Volume 2, Issue 12, December 2016 ISSN NO: 2454 - 7514

North Asian International Research Journal Consortium

North Asian International Research Journal

Of

Science, Engineering and Information Technology

IRJIF IMPACT FACTOR: 3.821

NAIRJC JOURNAL PUBLICATION

North Asian International Research Journal Consortium

Welcome to NAIRJC

ISSN NO: 2454 -7514

North Asian International Research Journal of Science, Engineering & Information Technology is a research journal, published monthly in English, Hindi. All research papers submitted to the journal will be double-blind peer reviewed referred by members of the editorial board. Readers will include investigator in Universities, Research Institutes Government and Industry with research interest in the general subjects

Editorial Board

Address: -North Asian International Research Journal Consortium (NAIRJC) 221 Gangoo, Pulwama, Jammu and Kashmir, India - 192301, Cell: 09086405302, 09906662570, Ph. No: 01933-212815, Email: nairjc5@gmail.com, nairjc@nairjc.com, info@nairjc.com Website: www.nairjc.com

IRJIF IMPACT FACTOR: 3.821

PROPERTIES OF OUTLIERS IN HIGH DIMENSIONAL DATA WITH HUB AWARENESS

A SOMAKALA*

*M.Tech Student, Department of CSE, JNTUA College of Engineering, Ananthapuramu Andhra Pradesh, India

P DILEEP KUMAR REDDY**

**Lecturer, Department of CSE, JNTUA College of Engineering, Ananthapuramu Andhra Pradesh, India

Abstract— The outlier detection, for its importance for its accuracy is increasing day by day. Detecting outlier in high dimensional data has lot of challenges and properties. Distance concentration is one of the challenges faced, where distances between data points will have no meaning. Hubness is a phenomenon which is observed in the high dimensional data. These hubness awareness can be said as informativeness, where information of each data point is known in the whole data set. A considerable amount of research is done in this field. Few salient features like distance based methods will be discussed in the following paper. Reverse nearest neighbors' uses in emergence of hubs and their properties will be discussed.

Keywords—Informativeness, hubness, high dimensional data, outlier, reverse nearest neighbors

I. INTRODUCTION

Detection of outliers plays a key role in pattern recognition. Any point is a part which is different from whole data set is an outlier. To get a part which is apart from other parts of any dataset we follow few techniques in order to get the data sets which are apart from other points.

Example: Let us suppose we have a dataset of National Basket-Ball Association players list from 1978-2003 consisting of all star players. Among them which players has an edge, they can be considered as outliers. Outlier detection methods helps in getting these data points as output. To get outliers in above high dimensional data hubness approach can be used.

Outlier detection in high dimensional data [1] [6] with distance based methods faces a lot of contests, one among them is distance concentration [7]. This

is feature where distances between high dimensional data becomes useless as dimensionality increases. For any datasets standard deviation and mean are the parameters considered to determine the accuracy of a method.

In high dimensional data [2] [3] when dimensionality increases it is facing distance concentration effect. This can be further explained as normally in any dataset standard deviation among the points remains unchanged, where-as the mean increases as number of data points increases, but due to the effect of distance concentration, standard deviation and mean are increasing and effecting the outlier detection efficiency.

Few of the obstacles faced in the high dimensional data are overcame with the properties of hubs. Hubness [4][6] is a phenomenon which is commonly observed with the high dimensional data. Mathematically it can be represented as all the data points which has a standard deviation is less than one and equal to one are considered as hubs and all other points whose standard deviation is greater than one are antihubs. The number of points that belongs to a hub can be considered as neighborhood score Nk. All the points which are regular are considered and represented as Nk.

Unsupervised, Semi supervised and supervised data learnings are basic machine learning techniques. Unsupervised learning costs less in practical uses, but semi supervised and supervised learning has its own usefulness, but may cost more in both time and structure. Every dataset has its unique features if we are able to get that informativeness then overall quality of detection of outliers can be increased.

Getting neighborhood score is one such method used to get the informativeness of a data point in a data set. Reverse nearest neighbor algorithm [1] is where we calculate, let x be a point k be the number of occurrences then Nk is the count of how many times x is among k nearest neighbors for every point in a given dataset. Nk is the neighborhood score. For all the points in the dataset the data points with highest Nk count belong to a hub. In any given dataset few points has more information than other points, by virtue these points occur less frequently than other points. These data points will be the point of interest in this paper.

The input data for these systems are collected from any standard dataset. This Data is reprocessed such data before forwarding for future steps. Data mining techniques such as data integration, cleaning, transformation will be used to reprocess large dataset contents and to generate required clean data. From there outliers are calculated.

II. RELATED WORK

Nenad Tomaˇsev et.al in [] worked on getting the informativeness of a data point in hub awareness approach. They calculated outliers in a probabilistic approach [5] to form hubs. Here labels are given to hubs. It just says that a point if belonging to a hub has a label. But the output which are outliers of a dataset do not have any labels. Only hubs are formed with supervised or semi supervised learning.

MiloˇsRadovanovi´c et al [7] in this paper explained about how hubs ae emerged. The need for the emergence of hubs in the high dimensional data are discussed in detail. They have discussed how hubs and anti-hubs with the help of reverse nearest neighbors are formed and outliers are determined.

Influence outlier detection [3] proposed on a symmetric neighbourhood relationship measure considers both neighbours and reverse neighbours of an object when estimating its density distribution .Uses when outliers are in the location where the density distributions in the neighbourhood are significantly different. They have mainly concentrated on the densities of the locality. These methods are very useful in determining, on a world map.

MiloˇsRadovanovi´c et al [1] in their work they detected outliers for unsupervised data with reverse nearest neighbors using ODIN method. They have proposed a unifying view of the role of reverse nearest neighbor counts in unsupervised outlier detection of how unsupervised outlier detection methods are affected with the higher dimensionality. These parameters are extended for large values of k. Relationship between hubness and sparsity are explored. Mainly they cleared about how properties of data and type of outliers are interpreted. This helped in increase in reach of reverse nearest neighbors. In this paper we are improvising by doing classification with semi supervised data.

III. PROPERTIES OF OUTLIERS IN HIGH DIMENSIONAL DATA WITH INFORMATIVENESS

Various studies are done in the field of outlier detection, deducting various properties of outliers for hubness and sparsity, skewness properties are studied. Informativess represents a self-learning approach to learn from existing data in a dataset. For supervised and semi supervised machine learnings, above properties are exploited. These properties are discussed below.

A. DISTANCE BASED AND ANGLE BASED METHODS

Distance based methods are very common methods to get outliers of any datasets. Nearest Neighbors (NN), k-Nearest Neighbors (kNN) are few methods to detect outliers. Usually in these methods Euclidean distance is calculates. Depending

upon all the Euclidean distances calculated, classes are assigned to a point an outliers are determined.

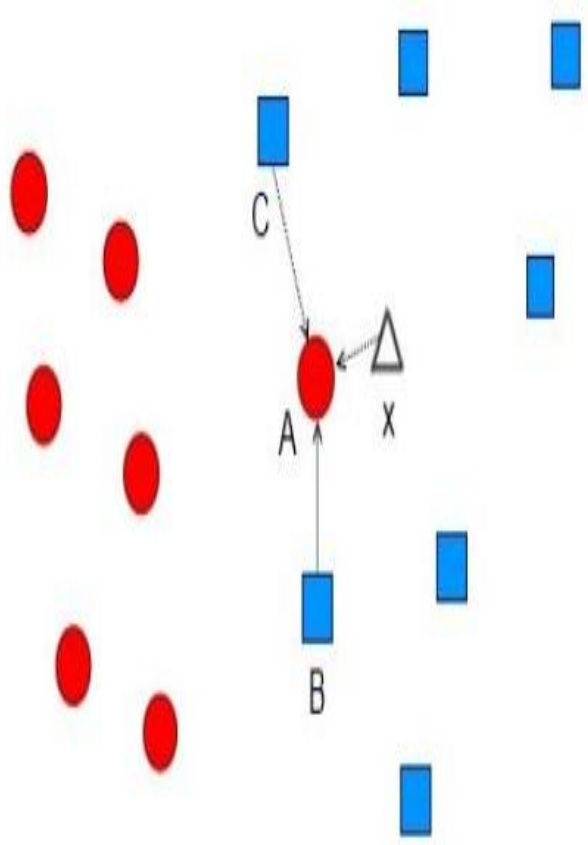

Figure 1: Euclidean distances are calculated from point A

Among these Reverse Nearest Neighbors [1] (RNN) which has properties of getting to how many points in a data set our point of interest is near can be calculated. With the help of this usually neighborhood score is calculated, which is important to get hubs with high dimensional data. If the neighborhood score is less than a threshold value in a dataset, then these points belongs to a hub and all the other points belongs to an antihub. Few data points that do not belong to any hub will be outliers.

In some methods like local outlier factor [8][12] (LOF), outliers are determined according to the density of the data set. In any given dataset, if all points are plotted, we can observe that density may vary from a grid to grid. If density is low then the points in that grid will be outliers.

Angle based approaches [9] are also observed in previous studies. One among them is Angle Based Outlier Detection. With this depending upon the differences and variations in the angle between the vectors, outliers are detected. Whatever may these distance based methods are, there will always be shared neighbor problem when determining the outliers.

B. HUBNESS AND SPARSITY

Hubness helps in getting the information which is similar to the others, i.e., when a set of songs is considered, hubs with hubs we see which songs are similar, other songs which do not belong to any hub, which can be said as the songs that do not confine to the pattern of songs present in the list will be outliers.

One of the major properties that is observed in hubs is few data points have more information than others, there by being less frequently occurred events, which are antihubs [1]. It can be explained as the points which occurs less frequently has more information than the points which occur frequently.

This feature is exploited in getting the informtiveness [10] [11] of a data point in a data set. It is further discussed in the following sections in detail.

Previously we have discussed a few properties of hubs and anti-hubs depending upon its standard deviation and mean in high dimensional data. All datasets may not have equal sparsity. Some datasets may have very sparse data points and some may not.

Sparsity [1] effects the determination of outliers depending upon the values of k. If k<<n then, we get a situation of strong hubness which results in making N_k, neighborhood score as zero making every data point a potential outlier. But these properties of sparsity are not effecting the hubness from the data, hence this can be eliminated by simply taking k values which are moderate. Skewness [1] is a values of percentage of outliers in overall data set. This does not affect any properties of a hub.

C. INFORMATIVENESS

Any event which occurs on a regular basis is just an information. Like sun rising in the east. But Mars coming nearer to earth occurs rarely making it a surprised event, containing the information which we need. There by making events which do not occur on a regular basis contains more information than the events which do occur regularly.

Informativess is a mathematical representation of mod of logarithm of probability of a data point occurring among in a data set. It is always a positive value and it can have a value of zero. But getting zero information do not have any meaning.

Need for informativeness lies in making decisions based upon the data which is already available in the given dataset. When the concept of hubs and anti-hubs are used not all points belongs to either of them. Few may fall in between them. For these to be placed correctly informativeness approach can be used to handle them. Probabilistic approaches can be followed to get the accurate results.

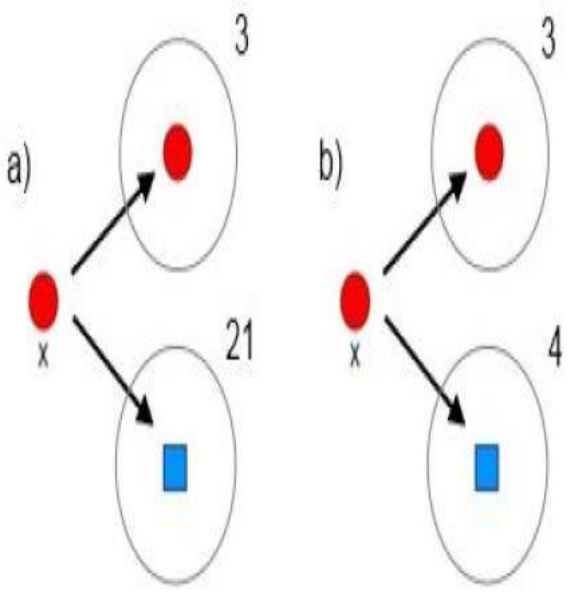

Figure 2: Explaining how any data point is assigned to a hub depending upon the informativeness

In the above figure 2, informativeness of x is calculated with respect to the data already available and in the hubs it depicted where there is a lot of difference in the informativeness and not much difference.

IV. CONCLUSIONS AND FUTURE WORK

Various methods to detect outliers in high dimensional data with hubness aware in formative-ness is discussed in the paper. These properties can be used at different fields to get the results required. These properties and techniques can be further studied and extended for different values of n, k and dimensions.

Outlier detection has very much prominence in various fields like fraud detection, medical diagnosis, hacking, face detection and various other fields. In present days importance is given to accuracy in getting results. Hence methods can be developed to increase the accuracy of detecting outliers in high dimensional data. Distance based methods can be further extended. Studies can be further explored in determining the properties of hubs and to exploit these features to get the outliers in much dramatic fashion. Relations between sparseness of data points in a data set can be drawn further.

REFERENCES

[1] Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovi" Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection" in IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 5, May 2015

[2] S. Ramaswamy et al. "Efficient algorithms for mining outliers from large data sets."SIGMOD Record, 29(2):427–438, 2000.

[3] M. Radovanović et al. Hubs in space: "Popular nearest neighbors in high-dimensional data". Journal of Machine Learning Research, 11:2487–2531, 2010.

[4] N. Toma˘sev, M. Radovanovi´c, D. Mladeni´c, and M. Ivanovi´c, "Hubness-based fuzzymeasures for high dimensional k-nearest neighbor classification," in Machine Learningand Data Mining in Pattern Recognition, MLDM conference, 2011.

[5] "A probabilistic approach to nearest neighbor classification: Naive hubnessbayesian k-nearest neighbor," in Proceeding of the CIKM conference, 2011.

[6] M. Radovanovi´c, A. Nanopoulos, and M. Ivanovi´c, "Nearest neighbors in highdimensionaldata: The emergence and influence of hubs," in Proc. 26th Int. Conf.on Machine Learning (ICML), 2009, pp. 865–872.

[7] A.C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distancemetrics in high dimensional spaces," in Proc. 8th Int. Conf. on Database Theory (ICDT), 2001, pp. 420–434.

[8] H.-P. Kriegel, P. Kr€oger, E. Schubert, and A. Zimek, "LoOP: Localoutlier probabilities," in Proc 18th ACM Conf. Inform. Knowl. Manage., 2009, pp. 1649–1652.

[9] H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlierdetection in high-dimensional data," in Proc 14th ACM SIGKDDInt. Conf. Knowl. Discovery Data Mining, 2008, pp. 444–452.

[10] N. Tomasev and D. Mladenic, "Nearest neighbor voting in highdimensional data: Learning from past occurrences," Comput. Sci. Inform.Syst., vol. 9, no. 2, pp. 691–712, 2012.

[11] N. Tomasev, M. Radovanovic, D. Mladenic, and M. Ivanovic,"The role of hubness in clustering high-dimensional data," IEEETrans. Knowl. Data Eng., vol. 26, no. 3, pp. 739–751, Mar. 2014.

[12] K. Zhang, M. Hutter, and H. Jin, "A new local distance-based outlierdetection approach for scattered real-world data," in Proc 13thPacific-Asia Conf. Knowl. Discovery Data Mining, 2009, pp. 813–822.

Publish Research Article

Dear Sir/Mam,

We invite unpublished Research Paper,Summary of Research Project,Theses,Books and Book Review for publication.

Address:- North Asian International Research Journal Consortium (NAIRJC)
221, Gangoo Pulwama - 192301
Jammu & Kashmir, India
Cell: 09086405302, 09906662570,
Ph No: 01933212815
Email:- nairjc5@gmail.com, nairjc@nairjc.com , info@nairjc.com
Website: www.nairjc.com