Improvised K-Means Algorithm

Saranya Pachamuthu, Atul Agarwal and Sahil
Department of Computer Science and Engineering
SRM University

saranya.p@srmuniv.ac.in
atul_agarwal@srmuniv.edu.in
sahil_ramchander@srmuniv.edu.in

***Abstract*—Improving upon the K-means algorithm is the main method used in data analytics. The k-means clustering algorithm is the main technique that is designed for various practical applications. But originally the k-means algorithm was computationally expensive and the final storage directly correlates to the correction of the initial Centroids, which are elected by the algorithm at random. Many modifications have already been suggested to improve upon the performance of k-means, but a large number of them requires supplementary inputs as inception values for the quantity of data points available in a set collected. This article proposes a different approach to find the best initial centroids and presents an even more efficient way to allocate data points to appropriate clusters thus reducing the complexity of time. The proposed algorithm's implementation is well facilitated and improve the K-means efficiency which requires a simple data structure to maintain certain information in every single one of the iterations to be used in next.**

Keywords = Clustering, Data Analytics, K - Means, Centroids, Optimization, Data Accuracy

1. INTRODUCTION

[i]In today's world rapid advancements in science and technology chiefly in the field of scientific data collection methods of high dimensionality, insensitivity to order of attributes, have resulted in the large-scale accumulation of promising interoperability and usability and thus a need for handling such data is much required. [ii]Cluster analysis is a technique pertaining to diverse fields of computer science and one of the leading data analytics tool obtainable in the data mining technology today. Owing to the development of novel Clustering algorithms which are mainly divided into two techniques for generating and accumulating data, the rate of these categories: Hierarchical algorithms and Partition growth of scientific databases has become tremendous algorithms.

[iii]Current hierarchical clustering algorithms separates the data set into smaller subsets and inferences the hierarchical information from it by using traditional database fashion; hence it is practically impossible to extract any useful information from the provided data. A partition clustering algorithm partitions the analysis techniques for the same aforementioned objective.

[iv]The effective mining methods are those which convert the given data sets into desired number of sets in least number of steps possible. It is thus absolutely necessary to infer as much of the implicit information available and just for that, numerous methods have been already proposed to solve large databases clustering problem among which one of the clustering methods is k-means clustering algorithm developed by Mac Queen Applications in data mining, statistical data analysis, and data in 1967.

[v]The tolerance, robustness and ease of use of k-means clustering algorithm in compression and vector quantization is what made this algorithm to be used in several fields where segregation of data into groups of similar objects is required.

[vi]Once the algorithm is applied, each group of the k-means clustering algorithm that are created should mainly consists of the objects that are relatable between themselves in some way or the other, since its intelligence towards the

cluster massive data rapidly and dissimilar to objects of other groups it is widely used in all kinds of fields. From the efficient, but computational complexity of the machine's learning perspective, Clustering can be viewed as original k-means algorithm is very high, especially for unsupervised learning of concepts. [vii]The machine learning can be achieved by using this algorithm as the clustering doesn't depend on various types of clusters depending on random variables over predefined classes and training samples simultaneously choosing initial centroids and thus there is an effectiveness in classifying the data objects.

[viii]Two types of clustering have been studied - clustering the documents on the basis of the distributions of words that co-occur in the documents, and clustering the words using the distributions of the documents in which they occur. In this algorithm we have used a double-clustering approach in which we first cluster the words and then use this word-cluster to cluster of the documents. The clustering of words reduces the feature space and thus reduces the noise and increases the time efficiency. The goal of a document clustering scheme is to minimize intra-cluster distances between documents, while maximizing inter-cluster distances (using an appropriate distance measure between documents).

In general, this algorithm can be used for clustering of objects based on their features and characteristics. A recently introduced principle, termed the information bottleneck method is based on the following simple idea. [ix]Given the empirical joint distribution of two variables, one variable is compressed so that the mutual information about the other is preserved as much as possible. Here the two variables are the object and the features. First, the features are clustered to preserve the information of objects and then these clusters are used to reduce the noise in the object graph. By clustering the articles, we could reduce our domain of search for recommendations as most of the users had interest in let's say news corresponding to a few number of clusters. This improved our time efficiency to a great extent. Also, we could identify the articles of same news from different sources.

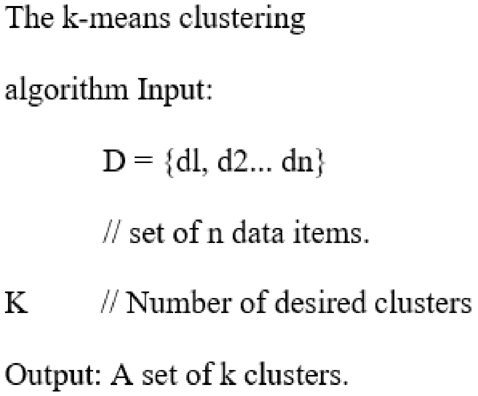
In clustering, it is the dispersal and the nature of data that will regulate cluster relationship, in opposition to the classification where the classifier learns the association between objects and classes from a so-called training set, i.e. a set of data correctly labeled by hand, and then replicates the learnt behavior on unlabeled data.

2. K-MEANS CLUSTERING ALGORITHM (The Existing Algorithm)

[x]Our process is to create k disjoint clusters from a given set of data, in which the value of k is already set. The algorithm has two phases: The first consists in defining k centroids, one for each group. The next step is to take every point fitting to the data establish it and associate it with the closest centroid according to their Euclidean distance. [ix]Which is generally considered as the distance between data points and centroids. After all points are included in some clusters, the first step is completed and the initial group is ready. Now we must recalculate the new centroids, since the addition of new points can lead to a change in the group's centroids. Once you find new k centroids and a new link will be created between them, data points and the new centroid closer, generating a cycle.

As a result of this cycle, k centroids can change their position gradually. In the end, a situation is achieved where the centroids no longer move. This indicates the junction criterion for grouping and thus a natural stopping point of the process.

Algorithm:



The k-means clustering

algorithm Input:

D = {dl, d2... dn}

// set of n data items.

K // Number of desired clusters

Output: A set of k clusters.

Steps:

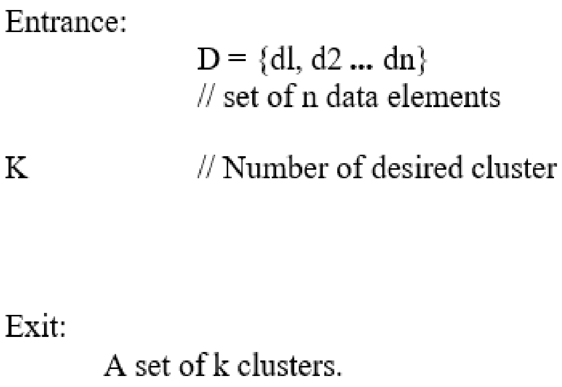
I. Randomly select k data-items from D as primary centroids.

II. Reiterate to allocate each item di to the cluster which has the nearby centroid; Compute new mean for each cluster;

III. Till junction criteria was met.

[xi]The K-means algorithm seems to provide us with partitions which are efficient in terms of variance that is verified to major extend by mathematical analysis and practical experience and the k-means procedures are comparatively easily programmed and thus it is computationally economical which makes it feasible to process huge samples on a normal personal computer.

3. ENHANCED ALGORITHM (Improvised K-Means Algorithm)



```
Entrance:
                D = {dl, d2 ... dn}
                // set of n data elements

K               // Number of desired cluster

Exit:
    A set of k clusters.
```

Steps 1: fix the initial centroids for the groups using the algorithm.

Step 2: allocate each of the data point to the suitable clusters for using the algorithm

Steps 3: calculate the distance between the data point and the closest center chosen in the last iterations for each of the data points.

Steps 4: select another data point (new) as a center using a weighted probability distribution.

Steps 5: repeat the above steps until the centers are chosen and the proceed with the standardized format of k-means algorithm.

4. RESULTS

The k-Means algorithm is advanced with a first the paradigm the standard, and then it is followed by our improved k-means algorithm. The improved k- sign algorithm can be used to regulate the cluster's centroid. The investigative results are deliberated for the K-means the algorithm take the duration for which the complexity is greater in various data sets. The provided clusters from the normal K-Means algorithm is presented. The normal distribution of data points is taken to easily implement and take the steps of convenience for our data sets. The number of clusters and data the points are given by the user during the execution of Program. The number of data points is let's say a 1000 then the number of the cluster data is 10 (k = 10).

The algorithm is repeated in order to assign centroids for efficient output. The cluster's centroids are calculated for each cluster's average value and so the clusters are formed contingent on the distance between the data points. For diverse input data points, the algorithm provides different types of output. Improved k-means is better than k-means in experimental results. In the cluster size it must be diverse in a different run.

The first cluster size in run l is ingenious out of 99 in their quality of cluster size. The improvement sign can be seen by comparing the average time for five different runs and clearly the average time for our algorithm (43.19) is less than the average time taken by the standard k-means (62.2).

5. CONCLUSION

Time complexity is measured by CPU's time elapsed working on each of the algorithms. Usually, the time complexity is highly dependent on the processor and thus, differs from one processor to another, which in turn relies on its speed and thus the whole system.

As, our derived algorithm works well for decision making spherical shaped clusters in various types of data points. The advantage of the K-Means algorithm is its favorable execution time whereas its drawback is that the user has to know beforehand that how many clusters are to be searched for. Our examination of the proposed process shows that the k- means is favorable for smaller data sets and our algorithm is preferred where there are huge data sets.

6. ACKNOWLEDMENT

We would like to take this opportunity to express our deepest appreciation to our guide and mentor Mrs. P. Saranya. Her invaluable and continual guidance and encouragement providedus with an exceptional atmosphere which gravely helped us in doing our research. All through our work, in spite of her hectic schedule, she gave us admirable support for completing this research work.

7. FUTURE WORK

Our method's ability, according to our understanding, gains a considerable improvement over randomly chosen initiation points is majorly due to our capability to evade the empty clusters' problem that affects the standard K- Means. Because throughout the refinement process we reorganize the null clusters to distant points and iterate the K-Means algorithm and thus an initiation point attained by our improvement method is less probable to lead the following clustering algorithm to a "bad" solution.

But there can be various ways to improve the existing algorithm further more depending on the data set as well as the available resources.

Our proposed method can be normalized using the techniques mentioned above, to the common type of data set available and used by the data scientists around the world. There is still a lot of work to be done in this field in order to achieve the maximum efficiency possible for better results. And also, there is always a probability that there may be some other algorithm that can be optimized so much that it can outperform k-means or a new algorithm in itself may be created for more accurate and optimized solutions also but that's a very unlikely case.

8. REFERENCES

i. J. Banfield and A. Raftery, "Model-based gaussian and non-Gaussian Clustering", Biometrics, vol. 49: 803-821, pp. 15-34, 1993.

ii. C. Bishop, 1995. Neural Networks for Pattern Recognition. Oxford University Press.

iii. P. S. Bradley, O. L. Mangasarian, and W. N. Street. 1997. "Clustering via Concave Minimization", in Advances in Neural Information Processing Systems 9,M. C. Mozer, M. I. Jordan, and T. Petsche (Eds.) pp 368-374, MIT Press, 1997.

iv. P. S. Bradley, U. Fayyad, and C. Reina, "Scaling Clustering Algorithms to Large Databases", To appear,Proc. 4th International Conf. on Knowledge Discovery and Data Mining (KDD-98). AAAI Press, Aug. 1998.

v. P. Cheeseman and J. Stutz, "Bayesian Classification (AutoClass): Theory and Results", in [FPSU96], pp. 153-180. MIT Press, 1996. 5

vi. A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM algorithm". Journal of the Royal Statistical Society, Series B, 39(1): 1-38,1977.

vii. R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis. New York: John Wiley and Sons. 1973

viii. U. Fayyad, D. Haussler, and P. Stolorz. "Mining Science Data." Communications of the ACM 39(11), 1996. Fayyad, U., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (Eds.) Advances in Knowledge Discovery and Data Mining. MIT Press, 1996.

ix. U. Fayyad, C. Reina, and P. S. Bradley, "Refining Initialization of Expectation Maximization Clustering Algorithms", To appear, Proc. 4thIntemational Conf. on Knowledge Discovery and Data Mining (KDD-98). AAAI Press, Aug. 1998.

x. D. Fisher. "Knowledge Acquisition via Incremental Conceptual Clustering". Machine Learning, 2:139-172,1987.

xi. E. Forgy, "Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications", Biometrics 21:768.1996