Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 7, July 2013, pg.161 – 164

RESEARCH ARTICLE

Selection of Initial Centroids for k-Means Algorithm

Anand M. Baswade[1], Prakash S. Nalwade[2]
[1]M.Tech, Student of CSE Department, SGGSIE&T, Nanded, India
[2]Associate Professor, SGGSIE&T, Nanded, India

[1] anandbaswade08@gmail.com; [2] psnalwade@yahoo.com

Abstract— Clustering is one of the important data mining techniques. k-Means [1] is one of the most important algorithm for Clustering. Traditional k-Means algorithm selects initial centroids randomly and in k-Means algorithm result of clustering highly depends on selection of initial centroids. k-Means algorithm is sensitive to initial centroids so proper selection of initial centroids is necessary. This paper introduces an efficient method to start the k-Means with good initial centroids. Good initial centroids are useful for better clustering.

Key Terms: - Data mining; clustering; k-Means

I. INTRODUCTION

Data mining [3] is a process that uses various techniques to discover “patterns” or “knowledge” from data. Classification, clustering, association rule mining these are some of the data mining techniques. In which clustering is collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to another clusters. Which means cluster analysis is used for finding groups of objects such that the objects in a group will be similar to one another and different from the objects in other groups. Clustering is unsupervised learning technique. Unsupervised learning means learning without prior knowledge about the classification of sample.

II. K-MEANS FOR CLUSTERING

k-Means [1] algorithm is type in partitioning method. In case of partitioning method there is partitioning of n-objects into k-clusters. It is typical clustering approach via partition datasets iteratively. In this there is division of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one cluster. The concept is use K as a parameter, Divide n object into K clusters, to create relatively high similarity in the cluster, relatively low similarity between clusters.

A. K-Means steps

Input: number of clusters K and dataset.

a. Initially take any k objects as centroids.
b. Find distance of all objects from those k centroids, less the distance the object is in that centre of centroids.
c. Now find the centroids from the objects which are in that clusters.
d. Repeat step 2 and step 3 until the value of centroids is same.(that is no change in value of centroids).

Output: K-clusters.
To understand k-Means algorithm, details for k-mean [1, 2, 3] algorithm is given below:

Let $X_i = \{X_1, X_2, X_3, \dots \dots X_n\}$ be the *n* objects.

$X_{i,m} = \{x_{i,1}, \ x_{i,2}, \dots \dots x_{i,m}\}$ be the *m* variables.

$Z_l = \{Z_1, Z_2, \dots \dots Z_k\}$ be the *k* centroids.

$u_{i,l} = \begin{cases} 1 \\ 0 \end{cases}$; $u_{i,l}$ is *n x k* matrix and it is 1 if object *i* allocate in cluster *l.*

B. k-MEANS ALGORITHM

Input: k, Data.
Output: n objects into k clusters.
Step 1. Choose k random initial centroids in the input space.

Step 2. Assign the cluster centres Z_l to those positions.

Step 3. For each $X_i \in$ Data

-compute distance d($x_{i,j}$, $z_{l,j}$) for each Z_l

$$d(x_{i,j}, z_{l,j}) = \sqrt{\sum_{j=1}^{m} |x_{i,j} - z_{l,j}|^2}$$

-Assign X_i to the cluster with the minimum distance.

$$u_{i,l} = 1 \text{ If}$$
$$\sum_{j=1}^{m} d(x_{i,j}, z_{l,j}) \leq \sum_{j=1}^{m} d(x_{i,j}, z_{t,j}) \quad \text{for } 1 \leq t \leq k$$
$$u_{i,t} = 0 \ \text{ for } t \neq l$$

Step 4. For each Z_l Move the position of Z_l to the mean of the points in that cluster

$$Z_{l,j} = \frac{\sum_{i=1}^{n} u_{i,l}\, x_{i,j}}{\sum_{i=1}^{n} u_{i,l}}$$
$$\text{for } 1 \leq l \leq k \text{ and } 1 \leq j \leq m.$$

Step 5. Stopping criteria

-No (or minimum) re-assignments of data points to different clusters. i.e. $u_{i,l}$ Remain unchanged. OR

-No (or minimum) change of centroids. i.e. Z_l Remain unchanged.

C. Performance

I. Strength of k-Means algorithm:

i. Relatively efficient O(knt) where k-number of clusters, n-number of objects, t-number of iteration.
ii. Easy to implement and understand.
iii. Objects automatically assigns to clusters.
iv. Often terminate at local optimum.

II. Weakness of k-Means algorithm:

i. User need to provide input k as number of clusters.(need to specify k)
ii. Different initial k objects may produce different clustering results.
iii. Unable to handle noisy data and outlier.
iv. Not suitable for non-convex shapes.
v. Does not apply directly to categorical data

III. New Proposed Method for Selection of Initial Centroids

In traditional k-Means algorithm starting initial points are selected randomly and result of traditional k-Means is highly depend upon selection on initial centroids so In case of traditional k-Means one of the major weakness is different initial k objects may produce different clustering results. New method is proposed in this paper to select better initial centroids rather than selecting initial centroids randomly.

New proposed method:

Step 1: From n objects calculate a point whose attribute values are average of n-objects attribute values.so first initial centroid is average on n-objects.

Step 2: select next initial centroids from n-objects in such a way that the Euclidean distance of that object is maximum from other selected initial centroids.

Step 3: repeat step 2 until we get k initial centroids.
From these steps we will get initial centroids and with these initial centroids perform k-Means algorithm.

IV. Implementation Details

We have implemented the improved k-Means clustering algorithm in MATLAB with datasets from UCI machine learning repository. This method may not work for some datasets and it may not give required results for larger value of k. The proposed method performs well when value of k is small. Here we used heart disease dataset and the value of k=2 and then at the beginning we use traditional k-Means algorithm and then we used new proposed method for selection of initial centroids instead of selecting initial centroids randomly. By using new approach we obtained good clustering results. The new method of selection of initial centroid is better than selecting the initial centroids randomly.

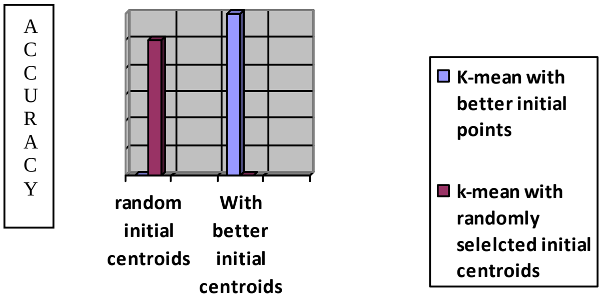

Fig 1: Accuracy for k-Means and k-mean with better initial centroids

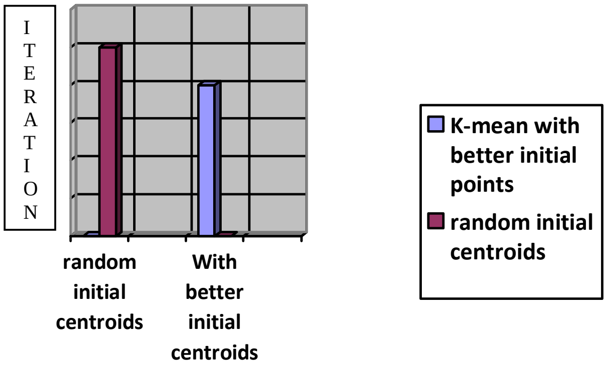

Fig 2: Number of Iteration for k-mean and k- mean with better initial centroids

V. Conclusions

In this paper we have proposed a simple idea for selection of initial centroid that make k-Means more efficient and produce good quality clusters. As the traditional k-Means algorithm cluster results are highly depends on initial centroids and the k-Means algorithm selects initial centroids randomly so there is need to selects initial centroids properly to improve performance. The proposed method selects initial centroids for k-Means algorithm to improve clustering results. Finally we can say that proposed method for selection of initial centroid is better than selecting initial centroids randomly.

References

[1] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observation," Proc. Fifth Berkeley Symp. Math. Statistica and Probability, pp.281-297,1967.

[2] Huang, "Extensions to the k-Means Algorithms for Clustering Large Data Sets with Categorical Values," Data Ming and Knowledge Discovery, vol. 2, no. 3, pp. 283-304, 1998.

[3] Joshua Zhexue Huang , Michael K. Ng , Hongqiang Rong , Zichen Li,"Automated Variable Weighting in k-Means Type Clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, v.27 n.5, p.657-668, May 2005.

[4] Jiawei Han and Micheline Kamber, Data Mining:Concepts and Techniques (Second Edition. Jim Gray, Series Editor, Morgan Kaufmann Publishers, March 2006).