EECS 4412 Data Mining
Assignment 3 Question 1 Report

Name: Chun Ho Li
Student ID: 214908800

Date: 12 April 2021

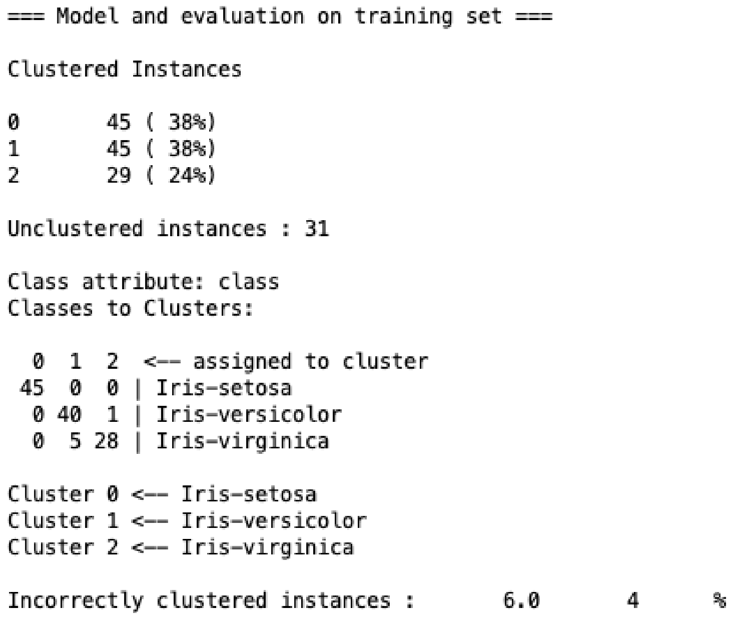
1. Clustering

1. Parameterize DBSCAN for Iris Plant dataset

Parameters:

Distance Function – Manhattan Distance
Epsilon – 0.19
minPoints – 4

Result:



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

Clustered Instances

0       45 ( 38%)
1       45 ( 38%)
2       29 ( 24%)

Unclustered instances : 31

Class attribute: class
Classes to Clusters:

  0  1  2  <-- assigned to cluster
 45  0  0 | Iris-setosa
  0 40  1 | Iris-versicolor
  0  5 28 | Iris-virginica

Cluster 0 <-- Iris-setosa
Cluster 1 <-- Iris-versicolor
Cluster 2 <-- Iris-virginica

Incorrectly clustered instances :       6.0      4      %
```

Discussion:

Out of 150 instances, there are 6 instances incorrectly clustered which is 4% of error with 31 instances classified as noisy data. In each of the three clusters, all instances are classified correctly with the setosa class, while the other two classes versicolor and virginica has some overlapped clustering. DBSCAN is very sensitive to parameters, even a slight tweak in epsilon result in huge change in error rate, a smaller number of epsilon and minPoints will result in more clusters generated and more sensitive in clustering. As a result of fine tuning and clusters generated, I can conclude that the attributes data of versicolor and virginica are very similar, containing some overlapping and in convex shape, while setosa are very different.

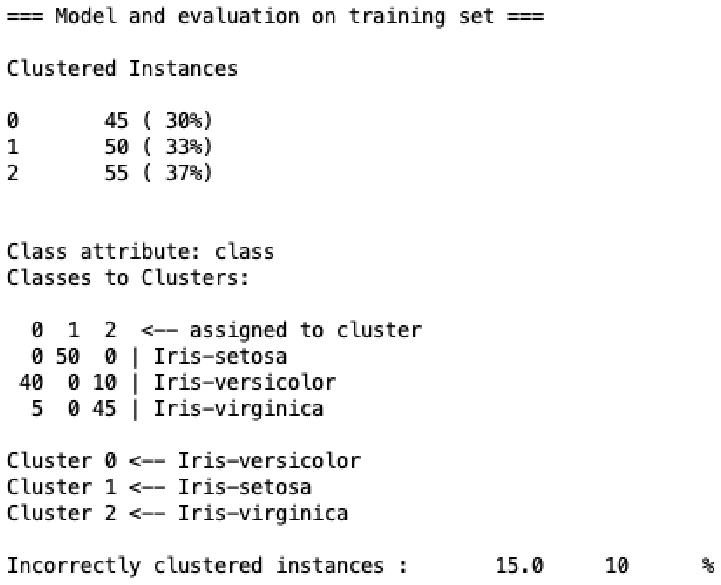
2. Other clustering methods

Simple K mean clustering

parameters:

k – 3
distance function – Manhattan Distance
initialization method – Canopy
seed – 7

Result:



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

Clustered Instances

0       45 ( 30%)
1       50 ( 33%)
2       55 ( 37%)


Class attribute: class
Classes to Clusters:

  0  1  2  <-- assigned to cluster
  0 50  0 | Iris-setosa
 40  0 10 | Iris-versicolor
  5  0 45 | Iris-virginica

Cluster 0 <-- Iris-versicolor
Cluster 1 <-- Iris-setosa
Cluster 2 <-- Iris-virginica

Incorrectly clustered instances :       15.0     10      %
```

Discussion:

While the number of incorrectly cluster instances are more than DBSCAN, all 150 instances are being clustered without being classified as noisy data. However, we have to know the number of clusters beforehand with K-mean, and it is very sensitive to outliers such as overlapping instances with versicolor and virginica. Again, there are some overlapping in between versicolor and virginica while all setosa instances are clustered correctly. K-means also work well when the clusters are in convex shape.