SURVEY PAPER ON OUT LIER DETECTION USING FUZZY LOGIC BASED METHOD

Deepa Verma, Rakesh Kumar and Akhilesh Kumar

Department of Information Technology,
Rajkiya Engineering College, Ambedkar Nagar (U.P) – 224 122, India

ABSTRACT

Fuzzy logic can be used to reason like humans and can deal with uncertainty other than randomness. 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. fuzzy logic system is very promising, since they exactly tackle the situation associated with outliers. Fuzzy logic that addresses the seemingly conflicting goals (i) removing noise, (ii) smoothing out outliers and certain other salient feature. This paper provides a detailed fuzzy logic used for outlier detection by discussing their pros and cons. Thus this is a very helpful document for naive researchers in this field.

KEYWORDS

Data mining, fuzzy logic, Outlier Detection. Artificial Intelligent Information Systems

1. INTRODUCTION

In a sequence of measurements or generic investigational data the outlier is defined as an observation that deviates too much from other observations and it is implicit to be generated by a mechanism different from other observations. The inlier, on the other hand, is defined as an observation that is explained by underlying probability density function. The exact definition of an outlier depends on the context [1]. Several definitions have been given and can be roughly grouped into six categories [2]: I) distribution based, II) depth-based, III) distance-based, IV) clustering based, V) density-based VI) categorical-based. As a low local density of other data on the observation is an indication of a possible outlier, all the above listed definitions substantially come from different strategies for density estimation. The presence of the outliers is not always a drawback. For instance, in clustering applications, outliers are considered as noise observations that should be removed in order to perform a more reliable clustering, while in data mining applications, the detection of anomalous patterns in data is more interesting than the detection of inliers. In some applications in fault diagnosis, the outliers can be the indicators of a faulty in the monitored system, thus their presence must be carefully pointed out. When the experimental data are used to validate a model and/or optimize some of its internal parameters, the presence of outliers should be avoided and, for this reason, usually a preprocessing of the data is mandatory in order to filter out such anomalous observation. In the steelmaking field a large amount of data are collected concerning the chemical composition of a steel cast and its derived products at the different stages of the manufacture as well as process variables, measurements and analyses concerning the properties of a each particular manufactured product. While process variables are normally extracted in an automatic way from the plant information system, some of the results of the chemical analyses, mechanical tests and inspections are still manually input by the technical personnel. Errors are thus possible on the values that are recorded by hand, some of which are quite easily recognized (for instance typographic errors) but most of them are not so easy to point out. Moreover, anomalous events or faults in the measuring devices as well as in the machinery

DOI: 10.5121/ijci.2017.6204

itself can results in measurements that are indeed anomalous but do not differ in the order of magnitude with respect to standard data. Thus their detection cannot be performed via trivial tests and/or visual inspection, but rather through the implementation of suitable detection strategies, capable of pointing out anomalous data in an efficient and rapid way. Efficiency and speed are often conflicting requirements especially when treating big amount of data, as actually in the case of steelmaking industry us. **Fuzzy logic** is an approach to computing based on "degrees of truth" rather than the usual "true or false" (1 or 0) Boolean **logic** on which the modern computer is based. Fuzzy logic is capable of supporting human type of reasoning in natural form to a considerable extent. It does so by allowing partial membership for data items in fuzzy subsets. Integration of fuzzy logic with data mining techniques has become one of the key aspects of soft computing in handling the challenges posed by the massive collection of natural data. The fuzzy set is different from a crisp set in that it allows the elements to have a degree of membership. The core of a fuzzy set is its membership function which defines the relationship between a value in the set's domain and its degree of membership in the fuzzy set. The relationship is functional because it returns a single degree of membership for any value in the domain.

2. LITERATURE SURVEY

Crisp set can be used to handle structured data which is generally free from outlier points. But it is difficult to handle formless 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 level 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.

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(FCM) 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.

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.

3. Different Aspects Of An Outlier Detection Problem

In this section discusses the different aspects of outlier detection. As mentioned earlier, a specific formulation of the problem is determined by several different factors such as the input data, unsupervised data, the availability (or unavailability) of other resources as well as the constraints and requirements induced by the application domain. This section brings forth the richness in the problem domain and motivates the need for so many diverse techniques.

(i) input data: a key component of any outlier detection technique is the input data in which it has to detect the outliers. Input is generally treated as a collection of data objects or data instances (also referred to as record, point, vector, pattern, event, case, sample, observation, or entity) [tan et al. 2005a]. Each data instance can be described using a set of attributes (also referred to as variable, characteristic, feature, field, or dimension). The data instances can be of different types such as binary, categorical or continuous. Each data instance might consist of only one attribute (univariate) or multiple attributes (multivariate). In the case of multivariate data instances, all attributes might be of same type or might be a mixture of different data types.

(ii)Type of Supervision: Besides the input data (or observations), an outlier detection algorithm might also have some additional information at its disposal. A labeled training data set is one such information which has been used extensively (primarily by outlier detection techniques based on concepts from machine learning [Mitchell 1997] and statistical learning theory [Vapnik 1995]). A training data set is required by techniques which involve building an explicit predictive model. The labels associated with a data instance denote if that instance is normal or outlier1. Based on the extent to which these labels are utilized, outlier detection techniques can be divided into three categories

(ii)a. Supervised outlier detection techniques. First category techniques assume the availability of a training data set which has labeled instances for normal as well as outlier class. Typical approach in such case is to build predictive models for both normal and outlier classes. Any unseen data instance is compared against the two models to determine which class it belongs to. Supervised outlier detection techniques have an explicit notion of the normal and outlier behavior and hence accurate models can be built. One drawback here is that accurately labeled training data might be prohibitively expensive to obtain. Labeling is often done manually by a human expert and hence requires a lot of effort to obtain the labeled training data set. Certain techniques inject artificial outliers in a normal data set to obtain a fully labeled training data set and then apply supervised outlier detection techniques to detect outliers in test data [Abe et al. 2006].

(ii)b. Semi-Supervised outlier detection techniques. Second category of techniques assume the availability of labeled instances for only one class. it is often difficult to collect labels for other class. For example, in space craft fault detection, an outlier scenario would signify an accident, which is not easy to model. The typical approach of such techniques is to model only the available class and decare any test instance which does not fit this model to belong to the other class.

(ii)c. Unsupervised outlier detection techniques. The third category of techniques do not make any assumption about the availability of labeled training data. Thus these techniques are most widely applicable. The techniques in this category make other assumptions about the data. For example, parametric statistical techniques, assume a parametric distribution of one or both classes of instances. Similarly, several techniques make the basic assumption that normal instances are far more frequent than outliers. Thus a frequently occurring pattern is typically considered normal while a rare occurrence is an outlier. The unsupervised techniques typically suffer from higher false alarm rate, because often times the underlying assumptions do not hold true.

4. OUTLIER DETECTION METHODS IN INDUSTRIAL SENSOR DATA

The occurrence of outliers in industrial data is handle rank deficiency. Comparative studies of these methods is often the rule rather than the exception. Many standard outlier detection methods fail to detect outliers in industrial data because of the high data dimensionality. These problems can he solved by using robust model-based methods that do not require the data to be of full rank. Jordaan and Smits [5] propose a robust model-based outlier detection approach that exploits the characteristics of the support vectors extracted by the Support Vector Machine method [6]. The method makes use of several models of varying complexity to detect outliers based on the characteristics of the support vectors obtained from SVM-models. This has the advantage that the decision does not depend on the quality of a single model, which adds to the robustness of the approach. Furthermore, since it is an iterative approach, the most severe outliers are removed first. This allows the models in the next iteration to learn from cleaner" data and thus reveal outliers that were masked in the initial model. The need for several iterations as well as the use of models, however, makes the on-line application of this method difficult for the not negligible computational burden. Moreover, if the data to be on-line processes come from dynamic systems, that tend to change their conditions through time, although not rapidly, the models update is required and this can furtherly slow the outlier detection. On the other hand, when low computational burden is required due to strict time constraints, there is a quite wide variety of fast methods. One of the most commonly used is the Grubbs test [7], a standard and widely applied statistical procedure, which however, to be appropriately applied, requires that data distribution can be reasonably approximated by a normal distribution, as it takes into account only the mean and standard deviation of the data distribution. Also in [8] a fast outlier detection method is proposed which uses the local correlation integral (LOCI). This method is highly effective for detecting outliers and groups of outliers, although it takes into consideration only the number of neighboring points. Neural Networks have been successfully applied for outliers detection, often coupled with techniques such as Principal Component Analysis (PCA) [9] and Partial Least Squares (PLS) [10]: for instance in [11] outliers are found by investigating points at the edges of previously constructed clusters. The neural networks used here have class coded output nodes. Outliers should not activate any output node. If the maximal output value of the neural network is less than a rejection threshold, then the input pattern is rejected.

5. APPLICATIONS OF OUTLIER DETECTION

The ability to detect outliers is a highly desirable feature in application domains for a variety of reasons. In this section we discuss some of the popular applications of outlier detection. For each application domain we discuss the significance of outliers, the challenges unique to that domain and the popular approaches for outlier detection that have been adopted in that domain.

(i)Intrusion Detection: Intrusion detection refers to detection of malicious activity (break-ins, penetrations, and other forms of computer abuse) in a computer related system [Phoha 2002]. These malicious activities or intrusions are very interesting from a computer security perspective. An intrusion is different from the normal behavior of the system. This property allows a direct formulation of the problem as an outlier detection problem. Outlier detection techniques have been widely applied for intrusion detections.

(ii) Fraud Detection: Fraud detection refers to detection of criminal activities occurring in commercial organizations such as banks, credit card companies, insurance agencies, cell phone companies, stock market etc. The malicious users might be the actual customers. of the organization or might be posing as a customer (also known as identity theft). The fraud occurs when these users consume the resources provided by the organization in an unauthorized way. The organizations are interested in immediate detection of such frauds to prevent economic losses.

6. ADVANTAGES AND DISADVANTAGES OF FUZZY LOGIC BASED METHOD

Advantages:

1. 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.
2. Fuzzy Logic allows you to model in a more intuitive way complex dynamic systems.
3. Fuzzy logic is capable of at the bottom of, human type reasoning in natural form by allowing partial membership for data items in fuzzy subsets [29].
4. Fuzzy logic can handle uncertainty at various stages.
5. With this approach knowledge can be expressed in terms of If–THEN rules.

Disadvantage:

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.

7. CONCLUSION AND FUTURE SCOPE

In this survey we have discussed different ways in which the problem of outlier detection has been formulated in literature. outlier detection algorithms are available which are based on fuzzy logic and 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 domain. But very few have used this hybrid approach for outlier detection. Thus, there is still a wide scope to apply -neural techniques in this area to improve the quality and required time of outlier detection process. In this paper we discussed the techniques and methods used by different researchers and basic introduction and algorithms are discussed in brief. At the end, we conclude that the outlier detection is must to increase the speed of processing of data and also helpful in reduction of processing cost of data. We have to compare all the techniques which are effectively used till the time and get the best one as the result of this comparison and reliable to implement. Try to achieve the objectives which are defined previously.

REFERENCES

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

[2] Chiu, S. L. (1994). Fuzzy model identification based on cluster estimation. Journal of Intelligent and Fuzzy Systems, 2, 267–278

[3] Hodge, V. and J. Austin, A Survey of Outlier Detection Methodologies, Artificial Intelligence Review, Vol. 22, 2004, pp. 85–126.

[4] Victoria J. Hodge and Jim Austin, A Survey of Outlier Detection Methodologies, Kluwer Academic Publishers, 2004.

[5] L.Zadeh, Fuzzy sets, Inform. And control, vol.8, pp. 338-353, ,1965.

[6] Sankar K. Pal, P. Mitra, Data Mining in Soft Computing Framework: A Survey, IEEE transactions on neural networks, vol. 13, no. 1, January 2002.

[7] E. Cox, Fuzzy Modeling And Genetic Algorithms For Data Mining And Exploration, Elsevier, 2005.

[8] Bezdek, J, L. Hall, and L. Clarke, Review of MR Image Segmentation Techniques Using Pattern Recognition, Medical Physics, Vol. 20, No. 4, 1993, pp. 1033–1048.

[9] Pham, D, Spatial Models for Fuzzy Clustering, Computer Vision and Image Understanding, Vol. 84, No. 2, 2001, pp. 285–297.

[10] Rignot, E, R. Chellappa, and P. Dubois, Unsupervised Segmentation of Polarimetric SAR Data Using the Covariance Matrix, IEEE Trans. Geosci. Remote Sensing, Vol. 30, No. 4, 1992, pp. 697–705.

[11] Al- Zoubi, M. B., A. Hudaib and B Al- Shboul A Proposed Fast Fuzzy C-Means Algorithm, WSEAS Transactions on Systems, Vol. 6, No. 6, 2007, pp. 1191-1195.

[12] Maragos E. K and D. K. Despotis, The Evaluation of the Efficiency with Data Envelopment Analysis in Case of Missing Values: A fuzzy approach, WSEAS Transactions on Mathematics, Vol. 3, No. 3, 2004, pp. 656-663.

[13] Binu Thomas and Raju G, A Novel Fuzzy Clustering Method for Outlier Detection in Data Mining, International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009.

[14] Moh'd Belal Al-Zoubi, Ali Al-Dahoud, Abdelfatah A. Yahya, New Outlier Detection Method Based on Fuzzy Clustering, WSEAS transactions on information science and applications, Issue 5, Volume 7, May 2010, pp. 681 – 690.

[15] Hawkins, S.; He, X.; Williams, G.J. & Baxter, R.A., Outlier detection using replicator neural networks. Proceedings of the 5th international conference on Knowledge Discovery and Data Warehousing, 2002.

[16] R. Hecht-Nielsen. Replicator neural networks for universal optimal source coding. Science, 269 (1860-1863), 1995.

[17] Williams, G.; Baxter, R.; He, H. & Hawkison,S., A comparative study of RNN for outlier detection in data mining, Proceedings of the IEEE International Conference on Data Mining, pp. 709–712, 9-12 December 2002, Australia.

[18] Nag, A.K.; Mitra, A. & Mitra, S., Multiple outlier Detection in Multivariate Data Using Self-Organizing Maps Title, Computational Statistical, N.20, 2005, pp.245-264.

[19] Peng Yang, Qingsheng Zhu and Xun Zhong, Subtractive Clustering Based RBF Neural Network Model for Outlier Detection, Journal Of Computers, Vol. 4, No. 8, August 2009, pp. 755-762.

[20] N. P. Jawarkar, R. S. Holambe and T. K. Basu, Use of fuzzy min-max neural network for speaker identification, Proc. IEEE Int. Conference on Recent Trends in Information Technology (ICRTIT 2011), MIT, Anna university, Chennai, Jun.-3-5, 2011.

[21] S. S. Panicker, P. S. Dhabe, M. L. Dhore, Fault Diagnosis Using Fuzzy Min-Max Neural Network Classifier, CiiT International Journal of Artificial Intelligent Systems and Machine Learning, Issue Jul. 2010. http://www.ciitresearch.org/aimljuly2010.html.

[22] M. Mohammadi, R. V. Pawar, P. S. Dhabe, Heart Diseases Detection Using Fuzzy Hyper Sphere Neural Network Classifier, CiiT International Journal of Artificial Intelligent Systems and Machine Learning, Issue July 2010. http://www.ciitresearch.org/aimljuly2010.html.\

[23] Wang G., Jinxing Hao, Jian Ma, Lihua Huang, A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering. Expert Systems with Applications (2010), doi:10.1016/j.eswa.2010.02.102

[24] Gath, I and A. Geva, Fuzzy Clustering for the Estimation of the Parameters of the Components of Mixtures of Normal Distribution, Pattern Recognition Letters, Vol. 9, 1989, pp. 77-86.

[25] Cutsem, B and I. Gath, Detection of Outliers and Robust Estimation using Fuzzy Clustering, Computational Statistics & Data Analyses, Vol. 15, 1993, pp. 47-61.

AUTHORS

Akhilesh Kumar graduated from Mahatma Ghandhi Mission’s college of Engg. and technology, Noida, Uttar Pradesh in Computer Science & Engineering in 2010. He has been M.Tech in the department of Computer Science & Engineering, Kamla Nehru Institute of Technology, Sultanpur (Uttar Pradesh). SinceAugust2012, he has been with the Department of Department of Information Technology, Rajkiya EngineeringCollege, Ambedkar Nagar, as an Assistant Professor.His area of interests includes Computer Networks and Mobile ad-hoc Nerwork.

Rakesh Kumar was born in Bulandshahr (U.P.), India, in 1984. He received the B.Tech. degree in Information Technology from Kamla Nehru Institute of Technology, Sultanpur (U.P.), India, in 2007, and the M.Tech. degrees in ICT Specialization with Software Engineering from the Gautam Buddha University, Greater Noida, Gautam Budh Nagar, Uttar Pradesh, India, in 2012.In 2007, he joined the Quantum Technology, New Delhi as a Software Engineer and Since August 2012, he has been with the Department of Department of Information Technology, Rajkiya Engineering College, Ambedkar Nagar, as an Assistant Professor. His current research interests include Computer Network, Multicast Security, Sensor Network and data mining. He is a Life Member of the Indian Society for Technical Education (ISTE), and he is a Nominee Member of Computer society of India.