[[1]](#footnote-0)

**IMPROVISED KMEANS EFFICENCY**

By:- Atul Agarwal and Sahil

***Abstract*—**Improving Kmeans analysis is one of the main methods of data analysis and the k-means clustering algorithm. The main technique that is used for many practical applications. But the original k-means algorithm is computationally expensive and the final storage depends a lot on the correction of the initial Centroids, which are selected at random. Many improvements have already been proposed to improve performance of k-means, but most of require supplementary inputs as inception values for the number of data points in a set. This article proposes a new method to find the best initial centroids and provide an efficient how to assign data points to appropriate clusters. Reduce the complexity of time. This algorithm is easy to implement, and improve the Kmeans efficiency which requires a simple data structure to maintain certain information in each iteration to be used in next iteration.

**INTRODUCTION**

Advances in scientific data collection methods high dimensionality, insensitivity to order of attributes, have resulted in the large scale accumulation of promising interoperability and usability. Cluster analysis is a data pertaining to diverse fields of science and one of the primary data analysis tool in the data mining technology. Owing to the development of novel Clustering algorithms are mainly divided into two techniques for generating and collecting data, the rate of categories: Hierarchical algorithms and Partition growth of scientific databases has become tremendous algorithms. A hierarchical clustering algorithm divides Hence it is practically impossible to extract useful the given data set into smaller subsets in hierarchical information from them by using conventional database fashion. A partition clustering algorithm partition the analysis techniques.

Effective mining methods are data set into desired number of sets in a single step. Absolutely essential to unearth implicit information from numerous methods have been proposed to solve huge databases clustering problem. The most popular clustering method CLUSTERING is an important tool for a variety of is k-means clustering algorithm developed by Mac Queen Applications in data mining, statistical data analysis, data in 1967. The easiness of k-means clustering algorithm compression and vector quantization. Clustering is a made this algorithm used in several fields division of data into groups of similar objects.

Each group the k-means clustering algorithm is more prominent consists of objects that are similar between themselves since its intelligence to cluster massive data rapidly and dissimilar to objects of other groups. From the efficiently. But the computational complexity of the machine learning perspective, Clustering can be viewed as original k-means algorithm is very high, especially for unsupervised learning of concepts. Unsupervised large data sets. Moreover, this algorithm results in machine learning means that clustering does not depend different types of clusters depending on the random on predefined classes and training examples while choice of initial centroids. The effectiveness of a classifying the data objects.

**KMEANS CLUSTERING ALGORITHM**

Our process is to classify a given set of data in k number of disjoint clusters, in which the value of k is set go ahead .The algorithm consists of two separate phases: The first phase consists in defining k centroids, one for each group. The next step is to take every point belonging to the data establish it and associate it with the nearest centroid Euclidean distance .It is generally considered that determines the distance between data points and centroids. When all the points are included in some clusters, the first step is completed the initial group is ready. At this point we must recalculate the new centroids, since the inclusion of new points can lead to a change in the group's centroids. Once you find new k centroids, 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 them position gradually. In the end, a situation be reached where the centroids no longer move. This indicates the convergence criterion for grouping.

**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;
3. Until convergence criteria is met.

K-means appears to give partitions which are reasonably efficient in the sense of within class variance, corroborated to some extend by mathematical analysis and practical experience. Also, the k-means procedure is easily programmed and is computationally economical, so that it is feasible to process very large samples on a digital computer.

**ENHANCED ALGORITHM (Improvised Kmeans Algorithm)**

Entrance:

D = {d1, d2... dn} // set of n data elements

K // Number of desired clusters

Exit:

A set of k clusters.

Steps:

Phase 1: determine the initial centroids of the groups for using algorithm 3.

Step 2: assign each data point to the appropriate clusters for using algorithm 4

**RESULTS**

The k-Means algorithm is advanced with a first the paradigm, followed by an advanced k-means algorithm. The improved k-sign algorithm can be used to determine the cluster centroid. The investigative results are discussed for the Kmeans the algorithm must take the time for which the complexity is greater different data set. The resulting clusters of the normal K-Means distribution the algorithm is presented. The normal distribution data Points are taken to easily implement and take the steps of convenient 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 1000 and the number of the cluster data is 10 (k = 10).

The algorithm is repeated to allocate times for efficient output. The cluster centres (centroids) are calculated for each cluster average value and cluster are formed depending on the distance between the data points. For different 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 different in a different run. The k-indicates that the algorithm for the elapsed time is 2343.3ms. So the first cluster size is 84 in run1. It can be measures of the quality of the clusters.

The efficient method has passed the time is 62.2ms. So the first cluster size in run1 is ingenious out of 99 their quality of cluster size. The method of execution can be performing five average run. The time can be taken from 2116.26 in k means algorithm. The improved k-sign method can be taken the average time is 43.194. The average time improved it is less than k-means.

**CONCLUSION**

The time complexity can be calculated by CPU elapsed time for different two algorithms. As a rule the time complexity varies from one processor to another processor, which depends on the speed and the type of the system. Thus, our derived algorithms work well for decision spherical-shaped clusters in different type of data points. The advantage of the K-Means algorithm is its favourable execution time. Its drawback is that the user has to know in advance how many clusters are searched for. From the experimental results (by many execution of the programs), it is practical that K-Means algorithm is efficient for smaller data sets and advanced Kmeans algorithm seems to be efficient for huge data sets than K-means algorithm.

References

[BR93] J. Banfield and A. Raftery, “Model-based

gaussian and non-Gaussian Clustering”, Biometrics, vol.

49: 803-821, pp. 15-34, 1993.

[B95] C. Bishop, 1995. Neural Networks for Pattern

Recognition. Oxford University Press.

[BMS97] 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.

[BFR98] 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.

[CS96] P. Cheeseman and J. Stutz, “Bayesian

Classification (AutoClass): Theory and Results”, in

[FPSU96], pp. 153-180. MIT Press, 1996.

[DLR77] 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.

[DH73] R.O. Duda and P.E. Hart, Pattern Classification

and Scene Analysis. New York: John Wiley and Sons.

1973

[FHS96] U. Fayyad, D. Haussler, and P. Stolorz. “Mining

Science Data.” Communications of the ACM 39(11),

1996.

[FPSU96] Fayyad, U., G. Piatetsky-Shapiro, P. Smyth,

and R. Uthurusamy (Eds.) Advances in Knowledge

Discovery and Data Mining. MIT Press, 1996.

[FRB98] U. Fayyad, C. Reina, and P. S. Bradley,

“Refining Initialization of Expectation Maximization

Clustering Algorithms”, To appear, Proc. 4th

International Conf. on Knowledge Discovery and Data

Mining (KDD-98). AAAI Press, Aug. 1998.

[F87] D. Fisher. “Knowledge Acquisition via Incremental

Conceptual Clustering”. Machine Learning, 2:139-172,

1987.

[F65] E. Forgy, “Cluster analysis of multivariate data:

Efficiency vs. interpretability of classifications”,

Biometrics 21:768. 1965.

1. [↑](#footnote-ref-0)