Evaluation of Enhanced K- MEAN Algorithm to the Student Dataset

R.Ranga Raj
Department of Computer Science, Hindusthan College of Arts and Science, Coimbatore
Email: rraj75@rediffmail.com

Dr.M.Punithavalli
Director, Department of Computer Studies Sri Ramakrishna Engineering College, Coimbatore
-

ABSTRACT

Conventional database querying methods are inadequate to extract useful information from huge data banks. Cluster analysis is one of the major data analysis methods and the k-means clustering algorithm is widely used for many practical applications. In this paper, the enhanced k-mean algorithm applied to the huge student dataset to find out the different categories and group them.

Keywords**: Clustering, k-means algorithm, enhanced k-means algorithm**

Date of Submission: July 07, 2012 Date of Acceptance: September 01, 2012

1. Introduction

Cluster analysis [1] is one of the major data analysis methods which is widely used for many practical applications in emerging areas like Bioinformatics [2], [3]. Clustering is the process of partitioning a given set of objects into disjoint clusters. This is done in such a way that objects in the same cluster are similar while objects belonging to different clusters differ considerably, with respect to their attributes. The k-means algorithm [1, 5, 6,] is effective in producing clusters for many practical applications. This algorithm results in different types of clusters depending on the random choice of initial centroids.

2. The K-Mean Clustering Algorithm

The algorithm consists of two separate phases: the first phase is to define k centroids, one for each cluster. The next phase is to take each point belonging to the given data set and associate it to the nearest centroid. Euclidean distance is generally considered to determine the distance between data points and the centroids. When all the points are included in some clusters, the first step is completed and an early grouping is done. At this point we need to recalculate the new centroids, as the inclusion of new points may lead to a change in the cluster centroids.

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.
```

3. Related Work

A variant of the k-means algorithm is the k-modes [4, 7] method which replaces the means of clusters with modes. Like the k-means method, the k-modes algorithm also produces locally optimal solutions which are dependent on the selection of the initial modes.

The original k-means algorithm is computationally very expensive because in each iteration computes the distances between data points and all the centroids.

4. Enhanced Approach

In the enhanced clustering method discussed in this paper [8] both the phases of the original k-means algorithm are modified to improve the accuracy and efficiency. In the enhanced method data points are assigned to the clusters.

Algorithm 2: Assigning data point to cluster

```
Input:
  D={d1,d2,….,dn}//set of n datapoints
  C={c1,c2,…..,cn}//set of k clusters
Output:
  A set of k clusters
Steps:
     1. Compute the distance of each data- point di
  (1<=i<=n) to all the centroids cj (1<=j<=k) as d(di, cj);
     2. For each data-point di, find the closest
  centroid cj and assign di to cluster j.
     3. Set ClusterId[i]=j; // j:Id of the closest cluster
     4. Set Nearest_Dist[i]= d(di, cj);
     5. For each cluster j (1<=j<=k), recalculate the
  centroids;
     6. Repeat
     7. For each data-point di,
       7.1 Compute its distance from the centroid of
       the present nearest cluster;
       7.2 If this distance is less than or equal to the
       present nearest distance, the data-point stays in the
       cluster;
       Else
         7.2.1 For every centroid cj (1<=j<=k)
         Compute the distance d(di, cj);
         End for;
         7.2.2 Assign the data-point di to the cluster
         with the nearest centroid cj
         7.2.3 Set ClusterId[i]=j;
         7.2.4 Set Nearest_Dist[i]= d(di, cj);
     End for;
     8. For each cluster j (1<=j<=k), recalculate the
     centroids;
  Until the convergence criteria is met.
```

In the above algorithm Euclidean distance used for determining the closeness of each data point to the cluster centroids. The distance between one vector X = (x1, x2 ...xn) and another vector Y = (y1, y2 ….yn) is obtained as

$$d(X, Y) = \sqrt{(x1-y1)2 + (x2-y2)2 + ... + (xn-yn)^2}$$

The distance between a data point X and a data-point set D is defined as d(X, D) = min(d (X, Y),whereY ∈ D).

The initial centroids of the clusters are given as input to the second phase, for assigning data-points to appropriate clusters. The steps involved in this phase are outlined as Algorithm 2.

5. Experiment and Data Analysis

To evaluate the enhanced k-means algorithm, a student assessment real time data set has been created based on their academic performance and other activities. The data set includes three years real time of student's records. The below table shows the results obtained out of the data sets. The number of clusters is taken as 6.

No. of Groups	No. of Students
1	73
2	35
3	49
4	68
5	54
6	91

TABLE 1: Student Assessment Table

The Graphical representation of the above student assessment table as follows.

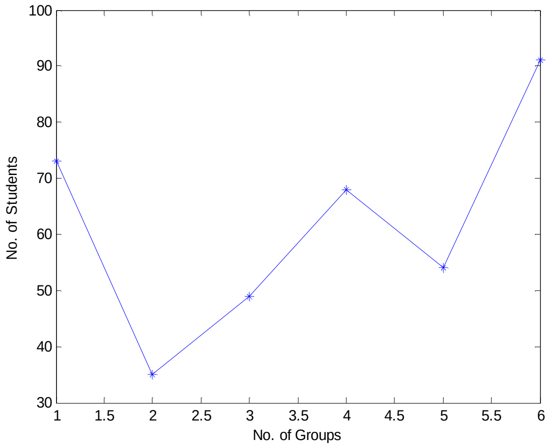

Fig 1: Graph of Student Assessment

In the above graph X axis represent the number of groups (clusters) and Y axis represent the number of students.

6. Conclusion

While implementing the enhanced k-means algorithm, the following issues raised such that due to the random selection of the data set the duplicate value exists. And also the consistency of the student data set is questioned. Further the number of desired clusters is still required to be given as an input, regardless of the distribution of the data points. In future, the work is focused to over come the issues stated.

References

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

[2]. Amir Ben-Dor, Ron Shamir and Zohar Yakini, “Clustering Gene Expression Patterns,” *Journal of Computational Biology*, 6(3/4): 1999

[3]. 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.

[4]. Chaturvedi J. C. A, Green P, “K-modes clustering,” *J. Classification*, (18):35–55, 2001.

[5]. 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.

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

[7]. Margaret H. Dunham, *Data Mining-Introductory and AdvancedConcepts*, Pearson Education, 2006.

[8]. K.A.Abdul Nazeer, M.P.Sabestian *‘Improving the Accuracy and Efficiency of the K-means Clustering Algorithm’,WCE 2009 London, UK.*