Integrated Intelligent Research (IIR) International Journal of Data Mining Techniques and Applications
Volume 5, Issue 1, June 2016, Pages: 11-14
ISSN: 2278-2419

Survey on Outlier Detection for Support Vector Machine

Vijaya Shambharkar, Vaishali Sahare
Department of Computer Science & Engineering,G. H. Raisoni Institute of Engg & Technology for Women, Nagpur
Email: shambharkar.vijaya@gmail.com, vaishali.sahare@raisoni.net

Abstract - Outlier is the data object which does not comply with the general behaviour or model of data. Which are grossly different from entire set of data. From large data set detecting outliers present different challenge resulting from curse of dimensionality. As the data size is double every year, there is a need to detect outlier in large datasets as early as possible. If there are lot of outliers in data set there might de misclassification of data and outlier data might be classified as normal data. More contrasting outlier score gives by SVM in high dimensional data in which training the data set is relatively easy. SVM mainly focusing on high dimensionality of data, this method will be allowed to use a training data set to train the classifier while detecting outliers from high dimensional data.

Keywords - Outlier detection, high dimensional data. Curse of dimensionality, SVM.

I. INTRODUCTION

OUTLIER (anomaly) is the data objects which don't comply with general behaviour, despite the lack of mathematical definition of outliers and in practice which is widely use in practice. Outlier detection task can be categorised as supervised, semi-supervised and unsupervised,[1] depending on the presence of labels and regular instances. Among these categories, in propose approach use supervised method.In various knowledge domains outlier detection is a subject of interest. This process is a subject of increasing interest among analysts. As the size of data is doubling Prediction accuracy of outlier detection technique is generally high and robust, it works well training examples contain errors for fast evaluation of the learned target function. Clustering is very effective method for finding outliers, in various applications only clustering is used but only clustering is not sufficient for detecting and analysing the outliers because whenever we deal with large dataset possibility of detecting outlier as normal data. To deal with high dimensional data scalable model based clustering is required.

A. DEFINING OUTLIER

An outlier is a point that considerably dissimilar or deviates with the overall data set. Fig 1 illustrate outliers in 2-dimentional data set, N1 and N2 are two clusters in which most of the points lie.[2] Points that are away from clusters e.g. points O1, O2 and points in O3 are outliers.

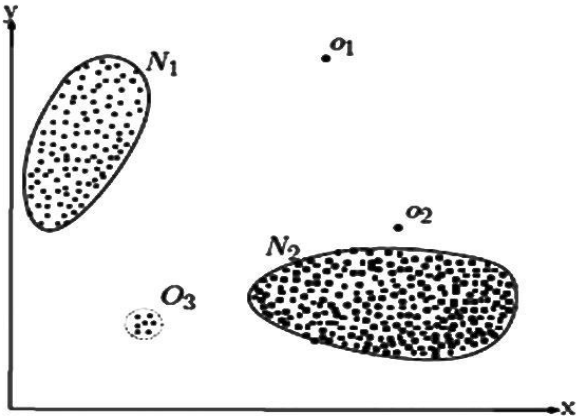

Figure 1: Example of outlier, points O1, O2 and O3 are the outliers in the figure. [2]

B. OUTLIER DETECTION SYSTEM

In data mining outlier detection system play important role, it is very crucial to find outliers in high dimensional data. It can be applicable in various domains such as, froad detection, network intrusion, weather prediction, medical investigation etc., due to its inherent importance in various domain considerable research efforts have been conducted in the past decade. No of outlier detection techniques have been proposed that use different mechanism and algorithm.

a. Outliers can be finding without prior knowledge of dataset. Called learning process in which data treat as a static distribution which finds remote points as potential outliers. The data set might be subdivided to improve outlier detection. It also assumes that normal instances of data are much more frequent than anomalous instances. A drawback of this approach is need dynamic and very large database.

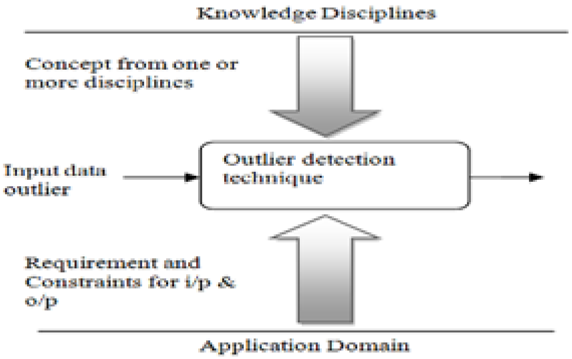
Knowledge Disciplines
Concept from one or more disciplines
Input data outlier
Outlier detection technique
Requirement and Constraints for i/p & o/p
Application Domain

Fig2: A general view of an outlier detection system [5]

b. In this approach assumed that the dataset only has labelled instances of 'normal' class. They do not require labels for anomalous class e.g. in case of safely critical systems where an anomaly would mean an accident which would be hard to model. Typically a model is built for normal behaviour and then is used to identify anomalies.

c. In this approach labelled instances are availability for both normal and anomalous class. When a new instance of data is encountered it is compared to the model to determine its class. To obtained labels for outlier classes is hard.

C. WHY WE NEED TO FIND OUTLIER

If there are a lot of outliers in the dataset there is possibility of misclassification of data and outliers might be classified as normal data. i.e. false negative rate will be very high, also computational complexity is very high. In recent years, outlier detection has attention in the society as well as in information industry due to extensive availability of usually large amount of data. There is a need for converting such data into helpful knowledge and information. The knowledge and information gained can be used for applications widespread from fraud detection, medical diagnosis, instruction detection Outlier indicate bad data. For example, coded data is incorrect .If outlying point is erroneous, then from analysis outlier score can be detected and also corrected if feasible. In some cases, it is not possible in some cases for us to detect an outlier point which is accurse due to number of variation. .However if in a data set contains multiple outliers we requires robust statistical techniques.

D. TYPES OF OUTLIER

There are three types of outliers in data set according to their nature they can be categorised in Point Outliers, Contextual Outliers and Collective Outliers.

a. **Point outlier**: If each single data point can be assume as an outlier with respect to the remaining data set then that instances known as point outliers. For example in fig 1 point O1, O2 and points in region O3 are outside or away from the clusters, and hence that points are outliers.

b. **Contextual outliers**: contextual outlier is a mini set of objects that contribute robust similarity with significantly larger set of objects on some attributes, but differ dramatically on some other attributes.

c. **Collective Outliers**: the subset of data objects which deviate collectively with respect to the entire data set even if data points are not outliers that can also consider as outliers.

E. OUTLIER DETECTION METHODS

In research field outlier detection methods deals with high dimensional and low dimensional data set, methods can be applicable on the nature of data.it can be divide in four types.

Statistical method: This method rely on the distance of data to fit the dataset. For Single dimensional data set statistical method are use, for real time data set it more suitable. it is model based technique which can divide in parametric, non parametric and depth based categories based on how the probability distribution model is built this. In parametric categories either considering known underline distribution of observation at least they rely on statistical estimates of unknown distribution parameters. In depth based categories data objects are arranging in convex hull layers with respect to peeling depth in the data space [3].

a. **Distance based method:** In machine learning and data mining field distance based techniques are windy used and it totally depends on the local neighbourhood of the data points. Need a distance computation among two data points if data is high dimensional then huge amount of computation required which will increases the computational cost. This method is also known as Nearest Neighbour Analysis and it can be used for classification, clustering along with most importantly for outlier detection [2].The most significant feature of the distance based outlier detection technique is that they have an explicit notion of proximity that is defined in the terms of a distance and similarity measure in between any two individual data set or a sequence of instances. In this method distance among two instances can be computed by using Lp metrix similarly Euclidean distance metrics, some non-metric distance functions are also used for making the distance based definitions of outliers very general.

b. **Density based method:** density based method having more complex mechanism also as compare to distance based method, to model the outlierness it computationally more crutial.it find out Local density of its nearest neighbour and higher modelling capacities to find outlier at the same time it require expensive computation [2].In density based method author proposed a local outlier factor (LOF) technique. The LOF method is indirect path of finding outliers. The basic idea of this technique is that, dividation of distances among a data points and all other points will analogous to the cumulative distance dividation for all pairwise distance if there are many other near-by points. There may be possible, points that are true anomalies and for which the peak in their distance dividation might exact same with the peak in the cumulative distribution. For any given data point outlier score can be assigned called as Local outlier factor (LOF), depending on its distance from its local neighbourhood.

c. **Clustering based method:** Among all the outlier detection methods Cluster based method are constantly most probably used for low dimensional dataset in the clustering. So many data mining algorithms find the outlier by clustering dataset themselves and detecting outliers which do not lie in the cluster or far away from the cluster. This method can be divided in two categories Partitioning Clustering Method and hierarchical clustering. In partitioning method data set can be split in to clusters in specific no, every cluster is denoted by k is specified by human users. It usually starts with the starting position then the objective function is optimized until the data set maximizes the optimal value of the data set. In the Partitioning methods, various centroid based methods like k-mean k-medoide are used. Another important category of clustering is hierarchical method in which entire data set is further breakdown into various subsets or small dataset.

II. LITERATURE SURVEY

In the literature survey, the survey of different outliers detection methods designed for detection is studied. We examined that however when dimension (attributes)are less it is for detection but when dealing with high dimensional data it is more complex, and requires special attention in performance when detection such as speed and accuracy of method to be used. From the abow methodology distance based method is more suitable foe high dimensional data set.

A. Basic of SVM

In machine learning technique Support Vector Machines (SVM) is today's popular technique which is used in a various applications. This includes for example handwritten digit recognition, object recognition, speaker identification, text categorization and also anomaly detection. In those applications, SVM give good result in terms of the generalization error**.**[5] SVM classifier perform the statistical analysis using terms frequency as input to the support vector machine calculations. For separation of two or multiple classes it uses hyperplan. Eventually, the calculation retained highest margin; where "margin" is define as the shortest distance from sample point to the hyperplan. The sample points that are from the margin are known as support vectors and to forms SVM model.

B. SVM model

SVM is a supervised machine learning technique in which data can be arrange based on prior knowledge or basic information to get from row data.[5] SVM can build a model for prediction of new example, In many application for separating data it play important role.

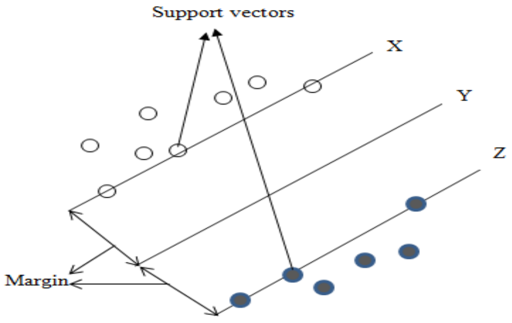

Fig3: SVM Model [5]

The fig3 represent the simple model of SVM technique. In model two patterns are present; the aim of this model is to break up these two patterns. It contains of three lines X, Y and Z the line Y is called as margin of separation. X and Z are the lines on the both sides of the line of margin. These three lines create the hyper plane that differentiates the given pattern and the patterns on X and Z lines called support vectors. The perpendicular distance between XY and YZ known as margin. For accurate classification SVM can maximize the margin and gives better classification result and hence minimize the error rate.

C. SVM classification technique

The support vector machine mainly use for classification of pattern that means this algorithm is mostly used for classifying the various types of patterns. Now, there is different type of patterns i.e. Linear and non-linear. Linear patterns are patterns that are easily distinguishable or can be easily separated in low dimension whereas non-linear patterns are the patterns that are not easily distinguishable or cannot be easily separated and hence these type of patterns need to be further manipulated so that they can be easily separated. Basically, for classification used optimal hyper plane, for linearly separable patterns.[6] and it is the major concept behind SVM. The optimal hyper plane is selected from the set of hyper planes for classification of patterns that maximizes the margin of the hyper plane i.e. the distance from nearest point of each patterns to the hyper plane. Maximize the margin so that it can correctly classify the given patterns is the main objective of SVM i.e. larger the margin size more accurately it classify the patterns.

III. PROPOSED METHOD

In the proposed SVM outlier detection method, a supervised learning approach can be used to detect anomalies it mainly focus on effect of high dimensional data set. In which feature extraction and classification are jointly perform. A feature is a combination of attributes that is of spatial interest and capture important characteristics of data. A training dataset contain known instances and which instances is used to learn a classification model. The learned model is then applied on the test dataset in order to classify labelled records into normal and anomalous records.Optimal hyper plane which is main object in svm use for separating the patterns this plane comprises of three lines, mainly marginal line and two other lines on either side of marginal lines where support vectors are located. When dimensions are increase svm can maximize the margin, it does not suffer from the curse of dimensionality so the classification rate is higher and it will improve the accuracy. Complexity parameter is also use by the SVM. Different pattern are easily separable using plan and lines.

IV. CONCLUSION

This paper describes existing outlier detecting techniques adopted for detecting outliers, different techniques depend upon different assumption and classification. SVM algorithm is efficient and fast. The Proposed SVM algorithm will improve the accuracy of outlier detection and decreasing false negative rate. It will identify the Support Vectors of a given set of points and focusing on effect of high dimensionality on supervised outlier detection.

References

[1] Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic "Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 5, MAY 2015 1369

[2] Kamal Malik1 H.Sadawarti2, Kalra G.S3.,Member IEEE, "comparative analysis of outlier detection techniques" Volume 97– No.8, July 2014

[3] Abhishek B. Mankar1, Namrata Ghuse2 "A Review on Detection of Outliers Over High Dimensional Streming Data Using Cluster Based Hybrid Approaches" Volume 3 Issue 11, November 2014

[4] S. Zhang, Ji Zhang “Advancements of Outlier Detection: A Survey” 04 February 2013

[5] Ashis Pradhan1 “Support vector machine- A survey” ISSN 2250-2459, Volume 2, Issue 8, August 2012

[6] Ms. Snehal S. Joshi1, Mr. Navnath D. Kale2 “Survey: Support Vector Machine and Its Deviations in Classification Techniques” Volume 4, Issue 12, December 2014

[7] P. Julia Grace, G. Jeyakumar “ A Statistical Decision making approach for disease Diagnosis” vol 01,Issue 02, December 2012 International Journal of Data Mining Techniques and Applications ISSN 2278-2419