

Experiment:-5

- Description:- K-means algorithm aims to partition n observation into " k clusters" in which each observation belongs to cluster with nearest mean, serving as a prototype of cluster. The result is partitioning of data into various cells.

- Dataset:- Iris dataset.

- Steps:-

- ① Create a csv file
- ② Now open weka explorer & then select all attribute in table
- ③ Select cluster tab in tool & choose normal K-means technique to see result.

- Result:- In this experiment, we have successfully implement K-means algo.

* Viva-Question:-

Q1 Enumerate the strategy of K-means algorithm implementation on any unlabelled dataset.

-
- ① Initialize: Randomly select k -centroid
 - ② Assign: Assign each data point to nearest centroid.
 - ③ Update: Recalculate centroid based on assigned points
 - ④ Iterate: Repeat step 2 & 3 until centroid stabilize are reached

EXPERIMENT 5

Weka Explorer

PreprocessClassifyClusterAssociateSelect attributesVisualize

Clusterer

ChooseSimpleKMeans-init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Cluster mode

Use training set

Supplied test set

Percentage split

Classes to clusters evaluation

Store clusters for visualization

Set...

% 66

(Nom) class

Ignore attributes

Start

Stop

Result list (right-click for options)

08:45:41 - SimpleKMeans

Clusterer output

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: iris

Instances: 150

Attributes: 5

sepalength

sepalwidth

petallength

petalwidth

Ignored: class

Test mode: Classes to clusters evaluation on training data

=== Clustering model (full training set) ===

KMeans

=====

Number of iterations: 6

Within cluster sum of squared errors: 6.998114004826762

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4

Cluster 1: 6.2,2.9,4.3,1.3

Cluster 2: 6.9,3.1,5.1,2.3

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute Full Data Cluster#

(150.0) (61.0) (50.0) (39.0)

=====

sepalength 5.8433 5.8885 5.006 6.8462

sepalwidth 3.054 2.7377 3.418 3.0821

petallength 3.7587 4.3967 1.464 5.7026

petalwidth 1.1987 1.418 0.244 2.0795

=====

Time taken to build model (full training data) : 0.01 seconds

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

Clustered Instances

0 61 (41%)

1 50 (33%)

2 39 (26%)

Class attribute: class

Classes to Clusters:

0 1 2 <-- assigned to cluster

0 50 0 | Iris-setosa

47 0 3 | Iris-versicolor

14 0 36 | Iris-virginica

Cluster 0 <-- Iris-versicolor

Cluster 1 <-- Iris-setosa

Cluster 2 <-- Iris-virginica

Status

OK

Incorrectly clustered instances : 17.0 11.3333 %

- ① Customer segmentation
② Image Compression
③ Anomaly detection
④ document clustering
⑤ Biological Data analysis.

Q2
→ Value of k in k -means can optimize using Elbow method or gap Statistic.

JIET
UNIVERSE