I.J. Intelligent Systems and Applications, **2018, 7, 48-57**
Published Online July 2018 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2018.07.05

Efficient Clustering Algorithm with Enhanced Cohesive Quality Clusters

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

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

Received: 23 June 2017; Accepted: 15 September 2017; Published: 08 July 2018

***Abstract*—**Analyzing data is a challenging task nowadays because the size of data affects results of the analysis. This is because every application can generate data of massive amount. Clustering techniques are key techniques to analyze the massive amount of data. It is a simple way to group similar type data in clusters. The key examples of clustering algorithms are k-means, k-medoids, c-means, hierarchical and DBSCAN. The k-means and DBSCAN are the scalable algorithms but again it needs to be improved because massive data hampers the performance with respect to cluster quality and efficiency of these algorithms. For these algorithms, user intervention is needed to provide appropriate parameters as an input. For these reasons, this paper presents modified and efficient clustering algorithm. This enhances cluster's quality and makes clusters more cohesive using domain knowledge, spectral analysis, and split-merge-refine techniques. Also, this algorithm takes care to minimizing empty clusters. So far no algorithm has integrated these all requirements that proposed algorithm does just as a single algorithm. It also automatically predicts the value of k and initial centroids to have minimum user intervention with the algorithm. The performance of this algorithm is compared with standard clustering algorithms on various small to large data sets. The comparison is with respect to a number of records and dimensions of data sets using clustering accuracy, running time, and various clusters validly measures. From the obtained results, it is proved that performance of proposed algorithm is increased with respect to efficiency and quality than the existing algorithms.

***Index Terms*—**Clustering, Cluster, Massive Data, k-means, Cohesive, Quality, Validity Measures.

I. INTRODUCTION

An increasing number of standard and modern applications generate the massive data with respect to more number of records and dimensions. So, analyzing this type of data is a challenge in the field of research in data clustering area. It is necessary to understand the capability of clustering algorithms before applying it on large data. Also, preprocessing of data plays a vital role in data clustering. Data clustering is simply the grouping of similar kind of data in one group and dissimilar in another group. These algorithms have following categories: Partitioned based, Hierarchical based, Density based, Grid and Model based. k-means, k-medoids, hierarchical, DBSCAN are fall in the above categories. These methods are widely used in the field of science and technology domains [29-34][41-42][48-52].The researchers have made a lot of improvements in these clustering algorithms in order to overcome the problems with respect to large data, efficiency, quality of clusters empty clusters, the optimal value of k and initial centroids. Still, there is a scope to improve these algorithms. In partitioned based clustering, the data is simply partitioned into k number of partitioned. When data is clustered into clusters in a hierarchical manner, it is known as hierarchical clustering. The clustering which is based on density regions, connectivity and boundary of data objects is known as density based clustering. When datasets are divided into a number of grids then this approach is called grid based approach. Lastly, model based is where clustering is done using some mathematical models. From the literature, this paper has identified some criteria to compare the various clustering algorithms namely the volume of data, the velocity of data and variety of data for the clustering algorithms. These three V's generally affects the performance of algorithms. From the survey, this paper has identified requirements of clustering algorithms for massive data. The requirements are as follows: Firstly, these clustering algorithms should be more efficient with respect to running time and quality of clusters. Secondly, user input parameters in algorithms must be minimum.

For above reasons, this paper proposes efficient clustering algorithm using spectral analysis, domain knowledge, and split-merge-refine approach. This enhances the efficiency, quality and minimizes empty clusters. The performance of this algorithm is checked on 10 real small to large data sets. This paper is organized as

follows: Section II summarizes literature surveys and related work from the year 1999 to 2017. Section III describes the working of some standard clustering algorithms from all the above categories. Section IV covers the proposed clustering algorithms. Section V covers experimental results analysis of standard and proposed clustering algorithms on various data sets using validity measures. Section VI covers conclusion and future work of this paper.

II. Related Work

One of the simple efficient filtering algorithms is Lloyd's k-means clustering. The implementation of this algorithm is easy because it uses k-dimensional tree data structure for data clustering. But this algorithm can be improved further [1]. The appropriate value of k and initial selection plays a vital role in improving the quality of clusters. In order to overcome these problems associated with the initial selection, the new greedy initialization method is used. These select suitable initial centroids to form more compact and separated clusters [2]. This new method takes more time to search centroids. The distance measures play a vital role in the process of clustering. New distance measure along with new clustering algorithm which is based on the k-means algorithm is known as the circular k-means [3]. This algorithm is used to cluster the vectors of directional information. Selecting the appropriate value of k is difficult for clustering. Also, checking the validity of final clusters is challenging. For these two reasons, validity measures are useful. Nine validity measures based on a property of symmetry are used to estimate the value of k and validity of clusters. These measures include Davies–Bouldin, Dunn index, Point symmetry, I, Xie–Beni, FS, K, and SV index [4]. One of the popular clustering algorithms is k-means. For the k-mean, the user needs to give the value of k and initial centroids, which affects the overall quality of clusters. Also, results of k-means may be affected by noise in data. Density based noise detection can be used to filter noise. This helps in getting more accurate results [5].

The k-means is known for the lower computational cost. The use of appropriate data plays a major role in clustering. Imbalanced data distribution definitely affects the performance of k-means clustering. This type of data always produces clusters of the same structure. This problem mainly occurs in fuzzy type k-means than hard type k-means [6]. Random initialization in k-means, yield the poor clusters. The density of data objects and k-nearest can be combined to get more accurate clusters. This combination can be used for initialization of centroids to improve the performance over the k-means clustering [7]. The tremendous growth of data generated by applications hampers the performance of clustering algorithms. Such data require new efficient k-means clustering algorithm. The k-means++ solved the problems of initial centroids but did not work well for large data. The competitive k-means [8] solved the problems. From the survey, authors [9] summarized the various points regarding clustering algorithms. Effective clustering algorithms must possess the following properties: Algorithm generates the clusters of arbitrary shapes. It should be able to handle large and high dimensional data. The algorithm must detect and remove outliers from data. Also, it should be worked on numerical as well as non-numerical types of data. The k-means and fuzzy k-means can be used for a variety of applications. The fuzzy k-means is used in Geographic Information System to cluster the Hot Spot areas [10]. In terms of machine learning, clustering algorithms are the unsupervised type of algorithms. New algorithm [11] is more efficient and produces quality clusters. But this efficient algorithm needs improvements for large data. The k-means produces compact and more separated clusters. The authors in the paper [12] present various clustering algorithms using k-means by combining both separation and compactness which helps in producing high-quality clusters. The authors have [13] surveyed about the existing clustering algorithms and analyzed the various points regarding same. The points are as follows: All validity measures cannot be applied to all clustering algorithms. The Expectation-Maximization and Fuzzy clustering are efficient clustering algorithms. All clustering algorithms produce unstable clusters. Cluster refinement is an important step in clustering to enhance the clusters quality. Fuzzy clustering and proximity can be used to cluster's the data to avoid the cluster's membership conflict [14]. The clustering methods can be widely used in computer science domains like networking and wireless. Authors in the paper [15] presents enhance k-means clustering using Dijkstra algorithm to optimize energy consumption in the wireless network. Selecting appropriate candidates for the initial centroids is essential for clustering quality and the performance. Authors in the paper [16] have proposed a hybrid evolutionary model. It has Meta heuristic methods to identify the appropriate candidates for initial centroids in *k*-means. The accuracy of k-means is also affected by the value of k because this value has to provide at the beginning of clustering. The exact value can be predicted by using backtracking method and enhanced Euclidian distance [17]. The clustering algorithm is said to be robust if it can handle noisy data and produce the good quality clusters [18].There are various types of in fuzzy c-means clustering. All these types are robust clustering algorithms because they can easily handle noisy data. Out of the all clustering algorithms, k-means, and k-medoids are widely used algorithms but are not efficient. This is because cluster's quality is not good for large data sets. Authors [19] proposed a little efficient algorithm to enhance cluster quality by using clusters aggregation and spectral analysis. Clusters validity measures are used as the fitness functions to evaluate the quality clusters which are produced. Most of the measures are data dependent because they are designed to address certain types. These measures include Dunn, Davies Bouldin, Calinski Harabasz, CS, I, and the Silhouette score [20]. Authors in the paper [21] propose a novel technique to predict the value of k and initial centroids with improved quality

clusters. Paper [22] consists survey regarding the existing and improved clustering algorithms and presents scope for further improvements.

Authors propose [27] method to optimize a number of clusters k, with minimum time complexity. This reduces the effort required for each iteration by decreasing re-clustering of data objects. The different distance measures are used to track the effect on the computational time required per iteration. Therefore, this algorithm may produce less reliable clusters of data objects [28]. This paper presents enhanced method to select k and initial centers using weighted mean. This method is better in terms of mathematical computation and reliability. The performance of the standard k-means algorithm will be affected by the selection of the initial centers and converges to local minimum problems [29]. This paper proposes a new algorithm for initialization of the k-means to converge into a better local and also to produce better clusters.

From the above literature, it is observed that very few papers are focusing on all the aspects related to enhancement in a single algorithm. Hence, this paper is integrating more than two features in the single algorithm.

III. Candidate Clustering Algorithms

As per the above literature, clustering algorithms are partition, hierarchical, density, grids and model based. In each of the above categories, some candidate clustering algorithms are studied in this section. This paper studies algorithm from various categories analyzes their strong and weak points. This is done by applying and measuring performance on various data sets from the kaggle with various validity measures. Then these algorithms can be used to compare with the proposed efficient clustering algorithm. The summary of these clustering algorithms with their basic steps and their strong and weak points are given as follows:

A. k-Means Clustering

The k-means belongs to partition based clustering. Inputs for k-means are a number of clusters and it selects initial centroids randomly. The algorithm is divided into following three steps:

1. Read data and value of k.
2. Data assignment steps using distance measures.
3. Updating centroids and clusters.
4. Forming final clusters.

The value of k should be provided at the beginning of clustering. Poor qualities clusters are generated because initial centroids are selected randomly.

B. Partition Around Medoids

Partition around medoids or PAM is partition algorithm to find a sequence of objects called medoids. It is centrally located in clusters with an average distance of objects to their closest selected object to be minimum. Steps of algorithms are as follows:

1. k-objects selection as medoids
2. Swapping the k-selected objects with unselected objects.

This algorithm can be used for the numerical type of data.
Quality can be improved further.

C. The Balanced Iterative Reducing and Clustering using Hierarchies

BIRC or Balanced Iterative Reducing and Clustering using Hierarchies is the example of the hierarchical type of clustering algorithm [44]. This algorithm constructs clustering features tree from data and leaf nodes are clustered. Steps of this algorithm are as follows:

1. Scans the data sets to construct features tree.
2. Apply the clustering to cluster the leaf nodes.

The quality of clusters generated by the BIRC is not good.

D. The Clustering Using Representatives

CURE or Clustering Using Representatives is the example of the hierarchical type of clustering algorithm [44]. Steps of this algorithm are as follows:

1. Read data sets and create the p partitioned.
2. create the representative points for k clusters.

CURE has a high run time complexity for big data.

E. The Density Based Spatial Clustering of Applications with Noise

DBSCAN or Density Based Spatial Clustering of Applications with Noise belongs to density clustering algorithm. For this algorithm, the user has to provide minimum points and radius. Steps of this algorithm are as follows:

1. Select any data objects as a unvisited point.
2. Identify the neighborhoods of the point.
3. Clusters the points
4. Identify the other unvisited.

This algorithm is only applicable for the specific type of data sets.

F. CLIQUE Clustering

It is grid based type algorithm of clustering used to find subspace clusters of data sets. This algorithm is divided into following steps:

1. Finding the dense area of data sets.
2. Generate the k-dimensional cells.
3. Eliminates the low-density cells.
4. Cluster the high-density cells.

It is necessary to give more parameters to work correctly.

G. Expectation Maximization

EM or Expectation Maximization is standard model based approach to clustering. This algorithm is divided into following steps:

1. Assigning the each data objects hypothetically to one of the clusters.
2. Update the hypothesis and assign data objects to new clusters.

EM takes more running time for to cluster the data sets.

IV. Proposed Efficient Clustering Algorithm

From surveys of modified and standard methods of the clustering algorithm, this paper presents efficient clustering algorithm with following features.

1. It is uses improved k-means clustering algorithm [21] to automatically predict the value of a number of clusters and appropriate initial centroids.
2. This algorithm is making use of split and merges technique to clusters large data.
3. This algorithm also removes the empty clusters.
4. Refinement step is added to form more cohesive clusters.
5. The algorithm is more efficient and produces high-quality clusters.

This algorithm consists of following three major steps:

1. Predicting value of k and initial centroids using domain knowledge and spectral analysis[19][21]
2. Forming the intermediate clusters.
3. Refining the clusters to form final clusters.

Fig.1 shows the flow of proposed algorithm.

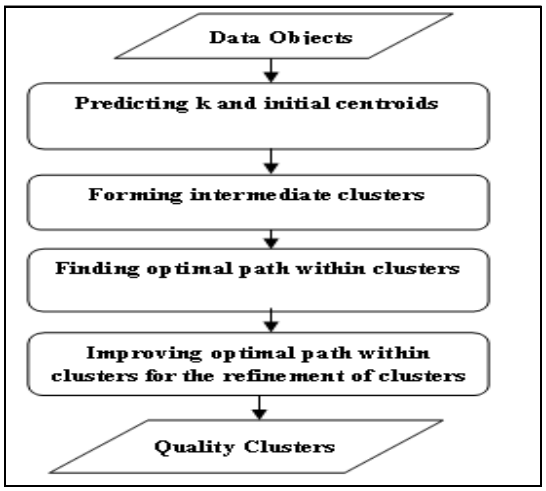

Fig.1. Flow of Proposed Algorithm

Detailed algorithm is given as follows:

Input: Data Objects
Output: Quality k-Clusters

Phase 1: Predicting k and initial centroids

1. Scan the input data and estimate the value of k by understanding and analyzing properties of data objects using domain knowledge and spectral analysis.
2. Select only required attributes of data from the data sets using above analysis.
3. Use improved clustering algorithm to determine initial centroids and form the clusters.

Phase 2: Forming intermediate clusters

1. Check if the k is sufficient to apply fine tuning, if yes, fine tune k by reducing it.
2. Find the cluster with maximum negative impact, i.e., maximum within-SS.
3. Split the cluster into 2 new clusters and replace it in the list of clusters.
4. Calculate the accuracy of the newly formed clusters.
5. Repeat steps 2, 3 and 4 until the accuracy is significantly increasing.
6. If the new accuracy has no significant improvement, consider the previous instance of clusters list as the final clusters.
7. Create clusters using above methods.

Phase 3: Improving and finding optimal path within clusters for the refinement of clusters

1. Identify outliers and form them into one cluster.
2. Reduce the clusters by eliminating empty clusters.
   1.1. Scan all the clusters and for each cluster check if there exist some points in it.
   1.2. If for any cluster number of point is zero, then remove the cluster and reduce the value of k by one.
3. Take all the formed clusters.
4. For every element in the cluster 'A':
   1. Take element of this cluster A_i
   2. For every element in the other cluster 'B':
      1. Take the distance

$$D = (A_i - B_i)2 \quad (1)$$

      2. Store this distance in Distances [] Array.
5. The minimum (Distances []) is the distance between cluster A and B.
6. Store this minimum distance between these two clusters.
7. Repeat steps 2, 3 and 4 for all cluster pairs possible.
8. Find out the cluster pair with a minimum distance between each other, say pair P.
9. Merge these clusters and calculate new accuracy.
10. Repeat this until the accuracy significantly increases.

11. If the new accuracy has no significant improvement, consider the previous instance of clusters list as the final clusters.
12. Stop when criteria met.

In this algorithm, there are three phases. The first phase predicts the value of k and initial centroids. The second phase uses split and merges techniques to form intermediate clusters based on predicted value of k. And the third step is responsible for refining the clusters with no empty clusters and high cohesive clusters.

V. Experimental Results

The clustering algorithms[25] such as k-means, PAM, Hierarchical clustering, DBSCAN and proposed clustering algorithms are applied to various data sets. Also proposed algorithm is compared with R-k-means clustering which uses Lloyd's algorithm. And then comparative experimental results are presented in this section. For the experiments, this paper is using the accuracy of clustering, running time of algorithms, silhouette, Dunn, DB and CH scores to compare the performance of algorithms. Details of these measures are given in paper [4][13][40]. The value of accuracy, Silhouette, Dunn, and CH should be high whereas the value of running time and DB index should be low for better clustering. For the maximum data sets, all the values of above measures are getting optimal for proposed clustering algorithm.

A. Data Sets

This paper uses 10 real data sets from kaggle site. These data sets include small to large data in size. Table 1 shows the details of data sets used.

Table 1. Data Sets Used

SN	Data Sets	Number of instances	Number of attributes
1	Accident	2057	15
2	Airline clusters	3999	7
3	Breast Cancer	569	30
4	Cities	493	10
5	Diamond	3089	11
6	Judges Rating	43	13
7	Rating1	2105	27
8	Salary	3123	6
9	Galaxy	3462	65
10	Voting	1076	5
11	Sensor1	8845	78

Table 2 shows results of standard k-means and proposed clustering algorithm with respect to efficiency.

Fig.2 to Fig.9 shows the performance of proposed, existing clustering algorithm and R-k-means clustering algorithms. Table 3 shows results of standard k-means and proposed clustering algorithm with respect to quality of clusters.

Table 2. The efficiency of Standard k-means Vs. Proposed Clustering

Data sets	Algorithm	Accuracy	Running Time
Accident	Proposed Algo.	95.4	0.000581
	Standard k-means	93.81	0.000088
Airline Clusters	Proposed Algo.	98.7	0.00212
	Standard k-means	98.4	0.011
Breast cancer	Proposed Algo.	97.43	0.000551
	Standard k-means	96.65	0.000015
Cities	Proposed Algo.	99.31	0.000521
	Standard k-means	97.5	0.000061
Diamond	Proposed Algo.	98.25	0.000249
	Standard k-means	98.06	0.000669
Judges rating	Proposed Algo.	97.15	0.0011
	Standard k-means	90.94	0.00026
Rating 1	Proposed Algo.	79.37	0.000463
	Standard k-means	61.5	0.000039
Salary	Proposed Algo.	96.39	0.000666
	Standard k-means	95.8	0.000034
Galaxy	Proposed Algo.	99.86	0.00134
	Standard k-means	99.67	0.0122
Voting	Proposed Algo.	96.12	0.000559
	Standard k-means	95.44	0.000793
Sensor 1	Proposed Algo.	95.01	0.000891
	Standard k-means	85.67	0.0138

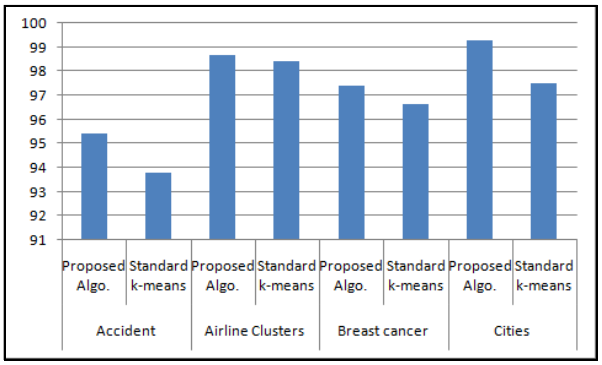

Fig.2. Accuracy of Proposed Algorithm

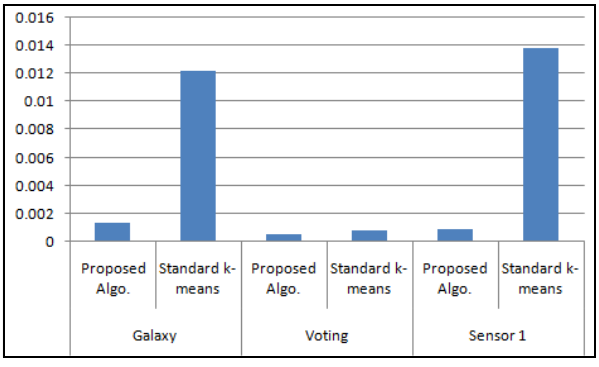

Fig.3. Running Time of Proposed Algorithm

Table 3. Clusters Quality of Proposed Clustering

Data sets	Algorithm	Silhouette Score	Dunn Score	CH Score	DB Score
Accident	Proposed Algo.	0.43	0.0025	3258.9	0.1
	Standard k-means	0.46	0.002	2434	0.77
Airline Clusters	Proposed Algo.	0.29	0.00078	0.0003	0.01
	Standard k-means	0.28	0.0003	2051.93	0.94
Breast cancer	Proposed Algo.	0.44	0.0079	1617.68	0.08
	Standard k-means	0.4	0.0071	1265.64	0.78
Cities	Proposed Algo.	0.35	0.0010	3224.34	0.07
	Standard k-means	0.34	0.0007	888.16	0.82
Diamond	Proposed Algo.	0.57	0.0060	19190.56	0.06
	Standard k-means	0.56	0.0026	19188.48	0.49
Judges rating	Proposed Algo.	0.69	0.33	96.17	0.13
	Standard k-means	0.23	0.27	34.04	0.88
Rating 1	Proposed Algo.	0.19	0.050	619.0	0.5
	Standard k-means	0.15	0.026	621.89	1.9
Salary	Proposed Algo.	0.36	0.0013	1071.01	0.02
	Standard k-means	0.35	0.0007	989.83	0.91
Galaxy	Proposed Algo.	0.53	0.01	91385.11	0.64
	Standard k-means	0.35	0.01	33459.76	0.88
Voting	Proposed Algo.	0.64	0.011	468.29	0.03
	Standard k-means	0.24	0.0038	388.16	1.0
Sensor 1	Proposed Algo.	0.55	0.0005	4194.5	1.11
	Standard k-means	0.5	0.0004	2063.7	0.88

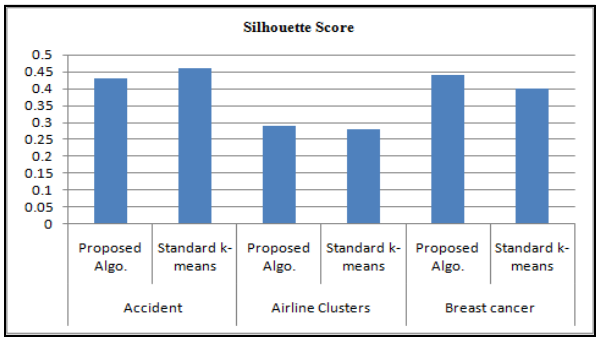

Fig.4. Silhouette Score of Proposed Algorithm

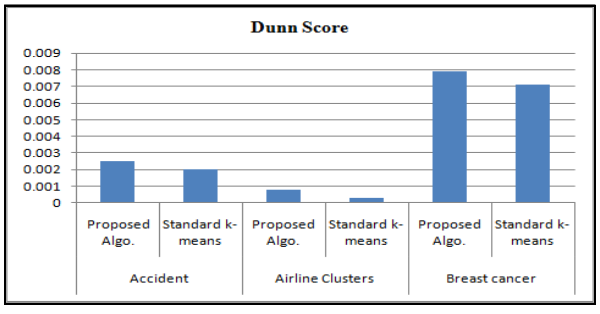

Fig.5. Dunn Score of Proposed Algorithm

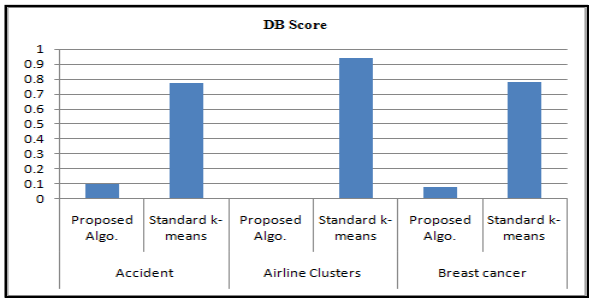

Fig.6. DB Score of Proposed Algorithm

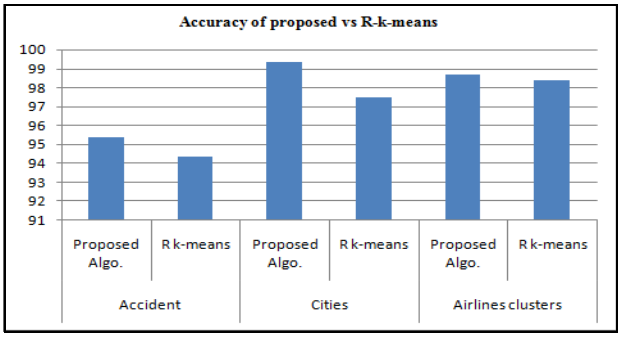

Fig.7. Accuracy of Proposed Algorithm

Table 4. Proposed Algorithm Vs. R-k-means

Data sets	Algo.	Accuracy	Run Time	Silhouette Score	Dunn Score	CH Score	DB Score
Accident	Proposed Algo.	95.4	0.000581	0.43	0.0025	3258.9	0.1
	R k-means	94.38	0.222	0.33	0.0033	2644	0.83
Cities	Proposed Algo.	99.31	0.000520	0.35	0.0010	3224.34	0.07
	R k-means	97.5	0.055	0.32	0.00079	908.29	0.92
Airlines clusters	Proposed Algo.	98.7	0.00212	0.29	0.00078	2712	0.01
	R k-means	98.4	0.89	0.27	0.00029	2147	1.04

Table 4 shows the R programming k-means and proposed clustering algorithm. R tool uses Lloyd's algorithms in k-means [38-39].

Table5 shows the comparison of proposed clustering and other existing clustering algorithms.

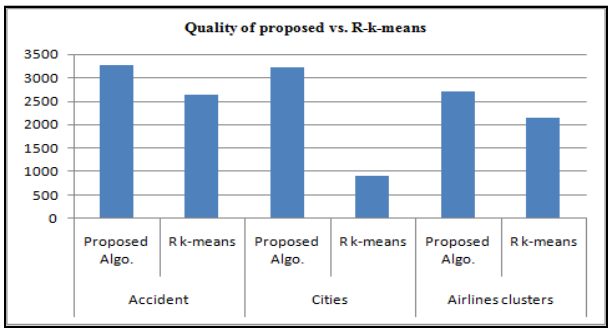

Fig.8. CH Score of Proposed Algorithm

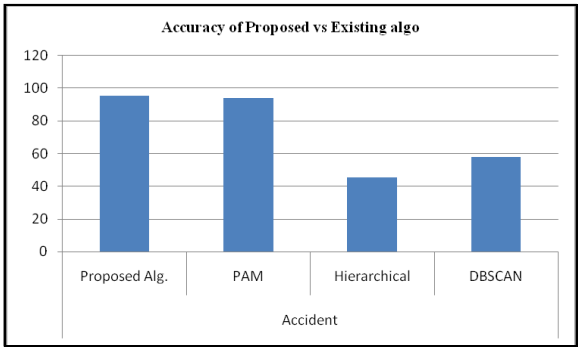

Fig.9. The accuracy of Proposed Vs. Existing Algorithms

Table 5. Existing Algorithms Vs. Proposed Algorithm

Data sets	Algorithms	Accuracy	Running Time	Silhouette Score	Dunn Score
Accident	Proposed Alg.	95.4	0.00058	0.43	0.0025
	PAM	93.97	10.18	0.43	0.001
	Hierarchical	45.43	0.0018	-0.027	0.0013
	DBSCAN	57.70	0.0076	-0.57	4236599
Cities	Proposed Alg.	99.31	0.00125	0.35	0.023
	PAM	97.11	0.0023	0.33	0.020
	Hierarchical	96.91	0.0032	0.30	0.021
	DBSCAN	80.01	0..0033	0.31	0.22
Airline Clusters	Proposed Alg.	96.22	2.27	0.33	0.00032
	PAM	96.22	2.27	0.33	0.00032
	Hierarchical	97.12	2.12	0.33	0.0032
	DBSCAN	95.12	1.23	0.30	0.0021

From the above results it is observed that as a number of instances in data sets increases, accuracy is also increased. Fig.10 shows this trend.

From the Table 3, it is observed that higher the number of instances of data sets then lower the value of silhouette score. This trend is shown in the Fig.11.

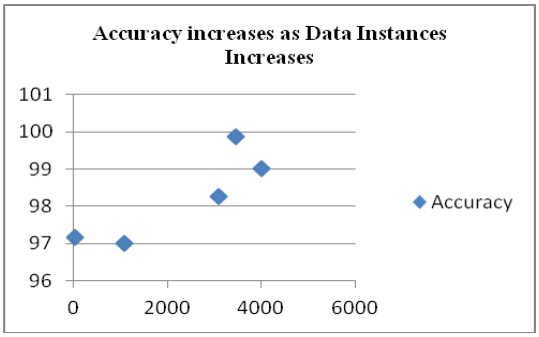

Fig.10. Accuracy increases as Data Instances Increases

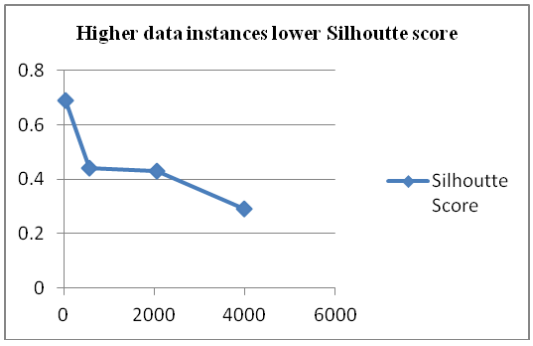

Fig.11. Higher Data Instances Lower Silhouette Score

VI. Conclusions

Clustering group large data into segments or clusters. The well-known clustering algorithms are the k-means, k-medoids, c-means, hierarchical and DBSCAN [35-38]. But these algorithms suffer from scalability problems and need improvements in it for the massive data. For the massive data, the performance of these algorithms is degraded. Therefore it produces bad cluster quality and takes more time for clustering. Hence massive data need some efficient clustering algorithm. By considering these gaps this paper is proposing efficient clustering algorithms using spectral analysis and split and merges technique. This paper is focusing not only clusters quality but also the efficiency of the clustering process. The algorithm is implemented using Python programming language. The performance of proposed clustering is checked on 11 real data sets with different size using six performance evaluation matrices. The performance is compared with standard clustering algorithms and R tool k-means clustering. The R tool is using Lloyd's Clustering algorithm for initial centers selection [33][38-39]. From the experiments on these data sets and algorithms mentioned above, the performance of proposed algorithm is better than these clustering algorithms. The values of all the performance matrices for the proposed algorithm on almost of all data sets and algorithms are getting optimal (approximately increased by 10-15%). The future scope of this work is further improvement in proposed clustering algorithm and applying on more real data sets.

References

[1] Tapas Kanungo David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, "An Efficient k-Means Clustering Analysis and Implementation", IEEE transactions on pattern analysis and machine intelligence, vol. 24, no. 7, July 2002.

[2] Wei Zhong, Gulsah Altun, 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 Nanobioscience, Vol. 4, No. 3, September 2005.

[3] Dimitrios Charalampidis, "A Modified K-Means Algorithm for Circular Invariant Clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 12, December 2005.

[4] Sriparna Saha, and Sanghamitra Bandyopadhyay,"Performance Evaluation of Some Symmetry-Based Cluster Validity Indexes", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 39, No. 4, 2009.

[5] Juntao Wang and Xiaolong Su, "An improved K-Means clustering algorithm", IEEE 3rd International Conference on Communication Software and Networks, 2011.

[6] Jiye Liang, Liang Bai, Chuangyin Dang, and Fuyuan Cao," The K-Means-Type Algorithms Versus Imbalanced Data Distributions", IEEE Transactions On Fuzzy Systems, Vol. 20, No. 4, August 2012.

[7] Mohamed Abubaker and Wesam Ashour," Efficient Data Clustering Algorithms: Improvements over Kmeans ", I.J. Intelligent Systems and Applications, 37-49,2013.

[8] Rui Máximo Esteves, Thomas Hacker, and Chunming Rong,"Competitive K-means", IEEE International Conference on Cloud Computing Technology and Science,2013.

[9] Rui Xu, and Donald Wunsch II," Survey of Clustering Algorithms ", IEEE Transactions on Neural Networks, Vol. 16, No. 3, May 2005.

[10] Ferdinando Di Martino, Vincenzo Loia, and Salvatore Sessa, "Extended Fuzzy C-Means Clustering in GIS Environment for Hot Spot Events", B. Apolloni et al. (Eds.): KES 2007/WIRN 2007, Part I, LNAI 4692, pp. 101–107, Springer-Verlag Berlin Heidelberg 2007.

[11] Bikram Keshari Mishra,Nihar Ranjan Nayak,Amiya Rath,Sagarika Swain ," Far Efficient K-Means Clustering Algorithm " , ICACCI-12 ,August 2012.

[12] Xiaohui Huang, Yunming Ye, and Haijun Zhang," Extensions of Kmeans-Type Algorithms: A New Clustering Framework by Integrating Intracluster Compactness and Intercluster Separation", IEEE Transactions on Neural Networks and Learning Systems, Vol. 25, No. 8, August 2014.

[13] Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y. Zomaya, Sebti Foufou, And Abdelaziz Bouras," A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis ", IEEE Transactions On Emerging Topics In Computing,2014.

[14] Michał Kozielski and Aleksandra Gruca," Soft approach to identification of cohesive clusters in two gene representations", Procedia Computer Science 35, 281 – 289, 2014.

[15] G.Sandhiya and Mrs.ramyajothikumar,"Enhanced K-Means with Dijkstra Algorithm for", 10th International Conference on Intelligent Systems and Control, 2016.

[16] Jeyhun Karimov and Murat Ozbayoglu," Clustering Quality Improvement of k-means using a Hybrid Evolutionary Model", Procedia Computer Science 61, 38 – 45, 2015.

[17] Vikas Verma, Shweta Bhardwaj, and Harjit Singh," A Hybrid K-Mean Clustering Algorithm for Prediction Analysis", Indian Journal of Science and Technology, Vol 9(28), DOI: 10.17485/ijst/2016/v9i28/98392, July 2016.

[18] Shashank Sharma, Megha Goel, and Prabhjot Kaur," Performance Comparison of Various Robust Data Clustering Algorithms", I.J. Intelligent Systems and Applications, 63-71, MECS, 2013.

[19] 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.

[20] Rui Xu, Jie Xu, and Donald C. Wunsch, II," A Comparison Study of Validity Indices on Swarm-Intelligence-Based Clustering ", IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 42, No. 4, August 2012.

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

[22] Anand Khandare and 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, Advances in Intelligent Systems and Computing 434, DOI 10.1007/978-81-322-2752-6_48, 2016.

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

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

[25] https://www.kaggle.com/datasets.

[26] http://scikitlearlearn.org/stable/modules/generated/sklearn.cluster.KMeans.html.
[27] Preeti Jain, Dr. Bala Buksh,"Accelerated K-means Clustering Algorithm ", I.J. Information Technology and Computer Science, 39-46 DOI: 10.5815/ijitcs.2016.10.05 ,MECS, 2016.
[28] Aleta C. Fabregas, Bobby D. Gerardo, Bartolome T. Tanguilig III,"Enhanced Initial Centroids for K-means Algorithm " ISSN: 2074-9007 (Print), ISSN: 2074-9015 (Online) DOI: 10.5815/ijitcs, MECS, 2017.
[29] P.SIVAKUMAR, Dr.M.RAJARAM,"Efficient and Fast Initialization Algorithm for K-means Clustering", I.J. Information Technology and Computer Science, 1, 19-24 DOI: 10.5815/ijitcs.2012.01.03,MECS, 2012.
[30] Yugal Kumar, G. Sahoo, "A Review on Gravitational Search Algorithm and its Applications to Data Clustering & Classification", I.J. Intelligent Systems and Applications, 2014, 06, 79-93 DOI: 10.5815/ijisa.2014.06.09, MECS, 2014.
[31] Handayani Tjandrasa, Isye Arieshanti, Radityo Anggoro, "Classification of Non-Proliferative Diabetic Retinopathy Based on Segmented Exudates using K-Means Clustering", I.J. Image, Graphics and Signal Processing, 1, 1-8 DOI: 10.5815/ijigsp.2015.01.01, MECS, 2015.
[32] Muhammad Ali Masood, M. N. A. Khan, "Clustering Techniques in Bioinformatics ", I.J. Modern Education and Computer Science, 2015, 1, 38-46 DOI: 10.5815/ijmecs.2015.01.06, MECS, 2015.
[33] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, Angela Y. Wu"An Efficient k-Means Clustering Algorithm: Analysis and Implementation", IEEE Transactions on Pattern Analysis and Machine Intelligence archive Volume 24 Issue 7, Page 881-892,2002.
[34] Purnawansyah, Haviluddin, "K-Means clustering implementation in network traffic activities", International Conference on Computational Intelligence and Cybernetics,10.1109/CyberneticsCom.2016.7892566, 2016.
[35] Chang Lu, Yueting Shi, Yueyang Chen, " Data Mining Applied to Oil Well Using K-Means and DBSCAN", 7th International Conference on Cloud Computing and Big Data, 10.1109/CCBD.2016.018,2016.
[36] Yohwan Noh, Donghyun Koo, Yong-Min Kang, DongGyu Park, DoHoon Lee, "Automatic crack detection on concrete images using segmentation via fuzzy C-means clustering", International Conference on Applied System DOI:10.1109/ICASI.2017.7988574,2017.
[37] Kai-Shiang Chang, Yi-Wen Peng, Wei-Mei Chen, "Density-based clustering algorithm for GPGPU computing", International Conference on Applied System Innovation, DOI: 10.1109/ICASI.2017.7988545, 2017.
[38] Dilmurat Zakirov, Aleksey Bondarev, Nodar Momtselidze, "A comparison of data mining techniques in evaluating retail credit scoring using R programming", Twelve International Conference on Electronics Computer and Computation, DOI:10.1109/ICECCO.2015.7416867, 2015.
[39] Tran Duc Chung, Rosdiazli Ibrahim, Sabo Miya Hassan, "Fast approach for automatic data retrieval using R programming language",2nd IEEE International Symposium on Robotics and Manufacturing Automation, DOI: 10.1109/ROMA.2016.7847824, 2016.
[40] M. Arif Wani, Romana Riyaz, "A new cluster validity index using maximum cluster spread based compactness measure", International Journal of Intelligent Computing and Cybernetics, ISSN: 1756-378X, 2016.
[41] Deepali Aneja, Tarun Kumar Rawat, "Fuzzy Clustering Algorithms for Effective Medical Image Segmentation", I.J. Intelligent Systems and Applications, 11, 55-61 DOI: 10.5815/ijisa.2013.11.06 , MECS ,2013.
[42] J Anuradha, B K Tripathy, "Hierarchical Clustering Algorithm based on Attribute Dependency for Attention Deficit Hyperactive Disorder", I.J. Intelligent Systems and Applications, 06, 37-45, DOI: 10.5815/ijisa.2014.06.04, MECS, 2014.
[43] SudiptoGuha,Rajeev Rastogi,KyuseokShim, "Cure: an efficient clustering algorithm for large databases",DOI: 10.1016/S0306-4379(01)00008-4 ,Elsevier, 2001.
[44] Tian Zhang, Raghu Ramakrishnan, Miron Livny," BIRCH: an efficient data clustering method for very large databases", SIGMOD '96 Proceedings of the 1996 ACM SIGMOD international conference on Management of data Pages 103-114,1996.
[45] https://www.python.org.
[46] https://www.programiz.com/python-programming.
[47] Brian S. Everitt, Sabine Landau, Morven Leese, "Cluster Analysis ", 4th Wiley Publishing ISBN:0340761199 9780340761199,2009.
[48] Fareeha Zafar, Zaigham Mahmood, "Comparative analysis of clustering algorithms comprising GESC, UDCA, and k-Mean methods for wireless sensor networks ", URSI Radio Science Bulletin Volume:84, Issue:4, 10.23919/URSIRSB.2011.7909974, 2011.
[49] Xiaohui Huang, Yunming Ye, Haijun Zhang, "Extensions of Kmeans-Type Algorithms: A New Clustering Framework by Integrating Intracluster Compactness and Intercluster Separation ", IEEE Transactions on Neural Networks and Learning Systems Volume: 25, Issue: 8, 10.1109/TNNLS.2013.2293795, 2014.
[50] Jianyun Lu, Qingsheng Zhu,"An Effective Algorithm Based on Density Clustering Framework", IEEE Wireless Communications Letters, Volume: 5, Issue: 6, DOI: 10.1109/LWC.2016.2603154,2016.
[51] Yuan Zhou, Ning Wang; Wei Xiang,"Clustering Hierarchy Protocol in Wireless Sensor Networks Using an Improved PSO Algorithm", IEEE Access, Volume: 5, DOI: 10.1109/ACCESS.2016.2633826,2016.
[52] Neha Bharill, Aruna Tiwari, Aayushi Malviya,"Fuzzy Based Scalable Clustering Algorithms for Handling Big Data Using Apache Spark" IEEE Transactions on Big Data, Volume: 2, Issue: 4, Pages: 339 - 352, DOI: 10.1109/TBDATA.2016.2622288, 2016.

Authors' Profiles

Anand Khandare graduated from Sant Gadge Baba Amravati University, Amravati in Computer Science and Engineering in 2005. He got his Master from Mumbai University in 2010-11. He is now pursuing his Ph.D. from Sant Gadge Baba Amravati University. Currently, he is working as Assistant Professor at Thakur college of Engineering and Technology, Mumbai University. He has 11 years of teaching experience in the same institute. He has published more than 10 papers in international journal and conference. 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 graduated from Sant Gadge Baba Amravati University, Amravati in Computer Science and Engineering. He got His Master and Ph.D. degree form the same university. Currently he is working as Professor in Computer Science and Engineering at PRMIT &R, Badnera, and Amravati. He has more 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, Abrar Alvi, "Efficient Clustering Algorithm with Enhanced Cohesive Quality Clusters", International Journal of Intelligent Systems and Applications(IJISA), Vol.10, No.7, pp.48-57, 2018. DOI: 10.5815/ijisa.2018.07.05

