Finding Anomaly with Fuzzy Rough C-Means Using Semi-Supervised Approach

Gadekar S. S.[1], Prof. Shinde S. M.[2]

[1]Savitribai Phule University of Pune, Maharashtra, India

[2]H.O.D of Computer Science and Engineering, JS College of Engineering, Pune, Maharashtra, India

Abstract: ***Outlier detection is initial step in various data-mining applications. This methods have been suggested for number of applications, such as credit card fraud detection, clinical trials, voting irregularity analysis, data cleansing, network intrusion, severe weather prediction, geographic information systems, athlete performance analysis, and other data-mining tasks proposed algorithm. In this proposed system combines the fuzzy set theory, rough set theory and semi-supervised learning to detect outliers and is a new try in area of outlier detection for semi-supervised learning. Without considering those points located in lower approximation of a cluster, proposed algorithm only need discuss the possibility of the points in boundary to be assigned as outliers and has many advantages over SSOD. Proposed algorithm uses labelled normal and outliers and as well as samples without labels and can improve outliers detection accuracy and reduce false alarm rate under the guidance of labeled samples. Proposed algorithm will be applied to many outlier detection fields which has only partially labeled samples, especially that does not make a certain judgment in uncertain conditions. But, the results depend on selection of number of cluster c, initial canter of cluster, parameters, proposed algorithm usually also stops on a local minimum. So, during the process, it must carefully select initial canters and other parameters. The proposed system proposes the technique that may add parameters to speed up the technique.***

Keywords: Pattern recognition; Outlier detection; Semi-supervised learning; Rough sets; Fuzzy sets; C-means clustering

1. Introduction

“An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”. Outliers are also referred to as abnormalities, discordant, deviants, or anomalies in the data mining and statistics literature. In most applications, the data is created by one or more generating processes, which could either reflect activity in the system or observations collected about entities. When the generating process behaves in an unusual way, it results in the creation of outliers. Therefore, an outlier often contains useful information about abnormal characteristics of the systems and entities, which impact the data generation process.

In this paper, inspired by the SSOD and FRCM algorithm, authors proposed algorithm which combines the fuzzy set theory, rough set theory and semi-supervised learning to detect outliers. Proposed algorithm utilizes known normal points and known outliers to partially supervise outlier detection and inherits SSOD algorithm by minimizing the objective function that takes clustering result, outlier assignments as well as mislabel punishment into consideration. Proposed algorithm also introduces FRCM clustering algorithm based on fuzzy and rough set theory in the semi-supervised outlier detection. Due to the lower approximation of each cluster belongs to the cluster without doubt, proposed algorithm only need discuss the possibility of the points in boundary to be assigned as outliers. The computation cost is inexpensive. Experimental results on artificial and real data sets show that the proposed algorithm works better than SSOD algorithm. Anomaly-based intrusion detection using fuzzy rough clustering is proposed by Chimphlee et al. [4], which is the application of FRCM to anomaly intrusion detection.

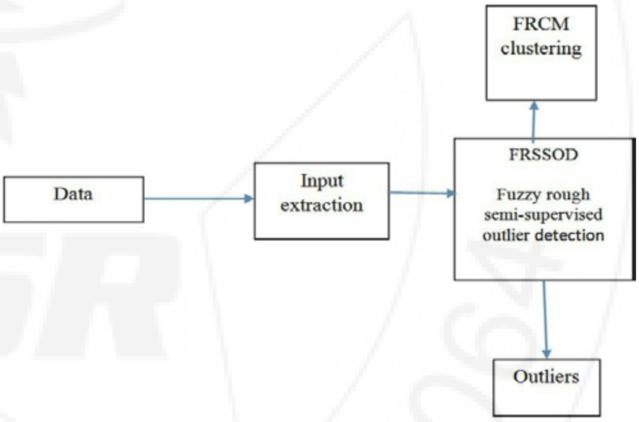

Figure 1.2: Block diagram

2. Literature Review

2.1 Novel Framework for Incorporating Labelled Examples into Anomaly Detection

In literature it has been presented that a principled approach for incorporating labeled examples into an anomaly detection task. It has been demonstrated that, with the addition of labeled examples, the anomaly detection algorithm can be guided to develop better models of the normal and abnormal behavior of the data, thus improving the detection rate and reducing the false alarm rate of the algorithm. A framework based on the finite mixture model is introduced to model the data as well as the constraints imposed by the labeled examples. Anomalies or outliers are aberrant observations whose characteristics deviate significantly from the majority of the data. Anomaly detection has huge potential benefits in a variety of applications, including the detection of credit card frauds, security breaches, network intrusions, or failures in mechanical structures. The labeled examples are used to guide the anomaly detection algorithm towards distinguishing data points that are hard to classify (e.g., anomalies that are located in high density regions or normal points that are located in low density regions).

2.2 Distance-Based Outliers: Algorithms and Applications

In literature it has been presented work that has been concerned with finding outliers (exceptions) in large, multidimensional datasets. The identification of outliers can lead to the discovery of truly unexpected knowledge in areas such as electronic commerce, credit card fraud, and even the analysis of performance statistics of professional athletes. Existing methods that we have seen for finding outliers can only deal efficiently with two dimensions/ attributes of a dataset. In literature it has been studied that the notion of D (distance-based) outliers. Specifically, It has been shown that (i) outlier detection can be done efficiently for large datasets, and for k-dimensional datasets with large values of k (e.g., k 5); and (ii), outlier detection is a meaningful and important knowledge discovery task.

2.3 Fuzzy Clustering Using Kernel Method

Classical fuzzy C-means clustering is performed in the input space, given the desired number of clusters. Although it has proven effective for spherical data it fails when the data structure of input patterns is non-spherical and complex. In literature it has been presented that a novel Kernel based fuzzy C-means clustering algorithm Its basic idea is to transform implicitly the input data into a higher dimensional feature space via a nonlinear map ,which increases greatly possibility of linear separately of the patterns in the feature space, then perform FCM in the feature space. Another good attribute of KFCM is that it can automatically estimate the number of clusters in the dataset.

2.4 Entropy-Based Outlier Detection

This section presents a modified entropy-based outlier detection algorithm, using both positive and negative data. The suggested algorithm is inspired by the LSA algorithm The aim of LSA algorithm is to find k expected outliers in the data set. This algorithm takes the number of outliers (assumed to be k) as input and iteratively improves the value of object function. The main phases of LSA algorithm are as follows:

1. In the iteration process, for each point labeled as non-outlier, its label is exchanged with each of the k outliers. Subsequently the entropy objective is re-evaluated.
2. If the entropy decreases, the label of the outlier point that achieves the lowest entropy is changed with the non-outlier point. Then the algorithm proceeds to the next point.
3. When all non-outlier points have been checked for possible replacement, one sweep is completed.
4. If at least one label was changed in a sweep, a new sweep is initiated.
5. The algorithm terminates when a full sweep does not change any labels. This is a local optimum that is reached.

3. Proposed System

The proposed system inspired by the SSOD and FRCM algorithm, it proposes the proposed algorithm which combines the fuzzy set theory, rough set theory and semi-supervised learning to detect outliers. proposed algorithm utilizes known normal points and known outliers to partially supervise outlier detection and inherits SSOD algorithm by minimizing the objective function that takes clustering result, outlier assignments as well as mislabel punishment into consideration. Proposed algorithm also introduces FRCM clustering algorithm based on fuzzy and rough set theory in the semi-supervised outlier detection.

3.1 Fuzzy Nearest Neighbour Algorithms

Many proposals have been presented since the publication of the first works in the field. These proposals focus not only on improvements to the classical model, but also other topics such as the use of different extensions of fuzzy sets, the addition of a pre-processing mechanism based on data reduction, or the development of real-world applications, describing instances of problems tackled successfully by fuzzy nearest neighbour techniques. This section is devoted to surveying relevant works in these directions, describing the key elements of each approach. After the survey, several common properties of the methods are identified and described. These properties are used to characterize the main approaches surveyed, providing with an insight into the existing differences in the design of the methods.

3.1.1 A Survey On Fuzzy Nearest Neighbour Classification

Since the presentation of the very first proposals, fuzzy nearest neighbor classification has become a distinctive area within the field of nearest neighbor classification and instance based learning. The addition of FST based mechanisms to the traditional approaches has enabled very accurate and flexible classification models to be defined, with outstanding results when applied to supervised learning problems. Through this section, both classical approaches and new extensions will be surveyed, including proposals based on possibility theory, intuitionistic sets, fuzzy rough sets and data pre-processing. A description of other interesting proposals using both nearest neighbor classification and fuzzy sets is also included. The survey is finished with several remarkable examples of applications of fuzzy nearest neighbor classification to real-world scenarios.

3.1.2 Nearest Neighbour Algorithms Based On Fuzzy Sets Theory

It is based on a learning scheme of class memberships, providing each training instance with membership array which defines its fuzzy membership to each class. After the learning process, the final classification is performed similarly to k-NN, but every neighbor uses its membership array for the voting rule, instead of just giving one vote as in the crisp k-NN. Fuzzy KNN, the classifier described in this work, has been the baseline of many advanced methods hybridizing FST and k-NN classifiers. Furthermore, there are plenty of applications in many fields of research based on this model, mainly due to its good behaviour when tackling supervised learning problems.

4. Conclusion

Proposed algorithm in this paper combines the fuzzy set theory, rough set theory and semi-supervised learning to detect outliers and is a new try in area of outlier detection for semi-supervised learning. Without considering those points located in lower approximation of a cluster, proposed algorithm only need discuss the possibility of the points in boundary to be assigned as outliers and has many advantages over SSOD. Proposed algorithm uses labeled normal and outliers and as well as samples without labels and can improve outliers detection accuracy and reduce false alarm rate under the guidance of labeled samples.

5. Future Work

In the future, we will research these problems. In addition, the speed of the algorithm is depended on the times of iteration, the speed of fuzzy rough C-means clustering, and the size of unlabelled samples, we also will pursue to improve the speed of this algorithm by taking some measures in the future. Proposed algorithm will be applied to many outlier detection fields which has only partially labeled samples, especially that does not make a certain judgment in uncertain conditions.

References

[1] V. Barnett, T. Lewis, Outliers in Statistical Data, 3rd ed., John Wiley & Sons, New York, 1984.

[2] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

[3] M.M. Breunig, H.P. Kriegel, R.T. Ng, J. Sander, Lof: identifying density-based local outliers, in: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, vol. 29, issue 2, ACM Press, Dalles, TX, 2000, pp. 93-104.

[4] W. Chimphlee, A.H. Addullah, M.N.M. Sap, S. Srinoy, S. Chimphlee, Anomaly-based intrusion detection using fuzzy rough clustering, in: 2006 IEEE International Conference on Hybrid Information Technology, vol. 1, issue 9-11, IEEE, Springer, Cheju Island, Korea, 2006, pp. 329-334.

[5] J.C. Dunn, Some recent investigations of a new fuzzy partition algorithm and its application to pattern classification problems, Journal of Cybernetics 4 (1974) 1-15.

[6] J. Gao, H.B. Cheng, P.N. Tan, A novel framework for incorporating labeled examples into anomaly detection, in: Proceedings of the Second SIAM International Conference on Data Mining, Bethesda, Maryland, USA, 2006, pp. 593-597.

[7] J. Gao, H.B. Cheng, P.N. Tan, Semi-supervised outlier detection, in: Proceedings of the ACM Symposium on Applied Computing, vol. 1, ACM Press, Dijon, France, 2006, pp. 635-636