International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-04, Issue-07, Oct 2018

A Comparative Study Of Clustering Algorithms For Outlier Identification

Priya. M[1], M. Karthikeyan[2]

[12]Department of Computer and Information Science, Annamalai University.

[1]mpriyaau@gmail.com, [2]karthiaucse@gmail.com

Abstract- The outlier analysis is an important task in data mining and cluster analysis is the most important technique. Clustering is the grouping of similar objects and the algorithms based on their similarity. In this paper focused on a comparison of clustering algorithms for outlier detection. The comparison has been made to detect outliers in the datasets by using K-Means, DBScan, Expectation Maximization, and Hierarchical clustering algorithms. These algorithms are implemented for the healthcare data sets such as breast cancer dataset, diabetes dataset, and liver dataset. The experimental result shows that for outlier identification and detection of clustering algorithms give most consistent and robust result comparing to other algorithms.

Keywords — Data mining, Outlier analysis, Clustering algorithm, K-Means, DBScan, Expectation Maximization, Hierarchical clustering.

I. INTRODUCTION

Data mining is an important technique of extracting and identity patterns from large amount of data. There are several techniques and algorithms are used to extract the hidden patterns from the datasets. Based on the patterns, the data mining task is to be classified into classification, association summarization clustering, etc. Data mining is an intelligence discipline and computational contribute tools for data into information, discovery of new knowledge and decision making. The mining process is an important part of verification and validation of patterns of the data.

Outlier analysis is an important task in data mining called to as outlier mining. It is used in many applications such as stock market analysis, fraud detection, marketing, intrusion detection, network sensors, medical diagnosis etc. An outlier, according to Hawkins, is "an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" [1]. By definition, outliers are representing a small portion and rare occurrences of the data. Detect and remove the outliers the approaches can be classified into Clustering-based approaches, Distribution-based approaches, Distance based approaches Depth-based approaches, and Density-based approaches. In all the approaches, the algorithm to detect outliers consists of two steps; the first to identify a profile around a data set. In the second step, a data point is analyzed and is identified as outlier when its attributes are different from the other attributes. All the approaches are assumed that the normal objects will be similar, while the anomalies will be different. Several measures are used to detect out the deviation of a data point from other points which tells the outlierness of a data point. As we know that outliers are in very few numbers in a data set, by removing the points which are outliers, so the computation time can be reduce.

Cluster analysis or clustering is a group of observations into subgroup called clusters so that the observations in the same clusters are similar. It is an unsupervised learning method and useful technique for the discovery of data distribution and patterns in the original data. Clustering algorithms are used for outlier detection, where outliers objects that are "far away" from any cluster. The purpose of clustering is to group together data objects that are close to one another. The knowledge obtained from data mining is useful when applied can directly help to serve the patients better for health care providers [2]. In this work, identify the outliers using clustering algorithms.

The rest of the paper is organized as follows. Section 2 describes discussion on the related works of clustering algorithms. Section 3 presents a detailed description of clustering algorithms. In section 4 gives some experimental results to show the performance of algorithm. Conclusion is given in Section 5.

II. REVIEW OF LITERATURE

A. Loureiro et al., proposed the outlier detection using clustering methods and describes the application of hierarchical clustering methods to the outlier detection task. This paper is to detect the erroneous of foreign trade transactions in data collection [3]. S.J. Redmond et al., described a method for initializing the K-means clustering algorithm using a *kd*-tree. The method describes the use of a *kd*-tree to perform a density estimation of the objects at various locations and to choose *k* seeds for the K-means

 DOI : 10.18231/2454-9150.2018.0975

algorithm [4]. M. F. Jiang et al., presented Two-phase clustering process for outlier detection. They assumed outliers behavior that they are do not belongs to any cluster and they are different from other members or belong to small clusters [5]. S. Vijayarani and Nithya,S., presented an efficient clustering algorithm for outlier detection. This paper discuss about the clustering techniques for outlier detection. They have described a new methodology for outlier detection. The result this paper shows that algorithm ECLARANS improves the accuracy of outlier detection and reduces the time complexity when compared with other algorithms [6]. S. Jiang el al., proposed a clustering-based outlier detection method (CBOD). This paper is used to detect outliers in two-step process. The first step used a one-pass clustering algorithm to cluster objects in the dataset and detect all small clusters as outliers and the second step as to determine a outlier factor of the outliers in the large clusters [7]. M. Ester et al., proposed to discover outliers using density based clustering algorithms in large spatial databases [8]. Breunig et al. developed the local outlier factor (LOF), for each object in the data set, which is one of the most frequently used methods in outlier detection and identify its degree of outlierness[9].

From the review of related literature, we proposed on a comparative study of clustering algorithms for outlier identification. The result of experiment shows the outlier identification. This work makes use of health care data sets such as breast cancer dataset, diabetes data set and liver data set.

III. Clustering Algorithms

Clustering or cluster analysis is defined as a grouping of data object into sub groups called clusters so that object belongs to the same class. It is an unsupervised learning in data mining. Clustering algorithms can be classified into different groups such as partitioning methods, hierarchical methods, and density based methods and spectral methods.

A. K-Means Clustering Algorithm

K-means clustering algorithm is widely used and well known partitioning based method. It is one of machine learning algorithm which can be used to recognize groups of similar object automatically. This algorithm classifies object to a pre-determined number of clusters by the user. Let a given data set, *D* of object *n* and the integer number *k*, the K-means clustering algorithm search and partition of *D* into *k* clusters,$C_1, C_2, \ldots, C_k$, that is, $C_i \subset D$ and $C_i \cap C_j = \emptyset$ for $(1 \leq i, j \leq k)$. The k-means algorithm is to identity groups of data are "closed-by" to each other. Cluster similarity is measured to take a *k* objects as centroids (mean value of object) randomly and choose each cluster as for as from each other.

The K-means algorithm proceeds by initialize the *k* value of the objects which represents cluster mean or center between the object. Based on the square of the Euclidean distance from the cluster objects and cluster mean, the remaining objects are assigned to the existing clusters. Choosing the closest to each object then computes the new mean of each cluster, so as the cluster center is updated. The result of square function is to makes *k* clusters. The quality of cluster C_i is to be measured by the within cluster variation, which is the sum of *squared error* between all objects in C_i and the centroid c_i, defined as the equation (1).

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2 \quad (1)$$

where *E* is the sum of square error for all data objects, *p* is the given data objects, c_i is mean(centroid) of cluster C_i. The algorithm steps are given below,

***Input*:** *D* contains *n* objects of a data set and *k* the clusters number.

Algorithm:

Step 1: *Initialize:* Select *k* object randomly to initialize the cluster.

Step 2: *Assign:* Find the cluster center which is closest for each object based on centroid (Mean value) of the objects and assigns the object is the most similar.

Step 3: *Update mean:* Compute the new mean value of each input data object that is to update the mean value of cluster center.

Step 4: Repeat the step 2 and step 3 until no change in the mean value and centroid position.

***Output*:** Set of *k* the clusters.

B. Density Based Clustering Algorithm

DBSCAN (Density Based Spatial Clustering of Application with Noise) is a density based clustering algorithm. This algorithm have been developed to discover arbitrary shape clusters with sufficient high density region with specified by the radius of data object and separated by low density region that represents noise. The data objects with high density above the determined threshold are constructed as a cluster. The DBScan algorithm has the ability to discover clusters with arbitrary shape such as oval, linear, concave etc., moreover, it does not need to require the specified number of clusters. The clusters are defined as density connected regions with respect to *ε*- neighborhood of objects (*radius*) and least minimum points (*MinPts*) of the object. The connectivity and density are measured by local distribution of nearest nieghbors. For each data point, the cluster is determined by the *ε*- neighborhood of object (*Eps*) and check if it contains greater than *MinPts* of data object. The DBScan clustering algorithm steps are shown below,

***Input*:** *D* as a data set and *Eps* - *ε* - neighborhood points, *MinPts* – minimum points

Algorithm:

 DOI : 10.18231/2454-9150.2018.0975

Step 1: Initialize *Eps* and *MinPts*.

Step 2: *Search:* find the cluster for the density connected region with corresponding given *Eps* and *MinPts*.

Step 3: Check whether if *Eps* of data point is greater than the *MinPts*, then new cluster is formed.

Step 4: Repeat the step 3 then it merge a few density reachable clusters until no new data points can be joined to any clusters.

***Output*:** Arbitrary shaped clusters.

C. EM Clustering Algorithm

The **EM (Expectation Maximization)** algorithm is to be a general method for locating the maximum-likelihood estimate of the parameters of an distribution from a given knowledge set once the information is incomplete or has missing values. In statistics, expectation maximization (EM) algorithm is an iterative method to find maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. There are two main applications of the EM algorithm, the primary one happens once the information gets from the observation method. The other happens once optimizing the likelihood operation can be implicit by assuming the existence values for added however missing (or hidden) parameters. The EM iterations between acting an expectation (E) step, which creates an operation for the expectation of the log-likelihood, evaluated using this estimate for the parameters, and maximization (M) step, that computes parameters increasing the expected log-likelihood found on the E step. These parameter-estimates are then used to describe the distribution of the latent variables within the next E step.The basic steps for the algorithm is shown below.

***Input*:** Data set and the clusters number.

Algorithm:

Step 1: Initialize the model's parameters and a probability distribution is created. This is the "E-Step" for the "**E**xpected" distribution.

Step 2: Newly observed object is fed into the model.

Step 3: The probability distribution from the E-step is tweaked to include the new data. This is called the "M-step."

Step 4: Steps 2 through 4 are repeated until stability (i.e. a distribution that doesn't change from the E-step to the M-step) is reached.

***Output*:** Set of clusters.

D. Hierarchical Clustering Algorithm

Hierarchical clustering is a method of cluster analysis in which hierarchy of clusters is created in such a way that the data objects in clusters are decomposed based on some criteria. The clusters thus obtained in hierarchy are known as dendogram that shows how the clusters are related to each other. There are mainly two approaches to generating a hierarchical clustering: (i) **Agglomerative:** In this algorithm starts with the data points as individual clusters and at every step, merge the closet pair to form a cluster. This process is repeated until single cluster is formed having all the points. (ii) **Divisive:** In divisive, all of the data objects are used to form one initial cluster. The cluster is split according to some principle, such as the maximum Euclidean distance between the closest neighboring objects in the cluster. This process is repeated until each cluster contains only a single object.

Hierarchical Cluster implements agglomerative (bottom-up) generation of hierarchical clusters. We tend to begin by taking each data point as a cluster. Then, merge two clusters at a time. So as to make a decision those two clusters to merge, we tend to compare the pairwise distances between any two clusters and choose a pair with the minimum distance. Once we tend to merge two clusters into a much bigger one cluster, a new cluster is made. Hierarchical clustering algorithm steps are given below.

***Input*:** Data Set.

Algorithm:

Step 1. Start by assigning each item to a cluster, each containing just one item that is level 0 and sequence number 0.

Step 2. Find the smallest dissimilar pair of clusters within the current clustering, according to wherever the minimum is over all pairs of clusters within the current clustering.

Step 3. Increment the sequence number by 1. Merge clusters into a single cluster to form the next clustering.

Step 4. Compute distances between the new cluster and each of the old clusters.

Step 5. Repeat steps 2 and 4 until all objects are clustered into a single cluster.

***Output*:** Set of clusters.

IV. RESULTS AND DISCUSSIONS

In this section, we focused analysis the outlier identification performance using clustering algorithms. In machine learning most of the work related to the domain of health care diagnosis has concentrated on the data set in the UCI repository. For experiments, we have used health care data sets such as breast cancer dataset, diabetes data set and liver disorder data set. The outlier identification is performed on the proposed method to real world datasets with different characteristics.

A. Dataset Description

Breast Cancer dataset which has been used for records the measurements for breast cancer cases. The data set contains the Dimensions of 699 data sample medical records (objects). Each record contains 11 attributes which are considered as risk factors for the occurrence of cancer. There are two classes to diagnosis, Benign and Malignant

 DOI : 10.18231/2454-9150.2018.0975

(malignant means cancerous and benign means non-cancerous). The objects labeled Benign as normal data and the malignant class considered as outliers. The dataset have 458 benign diagnosis records as normal objects and 241 malignant diagnosis records as outlier objects. The figure 1 shows the histogram evaluation of all attributes for breast cancer dataset.

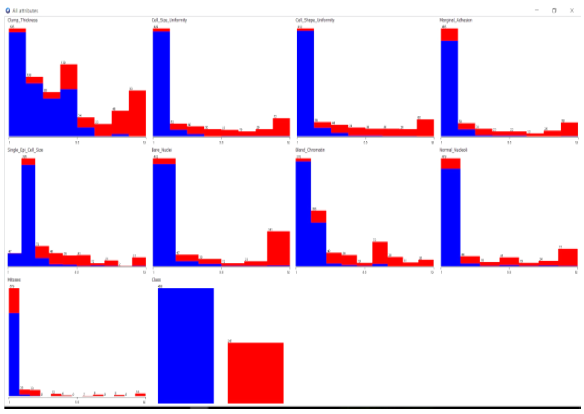


Figure 1 Histogram for breast cancer dataset

Diabetes data set which has been used for the records of Diabetes diagnosis. The data set contains the Dimensions of 768 data sample medical records (objects). Each record contains 8 attributes which are considered as factors for the occurrence of diabetes. The output class variable labeled as 0 or 1(class value 1 is for diabetes and 0 is for non-diabetes). The dataset have 500 records as normal objects and 268 records as outlier objects. The figure 2 shows the histogram evaluation of all attributes for diabetes dataset.

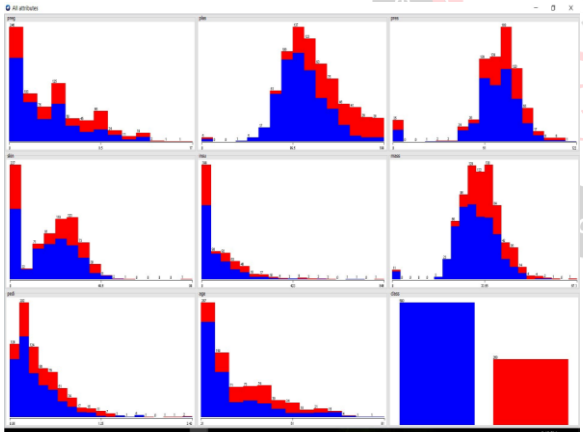


Figure 2 Histogram for diabetes dataset

Liver dataset has been used for the records of liver disease. The data set contains the Dimensions of 345 data sample medical records (objects). Each record contains 7 attributes; the first 5 variables are all blood tests which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. Based on the variable drinks number of pint alcoholic beverages drunk per day, the selector field is to split data into two class sets labeled as 1 or 2 that is presence or absence of a liver disorder. The dataset have 145 records as normal objects and 200 records as outlier objects. The figure 3 shows the histogram evaluation of all attributes.

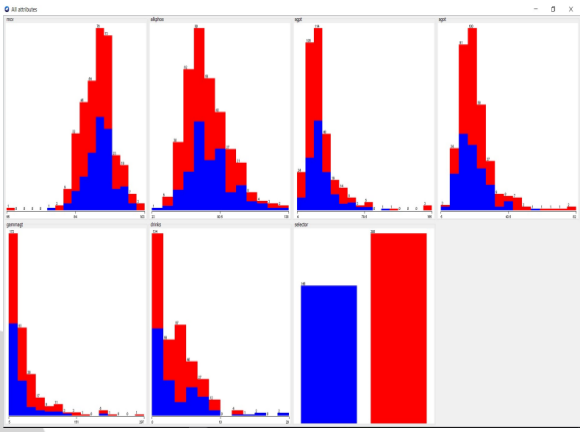


Figure 3 Histogram for liver dataset

B. Result Analysis

Data mining technique detect the relevant patterns or information from the raw data, using the data mining algorithms. Different clustering algorithms are used for the data mining technique. Each algorithm gives a unique result from the input data. These algorithms are compared with healthcare diagnosis dataset which produce the result as tested positive, tested negative for the affected and not affected by the disease. All three data sets can be clustered with two clusters, cluster 0 and cluster 1. The experiments performed on the datasets and gave the result as shown below. The table 1 shows the experimental result for breast cancer dataset. The K- means algorithm is to cluster the total objects into two clusters such cluster 0 as 35% objects as outliers and cluster 1 as 65% objects as normal objects. DBScan and Hierarchical algorithm is to cluster the total objects into two clusters such cluster 0 as 66% objects as normal objects and cluster 1 as 34% objects as outliers

Table 1. Result Analysis for breast cancer dataset

Datasets	No. of objects	Clustering Algorithms			Clustered Objects	
					Cluster 0 tested negative	Cluster 1 tested positive
Breast cancer	699	K –Means	No. of iterations	5	246	453
			Sum of squared error	259.92		
		DBScan	ε- neighborhood of object *(Eps)*	0.9	458	238
			Minimum Points *(MinPts)*	5		Unclustered-3
		EM	No. of iterations	3	276	423
		Hierarchical	No. of Clusters	2	458	241

 DOI : 10.18231/2454-9150.2018.0975

The table 2 shows the experimental result for Diabetes dataset. The K- means algorithm is to cluster the total objects into two clusters such cluster 1 as 35% objects as outliers and cluster 0 as 65% objects as normal objects. DBScan and Hierarchical algorithm is to cluster the total objects into two clusters such cluster 0 as 35% objects as outliers and cluster 1 as 65% objects as normal objects. EM algorithm takes 56% outliers and 44% normal objects.

Table 2. Result Analysis for Diabetes dataset

Datasets	No. of objects	Clustering Algorithms			Clustered Objects	
					Cluster 0 tested negative	Cluster 1 tested positive
Diabetes	768	K –Means	No. of iterations	4	500	268
			Sum of squared error	149.51		
		DBScan	ε- neighborhood of object *(Eps)*	0.8	268	500
			Minimum Points *(MinPts)*	6		
		EM	No. of iterations	46	432	336
		Hierarchical	No. of Clusters	2	268	500

The table 3 shows the experimental result for Liver disorder dataset. The K- means algorithm is to cluster the total objects into two clusters such cluster 0 as 81% objects as outliers and cluster 1 as 19% objects as normal objects. DBScan and Hierarchical algorithm is to cluster the total objects into two clusters such cluster 0 as 42% objects as outliers and cluster 1 as 58% objects as normal objects. EM algorithm takes 77% outliers and 23% normal objects.

Table 3. Result Analysis for Liver dataset

Datasets	No. of objects	Clustering Algorithms			Clustered Objects	
					Cluster 0 tested negative	Cluster 1 tested positive
Liver	345	K –Means	No. of iterations	9	280	65
			Sum of squared error	174.99		
		DBScan	ε- neighborhood of object *(Eps)*	0.7	145	200
			Minimum Points *(MinPts)*	3		
		EM	No. of iterations	18	266	79
		Hierarchical	No. of Clusters	2	145	200

C. Outlier Identification

After applying all the clustering algorithms, then we analyses and find out the best clustering algorithm for outlier identification. Table 4 shows that comparison results for identifying outliers.

Table 4. Number of outliers identified.

Datasets	Outliers identified by algorithms			
	K-Means	DBScan	EM	Hierarchical
Breast Cancer	246	241	276	241
Diabetes	268	268	336	268
Liver	280	200	266	200

We note that using DBScan algorithm and hierarchical clustering algorithms has to identifying outliers efficiently better than others. In breast cancer dataset K-means detected 246 outliers, DBScan detected 238 outliers and 3 as unclustered objects, it also considered as outliers, EM algorithm identified 276 as outliers and hierarchical algorithm identified 241 as outliers. In diabetes dataset K-means, DBScan and hierarchical algorithms are correctly identified 268 objects outliers and EM algorithm identified 336 as outliers. In liver cancer dataset K-means detect 65 as outliers, DBScan and hierarchical algorithms are correctly detected 200 outliers and EM algorithm identified 266 as outliers. DBScan and hierarchical clustering algorithms can be identifying outliers objects as same as for original number of outliers objects in the dataset. Thus the DBScan and hierarchical algorithms improves accuracy of identifying the outliers compared with other algorithms. The figure 4 shows that the number of outliers identified by the clustering algorithms.

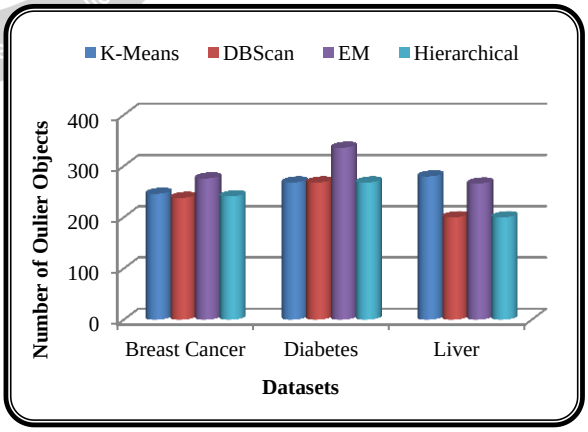


Figure 4 Outlier Identification

D. Time Complexity

The time complexity factor performance of clustering algorithms can be measured in the time taken to build the model for identifying outliers in each data set. Table 5 shows that time efficiency for clustering algorithms. In the breast cancer data set, K-means algorithm and DBScan algorithm has to take 0.04 seconds, EM algorithm as take

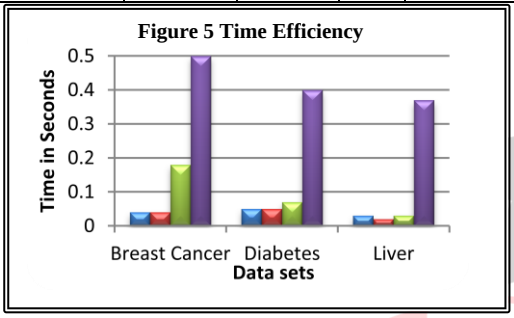
 DOI : 10.18231/2454-9150.2018.0975

0.18 seconds and hierarchical algorithm has to take 0.5 seconds. In diabetes data set, both the algorithms take same time as 0.05 seconds, EM algorithm as take 0.07 seconds and hierarchical algorithm has to take 0.4 seconds. In liver data set, K-means algorithm and EM algorithm has to take 0.03 seconds and DBScan algorithm as 0.02 seconds and hierarchical algorithm has to take 0.37 seconds. Time efficiency of the clustering algorithm for each data set is shown in figure 5. Comparing the performance of time, we note that using DBScan algorithm takes better time to taken than other algorithms.

Table 5. Time Efficiency

Datasets	Time taken to build the model in Seconds			
	K-Means	DBScan	EM	Hierarchical
Breast Cancer	0.04	0.04	0.18	0.5
Diabetes	0.05	0.05	0.07	0.4
Liver	0.03	0.02	0.03	0.37

Figure 5 Time Efficiency



V. CONCLUSION

In this paper, outlier identification and performance using various clustering algorithms has been discussed. Comparison has been made to identify the outliers for the healthcare data sets such as breast cancer dataset, diabetes dataset, and liver dataset using k-means, DBScan, EM and hierarchical clustering algorithms. With the help of tables and figures the analysis result of various algorithms used to form the cluster are shown. Every algorithm has their own importance and we use them on the behaviour of the data. The DBScan and hierarchical algorithms were identifying the outlier objects correctly. Moreover, the DBScan algorithm takes better time performance. Thus the experimental results demonstrate that outlier identification accuracy is efficiently to compare the clustering algorithms and it yields better results in outlier detection. Comparing the clustering algorithms, we note that using the DBScan algorithm takes better results than other algorithms. Outlier detection is very important techniques in various fields, like medical and public health outlier detection, credit card fraud detection, fault diagnosis in machines, weather prediction, network robustness analysis, and etc. In future clustering based on non-hierarchical techniques can also be used to validate results.

REFERENCES

[1] Hawkins. D.M, "*Identification of Outliers"*, Chapman and Hall, London, 1980.

[2] Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang, Lei Hua, August 2012, "Data Mining in Healthcare and Biomedicine: A Survey of the Literature", *Journal of Medical Systems*, Vol. 36(4), pp 2431-2448.

[3] Loureiro,A., Torgo, L. and Soares, C. (2004), "Outlier Detection using Clustering Methods: A Data Cleaning Application", *In Proceedings of KDNet Symposium on Knowledge-Based Systems for the public Sector*. Bonn, Germany.

[4] Redmond S.J., Heneghan,C., 2007,"A method for initialising the K-means clustering algorithm using kd-trees", *Pattern Recognition Letters*, Vol. 28, pp. 965–973.

[5] Jiang,M.F., Tseng,S.S., 2001, "Two-phase clustering process for outliers detection, *Pattern Recognition Letters*, Vol 22, No.6-7, pp 691-700.

[6] Vijayarani,S., Nithya,S., October 2011, "An efficient clustering algorithm for outlier detection", IJCA 0975-8887, Vol. 32(7).

[7] Jiang, S. and An, Q., (2008), "Clustering-Based Outlier Detection Method", *Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, Vol. 2, Pp.429-433.

[8] Ester,M., Kriegel,H.P., Sander,J., Xu,X., 1996, "A density-based algorithm for discovering clusters in large spatial databases with noise", *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Vol. 96, pp. 226–231.

[9] Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; Sander, J. (2000), "LOF: Identifying Density-based Local Outliers". Proceedings of the 2000, *ACM Sigmod Record, ACM*. Vol. 29(2): PP 93–104.

[10] Han. J, M. Kamber , J. Pei , "*Data Mining: Concepts and Techniques*" , Elsevier, 2011.

[11] Barnett.V , Lewis. T, "*Outliers in Statistical Data*", 3, Wiley New York, 1994.

[12] C. Aggarwal and P. Yu, (2001), "Outlier Detection for High Dimensional Data". *In Proceedings of the ACM SIGMOD International Conference on Management of Data*, Volume 30, Issue 2, pages 37 – 46.

[13] Duan. L., Xu. L., Liu. Y., Lee. J., 2009, "Cluster-based outlier detection", *Annals of Operations Research*, 168 (1) pp. 151–168.

[14] P. Vijaya, M.N. Murthy and D. K. Subramanian. Leaders-sub leaders, "An efficient hierarchical clustering algorithm for large data sets",*Pattern Recognition Letters* (2004) 505-513.

 DOI : 10.18231/2454-9150.2018.0975