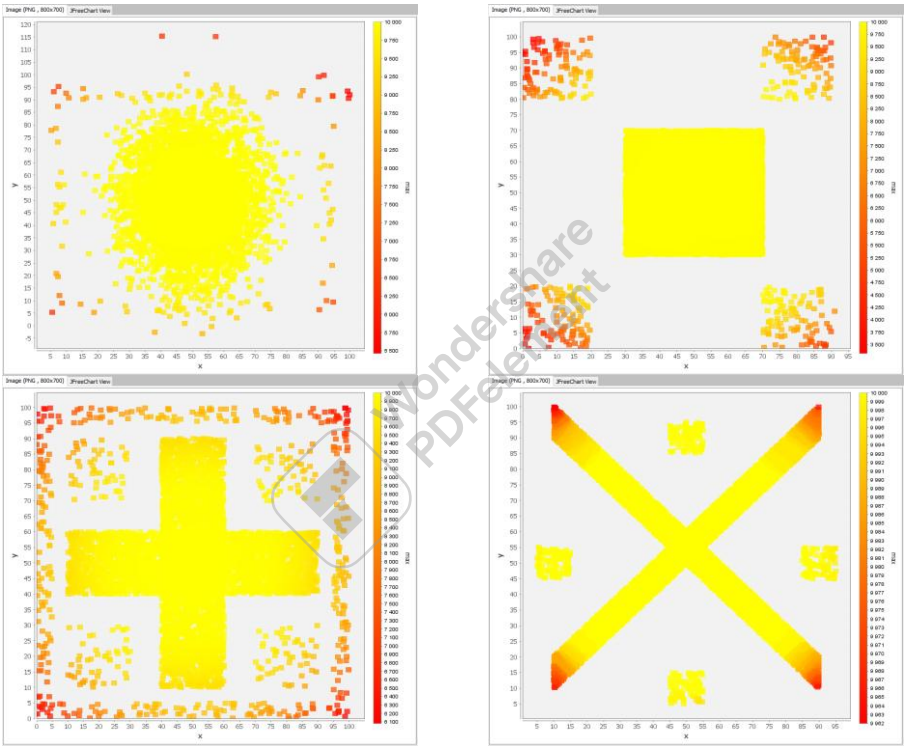

Fig. 2. Values determining the degree of anomaly when applying the maximum function to aggregation

Source: own study

The values of the ACC and PREC measures for the compared methods are presented in the Tables 2 and 3.

Table 2. ACC measure values

Database	RCOD FR	RCOD DT	EE	GM	IF	LOF	SVM
Artificial 1	0.993	0.992	0.994	0.993	0.994	0.969	0.992
Artificail 2	1	1	1	1	1	0.929	0.961
Artificial 3	0.999	0.999	0.988	0.987	0.996	0.894	0.924
Artificial 4	1	0.999	0.961	0.961	0.974	0.962	0.977
Annthyroid	0.911	0.885	0.92	0.882	0.899	0.881	0.868
Arrhythmia	0.853	0.845	0.845	0.819	0.843	0.779	0.735
Breastw	0.956	0.9478	0.933	0.933	0.939	0.461	0.388
Cardio	0.979	0.967	0.887	0.809	0.908	0.843	0.891
Cover	0.990	0.988	0.981	0.982	0.982	0.981	x
Glass	0.935	0.954	0.930	0.925	0.930	0.939	0.925
Ionosphere	0.952	0.898	0.917	0.883	0.758	0.863	0.698
Letter	0.932	0.928	0.897	0.911	0.885	0.936	0.884
Lympho	1	0.946	0.959	0.959	0.986	0.973	0.939
Mammography	0.967	0.967	0.954	0.966	0.964	0.958	0.964
Musk	1	1	0.999	0.988	0.998	0.947	0.937
Optdigits	0.992	0.993	0.942	0.943	0.943	0.947	0.946
Pendigits	0.993	0.989	0.959	0.957	0.971	0.958	0.959
Pima	0.611	0.591	0.664	0.655	0.674	0.509	0.609
Satellite	0.893	0.877	0.801	0.656	0.73	0.577	0.673
Satimage-2	0.998	0.998	0.991	0.985	0.996	0.978	0.977
Shuttle	0.934	0.994	0.963	0.98	0.993	0.87	0.913
Speech	0.981	0.975	0.968	0.969	0.968	0.97	0.967
Thyroid	0.969	0.967	0.983	0.964	0.977	0.952	0.96
Vertebral	0.742	0.7833	0.754	0.758	0.758	0.767	0.754
Vowels	0.945	0.967	0.935	0.957	0.944	0.954	0.943
Wbc	0.952	0.9521	0.937	0.939	0.952	0.91	0.913
Wine	0.985	0.969	0.93	0.853	0.86	0.845	0.891
Nad	0.997	0.993	0.542	0.489	0.510	0.533	x
Unsw0	0.676	0.977	0.646	0.645	0.581	0.557	x

Source: own study

Comparing the values of the ACC measure, it can be safely stated that the proposed solution does not differ from the effectiveness of other recognized methods. An in-depth analysis carried out on a large number of databases, consisting of data with different characteristics, allows for a thesis that the proposed method is effective and has the potential for further development.

Table 12. PREC measure values

Database	RCOD FR	RCOD DT	EE	GM	IF	LOF	SVM
Artificial 1	0.765	0.75	0.820	0.809	0.831	0.111	0.767
Artificail 2	1	1	1	1	1	0.078	0.488
Artificial 3	1	1	0.893	0.888	0.965	0.067	0.328
Artificial 4	1	1	0	0	0.332	0.03	0.409
Annthyroid	0.281	0.198	0.459	0.205	0.318	0.199	0.111
Arrhythmia	0.45	0.479	0.459	0.205	0.318	0.199	0.111
Breastw	0.949	0.932	0.904	0.904	0.912	0.23	0.126
Cardio	0.951	0.840	0.411	0.011	0.523	0.182	0.434
Cover	0.282	0.162	0.019	0.053	0.087	0.026	x
Glass	0	0	0.125	0.111	0.125	0.25	0.111
Ionosphere	0.930	0.877	0.888	0.84	0.664	0.81	0.579
Letter	0.267	0.395	0.18	0.283	0.008	0.49	0.071
Lympho	1	0.25	0.5	0.5	0.833	0.667	0.2
Mammography	0.326	0.326	0.008	0.269	0.232	0.093	0.224
Musk	1	1	0.979	0.814	0.969	0.156	0
Optdigits	1	0.911	0	0	0.013	0.087	0.053
Pendigits	0.906	0.761	0.103	0.045	0.353	0.071	0.09
Pima	0.416	0.388	0.519	0.506	0.534	0.296	0.44
Satellite	0.829	0.812	0.685	0.457	0.573	0.332	0.483
Satimage-2	0.973	1	0.629	0.371	0.845	0.114	0.07
Shuttle	0.860	0.961	0.744	0.858	0.949	0.091	0.391
Speech	0	0	0.033	0.05	0.033	0.1	0.016
Thyroid	0.419	0.4	0.656	0.28	0.538	0.022	0.196
Vertebral	0.053	0.167	0	0.033	0.033	0.067	0
Vowels	0.214	0.5	0.06	0.38	0.18	0.327	0.163
Wbc	0.571	0.667	0.429	0.45	0.571	0.19	0.2
Wine	0.833	0.8	0.556	0	0.1	0	0.3
Nad	0.995	0.992	0.542	0.469	0.491	0.514	x
Unsw0	0.950	0.983	0.723	0.722	0.672	0.654	x

Source: own study

The values of the PREC measure indicate, however, that the proposed solution is characterized by high stability in the correct classification of outliers. This is a very important property, especially if you plan to apply fuzzy set-based modifications.

Analyzing the results presented in Tables 2 and 3, it can be concluded that the proposed solution can easily compete with other recognized methods of detecting outliers. A thorough analysis of the considered measures allows to state that in most of the analyzed databases, the proposed solution has the highest values. Only in a few cases the introduced method differs slightly

from other methods. Usually, however, only one compared method is able to achieve measure values better than RCOD. In addition, it should be noted that the proposed algorithm is stable and every time returns the result of the classification, which is not always true in the case of SVM.

4. Conclusion and future work

The proposed solution for detecting outliers uses statistical data semantics and distribution properties. Through the proper selection of parameters classifying the elements as outliers, a tool was obtained, the effectiveness of which is comparable, or even better, than other recognized methods. In the further stages of developing the concept, it is planned to conduct in-depth research on increasing efficiency through more complex classification methods. In addition, it is planned to apply modifications using operations on fuzzy sets, in particular a good effect may be achieved by combining the proposed solution with anomaly detection techniques using information granules [24], [25].

Bibliography

[1] Aitkin M., Wilson G.T., Mixture models, outliers, and the EM algorithm, Technometrics, 1980, 22(3): 325–331.

[2] Abe N., Zadrozny B., Langford J., Outlier detection by active learning, In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, 504–509.

[3] Aggarwa C.C., Sathe, S., Theoretical foundations and algorithms for outlier ensembles, Acm sigkdd explorations newsletter, 2015, 17(1): 24–47.

[4] Akoglu L., Tong H., Koutra D., Graph based anomaly detection and descryption: a survey, Data mining and knowledge discovery, 2015, 29(3): 626–688.

[5] Albanese A., Pal S.K., Petrosino A., Rough sets, kernel set, and spatiotemporal outlier detection, IEEE Transactions on knowledge and data engineering, 2012, 26(1): 194–207.

[6] Amer M., Goldstein M., Abdennadher S., Enhancing one-class support vector machines for unsupervised anomaly detection, In: Proceedings of the ACM SIGKDD workshop on outlier detection and description, 2013, 8–15.

[7] Angiulli F., Pizzuti C., Fast outlier detection in high dimensional spaces, In European conference on principles of data mining and knowledge discovery, 2002, 15–27.

[8] Breunig M.M., Kriegel H.P., Ng R.T., Sander J., LOF: identifying density-based local outliers, In Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 2000, 93–104.

[9] Chandola V., Banerjee A., Kumar V., Anomaly detection: A survey, ACM computing surveys (CSUR), 2009, 41(3), 1–58.

[10] Chen Y., Miao D., Wang R., Outlier detection based on granular computing, In International conference on rough sets and current trends in computing, 2008, 283–292.
[11] Chimphlee W., Abdullah A.H., Sap M.N.M., Srinoy S., Chimphle, S., Anomaly–based intrusion detection using fuzzy rough clustering, In 2006 International Conference on Hybrid Information Technology, 2006, 329–334.
[12] Chouhan N., Khan A., Network anomaly detection using channel boosted and residual learning based deep convolutional neural network, Applied Soft Computing, 2019, 83, 105612.
[13] De la Hoz E., De La Hoz E., Ortiz A., Ortega J., Martínez-Álvarez A., Feature selection by multi-objective optimisation: Application to network anomaly detection by hierarchical self-organising maps, Knowledge-Based Systems, 2014, 71, 322–338.
[14] Duraj A., Szczepaniak P.S., Ochelska-Mierzejewska J., Detection of outlier information using linguistic summarization, In Flexible query answering systems, 2015, 101–113.
[15] Fanaee H., Gama J., Tensor-based anomaly detection: An interdisciplinary survey, Knowledge-Based Systems, 2016, 98, 130–147.
[16] Gómez J., González F., Dasgupta D., An immune-fuzzy approach to anomaly detection, In The 12th IEEE International Conference on Fuzzy Systems, 2003. FUZZ'03. 2003, 1219–1224.
[17] Habeeb R.A.A., Nasaruddin F., Gani A., Hashem I.A.T., Ahmed E., Imran M., Real–time big data processing for anomaly detection: A survey. International Journal of Information Management, 2019, 45, 289–307.
[18] Hoang X.D., Hu J., Bertok P., A program-based anomaly intrusion detection scheme using multiple detection engines and fuzzy inference, Journal of Network and Computer Applications, 2009, 32(6): 1219–1228.
[19] Izakian H., Pedrycz W., Anomaly detection in time series data using a fuzzy c-means clustering, In 2013 Joint IFSA world congress and NAFIPS annual meeting (IFSA/NAFIPS), 2013, 1513–1518.
[20] Jiang F., Chen Y.M., Outlier detection based on granular computing and rough set theory, Applied intelligence, 2015, 42(2): 303–322.
[21] Karczmarek P., Kiersztyn A., Pedrycz W., Fuzzy set–based isolation forest, In 2020 IEEE International Conference on Fuzzy Systems (FUZZ–IEEE), 2020, 1–6.
[22] Karczmarek P., Kiersztyn A., Pedrycz W., Al E., K–Means–based isolation forest. Knowledge-Based Systems, 2020, 195, 105659.
[23] Keller F., Muller E., Bohm K., HiCS: High contrast subspaces for density-based outlier ranking, In 2012 IEEE 28th international conference on data engineering, 2012, 1037–1048.

[24] Kiersztyn A., Karczmarek P., Kiersztyn K., Pedrycz W., The Concept of Detecting and Classifying Anomalies in Large Data Sets on a Basis of Information Granules, In 2020 IEEE International Conference on Fuzzy Systems (FUZZ–IEEE), 2020, 1–7.
[25] Kiersztyn A., Karczmarek P., Kiersztyn K., Pedrycz W., Detection and Classification of Anomalies in Large Data Sets on the Basis of Information Granules, IEEE Transactions on Fuzzy Systems, 2021.
[26] Knorr E.M., Ng R.T., Tucakov V., Distance-based outliers: algorithms and applications, The VLDB Journal, 2000, 8(3): 237–253.
[27] Lazarevic A., Kumar V., Feature bagging for outlier detection, In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 2005, 157–166.
[28] Li J.G., Hu X.G., Efficient mixed clustering algorithm and its application in anomaly detection. Jisuanji Yingyong, Journal of Computer Applications, 2010, 30(7), 1916–1918.
[29] Li J., Pedrycz W., Jamal I., Multivariate time series anomaly detection: A framework of Hidden Markov Models, Applied Soft Computing, 2017, 60, 229–240.
[30] Lin S.W., Ying K. C., Lee C.Y., Lee Z.J., An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection, Applied Soft Computing, 2012, 12(10): 3285–3290.
[31] Liu F.T., Ting K. M., Zhou Z.H., Isolation forest, In 2008 eighth IEEE international conference on data mining, 2008, 413–422.
[32] Liu F.T., Ting K.M., Zhou Z.H., Isolation-based anomaly detection, ACM Transactions on Knowledge Discovery from Data (TKDD), 2012, 6(1): 1–39.
[33] Malhotra P., Vig L., Shroff G., Agarwal P., Long short term memory networks for anomaly detection in time series, In Proceedings, 2015, 89–94.
[34] Micenková B., McWilliams B., Assent I., Learning outlier ensembles: The best of both worlds-supervised and unsupervised, In Proceedings of the ACM SIGKDD 2014 Workshop on Outlier Detection and Description under Data Diversity (ODD2), 2014, 51–54.
[35] Moshtaghi M., Bezdek J. C., Leckie C., Karunasekera S., Palaniswami M., Evolving fuzzy rules for anomaly detection in data streams, IEEE Transactions on Fuzzy Systems, 2014, 23(3): 688–700.
[36] Östermark R., A fuzzy vector valued KNN-algorithm for automatic outlier detection, Applied Soft Computing, 2009, 9(4), 1263–1272.
[37] Ramaswamy S., Rastogi R., Shim K., Efficient algorithms for mining outliers from large data sets, In Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 2000, 427–438.
[38] Rayana S., Akoglu L., Less is more: Building selective anomaly ensembles, ACM transactions on knowledge discovery from data, 2016, 10(4), 1–33.

[39] Rayana S., ODDS Library. Stony Brook University, Department of Computer Sciences, 2016, Available: http://odds.cs.stonybrook.edu.
[40] Rousseeuw P.J., Driessen K.V., A fast algorithm for the minimum covariance determinant estimator, Technometrics, 1999, 41(3): 212–223.
[41] Sathe S., Aggarwal C., LODES: Local density meets spectral outlier detection, In Proceedings of the 2016 SIAM international conference on data mining, 2016, 171–179.
[42] Schölkopf B., Platt J.C., Shawe-Taylor J., Smola A.J., Williamson R.C., Estimating the support of a high-dimensional distribution, Neural computation, 2001, 13(7): 1443–1471.
[43] Scitovski R., Sabo K., DBSCAN-like clustering method for various data densities, Pattern Analysis and Applications, 2020, 23(2): 541–554.
[44] Tan S.C., Ting K.M., Liu T.F., Fast anomaly detection for streaming data, In Twenty-Second International Joint Conference on Artificial Intelligence, 2011.
[45] Ting K.M., Tan S.C., Liu F.T., Mass: A new ranking measure for anomaly detection, Gippsland School of Information Technology, Monash University, 2009.
[46] Ting K.M., Zhou G.T., Liu F.T., Tan J.S.C., Mass estimation and its applications, In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, 989–998.
[47] Tsang C.H., Kwong S., Wang H., Genetic–fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection, Pattern Recognition, 2007, 40(9): 2373–2391.
[48] Wang B., Mao Z., Outlier detection based on Gaussian process with application to industrial processes, Applied Soft Computing, 2019, 76, 505–516.
[49] Wilbik A., Keller J.M., Bezdek J.C., Linguistic prototypes for data from eldercare residents, IEEE Transactions on Fuzzy Systems, 2013, 22(1): 110–123.
[50] Zhou C., Paffenroth R.C., Anomaly detection with robust deep autoencoders, In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017, 665–674.
[51] Zhu X., Pedrycz W., Li Z., Granular models and granular outliers, IEEE Transactions on Fuzzy Systems, 2018, 26(6): 3835–3846.
[52] Zimek A., Gaudet M., Campello R.J., Sander J., Subsampling for efficient and effective unsupervised outlier detection ensembles, In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013, 428–436.