IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

Unsupervised Distance-Based Outlier Detection using Reversible KNN with Fuzzy Clustering

S. Vasuki

Dept. of Computer Applications, .J .J College of Arts and Science (Autonomous), Pudukkottai, India

Corresponding Author: vasuki_msitm@yahoo.co.in, 9443411125

DOI: https://doi.org/10.26438/ijcse/v7i6.11951199 | Available online at: www.ijcseonline.org

Accepted: 23/Jun/2019, Published: 30/Jun/2019

***Abstract*:** The detection of outliers in high-dimensional data raises some of the challenges of "dimension curse". A major point of view is that the concentration of distances, that is, the distance trends in high-dimensional data becomes illegible, making it difficult to detect outliers by marking all points as values by a distance-based approach. In this paper, implement that the idea of distance-based methods can produce more contrast outliers in high-dimensional environments to provide evidence to support the idea that this view is too simple. In addition, we show that high dimensions can have different effects when there is no oversight to re-examine the concept of a more recent inverse neighbor in the context of atypical detection. It has recently been observed that the distribution of the inverse neighborhood count of points deviates in a high dimension, which causes a phenomenon called a hubness. This work provide information on how some antihubs rarely appear in the k-NN list at other points, and explain the connection between antihubs, outlier values and existing unsupervised outlier detection methods. In evaluating the classical approach to k-NN, angle-based techniques are designed for high-dimensional data, local outliers based on density, and various methods based on anti-sheathing. Combining and real-world data, this work provide new information about the utility of reverse neighborhood counting to detect outliers without supervision.

Keywords: Clustering, data mining, fuzzy c-means, outliers, unsupervised learning.

I. INTRODUCTION

Data mining can be defined as an activity that allows the new non-trivial extraction of information in large databases. Our goal is to discover the use of machine learning techniques combining statistical and database technology to extract hidden patterns, trends, data, or other unexpected delicate relationship. This emerging discipline is a continuous data mining of large data sets in a wide variety of business environments, scientific and engineering applications, and spatio-temporal data mining. For sequential data, we refer to the requirement to provide data related to the indicator.

An outlier (abnormal) is detected to identify a behavioral pattern in which the establishment of the task does not conform to the rule. Benefits are very atypical values because they can exist in several ways, such as detecting intruders and fraud, as well as medical diagnosis of critical and actionable information. Depending on the tag and/or periodic instance, the detection of outlier value tasks can be classified as the presence of supervised, semi-supervised, and unsupervised outliers. Among these classes, unsupervised methods are widely used for other types of requirements, usually accurate and represent labels. Unsupervised methods primarily detect outliers based on measures of distance or similarity.

Based on the availability of these tags, the data anomaly detection operation is one of three models:

1) Form a monitoring of abnormalities under the supervision of the case to consider ways and means of marking the availability of normal and abnormal categories of training data sets.

2) The semi-supervised anomaly detection mark formed under normal conditions, considering the form and supervision means does not require the availability of such an exceptional tag.

3) Training data is not required for detecting anomalies in an unsupervised mode of operation.

II. LITERATURE SURVEY

Milos Radovanovi et al. [1] described evidence supporting the view that this view is too simple. They provide information about the infrequent occurrence of certain points (rebels) in the k-NN list at other points and explain the connection between rebellion, outliers and existing unsupervised outlier detection methods.

Ranjita Singh et al. [2]. A prominent histogram method is proposed to analyze the eigenvalues of the data set in detail. It also proposes a fuzzy mining algorithm based on Apriori Tid method to find fuzzy association rules from a given quantitative transaction. Due to the fuzzy rule mining method, it activates a small number of false positives when detecting abnormalities in a large database, and has a high true positive rate value and a low false negative rate value.
Colin Chen et al. [3]. An overview of robust regression methods is provided, the ROBUSTREG process is described, and the process is used to adjust the regression model and display outliers and leverage points. They also analyzed the scalability of the ROBUSTREG process for applications in data cleansing and data mining.

J. Michael Antony Sylvia et al. [4]. It recommends detecting unsupervised anomalies in high-dimensional data. The detection of anomalies in high-dimensional data shows that as the dimension increases, there is a center and rebellion. A hub is a frequently occurring point in the list of recent neighbors. Antihubs are points that don't often appear in the kNN list.

Jayshree S. Gosavi et al. [5]. The purpose of the proposed work is to develop and compare some methods for detecting outliers and propose ways to improve them. The proposed work includes details of the development and analysis of atypical value detection algorithms, such as local anomaly factors (LOF), atypical value factors (LDOF) based on local distance, and are affected.

Pamula et al. [6]. An effective outlier detection method is proposed. By applying the K-means algorithm, data instances that are unlikely to be atypical value candidates are identified by using the radius of each group, and these data instances are eliminated from the data set. This study establishes the concept of fuzzy approximation c-means (FRCM) to analyze groupings.

III. METHODOLOGY

A. Unsupervised Learning

In machine learning, unsupervised learning refers to the problem of trying to find hidden structure in unlabeled data. Unsupervised learning seems to be much more difficult: the goal of the computer is to learn how to do something that we didn't tell you how to do. In fact, there are two ways to learn unsupervised. The first method is to teach the agent without a clear classification, but to express success through some kind of reward system.

B. Nearest Neighbour Classifier

In pattern recognition, the nearest neighbor algorithm (KNN) is a method of classifying objects according to training samples closest to the feature space.

K-NN is an instance-based or lazy learning-based method where functions are only localized and all calculations are deferred to classification. It is called lazy because it does not have any training phase or minimal training phase. All the training data is needed during the testing phase and it uses all the training data. So if we have large number of data set then we need special method to work on part of data which is heuristic approach.

C. Fuzzy C-Means

Fuzzy c-means (FCM) is a clustering method that allows a portion of data to belong to two or more clusters.
The algorithm works by assigning membership to each data point corresponding to each grouping center based on the distance between the grouping and the data points.
The more data is near the center of the cluster, the more itbelongs to a particular cluster center. Obviously, the sum of the members of each data point must be equal to 1.

D. Data Partitioning

The pre-processed data is divided into the number of clients from the central administrator node (ie, the server) based on the data requests made by the required number of clients. Each client will process this partition data to identify outliers based on the applied algorithmic policies.

E. Outlier Detection

Outlier values will initially be applied to distributed clients and identification techniques that detect outliers, and the results of the program will be integrated into the final stages of server computing. To this end, the KNN methods of the proposed anomaly detection algorithm are ABOD and INFLO.

IV. ALGORITHM USED

A. Unsupervised Learning

Unsupervised learning is an automatic learning task used to infer functions that describe hidden structures from unmarked data. This learning method is deeply related to the problem of density estimation in statistics.

B. *Introduction to KNN*

K-NN is an instance-based or lazy learning-based learning where functions are only localized and all calculations are deferred to classification. It is called lazy because it does not have any training phase or minimal training phase.

The detailed steps of K - Nearest Neighbor Algorithm

- For each example of training data(x, f(x)), add the training data to the list of training examples.
- Given a query instance x_q to be classified,
- Let $x_1,x_2...x_k$ denote the k instances from training examples that are nearest to x_q .

- Return the class that represents the maximum of the k instances

C. *Fuzzy Algorithm*

In this algorithm, two preliminary operations must be performed: centroid calculation, which is simply the average vector of the data distribution, and the grouping operation, which is performed by the FCM.

1. Define terms and language variables (initialization).
2. Build member functions (initialization)
3. Build a rule base (initialization)
4. Use the membership function to convert the fuzzy value to a fuzzy value (fuzzy)
5. Evaluate the rules in the rule based member function (reasoning)
6. Combine the results of each rule (inference)
7. Convert the outcome data to a non-diffuse value (defuzzification)

V. PROBLEM STATEMENT

A. Previous Studies

Depending on the tag and/or periodic instance, the detection of outlier tasks can be classified as the presence of supervised, semi-supervised, and unsupervised outliers. Unsupervised methods are widely used in these categories because many other categories require accurate and representative labels that are often expensive. Unsupervised methods include distance-based functions or similarity measures to detect outliers based primarily on distance. It is generally accepted that long distances are meaningful because of the "curse of dimensions", and the distance of the server is increasingly difficult to identify a given dimension due to the measurement of distance. In the distance measured concentrations of the effect of outliers unsupervised means for becoming a high-dimensional space almost as good as each point.

B. Proposed Work

The key is to understand how to increase the dimensions Anomaly Detection. As interpreted by the real challenge of "dimension curse" brought different and each point has become an outsider in almost a good high-dimensional space of the generally accepted view. We will provide further evidence that challenges this view of the (re) test method of motivation. Restore the most recent count neighbor in the past they have proposed a method to express the data points outlierness, but no vision, in addition to the basic instincts are provided why these counts should represent a significant outlier scores. Recent observations restored to the neighbor count increased data dimension worth considering re-value anomaly detection tasks affected. This work established a technique in which the concept of concentrators, especially the antihub algorithm (points with low concentrators), was integrated into the result set obtained from techniques such as KNN and fuzzy C-means (FCM) to detect outliers. The type value is mainly to reduce the computational time. It compares the results of all the techniques by applying it on three different real data sets. The experimental results show that KNN Antihub significantly reduces computation time compared to Antihub and FC Antihub in all comparisons. The conclusion is that when Antihub is applied to KNN, it performs better.

VI. IMPLEMENTATION

A. *Method*

Our experimental evaluation found that the two methods described in the previous section, showing $AntiHub_k$ and $AntiHub^2{}_k$, where k is the number of nearest neighbors is used. We will always take the Euclidean distance. For convenience, K may be referred to as a fraction of the size of n data sets.

B. Data Sets

In this experiment, Breast Cancer Wisconsin Diagnostic dataset (WDBC) and Breast Cancer Wisconsin Prognostic dataset (WPBC) are used All the algorithms of proposed method are implemented in MATLAB (R209a). Data is collected from UCI Machine Learning Repository.

WDBC

This data set contains 569 medical diagnostic records, each with 32 features of attributes (ID, decisions attribute (diagnosis), and 30 real valued input features). The diagnosis is binary: Benign and Malignant.

WPBC

This data set contains 198 instances and 33 features.

The attributes of this dataset are almost identical to those of WDBC, but they also have three additional features: time, tumor size, and lymph node status. The result is binary: recurring and non-recurring.

C. Pre-processing

The missing values are replaced with appropriate values by filling the corresponding mean-mode value. All features are expressed as actual value measurements, but they must be discretized according to the purpose of the approximate set theory. By applying equal width binning with the number of bins 5, the dataset is discretized and new dataset with crisp values are produced.

D. Outlier Detection

Distance based approach is applied in each cluster to findthe data points those are closest to the centroid and theyare pruned. Finally K-nearest neighbour is applied forremaining data points and outliers are detected based on top-n fashion distance approach. The details of outliers in different datasets are summarized in table 1 and table 2.

Table 1. Outlier Detection of WDBC Dataset

No .of Data Points in WDBC	No. of Outliers
316	13
253	11

Table 2. Outlier Detection of WPBC Dataset

No .of Data Points in WPBC	No. of Outliers
97	2
101	2

Table 3. Performance Measure for Proposed Method

S. No	Algorithm	Consider for All features		Feature Subset for Proposed method	
		Accuracy %	Time in sec	Accuracy %	Time in sec
1	KNN	96.4912	0.39	98.1982	0.6
2	Proposed Method	98.3684	0.2	99.0991	0.5

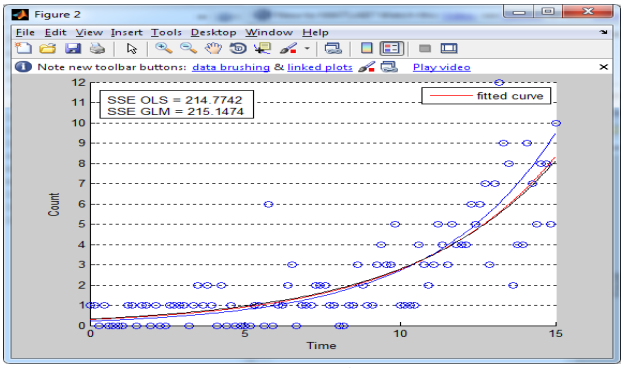

Fig. 1 Feature selection set for outlier detection

Most of the methods designed in existing algorithms use the training data available at the beginning of the learning process for feature selection. The proposed method applies feature selection and eliminates outlier data points in a natural data points. Therefore, our method generates different feature subsets, which reduces the computational complexity of the classification algorithm.

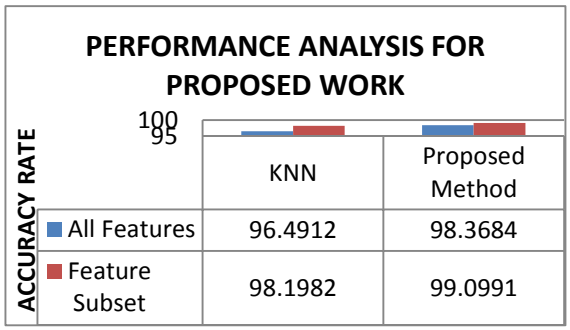

Fig 2. performance analysis for KNN and FCM

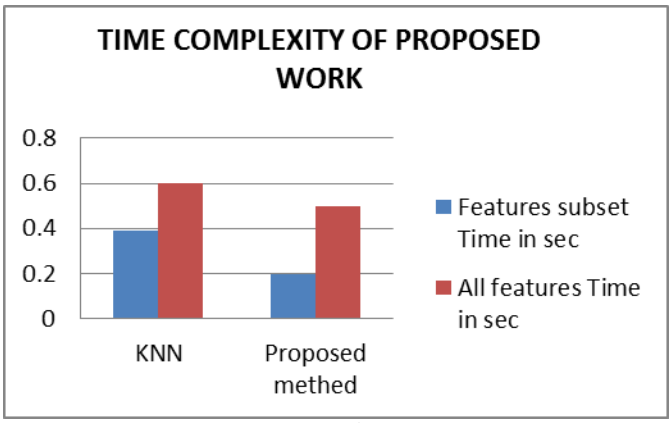

Fig3. Time complexity of proposed method

The proposed method consists of three phases. In the first phase, a set of patterns are classified by FCM clustering. Binary classification is the task of classifying the members of a given data set into two groups on the basis of whether they have some property or not. The binary classification task in the context of medical domain is to differentiate between normal and abnormal situations. In the second phase, outliers are constructed by a distance-based technique, and finally rough set feature selection is applied to find minimal feature subset for classification. The proposed method is illustrated using below specified Fig.

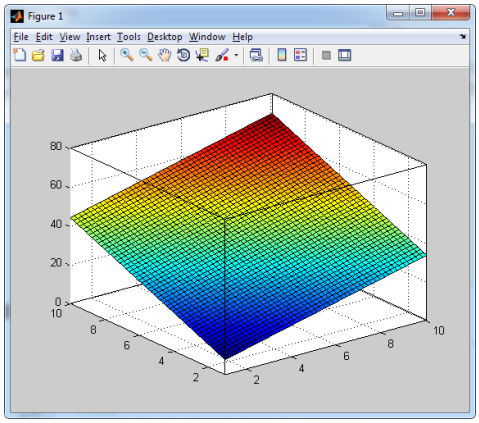

Fig.4 Performances analysis for proposed method

VII. Conclusion and Future Work

This work presents an efficient hybrid method for rough set feature selection based on KNN with FCM clustering and distanced based outlier detection. The entire model has been implemented on breast cancer data sets. Initially, FCM clustering is used to generate the partition and then by applying the distance based outlier, deviating data points have been removed. Finally, minimal feature subset has been obtained by applying degree of dependency based approach of rough set theory. Traditional feature selection algorithms find feature subset using whatever training data is given to them. The proposed method promotes the idea to actively

select features from natural grouping of data and it avoids anomalous data points. Hence, the reduct obtained by our method has a positive impact on the results of classification algorithms while compared to other feature selection methods. We also affirm that the KNN with FCM algorithm is the best performing algorithm which provides 100 percent and nearly 93 percent accuracy in classifying the WDBC and WPBC data sets respectively.

High k values may be useful, but: cluster boundaries may cross, producing insignificant results for detecting local outliers. Therefore, it is necessary to determine the best neighborhood size in the future. Computational complexity occurs; the approximate search/indexing method of NN is no longer valid. Then it is possible to solve this problem for big k. In the future work, the method of detecting (semi-) supervised outlier values is extended. Explore the relationship between intrinsic dimensions, distance concentration, reverse center and their impact on subspace methods to detect outliers. Investigate secondary distance/similarity measures, such as distance from shared neighbors.

REFERENCES

[1]. Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection by Milos Radovanovi ˇ c, Alexandros Nanopoulos, and Mirjana Ivanovi, IEEE Transactions On Knowledge And Data Engineering, Revised October 2014

[2]. An Efficient Anomaly Detection System Using Featured Histogram and Fuzzy Rule Mining by Ranjita Singh, Sreeja Nair., January 2014 ISSN: 2277 128X Volume 4, Issue 1, January 2014 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering

[3]. Robust Regression and Outlier Detection with the ROBUSTREG Procedure by Colin Chen, SAS Institute Inc., Cary, NC, Paper 265-27, Feb 2013, IEEE Trans. Automat Control 19, 716–723.

[4]. Recursive Antihub[2] Outlier Detection In High Dimensional Data by J.Michael Antony Sylvia, Dr.T.C.Rajakumar. Vol-2, Issue-8 PP. 1269-1274, 3 0 A u g u s t 2 0 1 5, Global Journal of Advanced Research(GJAR) Vol-2, Issue-8 PP. 1269-1274 ISSN: 2394-5788

[5]. Unsupervised Distance-Based Outlier Detection Using Nearest Neighbours Algorithm on Distributed Approach: Survey by Jayshree S.Gosavi , Vinod S.Wadne, (An ISO 3297: 2007 Certified Organization) IJIRCCE ,Vol. 2, Issue 12, December 2014

[6]. Pamula, Rajendra, Jatindra Kumar Deka, and Sukumar Nandi. "An outlier detection method based on clustering." Emerging Applications of Information Technology (EAIT), 2011 Second International Conference on. IEEE, 2011.

[7]. W. Jin, A. K. H. Tung, and J. Han. Finding top-n local outliers in large database. In 7th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pages 293–298, 2001.

[8]. E.M. Knorr and R. T. Ng. Algorithms for mining distancebased outliers in large datasets. In Proceedings 24th Int. Conf. Very Large Data Bases, pages 392–403, New York, USA, 1998.

[9]. Lee, S. J. Stolfo, and K. W. Mok. A data mining framework for building intrusion detection models. In IEEE Symposium on Security and Privacy, pages 120–132, May 1999.

[10]. J. Liu and P. Gader. Outlier rejection with mlps and variants of RBF networks. In Proceedings of The 15th Int. Conf. on Pattern Recognition, pages 680–683, 2000.

[11]. S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In Proceedings of the 2000 ACM SIGMOD Int. Conf. on Management of Data, pages 427–438, Dallas, Texas, May 2000.

[12]. P. J. Rousseeuw and A. M. Leroy. Robust Regression and Outlier Detection. John Wiley and Sons, New York, October 1987.

[13]. G. Williams, R. Baxter, H. He, S. Hawkings, and L. Gu. A comparative study of RNN for outlier detection in data mining. In Proceedings of the 2nd IEEE Int. Conf. on Data Mining, Maebashi City, Japan, December 2002.

[14]. K. Yamanishi, J. Takeuchi, G. Williams, and P. Milne. Online unsupervised outlier detection using finite mixtures with discounting learning algorithm. In Proceedings The Sixth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pages 320–324, August 2000.

[15]. T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: A new data clustering algorithm and its applications. Data Mining and Knowledge Discovery, 1(2):141–182, 1997

Authors Profile

Dr. S.Vasuki, the educational qualification of author is Ph.D in computer science in September 2018 in J.Jcollege of arts and science(Autonomous),pudukkottai affliated to Bharathidasan University, tiruchirappalli and completed State level Eligiility Test(TNET) in 2018 conducted by Mother Terasa University Tamilnadu,India. ,M.Phil.in computer science done in AlgappaUniversity, Karaikudi,Tamilnadu, India. In the year of April 2008.P.G degree M.S(IT&M) in Ayya Nadar Janaki Ammal College, Sivakasi,Tamilnadu, India, in the year of April 2003.The author's major area of interest is data mining. She presented and participated in various International and national conferences. she has 12 years teaching experience and 10 years research experience .

