Performance Improvements of k-means Clustering Algorithm

Rakesh P. Badgujar[1], SachinPatel[2], Vijay Bircha[3]

Research Scholar[1], Asst. Professor[2], Professor, Head of Computer Science and Engineering Department[3]
Computer Science & Engineering Department[1],
Swami Vivekanand College of Engineering Indore, India[1]

Abstract: On the basis of the similarity between the instances of the data, grouping or clustering the instances of the dataset regardless of its size is considered as an important part of data mining. There are several efficient algorithms for clustering including K-means clustering algorithm. This algorithm divides the entire dataset into user-defined, k number of clusters. A cluster is a group of similar kind of instances. The aim is to analyze performance in execution time for k-means algorithm. But it is observed that the algorithm takes longer time to produce results. Hence to minimize the overall execution time, a ranking method is used in combination with k-means clustering algorithm. The comparison in execution time for both the approaches is also shown and it is observed that the k-means algorithm with ranking method is time efficient than that of the k-means algorithm alone.

***Keywords*– Data Mining, K-means algorithm, Data Clustering, Ranking Approach**

I. INTRODUCTION

Advanced technologies have resulted in collection of huge data useful for predictive analysis for researches. From huge datasets, it becomes difficult for traditional mining methods to find out relevant information. Analyzing big data is highly essential for the sake of atomization in critical commercially practical applications [1]. K-means clustering method is most widely used to cluster such large datasets for commercial applications as well. Clustering is a method of dividing large data into k groups; these groups form the instances of same kind. Clustering does not depend on the prior knowledge for grouping; rather it learns directly from the dataset given and then clusters the instances of dataset into k groups. The efficiency of k-means algorithm also depends on the centroid selection [2]. There are several methods introduced so as to improve overall algorithm efficiency. Clustering algorithms necessarily demands efficiency as well as cluster quality. The prediction of cluster labels is quite impossible but it can very finely partition data into groups. Many algorithms for clustering depend on heuristic and mathematical formulae. Many solutions to clustering include mathematical programming for implementing various clustering approaches. Most algorithm fail in scalability when the size and dimension of the dataset increases. Section II discusses the theoretical background for k-means clustering algorithm. It describes the author's work for k-means clustering algorithm and its accuracy. Section III describes the overall system working steps to closely understand the algorithmic steps thoroughly. The paper also evaluates the results as well as discussions for the developed system. Finally, the work is concluded and future work is also discussed.

II. THEORETICAL BACKGROUND

Abdul et. al [1] researched for improving traditional k-means clustering method by systematically assigning centroid positions. This approach helped attaining high accuracy in clustering process for even large datasets. Once the centroid is initially selected it is iteratively refined to evaluate other better centroids and this results in more execution time till the convergence is satisfied. Hence this work shows that the centroid selection initially at good level can reduce much time.

Napoleon et. al [2] used uniform distribution method to classify data point within the cluster. This method reduced the time complexity of the traditional approach. This method proved to be helpful in generating the best quality clusters at the end. It reduces overall elapsed time to select initial centroids of each cluster.

Madhuri et. al [3] showed that the iterative k-means approach finds better centroids initially than that of random selection which results in better cluster quality.

Paul et. al [4] used an approach to select centroids so precisely for the very first time, that the algorithm takes lesser time so as to iterate and find the final centroids. The time is reduced if the centroids at initial are selected finely for every cluster.

III. PROPOSED WORK

This section clearly describes the overall working of the system. The very first step of the system is indexing and then testing. At testing phase, a sample image is uploaded to the system and the output is the similar image. At the indexing phase, all the images in the dataset are clustered using k-means clustering algorithm. Then at the searching part, ranking approach is used to search the image similar within the cluster. Following figure represents the overall working of the system.

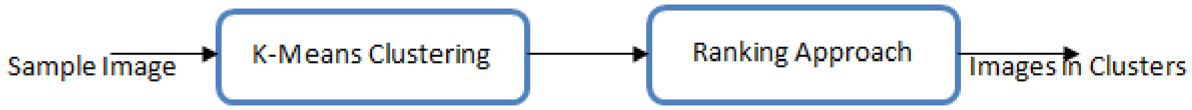


Fig. 1 System Overview

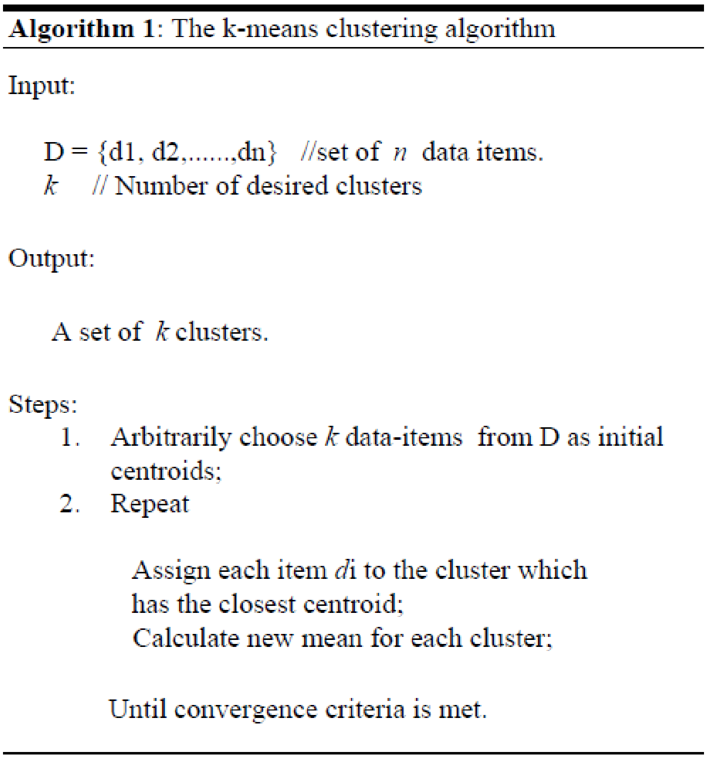
K-Means Clustering:

The k-means clustering algorithm is very advantageous for generating groups from scattered data for all the commercial applications. This algorithm is highly complex in computation when data is huge [5]. Moreover, the accuracy of clustering outcome depends solely on the process of centroid selection. The system uses k-means clustering to cluster the images for the user-defined number of clusters. Many researchers have conducted study to improve the performance of k-means clustering algorithm. K-means algorithm works in two stages [6]:

1. Random selection of k centroids.

2. Place each instance to the cluster of the nearest centroid. The distance of the centroid and the data point is calculated using Euclidean distance formulae.

When the entire data instance is included in a cluster, new centroid is recalculated. This may result in change of centroids for other iterations too. Once the centroid does not move for next loops, the convergence situation is met. Pseudo code for the algorithm is as written below as Algorithm 1. K-means algorithm generally produces accurate results but the time of execution was suffering when the large dataset is uploaded. The major point of concern remains while selecting centroid points initially. Hence this algorithm is computationally very expensive in terms of execution time for huge data.



Algorithm 1: The k-means clustering algorithm

```
Input:

    D = {d1, d2,......,dn}   //set of n  data items.
    k    // Number of desired clusters

Output:

    A set of  k clusters.

Steps:
    1.  Arbitrarily choose k data-items  from D as initial
        centroids;
    2.  Repeat

          Assign each item di to the cluster which
          has the closest centroid;
          Calculate new mean for each cluster;

        Until convergence criteria is met.
```

Elaboration of algorithmic Steps for K-means clustering is as listed below [7, 8, 9]:

The algorithm consists of four steps, initialization, classification, recalculation of centroid and condition of convergence.

Initialization: At the initial step, number of clusters, k is defined first as well as random selection of centroid for each cluster is done at this stage.

Classification: For each other data points that are not initialized as the centroid of the cluster, classification is performed in order to assign these data points to the clusters that are nearest its respective centroids. Euclidean distance is calculated for the data points to the centroid of each cluster. The data point that belongs to the cluster that is nearest to the centroid of the cluster is placed inthose respective clusters.

Centroid Recalculation: Once all the clusters are formed, a new centroid is calculated iteratively.

Condition of Convergence: Stopping condition is when the recalculation of centroids does not change any value. Stopping condition is achieved also when no data point is moved between the clusters. Final convergence appears when a threshold value is met.

The algorithm repeats the centroid calculation till the convergence condition is satisfied.

Advantages of k-means clustering method:

1. Highly robust and fast in execution for limited data instances.
2. Clusters formed finally are all disjoint sets hence no overlapping is observed.

Disadvantages of k-means clustering method:

1. Algorithm is quite complex when it calculates Euclidean distance for all the iterations
2. Centroids are to be predefined manually.
3. It may also generate empty clusters.

System Flow Chart to show the ranking approach:

1. Training image folder is provided for indexing at first.
2. Number of clusters, k is determined by the user.
3. Random centroid is also predefined before algorithm steps startup.
4. Euclidean distance for each data point is calculated so as to classify each data point.
5. Ranking approach is also applied for more accurate cluster formation.
6. All the k clusters are formed finally.

Ranking method is used with k-means clustering with the aim of fastening the traditional algorithm for huge data.

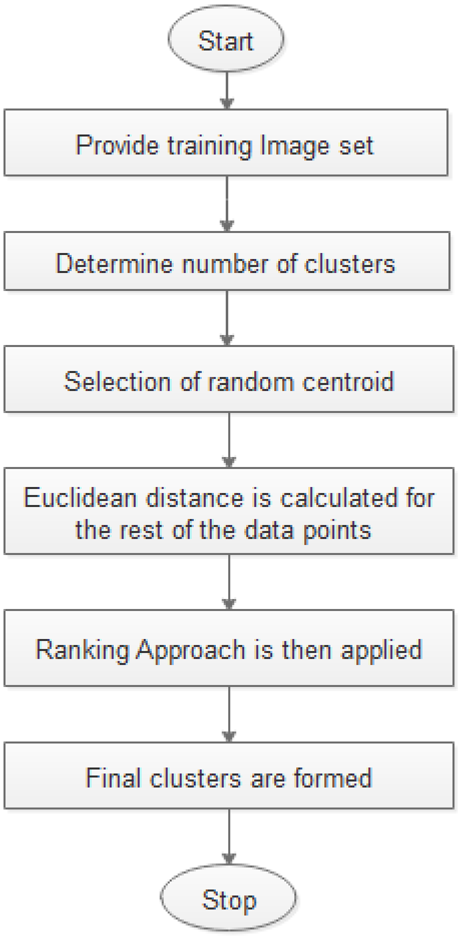

Start
Provide training Image set
Determine number of clusters
Selection of random centroid
Euclidean distance is calculated for the rest of the data points
Ranking Approach is then applied
Final clusters are formed
Stop

Fig. 2 System Flow Chart

IV. Research Objective

The aim of this project is to implement a new kind of add-on to the existing k-means clustering algorithm so as to get faster solution to the large set of records. It is observed that the overall execution time is comparatively slower when it comes to execution of large set of records. This project work aims to use ranking method in combination with k-means clustering algorithm so as to get better results in terms of time comparison.

V. Experimental Results And Discussion

This section describes the comparison of traditional k-means clustering and its combination with ranking method in regards to the execution time. For 10 set of datasets, for different number of records, the execution time is monitored thoroughly. The number of records taken is from 50 to 500. The execution time comparison is performed for both the approaches: traditional k-means clustering algorithm and its combination with ranking method. As shown in Table 1, the k-means clustering takes comparatively more time than that of the ranking approach.

Table 1 Comparison of execution time for K-means clustering and ranking approach

Records	**Execution Time for traditional k-means Clustering Algorithm**	**Execution Time for Ranking approach in combination of k-means clustering**
50	98	91
100	132	121
150	198	167
200	209	190
250	287	267
300	309	310
350	380	326
400	390	376
450	467	422
500	487	476

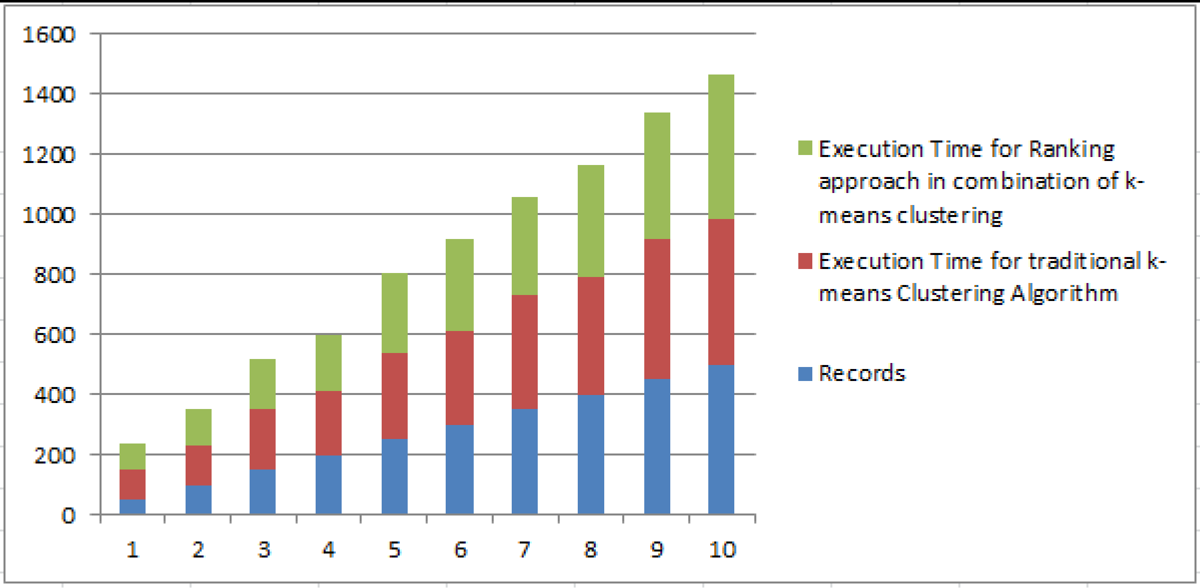


Fig. 3 Experimental Results in comparison of both the approaches

Fig. 3 represents the graphical representation of execution time. As the figure describes, the execution time for the ranking method is comparatively less than that of k-means clustering algorithm. The difference is suitably noted in the figure below. The number of records varied for experiments were 50 to 500. The motive was to clearly show that the ranking method can work better than traditional k-means clustering algorithm.

VI. Conclusions and Future Scope

This paper implements the ranking method in combination with traditional k-means clustering method. The execution time is the point of focus to analyze the efficiency of both the approaches. As stated earlier, traditional k-means clustering algorithm works like charm for the limited number of records but the hectic part is to make the algorithm work faster with huge data. Since traditional k-means clustering cannot work fine with large data our idea is to use ranking method in combination so as to fasten the clustering process for large datasets. The experiments clearly show that the ranking method is highly beneficial with k-means clustering algorithm for huge data as well.

Future scope is to focus on space complexity as well and to find another way to minimize space complexity of the application. Since these days computers are of high configuration does not really care about space but there are certain heavy applications that uses more memory space and at that time, space complexity becomes highly important.

References

[1] K. A. Abdul Nazeer & M. P. Sebastian "Improving the Accuracy and Efficiency of the K-Means Clustering Algorithm". Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, London, U.K, July 1 - 3, 2009.

[2] D. Napoleon & P. Ganga Lakshmi, "An Efficient K-Means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points", IEEE, 2010.

[3] Madhuri A. Dalal & Nareshkumar D. Harale "An Iterative Improved k-means Clustering" Proc. of Int. Conf. on Advances in Computer Engineering, 2011.

[4] Paul S. Bradley & Usama M. Fayyad, "Refining Initial Points for K-Means Clustering", 15th International Conference on Machine Learning, ICML98.

[5] Osama Abu Abbas "Comparison of various clustering algorithms" The International Arab Journal of Information Technology, Vol. 5, No. 3, July 2008.

[6] Jirong Gu & et.al, "An Enhancement of K-means Clustering Algorithm ", IEEE International Conference on Business Intelligence and Financial Engineering, 2009.

[7] Dost Muhammad Khan & Nawaz Mohamudally "A Multiagent System (MAS) for the Generation of Initial Centroids for k-means clustering Data Mining Algorithm Based on Actual Sample datapoints", IEEE, 2009.

[8] Malay K. Pakhira, "Clustering Large Databases in Distributed Environment ", IEEE 2009 WEE International.

[9] Shi Na & et.al,"Research on k-means Clustering Algorithm", IEEE Third International Symposium on Intelligent Information Technology and Security Informatics, 2010.

[10] Jaehui Park, Sang-goo Lee "Probabilistic Ranking for Relational Databases based on Correlations" ACM 2010.

[11] Amir Ben-Dor, Ron Shamir and Zohar Yakini, "Clustering Gene Expression Patterns," *Journal of Computational Biology*, 6(3/4): 281-297, 1999

[12] Chaturvedi J. C. A, Green P, "K-modes clustering," J. Classification, (18):35–55, 2001.

[13] Daxin Jiang, Chum Tong and Aidong Zhang, "Cluster Analysis for Gene Expression Data*," IEEE Transactions on Data and Knowledge Engineering*, 16(11): 1370-1386, 2004.

[14] Fahim A.M, Salem A. M, Torkey A and Ramadan M. A, "An Efficient enhanced k-means clustering algorithm," *Journal of Zhejiang University*, 10(7):1626–1633, 2006.

[15] Huang Z, “Extensions to the k-means algorithm for clustering large data sets with categorical values”, *Data Mining and Knowledge Discovery*, (2):283–304, 1998.

[16] Jiawei Han M. K, “Data Mining Concepts and Techniques”, *Morgan Kaufmann Publishers*, An Imprint of Elsevier, 2006.

[17] Margaret H. Dunham, “Data Mining- Introductory and Advanced Concepts”, *Pearson Education*, 2006.

[18] McQueen J, “Some methods for classification and analysis of multivariate observations,” *Proc. 5th Berkeley Symp. Math. Statist. Prob.*, (1):281–297, 1967.

[19] Pang-Ning Tan, Michael Steinback and Vipin Kumar, “Introduction to Data Mining”, *Pearson Education*, 2007.

[20] Stuart P. Lloyd, “Least squares quantization in pcm,” *IEEE Transactions on Information Theory*, 28(2): 129-136.

[21] Yuan F, Meng Z. H, Zhang H. X and Dong C. R, “A New Algorithm to Get the Initial Centroids,” *Proc. of the 3rd International Conference on Machine Learning and Cybernetics*, pages 26–29, August 2004.