International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 11 | Nov -2016 www.irjet.net p-ISSN: 2395-0072

COMPARISON OF CLUSTERING ALGORITHMS BASED ON OUTLIERS

Shivanjli Jain[1], Amanjot Kaur Grewal[2]

[1]*Research Scholar, Punjab Technical University, Dept. of CSE, Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib, Punjab, India*

[2]*Assistant Professor, Punjab Technical University, Dept. of CSE, Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib, Punjab, India*

Abstract: *Data mining, in general, deals with the discovery of non-trivial, hidden and interesting knowledge from different types of data. With the development of information technologies, the number of databases, as well as their dimension and complexity, grow rapidly. It is necessary what we need automated analysis of great amount of information. The analysis results are then used for making a decision by a human or program. One of the basic problems of data mining is the outlier detection. The outlier detection problem in some cases is similar to the classification problem. For example, the main concern of clustering-based outlier detection algorithms is to find clusters and outliers, which are often regarded as noise that should be removed in order to make more reliable clustering. In this research, the ability to detect outliers can be improved using a combined perspective from outlier detection and cluster identification. In proposed work comparison of four methods will be done like K-Mean, k-Mediods, Iterative k-Mean and density based method. Unlike the traditional clustering-based methods, the proposed algorithm provides much efficient outlier detection and data clustering capabilities in the presence of outliers, so comparison has been made. The purpose of our method is not only to produce data clustering but at the same time to find outliers from the resulting clusters. The goal is to model an unknown nonlinear function based on observed input-output pairs. The whole simulation of this proposed work has been taken in MATLAB environment.*

Keywords: ***Outliers, Data mining, Clustering, K-mean, K-medoid, DBSCAN, Iterative k-mean***

1. Introduction

Outlier refers to the additional data which occur in the dataset when the clustering is done. Outliers are patterns in data that do not conform to a well-defined notion of normal behavior.
Such data objects, which are grossly different from or inconsistent with the remaining data, are called outliers as shown in figure 1.

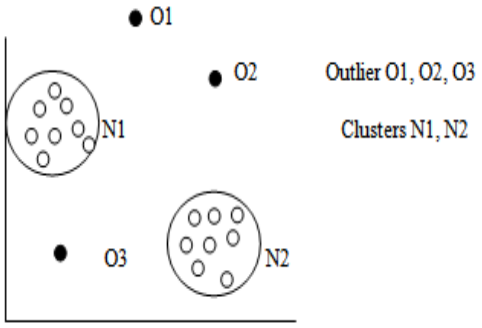

Figure 1: Outliers

Clustering is a data mining technique of grouping set of data objects into multiple groups or clusters so that objects within the cluster have high similarity, but are very dissimilar to objects in the other clusters. A software engine can scan large amounts of data and automatically report interesting patterns without requiring human intervention [20, 21].

Four algorithms are proposed related to supervised and unsupervised learning algorithms. The algorithms propose are: K-mean algorithm, K-medoids, DBSCAN (Density based clustering algorithm) and Iterative K-mean algorithm.

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The K-means module can be summarized as:

i. Find cluster values.
ii. Assign each document the cluster values that have low distance value.
iii. Foll***owing equation as shown below can be used to measure the cluster;
iv. $v_o = \frac{1}{m_k}\sum f_k$ (1.1)
v. Where$f_k$denotes the document vectors that belong to cluster; v_o stands for the centroid vector;$m_k$is the number of document vectors belong to cluster
vi. Repeating until good results are not obtained.

To reduce the complexities of K-means and to improve the detection rate, K-Medoids algorithm has been proposed. The K-medoid algorithm can be summarized as:

Begin

i. Obtain features of high priority than less priority from cluster size.
ii. Take medoids and use Euclidean distance to measure the dissimilarity between the clusters. After this sort out the clusters in an ascending order.
iii. Map each object with medoid that has close value and also find the optimal value from large number of objects
iv. Exchange the current medoid with the medoid that has minimum valu of dissimilarity.
v. Again Exchange the current medoid with the medoid that has minimum value of dissimilarity. But if the value is same as the previous then algorithm will be stopped otherwise repeat step 4.
vi. End

Density based clustering algorithm (DBSCAN) is the most widely used density based algorithm. It uses the concept of density reach ability and density connectivity [10]. Steps involved in implementing DBSCAN are as follows:

i. Start with an arbitrary starting point that has not been visited.
ii. Extract the neighborhood of this point using ε (All points which are within the ε distance are neighborhood).
iii. If there is sufficient neighborhood around this point then clustering process starts and point is marked as visited else this point is labeled as noise (Later this point can become the part of the cluster).
iv. If a point is found to be a part of the cluster then its ε neighborhood is also the part of the cluster and the above procedure from step 2 is repeated for all ε neighborhood points. This is repeated until all points in the cluster is determined.
v. A new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.
vi. This process continues until all points are marked as visited.

The *k*-means approach to clustering performs an iterative alternating fitting process to form the number of specified clusters, known as Iterative k-mean algorithm. The *k*-means approach is a special case of a general approach called the *EM algorithm*; *E* stands for Expectation (the cluster means in this case), and *M* stands for maximization, which means assigning points to closest clusters in this case.

2. A Glance of existing techniques

Saif et.al, presented data analysis techniques for extracting hidden and interesting patterns from large datasets. DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a pioneer density based algorithm. Many researchers attempted to enhance the basic DBSCAN algorithm, in order to overcome these drawbacks, such as VDBSCAN, FDBSCAN, DD_DBSCAN, and IDBSCAN. In this study, author survey over different variations of DBSCAN algorithms that were proposed so far. These variations are critically evaluated and their limitations are also listed. Haizau et.al proposed a robust method for robust local outlier detection with statistical parameters, which incorporates the clustering, based ideas in dealing with big data. . The experimental results demonstrate the efficiency and accuracy of the proposed method in identifying both global and local outliers, Moreover, the method also proved more robust analysis than typical outlier detection methods, such as LOF and DBSCAN. Hans et.al, propose a novel outlier detection model to find outliers that deviate from the generating mechanisms of normal instances by considering combinations of different subsets of attributes, as they occur when there are local correlations in the data set. This model enables to search for outliers in arbitrarily oriented subspaces of the original feature space. Abir et.al, proposed that clustering algorithms to determine the critical grouping in a set of unlabeled data. Lot of clustering works engages input number of clusters which is severe to find out. Yiang et.al, showed that The K-Medoids clustering algorithm solves the problem of the K-Means algorithm on processing the outlier samples, but it is not be able to process big-data because of the time complexity.

3. Simulation work

The predefined module of the current structure is required to be changed; therefore, comparison has been made based on four methods i.e. K-Mean, K-Medoids, Density based and iterative K-Mean. The main aim of proposed work is too made possible of outlier detection on text and compound data. In the research work, classification accuracy using clustering algorithm has been increased. The outliers need to detect for the better results with the compound data set. In the end, we present comparative method to detect the outlier like K-Mean, K-Medoids, Density based and iterative K-Mean and achieve more accuracy. For simulation, following parameters are considered:

i. Precision: In proposed work we obtained precision value:

$$\text{Precision} = \frac{True\ Positve}{True\ Positve+False\ Positive}$$

ii. Recall: In proposed work we obtained recall value:

$$\text{Recall} = \frac{True\ Positve}{True\ Positve+False\ Negative}$$

iii. Efficiency: In proposed work we obtained efficiency value:

$$\text{Efficiency} = \frac{\text{Precision}}{\text{Recall}} * 100$$

iv. F-measure: In proposed work we obtained f-measure value:

$$\text{F Measure} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$$

v. Accuracy: Accuracy is a general term used to describe how accurate a biometric system performs. Its formula is given as:

$$\text{Accuracy} = \frac{True\ Positve+True\ Negative}{True\ Positve+False\ Positive+\ True\ Negative+False\ Negative} * 100$$

vi. Cost

Cost is a general term used to describe how much cost is suitable for a system. It depends on the processing time and outlier.

4. Methodology

Following are the steps as methodology for the execution of the work:

Step 1: Firstly initiate the dataset that contain mixed elements or data.

Step 2: Apply K-mean on the dataset to formulate the clusters.

Step 3: Clusters are created using k-means now sort the cluster to easily find out the outlier if any.

Step 4: Now check the Semantic outlier using Outlier Detection Integrating Semantic Knowledge.

Step 5: Apply K-mediods on the dataset to formulate the clusters.

Step 6: Clusters are created using k-medoids now sort the cluster to easily find out the outlier if any.

Step 7: Now check the Semantic outlier using Outlier Detection Integrating Semantic Knowledge.

Step 8: Apply density method on the dataset to formulate the clusters.

Step 9: Clusters are created using density method now sort the cluster to easily find out the outlier if any.

Step 10: Now check the Semantic outlier using Outlier Detection Integrating Semantic Knowledge

Step 11: Apply iterative k-mean on the dataset to formulate the clusters.

Step 12: Clusters are created using iterative k-mean now sort the cluster to easily find out the outlier if any.

Step 13: Now check the Semantic outlier using Outlier Detection Integrating Semantic Knowledge.

Step 14: Calculate the results and compare the proposed work results using various metrics.

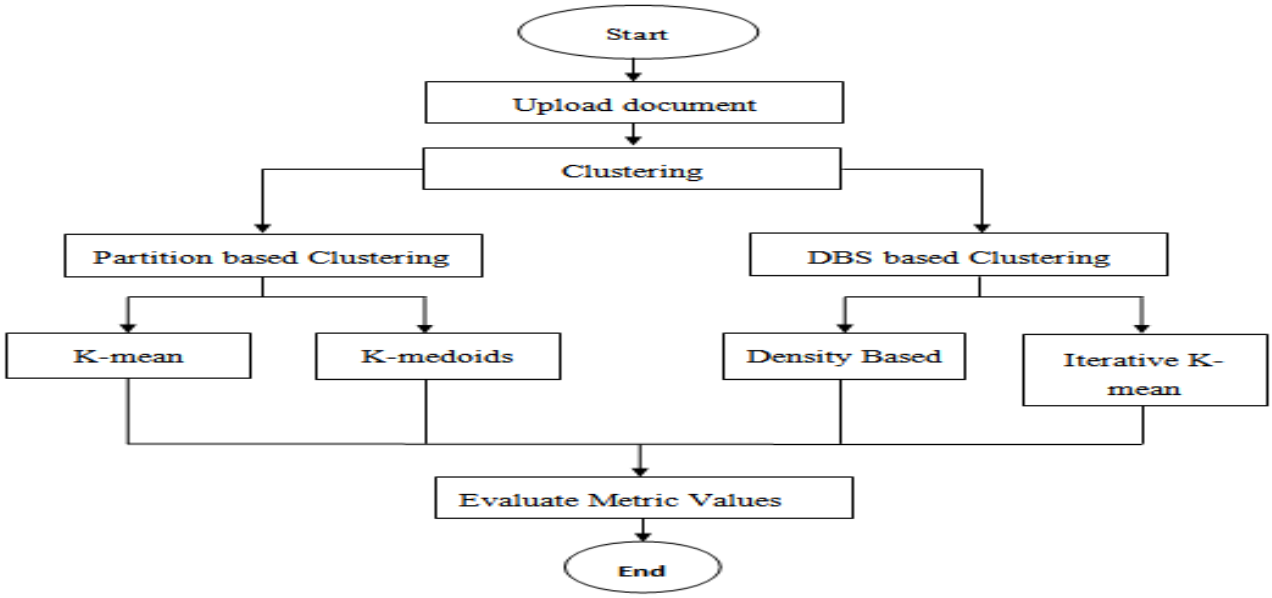

Figure **2:** Simulation flowchart

5. Simulation result

In this section, result and analysis of the research is done. The whole simulation is done in MATLAB.

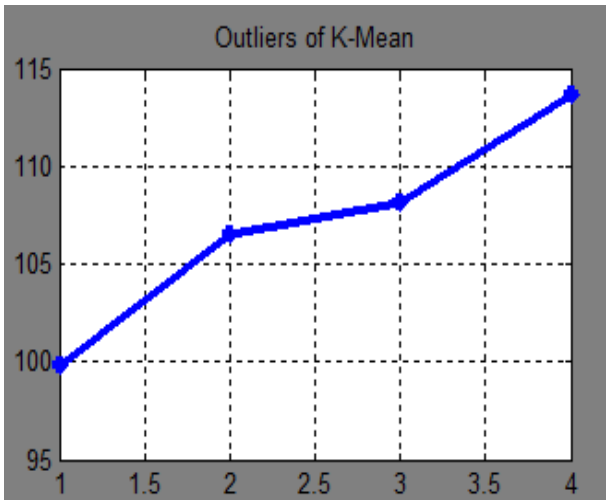

Figure 3: Outliers of K-Mean

Figure 3 shows the outliers found during K-mean algorithm utilized in data mining.

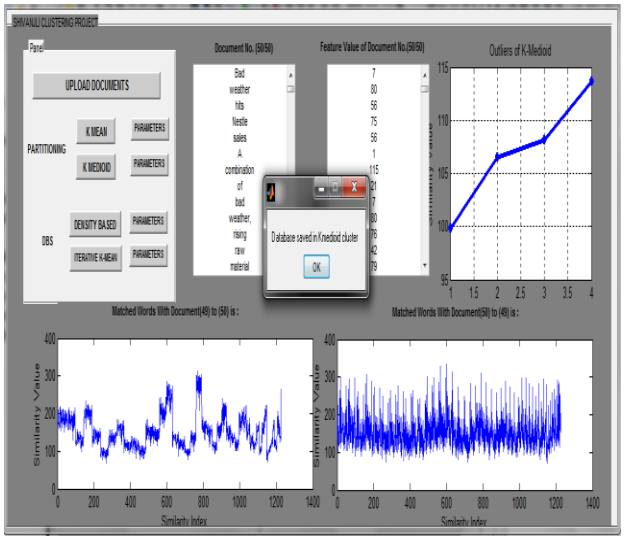

Figure 4: K –Medoid Utilization

Figure 4 shows the similarity found using K-Medoid algorithm. In above figure, matched words as well as similarity value is shown with feature vector representation.

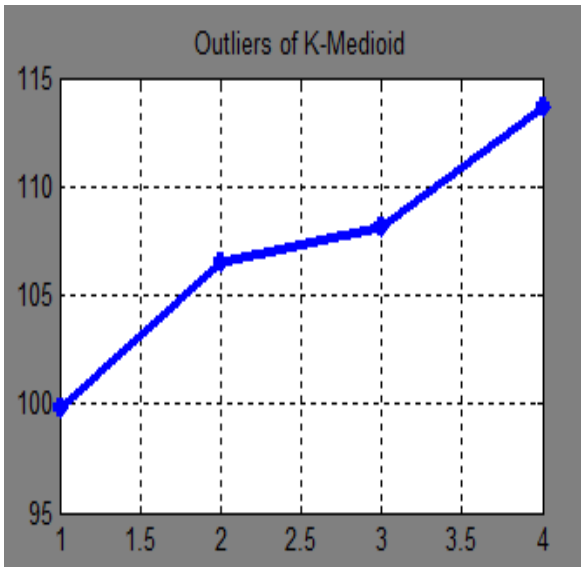

Figure 5: Outliers of K-Medoid

Figure 5 describes the outliers that are found using K-Medoid algorithm in data mining.

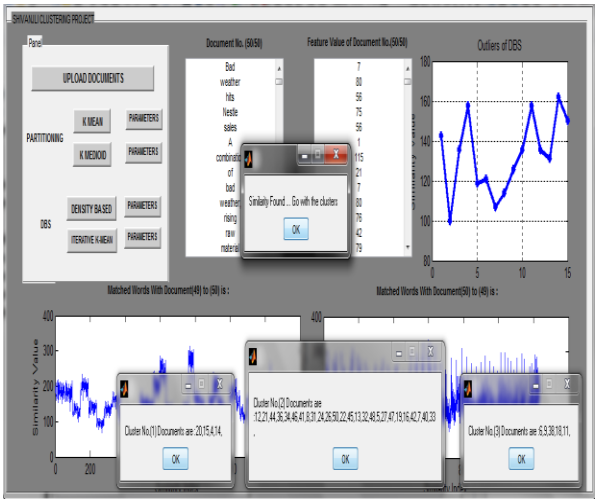

Figure 6: DBSCAN Utilization

Figure 6 shows the similarity that has been found using DBSCAN algorithm. In above figure matched words as well as similarity value has been shown with feature vector representation

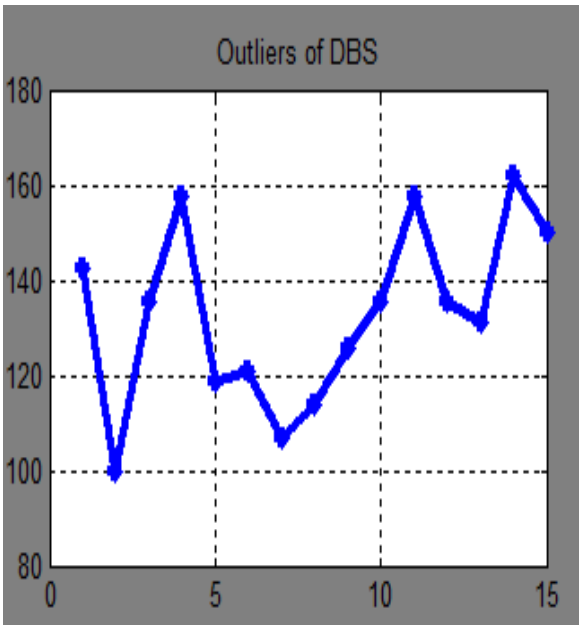

Figure 7: Outliers of DBSCAN

Figure 7 describes the outliers found in clustering using the DBSCAN algorithm.

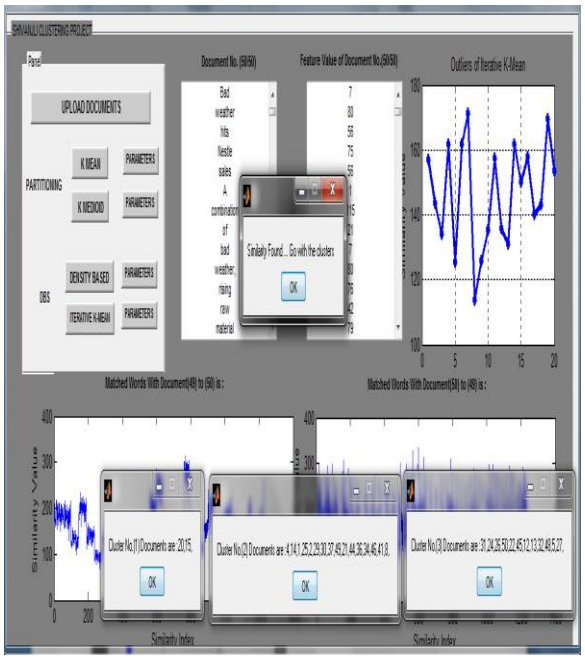

Figure 8: Iterative k Mean Utilization

Figure 8 shows the similarity that has been found using Iterative K-Mean algorithm. In above figure matched words as well as similarity value is shown with feature vector representation. With the help of given formula we calculate the parameters for proposed work.

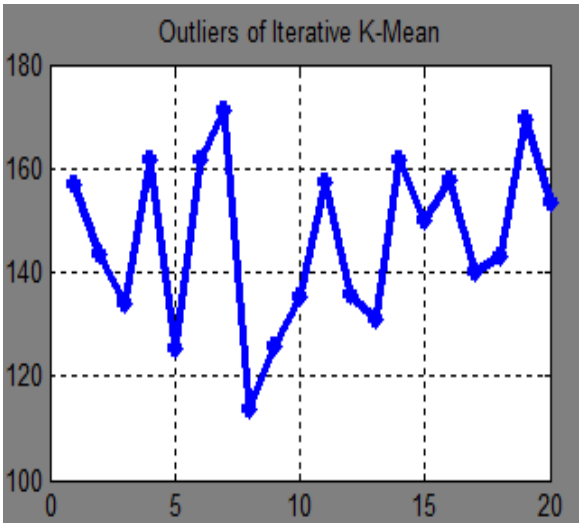

Figure 9: Outlier of Iterative K-Mean

Figure 9 describes the outliers found in the clustering using Iterative K-Mean algorithm.

Table 1: Comparison table of Accuracy, Precision, Recall, F-measure, Efficiency and cost w.r.t. to K-mean, K-medoid, DBSCAN and Iterative K-mean algorithms

Algorithm	Precision	Recall	F-Measure	Accuracy	Efficiency	Cost
K-Mean	.8	.92	.85	79.26	86.95	3.6
K-Medoids	.8	.92	.85	79.26	86.95	2.5
DBSCAN	.3	.7	.42	89.76	42.86	1.63
Iterative K-Mean	.4	.6	.48	84	66.66	5.31

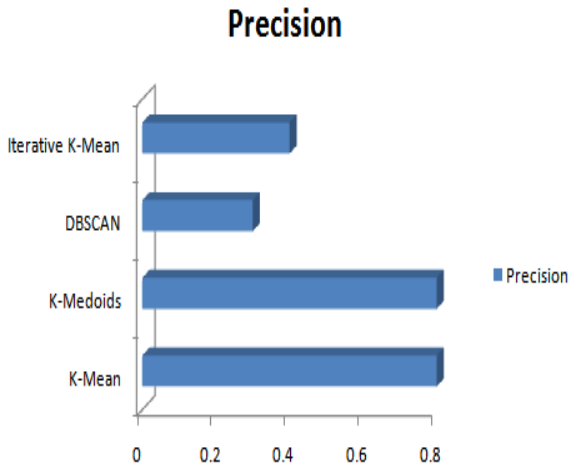

Figure 10: Precision graph w.r.t proposed algorithms

Figure 5.20 is for Precision graph. The comparison of proposed algorithms like Iterative k-mean, DBSCAN, K-medoids and K-mean is done in the above graph. It can be seen that DBSCAN has less value for precision as compare to other algorithms.

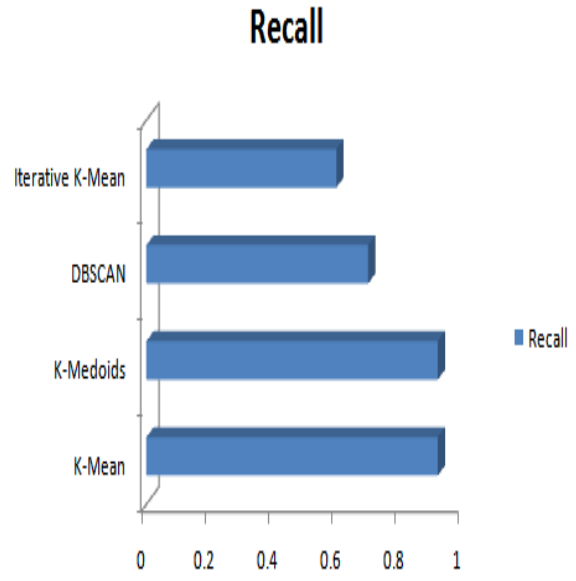

Figure 11: Recall rate graph w.r.t proposed algorithms

Figure 5.21 shows the Recall rate graph for Iterative K-mean, DBSCAN, K-medoids and K-mean algorithms. The value for Recall rate for K-medoids and K-mean is more as compare to Iterative k-mean and DBSCAN.

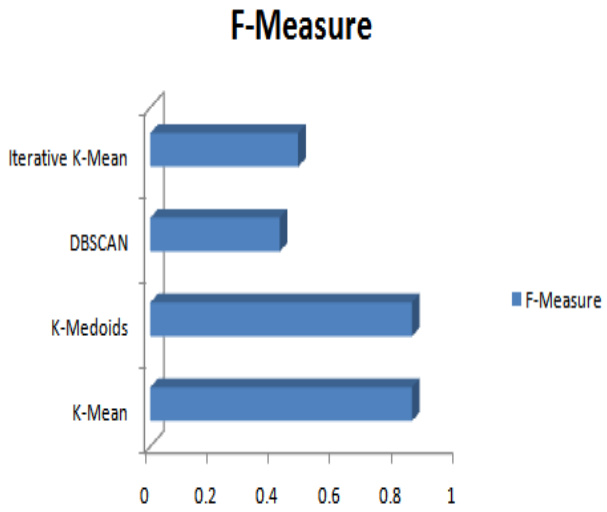

Figure 12: F-measure graph w.r.t proposed algorithms

Figure 5.22 shows the F-measure values for the proposed algorithms that are: Iterative K-mean, DBSCAN, K-medoids and K-mean. The value of F-measure for K-medoids and K-mean is more as compare to Iterative k-mean and DBSCAN. DBSCAN has less value as compare to other algorithms.

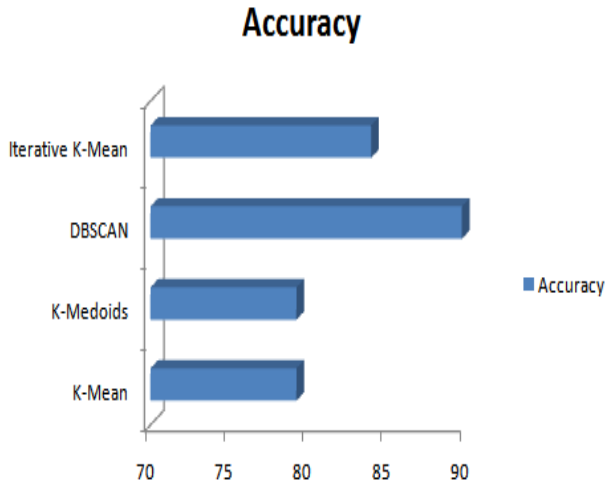

Figure 13: Accuracy graph w.r.t proposed algorithms

The graph for accuracy is shown in the above figure 5.23. The value for Accuracy for K-medoids and K-mean are less. DBSCAN has more accuracy as compare to other algorithms.

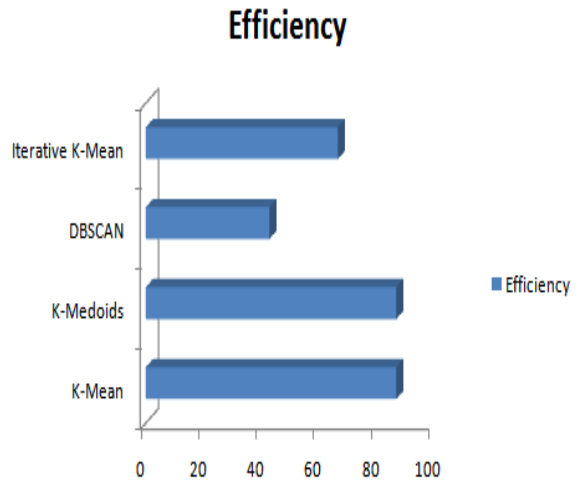

Figure 14: Efficiency graph w.r.t proposed algorithms

Figure 5.24 shows the efficiency for the proposed algorithms. Efficiency for K-medoids and K-mean are more. DBSCAN has less efficiency.

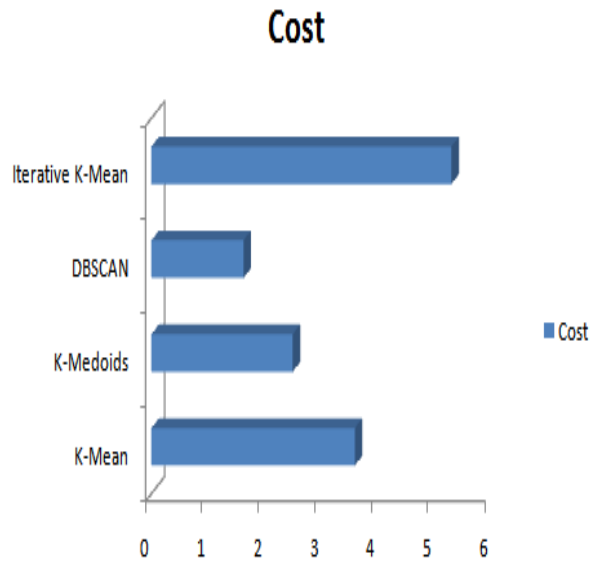

Figure 15: Cost graph w.r.t proposed algorithms

Graph for cost is shown in above figure 5.25. DBSCAN consumes less cost as compare to other algorithms. Iterative k-mean has more cost.

6. Conclusion and Future Scope

There are lot of methods for detecting the outlier in data mining. Every one focuses on trying different-different techniques to detect the outlier for better results. The proposed methods focus on detection of outlier and then compare them. In proposed work, four algorithms i.e k-mean, k-medoid, DBSCAN and Iterative K-Mean has been compared based on proposed dataset that contains only text document. From result simulation it has been found out that best algorithm is DBSCAN having accuracy of 89%.

The future work requires modifications that can make applicable for dataset that contain multiple symbols as well as text and date/time. The approach needs to be implemented on more complex dataset (dataset contains at least 20,000 records) and also focus on reduction of mean square error

References

[1] H. P. Kriegel, P. Kröger, E. Schubert and A. Zimek, "Outlier Detection in Arbitrarily Oriented Subspaces," 2012 IEEE 12th International Conference on Data Mining, Brussels, 2012, pp. 379-388.

[2] S. Wu and S. Wang, "Information-Theoretic Outlier Detection for Large-Scale Categorical Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 3, pp. 589-602, March 2013.

[3] H. N. Akouemo and R. J. Povinelli, "Time series outlier detection and imputation," 2014 IEEE PES General Meeting | Conference & Exposition, National Harbor, MD, 2014, pp. 1-5.

[4] H. Du, S. zhao and D. zhang, "Robust Local Outlier Detection," 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, 2015, pp. 116-123.

[5] H. B. M. Shashikala, R. George and K. A. Shujaee, "Outlier detection in network data using the Betweenness Centrality," SoutheastCon 2015, Fort Lauderdale, FL, 2015, pp. 1-5.

[6] S. U. Rehman, S. Asghar, S. Fong and S. Sarasvady, "DBSCAN: Past, present and future," Applications of Digital Information and Web Technologies (ICADIWT), 2014 Fifth

International Conference on the, Bangalore, 2014, pp. 232-238.

[7] Huan Yu and Wenhui Zhang, "DBSCAN data clustering algorithm for video stabilizing system," Mechatronic Sciences, Electric Engineering and Computer (MEC), Proceedings 2013 International Conference on, Shengyang, 2013, pp. 1297-1301.

[8] L. Meng'Ao, M. Dongxue, G. Songyuan and L. Shufen, "Research and Improvement of DBSCAN Cluster Algorithm," 2015 7th International Conference on Information Technology in Medicine and Education (ITME), Huangshan, 2015, pp. 537-540.

[9] Y. Yang, B. Lian, L. Li, C. Chen and P. Li, "DBSCAN Clustering Algorithm Applied to Identify Suspicious Financial Transactions," Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2014 International Conference on, Shanghai, 2014, pp. 60-65.

[10] Smiti and Z. Elouedi, "Dynamic DBSCAN-GM clustering algorithm," Computational Intelligence and Informatics (CINTI), 2015 16th IEEE International Symposium on, Budapest, 2015, pp. 311-316.

[11] Y. Jiang and J. Zhang, "Parallel K-Medoids clustering algorithm based on Hadoop," Software Engineering and Service Science (ICSESS), 2014 5th IEEE International Conference on, Beijing, 2014, pp. 649-652.

[12] E. C. de Assis and R. M. C. R. de Souza, "A K-medoids clustering algorithm for mixed feature-type symbolic data," Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on, Anchorage, AK, 2011, pp. 527-531.

[13] Ying-ting Zhu, Fu-zhang Wang, Xing-hua Shan and Xiao-yanLv, "K-medoids clustering based on MapReduce and optimal search of medoids," Computer Science & Education (ICCSE), 2014 9th International Conference on, Vancouver, BC, 2014, pp. 573-577.

[14] W. Aljoby and K. Alenezi, "Parallelization of K-medoid clustering algorithm," Information and Communication Technology for the Muslim World (ICT4M), 2013 5th International Conference on, Rabat, 2013, pp. 1-4.

[15] L. Cao, D. Yang, Q. Wang, Y. Yu, J. Wang and E. A. Rundensteiner, "Scalable distance-based outlier detection over high-volume data streams," 2014 IEEE 30th International Conference on Data Engineering, Chicago, IL, 2014, pp. 76-87.

[16] Paliwal P. et al., "Enhanced DBSCAN Outlier Detection", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013.

[17] Parimala M. et al., "A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases", International Journal of Advanced Science and Technology, Vol. 31, June, 2011.

[18] Pratap R. et al., "An Efficient Density based Improved K- Medoids Clustering algorithm", International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[19] Rafsanjani M. et al., "A survey of hierarchical clustering algorithms", The Journal of Mathematics and Computer Science Vol .5, No.3, (2012) pp. 229-240.

[20] Rajagopal Dr.Sankar, "CUSTOMER DATA CLUSTERING USING DATA MINING TECHNIQUE", International Journal of Database Management Systems Vol.3, No.4, November 2011.

[21] Ramageri B. et al., "DATA MINING TECHNIQUES AND APPLICATIONS", Indian Journal of Computer Science and Engineering Vol. 1, No. 4, pp.301-305.

[22] Ramakrishnan M. et al., "Modified K-Means Algorithm for Effective Clustering of Categorical Data Sets", International Journal of Computer Applications (0975 – 8887) Volume 89 – No.7, March 2014.

[23] Ramaswamy S. et al., "Efficient Algorithms for Mining Outliers from Large Data Sets", ACM SIGMOD Record, Volume 29 Issue 2, June 2000, pp. 427-438.

[24] http://archive.ics.uci.edu/ml/

[25] Kavita, Pallavi Bedi, "Clustering of Categorized Text Data Using Cobweb Algorithm", International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online) Vol. 3, Issue 3, pp.249-254.

[26] Ms. S. Prabha et al., "Analysis of Different Clustering Techniques in Data and Text Mining", International Journal of Computer Science Engineering, ISSN: 2319-7323 Vol. 3 No.02 Mar 2014, pp.107-116.