A Technique for Outlier Detection Based on Heuristic Possibilistic Clustering

Dmitri A. Viattchenin

***Abstract*—**The paper deals with the problem of outlier detection in the data set. A corresponding technique is proposed in the paper. The technique is based on the heuristic approach to possibilistic clustering. A brief review of basic concepts and algorithms of the heuristic approach to possibilistic clustering is provided. A general plan of the proposed technique for outlier detection in the data set is proposed. An illustrative example is given. Some preliminary conclusions are formulated.

***Keywords*—**heuristic possibilistic clustering, outlier detection, principal allotment among fuzzy clusters, typical point.

I. Introduction

The first subsection of the introduction includes a consideration of a problem of outlier detection by using fuzzy clustering methods. Related works are considered briefly in the second subsection.

A. Preliminary Remarks

Fuzzy clustering is used when the boundaries among the clusters are uncertain and confusing. Fuzzy clustering is well established area, and fuzzy clustering algorithms are standard tools in unsupervised machine learning and applied statistics.

A possibilistic approach to clustering was proposed by Krishnapuram and Keller [1] and the approach can be considered as a special case of fuzzy approach to clustering because all methods of possibilistic clustering are objective function-based methods. On the other hand, constraints in the possibilistic approach to clustering are less strong than constraints in the fuzzy objective function-based approach to clustering and values of the membership function of a possibilistic partition can be considered as typicality degrees. So, the possibilistic approach to clustering is more general and flexible approach to clustering than the fuzzy approach. Many fuzzy and possibilistic clustering algorithms could be found in the corresponding books, for example, in [2].

One of the basic problems of data mining is the outlier detection [3]. Detecting the outliers of a data set is an important research way for data cleaning and finding new useful knowledge in many research areas. Outliers are objects, which deviate significantly from the rest of the data, so that it seems they are determined by strange process. Of course, outliers are often bad data points. On the other hand, in many applications outliers contain important information and their correct identification is crucial. An illustrative example is a computer security intrusion detection system, which finds outlier patterns as a possible intrusion attempts. Intrusion detection corresponds to a suite of techniques that are used to identify attacks against computers and network infrastructures. Anomaly detection is a key element of intrusion detection in which perturbations of normal behavior suggest the presence of intentionally or unintentionally induced attacks, faults and defects. So, identifying outliers is an important step in data mining.

The aim of the presented paper is a consideration of the problem of discovering outliers in the data. A novel technique for outlier detection is proposed in the paper. However, the previous results should be considered in the first place.

D. A. Viattchenin, National Academy of Sciences of Belarus, Minsk, Belarus (e-mail: viattchenin@mail.ru).

B. Related Works

Let us consider some fuzzy clustering-based approaches and techniques for outlier detection in the data. These approaches consider clusters of small sizes as clustered outliers. In other words, clusters which containing significantly less objects than other clusters should be considered as outliers.

In the first place, the fuzzy kernel clustering algorithm with outliers was proposed in [4]. An outstanding property of the developed FKCO-algorithm is that the FKCO-algorithm cannot obtain a satisfying clustering performance, but can identify the outliers easily with the newly defined criteria.

In the second place, robust version of the well-known maximum entropy clustering algorithm [5] is presented in [6]. This RMEC-algorithm is based on Vapnik's ε-intensive loss function [7] and its principal advantage is capability of labeling outliers in a data set using the obtained weighting factors. It should be noted, that Vapnik's ε-intensive loss function was used by Łęski for developing the εPCM-algorithm of possibilistic clustering [8].

In the fourth place, a method to estimate the noise distance in the noise clustering on the preservation of the hypervolume of the feature space was proposed by Rehm, Klawonn and Kruse in [9]. The proposed approach is independent of the number of fuzzy clusters in the data set. Moreover, this approach is not only to reduce the influence of outliers, but also to identify them.

In the fifth place, the robust interval competitive agglomeration clustering algorithm is described in [10]. This RICA-algorithm is developed to overcome the problems of outliers, the unknown clusters number and the initialization of prototypes in the clustering procedure for the interval-valued data.

In the sixth place, a method for outlier detection is proposed in [11]. The method is based on fuzzy clustering. In particular, the FCM-algorithm is first performed, and then small clusters are detected and considered as outlier clusters. Other outliers are then detected based on computing differences values of the objective function when objects are temporally removed from the data set.

So, the contents of this paper are the following: in the second section basic concepts and procedures of the heuristic approach to possibilistic clustering are considered in brief, the third section includes describing a novel technique for outlier detection and an illustrative example, in the fourth section preliminary conclusions are discussed and some perspectives for future investigations are discussed.

II. An Outline for a Heuristic Approach to Possibilistic Clustering

The first subsection of the section consists in basic concepts of heuristic approach to possibilistic clustering. Heuristic algorithms of possibilistic clustering are enumerated in the first subsection.

A. Basic Definitions of the Heuristic Approach to Possibilistic Clustering

A heuristic approach to possibilistic clustering was proposed in [12] and the approach was developed in other publications [13], [14], [15], [16]. The essence of the proposed heuristic approach to possibilistic clustering is that the sought clustering structure of the set of observations is formed based directly on the formal definition of fuzzy cluster and possibilistic memberships are determined also directly from the values of the pairwise similarity of observations. A concept of the allotment among fuzzy α-clusters is basic concept of the approach and the allotment among fuzzy α-cluster is a special case of the possibilistic partition which was introduced in [1]. All basic definitions and notations of the heuristic approach to possibilistic clustering can be founded in [12].

Heuristic algorithms of fuzzy clustering display low level of complexity and high level of essential clarity. Heuristic clustering algorithms which are based on a definition of the cluster concept are called algorithms of direct classification or direct clustering algorithms. Let us remind the basic concepts of the heuristic approach to possibilistic clustering.

Let $X = \{x_1,...,x_n\}$ be the initial set of elements and $T : X \times X \to [0,1]$ some binary fuzzy relation on X with $\mu_T(x_i, x_j) \in [0,1]$, $\forall x_i, x_j \in X$ being its membership function. Fuzzy tolerance is the fuzzy binary intransitive relation which possesses the symmetry property and the reflexivity property.

Let α be the α-level value of the fuzzy tolerance T, $\alpha \in (0,1]$. Columns or rows of the fuzzy tolerance matrix are fuzzy sets $\{A^1,..., A^n\}$ on X. Let A^l, $l \in \{1,\dots,n\}$ be a fuzzy set on X with $\mu_{A^l}(x_i) \in [0,1]$, $\forall x_i \in X$ being its membership function. The α-level fuzzy set $A^l_{(\alpha)} = \left\{(x_i, \mu_{A^l}(x_i)) \mid \mu_{A^l}(x_i) \ge \alpha, x_i \in X\right\}$ is fuzzy α-cluster. So, $A^l_{(\alpha)} \subseteq A^l$, $\alpha \in (0,1]$, $A^l \in \{A^1,\dots, A^n\}$ and $\mu_{A^l}(x_i)$ is the membership degree of the element $x_i \in X$ for some fuzzy α-cluster $A^l_{(\alpha)}$, $\alpha \in (0,1]$, $l \in \{1,\dots,n\}$. The membership degree will be denoted μ_{li} in further considerations. Value of α is the tolerance threshold of fuzzy α-cluster elements. The membership degree of the element $x_i \in X$ for some fuzzy α-cluster $A^l_{(\alpha)}$, $\alpha \in (0,1]$, $l \in \{1,\dots,n\}$ can be defined as a

$$\mu_{li} = \begin{cases} \mu_{A^l}(x_i), & x_i \in A^l_\alpha \\ 0, & otherwise \end{cases}, \tag{1}$$

where the α-level $A^l_\alpha = \{x_i \in X \mid \mu_{A^l}(x_i) \ge \alpha\}$, $\alpha \in (0,1]$ of a fuzzy set A^l is the support of the fuzzy α-cluster $A^l_{(\alpha)}$, $A^l_\alpha = Supp(A^l_{(\alpha)})$. The membership degree defines a possibility distribution function for some fuzzy α-cluster $A^l_{(\alpha)}$, $\alpha \in (0,1]$, and this possibility distribution function is denoted $\pi_l(x_i)$.

Let $\{A^1_{(\alpha)},..., A^n_{(\alpha)}\}$ be the family of fuzzy α-clusters for some $\alpha \in (0,1]$. The point $\tau^l_e \in A^l_\alpha$, for which

$$\tau^l_e = \arg\max_{x_i} \mu_{li}, \ \forall x_i \in A^l_\alpha, \tag{2}$$

is called a typical point of the fuzzy α-cluster $A^l_{(\alpha)}$, $\alpha \in (0,1]$, $l \in [1,n]$. Obviously, a fuzzy α -cluster can have several typical points. That is why symbol e is the index of the typical point.

Let $R^\alpha_z(X) = \{A^l_{(\alpha)} \mid l = \overline{1,c}, 2 \le c \le n\}$ be a family of fuzzy α-clusters for some value of tolerance threshold α, which are generated by a fuzzy tolerance T on the initial set of elements $X = \{x_1,..., x_n\}$. If condition

$$\sum_{l=1}^{c} \mu_{li} > 0, \ \forall x_i \in X, \tag{3}$$

is met for all $A^l_{(\alpha)}$, $l = \overline{1,c}$, $c \le n$, then the family is the allotment of elements of the set

$X = \{x_1,...,x_n\}$ among fuzzy α-clusters $\{A^l_{(\alpha)}, l = \overline{1,c}, 2 \le c \le n\}$ for some value of the tolerance threshold α. It should be noted that several allotments $R^{\alpha}_z(X)$ can exist for some tolerance threshold α. So, symbol z is the index of an allotment.

Thus, the problem of cluster analysis can be defined as the problem of discovering the unique allotment $R^*_c(X)$, resulting from the classification process and detection of fixed or unknown number c of fuzzy α-clusters can be considered as the aim of classification.

B. Heuristic Algorithms of Possibilistic Clustering: A Brief Review

Direct heuristic algorithms of possibilistic clustering can be divided into two types: relational versus prototype-based. A fuzzy tolerance relation T matrix is a matrix of the initial data for the direct heuristic relational algorithms of possibilistic clustering and a matrix of attributes is a matrix of the initial data for the prototype-based algorithms. In particular, the group of direct relational heuristic algorithms of possibilistic clustering includes

- the D-AFC(c)-algorithm which is based on the construction of an allotment among an a priori given number c of partially separate fuzzy α-clusters [12];
- the D-PAFC-algorithm which is based on the construction of an principal allotment among an unknown minimal number of at least c fully separate fuzzy α-clusters [12];
- the D-AFC-PS(c)-algorithm which is based on the construction of an allotment among an a priori given number c of partially separate fuzzy α-clusters in the presence of labeled object [12];
- the D-AFC(α)-algorithm which is based on the construction of an allotment among an a priori unknown number c of partially separate fuzzy α-clusters with respect to the given minimal value α of tolerance threshold [13];
- the D-AFC(u)-algorithm which is based on the construction of an allotment among an a priori unknown number c of partially separate fuzzy α-clusters with respect to the given maximal number u of elements in every class [14].

Moreover, the FG-AFC-algorithm of heuristic possibilistic clustering based on fuzzy tolerance graph decomposition was proposed in [15].

On the other hand, the family of direct prototype-based heuristic algorithms of possibilistic clustering includes [12]

- the D-AFC-TC-algorithm which is based on the construction of an allotment among an a priori unknown number c of fully separate fuzzy α-clusters;
- the D-PAFC-TC-algorithm which is based on the construction of a principal allotment among an a priori unknown minimal number of at least c fully separate fuzzy α-clusters;
- the D-AFC-TC(α)-algorithm which is based on the construction of an allotment among an a priori unknown number c of fully separate fuzzy α-clusters with respect to the minimal value α of the tolerance threshold.

The hierarchical H-AFC-TC-algorithm which is based on the construction of a hierarchy of allotments among an a priori unknown number c of fully separate fuzzy α-clusters was also proposed in [12].

It should be noted, that these prototype-based heuristic algorithms of possibilistic clustering are based on the transitive closure of the initial fuzzy tolerance. New direct prototype-based heuristic algorithms of possibilistic clustering were proposed in [16] and the family of algorithms is based on the TAGA-algorithm which is calculating different kinds of transitive approximation of the initial fuzzy tolerance [17]. So, the family of prototype-based algorithms includes

- the D-AFC-TAGA-algorithm which is based on the construction of an allotment among an a priori unknown number c of fully separate fuzzy α-clusters;

- the D-PAFC-TAGA-algorithm which is based on the construction of a principal allotment among an a priori unknown minimal number of at least c fully separate fuzzy α-clusters;
- the D-AFC-TAGA(α)-algorithm which is based on the construction of an allotment among an a priori unknown number c of fully separate fuzzy α-clusters with respect to the minimal value α of the tolerance threshold.

All prototype-based heuristic possibilistic clustering algorithms based on a transitive closure of an initial fuzzy tolerance relation are particular versions of corresponding prototype-based heuristic possibilistic clustering algorithms which based on the calculation of a transitive approximation of a fuzzy tolerance.

III. A Novel Technique for Outliers Detection

The proposed technique for outlier detection based on a heuristic algorithm of possibilistic clustering is considered in the first subsection. The second subsection of the section includes a consideration of an illustrative example of application of the proposed technique for outlier detection.

A. A General Plan of the Proposed Technique

Let us consider in detail a plan of the technique for outlier detection which was outlined in [18]. The presented very simple technique is based on sequential application of the D-PAFC-TAGA-algorithm of heuristic possibilistic clustering to a data set.

The technique is based on the assumption, that the cardinality of the outlier class is given a priory by analyst, ϕ. Moreover, the cardinality of a support of each unique outlier is equal one. There is the following seven-step procedure for outlier detecting:

1. The initial data set $X = \{x_1,\ldots,x_n\}$ should be processed by the D-PAFC-TAGA-algorithm by choosing a suitable distance $d(x_i, x_j)$ for fuzzy sets;
2. The support $Supp(A^l_{(\alpha)}) = A^l_\alpha$, $l \in \{1,\ldots,c\}$, $\alpha \in (0,1]$ should be detected for each fuzzy α-cluster in the constructed allotment among fuzzy α-clusters $R^*_c(X)$;
3. The cardinality $card(A^l_\alpha)$ of each support A^l_α, $l \in \{1,\ldots,c\}$ should be calculated;
4. The following condition is checked:
   if the condition $card(A^l_\alpha) < \phi$ is met for some fuzzy α-cluster $A^l_{(\alpha)} \in R^*_c(X)$,
   then the fuzzy α-cluster $A^l_{(\alpha)} \in R^*_c(X)$ is the outlier class and the condition $x_i \in A^l_\alpha$ is also met and these elements can be considered as outliers;
5. The following condition is checked:
   if the condition $card(A^l_\alpha) = 1$ is met for the fuzzy α-cluster $A^l_{(\alpha)} \in R^*_c(X)$,
   then the corresponding object $x_i \in A^l_\alpha$, $i \in \{1,\ldots,n\}$ should be identified as an outlier,
   else go to step 6;
6. The detected small fuzzy α-cluster $A^l_{(\alpha)} \in R^*_c(X)$, or each fuzzy α-cluster $A^l_{(\alpha)}$ from the detected family of small fuzzy α-clusters $\{A^l_{(\alpha)}\} \subset R^*_c(X)$, $card(A^l_\alpha) > 1$ should be processed by the D-PAFC-TAGA-algorithm and go to step 4;
7. The following condition is checked:
   if cardinality of each sub cluster of the outlier class $A^l_{(\alpha)} \in R^*_c(X)$ is equal one,
   then stop,
   else go to step 6.

In other words, if a fuzzy α-cluster includes only its unique typical point then the fuzzy α-cluster can be interpreted as an outlier. The distance for fuzzy sets $d(x_i, x_j)$ and the cardinality of the outlier class ϕ are parameters for the described technique.

The effectiveness of the proposed technique can be illustrated by an example, which is presented in the next subsection.

B. An Illustrative Example

Let us consider the simple illustrative example which was considered by Rehm, Klawonn and Kruse in [9]. The synthetic data set contains two good clusters and some outliers. The data are presented in Table I.

TABLE I
THE SYNTHETIC DATA SET

Numbers of objects, i	$\hat{x}^1$	$\hat{x}^2$	Numbers of objects, i	$\hat{x}^1$	$\hat{x}^2$	Numbers of objects, i	$\hat{x}^1$	$\hat{x}^2$
1	-10.44	-1.33	16	-3.68	-1.73	31	27.98	-4.01
2	14.75	-2.09	17	-8.90	-3.05	32	38.29	-0.04
3	6.78	-1.03	18	2.24	-2.04	33	22.22	-3.63
4	5.25	-0.87	19	2.91	-7.08	34	32.33	-7.45
5	-3.84	1.09	20	4.28	1.14	35	51.01	4.35
6	-1.29	8.42	21	36.04	-1.82	36	37.20	-1.33
7	-6.86	0.60	22	31.40	6.71	37	29.25	-7.83
8	-6.34	3.25	23	38.79	5.04	38	39.43	-1.97
9	-4.47	10.40	24	26.78	1.07	39	33.58	-0.72
10	2.95	-1.70	25	29.72	-1.50	40	40.18	-11.67
11	9.21	3.65	26	33.74	1.28	41	30.00	-18.00
12	6.74	1.47	27	28.51	-0.95	42	-20.00	-1.00
13	-2.46	-4.25	28	41.17	-0.40	43	15.00	-20.00
14	-10.89	-12.67	29	42.47	3.50	44	25.00	-22.00
15	1.19	-11.89	30	36.18	-3.98	45	-5.00	-25.00

This data were pre-processed according to a formula:

$$x_i^{t_1} = \frac{\hat{x}_i^{t_1} - \min_i \hat{x}_i^{t_1}}{\max_i \hat{x}_i^{t_1} - \min_i \hat{x}_i^{t_1}}, \; t_1 = 1, \ldots, m_1, \; i, j = 1, \ldots, m_1 \tag{4}$$

and the normalized Euclidean distance [19] was selected as a parameter for the D-PAFC-TAGA-algorithm:

$$e(x_i, x_j) = \sqrt{\frac{1}{m_1} \sum_{t_1=1}^{m_1} \left(\mu_{x_i}(x^{t_1}) - \mu_{x_j}(x^{t_1}) \right)^2}, \; i, j = 1, \ldots, n. \tag{5}$$

The threshold ϕ was selected in the experiment to n/c, where the number of clusters c is equal 2. The proposed technique was applied to the presented data set and the result is shown in Figure 1.

In the first place, the outlier class $A_{(\alpha)}^2 = \{x_{41}, x_{43}, x_{44}, x_{45}\}$ was detected in the first iteration of the presented technique. This class was processed by the D-PAFC-TAGA-algorithm at next iterations of the proposed technique. At finally, the object x_{45} was identified as an outlier in the first place. The object x_{41} was identified as an outlier in the second place. Objects x_{43} and x_{44} were detected as outliers in the third and fourth place consequently. So, outliers were detected as an ordered sequence $x_{45} \prec x_{41} \prec x_{43} \prec x_{44}$. The objects of "good" clusters are denoted in Figure 1 by symbol ● and detected outliers are denoted in Figure 1 by symbol □.

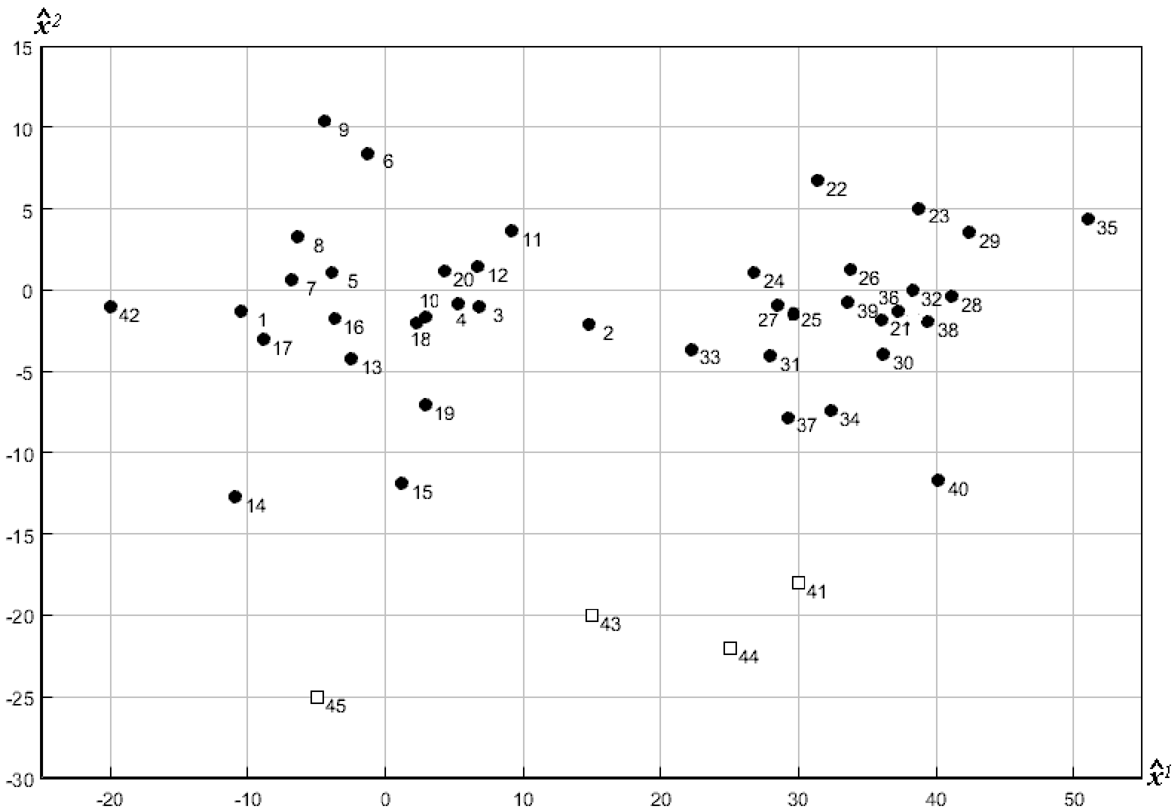

Fig. 1 The data set and detected outliers

IV. Conclusions

Preliminary conclusions are discussed in the first subsection of the section. The second subsection deals with the perspectives on future investigations.

A. Discussions

Outliers can contain information about abnormal behavior of the system and their correct identification is significant. So, the novel technique for outlier detection is proposed in the paper. The technique is based on sequential application of the D-PAFC-TAGA-algorithm to the data set. That is why the ordered sequence of detected outliers is a principal advantage of the proposed technique. This approach can be useful, for example, for assignment of air targets in anti-aircraft defense systems.

The results of application of the proposed technique to the Rehm, Klawonn and Kruse two-dimensional artificial data set [9] show that the technique is an effective tool for solving the outlier detection problem in the framework of exploratory data analysis.

B. Perspectives

Let us consider some perspectives for further investigations. Firstly, the proposed technique for outlier detection should be extended for a case of the large data set. Secondly, the sequential approach to outlier detection based on heuristic possibilistic clustering should be extended for a case of the relational initial data. Thirdly, the idea of sequential outlier detection can be used in the method of outlier detection in the interval-valued data [20].

These perspectives for investigations are of great interest both from the theoretical point of view and from the practical one, as well.

ACKNOWLEDGMENT

I am grateful to Prof. Janusz Kacprzyk and Dr. Jan W. Owsinski for their interest in my investigations. I also thank to Mr. Aliaksandr Damaratski for elaborating experimental software.

REFERENCES

[1] R. Krishnapuram and J.M. Keller, “A possibilistic approach to clustering”, *IEEE Transactions on Fuzzy Systems*, vol. 1 (2), pp. 98-110, 1993.

[2] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler, *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition*. Wiley: Chichester, 1999.

[3] V.J. Hodge and J. Austin, “A survey of outlier detection methodologies”, *Artificial Intelligence Review*, vol. 22 (2), pp. 85-126, 2004.

[4] H. Shen, J. Yang, and S. Wang, “Outlier detection by weighed mercer-kernel based fuzzy clustering algorithm”, *Iranian Journal of Electrical and Computer Engineering*, vol. 4 (2), pp. 129-136, 2005.

[5] R. Li and M. Mukaidono, “Gaussian clustering method based on maximum-fuzzy-entropy interpretation”, *Fuzzy Sets and Systems*, vol. 102 (2), pp. 253-258, 1999.

[6] S.T. Wang, F.L. Chung, Z. Deng, D. Hu, and X. Wu, “Robust maximum entropy clustering algorithm with its labeling for outliers”, *Soft Computing*, vol. 10 (7), pp. 555-563, 2006.

[7] V.N. Vapnik, *Statistical Learning Theory*. Wiley: New York, 1998.

[8] J. Łęski, “Robust possibilistic clustering”, *Archives of Control Sciences*, vol. 10 (3/4), pp. 141-155, 2000.

[9] F. Rehm, F.Klawonn, and R. Kruse, “A novel approach to noise clustering for outlier detection”, *Soft Computing*, Vol. 11 (5) pp. 489-494, 2007.

[10] J.-T. Jeng, C.-C. Chuang, and C.-J. Juan, “Robust interval competitive clustering algorithm with outliers”, *International Journal of Fuzzy Systems*, vol. 12 (3), pp. 227-236, 2010.

[11] M. Al-Zoubi, A. Al-Dahoud, and A.A. Yahya, “New outlier detection method based on fuzzy clustering”, *WSEAS Transactions on Information Sciences and Applications*, vol. 7 (5), pp. 129-136, 2010.

[12] D.A. Viattchenin, *A Heuristic Approach to Possibilistic Clustering: Algorithms and Applications*. Heidelberg: Springer-Verlag, 2013.

[13] D.A. Viattchenin, “A novel heuristic algorithm of possibilistic clustering for given minimal value of the tolerance threshold”, *Journal of Information, Control and Management Systems*, vol. 13 (2), pp. 161-174, 2015.

[14] D.A. Viattchenin, “Heuristic possibilistic clustering for detecting optimal number of elements in fuzzy clusters”, *Foundations of Computing and Decision Sciences*, vol. 41 (1), pp. 45-76, 2016.

[15] Viattchenin, D.A., Nikolaenya, E., Damaratski, A. “A fuzzy grapf-based heuristic algorithm of possibilistic clustering”, *Communications on Applied Electronics*, 3 (7), (2015), pp. 13-23.

[16] D.A. Viattchenin and A. Damaratski, “Direct heuristic algorithms of possibilistic clustering based on transitive approximation of fuzzy tolerance”, *Informatica Economicá*, vol. 17 (3), pp. 5-15, 2013.

[17] P. Dawyndt, H. De Meyer, and B. De Baets, “UPGMA clustering revisited: a weight-driven approach to transitive approximation”, *International Journal of Approximate Reasoning*, vol. 42 (3), pp. 174-191, 2006.

[18] D.A. Viattchenin, “Sequential outliers detection by using heuristic possibilistic clustering”, *Proceedings of the XI International Conference on Computer Data Analysis and Modeling (CDAM’2016)*, Minsk, Belarus, pp. 317-318, September 2016.

[19] A. Kaufmann, *Introduction to the Theory of Fuzzy Subsets*. New York: Academic Press, 1975.

[20] D.A. Viattchenin, “Detecting outliers in interval-valued data using heuristic possibilistic clustering”, *Journal of Computer Science and Control Systems*, vol. 5 (2), pp. 39-44, 2012.

