Sunil Bhutada Int. Journal of Engineering Research and Application *www.ijera.com*
ISSN : 2248-9622, Vol. 6, Issue 7, (Part -3) July 2016, pp.52-56

RESEARCH ARTICLE OPEN ACCESS

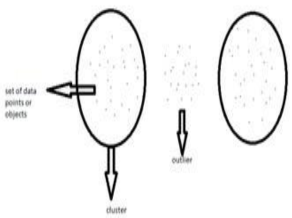
Outlier Detection Using Unsupervised Learning on High Dimensional Data

Sunil Bhutada[1], Anusha Velicheti[2]
[1]*Associate Professor In Sreenidhi Institute Of Science And Technology*
[2]*M.Tech Student In Sreenidhi Institute Of Science And Technology*

ABSTRACT
The outliers in data mining can be detected using semi-supervised and unsupervised methods. Outlier detection in high dimensional data faces various challenges from curse of dimensionality. It means due to the distance concentration the data becomes unobvious in high dimensional data. Using outlier detection techniques, the distance base methods are used to detect outliers and label all the points as good outliers. In high dimensional data to detect outliers effectively, we use unsupervised learning methods like IQR, KNN with Anti hub.
Index Terms: Outlier detection, unsupervised and semi-supervised learning, high dimensional data.

I. INTRODUCTION

OUTLIERS is the data objects that do not comply with the general behavior or model of the data, such data objects which are grossly different from or inconsistent with the remaining set of data are called outliers[1][3].



Outliers are of three types: supervised, semi-supervised and unsupervised.

1. Supervised learning is a task of machine learning which infer a utility from labeled training data. The data consist of a set of training examples. It consists a pair of an input object (vector) and a desired output value (supervisory signal). 2. Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses .The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data[6]. 3. Semi-supervised learning is a class of supervised learning tasks and techniques that also make use of unlabeled data for training small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data Supervised/unsupervised is the best way for basic data mining. There are 4 main tasks prediction, modeling, similarity and association. Prediction: If you are predicting a real number, it is called a regression. If you are predicting a whole number or a class, it is called a classification. Modeling: modeling is the same as prediction, but the model is comprehensible by humans, neural networks, and support vector machines work great, but do not produce comprehensible models, decision trees and classic linear regression are examples of easy to understand models. Similarity: If you are trying to find natural groups of attributes, it is called factor analysis. If you are trying to find natural groups of observation , it is called clustering. Association: It is much like correlation, but for enormous binary datasets.

1.1 Problem Statement

An outlier is an observation that appears to deviate markedly from other observations in the sample. Outlier may origin anomalous causes (the inconsistent data or any missing values). In some cases, it is not always possible to determine if an outlying point as bad data. Outliers may be due to random variation or may indicate something scientifically interesting. The following three issues regards to outliers:

Outlier labeling is a flag potential outlier for further investigation.

Outlier accommodation is used for robust statistical techniques that will not be unduly affected by outliers.

Outlier identification formally test whether observations are outliers. The outlier mining problem can be viewed as sub problems a. Define what data can be considered as inconsistent in a given data set. b. Find an efficient method to mine outliers so defined.

These problems can be solved by using following methods for outlier detection

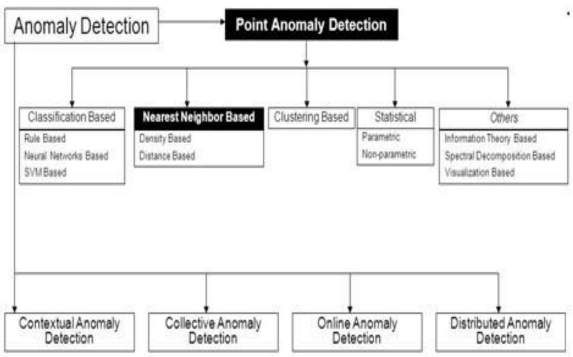
1. The statistical approach
2. The distance based approach
3. The density based local outlier approach
4. The deviation based approach.

When the outlier detection is risky that can be solved by distance based approach.

1.2 Theory

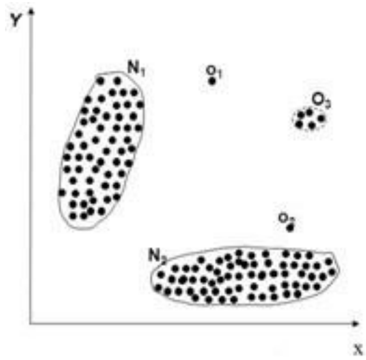
Outlier detection in high-dimensional data presents various challenges resulting from the curse of dimensionality. High dimensional data is the cluster analysis of data with anywhere from many dimensions. Curse of dimensionality arise when analyzing and organizing data in high dimensional spaces that do not occur in low dimensional settings. Some outliers can be designed to detect the presence of a single outlier while other tests are designed to detect the presence of multiple outliers. These can be grouped by the following characteristics:

a) What is the distributional model for the data? We restrict our discussion to tests that assume the data follow an approximately normal distribution. Is the test designed for a single outlier or is it designed for multiple outliers? If the test is designed for multiple outliers, does the number of outliers need to be specified exactly or can we specify an upper bound for the number of outliers? The following are a few of the more commonly used outlier tests for normally distributed data. This list is not exhaustive. The tests given here are essentially based on the criterion of distance based method. The distance based method is one of the methods in anomaly detection. Anomalies: the set of objects are considerably dissimilar from the remainder of the data. They occur relatively infrequently. When they do occur, their consequences can be quite dramatic and quite often in a negative sense. Anomaly is a pattern in the data that does not conform to the expected behavior. It is also referred as outliers, exceptions, peculiarities, surprise, etc. Anomalies translate to significant (often critical) real life entities like Cyber intrusions, Credit card fraud[2].



Type of Anomalies: Point Anomalies, Contextual Anomalies, Collective Anomalies.

Point Anomalies: An individual data instance is anomalous with respect to the data.



Output of Anomaly Detection:

Label: Each test instance is given a nor- mal or anomaly label. This is especially true of classification-based approaches. Score: Each test instance is assigned an anomaly score. It allows the output to be ranked and also it requires an additional threshold parameter.

II. OUTLIER DETECTION SCHEMES

Anomaly Detection Schemes: General Steps for anomaly detection is to build a profile of the normal behavior. Profile can be in patterns or summary statistics for the overall population. Use the normal profile to detect anomalies. Anomalies are observations whose characteristics differ significantly from the normal profile.

Anomaly detection methods: 1. Statistical- based 2. Model-based 3. Distance-based Outlier Detection

Distance based methods define outlier as an observation that is D min distance away from p percentage of observations in the dataset. The problem is then finding appropriate D min and p such that outliers would be correctly detected with a small number of false detections. This process usually needs domain knowledge. For measuring distance based outliers, we use KNN(K-Nearest neighbor). From different datasets, we identify outliers by forming clusters. The data which is outside the clusters are identified as outliers. Distance- based Approaches : Data is represented as a vector of features. Three major approaches
a. Nearest-neighbor based b. Density based
c. Clustering based.

Nearest-Neighbor Based Approach:

Compute the distance between every pair of data points. Nearest neighbor based outlier detection techniques can be broadly grouped into two categories based on how they compute the outlier score 1) Distance to Kth Nearest Neighbor Based[5] These techniques use the distance of a data instance to its kth nearest neighbor as the outlier score. 2) Relative Density Based These techniques compute the relative density of each data instance to compute its outlier score. 3) Using Other Manners - Additionally there are some techniques that use the distance between data instances in a different manner to detect outliers.

2.1 Nearest Neighbor Based Outlier Detection

Given two points p and q, we use dist (p,q) to denote the Euclidean distance between p and q. A k-Nearest- Neighbors query kNN(q,k,P) on a dataset P finds the set of k objects that are nearest to the query location q. Formally, an object p 2 P is in the result of kNN(q; k; P) if and only if it satisfies the following condition:

— $o \in P$ |dist(o, q) < dist(p, q)| < k. The idea of distance based outliers is extended by using the distance to the k-nearest neighbor to rank the outliers. Nearest Neighbor (NN) is finding the point closer to the query point in a given data set and a query point in k dimensional space. As dimensionality increases, the distance to the nearest data point approaches the distance to the farthest data point[3].

Definition Of Hubs And Antihubs

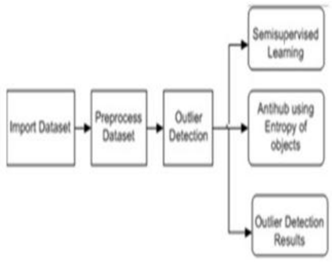
Definition 1: For $q \in (0, 1)$, hubs are the nq points $x \in D$ with the highest values of Nk(x).

Definition 2: For $p \in (0, 1)$, $p < 1-q$, antihubs are the np points $x \in D$ with the lowest values of Nk(x).

The anti hub method is unsupervised outlier detection method used for anomaly detection in high dimensional dataset. The data set exhibits that as dimensionality increases there exists hubs and anti hubs . Hubs are the point that frequently occurs in k-nearest neighbors. Anti hubs are the point that occurs infrequently in nearest neighbors list. Discrimination of outlier scores produced by Anti hub acquires longer period of time with larger number of iterations. The advantage of this method is, it gives more accurate result as compared to the unsupervised distance based method[7][8].

The method is implemented with four phases.
1) Import the data set.
2) Preprocess the data set. Here unsupervised learning approach is used.
3) Calculation of anti hub using the entropy of objects.
4) Outlier detection results.

Architecture:



2.1.1 Algorithm

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

Output: H(X) Entropy of objects.
1. 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
b. X is object

c. Calculate E threshold value e. Obtain Entropy
d. f. Detection of outlier set
3. Return the data set.

IQR [Inter Quartile Range] is another method to detect outliers.

2.2 IQR

The inter quartile range(IQR) is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.IQR is used in statistical analysis to help draw conclusions about a set of numbers. The IQR is often preferred over the range because it excludes most outliers.

2.2.1 Algorithm

1. Data arranged in order.
2. Calculate first quartile data (Q1)
3. Calculate third quartile data (Q3)
4. Calculate inter quartile data range (IQR)=Q3- Q1
5. Calculate lower boundary values = Q1-(1.5*IQR)
6. Calculate upper boundary values = Q3+(1.5*IQR)
7. Anything outside the lower and upper boundary value is an outlier.

III. CONCLUSION

The distance based outlier detection is per- formed on high dimensional datasets. The existence of hubs and anti hubs in high dimensional data is relevant to machine-learning techniques from various families: supervised, semi supervised, as well as unsupervised. Outlier scores also play an important role in outlier detection. This Paper provides a detailed survey of literature on distance based outlier detection. Based on the analysis, we illustrate the IQR and KNN with Anti Hub method for semi- supervised and unsupervised outlier detection. It improves accuracy and efficiency of distance based outlier detection.

REFERENCES

[1]. Jayshree S.Gosavi and Vinod.S. Wadne, Outlier Detection using Semi-supervised and Unsupervised Learning on High Dimensional Data, International Journal of Computer Application (2250-1797)Volume 5 No. 5, August 2015
[2]. V. Chandola, A. Banerjee, and V. Kumar, Anomaly detec- tion: A survey,60 ACM Comput Surv, vol. 41, no. 3, p. 15, 2009.
[3]. Ville Hautamaki, Ismo Karkk ainen and Pasi Franti, Outlier Detection Using k-Nearest Neighbour Graph, Proceedings of the 17th International Conference on Pattern Recognition (ICPR04) 1051-4651/04 IEEE
[4]. Hans-Peter Kriegel, Peer Krger, Arthur Zimek, Outlier Detection Techniques, The 2010 SIAM International Conference on Data Mining
[5]. Smita Patil , P. D.Chouksey, ”A Survey On: Distance Based Outlier Detection”, international Journal of Science and Re- search (IJSR)
[6]. Jayshree S.Gosavi , Vinod S.Wadne, ”Unsupervised Distance-Based Outlier Detection Using Nearest Neighbours Algorithm on Distributed Approach: Survey”, International Journal of Innovative Research in Computer and Communication Engineering
[7]. Mrs. R. Lakshmi Devi, Dr. R. Amalraj, ”An Efficient Unsupervised Adaptive Anti hub Technique for Outlier Detection in High Dimensional Data”, The International Journal Of Engineering And Science (IJES)Volume 4 Issue 11 Pages PP - 70-77 2015
[8]. J. Michael Antony Sylvia, Dr. T. C. Rajakumar, ”RECURSIVE ANTIHUB 2 OUTLIER DETECTION IN HIGH DIMEN- SIONAL DATA”.

Sunil Bhutada Graduated in B.E.(CSE) from Amravathi University in 1993. He received Masters Degree in M.Tech.(Software Engineering), from JNT University, Hyderabad, in 2006. He worked as Software Engineer thereafter and later shifted to academics in 1998. He is currently attached with Sreenidhi Institute of Science Technology in Hyderabad as Associate Professor in IT department. His areas of interest include Data Mining, Information Security, Information Retrieval System Presently he is pursuing Ph.D from

Jawaharlal Nehru Technological University, Hyderabad, India, in the field of Data Mining.

Anusha Velicheti is studying M.Tech (CNIS) in Sreenidhi Institute of Science Technology, Yamnampet, Gatkesar, Hyderabad. She completed B.Tech in CSE in the year 2012 from JNT University, Hyderabad. Her research interest include Data Mining.