# 

**Assignment No. 2**

## 1.1 Title:

Consider a suitable dataset. For clustering of data instances in different groups, apply different clustering techniques (minimum 2). Visualize the clusters using suitable tool.

## 1.2 Problem Definition:

Visualize the Cluster using Rapidminer tool

## 1.3 Prerequisite:

* + - Basic concepts of ETL.
    - Knowledge about Rapidminer tool

## 1.4 Software Requirements:

* + - Rapidminer tool

## 1.5 Hardware Requirement:

* + - PIV, 2GB RAM, 500 GB HDD, Lenovo A13-4089Model.

## 1.6 Learning Objectives:

Use Rapidminer functions to create K-means Clustering models and hierarchical clustering models

## 1.7 Outcomes:

Visualize the effectiveness of the K-means Clustering algorithm and hierarchical clustering using graphic capabilities in Rapidminer

## 1.8 Theory Concepts:

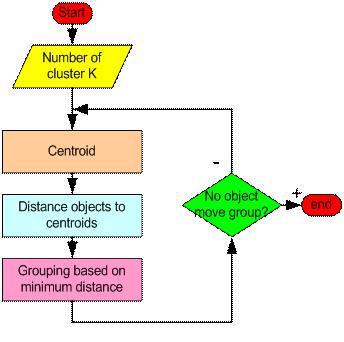
**What is K-means clustering?**

*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 are:

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. The "Choosing K" section below describes how the number of groups can be determined. 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.

**Steps to Perform K-Means Clustering**



## 

As a simple illustration of a k-means algorithm, consider the following data set consisting of the scores of two variables on each of seven individuals:

|  |  |  |
| --- | --- | --- |
| Subject | A | B |
|  |  |  |
| 1 | 1.0 | 1.0 |
| 2 | 1.5 | 2.0 |
| 3 | 3.0 | 4.0 |
| 4 | 5.0 | 7.0 |
|  |  |  |
| 5 | 3.5 | 5.0 |
| 6 | 4.5 | 5.0 |
| 7 | 3.5 | 4.5 |

This data set is to be grouped into two clusters. As a first step in finding a sensible initial partition, let the A & B values of the two individuals furthest apart (using the Euclidean distance measure), define the initial cluster means, giving:

|  |  |  |
| --- | --- | --- |
|  | Individual | Mean Vector |
|  |  | (centroid) |
| Group 1 | 1 | (1.0, 1.0) |
| Group 2 | 4 | (5.0, 7.0) |
|  |  |  |

The remaining individuals are now examined in sequence and allocated to the cluster to which they are closest, in terms of Euclidean distance to the cluster mean. The mean vector is recalculated each time a new member is added. This leads to the following series of steps:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Cluster 1 | | Cluster 2 | |
|  |  | Mean |  | Mean |
| Step | Individual | Vector | Individual | Vector |
|  |  | (centroid) |  | (centroid) |
| 1 | 1 | (1.0, 1.0) | 4 | (5.0, 7.0) |
| 2 | 1, 2 | (1.2, 1.5) | 4 | (5.0, 7.0) |
|  |  |  |  |  |
| 3 | 1, 2, 3 | (1.8, 2.3) | 4 | (5.0, 7.0) |
| 4 | 1, 2, 3 | (1.8, 2.3) | 4, 5 | (4.2, 6.0) |
| 5 | 1, 2, 3 | (1.8, 2.3) | 4, 5, 6 | (4.3, 5.7) |
| 6 | 1, 2, 3 | (1.8, 2.3) | 4, 5, 6, 7 | (4.1, 5.4) |

Now the initial partition has changed, and the two clusters at this stage having the following characteristics:

|  |  |  |
| --- | --- | --- |
|  | Individual | Mean Vector |
|  |  | (centroid) |
| Cluster 1 | 1, 2, 3 | (1.8, 2.3) |
| Cluster 2 | 4, 5, 6, 7 | (4.1, 5.4) |

But we cannot yet be sure that each individual has been assigned to the right cluster. So, we compare each individual’s distance to its own cluster mean and to that of the opposite cluster. And we find:

|  |  |  |
| --- | --- | --- |
|  | Distance to | Distance to |
| Individual | mean | mean |
| (centroid) of | (centroid) of |
|  |
|  | Cluster 1 | Cluster 2 |
| 1 | 1.5 | 5.4 |
|  |  |  |
| 2 | 0.4 | 4.3 |
|  |  |  |
| 3 | 2.1 | 1.8 |
| 4 | 5.7 | 1.8 |
| 5 | 3.2 | 0.7 |
| 6 | 3.8 | 0.6 |
| 7 | 2.8 | 1.1 |
|  |  |  |

Only individual 3 is nearer to the mean of the opposite cluster (Cluster 2) than its own (Cluster 1). In other words, each individual's distance to its own cluster mean should be smaller that the distance to the other cluster's mean (which is not the case with individual 3). Thus, individual 3 is relocated to Cluster 2 resulting in the new partition:

|  |  |  |
| --- | --- | --- |
|  | Individual | Mean Vector |
|  |  | (centroid) |
| Cluster 1 | 1, 2 | (1.3, 1.5) |
| Cluster 2 | 3, 4, 5, 6, 7 | (3.9, 5.1) |
|  |  |  |

The iterative relocation would now continue from this new partition until no more relocations occur. However, in this example each individual is now nearer its own cluster mean than that of the other cluster and the iteration stops, choosing the latest partitioning as the final cluster solution.

## Hierarchical Clustering

## What is Hierarchical clustering?

## Given a set of N items to be clustered, and an NxN distance (or similarity) matrix, the basic process of

## Johnson's (1967) hierarchical clustering is this:

## Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters equal the distances (similarities) between the items they contain.

## Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.

## Compute distances (similarities) between the new cluster and each of the old clusters.

## Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

## 

## 

1.9 Clustering using Rapidminer tool

This Operator performs clustering using the k-means algorithm.

## Description

This Operator performs clustering using the k-means algorithm. Clustering groups Examples together which are similar to each other. As no Label Attribute is necessary, Clustering can be used on unlabelled data and is an algorithm of unsupervised machine learning.

The k-means algorithm determines a set of k clusters and assignes each Examples to exact one cluster. The clusters consist of similar Examples. The similarity between Examples is based on a distance measure between them.

A cluster in the k-means algorithm is determined by the position of the center in the n-dimensional space of the n Attributes of the ExampleSet. This position is called centroid. It can, but do not have to be the position of an Example of the ExampleSets.

The k-means algorithm starts with k points which are treated as the centroid of k potential clusters. These start points are either the position of k randomly drawn Examples of the input ExampleSet, or are determined by the k-means++ heuristic if determine good start values is set to true.

All Examples are assigned to their nearest cluster (nearest is defined by the measure type). Next the centroids of the clusters are recalculated by averaging over all Examples of one cluster. The previous steps are repeated for the new centroids until the the centroids no longer move or max optimization steps is reached. Be aware that it is not ensured that the k-means algorithm converges if the measure type is not based on Euclidean Distance calculation (cause the recalculation of the centroids by averaging is assuming Euclidean space).

The procedure is repeated max runs times with each time a different set of start points. The set of clusters is delivered which has the minimal sum of squared distances of all Examples to their corresponding centroids.

## Differentiation

### k-Medoids



In case of the k-medoids algorithm the centroid of a cluster will always be one of the points in the cluster. This is the major difference between the k-means and k-medoids algorithm.

### k-Means (Kernel)



Kernel k-means uses kernels to estimate distances between Examples and clusters. Because of the nature of kernels it is necessary to sum over all Examples of a cluster to calculate one distance. So this algorithm is quadratic in number of Examples and does not return a Centroid Cluster Model (on the contrary the K-Means operator returns a Centroid Cluster Model).

## Input

* example set input*(Data Table)*



This input port expects an ExampleSet.

## Output

* cluster model*(Centroid Cluster Model)*



This port delivers the cluster model. It contains the information which Examples are part of which cluster. It also stores the position of the centroids of the clusters. It can be used by the Apply Model Operator to perform the specified clustering on another ExampleSet. The cluster model can also be grouped together with other clustering models, preprocessing models and learning models by the Group Models Operator.

* clustered set*(Data Table)*

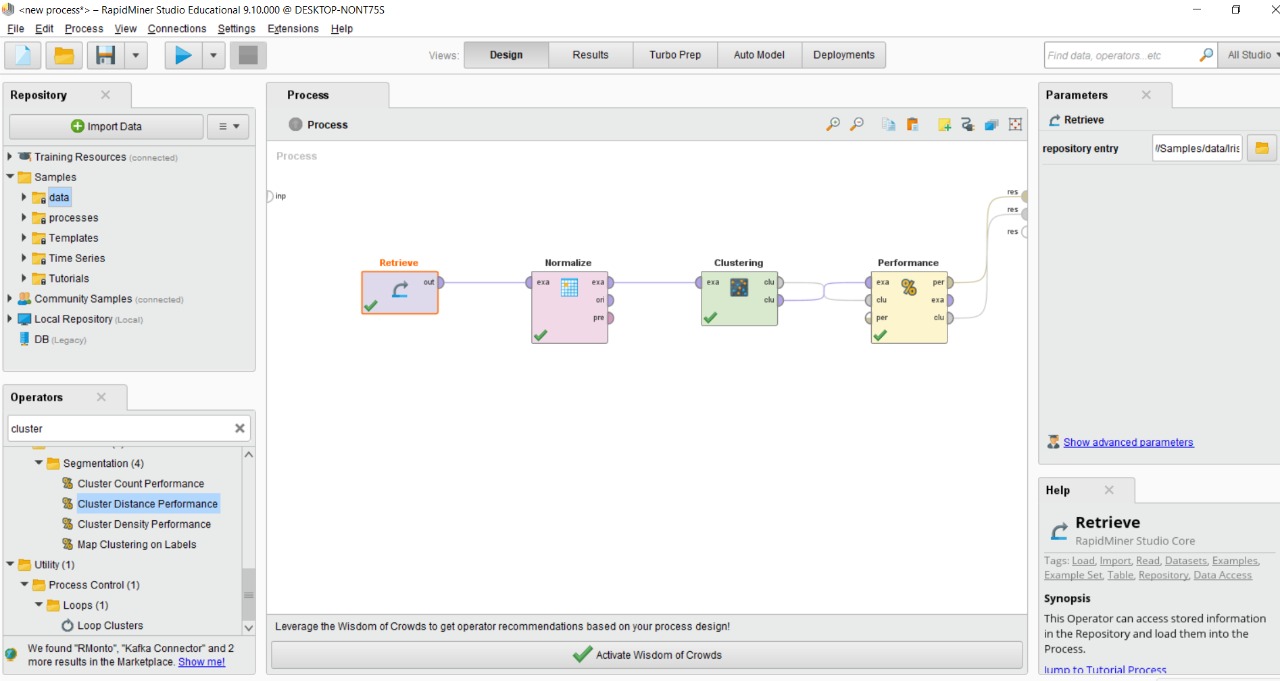


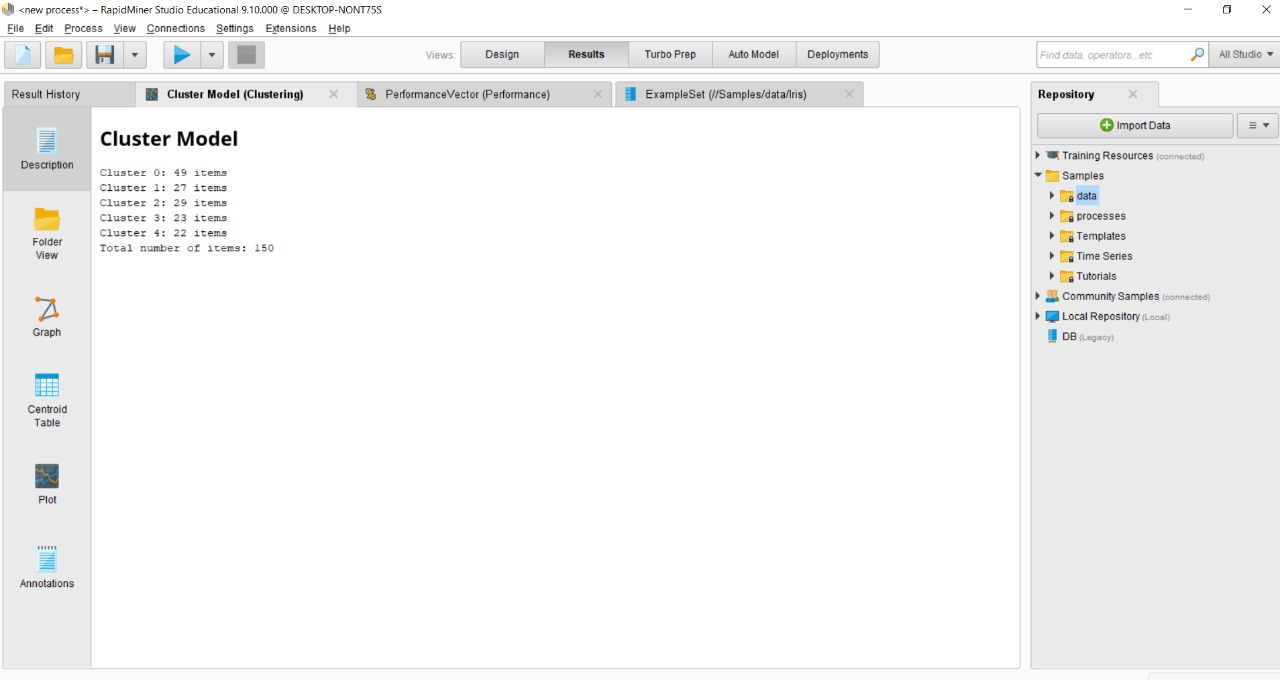
An Attribute 'id' with special role 'Id' is added to the input ExampleSet to distinguish Examples. Depending on the add cluster attribute and the add as label parameters an Attribute 'cluster' with special role 'Cluster' or 'Label' is also added. The resulting ExampleSet is delivered at this output port.

### Clustering of the Iris Data Set

In this tutorial process the Iris data set is clustered using the k-means Operator. The Iris data set is retrieved from the Samples folder. Also the label Attribute is also retrieved, but only for comparison of the cluster assignment with the class of the Examples. The label Attribute is not used in the Clustering.

## Snapshot of your assignment





**1.10 Assignment Question**

* + 1. What is difference between Supervised and Unsupervised Learning?
    2. What are different similarities between Kmean and KNN Algorithm?
    3. What is Euclidean distance? Explain with Suitable example?
    4. What is hamming distance? Explain with Suitable example?
    5. What is Chi Squre Distance? Explain with Suitable example?
    6. What are different types of Clustering?
    7. What is Weka Tool? Explain the Step to Perform Clustering on Sample data set?

**References**

1. https://docs.rapidminer.com/latest/studio/operators/modeling/segmentation/k\_means.html
2. http://www.rdatamining.com/examples/kmeans-clustering