IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 6, Ver. V (Nov.-Dec. 2016), PP 15-19
www.iosrjournals.org

Efficient Clustering Algorithm with Improved Clusters Quality

Mr. Anand Khandare[1], Dr. A.S. Alvi[2]
[1](PG Department of CSE, SGB Amravati University ,Amravati ,India)
[2](Department of CSE, PRMIT &R, Badnera, Amravati, India)

Abstract: Clustering *algorithm is data mining task where the value of a number of clusters and data objects are given as inputs to the algorithm to divide the data objects into k clusters. Clustering can be used in various applications like clustering online retailers, SMS, and email spam collection, human activity recognition and much more. There are huge researches and applications in the area of clustering. There are various clustering algorithms are available such as the k-means, k-medoids etc. The k-means is one of the widely used algorithms of clustering. But there are some problems of k-means related to efficiency and quality of cluster. This paper proposing little efficient clustering algorithm with improved cluster quality. For improvement in the cluster quality , this paper is using clustering aggregation and spectra analysis by understanding the properties of data before actual clustering. Then this modified algorithm is applied to online retails data set. And the results shows that proposed algorithm is little efficient and but producing quality clusters. The performance of proposed and standard k-means is compared by four performance metrics such Clustering Accuracy, Sum of Square Error, Compactness and Running Time.*

Keywords: *Clustering, Cluster quality, k-means, k-medoids, Sum of Square Error, Compactness*

I. Introduction

Now day's people are living in the world full of data. Every day, the online system gathers a large amount of information or data from customer's daily transactions with this system for further analysis and management of this information. One means to deal with this large data is to group this data into a set of clusters. The clustering plays an important role in this regard. Data clustering is the process of grouping similar data objects together into a number of clusters. The aim of clustering is to identify and classify objects into clusters that have the same meaning in the aspect of a particular problem [1].There are various categories of clustering algorithms such as distance based, hierarchical clustering, etc. Details of all clustering are given in paper [2]. The clustering is the technique of unsupervised learning as the without labeled data is given as the input to the clustering. And also clustering separate a finite unlabeled data set into finite clusters. The following figure shows the process of clustering.

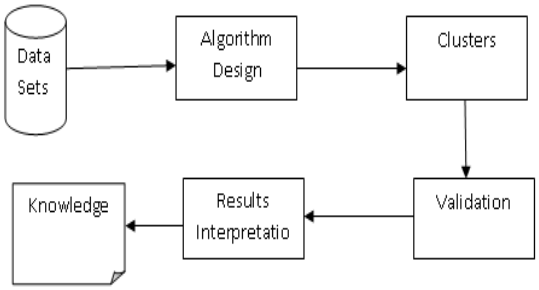

Fig 1: process of clustering

In this paper, several research papers on improving clustering algorithms are studied. From these papers, it is observed that still there is scope to improve the clustering algorithms efficiency and clusters quality. The paper proposed the clustering algorithm which is little efficient also producing quality clusters. For the experimentation, this paper is using a standard data set of online retails stores. This data set consists of eight attributes which are both numerical and text and approximately five lacks records of retails transactions. The main objective of this paper is designing a clustering algorithm which will produce quality clusters without compromising the efficiency of the algorithm. This paper is divided into five sections. The first section is covering background and introduction of clustering and our research and input data set. In the second section, paper presented a survey of literature. The third section is covering design and working of both standard and modified clustering algorithms. In next section, results and comparisons are given. And in last section, conclusion and future work of paper are stated.

II. Related work

Jeyhun Karimov and Murat Ozbayoglu [1] proposed the novel hybrid evolutionary model for the *k*-means clustering algorithm by using metaheuristic methods to identify the good initial centroids for the *k*-means clustering algorithm. The results indicate that the quality of cluster is improved by approximately thirty percent compared to the standard random selection of initial centroids. But still, there is scope for the improvement in the quality. Authors in the paper [3] proposed a method to determine the number of clusters ,k using spectra analysis techniques and then tested this algorithm on various data sets and observed that there is fluctuation in the value of k. Identifying the proper number of clusters from a dataset are two crucial issues in unsupervised learning. Authors proposed nine newly proposed PS-distance-based cluster validity indexes. These indexes exploit the property of PS to indicate both the appropriate number of clusters as well as the appropriate partitioning. The effectiveness of these nine newly developed indexes in comparison with the original nine cluster validity indexes are provided artificially generated and real-life datasets [4]. This paper [5] presented design and implementation of modified K-means algorithm .for the enhancement of quality of clusters, and execution time. Then the algorithm is applied on emotional intelligence data rather than to get meaningful clusters for analysis.

This analysis is used for decision making. For the shortcomings of the K-Means clustering algorithm, this paper [6] proposed an improved K-means algorithm using noise data Density-based detection method based on characteristics of noise data where the discovery and processing steps are added to the original k-means algorithm. In paper[7], BoostKCP, a simple but powerful heuristic method that has proved useful for reducing the unnecessary computation is proposed based on the properties of the Pearson correlation distance. Authors in the paper[8] studied various literature on improved k-means algorithms, summarized their shortcomings and identified scope for further enhancement to make it more scalable and efficient for large data. Also suggesting

Algorithm to avoid outliers using the concept of semantic analysis and AI. Paper [9] presents an improved version of the Moving KMeans algorithm known as Enhanced Moving K-Means algorithm. In this, members of the cluster with the highest fitness value are forced to become the members of the clusters with the smallest fitness value. Also, two versions of algorithms are proposed. In this paper [10], three new clustering algorithms by extending the existing k-means-type algorithms are proposed by integrating both intracluster compactness and inter-cluster separation. The properties and performances of these algorithms are checked on several synthetic and real-life data sets using four metrics accuracy, RandIndex, F-score, and normal mutual information. Today a major issue of clustering algorithms for big data is complexity.

Paper [11] introduces concepts and algorithms related to clustering, a concise survey of existing algorithms as well as providing a comparison. Authors said that no clustering algorithm performs well for all the evaluation criteria. Consensus clustering [12] is to find a single partitioning which agrees as much as possible with existing basic partitioning. Consensus clustering is the solution to finding cluster structures from heterogeneous data. This paper presents a study of K-means-based consensus clustering. This paper [13] proposed clustering technique Implemented in the quaternion domain for qualitative classification of E-nose data which is similar to the k-means clustering with better class reparability and higher cluster validity. But this technique requires more computations. This paper [14] proposed an extended Chaotic Particle Swarm Optimization algorithm which is called ECPSO for the optimum clustering. The proposed algorithm is an extended version of Particle Swarm Optimization, The ECPSO, enhanced the operators in the classical algorithm. The work in the paper [15] an enhanced parallel implementation of k-Means clustering using Cilk Plus and Open MP on the CPU and CUDA on the GPU.

III. Efficient k-Means

This paper selected k-means clustering algorithm for the enhancement. K-means is the clustering algorithm used to analyze the large volume of data generated by various modern applications. There are various problems of k-means clustering like efficiency, scalability and quality and accuracy. The aim of this paper is to modify the k-means clustering algorithm and developed an efficient algorithm with quality clusters.

1. Standard k-means

In the k-means clustering, given a set of n data points an integer k and the problem is to determine a set of k points. The k-means is a widely used method to partition a dataset into groups of clusters. But it requires expensive distance calculations of centroids and also chances of bad quality clusters. The summary of k-means is shown in following steps:

a. Choose the number of clusters k and input a dataset.
b. Randomly select initial candidates for the k cluster centers from the dataset.
c. Assign each object to the nearest cluster using a distance measure.
d. Re-compute the centroids of these k clusters to find new cluster centers.
e. Repeat these steps till condition met.

Figure 2 shows the working of the standard k-means clustering algorithm.

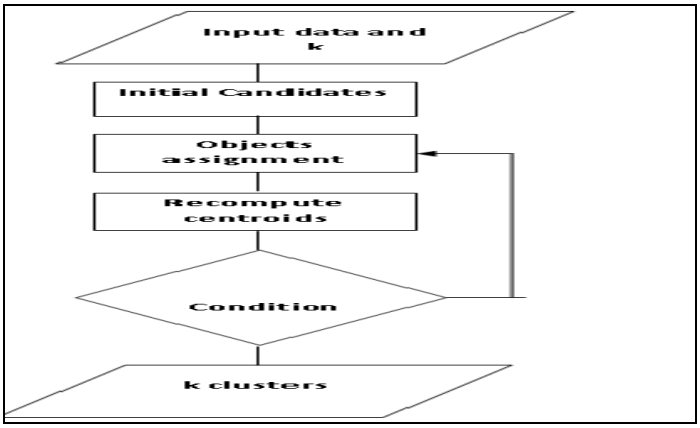

Fig 2: standard k-means

2. Efficient k-means

Because of randomly selecting initial centroids and user given value of k, the quality of clusters will be affected. Also, leads to degrading the efficiency of the algorithm. So for the better quality, it is necessary that algorithm should not depend on user given value of a number of clusters (k). And also there should be a standard way of selecting initial centroids. So by considering two objectives of efficiency and cluster quality, this paper proposed improved clustering algorithm. General steps of proposed clustering algorithm are shown as follows:

Input: Data Objects

Output: Quality k-Clusters

a. Scan the input data and estimate the value of k by understanding and analyzing properties (number or text of input data using domain knowledge and spectra analysis.
b. Select only required attributes of data from the data sets using above analysis.
c. Use standard k-means to
1. Select initial centroids from selected attributes.
2. Calculate distance and assign objects to clusters using initial centroids
3. Use the average of each cluster as initial centroids for the improved algorithm.
d. Create clusters by using these centroids.
e. Use clustering aggregation and
1. Check whether centroids are at beginning or middle or end of that clusters if it is not in the middle of the cluster, the adjustment should be done.
2. Check if some centroids are sufficiently close to each other to be clubbed into a single cluster, if so, combine the clusters and re-compute centroids.
3. Reform clusters by repeating step 1 and 2.
f. Create final clusters using above methods
g. Stop when condition met.

IV. Results and Comparative Analysis

For both algorithms, this paper is using a standard data set of online retail stores. This structure of this data set is shown in Table 1.

Table 1: Sample Data Set

	A	B	C	D	E	F	G	H
1	voiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
2	536365	85123A	WHITE HANGING H	6	12/1/2010 8:26	2.55	17850	United Kingdom
3	536365	71053	WHITE METAL LAN	6	12/1/2010 8:26	3.39	17850	United Kingdom
4	536365	84406B	CREAM CUPID HEA	8	12/1/2010 8:26	2.75	17850	United Kingdom
5	536365	84029G	KNITTED UNION FL	6	12/1/2010 8:26	3.39	17850	United Kingdom
6	536365	84029E	RED WOOLLY HOTT	6	12/1/2010 8:26	3.39	17850	United Kingdom
7	536365	22752	SET 7 BABUSHKA N	2	12/1/2010 8:26	7.65	17850	United Kingdom
8	536365	21730	GLASS STAR FROST	6	12/1/2010 8:26	4.25	17850	United Kingdom
9	536366	22633	HAND WARMER UN	6	12/1/2010 8:28	1.85	17850	United Kingdom
10	536366	22632	HAND WARMER RE	6	12/1/2010 8:28	1.85	17850	United Kingdom

This data set consists of eight attributes and approximately five lacks of instances of online retails. Firstly this paper applied standard k-means clustering for this data set. This algorithm considered all the attributes of data set. So this will decrease the performance of the algorithm. And also affect the quality of clusters. For standard k-means, there is no provision for understanding the input. Hence only numerical attributes are considered Sample results of standard k-means clustering are shown in table 1.

Table 2: Results of standard k-means

Data set	**Number of clusters**	**Sample clusters based on stock code customers id are clustered**	**Observation**
Online retail stores data	4 (Given by user)	{12583, 16098, 17967, 13069, 17629,14189}, {15547,17629,14189,15203,16551,16607,13717,15150}, {17967,17198,15021,12748,16931,15514,15026,15752,16657,16657,15854}	Value of clusters k is given by user Some customers are assigned are in more than one clusters and also one cluster is empty.

Table 3: Results of standard k-means

Data set	**Number of clusters**	**Sample clusters based on stock code customers id are clustered**	**Observation**
Online retail stores data	Based on a range of values and numbers value of k is calculated automatically. Also using dimension reduction technique algorithm selected attributes like StockCode, Quantity, CustomerID.	**On Sample basis** {17850,16250,12431} {15100,15311,13705} {12583,17511,13767}	The value of clusters k is not given by the user. The quality of clusters is increased by using dimension reduction techniques. Based on customer and their maximum quantity clusters are formed.

To measures the effectiveness of clustering algorithms following measures can be is used.

- SSE: Sum of square error is the distance of each data object to its nearest centroid.
- Maximum Radius: Largest distance from a data objects to its cluster centroid.
- Average Radius: the sum of the largest distance from an object to its cluster centroid divided by the number of clusters.
- Rand measure: Rand index as a measure of the percentage of correct decisions made by the algorithm.

To measures the performance of standard and proposed algorithms, this paper are using four performance metrics. Their values are shown in Table 4.

Table 4: Results of efficient k-means

Algorithm	**Clustering Accuracy**	**Sum Squared Error**	**Compactness**	**Run Time**
Standard K-means	71.12	1.5	4.13	450.12
Efficient k-means	74.12	1.3	3.50	423.10

Table 4 shows the clustering accuracy of proposed algorithm is increased by approximately by 3%. Also, SSE value is lower than standard algorithm. The proposed algorithm is producing little compact clusters than the standard algorithm in less run time. Figure 3 is showing the performance of these two algorithms.

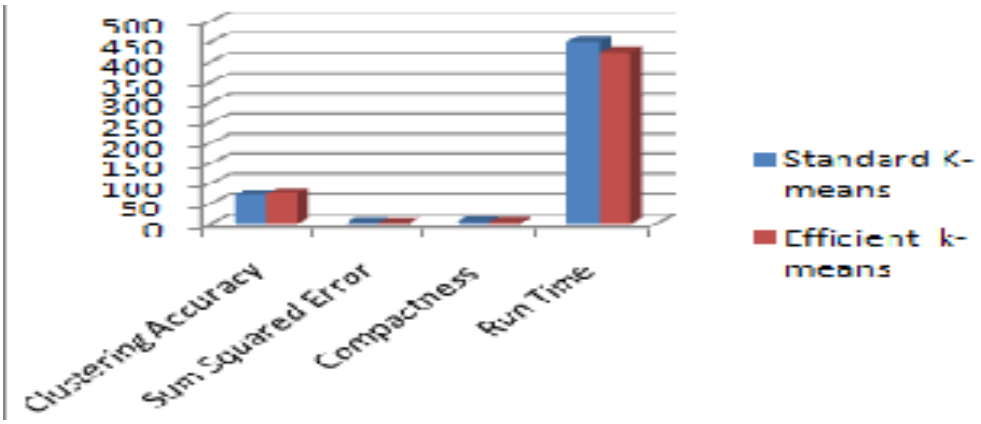

Fig 3: Comparative Analysis

V. Conclusion and future work

This paper studied various literature and proposed the little efficient clustering algorithm with more focus on improving the quality of clusters. This algorithm is using clustering aggregation and spectra analysis to analyze properties of the input data before applying clustering steps. Then standard k-means with some additional steps is used for the selection of initial centroids. Then this algorithm is applied to standard data set such as online retail stores. This data set consists of eight attributes and approximately five lacks records. Then proposed algorithm is compared with standard clustering algorithms and found that proposed algorithm is performing little better than standard algorithm. In this paper, two algorithms are compared with four measures of quality and performance. These measures are clustering accuracy, sum squared error, cluster compactness and run time. And it is found that the values of these measures are little better for the proposed algorithm than standard algorithm. Future work of this paper is to design scalable clustering algorithm which will work for large and more than one type of data sets.

References

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

[2] Rui Xu, Donald Wunsch,Survey of Clustering Algorithms, *IEEE Transaction on NEURAL NETWORKS, VOL. 16, NO. 3, (2005).*

[3] Wenyuan Li, Wee-Keong Ng,Ying Liu, Member, Kok-Leong Ong, Enhancing the Effectiveness of Clustering with Spectra Analysis, *IEEE Transaction on KNOWLEDGE AND DATA ENGINEERING, vol. 19, no. 7,* 2007.

[4] 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.

[5] 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.

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

[7] Kazuki Ichikawa , Shinichi Morishita,A Simple but Powerful Heuristic Method, for Accelerating k-Means Clustering of Large-Scale Data in Life Science, *IEEE/ACM Transaction on COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 11, NO. 4,* 2014.

[8] Anand Khandare , 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.

[9] 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.

[10] Xiaohui Huang, Yunming Ye, and Haijun Zhang,Extensions of Kmeans-Type Algorithms: A New Clustering Framework by Integrating Intracluster Compactness and Intercluster Separation, *IEEE Transaction on NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 25, NO. 8,* 2014.

[11] Adil Fahad1, Najlaa Alshatri1, Zahir Tari1, Abdullah Alamri1,Ibrahim Khalil1, Albert Y. Zomaya2,Sebti Foufou, Abdelaziz Boura , A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis, *IEEE Transaction on EMERGING TOPICS IN COMPUTING,* 2014.

[12] Junjie Wu,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.

[13] Ravi Kumar , Ramashraya Dwivedi,Ravi Kumar and Ramashraya Dwivedi, Quaternion Domain k-Means Clustering for Improved Real-Time Classification of E-Nose Data ,*IEEE SENSORS JOURNAL, VOL. 16, NO. 1*, 2016.

[14] Maryam Lashkari,Mohammad Hossein Moattar,The improved K-means clustering algorithm using the proposed extended PSO algorithm, *IEE International Congress on Technology, Communication and Knowledge* , 2015.

[15] Mohammed Baydoun; Mohammad Dawi; Hassan Ghaziri,Enhanced parallel implementation of the K-Means clustering algorithm *International Conference on Advances in Computational Tools for Engineering Applications Pages: 7 - 11, DOI: 10.1109/ACTEA.2016.7560102* , 2016.