An Efficient Approach for Outlier Analysis over Big Data

Deepa Verma  
Rajkiya Engineering College  
Ambedkar Nagar

Supriya Mishra  
Rajkiya Engineering College  
Ambedkar Nagar

ABSTRACT
Outlier analysis was used since a year decade for removing the irregular annotations from data sets. Outlier exists in data sets due to unexpected faults, activities, fake behavior, human intervention error, sensor error due to atmospheric interference etc. These detections are required to identify the system faults, irregular behavior of data in some specific application for decision making and to identify the frauds. The numbers of research have been proposed by the researchers but with limited solutions. Proposed detection techniques are not quite sufficient to resolve the current issues like performance and detect the accurate outlier from the big data. This paper entails an architecture and algorithm for outlier analysis over big data. We proposed an architecture and algorithm for outlier analysis which can be useful to resolve the current issues.

Keywords
Data String, Out layer detection, Data mining, Big data

INTRODUCTION
The most widely used field for research area is data mining techniques over big data, where extraction of interesting patterns, frequent patterns, list values, noisy data and data with irregular behavior mined from huge amount of data. As many applications and databases uses the information technology to ease the use of application but lots of problems remain exists as to efficiently extraction of missing and erroneous data from big data.

Outlier detection plays vital role in many applications such as network intrusion, medical diagnosis or fraud detection and needs to be more attention from the researchers. In most of the cases, it has been experienced that data mining and machine learning algorithm and different technique for statistical analysis are not giving the required result due to the presence of outliers. Correct identification and removal of outliers may improve the performance of these techniques. In some of the applications, data cleaning technique can be used for detecting and eliminating the outliers but with limited performance because different applications having different reasons for discovering outliers. Outlier detection can also be used in the mining of abnormal points for fraud detection, intrusion detection, stock market analysis, network sensors and marketing.

Generally, outlier is defined as data value which is out of place with respect to any rest of the data sets. Number of definitions has been proposed for outliers as irregular data observations is outlier. Another definition is identification of an observation which is extremely different from the other observations.

LITERATURE SURVEY
The result of outlier detection may be varies with different techniques in different domains. As the outlier detection technique depends on various factors such as input data distribution, data availability and constraints
over the application domain. Number of subsequent outlier techniques has been proposed by researchers is presented here in this section.

A Statistical Outlier Detection technique with Gaussian mixture model was proposed in [1]. Where each data point is given a formulated score and data point which have a high score declared as outlier. Detecting outlier based on the general pattern within data points was proposed by [2] where it combines a Gaussian mixture model and supervised method.

Depth based outlier detection has been discussed [3], where outlier detection search outliers at the border of the data space but independent of statistical distributions. These techniques are generally suited quantitative real-valued data sets or quantitative ordinal data distributions.

Deviation Based Outlier Detection method where a set of data points are given and outliers are detected as points that do not fit to the general characteristics of that set. So the variance of the set is minimized when removing the outliers. To deal with time series constraint oriented data, Sequential problem approach was proposed in [4] where outliers are identified by using normal features of data points and deviated features of data.

Explicit distance-based approaches proposed in [5] are based on the well known nearest-neighbor principle and uses a well-defined distance metric to detect outliers. They define outlier as the object which is greater in distance to its neighbors. The basic algorithm, the nested loop (NL) algorithm, calculates the distance between each pair of objects and then set as outliers those that are far from most objects. The NL algorithm has quadratic complexity, with respect to the number of objects, making it unsuitable for mining very large databases such as those found in government audit data, clinical trials data, and network data.

Aforesaid approach is further extended in [6]. In which they provide a prior consideration on distance of a point from its k the nearest neighbor. Where top k point are declared as an outliers. This approach alternatively proposed in [7] on the basis of outlier factor. Each data point is assigned formulated outlier factor computed as sum of distance from its k nearest neighbors. For detecting outliers linear time is used in [8] where data set get randomized for efficient search space. Recently we witnessed that a non parametric unsupervised based methods used for outlier detection which was proposed by a branch et al [9]. To address the uncertainty, temporal relation and transiency present within data distance based outlier detection for data stream method proposed (DBODDS) with the help of continuously adaptive data distribution function [10].

Density Based Outlier Detection method compares the density around a point with its local neighbors densities. The relative density of a point compared to its neighbors is computed as an outlier score. Density based outlier detection method uses density distribution of data points within data set. The idea of density based local outlier using comparison with density of local neighborhood was proposed [11].

Cluster analysis is popular unsupervised techniques to group similar data instances into clusters. Clustering partitions the data into groups, in which similar objects are contained. The assumed behavior of outliers is that they either do not belong to any cluster, or are forced to appear in a cluster where they are very different from other members or belong to very small clusters [12][13].

Streaming data uses a sliding window concept and different multi pass algorithms are used for detecting outliers within streaming data ([14],[15]). Some outliers were considered as inliers in other window, so this method is not efficient. Major problem is that sometimes outlier point may get classified as a inliers [15]. Choosing accurate window size in sliding window based outlier detection is required. Choice of sliding window is independent on data point used for implementation which gives poor result over outlier detection.

Outlier detection over time series data is mostly done using auto regression based outlier detection technique [16]. Auto-regression is also adopted for some outlier detection over streaming data [17], [18]. Outlier is detected using an estimated model and metric which computed based on comparisons.
PROBLEM DEFINITION AND OBJECTIVES
Presently, a few researches have been conducted on outlier data detection for large datasets. The proposed techniques for outlier detections are not up to the mark for accurate identification of outliers and also not maintaining the performance over big data.

The objective of the research work has to be setup as follows:
1. Architecture for integrating the data in data warehouse.
2. An efficient algorithm for outlier detection and analysis.

The proposed architecture and algorithm will be compared with the previously proposed techniques and technologies.

METHODOLOGY
Any research work begins with the literature review. One must be acquainted with the work already established in the same and related fields as well as be able to find out the difference of one’s work from other works.

The proposed research work entails critical review and examination of standards, algorithms and technologies at the preliminary stage and later a comparison of these against the fresh architecture and algorithms proposed on various performance benchmarks.

MATLAB (matrix laboratory) is a fourth-generation high-level programming language and interactive environment for numerical computation, visualization and programming. It allows matrix manipulations; plotting of functions and data; implementation of algorithms; creation of user interfaces; interfacing with programs written in other languages, including C, C++, Java, and FORTRAN; analyze data; develop algorithms; and create models and applications. It has numerous built-in commands and math functions that help you in mathematical calculations, generating plots, and performing numerical methods.

1) Dd tools
The dd tools Matlab toolbox provides tools, classifiers and evaluation functions for the research of one-class classification (or data description). The dd tools toolbox is an extension of the Prtools toolbox, more specifically, Prtools 5.0. In this toolbox Matlab objects for datasets and mappings, called prdataset and prmapping, are defined. dd tools uses these objects and their methods, but extends (and sometimes restricts) them to one-class classification.

2) Prtools
The dd tools toolbox build on top of Prtools, and therefore requires that Prtools is installed. Prtools allows training and evaluating classifiers and regressors in a simple manner. It contains numerous procedures for feature reduction, visualization and evaluation.

HYPOTHESIS
We have implemented our algorithm in MATLAB 7.10. To evaluate two performance factors such as clustering accuracy and outlier detection accuracy are used for analysis. We have used two biological data set, they are medical diagnosis and liver disorder. Detection rate refer to the ratio between the numbers of correctly detected outlier to the total number of actual outlier.

1. Medical diagnosis data set: In real World data repositories ,it is hard to find a data set for evaluating outlier detection algorithms, because only for very few real-world data sets it is exactly known which objects are really behaving differently. The diagnosis is binary: Benign and Malignant. There are two types of datasets so we are dividing datasets into 2 numbers of clusters.
2. Liver disorder data set: In this dataset only few numbers of outliers are detected at 75% from first cluster only one cluster is detected and in second cluster two outlier detected.

Fig 1: Clustering and outliers for cancer dataset

Fig 2: Clustering and outlier for liver disorder dataset

CONCLUSION
Now days, advances in hardware technology have allowed us to automatically record transactions of everyday life at fast rate. It leads to large amount of data which grow at an unbound less rate. Data streams are temporarily ordered, massive, fast changing and infinite sequence of data objects. Data stream clustering techniques are highly helpful to handle those data. The outlier detection is one of the challenging areas in data stream. By using data stream partial clustering and hierarchal clustering are helpful to detect out layer.

FUTURE SCOPE
The solution to issues in existing outlier techniques will leads the efficient utilization of technique by various departments such as income tax department, administration, weather forecasting department etc. to have the decision making.

REFERENCES


