International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

A Survey On: Distance Based Outlier Detection

Smita Patil[1], P. D.Chouksey[2]

[1, 2]Department of Computer Engineering, BSCOER, Pune, Maharashtra, India

Abstract: ***Data mining is technique & used for outlier Detection. Outlier is a data point which is different from the rest of data. Outlier Detection finds the pattern which is not similar to regular behavior. The entire methodologies for outlier detection can be broadly categorized as supervised outlier detection methods, semi-supervised outlier detection methods and unsupervised outlier detection methods. Unsupervised outlier detection methods have been proved to be prominent in most cases, where high dimensional data come in practice. Outlier detection can usually be considered as a pre-processing step for locating, in a data set, those objects that do not conform to well defined notions of expected behavior. In the previous work, Antihub method and unsupervised method is used for outlier detection. Here the distance based outlier detection is proposed by using antihub and semisupervised learning.***

Keywords: Outlier, K-NN, High dimensional dataset, Hubness, Antihub.

1. Introduction

Outlier detection is the method which identifying Patterns that do not conform to established standard behavior. Hawkins defines "the outlier as observation that deviates to large extent from the other observation which means that the pattern is generated by the different mechanism"

Outlier detection is the process of finding knowledge from large and multidimensional databases to learn the unexpected pattern and behavior of objects. The paper applies the OD on the k-dimentional dataset with k>=5. This approach uses the distance based outlier detection for multidimensional dataset.

Clustering is process of a collection of data into groups with respect to a distance or similarity measure. The objective of the clustering is to divide data into different groups by using their similarities. Data objects are added to the group with which its similarity is higher that the other groups. In data mining, clustering is used to discovery of the distribution of data and the detection of patterns.

Here authors have proposed a new clustering algorithm called C2P. This approach exploits index structures along the processing of closest pair queries in spatial databases. It combines the advantages of the hierarchical agglomerative and graph-theoretic clustering algorithms. The paper provides extension for large spatial databases and for outlier handling

The outlier detection techniques operate in one of the three modes are;

1) Supervised outlier Detection:
   These techniques are trained in supervised mode and consider the availability of labeled instances for normal as well as outlier classes in a training dataset.
2) Semi-supervised outlier Detection:
   This technique is trained in supervised mode and considers the availability of labeled instances for normal and do not require labels for the outlier class.
3) Unsupervised Outlier Detection:
   These techniques operate in unsupervised mode do not require training data from any class.

There are many more outlier detection techniques based on the nearest neighbor which considers that outlier object appears far from their nearest neighbor. Such methods base on a distance or similarity measure to search the neighbors with Euclidean distance. Numbers of neighbor-based OD methods include defining the outlier score of a point as the distance to its *k*th nearest neighbor

2. Literature Review

Here given evidence to support the opinion that distance-based methods can offer more contrasting outlier scores in high-dimensional dataset. Author also shows that high dimensionality can have a different impact [1], by reexamining the notion of reverse nearest neighbors in the unsupervised outlier-detection context.

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

Here provided the role of reverse nearest neighbor counts in problems concerning unsupervised outlier detection. The main focus is given on the unsupervised outlier-detection methods and the hubness phenomenon in high dimensionality.

Extended the work of antihubs to the large values of k and explored the relation between the hubness and data sparsity based on the unsupervised outlier detection. The extension of anthubs improves the discrimination in the outlier scores. The existence of hubs and antihubs in high-dimensional data is relevant to machine-learning techniques from various families: supervised, semi-supervised, as well as unsupervised. Here only unsupervised method is used, it does not give accurate result as compared to the other methods.

H.-P. Kriegel, M. Schubert, and A. Zimek[4] has proposed angle based outlier detection (ABOD). Outlier detection in high-dimensional data uses the variances of a measure over

angles between the different vectors of data objects. In ABOD technique, used the properties of the variances to actually take advantage of high dimensionality and found to be less sensitive to the increasing dimensionality of a data set. This technique is less efficient than the classic distance-based methods. The disadvantage is only angle based is used not the classic distance-based methods.

The LOF compare the local density of instances with the densities of its neighborhood instances. After that it assigns the outlier scores to given data objects. If LOF score equal to ratio of average local density of k nearest neighbor of instance and local density of data instance itself then data instance is considered to be normal and not as an outlier.

Local density of instances is computed by finding radius of small hyper sphere centered at the data instance after that dividing volume of k [5], i.e. k nearest neighbor and volume of hyper sphere. In this assign a degree to each object to being an outlier known as local outlier factor [5].

Objects are isolated depending on the surrounding neighborhood, instances lying in dense region are normal objects [5], if their local density is similar to their neighbors and objects are outlier if there local density lower than its nearest neighbor [5]. It is a critical or lengthy process as compared to the distance based methods.

The antihub2 method is unsupervised outlier detection method used for anomaly detection in high dimensional dataset. 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 k-nearest neighbors. Antihubs are the point that occurs infrequently in nearest neighbors list. In this paper authors have refined the antihub method to refine the outlier scores of a point produced by the antihub method by considering the nk scores of the neighbors of the data point.

Discrimination of outlier scores produced by Antihub2 acquires longer period of time with larger number of iterations [6]. Because of this recursive AntiHub2 method was introduced to improve the computational complexity of discriminating the outlier scores using less number of iterations to detect accurate outliers in high dimensional data [6].

Drawback:

It is used computational complexity for outlier so it is critical as compared to classic method. Outliers are detected using their distance to the neighboring data points. This approach is found to be more effective non-parametric outlier detection technique. In this paper author presented RBRP algorithm for mining distance based outliers from the high dimensional dataset. RBRP scales log-linearly as a function of the number of data points and linearly as a function of the number of dimensions [7].

Proposed approach uses recursive procedure known as divisive hierarchical clustering. In this process at each stage in the recursion, iteratively data is partitioned k partitions. Process start with k random centers, and assign each point to its closest center [7], creating k partitions. It is partition based method

3. Limitation

Current System is based on unsupervised outlier Detection. This method does not required any training for data set. So that there is no any condition for Distance But this method does not give the accurate result.

This method is not required for the accurate result.

4. Proposed System

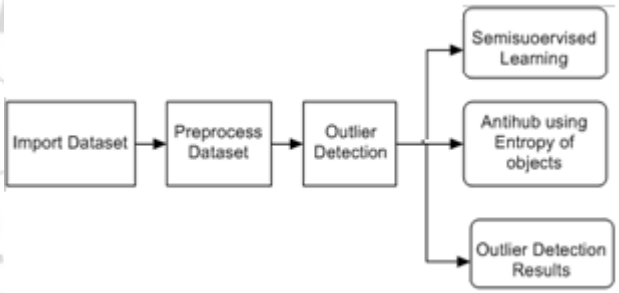
Proposed system uses the semi-supervised method which is used half training data. It gives more accurate result as compared to the unsupervised method.

The Proposed methodology for outlier detection is explained in this section. In the previous work, unsupervised distance based method used for outlier detection.

In the Proposed method semisupervised distance based outlier detection method is used. The advantage of this method is, it gives more accurate result as compared to the unsupervised distance based method.
The method is implemented with four phases.

1) In the first phase, import the data set.
2) In the second phase preprocess the data set. Here unsupervised learning approach is used. And calculation of Antihub using the entropy of objects.
3) Outlier detection results.

Architecture



Algorithm:

Input:
1. Training data set and objects.
2. Test data set.

Output:
H(X) –Entropy of objects.
Outlier Set.
1. Initialize Objects in the data set.
2. Do.
For each example data in the training set
a. T-Training data set
b. Outlier set
c. X is object
d. Calculate E threshold value
e. Obtain Entropy
f. Detection of outlier set

3. Return the data set.

5. Conclusion

The distance based supervised unsupervised etc approaches used for the outlier detection over high dimensional datasets. A Differet technique uses the different concepts such hubness, antihub sets to detect the outliers. Outlier scores also play an important role in outlier detection. This Paper presents a detailed survey of literature which was carried out ia data set for outlier detection. Based on the literature a new approach is proposed i.e. semisupervised learning method for outlier detection.

References

[1] Milos Radovanovi, Alexandros Nanopoulos and MirjanaIvanovi,"Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection", IEEE Transactions On knowledge And Data Engineering. Transactions, Vol. 27, No. 5, May 2015.
[2] Edwin, Raymond, "Distance based outliers: algorithms and applications", Springer- verlag, 2008.
[3] Alexandros Nanopoulos, Yannis Theodoridis, Yannis Manolopoulos, "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-based outlier detection in high-dimensional data, " in Proc 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2008, pp. 444–452.
[5] 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.
[6] 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.
[7] Amol Ghoting, Srinivasan Parthasarathy, and Matthew Eric Otey, "Fast Mining of Distance-Based Outliers in High-Dimensional Datasets" Springer, 2008.