Program – 7

AIM: To perform k-mean clustering on iris.arff in WEKA

Introduction and Theory

k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

The problem is computationally difficult (NP-hard); however, efficient heuristic algorithms converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both k-means and Gaussian mixture modeling. They both use cluster centers to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

The algorithm has a loose relationship to the k-nearest neighbor classifier, a popular machine learning technique for classification that is often confused with k-means due to the name. Applying the 1-nearest neighbor classifier to the cluster centers obtained by k-means classifies new data into the existing clusters. This is known as nearest centroid classifier or Rocchio algorithm.

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

It halts creating and optimizing clusters when either:

- The centroids have stabilized—there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved.

```
1 Initialize k means with random values
2
3 For a given number of iterations:
4     Iterate through items:
5         Find the mean closest to the item
6         Assign item to mean
7         Update mean
```

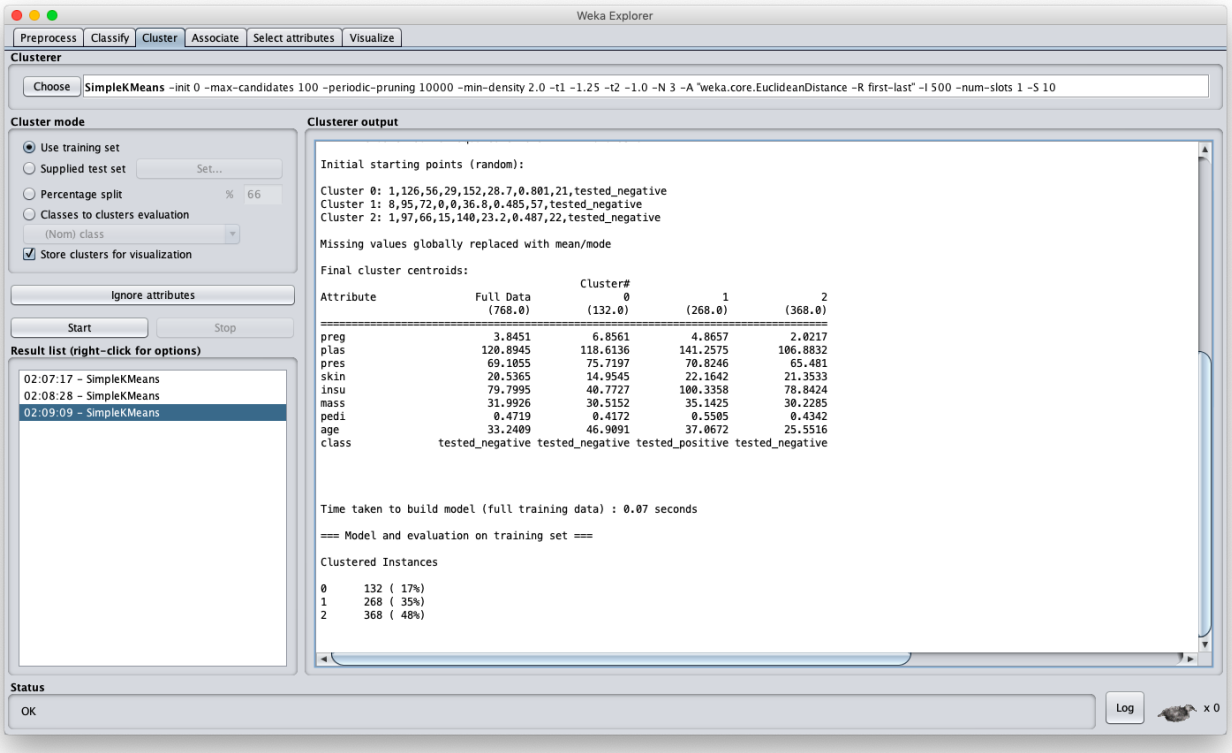
The k-mean pseudo-code

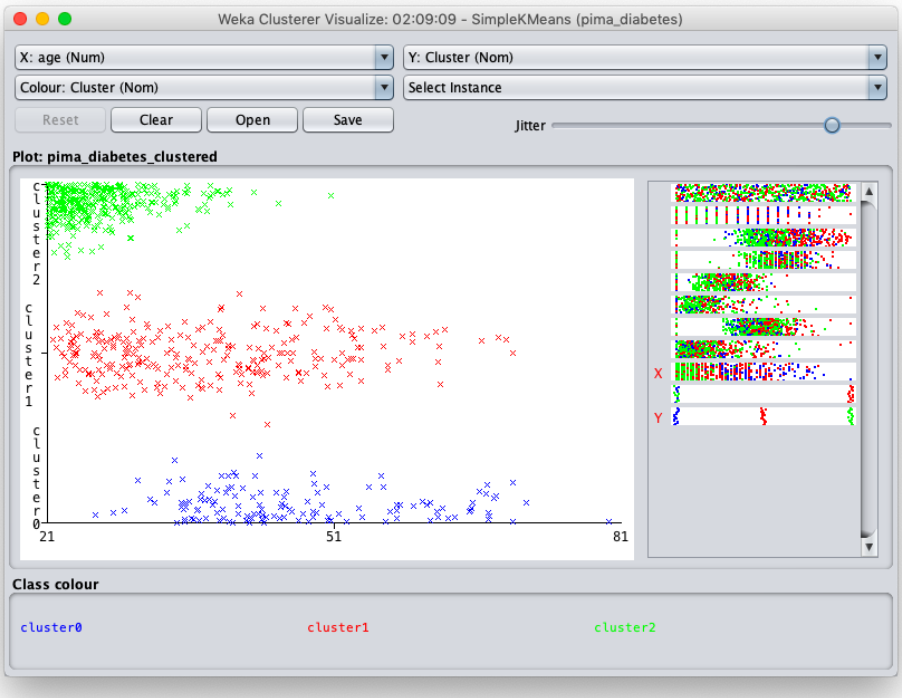
Procedure

1. Go to Weka Explorer.
2. Choose dataset in weka/data
3. Go to cluster tab
4. Choose cluster algorithm, in this case k-means
5. Click start.
6. Visualize the results

Program – 7

Results





Program – 7

Findings and Learnings:

1. K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem.
2. Weka software makes the implementation and visualization of K-means algorithms very simple and efficient.
3. It makes K-means easily available to non-programmer users as well.