A NOVEL APPROACH FOR K-NN ON UNSUPERVISED DISTANCE-BASED OUTLIER DETECTION

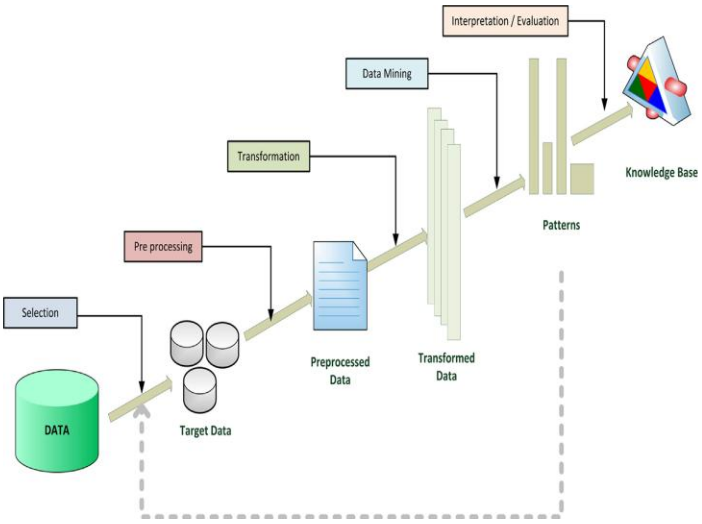
Syeda Khaja Momina Banu[1], P.Praveen[2]
[1]M.Tech in Computer Science and Engineering, [2]Asst.Prof.in Computer Science and Engineering
SR Engineering College, Warangal, Telangana, India

Abstract: Outlier detection in high-dimensional information presents different difficulties coming about because of the "scourge of dimensionality". A predominant view is that separation fixation, i.e., the inclination of separations in high-dimensional information to wind up incomprehensible, thwarts the discovery of exceptions by making separation based techniques mark all focuses as similarly great anomalies. In this paper, we give prove supporting the feeling that such a view is excessively basic, by exhibiting that separation based techniques can deliver all the more differentiating exception scores in high-dimensional settings. Moreover, we demonstrate that high dimensionality can have an alternate effect, by revaluating the thought of turnaround closest neighbors in the unsupervised exception recognition setting. It was as of late watched that the appropriation of focuses' switch neighbor include gets to be skewed high measurements, bringing about the wonder known as hubness. We give knowledge into how a few focuses (antihubs) seem occasionally in k-NN arrangements of different focuses, and clarify the association between antihubs, exceptions, and existing unsupervised anomaly discovery strategies. By assessing the great k-NN strategy, the edge based procedure intended for high-dimensional information, the thickness based nearby anomaly figure and impacted outlierness techniques, and antihub-construct strategies with respect to different engineered and genuine information sets; we offer novel understanding into the helpfulness of turnaround neighbour tallies in unsupervised Outlier discovery.

Keywords: Outlier detection, reverse nearest neighbors, high-dimensional data, distance concentration, Unsupervised learning.

I. INTRODUCTION

Data mining is called as data or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information from large amount of databases. Our goal is to discover the use of database technology to extract the hidden patterns, trends, data, or other unexpected delicate relationship. This emerging discipline in today's the business environment is extensive and diverse , 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. Outlier's detection means to identify the establishment of task which does not meet regular patterns of behaviour. Strong Outliers can be in different areas, such as intrusion and fraud detection and medical diagnosis of critical and actionable information. Task detection of outliers can be categorized into three types as unsupervised, supervised and semi supervised. Unsupervised is widely used for other type of needs, accurate and representative of a label to get expensive. Unsupervised methods mostly based on a distance measure similarities to detect outliers.



SYSTEM ARCHITECTURE

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

- The Supervision anomaly detection is formed under the supervision of the case to consider ways and means of marking of availability of normal and abnormal categories of training data set.
- The Semi-supervised anomaly detection mark formed under the normal circumstances, to consider ways and means of supervision does not require such exception label availability.
- The Anomaly detection, in an unsupervised mode operation technique does not require training data.

II. LITERATURE SURVEY

Milos Radovanovi c et al [1] describes the confirmation supporting the sentiment that such a view is excessively basic. They give knowledge into how a few focuses (antihubs) seem rarely in k-NN arrangements of different focuses, and clarify the association between antihubs, exceptions, and existing unsupervised anomaly discovery techniques.

N.Tomasev et al [2]. Proposed another calculation, Hubness data k-closest neighbor (HIKNN), which presents the k-event in development into the hubness-mindful k-closest neighbor voting system. Our assessment on high dimensional information indicates noteworthy enhancements over both the fundamental k-closest neighbor approach and all beforehand utilized hubness mindful methodologies. Closest Neighbor Voting in High Dimensional information gaining from past events.

M.Ivanovi et al [3]. We demonstrate that hubness, i.e., the propensity of high dimensional information to contain focuses (centers) that regularly happen in k-closest neighbor arrangements of different focuses, can be effectively abused in grouping. We approve our theory by exhibiting that hubness is a decent measure of point centrality inside a high dimensional information group and by proposing a few hubness-based bunching calculations, demonstrating that real center points can be utilized viably as bunch models or as aides amid the hunt down centroid based group setups.

J.Michael Antony Sylvia et al [4]. Proposes the unsupervised inconsistency discovery in high dimensional information. Inconsistency location in high dimensional information displays that as dimensionality increments there exists centers and antihubs. Centers are focuses that every now and again happen in k closest neighbor records. Antihubs are focuses that occasionally happen in kNN records.

JayshreeS.Gosavi et al [5]. Proposed work goes for creating and looking at a portion of the unsupervised exception discovery techniques and proposes an approach to enhance them. This proposed work goes in insights about the advancement and investigation of anomaly discovery calculations, for example, Local Outlier Factor (LOF), Local Distance-Based Outlier Factor (LDOF), and Influenced.

Pamula et al [6]. Proposes an effective anomaly recognition technique by applying K-implies calculation to perceive information cases which are not likely possibility for anomalies by utilizing the span of every group and expel those information occasions from the dataset.

III. METHODOLOGY

A. Unsupervised Learning

In machine learning, unsupervised learning alludes to the issue of attempting to discover shrouded structure in unlabeled information. Unsupervised learning appears to be much harder: the objective is to have the PC figure out how to accomplish something that we don't let it know how to do! There are really two ways to deal with unsupervised learning. The main approach is to instruct the operator not by giving unequivocal orders, but rather by utilizing some kind of reward framework to show achievement.

B. KNN

KNN is a standout amongst the most straightforward and straight forward information mining methods. It is called Memory-Based Classification as the preparation cases should be in the memory at run-time, when managing ceaseless qualities the contrast between the properties is figured utilizing the Euclidean separation.

C. Closest Neighbor Classifier

In example acknowledgment, the k-closest neighbor calculation (KNN) is a technique for arranging objects in view of nearest preparing cases in the element space. K-NN is a kind of occasion based learning, or lethargic realizing where the capacity is just approximated locally and all calculation is conceded until order. It is called languid in light of the fact that it doesn't have any preparation stage or negligible preparing stage. All the preparation information is required amid the testing stage and it utilizes all the preparation information. So in the event that we have expansive number of information set then we require unique strategy to chip away at a portion of information which is heuristic approach.

D. Baye's Theorem

Bayes' Theorem is a hypothesis of likelihood hypothesis initially expressed by the Reverend Thomas Bayes. It can be viewed as a method for seeing how the likelihood that a hypothesis is valid and how is influenced by another bit of proof. It has been utilized as a part of a wide assortment of settings, running from sea life science to the advancement of "Bayesian" spam blockers for email frameworks. In this equation, T remains for a hypothesis or speculation that we are keen on testing, and E speaks to another bit of proof that appears to affirm or disconfirm the hypothesis. For any suggestion S, we will utilize P(S) to remain for our level of conviction, or "subjective likelihood," that S is valid. Specifically, P(T) speaks to our best gauge of the likelihood of the hypothesis we are thinking about, preceding thought of the new bit of proof. It is known as the earlier likelihood of T.

E. Data Partitioning

The pre-handled information is isolated into number of customers from focal director hub i.e. server according to the information ask for made by fancied number of customers. This parcelled information will be then prepared by individual customers to distinguish anomalies in view of connected calculation methodology.

F. Anomaly Detection

Anomalies will at first apply to disseminated customer and the distinguished exception recognizable proof innovation; program results will be coordinated into server processing stray last stage. To this end, key proposals abnormality identification calculation KNN technique is ABOD and INFLO.

IV. ALGORITHM USED

A. Unsupervised Learning

Unsupervised learning is the machine learning errand of surmising a capacity to portray concealed structure from unlabeled information. Unsupervised learning is firmly identified with the issue of thickness estimation in insights.

B. Introduction to KNN

K-NN is a kind of example based learning, or apathetic realizing where the capacity is just approximated locally and all calculation is conceded until arrangement. It is called languid in light of the fact that it doesn't have any preparation stage or insignificant preparing stage. K - Nearest Neighbor Algorithm

•For every preparation case <x,f(x)>, add the case to the rundown of preparing cases.

• Given an inquiry occasion xq to be ordered,

•Let x1, x2...xk indicate the k occurrences from preparing cases that are closest to xq

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 outlier's 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 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 AntiHubk and AntiHub2k, 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 Set

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, decisionsattribute (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.

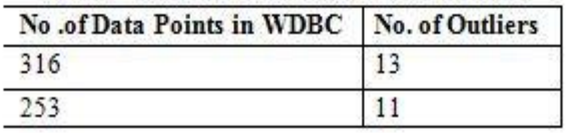
C. Pre-processing

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

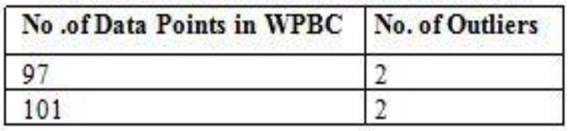
Distance based approach is applied in each cluster to find the data points those are closest to the centroid and they are pruned. Finally K-nearest neighbour is applied for remaining data points and outliers are detected based on top-n fashion distance approach. The details of outliers are summarized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

Table 1 Outlier Detection of WDBC Dataset

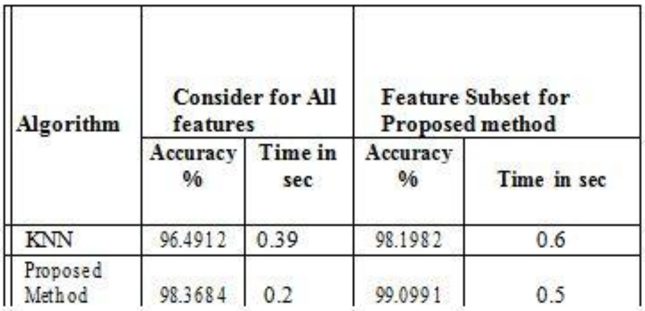


No .of Data Points in WDBC	No. of Outliers
316	13
253	11

Table 2 Outlier Detection of WPBC Dataset



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



Algorithm	Consider for All features		Feature Subset for Proposed method	
	Accuracy %	Time in sec	Accuracy %	Time in sec
KNN	96.4912	0.39	98.1982	0.6
Proposed Method	98.3684	0.2	99.0991	0.5

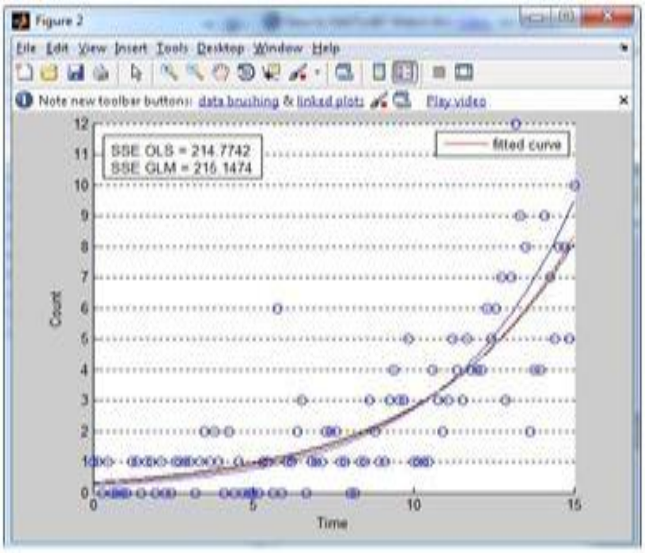
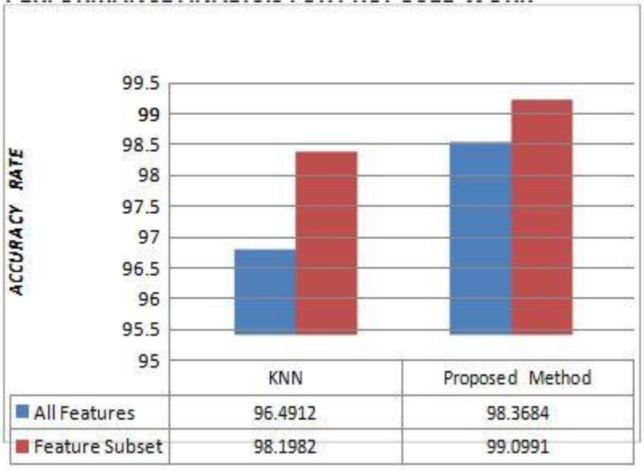


Figure 2
File Edit View Insert Tools Desktop Window Help
Note new toolbar buttons: data brushing & linked plots Play video
SSE OLS = 214.7742
SSE GLM = 216.1474
fitted curve
Count
Time
0
1
2
3
4
5
6
7
8
9
10
11
12
0
5
10
15

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 availableat 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.

PERFORMANCE ANALYSIS FOR PROPOSED WORK



	KNN	Proposed Method
🟦 All Features	96.4912	98.3684
🟥 Feature Subset	98.1982	99.0991

Fig 2. Performance analysis for KNN

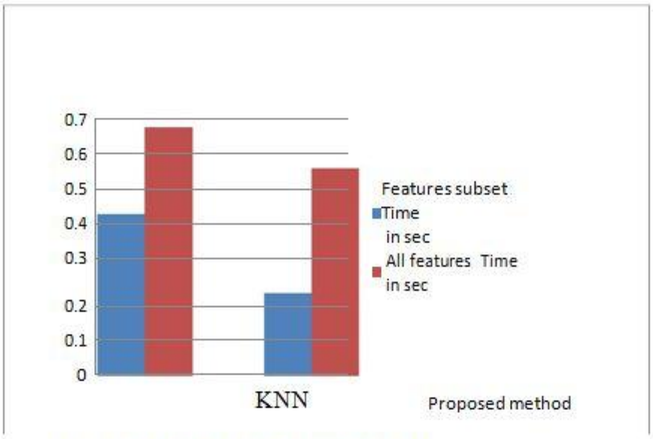


Fig 3: Time complexity of proposed method

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.

REFERENCES

[1] Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection by Milos Radovanovic, AlexandrosNanopoulos, and MirjanaIvanovi,

[2] IEEE Transactions On Knowledge And Data Engineering, Revised October 2014.

[3] An Efficient Anomaly Detection System Using Featured Histogram and Fuzzy Rule Mining by Ranjita Singh, Sreeja Nair., January 2014 ISSN: 2277 128X.

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

[5] Recursive Antihub2 Outlier Detection In High Dimensional Data by J.Michael Antony Sylvia, Dr.T.C.Rajakumar. Vol-2, Issue-8 PP. 1269-1274, 30August 2015.

[6] Unsupervised Distance-Based Outlier Detection Using Nearest Neighbours Algorithm on Distributed Approach: Survey by JayshreeS.Gosavi ,VinodS.Wadne, (An ISO 3297: 2007 Certified

Organization) IJIRCCE ,Vol. 2, Issue 12, December 2014.

[7] 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.

Syeda Khaja Momina Banu currently pursuing M.Tech in Department of Computer Science and Engineering at SR Engineering College, Warangal, Telangana, India. Research interest include in the area of Network Security, Data Mining etc.

P.Praveen is an assistant professor in Department of Computer Science and Engineeringat Sr Engineering College, Warangal He holds M.Tech. Degree in Computer Science and Engineering in 2010 from JNTU Hyderabad. His Pursuing PhD in data mining in Kakatiya University, His research activities are in the area of data mining, in particular clustering