A New Efficient Approach towards k-means Clustering Algorithm

Pallavi Purohit
Department of Information Technology, Medi-caps Institute of Technology, Indore

Ritesh Joshi
Department of Master of Computer Application, Medi-caps Institute of Technology, Indore

ABSTRACT
K-means clustering algorithms are widely used for many practical applications. Original k-mean algorithm select initial centroids and medoids randomly that affect the quality of the resulting clusters and sometimes it generates unstable and empty clusters which are meaningless. The original k-means algorithm is computationally expensive and requires time proportional to the product of the number of data items, number of clusters and the number of iterations.

The new approach for the k-mean algorithm eliminates the deficiency of exiting k mean. It first calculates the initial centroids k as per requirements of users and then gives better, effective and good cluster without scarifying Accuracy. It generates stable clusters to improve accuracy. It also reduces the mean square error and improves the quality of clustering. We also applied our algorithm for the evaluation of student's academic performance for the purpose of making effective decision by the student councilors.

Keywords
Cluster analysis, Centroids, K-mean.

1. INTRODUCTION
Unsupervised learning is the part of machine learning whose purpose is to give the ability to machine to find some hidden structure within data. Typical task in unsupervised learning include the discovery of "natural" clusters present in the data, finding a meaningful low dimensional representation of the data or learning explicitly a probability function that represents the true distribution of the data. The clustering problem is classical problem of database, knowledge discovery, artificial intelligence and theoretical literature is use to find similar groups of record from very large datasets [6]. Given a training data set, the goal of a clustering algorithm is to group similar data points in the same cluster while putting dissimilar data points in different clusters. Clustering is used in a wide variety of fields: biology, statistics, pattern recognition, information retrieval, machine learning, psychology, and data mining. For example, it is used to group related documents for browsing, to find genes and proteins that have similar functionality, to find the similarity in medical image database, or as a means of data compression. Clustering is an important branch of pattern recognition, and it aims at modeling fuzzy (i.e., ambiguous) unlabeled pattern efficiently [1].

There are a number of clustering methods which can be

classified into following categories: Partitioning methods, Hierarchical methods, Density-based methods, Grid-based methods, Model-based methods [10]. Each of these methods handles some issues related to clustering but, there is not a single universal clustering algorithm that can handle all the issues related to it [9]. With regard to the problem of partitioning N objects into k classes, to get the best clustering is a NP-hard problem. It is a well-known fact that the standard k-means algorithm gets easily trapped in a local minimum.

In Section-2 we have describe procedure of cluster analysis. In section-3 we have described advantages and limitations of existing K-mean algorithm. In Section-4 we discuss a new approach of variation of k mean. In section Section-5 we discuss the performance study of existing k mean and variation of k mean. Finally Section-6, Section-7 and Section-8 contain the conclusion, future work and references respectively.

2. PROCEDURE OF CLUSTER ANALYSIS
Cluster analysis is mainly divided into four basic steps as shown in Figure: 1[3]

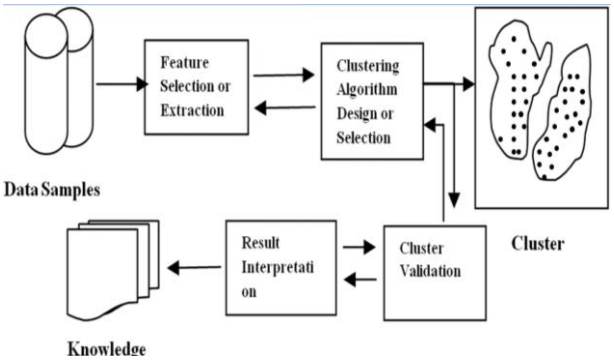

Figure 1: Clustering procedure steps

2.1 Feature Selection or Extraction
Feature selection is the process of discovering the most relevant attribute of a dataset to the data mining task. It is commonly used and powerful technique for reduction the dimensionality of a problem to more manageable task. Feature extraction utilizes some transformations to generate useful and novel features from original ones. It does not remove any of the original attribute from further consideration, This technique is best suited to dataset where most of the dimensions are relevant to the clustering task, but may are highly correlated or redundant. Generally the ideal features should be of use in distinguishing patterns belonging to different clusters, immune to noise, easy to extract and interpret [2].

2.2 Clustering Algorithm design and selection

In this step, the proximity (similarity or dissimilarity) measure and criterion function is selected. Proximity measure greatly affects the resulting clusters. Almost all clustering algorithm are explicitly or implicitly connected to some definition of proximity measure. Once the proximity measure is chosen, the criterion function is selected in order to optimize clustering problem, which is well defined mathematically (e.g. square error function). There are lots of clustering algorithms has been developed to solve different issues related to clustering in variety of fields, but there is no clustering algorithm that can be universally used to solve all problems. Therefore, it is important to carefully select and design the clustering algorithm which satisfies the characteristics of the specified problem.

2.3 Cluster Validation

It is difficult to identify that whether the clusters generated are of meaningful or just an artifact of an algorithm. Each clustering algorithm divides the given dataset into number of partition, without worrying about whether there exists any structure or not. Moreover, different clustering algorithm generates different result for the same dataset, and even some algorithm generates different result for different result for different set of parameters or different order of input data. Therefore there must be some evaluation standards and criteria to provide the user with the degree of confidence for the clustering results derived from the used algorithm.
There are three methods of validating criteria: [5]
External indices: based on prior knowledge and used as a standard to validate clustering solutions.
Internal indices: independent or prior knowledge. They examine the clustering structure directly from the original data.
Relative criteria: compares different clustering structure to decide which one may best reveal the characteristics of the objects.

2.4 Result Interpretation

The goal of the clustering algorithm is to extract the important hidden information from the original dataset and to provide user with meaningful insights. The result should be easily interpretable and usable by the user. The above Figure: 1 shows the feedback pathway, because it is possible that clustering algorithm may iterate for several times to find the optimal solution, or to find optimal value of parameters or select appropriate features.

3. REVIEW OF EXISTING K-MEAN CLUSTERING:

3.1 Distance Calculation

The distance between two points is taken as a common metric to assess the similarity among the components of a population. The most commonly used distance measure is the Euclidean metric which defines the distance between two points p = (p1, p2,…) and q = (q1,q2, ….) as

$$d = \sqrt{\sum (p - q)^2} \quad (1)$$

3.2 Cluster Seed

First document or object of a cluster is defined as the initiator of that cluster i.e. every incoming object's similarity is compared with the initiator. The initiator is called the cluster seed.

3.3 Existing K-mean

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result [2]. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point, this method needs to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function [9]

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad (2)$$

Where $\| x_i^{(j)} - c_j \|2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j , is an indicator of the distance of the n data points from their respective cluster centers [4].

The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4.Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

4. PROPOSED ALGORITHM

In proposed Algorithm of k-mean, for better result two main tasks are done. Instead of initial centroids are selected randomly, for the stable cluster the initial centroids are determined systematically. It calculates the Euclidean distance between each data point and selects two data-points between

which the distance is the shortest and form a data-point set which contains these two data-points, then we delete them from the population. Now find out nearest data point of this set and put it into new set. The numbers of elements in the set are decided by initial population and number of clusters systematically

Our Proposed Algorithm is as follow:

1. Set p = 1
2. Compute the distance between each data point and all other data- points in the set D
3. 3Find the closest pair of data points from the set D and form a data-point set Am (1<= p <= k) which contains these two data- points, Delete these two data points from the set D
4. 4Find the data point in D that is closest to the data point set Ap, Add it to Ap and delete it from D
5. 5Repeat step 4 until the number of data points in Am reaches (n/k+1)
6. If p<k+1, then p = p+1, find another pair of data points from D between which the distance is the shortest, form another data-point set Ap and delete them from D, Go to step 4
7. For each data-point set Am (1<=p<=k+1) find the arithmetic mean of the vectors of data points Cp(1<=p<=k+1) in Ap.
8. Select nearest object of each Cp(1<=p<=k+1) as initial centroid.
9. Compute the distance of each data-point di (1<=i<=n) to all the centroids cj (1<=j<=k+1) as d(di, cj)
10. For each data-point di, find the closest centroid cj and assign di to cluster j
11. Set ClusterId[i]=j; // j:Id of the closest cluster
12. Set Nearest_Dist[i]= d(di, cj)
13. For each cluster j (1<=j<=k+1), recalculate the centroids
14. Repeat
15. For each data-point di

15.1 Compute its distance from the centroid of the present nearest cluster

15.2 If this distance is less than or equal to the present nearest Distance, the data-point stays in the cluster, Else

15.2.1For every centroid cj (1<=j<=k+1) Compute the distance (di, cj); End for

15.2.2Assign the data-point di to the cluster with the nearest Centroid Cj

15.2.3 Set ClusterId[i] =j

15.2.4 SetNearest_Dist[i = d (di, cj); End for

16. For each cluster j (1<=j<=k+1), recalculate the centroids; until the convergence Criteria is met.

5. PERFORMANCE STUDY

Figure 3 shows the performance of accuracy study which has been carried out on same size of datasets. The accuracy of the model has been tested for both existing K -mean and new approach of K -means method. The experiment shows that the accuracy is significantly increase in new approach of k mean.

Dataset: Ecoli

Algorithm	Cluster seed	Mean Square Error	Accuracy (%)
K-mean	891350	93.57	80.95
	123456	81.18	82.44
	456539	60.66	88.69
	237854	61.26	91.36
New K-mean	-----	48.26	92.85

Table I Accuracy & MSE performance (Ecoli dataset)

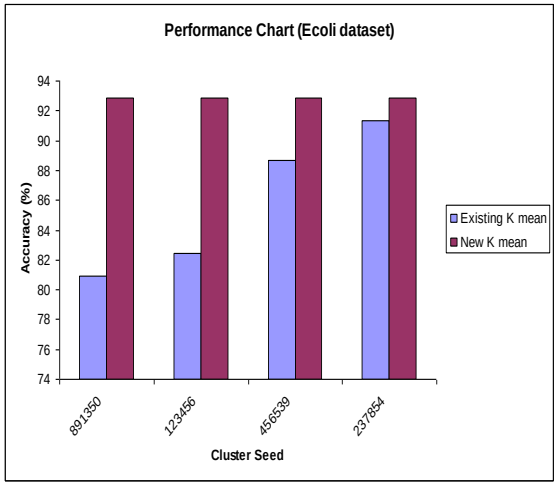

Figure 2. Accuracy performance chart (Ecoli Dataset)

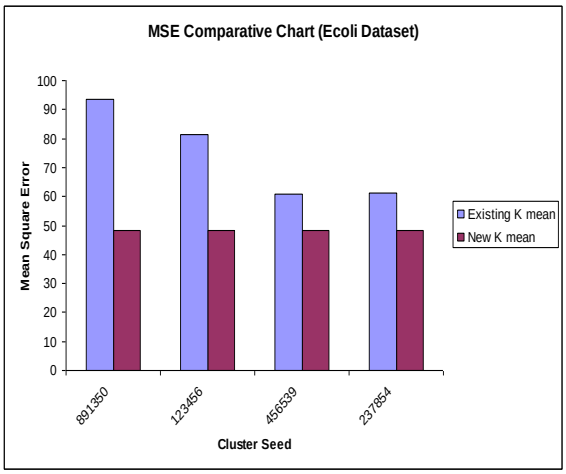

Figure 3 MSE Comparison chart (Ecoli Dataset)

the performance of accuracy study and Figure 3 shows mean square error comparison which has been carried out on Vehicle datasets.

Dataset: Vehicle

Algorithm	Cluster seed	Mean Square Error	Accuracy (%)
K- mean	123456	7085.63	74.72
	347698	5869.40	85.12
	763451	5816.62	81.54
	884712	5816.36	82.28
	995634	7029.56	72.61
New K-mean	-----	5849.38	87.86

Table II Accuracy & MSE performance (Vehicle dataset)

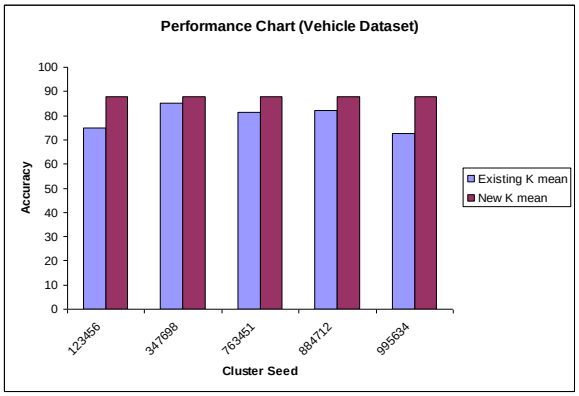

Figure 4 Accuracy performance chart (Vehicle Dataset)

6. CONCLUSION

A New k-mean algorithm which In new Approach of classical partition based clustering algorithm improve the execution time of k-means algorithm, with no miss of clustering quality in most cases. From our result we conclude that, the second proposed implementation of the k-means algorithm is the best one. From experiment we observe that proposed algorithm give more accuracy for dense dataset rather than sparse dataset.

7. REFRENCES

[1] Dechang Pi, Xiaolin Qin and Qiang Wang, "Fuzzy Clustering Algorithm Based on Tree for Association Rules", International Journal of Information Technology, vol.12, No. 3, 2006.

[2] Fahim A.M., Salem A.M., "Efficient enhanced k-means clustering algorithm", Journal of Zhejiang University Science, 1626 – 1633, 2006.

[3] Fang Yuag, Zeng Hui Meng, "A New Algorithm to get initial centroid", Third International Conference on Machine Learning and cybernetics, Shanghai, 26-29 August,1191 – 1193, 2004.

[4] Friedrich Leisch1 and Bettina Gr un2, "Extending Standard Cluster Algorithms to Allow for Group Constraints", Compstat 2006, Proceeding in Computational Statistics, Physica verlag, Heidelberg, Germany,2006

[5] J. MacQueen, "Some method for classification and analysis of multi varite observation", University of California, Los Angeles, 281 – 297.

[6] Maria Camila N. Barioni, Humberto L. Razente, Agma J. M. Traina, "An efficient approach to scale up k-medoid based algorithms in large databases", 265 – 279.

[7] Michel Steinbach, Levent Ertoz and Vipin Kumar, "Challenges in high dimensional data set", International Conference of Data management, Vol. 2,No. 3, 2005.

[8] Parsons L., Haque E., and Liu H., "Subspace clustering for high dimensional data: A review", SIGKDD, Explor, Newsletter 6, 90 -105, 2004.

[9] Rui Xu, Donlad Wunsch, "Survey of Clustering Algorithm", IEEE Transactions on Neural Networks, Vol. 16, No. 3, may 2005.

[10] Sanjay garg, Ramesh Chandra Jain, "Variation of k-mean Algorithm: A study for High Dimensional Large data sets", Information Technology Journal5 (6), 1132 – 1135, 2006.

[11] Vance Febre, "Clustering and Continues k-mean algorithm", Los Alamos Science, Georgain Electonics Scientific Journal: Computer Science and Telecommunication, vol. 4,No.3, 1994.

[12] Zhexue Huang, "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining".

[13] Nathan Rountree, "Further Data Mining: Building Decision Trees", first presented 28 July 1999.

[14] Yang liu, "Introduction to Rough Set Theory and Its Application in Decision Suppot System"

[15] Wei-YIn loh, "Regression trees with unbiased variable selection and interaction detection", University of Wisconsin–Madison.

[16] S. Rasoul Safavian and David Landgrebe, "A Survey of Decision Tree Classifier Methodology", School of Electrical Engineering ,Purdue University, West Lafayette, IN 47907.

[17] David S. Vogel, Ognian Asparouhov and Tobias Scheffer, "Scalable Look-Ahead Linear Regression Trees" .

[18] Alin Dobra, "Classification and Regression Tree Construction", Thesis Proposal, Department of Computer Science, Cornell university, Ithaca NY, November 25, 2002

[19] Yinmei Huang, "Classification and regression tree (CART) analysis: methodological review and its application", Ph.D. Student, The Department of Sociology, The University of Akron Olin Hall 247, Akron, OH 44325-1905,

[20] Yan X. and Han J. (2003), GSpan: Graph-Based Substructure Pattern Mining. Proc. 2nd IEEE Int.Conf. on Data Mining (ICDM 2003, Maebashi, Japan), 721–724. IEEE Press,Piscataway, NJ,USA.