Volume 6, Issue 10, October 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Supervised Distance-Based Outlier Detection

[1]B. Sivaiah, [2]Dr. A. P. Siva Kumar
[1]M.Tech Student, [2]Assistant Professor
[1, 2] Department of CSE, JNTUA College of Engineering, Anantapuramu, Andhra Pradesh, India

Abstract: ***Outlier detection in cluster data issue a challenge from the analyzing and organizing data in high-dimensional spaces. In general view is similarity concentration, the identification of similarity distances in clustering data to become difficult. The identification of outliers by using distances based methods produce more outlier scores. we provided in this paper a supervised similarity method for identifying the outliers scores in clustering high-dimensional data, how supervised method is used for the organizing dataset into different clusters. It shows the detected outliers in the graph. The experimental results shows that our paper effectively produce more outliers in the clustering high dimension data.***

Index terms- outlier, data mining, high-dimensional data, cluster, k-nn algorithm.

I. INTRODUCTION

Outlier finding is the assignment of identifying different patterns, There behaviour is different compare to other objects. In spite of absence of as mathematical representation of outlier and their identification is a generally connected hone. The enthusiasm for exceptions is strong since they may constitute basic and noteworthy data in different spaces. Finding of outliers in different domains like fraud detection, intrusion and medical diagnosis.

The finding of outliers to be classified as supervise, unsupervised and semisupervised ,the presence of labels for outliers and occurrences of anomalies. Among the above classes supervised techniques are used for the better results for finding the outliers in the different domains, because of the other categories requires exact and agent identity that regularly restrictively costly to acquire. Supervised strategies incorporate the separation based on a measure of separation (or) comparability keeping in mind the end goal to identifying outliers.

Outliers are different behaviour compare to the remaining objects in the high dimensional spaces. The below figure shows the how outliers are identifying in the clustering high dimensional spaces.

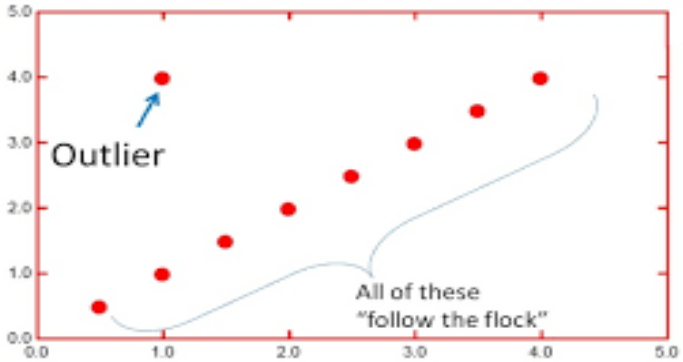

Fig.1.outlier that have a different behaviour with other points.

In the above diagram all points in the graph have same diagonal except point (1, 4).The point (1,4) is the outlier. All point in the diagonal is fallow the flock, this are representing as one group. The group of similar objects called cluster.

II. RELATED WORK

Nearest neighbour counts is a old technique for finding outliers of information, how ever this techniques representing the fundamental. They are do not effectively found important outlier scores. Recently they are identified that reverse neighbour counts are influenced by expanded dimensional information and will meet certain information and their reconsideration for the founding of outlier score they are introduced the "Outlier detection using K nearest neighbour graph". In reverse nearest-neighbour counts the hubness was observed with the k-occurrences. K occurrences defined has number of times point(x) appears in the k nearest neighbours of all remaining points in the dataset.

The ABOD(angle based outlier detection) system distinguishes outliers in high dimensional information by variances the different degrees of measure the outliers in the datasets. With the distance measured to find the closed points this expects the outliers in the datasets.

III. PROPOSED WORK

The detection of outlier in clustering data using the supervised distance based methods. In supervised mode we can detect outliers in clustered data as well as unstructured data. In the proposed system architecture the collection of large data se is given as the input to the proposed system. The information for the data is collected from different domains. We have to perform the pre-process operation before going to the next stage of the operation. Pre-processing is verifying the data that is come from original domains.

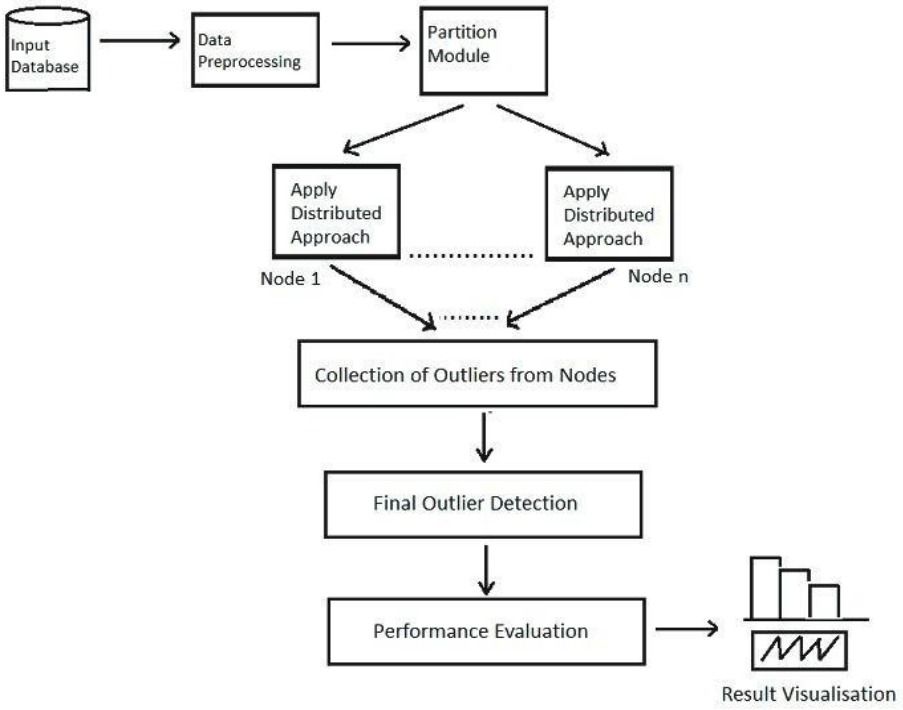

Fig.2. proposed system architecture.

The input dataset is being gone to the partition module. In partition module the data set is dividing into different clusters. In each cluster we have to find the infrequent occurrence of the items in the data set. We have to calculate the occurrence of the items in the dataset, calculate the support and confidence for the selected infrequent items.

1) Input data and Data Pre-processing:

Initial input data for this system will be gathered from standard dataset entry way i.e. The collected datasets may be available in their original ,uncompressed form, it is required to pre-process such information before sending to next steps, to pre-process the large datasets they are different methods available in the datamining, they are data integration, data transformation and data cleaning etc, and so on will be utilized and cleaned required information.

2) Data partitioning:

After success full completion of pre-processing we have to perform the partition of data into different clusters. partition of data into different clusters based on the their attribute values. These clustered data is pre-processed individuals to identify outliers based on applied algorithm method.

3) Outlier detection:

The proposed technique for identifying the outliers score will be first applied at distributed clusters and their results of identified outliers are collected and plot these outliers scores into the graph.

4) Performance evaluation and result visualization:

The detection of outliers by above methodology will be assessed on the basis of set evaluation parameters for their performance evaluation. This section gives the details about implemented system performance metrics and future implementation of system to be more understandable.

3.1 Antihub and Outliers:

Antihubs as an exceptional classification of focuses in high-dimensional spaces. Outliers identifying methods are classified into two types, they are local and global methodologies. outliers of a few information item can be founded on the local or global database of data objects. The difference between the local and global is exist entire variences and degrees the two contradicting extremes of local and global.

The antihubs and their development is a part of the analyzing and organizing data in high dimensional spaces, identified with different focus. This perspective will be for the most part to as numbers. The brief explanation for k-occurrences, hub and antihub.

1. K- occurrences:

Let dataset D € R^d is a set of n points. The point x which is x€ D, and the similarity measure. The k occurrences denoted Nk(x) is the number of items x found in the k-nn of remaining points in dataset D. K-nn is Nk(x) value and the count of point x within dataset D.

2. Hub:

Hubs are the points with the highest value of Nk(x) where x is the point in dataset D.

3. Antihubs:

Antihubs are the points with the lowest value of Nk(x) where x is the point in dataset D. Both antihubs and outliers are related. The below algorithm for detecting the outliers using the antihub method.

Input:
- Distance measure $dist$
- Ordered data set $D = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$, where $\mathbf{x}_i \in \mathbb{R}^d$, for $i \in \{1, 2, \ldots, n\}$
- No. of neighbors $k \in \{1, 2, \ldots\}$

Output:
- Vector $\mathbf{s} = (s_1, s_2, \ldots, s_n) \in \mathbb{R}^n$, where s_i is the outlier score of $\mathbf{x}_i$, for $i \in \{1, 2, \ldots, n\}$

Temporary variables:
- $t \in \mathbb{R}$

Steps:

1) For each $i \in (1, 2, \ldots, n)$
2) $t := N_k(\mathbf{x}_i)$ computed w.r.t. $dist$ and data set $D \backslash \mathbf{x}_i$
3) $s_i := f(t)$, where $f : \mathbb{R} \rightarrow \mathbb{R}$ is a monotone function

Algorithm 1. AntiHub method

3.2 K-NN algorithm:

k-nearest neighbour algorithm is one of the classic data mining method and its role is classification and similarity search. The k-nn returns the all objects that are matech the query object in data set D. The below diagram represents the k-nn algorithm.

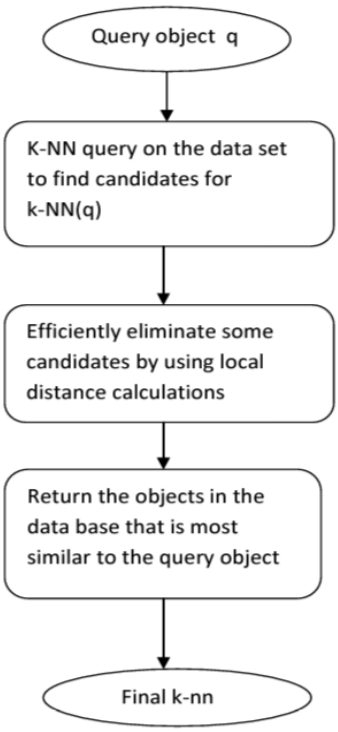

Fig.2. k- nn algorithm.

In the above diagram query object is given to the system ,the k-nn algorithm searches the all objects the are similar to the given query object and efficiently eliminate the some objects with the use of local distance calculations. The algorithm returns the objects the database that is most similar to the query object.

3.3 K-Occurrences:

k-occurrences identifying the number of occurrences in data set or database. It is a frequent item set mining, it identifying the frequent individual items In dataset and databases this are identified the frequent objects. The frequent

objects are identified by using the appriori algorithm ,we have to calculate the confidence and support for the infrequent items in a dataset. we have to maintain the high confidence and minimum support.

1. Support:

The X=>Y hold with support if the number of transactions or items in dataset D contains x+y.

Support= occurrence/total support.

Where occurrence is number of items found in frequent and total support is total number of transactions or related items.

2. Confidence:

The X=Y holds with certainty (or) confidence if the number of transactions (or) items in dataset D that contains X likewise contain Y.

Confidence=support(x+y)/support(x).

IV. CONCLUSIONS

We are provided in this paper a role of the antihub method ,k nearest neighbor and k-occurrences for the supervised outlier detection .we are explained the relationship between antihubs, outliers and hubs. The hubs, antihubs and outliers are machine learning concepts from different methods: supervise, unsupervised, semi supervised however in future work it would be interesting to semi supervised method.

REFERENCES

[1] Milo_s Radovanovi_c, Alexandros Nanopoulos, and Mirjana Ivanovic" Reversee nearest neighbours in unsupervised distance-based outlier detection,IEEE,vol 27,no 5,May 2015

[2] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms formining outliers from large data sets," SIGMOD Rec., vol. 29, no. 2,pp. 427–438, 2000.

[3] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: Algorithms and applications," VLDB J., vol. 8, nos. 3–4, pp. 237–253, 2000.

[4] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is"nearest neighbor" meaningful?" in Proc. 7th Int. Conf. DatabaseTheory, 1999, pp. 217–235.

AUTHOR'S PROFILE

B.Sivaiah received B.Tech degree in Computer Science Engineering from Rajeev Gandhi Memorial College of Engineering and Technology Nandyal, affiliated to JNTUA College of Engineering, Ananthapuramu, A.P, India, during 2010 to2013. Currently pursuing M.Tech in Artificial Intelligence at JNTUA College of Engineering, Ananthapuramu, A.P, India, during 2014 to 2016.Him area of interest is DataMining and Bigdata analytics.

Dr. A.P.Siva Kumar is an Assistant Professor of Computer Science and Engineering at Jawaharlal Nehru Technological University College of Engineering, Ananthapuramu. He obtained his Bachelor degree in Computer Science & Information Technology from Jawaharlal Nehru Technological University Hyderabad, Master of Technology in Computer Science from Jawaharlal Nehru Technological University Hyderabad and Ph.D. in cross lingual information Retrieval from Jawaharlal Nehru Technological University Ananthapuramu. He has published several Research papers in National \ International Conferences and Journals. His research interests includes Information Retrieval and Natural Language Processing.

