CSCI 4144 Project Proposal

Team Member:

Yucheng Liu B00590305

Yiying Zhang B00560058

Project Title:

Application of k-Means Clustering algorithm for prediction of Students' Academic Performance

DM method:

This paper presents k-means clustering algorithm as a simple and efficient tool to monitor the progression of students' performance in a higher institution. In addition, Euclidean distance measure of similarity is chosen to be used in the analysis of the students' scores.

First, Given a dataset with n data points $x_1, x_2, \dots, x_n$ such that each data point is in R^d. To solve the problem of finding the minimum variance clustering of the dataset into k clusters, we just find k points $\{m_j\}$ $(j = 1, 2, \dots, k)$ in R^d as cluster centroids such that

$$\frac{1}{n}\sum_{i=1}^{n}\left[min_j d^2(x_i, m_j)\right] \qquad (1)$$

is minimized, where $d(x_i, m_j)$ denotes the Euclidean distance between x_i and m_j.

The k-means algorithm provides an easy method to implement an approximate solution to Equation (1). It updates cluster centroids till a local minimum is found.

Step 1:	Accept the number of clusters to group data into and the dataset to cluster as input values
Step 2:	Initialize the first K clusters - Take first k instances or - Take Random sampling of k elements
Step 3:	Calculate the arithmetic means of each cluster formed in the dataset.
Step 4:	K-means assigns each record in the dataset to only one of the initial clusters - Each record is assigned to the nearest cluster using a measure of distance (e.g Euclidean distance).
Step 5:	K-means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic mean of all the clusters in the dataset.

The overall performance is evaluated by applying deterministic model

$$\frac{1}{N}\left(\sum_{j=1}^{N}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)\right) \qquad (2)$$

where N is the total number of students in a cluster and n is the dimension of the data. The group assessment in each of the cluster size is evaluated by summing the average of the individual scores in each cluster.

Application Scenario:

In order to increase the quality of the education, one effective way is to monitor the progress of every student's academic performance in order to find their characteristics then it is easier for teachers to make more specialized plan to different students. The most fundamental measurement of students is the GPAs, which can obviously tell teachers if students understand the material in taught in class or not and how well they understood it. However, it is too complex and cost too much time on teachers or other academic planners analyzing each student manually, so that they need a more powerful tool to help them cluster the students in a short time period. Applying the K-means method can effectively cluster all the students in to several groups according to student's GPAs. From the results, academic planners can change their plans immediately. Academic planners can make more suitable plans for the specific groups of student to enhance their academic performance.

Implementation Considerations:

Language: Python

Computer System: Linux

Other tools needed:

Spyder-Py2: used for programming and testing.