Assignment

Data Mining

Topic

K-Mean Algorithm Using Weka

Submitted by

Junaid Iqbal (70070342)

Master of Science in Computer Science
The University of Lahore
Lahore, Pakistan

THE UNIVERSITY OF LAHORE

January 2021

K-means clustering

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. A cluster refers to a collection of data points aggregated together because of certain similarities.

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.

Experiment:

Dataset: The Data set which I am using is Soybean Large Dataset which contain 35 attributes which are normalized.

The Dataset and the results are available on GitHub: (https://github.com/junaideqbal/data-mining/tree/main/assignment2)

Few Snapshots of results are given Below with 2, 3 and 4 number of clusters.

Number of Clusters 2:

```
=== Model and evaluation on training set ===

Clustered Instances

0      272 ( 40%)
1      411 ( 60%)
```

Number of Clusters 3:

```
=== Model and evaluation on training set ===

Clustered Instances

0      241 ( 35%)
1      402 ( 59%)
2       40 (  6%)
```

Number of Clusters 4:

```
=== Model and evaluation on training set ===

Clustered Instances

0      243 ( 36%)
1      284 ( 42%)
2       38 (  6%)
3      118 ( 17%)
```