**MACHINE LEARNING AND DATA MINING**

**PROJECT REPORT**

1. **Introduction**

Clustering is an unsupervised machine learning technique with a lot of applications in the areas of pattern recognition, image analysis, customer analytics, market segmentation, social network analysis, and more. A broad range of industries use clustering, from airlines to healthcare and beyond.

In this project, we will show you two efficient clustering algorithms and the differences between them.

1. **Problem**

The system needs to group the students based on a number of predefined attributes (e.g., age, gender, score, the number of registered courses for the semester, etc..).

Input: a file containing a set of vectors that present for students. Each element of a vector is equivalent to an attribute of a student.

Output: group of students after being clustered.

Approach: using K-means clustering or FCM clustering

1. **Solution**
   1. **K-means**
      1. **Introduction**

K-means is a crucial algorithm that is popularly used in machine learning, particularly in clustering problem. K-means belongs to the partition-based clustering and its main target is maximizing the distance between clusters and minimizing the distance between samples that are in the same cluster.

* + 1. **Algorithm**

Step 1: Choose randomly K centroids for K clusters. Each centroid represents for a cluster.

Step 2: Calculate the distance between all samples of the dataset and K centroids (usually Euclidean distance)

Step 3: Assign each sample to the cluster that is nearest to it.

Step 4: Re-calculate centroids as formular:

Where:

Centroid[k]: center of cluster k-th

x[i]: sample i-th in the cluster

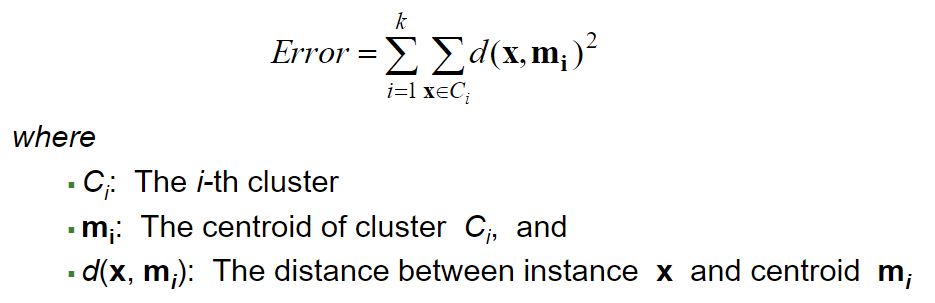
length[k]: the number of samples of cluster k-th

cluster[k]: cluster k-th

Step 5: if the algorithm converges, then stop; otherwise, go to step 2.

Convergence criterion:

* No (or insignificant) re-assignment of samples to different clusters, or
* No (or insignificant) change of centroids, or
* Insignificant decrease in the sum of error



The output of algorithm is several attributes. But there are 2 features we need t pay attention. The 2D array refers to the last centroids, the 1D array refers to the labels of each sample in the dataset.

Label[i] is an integer in range [0, k]

Where: k is the number of cluster

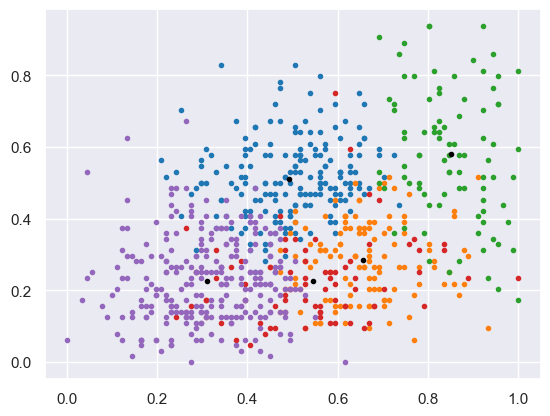
* + 1. **Program result**

Data before clustering

Data after clustering without any techniques

Data after clustering with normalization (using minmaxscalar)

Data after clustering with normalization and removing outliers



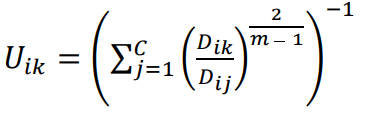
* + 1. **Summary**
       1. **Advantages**
* Easy to understand
* Easy to install and execute
  + - 1. **Disadvantage**
* Need to know the number of cluster and do not know how many clusters are optimal
* The last result depends on the random of the initial centroids
* Clusters need to have roughly the same number of points
* Clusters should be circular
* The algorithm can not be converge if there is a cluster lying inside another cluster
* Sensitive to outliers
  1. **Fuzzy-c-means (FCM)**
     1. **Introduction**

The idea of FCM is quite similar with K-means. It also belongs to the partition-based clustering and based on the separation among samples.

* + 1. **Algorithm**

Step 1: Choose randomly K centroids for K clusters. Each centroid represents for a cluster.

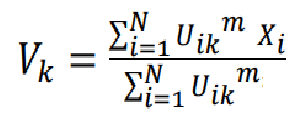
Step 2: Calculate the membership matrix as the formular:



Where:

uik: the characteristic value identifying element xi belonging to cluster k

Step 3: Re-calculate centroids as the formular:



Step 4: Repeat step 2, 3 until ‖ 𝑉 (𝑙) − 𝑉 (𝑙+1) ‖ < 𝜀

When the program stops, we receive:

* A 2D array refers to the centroids
* A 2D array refers to the characteristic value identifying each sample belonging to each cluster
  + 1. **Program result**
    2. **Summary**
       1. **Advantages**

Improve the quality of clustering comparing to K-means due to the complex of formulars

* + - 1. **Disadvantages**
* Need to know the number of cluster and do not know how many clusters are optimal
* The last result depends on the random of the initial centroids
* Sensitive to outliers
* The formulars are complex

1. **Comparation**

With 4 clusters:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Davies-Bouldin index | Root mean square standard deviation | Dunn index | R squared | Silhouette width criterion |
| K-means |  |  |  |  |  |
| FCM |  |  |  |  |  |

With 5 clusters:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Davies-Bouldin index | Root mean square standard deviation | Dunn index | R squared | Silhouette width criterion |
| K-means |  |  |  |  |  |
| FCM |  |  |  |  |  |

With 6 clusters:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Davies-Bouldin index | Root mean square standard deviation | Dunn index | R squared | Silhouette width criterion |
| K-means |  |  |  |  |  |
| FCM |  |  |  |  |  |