An Enhanced K-Means Clustering Algorithm to Improve the Accuracy of Clustering using Centroid Identification Based on Compactness Factor

Dr.M.Sakthi
Associate Professor and Head
Department of Computer Science
NGM College
Pollachi, Coimbatore (Dist.), Tamil Nadu.

Abstract---The Researchers find it difficult to extract information from a large data set through a standard function. It is found insufficient of standard functions to extract needed information. It has been considered that the k-means algorithm in the situation where the data is too enormous to be stored in main memory and must be retrieved sequentially, such as from a disk, and where it must be used as slight memory as possible. The k-means clustering also converges very quickly when it is employed to obtain data from huge data collections. It is also on other hand, k-means has some disadvantages too, and it includes affluent computation by getting cluster centers which are randomly selected at initial. It influences the two factors, performance of the algorithm and number of clusters initialization. In this paper an improved k-means algorithm in terms of data clash strainer mechanism is given. The data clash strainer mechanism is implemented through a function Regional Centroid Component (RCC) mechanism which is added to the standard k-means algorithm. This density based recognition mechanism is built on the properties of clash data. The clustering result is effectively enhanced by ignoring the clash data prior to the process of data clustering. Hence, the improved algorithm offers a great accuracy when compared to other existing cluster algorithms.

Keywords: Cluster, Data clash strainer, K-means, RCC.

1. Introduction

A vast amount of data is dealt in various fields and those big data are handled using data mining techniques to retrieve information. “We are living in the information age” is a popular saying; however, it is like actually living in the data age. Terabytes or petabytes of data pour into our computer networks, the World Wide Web, and various data storage devices every day from business and which is needed to be extracted in a useful manner to infer knowledge from it [1]. The technique of data mining involves the cluster analysis which is one of the main focuses of the present-day researchers.

Clustering is a fundamental method for appreciative and interpreting data that seeks to partition input objects into groups, known as clusters, such that objects within a cluster are similar to each other, and objects in different clusters are not. A clustering invention called k-means is simple, intuitive, and widely used in practice. Given a set of points S in a Euclidean space and a parameter k, the objective of k-means is to partition S into k clusters in a way that minimizes the sum of the squared distance from each point to its cluster center [3]. This circumstance causes the

formation of wide range of clustering algorithms such as COBWEB, DBSCAN, CURE, MEANS etc., [4].

This work introduces the method that avoids the arbitrary selection of options at initial and involves detection and eliminations of the identified far apart data collection from the clusters. It works initially to improve the performance of classic k-means clustering mechanism in terms of its accuracy and reduced complexity [14].

The rest of the paper is organized as follows; the section 2 gives the brief discussion on related works. In section 3 the basic nature of k-means clustering procedure is studied and proposed methodology based on the compactness based centroid detection technique is presented. Section 4 focuses on the comparison of proposed methodology with other existing clustering algorithms and gives the results of experiments. Finally section 5 concludes our work.

2.Related Works

Shorab et al., (2012) presents an empirical method to select the appropriate centroids at initial level in k-means clustering strategy and hence it tries to improve the algorithm in terms of its clustering accuracy as well as focused on the time of execution. Experimental results showed the better adeptness of the improved k-means clustering algorithm over the traditional k-means algorithm but it increases complexity of the clustering algorithm as the size of data set increases.

Cosmin M P, et al., (2014) reveals customer segmentation is done with data mining to know the customer characteristics information hidden inside. The way to find out the customer segments of a company is clustering analysis. Clustering is the process of forming segments of a set of data by measuring similarities between data with other data.

Patel and Prateek(2016), explores different kind of various problems using data mining clustering mechanism and the relationships between them. K-means clustering algorithms, hierarchical algorithm are discussed in this paper. The performance of this algorithm is compared in clustering process and gives the proposition about the suitability of such algorithms in different kind of states for the different datasets.

M A Syakuret al., (2018) says that the segmentation process puts customers in line with the characteristics of similar customer groups. Customer segmentation is a preparatory step to classify each customer according to a defined customer group. Customer segmentation based on market research and demography requires understanding the characteristics of all customers to be more effective.

Hong et al., (2019) proposed an improved k-means algorithm as the result of clustering reliability analysis and the proposed algorithm shows the stability and achieves better result when the solidity is uneven and there exist large difference in data clustering. Experimental results showed the ability of improved k-means algorithm in handling non-uniformed data set.

3. The Core Scheme of K-means Clustering Algorithm and The Proposed Methodology

The result of cluster analysis based on partition strategy the k-means methodology was derived [5]. This methodology requires arbitrary selection of "k" number of cluster centroids at initial. It also involves computation of distance between each selected centroid and each instances of organized data collection to find the nearest centroid, and also amend average distance of centroids. This process is repeated until standards or norms of the function met.

The mean squared deviation standard for clustering is calculated as follows,

$$D = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left\| l_{ij} - s_i \right\|^2$$

Where l_{ij} is instance of class I and s_i is centroid of class i. This methodology is illustrated in the figure 1. This K means clustering algorithm steps were given as follows. It involves the arbitrary selection of centroids, detection of data center point, calculating distance, forming clusters.

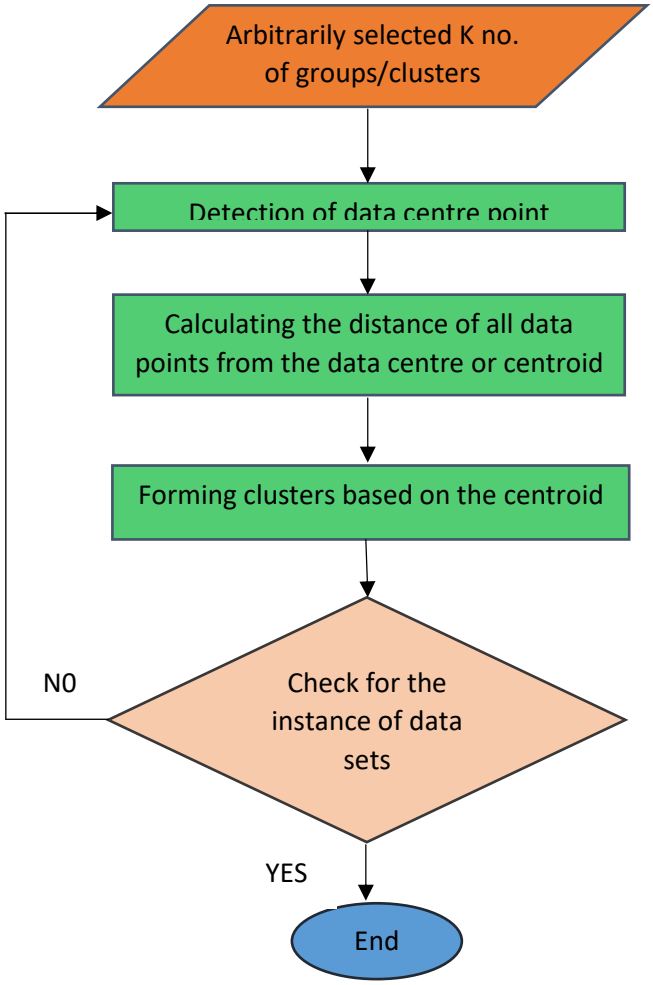


Figure 1. The fundamental plan or scheme of K-means algorithm

Input: P instances need to be cluster $\{a_1, a_2, \ldots\ldots, a_n\}$ and the k (no. of initial centroids)

Output: k centroids and the disagreement volume between each instances and its short-distant centroid neighbor.

K-means clustering Algorithm

Begin

Step 1: Randomly choose k instances as initial centroids ($c_1, c_2, \ldots\ldots, c_k$)

Step 2: Determine the distance between each instance a_i and each centroid and consequently connect each instance to the short-distant centroid neighbor by the following expression

$$dis(a_i, c_i) = \sqrt{\sum_{j=1}^{d} \left(x_{i1} - m_{j1}\right)^2},\ i = 1 \ldots . N, j = 1 \ldots . . k.$$

Where dis(a_i, c_i) is the distance of data i.e., i from cluster i.e., j.

Step 3: Determine the average of instances in each group or cluster i.e., new centroid using the formula below,

$$c_i = \frac{1}{N_i} \sum_{j=1}^{N_i} a_{ij}\ , i = 1,2, \ldots . k, \qquad N_i\ is\ the\ no. of\ instances\ of\ current\ cluster\ i$$

Step 4: Repeat the steps 2 and 3 until the standard function D met and produce ($c_1, c_2, \ldots\ldots, c_k$)

End

Figure 2.K-means clustering algorithm.

The intricacy of this K-means mechanism is expressed by the factors – arbitrary selected k number of clusters, number of repetition of the procedure and number of organized data instances [6].

3.1. Centroid recognition based on Compactness

The performance of the K-mean technique depends on the centroid selected at initial which greatly affects the result of the technique used. The outliers in the clusters away from the data compact region cause the newly founded centroid more deviated from data-compact region and so, it directly influences the final clustering result and the final result encounters the huge deviation from the actual[13]. To avoid such an outlier and to enhance the result it is better to discard the isolated data from our collection of data prior to the process of data clustering. The deviating level of each instance in organized data is determined using Regional Centroid Component (RCC) which involves the computation of distance of each instance from its short-distant centroid neighbor only after the completion of the process of producing k number of centroids and k number of shortest distance of each instance from its short-distant centroid neighbor. Finally, RCC detects the regional centroid as per the regional centroid component of each instance [7].

This **Regional Centroid Component (RCC)Detection** is illustrated in the following steps:

Step 1: computing k-short-distant centroid distance dis (d, i) where i is short-distant centroid neighbor, using the following formula

$\mathbf{dis(d,i)} = \sqrt{(\boldsymbol{a^1} - \boldsymbol{b^1})^2 + (\boldsymbol{a^2} - \boldsymbol{b^2})^2 + \cdots + (\boldsymbol{a^n} - \boldsymbol{b^n})^2}$ where n denotes data collection's dimension.

Step 2: compute the compactness of each instance d. The compactness of each instance d that express the scattering of short-distant data is determined using the corresponding k-short-distant cluster average.

$\mathbf{sda(p)} = \left(\frac{1}{k}\sum_{i=1}^{k} \boldsymbol{dis}(\boldsymbol{d}, \boldsymbol{i})\right)^{-1}$

Step 3: compute the regional centroid component of d

$$\boldsymbol{RCC(d)} = \frac{\sum_{i=1}^{k} \frac{\boldsymbol{sda(i)}}{\boldsymbol{sda(d)}}}{\boldsymbol{k}}$$

Step 4: in the case of RCC (d) << 1, where d is far apart data point, the instance p is discarded from the collection of required data and go to the initial step until no. of data collection leaves unaltered. And hence the newly created data collection is to be grouped.

Figure3. Regional Centroid Component (RCC)Detection

Where sda (i) is the regional compactness of k-short-distant centroid of d, sda (d) is regional compactness of d. RCC (d) express the scope of d as centroid. The RCC has the value of about one in Compactness dispensation data collection. The centroid component by which the centroid is differentiated is greater than others because the regional compactness of the centroids in the collection of data is much less than the regional compactness of its short-distant instances.

3.2. The Improved k-means clustering algorithm using Regional Centroid Component (RCC)

The mechanism is triggered by the process of elimination of far-apart data collection by employing the above said RCC based recognition strategy[12]. It ensures that the computation of initial centroid is free from the far-apart data collection instances and removes them in the determination of centroid. The improved k-means algorithm is executed on the newly selected data organization employing the RCC and illustrated in the following steps.

Input: P instances need to be cluster $\{a_1, a_2, \ldots\ldots, a_n\}$ and the k (no. of initial centroids)

Output: k centroids and the disagreement volume between each instances and its short-distant centroid neighbor.

Begin

Step 1: Rebuff on data collection w, compute RCC (d)

$$RCC(d) = \frac{\sum_{i=1}^{k} \frac{sda(i)}{sda(d)}}{k}$$

And if RCC (d) >>1 then, eliminate the far-apart data point d

Else retain the data collection d. and obtain the new data collection S.

Step 2: Determine the average of the data collection S by assuming the S as centroid at initial stage.

$$c_1 = \frac{1}{n}\sum_{i=1}^{n} a_i$$

Step 3: Discover the succeeding centroid with its distance from all other data set instances present in the group or class.

$$g_n = \sum_{i=1}^{N} \left(l_{k-1}^{j} - \left\| a_n - a_j \right\|^2, 0 \right)$$

$a_n$is representative data point for which the distance g_n is greatest. Determine the distance of centroid of each group or class from their class members and allocate it to the short-distant group or class.

Step 4: compute the average of instances in each group or class as new centroid.

Step 5: Redo the step 3 and 4 until the standard function D met, and p give (m_1, m_2, ….,m_k)

The performance of the proposed methodology in terms of its suitability and accuracy by comparing it with other existing clustering algorithms like mean shift clustering, DBSCAN (Density based Spatial Clustering Application Noise),Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM) and Agglomerative Hierarchical Clustering. The data sets from the UCI - one of the most popular neural network database- Abalone, Wine and Iris have been taken for our experiment. The table I gives this details in brief. The proposed work produces the better outcomes and offers optimal solution without avoiding the caliber of clustering. The experimental results prove the effectiveness of this improved version of k-means algorithm over all of other clustering strategies.

TABLE I. The Selected Data Sets

Data set	No. of objects	No. of properties	No. of groups
Abalone	15100	11	36
Iris	1400	24	6
Wine	3462	42	7

The figure.5 shows the accuracy level of the proposed work along with various clustering mechanism and it had achieved greater accuracy due to the introduction of Regional Compact Component strategy in traditional K-means algorithm.

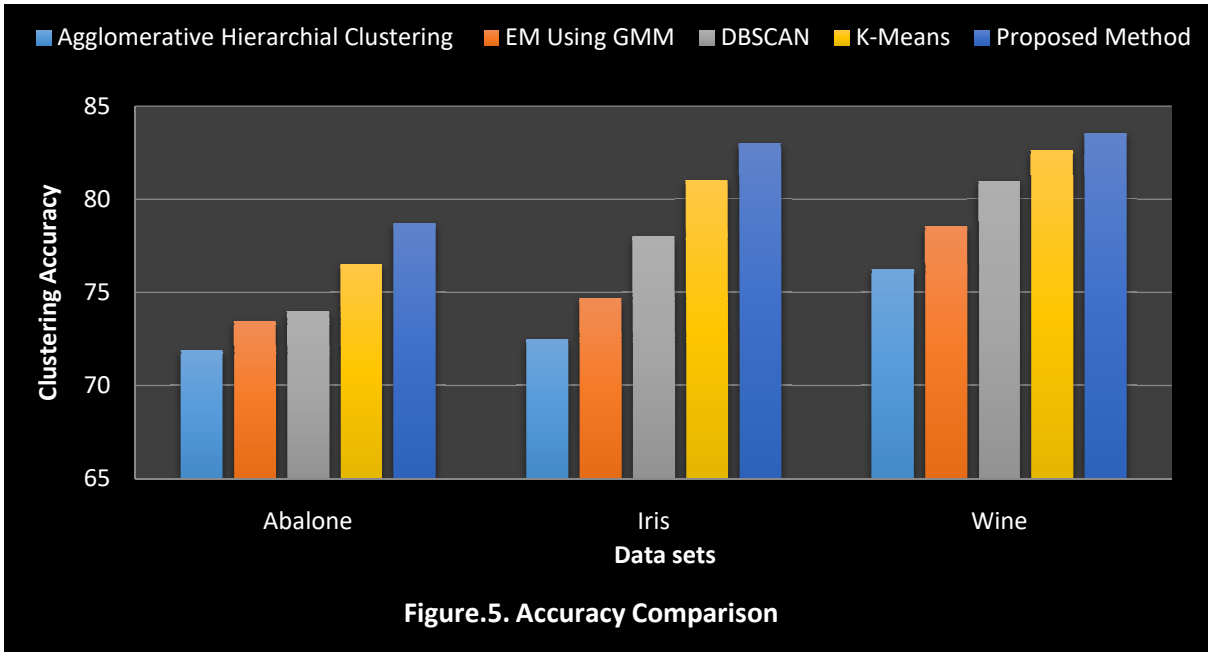
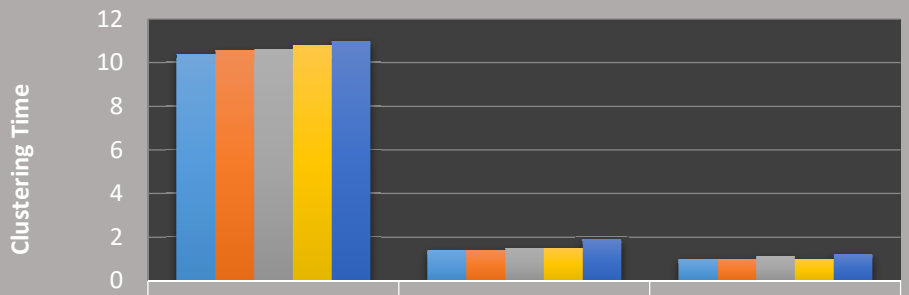


Figure.5. Accuracy Comparison

From the figure.5 it can be clearly noted that the accuracy level of the proposed RCC based improved K-means provides approximately10% higher level of accuracy when compared to other existing clustering algorithms. Furthermore, our proposed algorithm exhibits stability and parallelization efficiency and proves its greater usability.

Figure.6. Time Comparison



	Abalone	Iris	Wine
Agglomerative Hierarchial Clustering	10.39	1.399	0.994
EM Using GMM	10.55	1.4	1
DBSCAN	10.6	1.5	1.11
K-Means	10.8	1.5	1
Proposed Method	11	1.89	1.23

Clustering time comparison is also made with all other clustering strategies and it is illustrated in figure.6. Although the proposed algorithm's clustering time is greater than others, the difference in the time taken of other algorithms is not a considerable amount and proposed method's consumption closely resembles the others.

5. Conclusion

The arbitrary selection of "k" number of initial options in classical K-means algorithm makes it instable and increases complexity of the algorithm and hence its overall performance gets reduced in terms of accuracy. The problem is been overcome by introducing a novel compactness based Regional Centroid Component Recognition (RCC) function into the K-means clustering technique. As illustrated in experimental results it has been achieved the better accuracy than all of the other existing clustering algorithms in almost same amount of time taken by other algorithms.

References:

[1] Han, J., Kamber, M., & Pei, J. **"Data mining concepts and techniques"** third edition. *Morgan Kaufmann*, (2011).

[2] Muja, M., & Lowe, D. G **"Fast approximate nearest neighbors with automatic algorithm configuration."** *VISAPP (1)*, *2*(331-340), 2, (2009).

[3] Tan, Z., Jia, W., &Jin, W. Robust **"Adaptive beamforming using k-means clustering: A solution to high complexity of the reconstruction-based algorithm".** *Radioengineering*, *27*(2), 595-601, (2018).

[4] Cai, Q. Q., Cui, H. G., & Tang, H. **"Big data mining analysis method based on cloud computing".** In *AIP Conference Proceedings* (Vol. 1864, No. 1, p. 020028). AIP Publishing LLC, (2017, August).

[5] Dong, J., & Qi, M. **"K-means optimization algorithm for solving clustering problem".** In *2009 Second International Workshop on Knowledge Discovery and Data Mining* (pp. 52-55). IEEE , (2009, January).

[6] Huang, Z. **"Extensions to the k-means algorithm for clustering large data sets with categorical values".** *Data mining and knowledge discovery*, *2*(3), 283-304, (1998).

[7] Nazeer, K. A., & Sebastian, M. P. **"Improving the Accuracy and Efficiency of the k-means Clustering Algorithm".** In *Proceedings of the world congress on engineering* (Vol. 1, pp. 1-3). London: Association of Engineers, (2009, July)..

[8] Liu, B.**"A fast density-based clustering algorithm for large databases".** In *2006 International Conference on Machine Learning and Cybernetics* (pp. 996-1000), IEEE, (2006, August).

[9] Jianpeng Qi, Yanwei Yu, Lihong Wang**"An effective and efficient hierarchical k-means clustering algorithm".**In 2017,*International Journal of DistributedSensor Networks2017*, Vol. 13(8)

[10] M A Syakur, B K Khotimah, E M S Rochman and B D Satoto **Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile cluster** IOP Conference Series: Materials Science and Engineering, Volume 336, The 2nd International Conference on Vocational Education and Electrical Engineering (ICVEE) 9 November 2017

[11] Sujatha S and Sona A S **New Fast K-Means Clustering Algorithm using Modified Centroid Selection Method. International Journal of Engineering Research & Technology** International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 2, February- 2013

[12] InternatiSiwei Wang; Miaomiao Li; Ning Hu; En Zhu; Jingtao Hu; Xinwang Liu; JianpingYino**"K-Means Clustering With Incomplete Data".**In IEEE Access (Volume: 7)Page(s): 69162 – 69171,(19 April 2019)

[13] KamsonSirait , Tulus , Erna BudhiartiNababan**"K-Means Algorithm Performance Analysis With DeterminingThe Value Of Starting Centroid With Random And KD-TreeMethod".InJournal of Physics: Conference Series,** Volume 930, *International Conference on Information and Communication Technology* (IconICT) 25–26 August 2017, Medan, Sumatera Utara, Indonesia

[14] Cosmin M P, Marian C M, Mihai M **An Optimized Version of the K-Means Clustering Algorithm,** Proceedings of the 2014 Federated Conference on Computer Science and Information Systems (ACSIS) 2 p 695

[15] Singh H and Kaur K **New Method for Finding Initial Cluster Centroids in K-means Algorithm.** In International Journal of Computer Applications, July 2013,Volume 74– No.6