**K-means Clustering**

**Introduction:**

Today the important challenge that higher education faces, is reaching a stage to facilitate the universities in having more efficient, effective and accurate educational processes. Data mining is considered as the most suited technology appropriate in giving additional insight into the lecturer, student,teacher, manager, and other educational staff behavior and acting as an active automated assistant in

helping them for making better decisions on their educational activities. As discussed before, lack of deep and enough knowledge in higher educational system may prevent system management to achieve their quality objectives. Data mining technology can help bridging this knowledge gaps in higher educational system. Therefore the hidden patterns, association and anomalies, which are

discovered by some data mining techniques, can be used to improve the effectiveness, efficiency and the speed of the processes. As a result, this improvement may bring a lot of advantages to the higher educational system such as maximizing educational system efficiency, decreasing student's drop-out rate, increasing student's promotion rate, increasing student's retention rate, increasing student's transition rate, increasing educational improvement ratio, increasing student's success, increasing student's learning outcome, and reducing the cost of system processes. In order to achieve the above quality improvement, we need a data mining system that can provide the needed knowledge and insights for the decision makers in the higher educational system. We have analyzed such a system and in the next section we will present and describe an analysis model for this system.

**K-means:**

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm a

1. The centroids of the *K* clusters, which can be used to label new data
2. Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically.

Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.

The idea is to choose random cluster centers, one for each cluster. These centers are preferred to be as far as possible from each other. In this algorithm Euclidean distance measure is used between two multidimensional data points

X = (x1,x2,x3,…………xm )

Y = (y1,y2,y3,…………ym )

The Euclidean distance measure between the above points x and y are

described as follows:

D(X, Y) = ( ! ( xi - yi) 2)1/2

**Algorithmic steps:**

**Input**: D = {d1, d2, d3, ..., dn } // Set of n data points. K - Number of desired clusters

**Output**: A set of K clusters.

**Method:**

1. Select the number of clusters. Let this number be k.

2. Pick k seeds as centroids of the k clusters. The seeds may be picked

randomly unless the user has some insight into the data.

3. Compute the Euclidean distance of each object in the dataset from each of

the centroids.

4. Allocate each object to the cluster nearest, based on the distances

computed in the previous step.

5. Compute the centroids of the clusters by computing the means of the

attribute values if the objects are in each cluster.

6. Check if the stopping criterion has been met (e.g. the cluster membership

is unchanged). If not go to step 3.