International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 06
May 2017

An Evaluation ofDetection of Outliersby Reverse Nearest Neighbors Method

A.Prashanth[1], K.A.M.Sushma[2], P.Srinivas Rao[3]
[1]M.Tech ,CSE, Jayamukhi Institute Of Technological Sciences,Warangal,India
[2]Assistant professor,CSE, Jayamukhi Institute Of Technological Sciences,Warangal,India
[3]Associate professor,CSE, Jayamukhi Institute Of Technological Sciences,Warangal,India

ABSTRACT: Outlier Detection in high dimensional information goes into a rising system in today's inspection in the region of data mining.Data that is different or erratic from normal data set are recognized by outlier detection. Unusual data records because of some data errors can be treated as outliers typically detecting outliers and investigating large data sets recognizes the problem such as medical problems, a structural defect, and investigational errors. This paper focuses the different methods for detection of anomalies. In order to handle the difficulties related to outlier detection because of uncertain data, outlier detection technique based on the AntiHub term is used.

KEYWORDS- Data stream, Data mining, outlier detection.

I. INTRODUCTION

Inspite of the huge measure of information being gathered innumerous exploratory and business applications, specifc occasionsof hobbies are still very uncommon. These uncommon occasions,regularly called exceptions or irregularities, are characterizedas occasions that happen occasionally (their recurrence rangesfrom 5% to under 0.01% relying upon the application). Discoveryof exceptions (uncommon occasions) has as of late picked up agreat deal of consideration in numerous areas, extending fromvideo observation and interruption identifcation to fake exchangesand coordinate advertising. For instance, in video observationapplications, video directions that speak to suspicious and/orunlawful exercises (e.g. recognizable proof of movement violatorsout and about, discovery of suspicious exercises in the region ofarticles) speak to just a little divide of all video directions. Thus,in the system interruption discovery area, the quantity of digitalassaults on the system is regularly a little portion of the aggregatesystem movement. In spite of the fact that exceptions (uncommonoccasions) are by defnition rare, in each of these illustrations,their significance is entirely high contrasted with differentoccasions, making their identifcation critical. Information excavatingstrategies produced for this issue depend on both managed andunsupervised learning. Regulated learning routines commonlyassemble an expectation model for uncommon occasions inlight of named information (the preparation set), and utilize itto arrange every occasion [7-8]. The signifcant disadvantagesof regulated information mining strategies include: (1) need tohave marked information, which can be to a great degree tediousfor genuine applications, and (2) powerlessness to identifynew sorts of uncommon occasions. Interestingly, unsupervisedlearning systems commonly don't require marked information anddistinguish exceptions as information focuses that are altogetherdifferent from the typical (greater part) information in light ofsome measure [9].

These strategies are ordinarily called exception/irregularity recognition procedures, and their prosperity relies onupon the decision of closeness measures, highlight choice andweighting, and so on. They have the upside of distinguishing newsorts of uncommon occasions as deviations from typical conduct,yet then again they experience the ill effects of a conceivablehigh rate of false positives, basically since already concealed (yetordinary) information can be likewise perceived as exceptions/oddities. Regularly, information in numerous uncommon occasionsapplications (e.g. system movement observing, video observation,web use logs) arrives persistently at a tremendous pace in thisway representing a noteworthy test to break down it [9]. In suchcases, it is imperative to settle on choices rapidly and precisely.On the off chance that there is a sudden or startling change in thecurrent conduct, it is fundamental to distinguish this change asquickly as time permits. Expect, for instance, there is a PC in theneighborhood that uses just set number of administrations (e.g.,Web activity, telnet, ftp) through comparing ports. Every one ofthese administrations relate to specifc sorts of conduct in systemactivity information. On the off chance that the PC all of a suddenbegins to use another administration (e.g., ssh), this will positivelyresemble another sort of conduct in system activity information.Henceforth, it will be attractive to identify such conduct when itshows up particularly since it might frequently relate to unlawful ornosy occasions. Indeed, even for the situation when this particularchange in conduct is a bit much nosy or suspicious, it is imperativefor a security examiner to comprehend the system activity and toredesign the idea of the typical conduct. Further, on-line recognitionof irregular conduct and occasions additionally assumes a huge part.

II. RELATED WORKS

In recent time it is observed that the distribution of points'reverse-neighbor counts becomes skewed in highdimensions, which results in the phenomenon of hubness[1]. Authors also discussed that the how antihub appear veryinfrequently in k-NN lists of other points. They alsodiscussed the connection between the antihubs and existingunsupervised outlier detection [1].

Here provided the role of reverse nearest neighbor counts inproblems concerning unsupervised outlier detection. Themain focus is given on the unsupervised outlier-detectionmethods and the hubness phenomenon in highdimensionality.Extended the work of antihubs to the large values of k andexplored the relation between the hubness and data sparsitybased on the unsupervised outlier detection. The extensionof anthubs improves the discrimination in the outlier scores.The existence of hubs and antihubs in high-dimensional datais relevant to machine-learning techniques from various
families: supervised, semi-supervised, as well asunsupervised. Here only unsupervised method is used, itdoes not give accurate result as compared to the othermethods.

H.-P. Kriegel, M. Schubert, and A. Zimek[4] has proposedangle based outlier detection (ABOD). Outlier detection inhigh-dimensional data uses the variances of a measure overangles between the different vectors of data objects. InABOD technique, used the properties of the variances toactually take advantage of high dimensionality and found tobe less sensitive to the increasing dimensionality of a dataset. This technique is less efficient than the classic distancebased methods. The disadvantage is only angle based is usednot the classic distance-based methods.

The LOF compare the local density of instances with thedensities of its neighborhood instances. After that it assignsthe outlier scores to given data objects. If LOF score equal toratio of average local density of k nearest neighbor ofinstance and local density of data instance itself then datainstance is considered to be normal and not as an outlier.Local density of instances is computed by finding radius ofsmall hyper sphere centered at the data instance after thatdividing volume of k [5], i.e. k nearest neighbor and volumeof hyper sphere. In this assign a degree to each object tobeing an outlier known as local outlier factor [5].

Objects are isolated depending on the surroundingneighborhood, instances lying in dense region are normalobjects [5], if their local density is similar to their neighborsand objects are outlier if there local density lower than itsnearest neighbor [5]. It is a critical or lengthy process ascompared to the distance based methods.The antihub2 method is unsupervised outlier detectionmethod used for anomaly detection in high dimensionaldataset. Anomaly detection in high dimensional data exhibits that as dimensionality increases there exists hubs and antihubs [6]. Hubs are the point that frequently occurs in knearest neighbors. Antihubs are the point that occursinfrequently in nearest neighbors list. In this paper authorshave refined the antihub method to refine the outlier scoresof a point produced by the antihub method by consideringthe nk scores of the neighbors of the data point.Discrimination of outlier scores produced by Antihub2acquires longer period of time with larger number ofiterations [6]. Because of this recursive AntiHub2 methodwas introduced to improve the computational complexity

ofdiscriminating the outlier scores using less number of
iterations to detect accurate outliers in high dimensional data[6].

III. PROPOSEDAPPROACH

An outlier detector identifies statistics items that don'tconfirm an anticipated sample or different records items in factsstream. The detection of outlier enables in discovery ofsurprising information in records circulate. System version foroutlier detection technique is proven in fig.1. It suggests fundamentalphases for the procedure of identity of outliers and itsadditives. These are explained as follows:

A. Database series and pre-processing

For the input to the machine, datasets are collected fromthe UCI depository sets. In the device module widespreaddatabases are used and the databases are superviseddatabases. Supervised dataset incorporates class labelsin line with the data type. Class labels are assigned at thebasis of different attributes of the dataset. That mannerdatasets already characterized into different lessons basedat the dataype. As in keeping with the form of dataset elegance labelsaccording to magnificence kind are present. As part of preprocessing, the missing values inside the databases are filledwith the value zero or as null.

The first pre-processing technique used is data cleansing.Data cleaning is executed to discover the lacking values inthe data file. Also data cleansing is carried out to find outinconsistent information. The dataset used for outlier detectioncontained missing values and inconsistent information, for thesestatistics cleaning is achieved. As a few attribute does now notcontain any cost, for them as part of statistics cleansing,values are inserted like 0 and null.

The every other part of pre-processing is facts transformation.It consists of conversion of data values aptitude in the shape ofdataset into the data layout for destination device requiringrecords from source device consisting facts. Datatransformation carries information mapping, which maps informationfrom supply to vacation spot system. As the data is to be hadon UCI depository, the data set is mapped into thehotspot machine using data transformation technique.Hence, the datasets are equipped for the following step.

B. AntiHub1 Method

The approach AntiHub1 is primarily based on the ODIN method.ODIN method makes use of normal scores of outliers through analyzingthe closest neighbor be counted for the unique factor. Theordered facts set D is given as an enter. Number ofacquaintances and distance from unique point is supplied asinput. Temporary variable AntiHub rankings are used forcomparing cutting-edge discrimination score and raw outlierscore. For each enter ‘Sn’ is computed w.r.t. Dist and statisticsset (d/n).

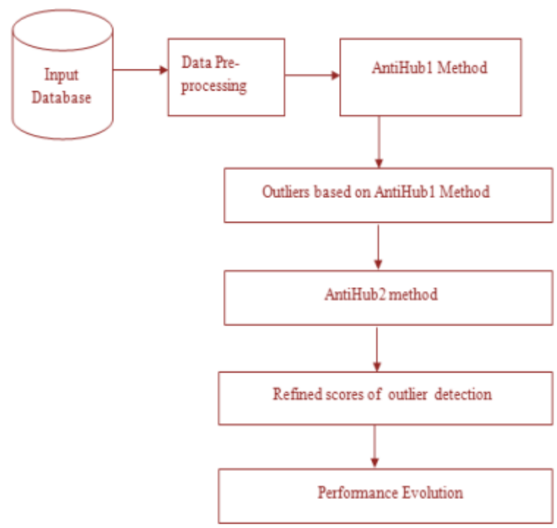

Fig. 1 Flowchart for outlier detection

The steps for AntiHub1 algorithm [13] are as follows:

1. For each point ‘i’ belongs to 1 to n,
2. Nx(x) = D\xi
3. s = f(t)

The steps for algorithm are described:

- Distance measured in the form of variable dist and ordered dataset as D is given as input for each point.
- For each point, distance from nearest neighbour is computed as Nx(x) with respect to data set and dist variable.

- The score of outlier for point ‘x’ is computed using monotone function and stored in the vector‘s’ as anoutput.

The Algorithm 1 AntiHub produces output in the formof vector as the outlier score of point x from data set D.Outliers are collected as output in the form of percentageoutlier values.

C. AntiHub2 Method

As normalization is not applied on the outlier score, theoutliers collected are not refined. To tackle weaknesses ofAntiHub1, a simple heuristic method AntiHub2 is appliedto the outlier score produced by the AntiHub1 method.AntiHub2 refines outlier scores produced by the AntiHub1method by considering the Nk scores of the neighbors of x.

For improvement in discrimination of scores that AntiHub2introduces compared to AntiHub, for each point x,AntiHub2 proportionally adds the sum of Nk scores of the knearest neighbors of x.

Input for AntiHub2 method contains measured distancefor each point, ordered data set, number of neighbours andratio of outliers for maximizing discrimination. Temporaryvariables are used for getting outlier score which are currentdiscrimination score, current raw outlier scores, antihubscore and sums of nearest neighbours’ scores. AntiHub2method is implemented for the refinement of the scores ofoutliers produced by AntiHub1 method.

The steps for AntiHub2 algorithm [13] are as follows:
1. For each point ‘i’ belongs to 1 to n,
2. AntiHub score a = AntiHub(D,k)
3. anni = summation of indices of k nearest neighbour
4. disc = 0
5. For each ‘i’ from 1 to n
6. ct = ai + ann
7. cdisc = discscore(ct,p)
8. If cdisc> disc
9. t=ct
10. s = f (t)
The steps for algorithm are described:

- Distance measured in the form of variable dist, ordered dataset as D and number of neighbours is given as input for each point.
- Ratio of outliers to maximize discrimination and search parameter for each step is initialized applicable to every point.
- Array for ordered dataset and number of neighbours is initialised as AntiHub(Dike).
- For each point sum of nearest neighbours’AntiHub scores as temporary variable ‘ann’ is calculated and stored. For each step from 0 to 1, loop is carried out, for each point.
- Raw outlier score ‘ct’ is calculated using proportion and point sum of nearest neighbours’AntiHub scores.
- Then the value raw outlier score ‘ct’ and ratio of outliers to maximize discrimination i.e. ‘p’ is transferred to temporary variable ‘disc’.
- Comparison for temporary variables ‘cdisc’ and ‘disc’ is carried out, and if cdisc>disc then current raw outlier score‘t’ is set as raw outlier score ‘ct’.

The score of outlier for point ‘x’ is computed usingmonotone function and stored in the vector‘s’ asan output.The second method considers the scores of neighboursfor point x. And then adds sum of scores of nearestneighbour. To find aggregate of neighbours’ scoressummation is calculated. The discrimination scorescompared using discScore parameter is provided to outputvector as an outlier score.

IV. CONCLUSION

In this paper, comparativeinfluences of basic parametersdependences on outlier detection are deliberated. Influencefor values and datasets and their inter-relationship are alsorecognized. To conclude, this paper aids to comprehend thefact about complete analysis of nature of task is to bemodelled prior to the algorithmic choice for outlierdetection.

REFERENCES

[1] Milos Radovanovi, AlexandrosNanopoulos andMirjanaIvanovi,”Reverse Nearest Neighbors

inUnsupervised Distance-Based Outlier Detection", IEEETransactions On knowledge And Data Engineering.Transactions, Vol. 27, No. 5, May 2015.

[2] Edwin, Raymond, "Distance based outliers: algorithmsand applications", Springer- verlag, 2008.

[3] AlexandrosNanopoulos, YannisTheodoridis, YannisManolopoulos, "C2P: Clustering based on Closest Pairs",Proceedings of the 27th VLDB Conference, Roma, Italy,2011.

[4] H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-basedoutlier detection in high-dimensional data, " in Proc 14thACM SIGKDD Int. Conf. Knowl. Discovery DataMining, 2008, pp. 444–452.

[5] K. Zhang, M. Hutter, and H. Jin, "A new local distancebased outlier detection approach for scattered real-worlddata, " in Proc 13th Pacific-Asia Conf on KnowledgeDiscovery and Data Mining (PAKDD), pp. 813–822.2009.

[6] J.Michael Antony Sylvia, Dr. T. C. Rajakumar Recursiveantihub "outlier Detection in High Dimensional Data."Vol-2, Issue-8 PP. 1269-1274 global journal of research,2015.012.

[7] W. Lee, S. Stolfo,"Data mining approaches for intrusiondetection", Proc. of the 7th USENIX security symposium,1998.

[8] E. Bloedorn, et al.,"Data Mining for Network IntrusionDetection: How to Get Started", MITRE Technical Report,August 2001.

[9] A.K. Jones, R.S. Sielken,"Computer System IntrusionDetection: A Survey. Technical report", University of VirginiaComputer Science Department, 1999.

[10] M. Masud, Q. Chen, L. Khan,J. Gao and J. Han "Classification andAdaptive Novel Class Detection of Feature Evolving Data Streams",IEEE Trans. Knowl. Data Eng., vol. 25, no. 7, July 2013.

[11] S. Ahmed Shaikh and H. Kitagawa "Continuous Outlier Detection onUncertain Data Streams", IEEE Ninth International Conference onIntelligent Sensors, Sensor Networks and Information Processing(ISSNIP) Symposium on Information Processing Singapore, 21–24April 2014 .

[12] Bo Liu, Yanshan Xiao, Philip S. Yu, ZhifengHao, and LongbingCao, "An Efficient Approach for Outlier Detection with ImperfectData Labels", IEEE Trans. Knowl. Data Eng., vol. 26, no. 7, July2014.

[13] Milos Radovanovic, AlexandrosNanopoulos, and MirjanaIvanovi"Reverse Nearest Neighbors in Unsupervised Distance-BasedOutlier Detection", IEEE Transactions On Knowledge And DataEngineering, Vol. 27, No. 5, May 2015.

[14] UCI Machine Learning Repository [Online]. Available:http://archive.ics.uci.edu/ml/datasets.html.

[15] Bo Liu, Yanshan Xiao, Philip S. Yu, ZhifengHao, and LongbingCao, "An Efficient Approach for Outlier Detection with ImperfectData Labels", IEEE Trans. Knowl. Data Eng., vol. 26, no. 7, July2014.