I. J. Computer Network and Information Security, **2017, 10, 57-65**
Published Online October 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2017.10.07

Performance Analysis of Improved Clustering Algorithm on Real and Synthetic Data

Mr. Anand Khandare
Department of CSE, SGB Amravati University Amravati, India
E-mail: anand.khandare1983@gmail.com

Dr. A. S. Alvi
Department of CSE, PRMIT &R, Badnera, Amravati, India
E-mail: abrar_alvi@rediffmail.com

Received: 15 June 2017; Accepted: 10 August 2017; Published: 08 October 2017

Abstract—Clustering is an important technique in data mining to partition the data objects into clusters. It is a way to generate groups from the data objects. Different data clustering methods or algorithms are discussed in the various literature. Some of these are efficient while some are inefficient for large data. The k-means, Partition Around Method (PAM) or k-medoids, hierarchical and DBSCAN are various clustering algorithms. The k-means algorithm is more popular than the other algorithms used to partition data into k clusters. For this algorithm, k should be provided explicitly. Also, initial means are taken randomly but this may generate clusters with poor quality. This paper is a study and implementation of an improved clustering algorithm which automatically predicts the value of k and uses a new technique to take initial means. The performance analysis of the improved algorithm and other algorithms by using real and dummy datasets is presented in this paper. To measure the performance of algorithms, this paper uses running time of algorithms and various cluster validity measures. Cluster validity measures include sum squared error, silhouette score, compactness, separation, Dunn index and DB index. Also, the k predicted by the improved algorithm is compared with optimal k suggested by elbow method. It is found that both values of k are almost similar. Most of the values of validity measures for the improved algorithm are found to be optimal.

Index Terms—Data mining, Clustering Algorithm, Validity Measure, Run time, Optimal Clusters.

I. INTRODUCTION

In the world of science &technology and the internet, daily data increases by units of terabytes. It is manually difficult to analyze and understand the hidden trends from this low to high dimensional data. Data mining is one of the ways to do so. This analysis may be from different thoughts or perspectives to summarize data into useful information. Data mining summarizes large data from different angles and categories.Then relates it to some current, past or future trends. It is observed from the literature that in a variety of areas and applications, the clustering algorithms are very popularly used [1]. Clustering accepts data sets that contain a large number of data items and produces groups of similar data objects. While forming the groups, the labels are not defined. Therefore, clustering belongs to unsupervised learning type. The best property of data clustering methods compared to other types of data mining is that it is used to manage the changes and identify the most useful features to separate one formed group from the other. Clustering can be used in real life areas such as psychology, biology, image processing and analyzing, economics, pattern recognition, bioinformatics, weather forecasting, etc. This paper has studied and implemented various standard and improved algorithms [1]. It has then simulated these algorithms on ten real datasets such as iris, salaries, wholesale, liver and university data sets and other two synthetic datasets [15]. The performance of improved and existing clustering algorithms is measured with more than five measures which are not done in any of the surveyed papers as yet.

The organization of this paper is as follows: Section I covers introduction, Section II presents a brief survey of the various literature. The third section covers standard clustering algorithms. In the fourth section, implementation and the results are discussed. In the last section, the conclusion and references are given.

II. RELATED WORK

Clustering algorithms are used in various domains such as the e-commerce, bioinformatics, image segmentation, speech recognition, financial analysis and fraud detection [1]. This paper presents a survey of various concepts and algorithms related to clustering. It has also designed improved k-means with some modification in finding k as well as initial centroid selection. Authors of the paper [2] present a brief summary of algorithms used to cluster the datasets from ranges of fields and applications. The clustering results and evaluation measures are presented in the paper. The k-means algorithm is one of the most well-known clustering algorithms. However, the processing performance of this algorithm can be degraded when it has to deal with big data. Authors suggest that a parallel algorithm with HADOOP can handle big and high dimensional data [3]. The paper presents two methods to improve the existing parallel version of the algorithm. First is the distance measure strategy and second is initial centroids selection strategy to minimize processing speed and increase stability. Paper [4] introduces two accelerated clustering algorithms using estimated subsample size and the novel stopping criterion. Authors in the paper [5] present a systematic study of k-means-based consensus clustering algorithm, identify necessary and sufficient conditions for the algorithms on both pure and noisy datasets.

The paper [6] presents efficient clustering algorithm by combining cluster aggregation with spectral analysis technique to improve cluster quality and efficiency. Authors in the paper [7] propose the modified k-means algorithm and then apply it on emotional intelligence data sets. This analysis is then used for decision making. A thorough survey of clustering algorithms and their related concepts are presented in paper [8]. Also, it focuses on some clustering algorithms that are the best for big data from the theoretical and empirical point of view. There are various cluster validity indices based on symmetry features. These are DB, DI, GDI, I, XB index, FS, K, and SV indexes [9]. Authors also suggest that an incorporation of the property of symmetry will improve the capabilities of these indices. The paper [10] has studied literature on improved k-means algorithms and presents the shortcomings and the scope where algorithms can be enhanced further. It also discusses the measures for distance, validity, stability as well as the algorithms for initial centroids selection to decide the value of k with minimum outliers.

An enhanced moving k-means is designed from the concepts of moving clustering algorithms [11] by some modifications in it. The authors in the paper [12] propose two novel enhanced algorithms such as geometric progressive fuzzy c-means and minimum sample estimate random fuzzy c-means by using some statistical techniques. This is to compute the size subsamples. To prevent uniform effect, paper [13] proposes concepts of multi-center clustering where multiple centers are used to represent the single cluster. It also proposes three subtypes of this algorithm using the global and multicenter approach. In the paper [14], authors present new centroids initialization approach to improving the basic k-means algorithm with high-quality clusters. Authors in papers [22-40] have tried to improve the clustering algorithms which are used in various domains like networking and biometrics. However, these algorithms can be improved further.

III. Clustering Algorithms

The goal of data clustering is to recover the appropriate number of clusters from the data sets. This is a challenging task in unsupervised learning for large data. Based on the working, clustering is divided into partitioning, hierarchical, density and much more. This section will discuss the working of existing clustering algorithms from all the above types.

A. K-means Clustering

The k-means need the data and value of k as an input and produces k clusters as an output [44]. The working of k-means is given as follows:

Input: Data, k
Output: k-clusters

1. Read dataset and randomly choose k initial means from data.
2. Find the distance between data objects and the initial mean.
3. Assign data object to cluster based on the minimum distance.
4. Find new means from clusters and repeat step2 to step 4
5. Stop when there is no change in clusters.

B. K-medoids Clustering

K-medoids is also known as Partition Around Medoids (PAM) because it uses medoids instead of mean [45]. The working of k-medoids is given as follows:

Input: Data, k
Output: k-clusters

1. Read data sets and choose k initial medoids.
2. Find the distance between data objects and initial medoids.
3. Assign data objects to closer medoids.
4. Check if any other data object is better medoids. If yes, change medoids and repeat the steps 2 to 4.
5. Stop when all the data objects are in clusters.

C. Hierarchical Clustering

The hierarchical clustering builds the hierarchy of clusters using dendrogram [46]. The working of k-medoids is given as follows:

Input: Data
Output: clusters

1. Read data objects and initially consider only one object in a single cluster.
2. Merge the objects based on minimum distance.
3. Repeat step 2 until all data objects are in a single cluster.

D. DBSCAN Clustering

DBSCAN is a density-based clustering algorithm to cluster the data objects with neighbor data objects [43].

Input: Datasets, eps, min-pts

Output: k clusters

1. Read data sets and select arbitrary starting object.
2. Find the neighborhood of this object using eps distance.
3. If there are sufficient neighborhoods around this object then clustering process will start and the object is marked as visited.
4. Otherwise, this data object is marked as noise data
5. If this object is found as a part of the cluster, its neighbors in the radius are also a part of the cluster.
6. Repeat the above procedure for all eps neighborhood objects and the objects are then marked as visited.

IV. Improved Clustering Algorithm

Standard k-means require the k value and they also select initial centroids randomly. This leads to bad quality clusters. This improved k-means automatically decides the value of k. It also calculates the initial centroids using arithmetic mean method [1]. Detailed steps and advantages of this algorithm are presented in the paper [1].

Input: Dataset
Output: k clusters

1. Read data objects and find a number of digits in the individual data objects.
2. Find a range of input data objects.
3. Calculate the difference of these above parameters.
4. Use this difference as the value of k.
5. Divide data objects into sub-array and split using maximum n/k elements into k initial clusters.
6. Use these as initial clusters and find the distance.
7. Check the distance of objects and decide whether it moves or it does not move in the other clusters.
8. Repeat step 6 and 7 till there is no change in the clusters.

This algorithm uses a novel technique to find the number of clusters using a range of inputs and number of digits in the input. This improved clustering algorithm requires no value of the number of clusters and also selects initial centroids using the split method. Hence, clusters produced by this algorithm are optimal and good quality clusters. The complexity of the algorithm is less. Hence, this algorithm is efficient. This algorithm predicts the appropriate value of k in the given data sets.

V. Results Analysis

A. Cluster validity and Datasets

From the literature, it observed that there is score to improve clustering algorithms. But from therotical perspective only, it is not suufient to categorize algorithms into efficient and non efficient category. The main focus of this work is to investigate the strong features of improved algorithm over the for existing algorithms. Hene the contribution of this work is summarizing features of algorithms in therotical and practical point of view. For the experiements, this paper implemented improved and existing clustering algorithm using python and R programming and used various real and symthetic data sets.

To evaluate the performance of improved clustering algorithm, this paper uses more than ten performance matrices. So far, none of the papers have made use of these many matrices. Maximum validity indexes are based on compactness and separation of clusters. The details of these validity indexes are as follows:

- **Within SS:** Sum Squared error within individual cluster measures the average distance between centroids and data objects in the cluster. Better clusters should have a lower value of within SS.
- **Between SS:** Sum squared error between clusters measures average distance between clusters. The better clusters should have a higher value of Between SS.
- **Total SS**: Total Sum Squared error indicates a total deviation in the clusters. It is the sum of Within SS and Between SS. It should be as low as possible.
- **Accuracy:** Clustering accuracy measures how correctly data objects are clustered with minimum deviation.
- **Silhouette index:** It measures the consistency within clusters by finding how well each object lies within its cluster. Its value is in between -1 to 1. More consistency clusters will yield a higher value.
- **Compactness:** Compactness measures the average distance between data objects in clusters. A lower value indicates more compact cluster.
- **Separation:** This measures the average distance between the pair of clusters. A Higher value indicates well-separated clusters.
- **Dunn Index:** It measures separation over the compactness. A higher value indicates well-separated clusters.
- **Time Complexity:** It measures the time required to cluster data objects in the given data sets.
- **Rand Index**: It measures the similarity between the two clustering results. Its value is 0 or 1. A higher value indicates that all data objects are correctly clustered.

All the validations cannot be applied to all the clustering algorithms. This paper uses these validity measures to evaluate the performance of improved clustering algorithm and some measures on existing algorithms. This paper makes use of real world data sets [15] [16] from UCI and Kdnuggets data sets. Also, some synthetic data sets are used. The details of these datasets are given Table 1:

Table 1. Data Sets Used

Number of variables	**Number of Observations**	**Datasets**	**SN**
8	100	iris	1
33	593	Census	2
7	109469	Salaries	3
3	44	Air Passenger	4
14	2198	University Ranking	5
8	440	Wholesale	6
33	5820	Students	7
11	582	Liver	8
6	50	Life cycle	9
10	4897	Shanghai	10
11	2603	Times data	11

B. *Clustering Results Analysis*

All the existing and improved algorithms are applied on datasets and the result is given in following sections. For the existing algorithms, data must be pre-processed. Then only on selected dimensions, these algorithms can be applied. From the experiments, it is found that the existing algorithms worked efficiently on preprocessed data.

1. K-means

Table 2. K-means Results

Dunn Index	**Silhouette Score**	**Accuracy (%)**	**Total-SS**	**Between-SS**	**Within-SS**	**Data sets**
0.09	0.39	91 (only 2 dimensions)	334.6	307.4	27.2	iris
0.0087	1.64	57.05	109866348686	62686224455	47180124231	wholesale
Not Working						Salary

The k-means clustering result on wholesale data is as shown in the Fig1.

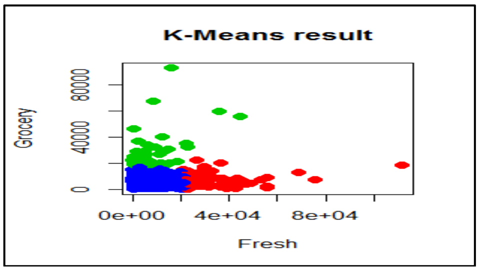

Fig.1. K-Means Clustering

Outliers the data objects having more distance from its center point. From the figure, it is observed that more outliers are present in clusters. Also, they are less separated clusters. For the large data (salary) k-means doesn't work properly. The k-means clustering algorithm doesn't work for large data. Because of this, the experiment in this paper applies k-means on salary data and algorithm does not work for this data. This observation is shown in Table 2.

2. K-medoids

Table 3. K-Medoids Results

Dunn Index	Silhouette Score	Accuracy (%)	DB index	diameter	Separation	Data sets
0.1	0.61	89.12 (only 2 dimensions)	0.60	2.0	0.20	iris

The clusters of the k-medoids clustering on iris data are as shown in Fig 2. The percentage of outliers in k-means and k-medoids are similar. Also, the Dunn score and Silhoutte score of both algorithms are getting aproximetly same for some selected data sets.

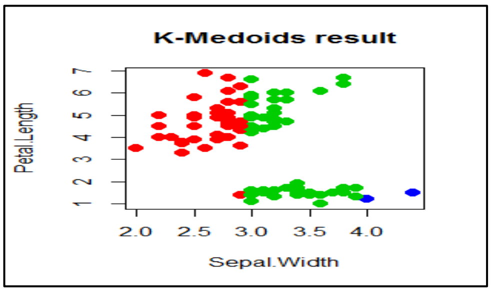

Fig.2. K-Medoids Clustering

From the Table 2 and Table 3, it is observed that the performance of both the algorithms is same. The accuracy of k-medoids is more than k-means clustering algorithm. Also, the quality of k-medoids is higher than k-means clustering. All the quality scores of k-medoids clustering algorithm are higher thank-means algorithm.

3. Hierarchical Clustering

Table 4. Hierarchical Clustering

Dunn Index	Silhouette Score	Accuracy (%)	DB index	Datasets
0.089	0.59	88 (only 2 dimensions)	0.60	Iris

Fig. 3 shows the results of hierarchical clustering algorithms

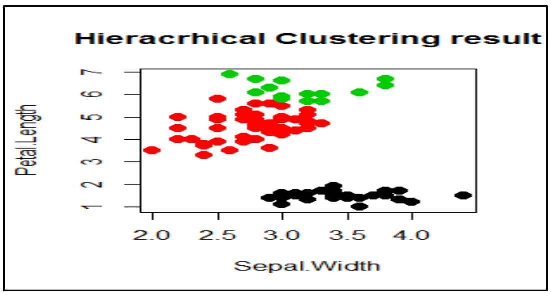

Fig.3. Hierarchical Clustering

4. DBSCAN Clustering

Table 5. DBSCAN Clustering

Dunn Index	Silhouette Score	Accuracy (%)	DB index	Data sets
0.50	0.60	90	0.33	Iris

Fig. 4 shows the results of DBSCAN clustering algorithm.

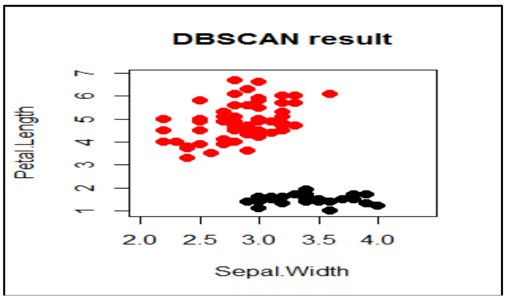

Fig.4. DBSCAN Clustering

5. Improved Clustering Algorithm

Improved clustering algorithm is applied to more than ten real data sets and two dummy data sets. Clustering results and performance analysis are discussed in this section. Table 6 shows the result analysis of this improved clustering algorithm. K-means does not work for salary data (large data) whereas improved clustering algorithm works for this data. For the experiments, this paper generates two synthetic or dummy data sets. It is observed that proposed algorithm works well with respect to quality and efficiency for these two synthetic data sets. The detailed factual analysis is shown in following Table 6.

Table 6. Improved Clustering Algorithm Results

Dunn Index	**Silhouette Score**	**Accuracy**	**Elbow method k**	**Predicted k**	**Data Sets**
0.11	0.32	95.1	8	9	Iris
0.04	0.45	95.66	7	7	Air Passenger
0.04	0.56	96.0	5	5	Students
0.06	0.25	85.96	11	12	University
0.04	0.2	95.42	54	56	Census
0.2	0.27	91.2	52	54	Wholesale
0.04	0.79	99.84	30	32	Life Cycle
0.03	0.3	92.59	36	39	Liver
0.03	0.32	95.47	28	29	Shanghai
0.04	0.49	90.85	20	21	Times
0.03	0.46	99.94	8	10	Annual Crime
0.002	0.35	95.5	58	60	Salaries
0.53	1.04	59.22	3	2	Dummy Data1
0.63	1.05	65.8	3	2	Dummy Data1

From the Table 6, it is observed that if a number of records in data set to increase, accuracy also increase for most of the data sets. This scenario is as shown in Fig 5.

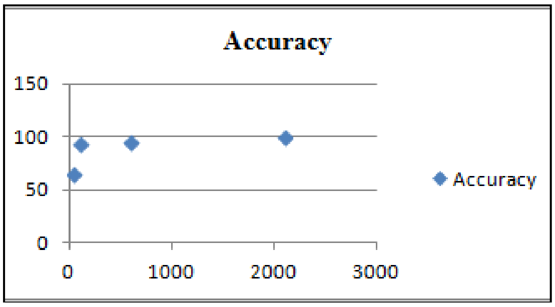

Fig.5. Accuracy of Algorithms

Fig. 6 shows the comparative analysis of algorithms with respect to accuracy. From the graph, it is observed that accuracy of our proposed algorithm is more than the existing algorithm clustering on the same data sets. The accuracy of improved clustering algorithm is increased by at least 10 percent. Approximately, the accuracy of four existing clustering is less 92 % where as the accuracy of improved clustering algorithm is more than 95%. For the most of the data large data sets, accuracy is greater than 90 %.

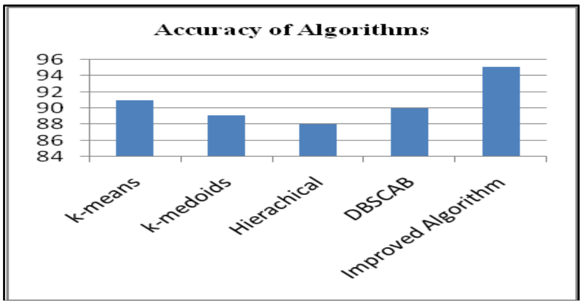

Fig.6. Accuracy of Algorithms

Fig. 7 shows the comparative analysis of the algorithms with respect to separation over compactness. The Dunn index is used to measure the quality of cluster with respect to separation over the compactness. The value of Dunn score should be high for good quality clusters. From the graph, it is observed that improved algorithm is more compact and separated than the k-means, k-medoids and the hierarchical algorithms. For some data sets, clusters produced by DBSCAN are well separated and compact because its Dunn score is high than existing and improved clustering algorithms.

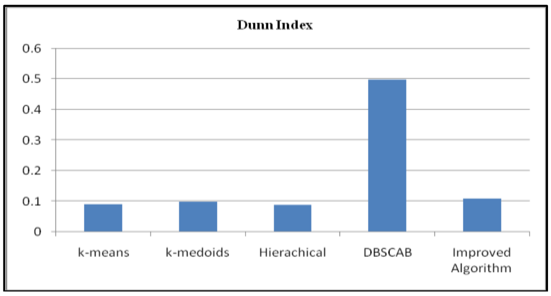

Fig.7. Quality of Clustering Algorithms

Fig.8 shows the silhouette score of the algorithms. The silhouette score measures the cohesion within the clusters.The higher value of silhouette score indicates the better clusters. From the graph, it observed that consistency of k-means and the improved algorithm is similar. For some data sets, the silhouette score for k-means and the improved clustering algorithm is coming same. For some data sets, silhouette score for the improved algorithm is little more.

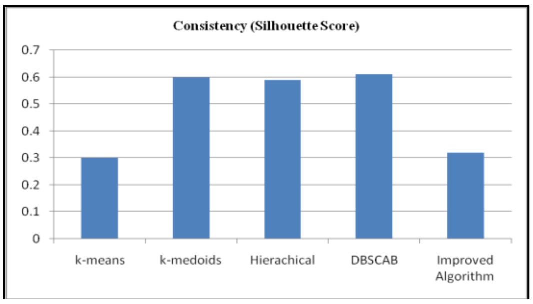

Fig.8. Consistency of Clustering Algorithms

Figure 9 shows the k predicted by improved clustering algorithm and the optimal value of k by elbow method. It is observed that both values k are approximately similar. The elbow method is used to find the value of k. Buts this method is more complex. Improved clustering algorithm is predicting k value for given data sets. This paper compared the k value predicted by the improved algorithm and k predicted by elbow method. And found that both the k values are approximately same. In some cases, k predicted by the algorithm is slightly higher than elbow method.

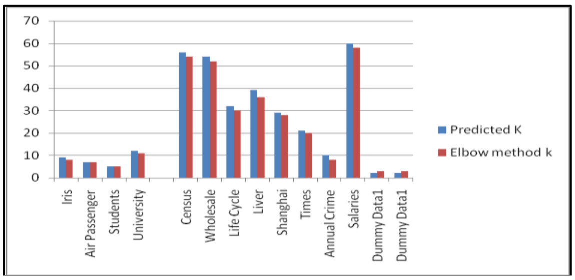

Fig.9. Predicted vs. Optimal k

Along with the Dunn, DB, and Silhouette score, the performance of algorithms is measured using running time required for executing the algorithms on the given datasets, Total sum squared error (SSE) in the clusters, compactness of the cluster and separation within clusters are produced by improved clustering. The values of these measures are tabulated in Table VII. From these values, it is observed that the values of some measures are optimal for improved clustering algorithm. Almost for all data sets, these values are getting better for improved clustering algorithm.

Table 7. Improved Clustering

Separation	**Compactness**	**Total-SS**	**Running Time**	**Datasets**
0.18	1.62	546.31	0.0003683	Iris
6.71	161.1	2308592.08	0.0001321	Air Passenger
3.32	1165.16	16428486540.67	0.0006263	Students
61.04	1105.93	1543475878.15	0.0001472	University
34087.9	9143101.64	4474433793416706.0	5.10	Census
897.29	56420.73	57595857524.96	0.0006204	Wholesale
11.67	318.5	48125221.89	0.0001947	Life Cycle
6.0	2309.76	102497725.13	3.60	Liver
1.8	662.13	907508930.44	0.0005696	Shanghai
83.7	258245.58	827522232448.56	0.0002363	Times
254.12	156701.20	6.492741929243135e+16	0.0006191	Annual Crime
365.67	160618.21	9410694270058.19	0.0007558	Salaries
1.08	2.03	315.96	0.0004970	Dummy Data1
0.63	1.76	200.91	0.0001620	Dummy Data1

The comparison of running time of the existing and improved algorithms is given the Fig.10. It is observed that running time of the improved algorithm is less. The improved algorithm is faster than existing clustering algorithms because running time required for improved algorithm is 50 5 less than existing algorithms. For the almost all data sets used,

Improved algorithm is taking less time to run than existing algorithms.

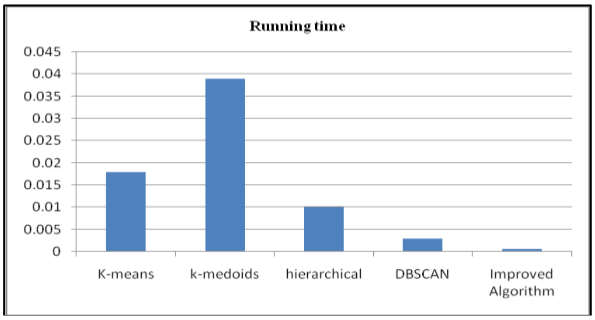

Fig.10. Running time of Algorithms

VI. Conclusion and Future Work

Clustering algorithms are algorithms used to cluster data from low to high dimensions. Also, they are used in various fields. This paper has studied and presented working of these algorithms. Also, has analyzed and presented a comparative analysis of the results of existing and improved clustering algorithms. This improved algorithm uses a novel technique to find the value of k and the initial centroids. The performance of these algorithms is measured using more than five validation measures and it is found that the performance of improved clustering algorithm is better than the other algorithms discussed in section III. For the performance analysis, 10 different real datasets are used from UCI and Kdnuggets machine learning website. Also, two generated datasets are used. Clustering performance is compared using more than 5 validation measures such as Sum Squared Error, silhouette score, Dunn index, DB index, compactness, and separation. For the improved clustering algorithms, the values of most of these measures are getting optimal over the other algorithms. Future work of this paper is to implement this algorithm for really challenging data sets and identify the meaning of full results.

References

[1] Mr. Anand Khandare, Dr. A.S. Alvi, “Clustering Algorithms: Experiment and Improvements”, IRSCNS, Springer, LNNS, July 2016.

[2] Rui Xu, Donald Wunsch II, “Survey of Clustering Algorithms, “IEEE transactions on neural networks, vol. 16, no. 3, May 2005.

[3] Qing Liao, Fan Yang, JingmingZhao,”An Improved parallel K-means Clustering Algorithm with MapReduce”, ICCT, pp 764-768, 2013.

[4] Jonathon K. Parker, Lawrence O. Hall, “Accelerating Fuzzy-C Means Using an Estimated Subsample Size”, IEEE trans on fuzzy systems, vol. 22, no. 5, Oct 2014.

[5] JunjieWu, Hongfu Liu, Hui Xiong, Jie Cao, Jian Chen,”K-Means-Based Consensus Clustering: A Unified View “, IEEE Transaction on knowledge and data engineering, vol. 27, no. 1, 2015.

[6] Mr. Anand Khandare, Dr. A.S. Alvi, “Efficient Clustering Algorithm with Improved Clusters Quality”, IOSR Journal of Computer Engineering, vol-18, pp. 15-19, Nov.-Dec. 2016.

[7] Mr. Anand D.Khandare, “Modified K-means Algorithm for Emotional Intelligence Mining”, International Conference on Computer Communication and Informatics (ICCCI -2015), Jan. 08 – 10, 2015.

[8] Adil Fahad1, Najlaa Alshatri1, Zahir Tari1, Abdullah Alamri, Ibrahim Khalil1, Albert Y. Zomaya, SebtiFoufou, AbdelazizBoura,” A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis “, IEEE Transaction on emerging topics in computing, 2014.

[9] Sriparna Saha, Sanghamitra Bandyopadhyay, “Performance Evaluation of Some Symmetry-Based Cluster Validity Indexes “, IEEE Transaction on systems, man, and cybernetics—part c: applications and reviews, vol. 39, no. 4, 2009.

[10] Mr. Anand Khandare, Dr. A.S. Alvi, “ Survey of Improved k-means Clustering Algorithms: Improvements, Shortcomings, and Scope for Further Enhancement and Scalability, Information Systems Design and Intelligent Applications, Springer, AISC, Pages 495-503,2016.

[11] Fasahat Ullah Siddiqui, Nor Ashidi Mat Isa, “Enhanced Moving K-Means (EMKM) Algorithm for Image Segmentation “, IEEE Transactions on Consumer Electronics, Vol. 57, No. 2, 2011.

[12] Jonathon K. Parker, and Lawrence O. Hall, “Accelerating fuzzy-c means using an estimated subsample size”, IEEE Trans. on the fuzzy system, vo. 22, no. 5, 2014.

[13] Jiye Liang, Liang Bai, Chuangyin Dang, and Fuyuan Cao, “The k-means-type algorithms versus imbalanced data distributions”, IEEE Trans. on fuzzy systems, vol. 20, no. 4, 2012.

[14] Wei Zhong, GulsahAltun, Robert Harrison, Phang C. Tai, and Yi Pan, “Improved K-Means Clustering Algorithm for Exploring Local Protein Sequence Motifs Representing Common Structural Property”, IEEE transactions on nanoscience, vol. 4, no. 3, Sept 2005.

[15] https://archive.ics.uci.edu/ml/datasets.html.

[16] http://www.kdnuggets.com/datasets/index.html.

[17] Weiguo Sheng, Shengyong Chen, Mengmeng Sheng, Gang Xiao, Jiafa Mao Yujun Zheng, "Adaptive Multi-Subpopulation Competition. Multi-Niche Crowding based Memetic Algorithm for Automatic Data Clustering", IEEE transactions on evolutionary computation, 2016.

[18] Qiuhong Li, PengWang, WeiWangHao Hu, Zhongsheng Li, Junxian Li, "An Efficient K-means Clustering Algorithm on MapReduce", DASFAA, LNCS 8421, pp. 357–371, 2014.

[19] https://www.rstudio.com.

[20] https://cran.r-project.org.

[21] *Qingshan Jiang, YanpingZhang, LifeiChen,”An Initialization Method for Subspace Clustering Algorithm”, .J. Intelligent Systems and Applications,* 2011, 3, 54-61, MECS, 2011.

[22] *Mohammed El Agha, Wesam M. Ashour, “*Efficient and Fast Initialization Algorithm for K-means Clustering*”, I.J. Intelligent Systems and Applications*, 2012, 1, 21-31, MECS.

[23] *Shashank Sharma, Megha Goel, Prabhjot Kaur,*

"Performance Comparison of Various Robust Data Clustering Algorithms", I.J. Intelligent Systems and Applications, 2013, 07, 63-71, *MECS.*

[24] B.K. Tripathy, Akash Goyal, Rahul Chowdhury, Patra AnupamSourav**,"**MMeMeR: An Algorithm for Clustering Heterogeneous Data using Rough Set Theory**",** I.J. Intelligent Systems and Applications, 2017, 8, 25-33, MECS.

[25] *Long Nguyen Hung, Thuy Nguyen Thi Thu, Giap Cu Nguyen, "*An Efficient Algorithm in Mining Frequent Itemsets with Weights over Data Stream Using Tree Data Structure*",* I.J. Intelligent Systems and Applications, 2015, 12, 23-31, MECS.

[26] *Zhengbing Hu, Yevgeniy V. Bodyanskiy, Oleksii K. Tyshchenko, Viktoriia O. Samitova, "*Fuzzy Clustering Data Given in the Ordinal Scale*",* I.J. Intelligent Systems and Applications, 2017, 1, 67-74, MECS.

[27] Manju Mam, Leena G, N S Saxena, "Improved K-means Clustering based Distribution Planning on a Geographical Network", I.J. Intelligent Systems and Applications, 2017, 4, 69-75, MECS.

[28] Sharfuddin Mahmood, Mohammad Saiedur Rahaman, Dr. Dip Nandi, Mashiour Rahman, "A Proposed Modification of K-Means Algorithm", I.J. Modern Education and Computer Science, 2015, 6, 37-42, MECS.

[29] Muhammad Ali Masood, M. N. A. Khan,"Clustering Techniques in Bioinformatics ", I.J. Modern Education and Computer Science, 2015, 1, 38-46, MECS.

[30] JinzhuHu, ChunxiuXiong, JiangboShu, XingZhou, Jun Zhu, "An Improved Text Clustering Method based on Hybrid Model", I.J.Modern Education and Computer Science, 2009, 1, 35-44, MECS.

[31] Prachi, Shikha Sharma, "Energy Efficient Clustering Protocol for Sensor Network", I. J. Computer Network and Information Security, 2016, 12, 59-66, MECS.

[32] *Sukhkirandeep Kaur, RoohieNaaz Mir,"* Clustering in Wireless Sensor Networks- A Survey*",* I. J. Computer Network and Information Security, 2016, 6, 38-51, MECS.

[33] *Mai Abdrabo, Mohammed Elmogy, GhadaEltaweel, Sherif Barakat,"* Enhancing Big Data Value Using Knowledge Discovery Techniques", I.J. Information Technology and Computer Science, 2016, 8, 1-12, MECS.

[34] Wei Zhong, G. Altun, R. Harrison, "Improved K-means clustering algorithm for exploring local protein sequence motifs representing common structural property", IEEE Transactions on NanoBioscience, Pages: 255 - 265, DOI: 10.1109/TNB.2005.853667, 2015.

[35] Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, "Semi-Supervised Linear Discriminant Clustering", IEEE Transactions on Cybernetics, Pages: 989 - 1000, DOI: 10.1109/TCYB.2013.2278466,2013.

[36] Aleta C. Fabregas, Bobby D. Gerardo, Bartolome T. TanguiligIII,"Enhanced Initial Centroids for K-means Algorithm ", I.J. Information Technology and Computer Science, 2017, 1, 26-33, MECS.

[37] Dazhao Cheng, JiaRao, YanfeiGuo,"Improving Performance of Heterogeneous MapReduce Clusters with Adaptive Task Tuning".

[38] IEEE Transactions on Parallel and Distributed Systems, Pages: 774 - 786, DOI: 10.1109/TPDS.2016.2594765.

[39] Orhan Kislal, Piotr Berman, Mahmut Kandemir, "Improving the performance of k-means clustering through computation skipping and data locality optimizations", Proceedings of the 9th conference on Computing Frontiers, ACM, 2012.

[40] JeyhunKarimov. Author links open the author workspace.MuratOzbayoglu, "Clustering Quality Improvement of k-means Using a Hybrid Evolutionary Model ", DOI: doi.org/10.1016/j.procs.2015.09.143,Elsevier, 2015.

[41] SalimaOuadfel, SouhamMeshoul, "Handling Fuzzy Image Clustering with a Modified ABC Algorithm ", I.J. Intelligent Systems and Applications, 2012, 12, 65-74, MECS.

[42] Vishwambhar Pathak, Dr. Praveen Dhyani, Dr. Prabhat Mahanti, "Autonomous Image Segmentation using Density Adaptive Dendritic Cell Algorithm ", I.J. Image, Graphics and Signal Processing, 2013, 10, 26-35, MECS.

[43] JunhaoGan, Yufei Tao, " DBSCAN Revisited: Mis-Claim, Un-Fixability, and Approximation", Proceeding SIGMOD '15 Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data Pages 519-530, 2015.

[44] Chen Li, YanfengZhang, MinghaiJiao, Ge Yu, "Mux-Kmeans: multiplex means for clustering large-scale data set ", ScienceCloud '14: Proceedings of the 5th ACM workshop on Scientific cloud computing June 2014.

[45] P. S. Bishnu, V. Bhattacherjee, "Application of K-Medoids with Kd-Tree for Software Fault Prediction ", ACM SIGSOFT Software Engineering Notes: Volume 36 Issue 2, March 2011.

[46] Chen Jin, ZhengzhangChen, William Hendrix, Ankit Agrawal, Alok Choudhary, "Improved Hierarchical Clustering for Face Images in Videos: Integrating positional and temporal information with HAC ", April 1, 2014.

[47] Jian YuHoukuan HuangShengfengTian, "Cluster Validity and Stability of Clustering Algorithms ", LNCS 3138, pp. 957–965, 2004.Springer-Verlag Berlin Heidelberg 2004.

[48] Ken-ichiFukuiMasayukiNumao, "Neighborhood-Based Smoothing of External Cluster Validity Measures ", PAKDD 2012, Part I, LNAI 7301, pp. 354–365, Springer-Verlag Berlin Heidelberg 2012.

[49] JunjieWu, Jian Chen, Hua Xiong, MingXie, "External validation measures for K-means clustering: A data distribution perspective ", DOI: doi.org/10.1016/j.eswa.2008.06.093.

[50] Hoel Le Capitaine, CarlFrelicot," A Cluster-Validity Index Combining an Overlap Measure and a Separation Measure Based on Fuzzy-Aggregation Operators", IEEE Transactions on Fuzzy Systems, Volume: 19, Issue: 3, June 2011.

Authors' Profiles

Anand Khandare has graduated from Sant Gadge Baba (SGB) Amravati University, Amravati in Computer Science and Engineering in 2005. He completed his Master's Degree from Mumbai University in Academic Year 2010-11. He is pursuing Ph.D. from Sant Gadge Baba Amravati University. Currently, he is working as an Assistant Professor at Thakur College of Engineering and Technology, Mumbai University. He has 11 years of teaching experience in the Institute. He has published more than 10 papers in international journals and conferences. He has also published C and C++ programming language books. His area of interest is machine learning and intelligent system. His interests also include web application development and mobile application development. He is a life time member of ISTE professional body.

Dr. A. S. Alvi has graduated from Sant Gadge Baba Amravati University, Amravati in Computer Science and Engineering. He got his Master's and a Ph.D. degree from the same university. Currently, he is working as a Professor in Computer Science and Engineering at PRMIT &R, Badnera, Amravati. He has more than 20 years of teaching experience. He has published more than 25 papers in international journals and conferences. His area of interest is Artificial intelligence and Algorithms. His interest also lies in Natural Language Processing. He is a Life time member of ISTE and IET professional bodies. He is also a research guide at SGB, Amravati University, Amravati.

How to cite this paper: Anand Khandare, A. S. Alvi,"Performance Analysis of Improved Clustering Algorithm on Real and Synthetic Data", International Journal of Computer Network and Information Security(IJCNIS), Vol.9, No.10, pp.57-65, 2017.DOI: 10.5815/ijcnis.2017.10.07