A Review of Outliers: Towards a Novel Fuzzy Method for Outlier Detection

A. Mazidi[1]*, F. Roshanfar[2], V. Parvin Darabad[2]

Abstract –Outliers and outlier detection are among the most important concepts of data processing in different applications. While there are many methods for outlier detection, each detection problem needs to be solved with the method most suited to its unique characteristics and features. This paper first classifies different outlier detection methods used in different fields and applications to provide a better understanding, and then presents a new fuzzy method for outlier detection. The proposed method uses the fuzzy logic and the local density to assign a point to data instances, and then determines whether a piece of data is normal or outlier based on the value of resulted membership function. Evaluation of the proposed outlier detection algorithm with synthetic datasets demonstrates its good accuracy; moreover, evaluation of the performance in solving real datasets show that the proposed method outperforms the k-means and K-NN algorithms.

Keywords: Outliers, Outlier Detection, Outlier Detection Applications, Fuzzy.

I. Introduction

The massive and growing volume of data stored in modern databases have spawned the need for robust methods of data analysis. One important subject, attracting attention, is the detection of inconsistent observations also known as outliers in different applications. An outlier is a piece of data that does not fit the expected pattern. According to Hawkins (1980),apiece of data is called an outlier when it exhibits large deviation from the rest of the data in the database, and this deviation is so much that it appears the data has to come from a different mechanism [1]. From another angle of view, within a database DB (*pct*, $d_{\min}$), outlier object *p* is a piece of data located at a distance greater than $d_{\min}$ from *pct* percentage of other objects in database. This definition leads to detection of only a certain type of outliers [2]. Fig.1 shows an instance of outliers versus normal data. In Figure (1), clusters N1 and N2consists of normal data, but objects and clusters denoted by O1, O2 and O3 have significant deviation from the rest of the data, thus they must be detected as outliers.

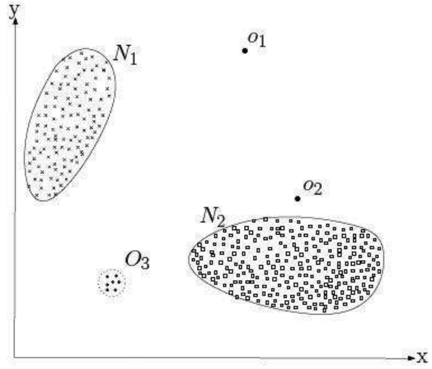

Fig. 1: Normal data and outliers [3]

The purpose of outlier detection methods is to find the patterns in which data do not behave as expected. Outlier detection have wide ranging application such as military monitoring and surveillance, intrusion detection in cyber security context, fraud detection for credit cards, health care services, and fault detection in critical systems. For example, an unusual traffic pattern in a computer network might mean a hacked computer is transferring confidential information out to the network from an unauthorized computer [4]. In the field of medical image processing, e.g., an unusual pattern in an MRI image can focus attention of the physician to a malignant tumor [5],outliers in the credit card transactions can indicate that the credit card is stolen [6], and unusual reading of sensors of a spacecraft can point to presence off laws in its components [7].

The presence of outliers, among others, can be the product of fraudulent activities such as fraud in credit cards,

1* **Corresponding Author :** Department of Computer Engineering, Faculty of Engineering, Golestan University, Gorgan, Iran. Email: Arash_mazidi_67@yahoo.com

2 Department of Electrical Engineering, Faculty of Engineering, Golestan University, Gorgan, Iran.

Received 2018.10.12 ; Accepted 2019.06.04

cyber intrusions, terrorist activities, and system failures. Outlier detection methods analyze the available data to acquire useful and discernable information about various aspects of studying the subject. This paper seeks to provide a better understanding of outliers and their applications and the related detection methods. Each application often produces their own characteristics. Hence, there are many different types of data, but one of the most widely used type is the data stream. Data streams possess a number of features which distinguish their processing from the other types of data. In this paper, the focus is to provide a novel method for outlier detection in data streams. The results of implementation will demonstrate the high accuracy of the proposed method.

In the rest of this paper, Section 2reviews the previous works on this subject. Section 3discusses the challenges ahead of effective outlier detection in different applications. Section 4describes the proposed algorithm, section 5, discusses the implementation and evaluation of proposed algorithm, and finally, section 6 concludes the paper and provides road map for future works.

II. Related Works

During the past decades there have been many studies on the outlier detection in different applications. In the following, we have provided the reader with a brief outline of most relevant researches of the outlier detection. Gogi *et al.* [8] assessed and compared distance and density based methods of outlier detection in network application. Fava *et al.* [9] classified, implemented and compared data mining-based fraud detection methods. Agreval & Han *et al.* [10] reviewed and assessed outlier detection methods for time-dependent data. Hudge & Austin [11] evaluated outlier detection methods in the field of machine learning and statistics. Malik *et al.* [12] presented numerical and symbolic methods of outlier detection. Aminesh *et al.* [13] developed a review on cyber intrusion detection methods. Zoriana *et al.* [14] used regression techniques and Euclidean distance for outlier detection. The new outlier detection method presented in[14] was based on the object's closest neighbors. This method was presented for the data processing of streams and used a sliding-window model for this purpose.

In [15], an outlier detection technique based on the entire dataset (global outliers) and the neighbor data (local outliers) was proposed. This technique is based on Global Deviation Factor (GDF) and Local Deviation Factor (LDF). GDF represents the deviation of a data point with respect to the entire data points while LDF relies on deviation of a data point from the recent data. Both factors are calculated from neighboring density. A data point is identified as an outlier if either its GDF or LDF is away from its average more than three times of its standard deviation. Using these measures, the user requires no longer to select cut-off limits. Last but not the least, Chawla *et al.* [16] presented the k-means and K-NN algorithms for outlier detection. They evaluated the algorithms on real and synthetic data and showed that the k-means algorithm has better accuracy than the K-NN algorithm..

III. Outlier Detection Applications

A. Intrusion detection

The Intrusion detection is one of the best approaches of detecting abnormal and atypical behavior in a computer system. The purpose of intrusion detection is to identify destructive behavior in computer related systems, e.g. computer intrusions and misuse of computer systems. Effective intrusion detection is faced with many challenges such as large volume of data; fast pace of the data streams and lack of access to data labels.

The intrusion detection systems are classified into host-based and network-based systems. The difference between the two systems is that, in the host-based system the outlier appears as a disruptive code and atypical behavior in the operating system. In the network-based system, on the other hand, outliers occur on network transmission data due to an attack on the network. Methods used in intrusion detection applications are listed in Table (1).

B. Fraud detection

Fraud is referred to the criminal behavior inside an economic organization such as banks, credit cards and insurance agencies, mobile phone companies and stores. A fraudulent user is a real user or individual who have illegally accessed identity of another user. The purpose of fraud detection activities is to pinpoint unauthorized usage of resources in order to prevent economic losses.

The general approach for fraud detection is to record and analyze all activities of customers and users, allowing deviations of activities to be spotted as an outlier or fraud. An important application in the topic is the credit card fraud detection, which is identification of fake credit cards or unauthorized usage of a valid credit card. Reader is referred to Table (1) for few methods used in fraud detection.

C. Mobile phone fraud detection

This problem is about mobile phone activities including dialed and received calls. The phone activities are usually shown as vector of call time and call location. Outliers and fraud in this application indicate long call duration to a certain place etc. The methods presented in Table (1) can be used for this aspect of fraud detection.

D. Fraud detection in insurance claims

The important issue of property insurance industry is fraud in damage claims and demand of damage compensations from the insurance agency. For example, in car insurance, individuals and manufacturers can use unauthorized and illegal access to tamper with claim processing systems.

Data that can be used for detecting these frauds are the evidence recorded by claimants. Fraud detection methods extract different features from these evidences. Usually insurance inspectors and experts assess user claims. The neural network is the basis of the most fraud detection methods used in this application [17].

E. Medical diagnosis

Information of patients such as age, blood types etc. are often stored as raw data. These data can be used to detect an outlier caused by different reasons such as atypical conditions of the patient or error in medical equipment. Patients who do not have any problem are usually labeled as such and this is why semi-supervised methods must be used for this application. In this application, outlier detection is of significant importance, as it deals with health and lives of patients; thus the methods used in this application must be very accurate. Some of the used methods in this field are shown in Table (1).

Table 1: applications of outlier detection methods

Intrusion detection	Host based	Statistical method (histogram)	A detection algorithm in linear time has been presented and a heuristic method has been used for parameters adjustment [18].
		hybrid models	A model with conventional and atypical training data has been created. Machine learning has been used to estimate distribution and statistical methods has been used for test [19].
		Neural networks	A neural network based software intrusion detection method has been presented [20].
		Support Vector Machines (SVM)	A probabilistic model on image space has been presented and its use for mammography image analysis has been demonstrated [2].
		Rule based system	Some models has been obtained which use system data to detect system intrusions [21, 22].
	Network based	Statistical method (histogram)	Atypical behavior of the program has been used identified with the use of system data [23]. A model with label data system has been created for intrusion detection [24].
		Statistical method without parameters	A system has been presented to estimate density without parameter based on typical data [25].
		Bayesian network	A high efficiency Bayes network model has been proposed for traffic analysis [26] and a system has been presented to detect intrusion in web based programs [27].
		Support Vector Machines (SVM)	A model with conventional and a typical data base has been created. Machine learning has been used to estimate distribution and statistical methods has been used for test [19].
		Rule based system	A system has been presented for showing the relationship between data mining and intrusion detection through analysis and mining of auditing data [28].
		Neural networks	Recurrent neural network has been presented as an algorithm for detecting outliers [29].
Fraud detection	Neural networks	An online system has been presented for fraud detecting in credit cards using neural classifier [30].	
	Rule based system	A hybrid method of data mining and neural network methods has been presented for detecting outlier with Low false alarm rate [31].	
	Cluster analysis	A method has been presented for detecting fraud behavior in credit cards using cluster analysis methods for data without labels [32].	
	Neighbor based	A method has been presented for fraud detection in credit cards using *k* neighbors [33].	
Mobile phone fraud detection	Statistical method (histogram)	A system has been presented for monitoring activities and identifying new and fascinating behavior using statistical method [34]. The fraud in mobile phones has been detected through a combination of visualization of phone information and mining algorithms, [35].	
	Parametric statistical method	A method has been presented for monitoring customer transaction in order to identify deviations from customer patterns [36].	

	Neural networks	A method has been presented through neural networks for fraud detection in mobile phones and change in pattern of phone use has led to the use of neural computing solutions [37].
Outlier detection in health and medicine	Parametric statistical method	A statistical method has been presented for detecting outliers and exceptions in large scale medical data stream [38].
	Bayesian network	To detect outbreaks, recent data has been compared with the model that was created with the base data. A method has been presented for creating base model by using Bayesian network [39].
	Neighbor based method	A method has been presented for detecting outliers in data stream such as heart rate data by using neighboring data [40].

F. Detection of industrial damage

Industrial units must constantly deal with components damaged because of repetitive use and wear and tear. Some damages should be identified quickly to avoid amplified or secondary damage. In this application, the basic data must be obtained by analysis of information collected from sensor nodes. Failures occurring in industrial units can be classified into mechanical failures (such as engine failure) and structural failure. For each set of failure, there are a number of detection methods which are shown in Table (2).

G. Image processing

The purpose of outlier detection in image processing is the identification of image changes over time and/or the identification atypical area in an image. Processing of satellite images, spectroscopy, mammography and video surveillance fall into this category. Here, Outlier is result of presence or movement of a foreign object and/or error in equipment. The main issue in this application is the large volume of input data (like a video). Table (2) shows the outlier detection methods in the field of image processing.

H. Outlier detection in text documents

Any data presented in form of text documents can have many aspects and characteristics. Outlier detection methods may be able to detect new topic, event or news in a story or a series of articles and papers. Another application of outlier detection in text is the identification of literary and scientific plagiarism. Table (2) lists some of the outlier detection methods applicable to text documents.

J. Sensor networks

In sensor networks, the data sent from various sensors distributed in environment in accordance with application, arrives in the center and is subjected to analysis. Outliers in this data indicate failure in one sensor or the occurrence of expected event in the sensor. For example, in the application of protecting military perimeters by means of sensors, data is sent to the center in succession and when a sensor observes something particular, it sends different data which indicates an event in the location. Outlier detection methods can be used to identify this event.

K. Other fields

Outlier detection methods are used in many different applications and some of these applications were mentioned. Table (3) shows a list of other applications that use outlier detection methods.

Table 2: applications of outlier detection methods

Detection of industrial damage	Mechanical damage	Parametric statistical method	In this application, data is in form of streams, the presented method is for detecting outliers in data stream and identifies the outlier in linear time and space [41, 42].
		Neural network	A 3 stage method has been presented for failure detection. It uses neural network for clustering and the analysis of distribution function [43].
		Rule based system	An algorithm has been presented based on association rule for outlier and failure detection in spacecraft systems [44].
	Structural damage	Statistical method (histogram)	A model has been created based on conventional data, and the data that falling outside the model perimeter has been identified as outlier [45].
		Combined method	An outlier detection method has been presented to identify outliers in the pressure sensor data of aircrafts [46].
		Neural networks	The neural network has been used to separate system's favorite changes from other changes such as failures and structural deterioration [47].
Image processing	hybrid methods	A probabilistic model on image space has been presented and its use for analyzing mammography image has been demonstrated [2].	

	Bayesian network	An algorithm has been presented = for classifying objects and detecting outliers in video surveillance [48].
	Support Vector Machine (SVM)	A Hybrid method has been presented for separating and segmenting audio signal using SVM [49].
	Neural networks	An outlier detection framework has been presented where neural network has been presented used as an adaptive classifier [50].
	Neighbor based methods	Each object has been assigned with a score based in its neighboring objects, which represent its outlierness [51].
Outlier detection in text documents	Parametric statistical method	A statistical methods has been presented for monitoring activities and identifying new behavior in the text [34].
	Support Vector Machine (SVM)	A SVM based method has been presented for classifying and retrieving data in the text [52].
	Data with no label	A fraud detection method has been presented detecting plagiarism and misuse in the unlabeled text [53].
Sensor network	Bayesian networks	A Bayesian network classifier has been presented and the data of network sensors has been allocated to conventional and atypical classes [54].
	Rule based system	A method with flexibility, reduced energy consumption, bandwidth etc. has been presented for detecting outliers in sensor data streams [55].
	Parametric statistical method	A method has been presented for detecting attacks on sensor networks [56, 57].
	Neighbor based	A method has been presented for detecting neighbor based and density based outliers in sensor networks [58].

Table 3: applications of outlier detection methods

Movement detection in robots	A method has been presented for audio and video tracking of the robot's movement [59, 60].
Traffic monitoring	A method has been presented for traffic monitoring by using the k-means clustering algorithm [61].
Error detection in web programs	A web based system has been modeled to a graph. Nodes represent services and edges represent their relations, and errors have thereby been identified [62].
Outlier detection in biology	Two methods have been presented for classifying genes [63], and for associative analysis of biology databases [64].
Outlier detection in census	A method has been presented for detecting distance based outliers on census data [65].
Dependence detection in criminal activity	A method has been presented for detecting criminal activity using associative and functional methods and by assigning points to data [66].
Outlier detection in astronomy	A system has been presented for detecting *k* outliers from astronomy data [67, 68].

IV. The Proposed Fuzzy Outlier Detection Method

The In this algorithm, each piece of data is assigned with as core or point which represents its presence or absence in the cluster. This score actually shows the amount of data density in database. This factor is local and is obtained based on the neighbors of each object. We will show that the score of all objects in the clusters is close to 1, and that objects outside the cluster will have a score close to zero.

Desired point for each data only depends on the value of parameter *Minpts*, which defines the number of close neighbors for each objects.

The presented algorithm is fuzzy (not binary), as it determines the membership in the cluster with respect to the obtained point for each data. The closer the membership is to one, the more definitive the membership is and the closer the membership is to zero, the more it shows non-membership.

In the following, the formula used to determine the point is defined as definitions based on a hierarchy.

Definition 1: *k*-distance of an object *p*

k-distance for object *p* is shown with *k*-distance(*p*) and is equal to distance(*p,o*) where *o* is an object in the database with below conditions:

1. For at least *k* objects $o' \in D \setminus \{p\}$ it holds that $d(p,o') \le d(p,o)$.
2. For at most *k*-1 objects $o' \in D \setminus \{p\}$ it holds that $d(p,o') \le d(p,o)$.

If we put *k* number of data inside a circle with data p as the center, the radius of the circle will be equal to the size of *k*-distance(*p*).

Definition 2: *k*-distance neighborhood of an object *p*

This definition includes all objects whose distance to *p* is not more than *k*-distance(*p,o*). Equation (1) shows that:

$$N_{\text{k-distance(p)}}(p) = \{ q \in D \backslash \{p\} \mid d(p, q) \leq \text{k-distance}(p) \} \quad (1)$$

Object *q* in equation (1) is called the *k* nearest neighbors of *p*. In definition 1, for better understanding, the size of *k*-distance(*p*) is proposed to be obtained with the use the radius of circle. Data located within and on the circumference of the circle are called nearest neighbors of *p*.

Definition 3: reachability distance of an object(*p*) w.r.t object(*o*)

This parameter shows the reachability distance of object *p* to object *o*, and can be calculated through equation (2).

$$\text{reach-dist } k(p, o) = \max \{ \text{k-distance}(o), d(p, o) \} \qquad (2)$$

Figure (2) shows the concept of reachability distance for *k*=3 (*k* is the number of nearest neighbors).

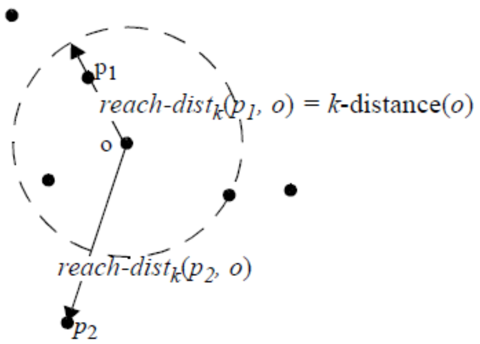

Fig. 2: reachability distance of object p_1 and p_2 to object *o* for *k*=3

As can be seen, for p_2 the reachability distance is equal to distance(p_2,o), and for p_1 distance is equal to 3-distance(o). The common density-based clustering algorithms use two parameters for determination of density. One parameter is *minpts* which determines the lowest value of neighboring objects and the other parameter determines the volume. These two parameters define the limits and boundaries for clustering algorithms. But the algorithm proposed in this study depends only on the value of *Minpts*.

Therefore we only keep the parameter *minpts* and use the values of reach-dist$_{MinPts}(p, o)$, for $o \in N_{MinPts}(p)$ for measuring the volume and determining the density in the neighborhood of object *p*.

Definition 4: local reachability density of an object

Local reachability density for object *p* is calculated through equation (3).

$$lrd_{MinPts}(p) = \frac{1}{\left(\frac{\sum_{o \epsilon N_{MinPts}(p)} reach\text{-}dist_{MinPts}(p,o)}{|N_{MinPts}(p)|}\right)} \quad (3)$$

This value is the inverse of reachability distance for near objects to object *p*, which can be seen in equation (3).

It should be noted that when the total reachability distances in neighboring objects, which can be seen in the denominator, is equal to zero the local density becomes infinite. This happens when there are at least *minpts* number of objects like object *p* in the database; in this case, the distances will be equal to zero. Here, we assume that there is no duplicate object and this situation never occurs.

Definition 5: point for object *p*

This point shows the membership value of object p in the clusters and is defined by equation (4).

$$Temp = \frac{\sum_{o \epsilon N_{MinPts}(p)} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{|N_{MinPts}(p)|} \qquad (4)$$

$$Point_{\,MinPts}(p) = \left\{\begin{matrix} \frac{1}{Temp} & Temp \geq 1 \\ Temp & Temp < 1 \end{matrix}\right\} \qquad (5)$$

This factor indicates that which objects are considered inside the cluster and which are not. When the value obtained from equation (5) is 1, the data is located in a location with high density and is considered to be normal data in the cluster. When the value is close to zero, the density around the data is low and there is an increased likelihood that data is not a member of the cluster that is shown in equation (6).

$$Point_{\,MinPts}(p) = \left\{\begin{matrix} 0 & \text{non} - \text{membership} \\ 1 & \text{crisp membership} \end{matrix}\right\} \qquad (6)$$

The membership of data in a cluster is determined by the use of a fuzzy threshold value. This fuzzy threshold value must be obtained experimentally and will be different for different applications. For example, in medical and sensitive applications this value must be close to one but in other applications this value can be set with greater flexibility.

There are two main advantages in the proposed algorithm:

1. *This algorithm identifies the clustering with respect to the density of the object's neighbors and is not a*

global model.

2. *This algorithm can cluster the data with any distribution and is not based on any particular data distribution.*

This algorithm does not need to determine the number of cluster at the beginning of the algorithm.

V. Implementation and Evaluation

Evaluation of algorithms needed some benchmark datasets. In this study, two datasets with normal and uniform distribution were created for this purpose. All results were obtained using a computer with Intel core 2 Duo T9300 processor with 2.5 GHz frequency, 4 GBs of RAM, and the final version of windows 7 operating system. The algorithms were implemented using C# in visual studio 2014. The dataset with normal distribution was created withMatlabR2014, and the dataset with uniform distribution was created with C# in VS.

The dataset with normal distribution included 500 data instances created with normal distribution $N_1(\mu_1,\Sigma_1)$, 500 data instances created with normal distribution $N_2(\mu_2,\Sigma_2)$ and 500 data instances created with normal distribution$N_3(\mu_3,\Sigma_3)$. This dataset also includes four low rate data batches with 25 data instances and normal distributions of $N_4(\mu_4,\Sigma_4)$ ،$N_5(\mu_5,\Sigma_5)$ ،$N_6(\mu_6,\Sigma_6)$ and $N_7(\mu_7,\Sigma_7)$. Distribution parameters of the batches are defined as follows.

$$N_1(\mu_1,\Sigma_1) \quad ; \quad \mu_1= [+1,+1] \quad ; \quad \Sigma_1= \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}$$

$$N_2(\mu_2,\Sigma_2) \quad ; \quad \mu_2= [-1,-1] \quad ; \quad \Sigma_2= \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}$$

$$N_3(\mu_3,\Sigma_3) \quad ; \quad \mu_3= [+1,-1] \quad ; \quad \Sigma_3= \begin{bmatrix} 0.03 & 0 \\ 0 & 0.03 \end{bmatrix}$$

$$N_4(\mu_4,\Sigma_4) ; \mu_4= [0,+1.5]; \Sigma_4= \begin{bmatrix} 0.002 & 0 \\ 0 & 0.002 \end{bmatrix}$$

$$N_5(\mu_5,\Sigma_5) ; \mu_5= [+1.5,0]; \quad \Sigma_5= \begin{bmatrix} 0.002 & 0 \\ 0 & 0.002 \end{bmatrix}$$

$$N_6(\mu_6,\Sigma_6) \quad ; \quad \mu_6= [0,0] \quad ; \quad \Sigma_6= \begin{bmatrix} 0.002 & 0 \\ 0 & 0.002 \end{bmatrix}$$

$$N_7(\mu_7,\Sigma_7) \quad ; \quad \mu_7= [-1,+1] \quad ; \quad \Sigma_7= \begin{bmatrix} 0.002 & 0 \\ 0 & 0.002 \end{bmatrix}$$

In this dataset, the goal is to identify data batches having allow density rate. Figure (3) shows a view of the introduced two-dimensional datasets.

In the dataset created with normal distribution, data instances of cluster N_1 were entered into the system and then data clusters N_4 and N_5 were distributed among data of cluster N_2 and entered into the system. Finally, data clusters N_6 and N_7were distributed among data of cluster N_3 and entered into the system.

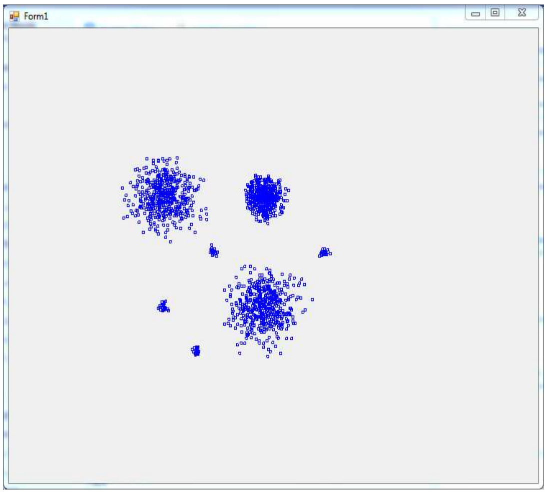

Fig. 3: Synthetic dataset with normal distribution

The other created dataset is the one with uniform distribution. This two-dimensional dataset has 1600 data instances including normal data and outliers. The dataset was created with C# programming language. The mentioned dataset includes seven clusters with uniform distribution but different density and dimensions. Outliers and atypical data are distributed with the probability of 0.02 among the normal data located in the clusters. The purpose of testing this dataset is to identify the distributed outliers and to test data with uniform distribution and different densities. Figure (4) shows a view of the introduced two-dimensional datasets.

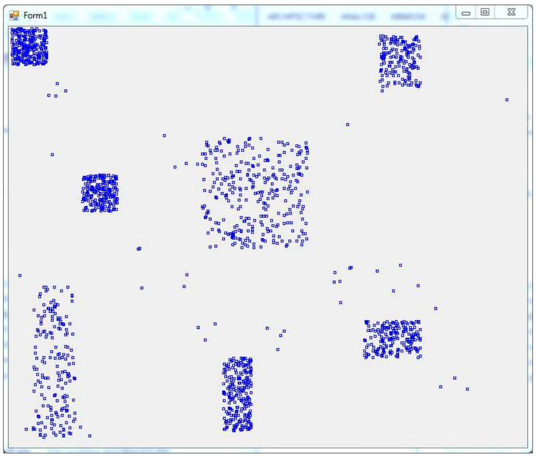

Fig. 4: Synthetic dataset with uniform distribution and different densities

Next, the presented algorithm was run on dataset with normal distribution. The result can be seen in figure (5). In this figure, identified normal data are shown with blue and identified outliers are shown with red.

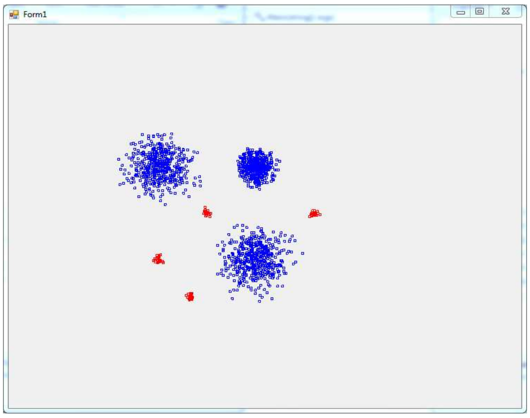

Fig. 5: The results of the proposed algorithm

As shown in figure (5), the algorithm was successful in identifying all outliers and minimized the false positive rate to zero.

The presented algorithm was also evaluated by implementing it on the other created dataset. The purpose of this dataset is to assess the accuracy of the proposed algorithm. In practice, the accuracy of algorithm, the false alarm rate and also the detection rate of the algorithm were evaluated. In figure (6), normal data is shown with blue and outliers are shown with red.

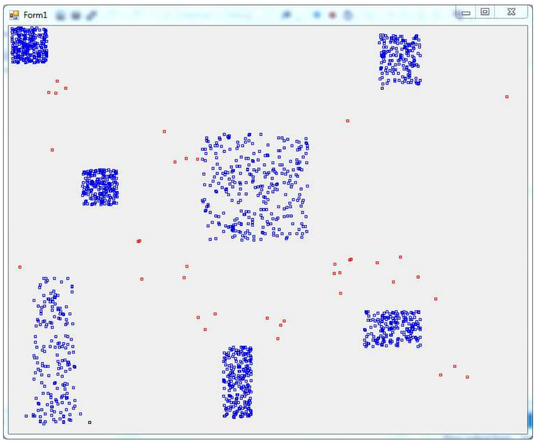

Fig. 6: The results obtained from the implementation of algorithm on dataset with uniform distribution

In another evaluation, the proposed algorithm was tested on real datasets KDD CUP 99 and Shuttle, and was compared with the k-means and K-NN algorithms.

The first real dataset is *kdd_cup* 99 that contains instances describing connections of sequences of *tcp* packets. Each instance is annotated with respect to being normal or an intrusion, and in the latter case, with the intrusion type. We experiment with a 10% sample, provided by the organizers, which contains 494 021 instances. In total, there are 23 classes, and 3 of them account for 98.3% of the whole dataset. We consider these three classes as non-outliers, and we target to discover all other 20 classes as outliers.

We use the shuttle dataset as second real dataset, which is publicly available in the UCI Machine Learning Repository [10]. The dataset contains 9 numerical attributes, and one categorical that can be interpreted as a class label. We use the training part of the dataset, which consists of 43500 instances. There are 7 distinct class labels. The three largest classes account for the 99.6% of the dataset. We consider these three classes as non-outliers, and the set to identify as outliers the rest four classes that account for 0.4% of the dataset.

To evaluate and compare the proposed algorithm with the other outlier detection algorithms, we used the precision criterion obtained through equation (7).

$$Precision = \frac{TP}{TP+FP} \qquad (7)$$

Where TP (True Positive) is the number of outliers that are identified correctly and FP (False Positive) is the number of normal data that are falsely identified as outlier.

Tables (4) and (5) compare results of the proposed algorithm and the algorithm presented in [16] obtained for datasets KDD CUP and Shuttle.

Table 4: The results obtained for real dataset KDD CUP by the proposed algorithm and the one presented in [16]

Algorithm	***k***	**Precision**
***k*-means [16]**	5	0.564
***k*-means [16]**	13	0.593
K-NN	10	0.241
K-NN	50	0.301
Proposed Algorithm	-	0.609

Table 5: The results obtained for real dataset Shuttle by the proposed algorithm and the one presented in [16]

Algorithm	***k***	**Precision**
***k*-means[16]**	10	0.155
***k*-means[16]**	20	0.172
K-NN [16]	10	0.114
K-NN [16]	50	0.155
Proposed Algorithm	-	0.194

Table (4) shows the results of implementing the proposed algorithm and the algorithms *k*-means and K-NN on real dataset KDD-CUP 99. The precision criterion calculated for the algorithm indicated that the proposed algorithm is more accurate and precise. The algorithms showed a relatively acceptable accuracy and outliers were detected rather well. On the other hand the K-NN algorithm has a high dependency on the number of neighbors (k). Table (5) shows the results of implementation on real dataset Shuttle. These results show that the precision is not as high as dataset KDD CUP and this is because the four classes inserted as outliers were not easily detectable and separable from the normal data in the other three classes.

While the precision achieved in the second data is relatively low, overall tests results show that the proposed algorithm has a better performance than the *k*-means and K-NN algorithms and can very well identify the outliers. Another problem is the high dependency of these two algorithms on the value of *k*. In the *k*-means algorithm, the dependency on the number and determination of clusters can affect and alter the results. The K-NN algorithm also has the same dependency on *k*, or the number of nearest neighbors to the object. But the proposed algorithm has no such dependency and achieves better results by using a fuzzy threshold value which can be determined experimentally for each specific application.

V. Conclusion and Future Works

Outliers are extremely valuable data and the process of outlier detecting has been the subject of numerous studies. This study introduced and classified outlier detection methods. The purpose of this study was to provide a better understanding about the outlier detection methods for those who tend to conduct research on the related application and methods. Some of the discussed methods are global and can be used for different problems but others are limited to specific applications. All these methods reclassified in tables (1), (2) and (3). Having access to this outlier classification for different fields and applications allows researchers to choose the desired field and conduct specialized research accordingly.

As shown in different applications, our proposition is the use of density based and neighbor based methods for application that need local data detection. Because in application, there is no need to assess the entire dataset and one should only assess the data around the target data; this approach allows us to not only increase the accuracy, but also to perform outlier detection in much shorter time. Next, this study presented a clustering fuzzy method for detecting local outliers. In this method, we use fuzzy logic and data scoring to determine the membership value of each data in the cluster or its non-membership, which represent the outlier nature of data. The presented algorithm was evaluated by implementing it on two created datasets with normal and uniform distribution, and the implementation results showed that it has an acceptable accuracy. The proposed algorithm was also implemented on real dataset and the results were compared with the results of two well-known algorithms. This comparison showed that the proposed algorithm outperforms the compared algorithm in outlier detection.

In future research we aim to evaluate the proposed algorithm in other practical applications and to developing it for stream datasets.

References

[1] D. M. Hawkins, Identification of Outliers, London: Chapman and Hall, 1980.

[2] E. M. Knorr and R. T. Ng, "A Unified Approach for Mining Outliers," in *Centre for Advanced Studies on Collaborative research*, Toronto, 1997.

[3] M. M. Breunig, H.-P. Kriegel, R. T. Ng and J. Sander, "LOF: Identifying Density-Based Local Outliers," *ACM Sigmod Record,* vol. 29, no. 2, pp. 93-104, 2000.

[4] V. Kumar, "Parallel and Distributed Computing for Cybersecurity," *IEEE Distributed Systems Online,* vol. 6, no. 10, 2005.

[5] C. Spence, L. Parra and P. Sajda, "Detection, Synthesis and Compression in Mammographic Image Analysis with a Hierarchical Image Probability Model," *Mathematical Methods in Biomedical Image Analysis,* pp. 3-10, 2001.

[6] S. Panigrahi, A. Kundu, S. Sural and A. Majumdar, "Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning," *Information Fusion,* vol. 10, no. 4, pp. 354-363, 2009.

[7] X. Yu-Zhen, Z. Yu-Lin and Y. Jian-Ping, "Detection of Outliers from Spacecraft Tracking Data using GP-RBF Network," *Acta Simulata Systematica Sinica,* vol. 2, 2005.

[8] P. Gogoi, D. K. Bhattacharyya, B. Borah and J. K. Kalita, "A survey of outlier detection methods in network anomaly identification," *The Computer Journal,* vol. 54, no. 4, pp. 570-588, 2011.

[9] P. Clifton, V. Lee, K. Smith and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," in *arXiv preprint arXiv*, 2010.

[10] M. Gupta, J. Gao, C. C. Aggarwal and J. Han, "Outlier Detection for Temporal Data: A Survey," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,* vol. 25, no. 1, 2013.

[11] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review,* vol. 22, no. 2, pp. 85-126, 2004.

[12] M. Agyeman, K. Barker and R. Alhajj, "A comprehensive survey of numeric and symbolic outlier mining techniques," *Intelligent Data Analysis,* vol. 10, no. 6, pp. 521-538, 2006.

[13] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest

technological trends," *Computer Network,* vol. 51, no. 12, pp. 3448-3470, 2007.

[14] Z. A. Bakar, R. Mohemad, A. Ahmad and M. M. Deris, "A Comparative Study for Outlier Detection Techniques in Data Mining," in *Cybernetics and Intelligent Systems*, 2006.

[15] S. Shiblee and L. Gruenwald, "An Adaptive Outlier Detection Technique for Data Streams," *Scientific and Statistical Database Management,* pp. 596-597, 2011.

[16] Chawla, Sanjay, and Aristides Gionis. "k-means-: A Unified Approach to Clustering and Outlier Detection." In *SDM*, pp. 189-197. 2013.

[17] P. L. Brockett, X. Xia and R. A. Derrig, "Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud," *Journal of Risk and Insurance,* vol. 65, no. 2, pp. 245-274, 1998.

[18] P. D'haeseleer, S. Forrest and P. Helman, "An immunological approach to change detection: Algorithms, analysis and implications," *Security and Privacy,* pp. 110-119, 1996.

[19] E. Eskin, "Anomaly Detection over Noisy Data Using Learned Probability Distributions," 2000.

[20] A. K. Ghosh, J. Wanken and F. Charron, "Detecting Anomalous and Unknown Intrusions Against Programs," in *Computer Security Applications Conference*, 1998.

[21] W. Lee, S. J. Stolfo and P. K. Chan, "Learning Patterns from Unix Process Execution Traces for Intrusion Detection," *AAAI Workshop on AI Approaches to Fraud Detection and Risk Management,* pp. 50-56, 1997.

[22] W. Lee, S. J. Stolfo and K. W. Mok, "Adaptive Intrusion Detection: a Data Mining Approach," *Artificial Intelligence Review,* vol. 14, no. 6, pp. 533-567, 2000.

[23] D. Anderson, T. F. Lunt, H. Javitz, A. Tamaru and A. Valdes, "Detecting Unusual Program Behavior Using the Statistical Component of the Next-generation Intrusion Detection Expert System (NIDES)," SRI International, Computer Science Laboratory, California, 1995.

[24] K. Yamanishi and J.-I. Takeuchi, "Discovering outlier filtering rules from unlabeled data: combining a supervised learner with an unsupervised learner," *seventh ACM SIGKDD international conference on Knowledge discovery and data mining,* pp. 389-394, 2001.

[25] D.-Y. Yeung and C. Chow, "Parzen-window network intrusion detectors," *16th International Conference on Pattern Recognition,* vol. 4, pp. 385-388, 2002.

[26] A. Valdes and K. Skinner, "Adaptive, Model-based Monitoring for Cyber Attack Detection," Springer, Berlin Heidelberg, 2000.

[27] A. Bronstein, J. Das, M. Duro, R. Friedrich, G. Kleyner, M. Mueller, S. Singhal and I. Cohen, "Self-Aware Services: Using Bayesian Networks for Detecting Anomalies in Internet-based Services," *IEEE/IFIP International Symposium on Integrated Network Management,* pp. 623-638, 2001.

[28] D. Barbará, J. Couto, S. Jajodia and N. Wu, "ADAM: a testbed for exploring the use of data mining in intrusion detection," *ACM Sigmod Record,* vol. 30, no. 4, pp. 15-21, 2001.

[29] G. Williams, R. Baxter, H. He, S. Hawkins and L. Gu, "A Comparative Study of RNN for Outlier Detection in Data Mining," in *IEEE International Conference on Data Mining*, 2002.

[30] J. R. Dorronsoro, F. Ginel, C. Sgnchez and C. S. Cruz, "Neural fraud detection in credit card operations," *IEEE Transaction on Neural Networks,* vol. 8, no. 4, pp. 827-834, 1997.

[31] R. Brause, T. Langsdorf and M. Hepp, "Neural Data Mining for Credit Card Fraud Detection," *11th IEEE International Conference on Tools with Artificial Intelligence,* pp. 103-106, 1999.

[32] R. J. Bolton and D. J. Hand, "Unsupervised Profiling Methods for Fraud Detection," *Credit Scoring and Credit Control VII,* pp. 235-255, 2001.

[33] V. R. Ganji and S. N. P. Mannem, "Credit card fraud detection using anti-k nearest neighbor algorithm," *International Journal on Computer Science & Engineering,* vol. 4, no. 6, pp. 1035-1039, 2012.

[34] T. Fawcett and F. Provost, "Activity monitoring: noticing interesting changes in behavior," *ACM SIGKDD international conference on Knowledge discovery and data mining,* pp. 53-62, 1999.

[35] K. C. Cox, S. G. Eick, G. J. Wills and R. J. Brachman, "Visual Data Mining: Recognizing Telephone Calling Fraud," *Data Mining and Knowledge Discovery,* vol. 1, no. 2, pp. 225-231, 1997.

[36] S. L. Scott, "Detecting network intrusion using a Markov modulated non homogeneous Poisson process," *Submitted to the Journal of the American Statistical Association,* 2001.

[37] P. Barson, S. Field, N. Davey, G. McAskie and R. Frank, "The Detection of Fraud in Mobile Phone Networks," *Neural Network World,* vol. 6, no. 4, pp. 477-484, 1996.

[38] E. Suzuki, T. Watanabe, H. Yokoi and K. Takabayashi, "Detecting interesting exceptions from medical test data with visual summarization," *Third IEEE International Conference on Data Mining,* pp. 315-322, 2003.

[39] W.-K. Wong, A. Moore, G. Cooper and M. Wagner, "Bayesian network anomaly pattern detection for disease outbreaks," *ICML,* pp. 808-813, 2003.

[40] J. Lin, E. Keogh, A. Fu and H. V. Herle, "Approximations to Magic: Finding Unusual Medical Time Series," *18th IEEE Symposium on Computer-Based Medical Systems,* pp. 329-334, 2005.

[41] E. Keogh, S. Lonardi and . B.-c. Chiu, "Finding Surprising Patterns in a Time Series Database in Linear Time and Space," *The eighth ACM SIGKDD international*

conference on Knowledge discovery and data mining, pp. 550-556, 2002.

[42] E. Keogh, J. Lin, S.-H. Lee and H. V. Herle, "Finding the most unusual time series subsequence: algorithms and applications," *Knowledge and Information Systems,* vol. 11, no. 1, pp. 1-27, 2007.

[43] S. Jakubek and T. Strasser, "Fault-diagnosis using neural networks with ellipsoidal basis functions," *American Control Conference,* vol. 5, pp. 3846-3851, 2002.

[44] T. Yairi, Y. Kato and K. Hori, "Fault Detection by Mining Association Rules from House-keeping Data," *International Symposium on Artificial Intelligence, Robotics and Automation in Space,* vol. 3, no. 9, 2001.

[45] G. Manson, G. Pierce and K. Worden, "On the long-term stability of normal condition for damage detection in a composite panel," *Key Engineering Materials,* vol. 204, pp. 359-370, 2001.

[46] S. J. Hickinbotham and J. Austin, "Novelty detection in airframe strain data," *15th International Conference on Pattern Recognition,* vol. 2, pp. 536-539, 2000.

[47] H. Sohn, K. Worden and C. R. Farrar, "Novelty Detection under Changing Environmental Conditions," *SPIE's 8th Annual International Symposium on Smart Structures and Materials,* pp. 108-118, 2001.

[48] C. P. Diehl and J. B. Hampshire, "Real-time object classification and novelty detection for collaborative video surveillance," *Neural Networks,* vol. 3, pp. 2620-2625, 2002.

[49] M. Davy and S. Godsill, "Detection of abrupt spectral changes using support vector machines. an application to audio signal segmentation," *ICASSP,* vol. 2, pp. 1313-1316, 2002.

[50] S. Singh and M. Markou, "An Approach to Novelty Detection Applied to the Classification of Image Regions," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,* vol. 16, no. 4, pp. 396-408, 2004.

[51] D. Pokrajac, A. Lazarevic and L. J. Latecki, "Incremental Local Outlier Detection for Data Streams," *IEEE Symposium on Computational Intelligence and Data Mining (CIDM),* pp. 504-515, 2007.

[52] L. M. Manevitz and M. Yousef, "One-Class SVMs for Document Classification," *Journal of Machine Learning Research,* vol. 2, pp. 139-154, 2001.

[53] D. Guthrie, Unsupervised Detection of Anomalous Text, Sheffield: Doctoral dissertation, University of Sheffield, 2008.

[54] D. Janakiram, V. A. M. Reddy and A. P. Kumar, "Outlier detection in wireless sensor networks using Bayesian belief networks," *First International Conference on Communication System Software and Middleware,* pp. 1-6, 2006.

[55] J. W. Branch, C. Giannella, B. Szymanski, R. Wolff and H. Kargupta, "In-Network Outlier Detection in Wireless Sensor Networks," *Knowledge and information systems,* vol. 34, no. 1, pp. 23-54, 2013.

[56] H. Song, S. Zhu and G. Cao, "Attack-resilient time synchronization for wireless sensor networks," *Ad Hoc Networks,* vol. 5, no. 1, pp. 112-125, 2007.

[57] A. Boukerche, H. A. Oliveira and E. F. Nakamura, "Secure Localization Algorithms for Wireless Sensor Networks," *Communications Magazine,* vol. 46, no. 4, pp. 96-101, 2008.

[58] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki and D. Gunopulos, "Online outlier detection in sensor data using non-parametric models," *32nd international conference on Very large data bases,* pp. 187-198, 2006.

[59] P. Crook, S. Marsland, G. Hayes and U. Nehmzow, "A Tale of Two Filters - On line Novelty Detection," *ICRA'02. IEEE International Conference on Robotics and Automation,* vol. 4, pp. 3894-3899, 2002.

[60] J. A. Ting, A. D`Souza and S. Schaal, "Automatic Outlier Detection : A Bayesian approach," *IEEE International Conference on Robatics and Automation,* pp. 2489-2494, 2007.

[61] G. Munz, S. Li and G. Carle, "Traffic Anomaly Detection Using K-Means Clustering," *GI/ITG Workshop MMBnet,* 2007.

[62] T. Ide and H. Kashima, "Eigenspace-based anomaly detection in computer systems," *tenth ACM SIGKDD international conference on Knowledge discovery and data mining,* pp. 440-449, 2004.

[63] K. Kadota, D. Tominaga, Y. Akiyama and K. Takahashi, "Detecting outlying samples in microarray data: A critical assessment of the effect of outliers on sample classification," *Chem-Bio Informatics,* vol. 3, no. 1, pp. 30-45, 2003.

[64] G. Atluri, R. Gupta, G. Fang, G. Pandey, M. Steinbach and V. Kumar, "Association Analysis Techniques for Bioinformatics Problems," *Bioinformatics and Computational Biology,* pp. 1-13, 2009.

[65] C.-T. Lu, D. Chen and Y. Kou, "Algorithms for Spatial Outlier Detection," *Third IEEE International Conference on Data Mining,* pp. 597-600, 2003.

[66] S. Lin and D. E. Brown, "An Outlier-based Data Association Method For Linking Criminal Incidents," *Decision Support Systems,* vol. 41, no. 3, pp. 604-615, 2006.

[67] H. Dutta, C. Giannella, K. D. Borne and H. Kargupta, "Distributed top-k outlier detection in astronomy catalogs using the DEMAC system," in *SDM*, 2007.

[68] Y.-X. Zhang, A. L. Lou and Y.-H. Zhao, "Outlier detection in astronomical data," *Astronomical Telescopes and Instrumentation,* pp. 521-529, 2004.

.