Performance Analysis of Big Data Using Outlier Detection Technique

Dimple Jeet Saini¹, Manisha Rajpoot², H.R Sharma
Mtech Scholar CT, PhD Scholar CSE, Dean, CSE
RCET Bhilai, Durg, RCET Bhilai, RCET bhilai
C.G., India, C.G., India, C.G., India

Abstract - Clustering and classification methods used in Data Mining. The goal is to provide a self-contained review of the concepts and the mathematics underlying clustering techniques. This data can be stored and maintained to generate information and knowledge. This information and knowledge has to be disseminated to every stake holders for the effective decision making process. Under this we utilize the concept of data pre-processing for outlier reduction. Here three algorithms were proposed by me namely K-mean, SOM(self organization mapping), and DBSCAN(Density Based Spatial Cluster Analysis with Noise) for detecting and removing outliers using a outlier score. By cleaning the dataset and clustering and classification based on similarity, we can remove outliers on the key attribute subset rather than on the full dimensional attributes of dataset.

Keyword— Datamining clustering, classification, Self organization mapping, DBSCAN, K-means

I. INTRODUCTION

Data warehousing is defined as a process of centralized data management and retrieval. Like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. With data mining, a retailer could use point-of-sale records of customer purchases to sent targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments. In the Architecture of data warehousing, there are many operation committed to handle the operation with mining process as compared to other it deprived the extraction transformation and load as it has operation layer integration layer data ware house and many database process can also conclude the server process through the consistent of enterprise process and processed work under the multilevel dimension process.

DATA MINING TECHNIQUES: There are several major data mining have been developing and using in data mining projects recently including association, classification, clustering, prediction, sequential patterns and decision tree. We will briefly examine those data mining techniques in the following sections.

Association: Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction. That’s is the reason why association technique is also known as relation technique. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together.

Classification: Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we develop the software that can learn how to classify the data items into groups.

Clustering: Clustering is a data mining technique that makes meaningful or useful cluster of objects which have similar characteristics using automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. To make the concept clearer, we can take book management in library as an example. In a library, there is a wide range of books in various topics available. The challenge is how to keep those books in a way that readers can take several books in a particular topic without hassle. By using clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label it...
with a meaningful name. If readers want to grab books in that topic, they would only have to go to that shelf instead of looking for entire library.

**K-MEAN** : This method for vector quantization originally occurred from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, it serves as a prototype of the cluster. This results in a partitioning of the dataset into Voronoi cells. The problem is computationally difficult in supervised learning but we can find the source for it and can have similar property but there are efficient heuristic algorithms that are commonly deploy and converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach by both algorithms. they both use cluster centers to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes and structure.

**DB SCAN**
Density-based spatial clustering of analysis with noise (DBSCAN) is a data clustering algorithm a set of points in some space, it groups together points that are closely packed together.

**II Literature Review**

Mrs. Bharti et al. [2014] the data mining technique should be able to identify the correlation between multiple attributes. Correlations Two types of data correlation appear at each sensor node. The first type is attribute correlation, that is, dependency among data attributes.

Hemlata Sahu et al. (2013) Clustering is unsupervised learning, where given data is categorized into subsets so that each subset represents a cluster which has distinctive properties. It has been considered a useful technique especially for applications that require scalability to large number of sensor nodes. Clustering also supports aggregation of data in order to summarize the overall transmitted data.

Mr. S. P. Deshpande et al. (2014) Data mining is the process of extracting application-oriented models and patterns with acceptable accuracy from a continuous, rapid, and possibly non-ended flow of data streams from sensor networks. In this case, whole data cannot be stored and must be processed immediately. Data mining algorithm has to be sufficiently fast to process high-speed arriving data. The conventional data mining algorithms are meant to handle the static data and use the multistep techniques and multi scan mining algorithms for analyzing static data-sets.

Boukerche et al. (2010) presented a distributed data extraction methodology to aggregate the data on sensor node which reduced the number of messages during transmission. The framework is evaluated on real dataset collected from a wireless sensor network sensors deployed in the Intel Berkeley Research Lab. Results show the strong correlation among some measurements, which is useful for anomaly detection.

Guralnik et al. (2011) use sequential pattern mining to learn typical behaviours of humans in their homes. Human behavior is inferred by using motion sensors, pressure pads, door latch sensors, and toilet flush sensors. These sequences are then analyzed by a human expert to identify complex behaviour models. These models can be used to select the appropriate response plan to the action of elderly.

Yoon and Shahabi (2012) present the clustered aggregation (CAG) algorithm that forms clusters of nodes sensing similar values within a given threshold (spatial correlation), and these clusters remain unchanged as long as the sensor values stay within a threshold over time (temporal correlation). CAG is a lossy clustering algorithm (most sensory readings are never reported) which trades a lower result precision for a significant energy, storage, computation, and communication saving.

**III PROBLEM STATEMENT & PROPOSED METHODOLOGY**

The outlier removal by density based method and kmean it provide less energy rate and value produced by data set is also not concerned as per the algorithm so it depicted as less efficient to detect the outlier value among cluster in centralized area of network as data set value for big data and real data are near but noisy so we proposed system will identify the value by unsupervised technique using K-mean ,SOM and DBSCAN to identify the removal of outlier from real data set by predictive clustering technique

The identification of distance measure : For numerical attributes, distance measures that can be used are standard equations like Euclidian, Manhattan, and maximum distance measure. All the three are special cases of Minkowski distance. But identification of measure for categorical attributes is difficult.

The number of clusters: Identifying the number of clusters is a difficult task if the number of class labels is not known beforehand. A careful analysis of number of clusters is necessary to produce correct results. Else, it is found that heterogeneous tuples may merge or similarly types tuples may be broken into many. This could be catastrophic if the approach used is hierarchical. Because in hierarchical approach if a tuples gets wrongly merged in a cluster that action cannot be undone. While there is no perfect way to determine the number of Clusters, there are some statistics.
that can be analyzed to help in the process. These are the Pseudo-F statistic, the Cubic Clustering Criterion (CCC), and the Approximate Overall R-Squared.

Lack of class labels: For real datasets (relational in nature as they have tuples and attributes) the distribution of data has to be done to understand where the class labels are assigned to exactly one of low, medium, and high. If there are missing, this is removed from dataset. If an attribute has too much missing values in which case methods have been suggested in Also, three cluster-based algorithms to deal with missing values have been proposed based on the mean-and-mode method in cluster.

Types of attributes in a database: The databases may not necessarily contain distinctively numerical or categorical attributes. They may also contain other types like nominal, ordinal, binary etc. So these attributes have to be converted to categorical type to make calculations simple.

ISSUES OF CLASSIFICATION ALGORITHM

Today Data Mining used in Many Application. So, There are various areas where datamining, classification used, but the main research issues and challenges are described below:

1) Data Cleaning: - Preprocess data in order to remove or reduce the noise (by applying smoothing techniques) and handle those missing values. (i.e. By replacing a missing value with the most commonly occurring value for that attribute, or with most probable value based on statistics) although most classification algorithms have some mechanisms for handling noise or missing data, this Step can help reduce confusion during learning.

2) Relevance analysis (feature selection): - Remove the insignificant or unnecessary attributes. Many of the attributes in the data may be irrelevant to the classification or prediction task. For example, data recording the day of the week on which a bank loan application was filed is unlikely to be relevant to the success of the application. Furthermore, other attributes may be unnecessary. Hence, relevance analysis may be performed on the data with the aim of removing any inappropriate or unnecessary attributes from the learning process. The time spent on relevance analysis, when added to the time spent on learning from the resulting “reduced” feature subset, should be less than the time that would have been sent on learning from the original set of features. Hence, such analysis can help to progress, classification efficiency and scalability.

3) Data transformation: - The Generalize and normalization of data. Numerical attribute income ⇒ categorical {Low, medium, high} Normalize all numerical attributes to [0,1]. The data can be generalized to higher-level concepts. This is particularly useful for continuous value of attributes. For example, numeric values for the attribute income may be generalized to discrete ranges such as low, medium, and high. Similarly, nominal-valued attributes, like a street, can be generalized to higher-level concepts, like city. Since Generalization abbreviate the original training data, minor input/output operations may Be involved during learning. Also, Large Database, Data Scalability Over fitting, Automation, accuracy, robustness, interpretability.

Unsupervised algorithms can detect outliers as a by-product of the clustering process. The Clustering algorithms works on the principle that objects that are far way from the centroid are treated as outliers. A Cluster-Based Outlier detection method. In the algorithm proposed, data is clustered using a K-Means algorithm and after sorting these objects with regard to the distance forms the outlier objects in the data. like that SOM used the byes process to analysis of weight and DBSCAN used the removal of outlier by its cluster approach.

K-MEAN

The most common algorithm uses an iterative refinement technique. Due to its ubiquity it is often called the k-means algorithm; it is also referred to as Lloyd’s algorithm, particularly in the computer science community. Given an initial set of k means m_1^{(0)},...,m_k^{(0)} the algorithm proceeds by alternating between two steps:

Assignment step: Assign each observation to the cluster whose mean yields the least within-cluster sum of squares (WCSS). Since the sum of squares is the squared Euclidean distance, this is intuitively the “nearest” mean. (Mathematically, this means partitioning the observations according to the Voronoi diagram generated by the means).

\[ S_i^{(t)} = \{ x_p : \| x_p - m_i^{(t)} \|^2 \leq \| x_p - m_j^{(t)} \|^2 \ \forall j, 1 \leq j \leq k \}, \]

where each \( x \) is assigned to exactly one \( S_i^{(t)} \), even if it could be assigned to two or more of them.

Update step: Calculate the new means to be the centroids of the observations in the new clusters.

\[ m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \]

ISSN: 2231-5381 http://www.ijettjournal.org
Since the arithmetic mean is a least-squares estimator this also minimizes the within-cluster sum of squares (WCSS) objective. The algorithm has converged when the assignments no longer change. Since both steps optimize the WCSS objective, and there only exists a finite number of such partitioning the algorithm must converge to a (local) optimum. There is no guarantee that the global optimum is found using this algorithm.

The algorithm is often presented as assigning objects to the nearest cluster by distance process aims at minimizing the WCSS objective, and thus assigns by "least sum of squares", which is exactly equivalent to assigning by the smallest Euclidean distance. Using a different distance function other than (squared) Euclidean distance may stop the algorithm from converging. Various modifications of k-means such as spherical k-means and k-medoids have been proposed to allow using other distance measures.

Given a set of observations \((x_1, x_2, \ldots, x_n)\), where each observation is a \(d\)-dimensional real vector, \(k\)-means clustering aims to partition the \(n\) observations into \(k\) \((\leq n)\) sets \(S = \{S_1, S_2, \ldots, S_k\}\) so as to minimize the within-cluster sum of squares (WCSS). In other words, its objective is to find:

\[
\text{arg min}_S \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2
\]

\[
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdOTSOM (SELF ORGANIZING MAP)

A distinguishing feature of SOM is that it uses a spacial organization on the centroids neurons lattice. Each centroid is assigned a pair of coordinates \((i,j)\). Sometimes, such a network is drawn with links between adjacent nodes, but can be misleading because the influence of one centroid on another is a neighborhood that is defined in terms of coordinates, not links. There are many types of SOM neural networks, but it will be focus on to two-dimensional SOMs with a rectangular or hexagonal organization of the centroids.

SOM is similar to K-means, there is a fundamental difference. Centroids used in SOM have a predetermined topographic ordering relationship. During the training process, SOM uses each data point to update the closest centroid and centroids that are nearby in the topographic ordering. The SOM centroids can also be thought of as the result of a nonlinear regression with respect to the data points. At a high level, clustering using the SOM technique consists of the steps described in Algorithm below:

Step 1: Initialize the centroids.
Step 2: repeat
Step 3: Select the next object.
Step 4: Determine the closest centroid to the object.
Step 5: Update this centroid and the centroids that are close, i.e., in a specified neighborhood.
Step 6: until The centroids don’t change much or a threshold is exceeded.
Step 7: Assign each object to its closest centroid and return the centroids and clusters.

DBSCAN (DENSITY BASED SPATIAL CLUSTER APPROACH FOR NOISE REMOVAL):

DBSCAN requires two parameters: \(\varepsilon\) (eps) and the minimum number of points required to form a dense or thick region (minPts). It starts with an arbitrary starting point that has not been visited. point's \(\varepsilon\)-neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is label as noise. Note that this point might later be found in a sufficiently sized \(\varepsilon\)-environment of a different point and hence be made part of a cluster.

If a point is found to be a dense part of a cluster, its is consider as \(\varepsilon\)-neighborhood is also part of that cluster. Hence, all points that are found within the \(\varepsilon\)-neighborhood are added, as is their own \(\varepsilon\)-neighborhood when they are also dense. This process continues until the density-connected cluster is completely found. Then, a new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.

In this system we will first select a data set. Our data set will be drift data with this we just identify the outlier with cluster and classification method by using the algorithm K-mean SOM, DBSCAN which work in unsupervised methodology we use basic step to continue the process which are as follows:

(1) Input the UCI data set value for drift data 
(2) full values must be pass in K-mean, SOM, DBSCAN algorithm 
(3) group data must be processed which is divided in to chunks 
(4) this data will be generated by both full and group data by using Matlab(803)2014

Figure 4.1 Flow chart of proposed system for removal of outlier.
Table algorithm for outlier removal from drift data

function [ndis rem] = outlier_removal(noc,cent,data,clas)

for j = 1 : size(data,1);
    a = cent(clas(j),:);
    b = data(j,:);
    dis(j) = euclid(a,b);
end
for i = 1 : noc
    ind = find(clas == i);
    sdis{i} = sort(dis(ind),'descend');
end
fra = 0.05;
for i = 1 : noc
    a = sdis{i};
    len = length(a);
    rem(i) = ceil(len * .05);
    if rem(i) == 0
        ndis{i} = a;
    else
        ndis{i} = a(1,1:end-rem(i));
    end
end

In the detection of outlier by unsupervised learning data stream is an unbounded sequence of data. As it is not possible to store complete data stream, for processing we divide it into data chunks of same size. Chunk size is specified by the user which depends upon the nature of data. It can be changed whenever required. When current data chunk is processed during that time incoming data is stored in buffer and later taken out as data chunk.

Values can be numeric and can be append on the data which we use to function the median of state by negative value and by positive values in the analytical parts.

In this figure the comparisons between two dataset is done where the full methodology is working on attributes and group data working with chunk data value which gives better result to remove outlier and noise to hold the rigidity the group data are evaluated in chunks in this true and false value.

IV. CONCLUSION

The goal of the algorithms presented in the paper is to improve the quality of data processing and to carry the underlying patterns in the data by reducing the impact of outliers at the pre processing stage. This outlier may be due to the noise and distortions in the data collection stage that consists of irrelevant or weak relevant data objects. From the algorithms, it is shown that by choosing a valid outlier score, the overall performance of the algorithm can be improved.
Analysis conducted using the three built-in drift datasets diabetes and sea data cancer data, shows the cluster-based outlier detection algorithm and classification with unsupervised method producing better accuracy than other detection method as the data carry huge identification can also be split into chunks so that the nearest value can sort the problem of outlier and noise.

**SCOPE OF FURTHER WORK**

This section discusses the few area where the current work can be taken further this work can be intended to be:-

The big data which carry noise and outlier can be removed by using cluster and classification algorithm it can also handle the bulky data with splitting the value into chunks.

The work can be handle for the science industry to handle the disease data set to remove outlier and noise in data mining issues

Attributed data and cluster performance can be check by using the algorithms and this technique lead to hold the neighbour weight of cluster nodes

In the future, we will work for devising new measures of outliers for streaming data and will apply the proposed framework to application specific real world data sets

We can also work further for the enhancement of neural network it can also lead the technology of artificial intelligence to enhance the weight and noise removal process.

**ACKNOWLEDGEMENT**

This work was supported by RCET, Bhilai, Chhattisgarh, INDIA. The author wishes to acknowledge all the open-access portals and websites that help in survey.

**REFERENCES**


**AUTHORS**

**First Author**  – Dimple Jeet Saini, M.Tech Scholar, RCET Bhilai  
**Second Author**  – Manisha Rajpoot,Ph.D Scholar, RCET Bhilai  
**Third Author**  – HR Sharma,Dean CSE RCET,Bhilai