PARALLEL ALGORITHMS FOR OUTLIER DETECTION IN HIGH-DIMENSIONAL DATA

Dr. C.P. Gupta and Arunima Sharma
Department of Computer Science and Engineering
University College of Engineering
Rajasthan Technical University, Kota, INDIA

Abstract - Outlier detection based on concept of deciphering different data by using conventional methods with modifications. This paper is based on different parallel approach for outlier recognition in High-dimensional (Mao et. al. 2015) Big-Data. The focal goal of Parallel Data mining is to discover significant patterns or rules in enormous datasets in real time. It is an inter-disciplinary area of research, which associate research from different areas such as machine learning, artificial intelligence (Hodge et. al. 2004), statistics, high performance computing, Big- Data and neural networks. Parallel Algorithms are best to detect outliers in a complex data set with a large future scope in present era.

Keywords - Outlier, Parallel algorithm, High-Dimensional data, Cluster, LOF, Map Reduce

INTRODUCTION

Precise identification of outliers has a significant role in statistical exploration. If conventional statistical prototypes are blindly applied to data encompassing outliers, the outcomes can be ambiguous from real one a lot. In addition, outliers themselves are habitually considered as the special data (Ben 2005) points of attention in many real-world situations and their identification is the main purpose of the exploration of data. Outlier Detection (Hautamaki et. al. 2004) is a wide exploration area that has engrossed important attention with several applications (Ngai et. al. 2011) in different areas of day to day life like discovering network invasion or credit card scam and others. There are numerous algorithms available for outlier detection that is somewhat effective to deal with different types of data but each data type has different algorithm because it need to operate differently than other type of data. Though, data sets available in the present day extant certain challenges to deal that make these current algorithms unreasonable or infeasible. Here quite a few outlier detection techniques are discussed in this paper which is addressing these challenges (Li et. al. 2015) associated to data sets presently existing to deal with. Some thoughtful methods are considered to identify outliers in large high-dimensional scattered data which comprise categorical or mixed-type characteristics. The latest approaches have runtime benefits and analogous accuracy compared to former approaches (Koufakou et. al. 2008), and is highly accessible with reverence to the amount of data points and features.
Figure 1 – Example of outlier (Red points) and clusters (Blue points) as scattered points.

PROBLEMS IN OUTLIER DETECTION

Outliers are always difficult to identify and analysed among several data values because of complexity of Big-data. But their absence can create problems for result. Each technique used for outlier detection identifies different values which make outlier detection difficult in field of statistics, because outlier for one technique maybe cluster (Christopher et. al. 2015) set for another one. Outliers are based on difference in data values but clusters are based on similarity among them, which cause set of outlier’s smalls than cluster set for a database. Graphical representation (Akoglu et. al. 2015; Zhang et. al. 1995; Shekhar et. al. 2001) of data is most suitable way for outlier identification among data points as scattered point. Another well-known problem in outlier detection is curse of dimensionality (Chandola et. al. 2009) occurs in large data set. It is caused because of variance and absence of data values in large database and it make difficult to reach at any point of conclusion.

CLASSIFICATION OF OUTLIER DETECTION APPROACHES

Like outliers, outlier detection approaches are also classified on basis of distance among data values, depth of data set, statics theory, dimension of data and other factors. Statistical tests are based on probability of distribution of data points. Depth based (Chauhan et. al. 2015) approach classify border values as outliers who are independent are present outside cluster density. Data which doesn’t belong to common characteristics means deviate is considered as outlier based on deviation from other datum values. Distance based approach is based of difference among neighbour (Brito et. al. 1997) data point. Density is well known factor to identify outliers. Density of outliers varies a lot from neighbour clusters density around it. Dimensionality of data increase as distance among data appoints increase. Due to sparse data most point behaves like an outlier in data set which reject concept based on density and distance for outlier detection.

Outlier detection on basis of data labelling is categorized as Supervised, Semi-supervised (Gao et. al. 2006) and Unsupervised outlier detection approaches. Labelled data is used in supervised approach as input to classify input set by using a function to identify outlier values which are present in input.. In Semi-Supervised Outlier detection technique analyses both labelled and unlabelled data parallel. Unsupervised (Campos et.al. 2016) outlier detection
approach is worked on unlabelled data and estimate output on based of unknown factors identified during processing used to create difference between outliers and rest of data.

![Classification of Outlier Detection Approaches](chart.png)

**PARALLEL ALGORITHMS FOR HIGH-DIMENSIONAL DATA**

An outlier detection based on perception of outlier detection unravelling set. Today’s High dimensional complex data is most difficult to handle to resolve it we need such techniques based on parallel approach for outlier’s identification. Main aim of Parallel Data mining (Aggarwal 2015) is to finding significant patterns or rules in large datasets. It is an interdisciplinary arena (Zanero et. al. 2004), which associate research from areas such as Big-data, machine learning, statistics, high performance computing (Billor et. al. 2000), statistical analysis and neural networks.

**MAP REDUCE ATTRIBUTE VALUE FREQUENCY (MR-AVF)**

MR-AVF is based on one of well-known MapReduce technique for parallel processing. AVF is scalable fast parallel technique for outlier identification for categorical data (Wu et. al. 2013). MapReduce assures load balancing and fault tolerance during implementation. Map Reduce is developed by Google whose main aim is to simplify processing of high dimensional (Aggarwal et. al. 2001) data on low cost cluster (Duan et. al. 2009) computers which contain several nodes to store and process data in distributed manner. Google develop it on Google file system (GFS) to reduce node failure by applying replication and speculative execution of data. It is fast, reliable, need minimum scanning, it has minimal user interference. Complexity of MR-AVF is linear which make it more useful for high dimensional data.

Existing Shortcoming in MR-AVF computation for each data point is complicated. Its performance is less than greedy technique on numerous dataset scans because greedy technique is based on several data scans (Xiong et. al. 2005), in which outliers are out of dataset.

**BORE (BAGGED OUTLIER REPRESENTATION ENSEMBLE)**

BORE is based on outlier scoring functions (OSFs) which are unsupervised (Zimek et. al. 2012) but the working of BORS is supervised which make it a semi-supervised (Vinueza et.
al. 2004) outlier detection technique. It perfectly measure disparity in labelled data is a predicted time in less price. OSF is used for data mapping to create a matrix whose each row is based on outlierness (Angiulli et. al. 2009) of data values. For BORE high dimensional data is easy to handle, because of allocation and calculation of outlierness can be done parallel. For conventional classifiers heterogeneity of outliers and imbalance in inherited data values is a major problem, which is resolved in BORE by Bagging and boot strapping aggregation.

In bagging assume Bags consists some sample data points which are need to be replaced, first find their average to create balance weight of each bag and in end each bag has equal number of outliers due to biased sampling but due to equivalence one outlier can be present in more than one bag. Its performance is high as compared to other approaches because of diverse set of outliers. It initiates the new research by combining outlier scoring function (Micenková et. al. 2015) in unsupervised manner. Shortcoming of BORE is the representation of arbitrary outliers in one bagging framework. It is not that much difficult for high dimensional data because classification and representation is done of each bag parallel, it is typical due to new instances derived at test time is problematic to handle.

CURE (CLUSTERING USING REPRESENTATIVES) ALGORITHM

CURE is a hierarchical approach to identify outlier by using tree for large database. CURE is combination of Partition and hierarchical clustering algorithm. When CURE is applied to Partitioning clustering, the summation of squared errors is seemed as large (Erfani et. al. 2016) variances in dimensions or geometrics of different clusters. When CURE is smeared to Hierarchical clustering, it processes the distance (Oku et. al. 2014) among (min, mean) distance, which work with altered shapes of clusters.

Running time of CURE Hierarchical clustering can be high if total numbers of levels are more, so to eliminate this problem in hierarchical clustering at each step centroid are merged with other clusters centroid. This enables correct identification of clusters and hidden complex outliers. Space complexity is O(n) and time complexity is O(n^2 log n) of CURE algorithm. Combination of Partition and hierarchical clustering algorithm make it easy to apply for large data set with high accuracy.

CLARANS (CLUSTERING LARGE APPLICATION BASED UPON RANDOMIZED SEARCH)

CLARANS is a randomized algorithm based on searching technique used to generate neighbouring nodes arbitrarily. When it find a neighbour node with suitable value it divide it in two sets and then start from new node else local minima (Gao et. al. 2010) is found start from beginning for next local minima. CLARANS always select data randomly at each step to avoid selection of same set in whole process. CLARANS is better than CLARA in terms of accuracy of result. As data increases the distance (Knorr et.al. 2000) among neighbour decreases which is a major advantage of local search in data set. Among several features of CLARANS one benefit of CLARANS clustering algorithm is that it gives best performance equated to other algorithms. It use diverse dimensions and categories of data sets and two performance aspects like clustering and outlier detection (Nguyen et. al. 2006, 2007) accurateness are used to check efficiency of outlier detection. Shortcoming of CLARANS is
that there are no predefined class label is present for the data values. In most cases outliers detected from a large volume of data increasing at a limitless rate.

**NDLOF, NDINFLO and NDDSNOF**

The local outlier factor (LOF) (Breunig et. al. 2000) is the first concept quantifying the degree of an object being an outlier. The LOF of an object is equal to the average of the ratios of the local reachability density of the neighbourhood and the local reachability densities of its neighbours. A parameter named MinPts is needed to determine the minimum size of the neighbours of the object. Obviously, the larger p’s local reachability density is, and the lower the local reachability densities of p’s MinPts-nearest neighbours are, the lower is the LOF value of p. INFLuenced Outlierness (INFLO) is presented to indicate the degree of an object being an outlier. Distance based outlier detection approaches are based on distance among data points and their neighbours. It has conventional theories for calculation of LOF, INFLO, and DSNOF. The new novel density based approaches guarantee the industrial process operating normally (Patel et. al. 2011) and mining algorithms give meaningful high quality output then conventional approaches. These novel density based approaches are ndLOF, ndINFLO, ndDSNOF which are based on new local density definition in the basis of the minimum hyper sphere (Yuan et. al. 2014) for outlier mining algorithm.

Figure 3 – Top-n Outliers detected in a dataset by ndLOF, ndINFLO, ndDSNOF, LOF, INFLO, and DSNOF.

**COMPARISON OF PROFICIENCIES IN OUTLIER RECOGNITION**

Categorized large data is complex to handle parallel with their time of arrival. They require large memory to save and fast to access (Hong et. al. 2015; Hubert et. al. 2015). Recent technologies which provide these types of facilities are cloud storage. For outlier detection in this area we require map-reduced technique for it. Map-Reduce are not efficient alone without some characteristics values. Therefore we use Attribute Value Frequency (AVF) approach with Map-Reduce technique which is known as MR-AVF Technique. It reduces complexity of large data sets by using parallel processing. Detecting a small number of outliers in a large dataset becomes complex in unsupervised outlier detection. In large datasets sometimes small values are ignored and not considered for processing which effect output of system a lot. To overcome this problem we use BORE. BORE is based on
unsupervised outlier scoring functions (OSFs) as features in a supervised learning framework which give optimized result.

Cluster based outlier detection has difficulty in extracting knowledge from nonstop fast growing data records parallel. For this type of datasets we use CLARANS and CURE Algorithms which are based on hierarchical and partition based techniques. Cluster based algorithms use k-neighbour based techniques. CLARANS is based on Random dynamic selection of data and CURE is based on Hierarchical methods to decompose a dataset. CURE is more efficient and also used for data streams in which different sizes and types of data sets are present.

CONCLUSION AND FUTURE DIRECTIONS

This paper gives a brief overview about various State of Art algorithms which are useful to handle high dimensional (Filzmoser et. al. 2008) data efficiently. In Future work we require to create such multi-dimensional (Filzmoser et. al. 2005) algorithms which work on expected probabilistic data values. Outlier Values are complex to analysis and because of unsupervised detection exact values of output can’t be detected so the value of output is considered as estimated value. To Increase importance of outlier detection Instead of Probabilistic output we require approximate output. Large database is complex to handle so we require partitioning it in small database but in dependent values it is complex and can’t be process parallel so such technique is required to partition data in independent subsets.

REFERENCES


