COMPARISON OF FUZZY - NEURAL CLUSTERING BASED OUTLIER DETECTION TECHNIQUES

Nilam Upasani
Research Scholar, Computer Engineering, Indian School of Mines, Dhanbad, India

Hari Om
Computer Engineering, Indian School of Mines, Dhanbad, India

ABSTRACT

Fuzzy logic can be used to reason like humans and can deal with uncertainty other than randomness. Ability to learn, adapt, fault tolerance and reason with available knowledge, are the distinguished features of neural networks. Outlier detection is a difficult task to be performed, due to uncertainty involved in it. The outlier itself is a fuzzy concept and difficult to determine in a deterministic way. The hybridization of fuzzy and neural soft computing system is very promising, since they exactly tackle the situation associated with outliers. This paper provides a detailed comparison of fuzzy and/or neural approaches used for outlier detection by discussing their pros and cons. Thus this is a very helpful document for naive researchers in this field.

Keywords: outlier detection; fuzzy logic; neural network; data mining; clustering and comparison;

I. INTRODUCTION

Business Intelligence is a de-facto standard that uses Data Mining techniques to manage and plan businesses. Mined knowledge is in the form of rules, association among the patterns, current and future customer trends. But there is a serious issue concerning with data mining, called outliers. In literature different definitions of outlier exist: the most commonly referred are [1]:

- “An outlier is an observation that deviates so much from other observations as to arouse suspicions that is was generated by a different mechanism “ (Hawkins, 1980).
- “An outlier is an observation (or subset of observations) which appear to be inconsistent with the remainder of the dataset” (Barnet & Lewis, 1994).
- “An outlier is an observation that lies outside the overall pattern of a distribution” (Moore and McCabe, 1999).
- “Outliers are those data records that do not follow any pattern in an application” (Chen and al., 2002).
An outlier in a set of data is an observation or a point that is considerably dissimilar or inconsistent with the remainder of the data” (Ramasmawy et al., 2000).

Outlier may occur due to data error, noise, anomaly, improper sensors or typo error. Outliers may cause incorrect knowledge discovery and obstruct the process of data mining and thus may badly affect the business profit. Recently, many application domains have recognized the direct mapping between the outliers in data and real world anomalies that are of great interest to data analyst. The outlier detection is searching for objects in the database that do not obey laws valid for the major part of the data. Outlier detection is an important pre-processing task [2]. It has several realistic applications in areas such as, image processing [3], remote sensing [4], fraud detection [5], identifying computer network intrusions and bottlenecks [6], criminal activities in e-commerce and detecting suspicious activities [7], Satellite image analysis, etc.

For any outlier detection technique, there are two ways to report the outliers which are scores and labels [8]. In scoring techniques, outlier score is assigned to each instance in the test data depending on the degree to which that instance is considered an outlier. Thus the output of such techniques is a ranked list of outliers. An analyst may choose to either analyze top few outliers or use a cut-off threshold to select the outliers. In label techniques, a label (normal or anomalous) is assigned to each test instance.

Outlier detection and clustering analysis are two greatly interconnected tasks, although they serve different purposes. Clustering locates the majority patterns in a data set and categorizes data accordingly, whereas outlier detection tries to capture those exceptional cases that deviate substantially from the majority patterns. Without intrinsic clustering in the data, no natural outliers will exist. Clustering is an important tool for outlier analysis. A number of clustering-based outlier detection techniques have been developed, most of which rely on the key assumption that normal objects belong to large and dense clusters, while outliers form very small clusters [9, 10].

Outlier detection methods are derived from three fields of computing which are statistics (proximity-based, parametric, non-parametric and semi-parametric), neural networks (supervised and unsupervised) and machine learning. The main advantage of statistical models is they are mathematically justified. They are commonly appropriate to quantitative real-valued data sets. In case of least quantitative ordinal data distributions, the ordinal data should be transformed to suitable numerical values for statistical (numerical) processing. This restricts their applicability and in case of complex data transformations it also increases the processing time. Much outlier detection has only focused on continuous real-valued data attributes whereas Machine Learning has mainly centered on categorical data. Machine learning techniques are robust, do not suffer the Curse of Dimensionality, work well on noisy data and have simple class boundaries compared with the complex class boundaries. Neural Network approaches are normally non-parametric and model based. They generalize well to unseen patterns, capable of learning complex class boundaries and have better reasoning ability along with fault tolerance. Much of the pattern recognition task is now implemented using fuzzy logic as it can be used to reason in imprecise situations and can handle uncertainty. If we hybridize fuzzy logic with the neural networks, it combines advantages of both. Such a hybrid system can also be used for Outlier Detection. As outlier detection and clustering analysis are two greatly interconnected tasks, this survey aims at providing an overview of the outlier detection techniques involving fuzzy and/or neural network approaches based on clustering, by focusing on their strengths and weaknesses.

This paper is organized as follows. Section II and III gives a comprehensive overview of the fuzzy clustering based outlier detection and neural network based outlier detection techniques. Advantages and disadvantages of these two approaches are discussed in Section IV. Section V points towards some hybridized fuzzy neural network approaches effectively used for pattern recognition, which can also be used for the task of outlier detection. Conclusions are given in section VI and references are cited at the end.
II. FUZZY CLUSTERING BASED OUTLIER DETECTION

Crisp set can be used to handle structured data which is generally free from outlier points. But it is difficult to handle unstructured natural data which often contain outlier data points. Fuzzy set (introduced by Zadeh [11]) is different from its crisp counterpart, as it allows the elements to have a degree of membership. Thus, to handle unstructured natural data which is qualitative and imprecise in nature, many of the data mining techniques are integrated with fuzzy logic [12]. In traditional clustering approaches each pattern belongs to one and only one cluster [13]. In case of fuzzy clustering, the membership function associates each pattern with all the clusters. Thus fuzzy clusters grow in their natural shapes. In this section some fuzzy clustering outlier detection techniques are analyzed.

A. Fuzzy C-Means algorithm (FCM) [14]:

FCM is one of the well known fuzzy clustering algorithms, and used in a wide variety of applications, such as medical imaging, remote sensing, data mining and pattern recognition [15,16,17,18,19]. Chiu, Yager & Filev have proposed Fuzzy c-means data clustering algorithm in which each data point belongs to a clustering to a degree specified by a membership grade [20]. Fuzzy c-means clustering has two processes: the calculation of cluster centers and the assignment of points to these centers using a form of Euclidian distance. This process is repeated until the cluster centers are stabilized. FCM partitions a collection of data points into fuzzy groups. Unlike k-means where data point must exclusively belong to one cluster center, FCM finds a cluster center in each group in order to minimize the objective function of the dissimilarity measure. FCM employs fuzzy partitioning so that a given data point can belong to several groups with the degree of belongingness specified by membership grades between 0 and 1. The membership of each data point corresponding to each cluster center will be calculated on the basis of distance between the cluster center and the data point, using exp. [1],

\[
\mu_j(x_i) = \left( \frac{1}{d_{ji}} \right)^{\frac{1}{m-1}}
\]

where

\(\mu_j(x_i)\) : is the membership of \(x_i\) in the \(j^{th}\) cluster
\(d_{ji}\) : is the distance of \(x_i\) in cluster \(c_j\)
\(m\) : is the fuzzification parameter
\(p\) : is the number of specified clusters
\(d_{ik}\) : is the distance of \(x_i\) in cluster \(C_k\)

The new cluster centers are calculated with these membership values using the exp. [2].

\[
c_j = \frac{\sum_{i} \mu_j(x_i)^{m} x_i}{\sum_{i} \mu_j(x_i)^{m}}
\]

where

\(C_j\) : is the center of the \(j^{th}\) cluster
\(x_i\) : is the \(i^{th}\) data point
\(\mu_j\) : the function which returns the membership
\(m\) : is the fuzzification parameter
More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one. After each iteration, the membership and cluster centers are updated.

Good characteristics of this algorithm are, it is simple to implement, robust in behavior, can handle multidimensional data, gives best result for overlapped data set and can model uncertainty within the data. The main drawback of this algorithm is that it is unable to handle outlier points effectively because of the restriction that the sum of membership values of a data point in all the clusters must be one. This tends to give high membership values for the outlier points. It is also incapable to calculate the membership value if the distance of a data point is zero. We get better result but at the cost of more number of iteration

B. Fuzzy clustering based algorithm proposed by Binu Thomas and Raju Ghave [21]:

They have proposed a novel fuzzy clustering method in which they have modified the membership function proposed by Fuzzy C-Means algorithm. In this method a very low membership value is assigned to the outlier points to handle unstructured natural data.

In FCM, many limitations arise because the membership of a point in a cluster is calculated based on its membership in other clusters. In this method, the membership of a point in a cluster depends only on its distance in that cluster calculated by a simple expression exp. [3].

\[
\mu_j(x_i) = \frac{\text{Max}(d_j) - d_{ji}}{\text{Max}(d_j)} \quad \ldots \ldots [3]
\]

Where
\[
\mu_j(x_i) : \text{is the membership of } x_i \text{ in the } j^{th} \text{ cluster}
\]
\[
d_{ji} : \text{is the distance of } x_i \text{ in cluster } c_j
\]
\[
\text{Max}(d_j) : \text{is the maximum distance in the cluster } c_j
\]

With this it is possible to calculate the membership value if the distance of a data point is zero which is the limitation of FCM. It is more proficient in handling the natural data with outlier points by allocating very low membership values to the outlier points. It is far better in the calculation of new cluster centers. This method has limitations in exploring highly structured crisp data which is free from outlier points.

C. Fuzzy clustering based algorithm proposed by Moh’d Belal et.al. [22]:

Moh’d Belal Al-Zoubi, Ali Al-Dahoud, Abdelfatah A. Yahya proposed an efficient method for outlier detection based on fuzzy clustering techniques. This technique is based on the key assumption that normal objects belong to large and dense clusters, whereas outliers form very small clusters. Here a small cluster is a cluster with fewer points than half the average number of points in the other clusters.

In this approach, FCM algorithm is performed first to produce a set of k clusters as well as the objective function (OF) and Compute the threshold (T) to determine outliers. Then small clusters are determined and the points that belong to these clusters are considered as outliers. After this if any outliers are left then those would be detected in the remaining clusters by temporary removing a point from the data set and re-calculating the objective function. If a noticeable change is observed in the Objective Function (OF), the point is considered an outlier. This approach is efficient to locate the outliers when applied to different data sets. However, the proposed method is very time consuming because the FCM algorithm has to run multiple times.
III. NEURAL NETWORK BASED OUTLIER DETECTION

Neural networks serve to learn, process, and predict the information using layers of interconnected units resembling the neurons that comprise the nervous system of the human brain. Neural networks have the advantage of powerful modeling ability. Neural network does not involve the complex process and has the advantage of requiring poor or no a priori assumption on the considered data. There are many outlier detection techniques based on neural networks.

A. Outlier Detection Using Replicator Neural Networks[23]

Hawkins S., H. X. Williams, G.J. and Baxter R.A. have proposed replicator neural network (RNN) based technique to find outliers in large multivariate databases. Earlier Replicator Neural Network was used for its data compression capabilities in several applications of image and speech processing [24, 25]. RNNs have a flexible, non-parametric representation of clusters and hence they have used it as a powerful tool for outlier detection [26]. In this approach, a feed-forward multi-layer neural network is used with three hidden layers sandwiched between the input and output layers which have n units each, corresponding to the n features of the training data. In order to minimize the average reconstruction error, the number of units in the three hidden layers is chosen experimentally across all training patterns. To develop a compressed and implicit model of the data during training, the input variables are also the output variables. A measure of outlyingness (called the Outlier Factor) of individuals is then developed as the reconstruction error of individual data points. The Outlier Factor of the $i^{th}$ data record $OF_i$ as the measure of outlyingness is defined by the average reconstruction error over all features (variables) as given in exp. [4].

\[
OF_i = \frac{1}{n} \sum_{j=1}^{n} (x_{ij} - o_{ij})^2 \quad \text{[4]}
\]

where, $n$ is the number of features, $x_{ij}$ is the input value and $o_{ij}$ is the input value. The OF is evaluated for all data records using the trained RNN.

Instead of the usual sigmoid activation function for the middle hidden layer a staircase-like function with parameters $N$ (number of activation levels) and $a3$ (transition rate from one level to the next) are employed. This activation function provides the data compression as it continuously divides distributed data points into a number of discrete valued vectors. The mapping to discrete categories naturally places the data points into a number of clusters which is essential for outlier detection. For scalability the RNN is trained with a smaller training set and then applied to all the data to calculate their outlyingness. Neural Network methods often have difficulty with such smaller datasets, though RNN performs satisfactory for both small and large datasets. But its performance degrades with datasets containing radial outliers and so it is not suggested for this type of dataset.

B. Multiple Outlier Detection in Multivariate Data Using Self-Organizing Maps Title [27]

Nag A.K., Mitra A. and Mitra S. proposed an artificial intelligence technique of self-organizing map (SOM) based non-parametric method for outlier detection. It can be used to detect outliers from large multidimensional datasets and also provides information about the entire outlier neighborhood.

SOM is a feed forward neural network that uses an unsupervised training algorithm, and through a process called self-organization, configures the output units into a topological representation of the original data (Kohonen 1997). SOM produces a topology-preserving mapping of the multidimensional data cloud onto lower dimensional visualizable plane and hence provides an easy way of detection of multidimensional outliers in the data, at respective levels of influence.
The outcome of the SOM algorithm is the set of weight vectors which are used to connect the multidimensional input vector space and the output grid. A trained weight vector associated with a grid on the output layer gives a unique identification of that unit. First the training process of winner selection and weight updation is done. Then the appropriate visualization of the topology of map is done and the outliers are detected by interpreting the results. Advantage includes the easy implementation as it does not require any priori assumption on the variable.

C. Subtractive Clustering Based RBF Neural Network Model for Outlier Detection [28]

Peng Yang, Qingsheng Zhu and Xun Zhong proposed a RBF neural network model using subtractive clustering algorithm for selecting the hidden node centers, which can achieve faster training speed. This method has adopted the Subtractive Clustering (SC) to determine the hidden node centers in RBF network since it allows a scatter input-output space partitioning, which can lead to a proper network structure and achieve computation efficiency. Accurate prediction is achieved by minimizing the variances of the nodes in the hidden layer. An RBF network is a three-layer feed forward neural network which consists of an input layer, a hidden layer and an output layer. The network is trained to implement a desired input-output mapping by producing incremental changes of the weights of the network.

Finally, according to the degree of outlier for each input instance candidate outliers can be obtained. Advantage of this method is that it can perform more accurate prediction with lower false positive rate and higher detection rate and it can be an effective solution for detecting outliers.

IV. ADVANTAGES AND DISADVANTAGES OF FUZZY - NEURAL CLUSTERING BASED TECHNIQUES

A. Advantages and disadvantages of fuzzy based techniques

• The use of a fuzzy set approach to pattern classification inherently provides degree of membership information that is extremely useful in higher level decision making.
• Fuzzy logic is capable of supporting, to a reasonable extent, human type reasoning in natural form by allowing partial membership for data items in fuzzy subsets [29].
• Fuzzy logic can handle uncertainty at various stages.
• With this approach knowledge can be expressed in terms of If–THEN rules.
• Disadvantage of fuzzy clustering is that with a growing number of objects the amount of output of the results becomes huge, so that the information received often cannot be worked up.

B. Advantages and disadvantages of neural networks based techniques

• Neural networks have an adaptive capability of learning new patterns and refine existing ones.
• These techniques are capable of learning complex class boundaries and have ability to learn complex nonlinear input-output relationships
• As neural networks are massively parallel in nature, Complex computational task can be performed in great speed.
• They are nonparametric that is it can work without any prior information available.
• Neural network have disadvantages such as lower detection precision in case of small learning sample size, weaker detection stability.
V. HYBRID APPROACHES INVOLVING BOTH FUZZY LOGIC AND NEURAL NETWORKS FOR PATTERN RECOGNITION

We have surveyed that the variations of fuzzy logic and neural network are best suitable for classification and clustering. This hybrid approach is used for speaker identification, real time heart diseases detection, face recognition, etc [30, 31, 32, 33]. However, not much work is done for using this hybrid approach for the outlier detection.

Wang G., Jinxing Hao, Jian Ma, Lihua Huang have proposed a novel approach for Intrusion Detection System to solve network security problems based on the hybrid approach [34]. The technique is called Fuzzy Clustering Artificial Neural Network (FC-ANN). The basic objective was to enhance the detection precision for low-frequent attacks and detection stability.

The general procedure of FC-ANN approach has three major stages. In the first stage, they have used one of the popular soft clustering techniques, fuzzy c-means clustering to divide the training data into several subsets. In the second stage, based on different training sets, different neural networks are trained. They have used classic feed-forward neural networks trained with the back-propagation algorithm for training. In the third stage, to eliminate the errors of different ANNs, a meta-learner, fuzzy aggregation module is used to learn again and combine the results of different ANN’s.

This method has employed divide and conquer approach. Fuzzy clustering is used to divide the whole training set into subsets which have less number and lower complexity. Thus the ANN can learn each subset more quickly, robustly and precisely so that low-frequent attacks can also be detected. The proposed technique is found better in performance compared to BPNN and other well-known methods such as decision tree, the naïve Bayes in terms of detection precision and detection stability.

VI. CONCLUSION AND FUTURE DIRECTIONS

In this survey we have discussed different ways in which the problem of cluster based outlier detection has been formulated in literature, and we attempted to provide an overview of the huge literature on different techniques. Different outlier detection algorithms are available which are based on fuzzy logic and neural network, which have their own advantages and disadvantages.

Some researchers have used hybrid approach, based on fuzzy logic and neural network to combine the advantages and to overcome the disadvantages of both the techniques for various domains which are mentioned in section V. But very few have used this hybrid approach for outlier detection. Thus, there is still a wide scope to apply modern fuzzy-neural techniques in this area to improve the quality and required time of outlier detection process.

REFERENCES


[29] Varun chandola, arindam banerjee and vipin kumar, outlier detection : a survey


