ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Outlier Detection Using Hub, Antihub & Semisupervised approach for Distance based Method

Smita S. Patil, Prof. P. D. Chouksey

Department of Computer Engineering, BSCOER Narhe, Pune, India

Department of Computer Engineering, BSCOER Narhe, Pune, India

ABSTRACT: Outlier Detection aims to find patterns in data that do not conform to expected behavior. Their importance in data is due to the fact that they can translate in to actionable information in a wide variety of applications. Several methods are investigated for outlier detection corresponding to categorical data sets. In the previous method, unsupervised technique is used where no training is provided. So the result are less accurate. In proposed method, semi-supervised learning approach is used where half training is provided to normal data set. Using Entropy of each object and E-threshold value Outliers are detected. The benefit is to get results accurate

KEYWORDS*:* Outlier,K -NN, High dimensional dataset, Hubness, antihub.

I. INTRODUCTION

Outlier detection is the process of finding outlying pattern from a given dataset. Outlier detection became important subject in different knowledge domains. Data size is getting doubled every years there is a need to detect outliers in large datasets as early as possible. In high-dimensional data outlier detection presents various challenges because of curse of dimensionality. By examining again the notion of reverse nearest neighbours in the unsupervised outlier-detection context, high dimensionality can have a different impact. In high dimensions it was observed that the distribution of points in reverse-neighbour counts becomes skewed. Detection of outliers in data defined as finding patterns in data that do not conform to normal behaviour or data that do not conformed to expected behaviour, such a data are called as outliers, anomalies, exceptions. Anomaly and Outlier have similar meaning. The analysts have strong interest in outliers because they may represent critical and actionable information in various domains, such as intrusion detection, fraud detection, and medical and health diagnosis. An Outlier is an observation in data instances which is different from the others in dataset. There are many reasons due to outliers arise like poor data quality, malfunctioning of equipment, ex credit card fraud. Data Labels associated with data instances shows whether that instance belongs to normal data or anomalous. Based on the availability of labels for data instance, the anomaly detection techniques operate in one of the three modes are -

i) Supervised Anomaly Detection, techniques trained in supervised mode consider that the availability of labelled instances for normal as well as anomaly classes in a training dataset.

ii) Semi-supervised Anomaly Detection, techniques trained in supervised mode consider that the availability of labelled instances for normal, do not require labels for the anomaly class.

iii) Unsupervised Anomaly Detection, techniques that operate in unsupervised mode do not require training data.

There are various methods for outlier detection based on nearest neighbours, which consider that outliers appear far from their nearest neighbours. Such methods base on a distance or similarity measure to search the neighbours, with Euclidean distance. Many neighbour-based methods include defining the outlier score of a point as the distance to its kth nearest neighbour (k-NN method), some methods that determine the score of a point according to its relative density, since the distance to the kth nearest neighbour for a given data point can be viewed as an estimate of the inverse density around it.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

II. RELATED WORK

In paper [1],authors has proposed unsupervised outlier detection method is used. The hubness concept is also used which is the increase in dimensionality that can be used in a standard technique for outlier detection. The hubness is explored in as a new aspect of the increase of dimensionality and by examining the hubness, authors show that it is an essential property of data distributions in high-dimensional data. Clustering is a method which is used to group similar objects in groups. So clustering is an important tool for outlier detection, it is focused along with hubness in this paper. This paper proposes a technique where the concept of hubness, mainly antihub algorithm is used.

The technique proposed for identifying outliers will be applied initially at distributed clients and their results of detected outliers would be integrated on server machine at final stage computation of outliers. To do this, the outlier detection strategies proposed are KNN Algorithm with ABOD and INFLO Method.

The Distributed approach proposed with above Method based on anomaly detection techniques based on nearest neighbor .In this technique [2] assumption is that normal data instances occur in dense neighbour hoods, while outliers occur far from their nearest neighbors. In this proposed work using concepts of nearest neighbor based anomaly detection techniques:(1) use the distance of a data instance to its kth nearest neighbors to compute the outlier score.(2) compute the relative density of each data instance to compute its outlier score.

In this paper Author[3] assesses several distance-based outlier detection approaches and evaluates them. They begin by surveying and examining the design landscape of extant approaches, while identifying key design decisions of such approaches. Then implement an outlier detection framework and conduct a factorial design experiment to understand the pros and cons of various optimizations proposed by us as well as those proposed in the literature, both independently and in conjunction with one another, on a diverse set of real-life datasets. The outcome of this study is a family of state of the art distance-based outlier detection algorithms. The combination of optimization strategies enables significant efficiency gains. Their factorial design study highlights the important fact that no single optimization or combination of optimizations (factors) always dominates on all types of data .

Author presents that[11] there are widely uses of KDD one of them is detection of criminal string, so the determination of such things is more interesting as compared to the usual patterns.

Here many scenarios of outlier detections are used for fact, it is necessary to assign the degree of outlier which means the local outlier factor. This is called the degree of local outlier factor. Here lot of give many properties. With the help of actual datasets, author shows that lof is useful for finding outliers.

This system is used for finding the outliers. For any object there are two types of objects.

The objects have classes and they are tight.

This is based on the parameter of minimum points of nearest points.

Author shows that [12]Hubs are that points which are frequently occurs in k nearest neighbour list & ant hubs are that points which are infrequently occurs in k nearest neighbor list. Outlier detection using Antihub2 method is more accurate. Discrimination of outlier scores produced by Antihub2 acquires longer period of time with larger number of loops.

Therefore Recursive AntiHub 2 method was introduced. It improve the computational complexity of discriminating the outlier scores with reduced quantity of iterations to detect the more prominent outlier in high dimensional data. This paper describes the advantages and disadvantages of supervised unsupervised and semisupervised methods.

Partition based algorithm is used to mine top n points as outliers. In partition based algorithm it first partitions the input points using some clustering based algorithm and computes lower and upper bounds on the distance of a point from the k^{th} nearest neighbor in each partition.

III. PROPOSED ALGORITHM

A. *Proposed Approach*

In our proposed work, we are concerned with employing supervision of limited amount of label information to detect outliers with high accuracy.

The semi-supervised method depends on the availability of the training dataset for normal observation.

In our approach we will implement the entropy based antihub algorithm to detect the outliers using the semi-supervised approach.

- We first imports the training set for normal data items.

DOI: 10.15680/IJIRCCE.2016. 0406377

- After that in our approach the entropy of all objects is calculated for outlier detection process.
- After calculating the entropy of all objects, the threshold entropy is determined by averaging the entropy of all objects.
- This threshold entropy act as boundary to detect the outliers and normal data items.

B. Architecture

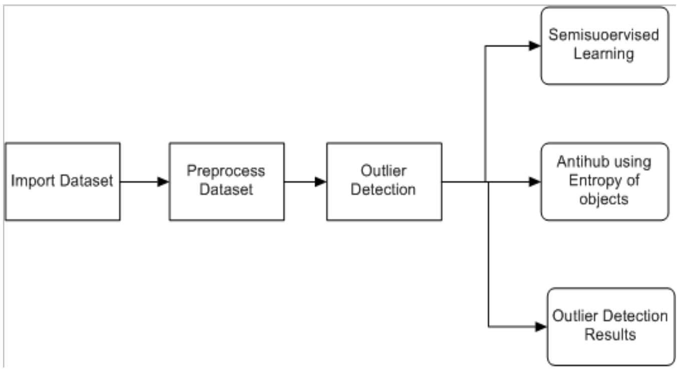

Fig. 1 Architecture Outlier Detection

We extend the existing antihub algorithm to the entropy based antihub algorithm. We will see the concept of entropy in following section.

C. Proposed Algorithm

Proposed algorithm works in following stages

1. Semi-supervised training:

As discussed in the literature we provide the training for the normal data item set N={x1,x2,….xm}. During the training the entropy of these normal data is calculated and the average entropy is calculated as explained above.

2. Antihub using entropy:

After training phase, the test data is given as input to the system. System first calculates the entropy of an object and compare with the threshold entropy. Accordingly the process is conducted till the all data objects finish with the processing.

3. Outlier detection:

In the last phase of detection the all items which belongs to the antihub are treated as outlier and given as out of the system.

D. Entropy Based Antihub algorithm:

Input: Training Dataset T={x_1, x_2...X_n}, Test dataset TD{X_1, $X_{2....}X_n$}.

Output: Outlier Set O= {O_1, O_2....O_n}

1. Import the training dataset *T*={X_1,X_2,…X_n}
2. For all objects in *T*
3. Calculate entropy of all training objects (normal)
4. End *for*
5. For all objects in *T*
6. $E_{Threshold} = \sum_{i=0}^{n} H\ Xi\ (Y)$
7. End for
8. Import test dataset *TD*={X_1,X_2,…X_m}
9. *For(i=1) to m*
10. Obtain the entropy of all test data objects
11. Compare the entropy of all objects with the $E_{Threshold}$
12. If (Hx(T_{Di}) > $E_{Threshold}$)
13. O=O U T D i
14. Else
15. N =N U TDi
16. End for

17. If termination condition is reached
18. Output outlier set *O and Normal Dataset N*

IV. MATHEMATICAL MODEL

Let S be the system such that,
S = S_1, I, O
Let S1 be the outlier detection system.

I represent the input to system.

O represent the output of the system.

S_1 = OS,AH ,T, TD, $E_{Threshold}$,H(x),X

I = T, TD, X ,$E_{Threshold}$,H(x)

O =H(x), $E_{Threshold}$, AH, OS

A) Obtain Entropy of objects:
Consider a set X containing n objects x_1; x_2; . . . ; x_n, each xi for 1<= i<= n being avector of categorical attributes [y_1; y_2; . . . ; y_m] , where m is the number of attributes, yj has a value domain determined by [y_1;j; y_2;j; . . . ; $y_{nj;j}$] (1 <=j <=m) and nj indicates the number of distinct values in attribute yj.

E is function to compute the entropy of the object.

H(x) = E(X);

Where,

H(x) is the entropy of the object.

X= Data object

B) Average Entropy:
Average is function which takes the entropy of all object and calculate the average entropy of the objects.

$E_{Threshold}$= Avg (H(x));

$E_{Threshold}$ = Average entropy of all objects.
C) Obtain antihub:
AntiHub is a function which takes training and test dataset as input and calculates the antihub set of objects.

AH = AntHub (T, TD,H(x), $E_{Threshold}$);

AH =Antihub objects set.

T =training data.

TD = Test Data.

D) Outlier Detection:
OD is a function which detects the outlier from the antihub set.

OS = OD (AH);
OS = Outlier set

V. EXPERIMENTAL RESULTS

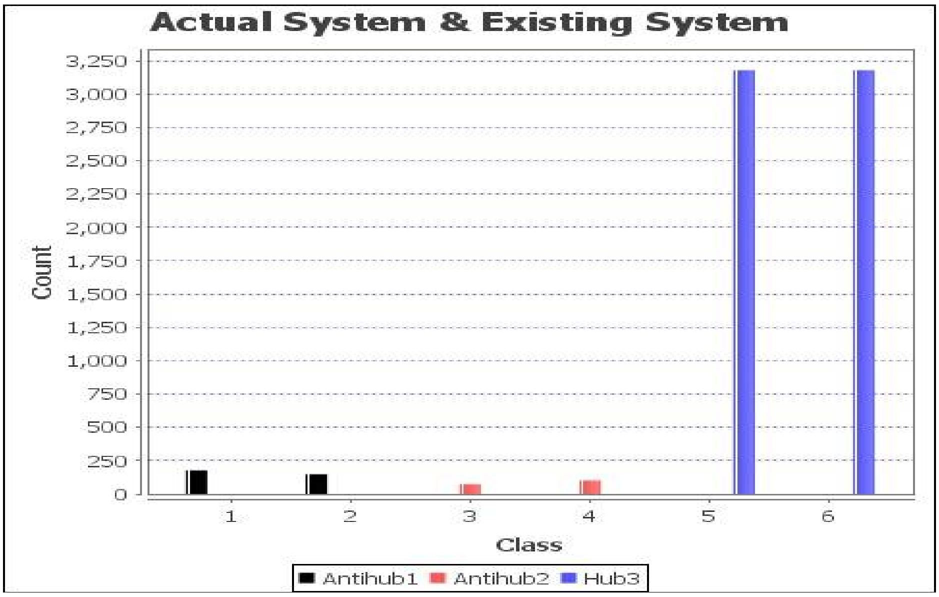

Fig .2 Comparison of Actual System and Existing System

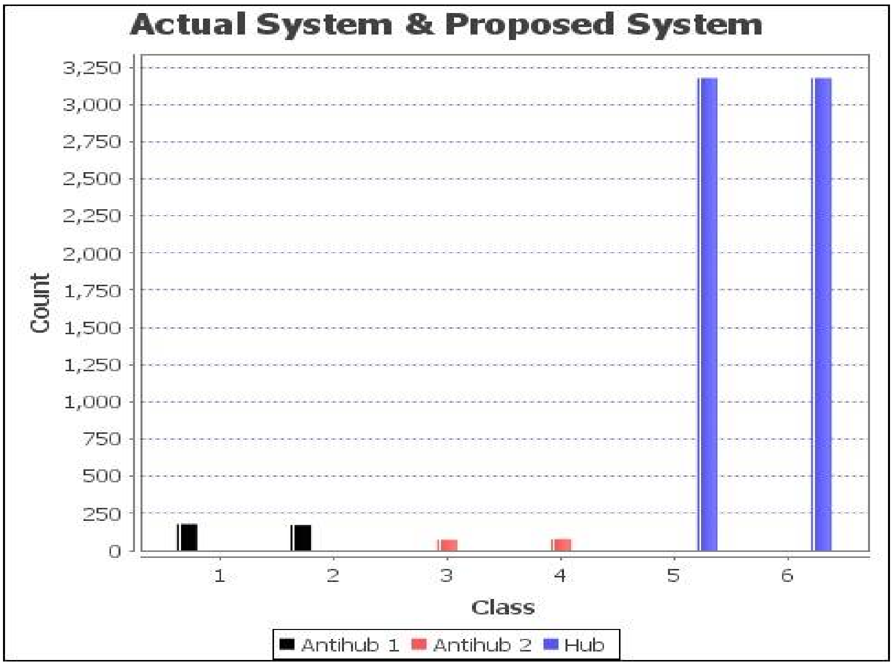

Fig. 3 Comparison of Actual and Proposed System

DOI: 10.15680/IJIRCCE.2016. 0406377

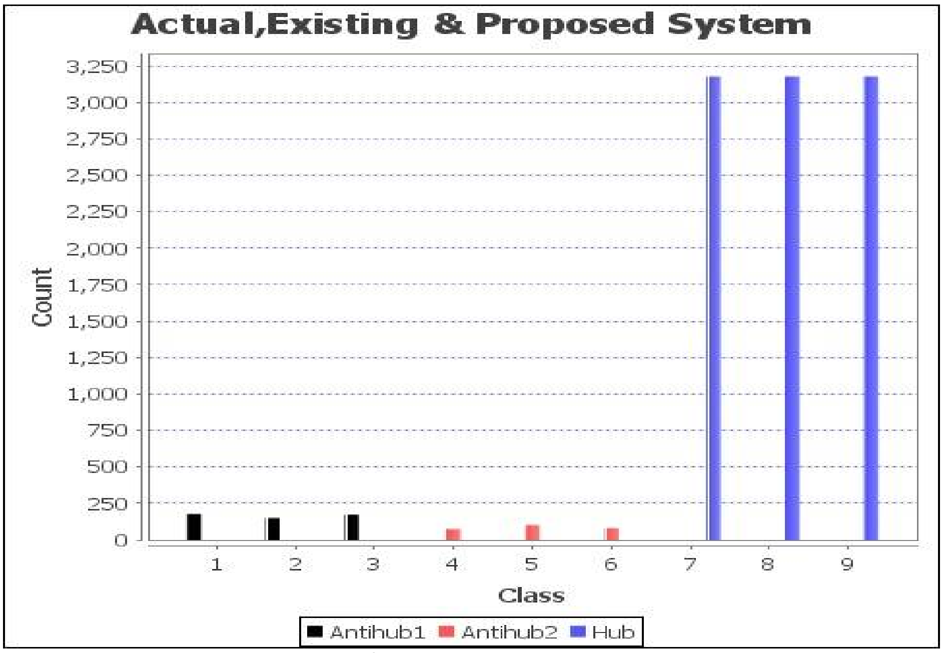

Fig. 4 Comparison of Actual, Existing and Proposed System

VI. CONCLUSION AND FUTURE WORK

Outlier detetion find out the unmatching patterns from data set. Different techniques use the different concepts such as hubness, antihub sets to detect the outliers.

In this paper Antihub Algorithm and semi-supervised methods are used. Outlier scores also plays an important role in outlier detection. The objective of this work is to achieve more accurate result. In proposed work by providing semisupervised training to data set, using entropy and Ethreshold value, we find out the more accurate result as compared to the existing system.

In future work, using supervised training implement outlier detection.

REFERENCES

1. Milos Radovanovi, Alexandros Nanopoulos and MirjanaI vanovi ,"Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection", IEEE Transactions On knowledge And Data Engineering. Transactions, Vol. 27, No. 5, May 2015.
2. Unsupervised Distance-Based OutlierDetection Using Nearest Neighbours Algorithm on Distributed Approach: Survey International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 12, December 2014.
3. Distance-Based Outlier Detection: Consolidation and Renewed Bearing Gustavo H. Orair Carlos H. C. Teixeira Department of Computer Science University dade Federal de Minas Gerais Wagner Meira Jr .Belo Horizonte, Brazil.Ye Wang Srinivasan ParthasarathyThe Ohio State University Columbus, USA.
4. Edwin, Raymond, "Distance based outliers: algorithms and applications", Springer- verlag, 2008.
5. Alexandros Nanopoulos,Yannis Theodoridis ,Yannis Manolopoulos ,"C2P: Clustering based on Closest Pairs", Proceedings of the 27th VLDB Conference, Roma, Italy, 2011.
6. H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlier detection in high-dimensional data," in Proc 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2008, pp. 444–452.
7. K. Zhang, M. Hutter, and H. Jin, "A new local distance-based outlier detection approach for scattered real-world data," in Proc 13th Pacific-Asia Conf on Knowledge Discovery and Data Mining (PAKDD), pp. 813–822. 2009.
8. J.Michael Antony Sylvia, Dr. T. C. Rajakumar Recursive antihub "outlier Detection in High Dimensional Data." Vol-2, Issue-8 PP. 1269-1274 global journal of research, 2015.
9. AmolGhoting, Srinivasan Parthasarathy, and Matthew Eric Otey, "Fast Mining of Distance-Based Outliers in High-Dimensional Datasets"Springer,2008.
10. "The Role of Hubness in Clustering High-Dimensional Data" Nenad Tomaˇ sev 1 , Miloˇ s Radovanovi´ c 2 , DunjaMladeni´ c 1 , and MirjanaIvanovi´ c 2.
11. Proc. ACM SIGMOD 2000 Int. Conf. On Management of Data, Dalles, TX, 2000 "LOF: Identifying Density-Based Local Outliers"

12. Recursive antihub2 outlier detection in high dimensional data Vol-2, Issue-8 PP. 1269-1274
13. International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163.
14. ”LOF: Identifying Density-Based Local Outliers” Proc. ACM SIGMOD 2000 Int. Conf. On Management of Data, Dalles, TX, 2000.
15. Ville Hautamäki, Ismo Kärkkäinen and Pasi Fränti University of Joensuu, Department of Computer Science Joensuu, Finland villeh, iak, franti @cs.joensuu.fi” Outlier Detection Using k-Nearest Neighbour Graph”