ISSN: 2321-7782 (Online)
Impact Factor: 6.047

Volume 4, Issue 7, July 2016

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Reverse Nearest Neighbours in Unsupervised Distance-Based Outlier Detection using FCM

N. Nagarathinam[1]
Research Scholar in Computer Science,
H.H.The Rajah's College(Autonomous),
Pudukkottai, India

K. Karpagam[2]
Research Guide, Assistant Professor of Computer Science,
H.H.The Rajah's College (Autonomous),
Pudukkottai, India

Abstract: ***Outlier detection in high-dimensional data presents various challenges resulting from the "curse of dimensionality". A prevailing view is that distance concentration, i.e., the tendency of distances in high-dimensional data to become indiscernible, hinders the detection of outliers by making distance-based methods label all points as almost equally good outliers. In this paper, we provide evidence supporting the opinion that such a view is too simple, by demonstrating that distance-based methods can produce more contrasting outlier scores in high-dimensional settings. Furthermore, we show that high dimensionality can have a different impact, by reexamining the notion of reverse nearest neighbors in the unsupervised outlier-detection context. It was recently observed that the distribution of points' reverse-neighbor counts becomes skewed in high dimensions, resulting in the phenomenon known as hubness. We provide insight into how some points (antihubs) appear very infrequently in k-NN lists of other points, and explain the connection between antihubs, outliers, and existing unsupervised outlier-detection methods. By evaluating the classic k-NN method, the angle-based technique designed for high-dimensional data, the density-based local outlier factor and influenced outlierness methods, and antihub-based methods on various synthetic and real-world data sets, we offer novel insight into the usefulness of reverse neighbor counts in unsupervised outlier detection.***

Keywords: Baye's theorem, clustering, data mining, fuzzy c-means, 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 in today's extensive and diverse business environment, science and engineering applications and Spatial-temporal data mining is all about data mining of large data sets continuous discovery. For sequential data, we mean that the requirement to provide data relating to indicators.

Outliers (abnormality) detection means to identify the establishment of a task does not meet the regular patterns of behavior. Interest is strong outliers, because they can be in various areas, such as intrusion and fraud detection and medical diagnosis of critical and actionable information. Task detection of outliers can be classified as supervised, semi-supervised and unsupervised presence of outliers according to the label and / or periodic instances. In these classes, unsupervised method is widely used for other types of needs, accurate and representative of a label often get expensive. Unsupervised methods mostly based on a distance measure or similarities to detect outliers.

Based on the availability of such tags data abnormality detection operation is one of three models:

1. Supervision anomaly detection is formed under the supervision of the case to consider ways and means of marking the availability of normal and abnormal categories of training data set.

2. Semi-supervised anomaly detection mark formed under normal circumstances, to consider ways and means of supervision does not require such exception label availability.

3. Anomaly detection, in an unsupervised mode operation technique does not require training data.

II. LITERATURE SURVEY

Milos Radovanovi c et al [1] describes the evidence supporting the opinion that such a view is too simple. They provide insight into how some points (antihubs) appear very infrequently in k-NN lists of other points, and explain the connection between antihubs, outliers, and existing unsupervised outlier-detection methods.

Ranjita Singh et al [2]. proposed a featured histogram approach that used for the detailed analysis of feature values of data set. He also proposes a fuzzy mining algorithm based on the Apriori Tid approach to find fuzzy association rules from given quantitative transactions. Due to fuzzy rule mining approach it triggers a small number of false positives the detection of anomalies in large databases with high value of true positive rate and low value of false negative rate.

Colin Chen et al [3]. provides an overview of robust regression methods, describes the procedure ROBUSTREG, and illustrates the use of the procedure to fit regression models and display outliers and leverage points. They also discusses scalability of the ROBUSTREG procedure for applications in data cleansing and data mining.

J.Michael Antony Sylvia et al [4]. proposes the unsupervised anomaly detection in high dimensional data. Anomaly detection in high dimensional data exhibits that as dimensionality increases there exists hubs and antihubs. Hubs are points that frequently occur in k nearest neighbor lists. Antihubs are points that infrequently occur in kNN lists.

Jayshree S.Gosavi et al [5]. proposed work aims at developing and comparing some of the unsupervised outlier detection methods and propose a way to improve them. This proposed work goes in details about the development and analysis of outlier detection algorithms such as Local Outlier Factor (LOF), Local Distance-Based Outlier Factor(LDOF), Influenced.

Pamula et al [6]. proposes an efficient outlier detection method by applying K-means algorithm to recognize data instances which are not probable candidates for outliers by using the radius of each cluster and remove those data instances from the dataset. The study establishes the idea of the fuzzy rough c-means (FRCM) to analyze clustering.

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 much harder: the goal is to have the computer learn how to do something that we don't tell it how to do! There are actually two approaches to unsupervised learning. The first approach is to teach the agent not by giving explicit categorizations, but by using some sort of reward system to indicate success.

B. KNN

KNN is one of the most simple and straight forward data mining techniques. It is called Memory-Based Classification as the training examples need to be in the memory at run-time, when dealing with continuous attributes the difference between the attributes is calculated using the Euclidean distance.

C. Nearest Neighbour Classifier

In pattern recognition, the k-nearest neighbour algorithm (KNN) is a method for classifying objects based on closest training examples in the feature space.

K-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until 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.

D. Baye's Theorem

Bayes' Theorem is a theorem of probability theory originally stated by the Reverend Thomas Bayes. It can be seen as a way of understanding how the probability that a theory is true and how is affected by a new piece of evidence. It has been used in a wide variety of contexts, ranging from marine biology to the development of "Bayesian" spam blockers for email systems.

$$P(T|E) = \frac{P(E|T) \times P(T)}{P(E|T) \times P(T) + P(E|\neg T) \times P(\neg T)}$$

In this formula, T stands for a theory or hypothesis that we are interested in testing, and E represents a new piece of evidence that seems to confirm or disconfirm the theory. For any proposition S, we will use P(S) to stand for our degree of belief, or "subjective probability," that S is true. In particular, P(T) represents our best estimate of the probability of the theory we are considering, prior to consideration of the new piece of evidence. It is known as the prior probability of T.

E. Fuzzy C-Means

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This algorithm works by assigning membership to each data point corresponding to each cluster centre on the basis of distance between the cluster and the data point. 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.

F. Data Partitioning

The pre-processed data is divided into number of clients from central supervisor node i.e. server as per the data request made by desired number of clients. This partitioned data will be then processed by individual clients to identify outliers based on applied algorithm strategy.

G. Outlier Detection

Outliers will initially apply to distributed client and the detected outlier identification technology, program results will be integrated into server computing stray final stage. To this end, strategic recommendations anomaly detection algorithm KNN method is ABOD and INFLO.

IV. ALGORITHM USED

A. Unsupervised Learning

Unsupervised learning is the machine learning task of inferring a function to describe hidden structure from unlabeled data. Unsupervised learning is closely related to the problem of density estimation in statistics.

B. Introduction to KNN

K-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. It is called lazy because it does not have any training phase or minimal training phase. **K - Nearest Neighbor Algorithm**

- For each training example <x,f(x)>, add the example to the list of training_examples.
- Given a query instance xq to be classified,
- Let x1,x2...xk denote the k instances from training_examples that are nearest to xq .

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

C. Fuzzy Algorithm

In this algorithm, two preliminary opeations must be performed: the calculation of the centroid, that is simply the mean vector of the data distribution, and the clustering operation, that is performed by means of the FCM.

1. Define the linguistic variables and terms (initialization)
2. Construct the membership functions (initialization)
3. Construct the rule base (initialization)
4. Convert crisp input data to fuzzy values using the membership functions (fuzzification)
5. Evaluate the rules in the rule base (inference)
6. Combine the results of each rule (inference)
7. Convert the output data to non-fuzzy values (defuzzification)

V. PROBLEM STATEMENT

A. Existing Work

Task detection of outliers can be classified as supervised, semi-supervised and unsupervised presence of outliers according to the label and / or periodic instances. Within these categories, unsupervised method is widely used, because more other categories require accurate and representative labels often get expensive. Unsupervised methods include methods based on a distance measure or similarity to detect outliers is largely based on distance. The generally accepted view is that, due to the "curse of dimensionality", long distance makes sense, because of the distance measurement, distance Serve becoming increasingly difficult to identify the dimensions set. 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 establishes a technique where the concept of hubness, especially the antihub (points with low hubness) algorithm is embedded in the resultant clusters obtained from techniques such as KNN and Fuzzy C Means (FCM) to detect the outliers mainly to reduce the computation time. It compares the results of all the techniques by applying it on three different real data sets. The Experimental results demonstrate that in all the comparisons, KNN Antihub provides a significant reduction in computational time than Antihub and FC Antihub. It is concluded that when the Antihub is applied into KNN, it outperforms well.

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 data set are nearly the same as WDBC yet it has three additional features Time, Tumor size and Lymph node status. The outcome is binary: Recurrent and Non-recurrent.

C. Pre-processing

The missing values are replaced with appropriate values by filling the corresponding mean-mode value. All features are represented in real valued measurement but they must be discretized for the purpose of rough 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 find the data points those are closest to the centroid and they are pruned. Finally K-nearest neighbour is applied for remaining data points and outliers are detected based on top-n fashion distance approach. The details of outliers 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

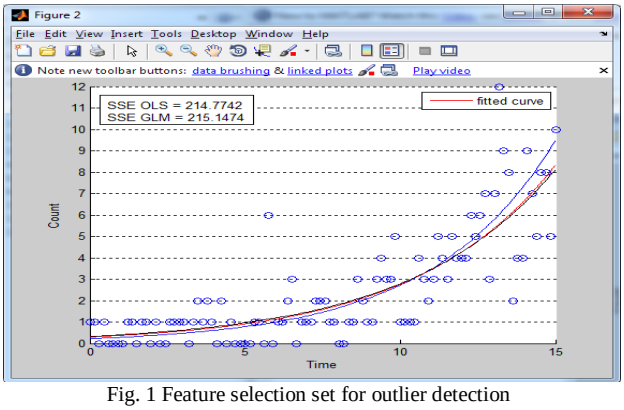

Fig. 1 Feature selection set for outlier detection

Most of the methods designed in existing algorithms use feature selection with the given training data which are available at the start of the learning process. The proposed method applies feature selection on natural grouping of data and it removes anomalous data points. Therefore, different feature subsets are generated by our method and they reduce the computational complexity of the classification algorithms.

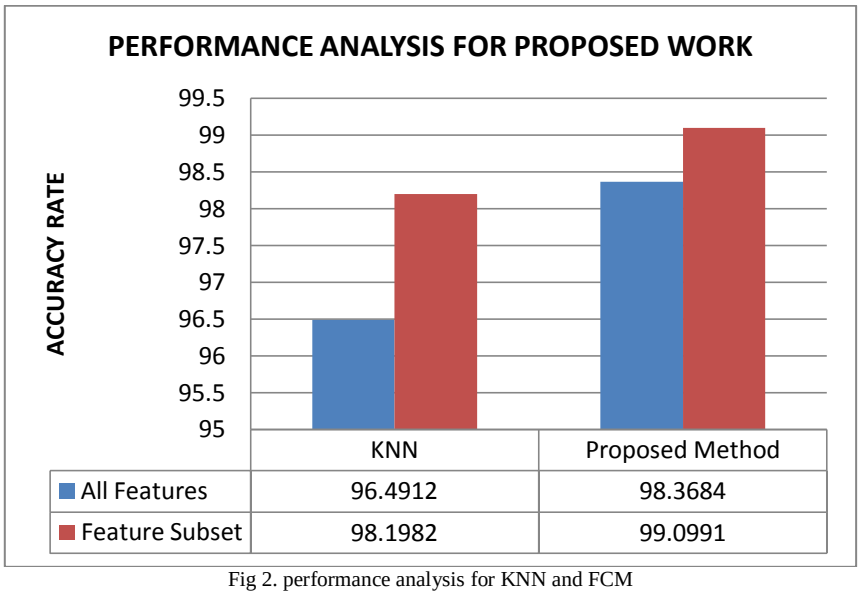

Fig 2. performance analysis for KNN and FCM

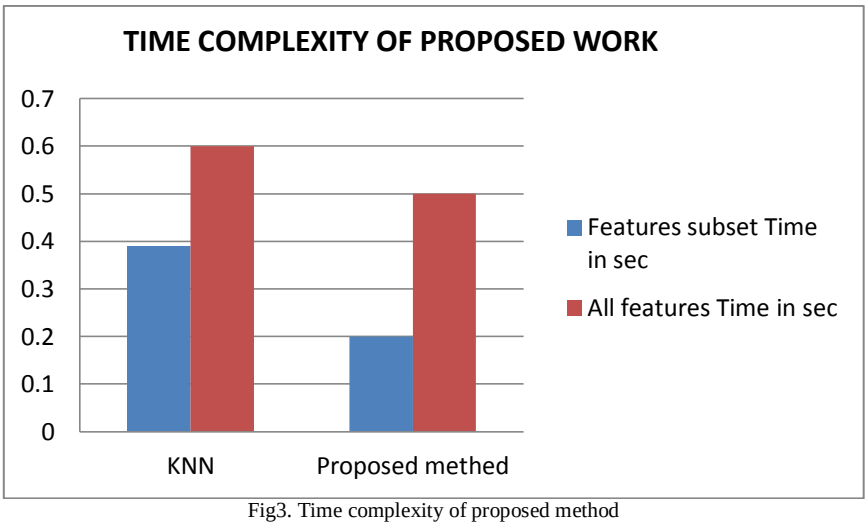

Fig3. Time complexity of proposed method

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 Fig.

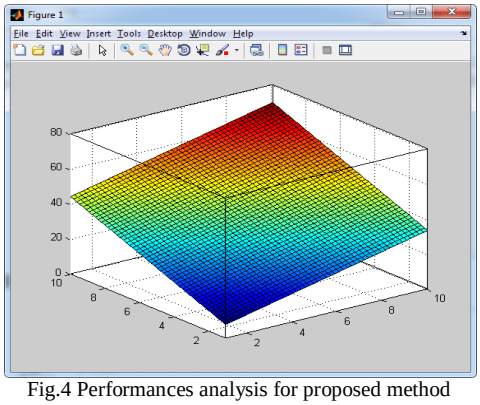

Fig.4 Performances analysis for proposed method

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 values of k can be useful, but: Cluster boundaries can be crossed, producing meaningless results of local outlier detection. So needs to determine optimal neighborhood size(s) in future. Computational complexity is raised; approximate NN search/indexing methods do not work anymore. So it is possible to solve this for large k. In future work Extend to (semi-)supervised outlier detection methods. Explore relationships between intrinsic dimensionality, distance concentration, (anti-)hubness, and their impact on subspace methods for outlier detection. Investigate secondary measures of distance/similarity, such as shared-neighbor distances.

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.
3. Robust Regression and Outlier Detection with the ROBUSTREG Procedure by Colin Chen, SAS Institute Inc., Cary, NC, Paper 265-27, Feb 2013
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.
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.

