EECS 4412: Data Mining
Assignment #3: Clustering

Name: Supriyo Ghosh
Student ID: 215318728

INTRODUCTION

In this report, we will explore different clustering methods available in Weka over the Iris Data Set. The data set includes three classes: *Iris-Setosa*, *Iris-Versicolour,* and *Iris-Virginica*. Each class has 50 samples and that makes a total of 150 instances. There are four non-class attributes: *sepallength*, *sepalwidth*, *petallength,* and *petalwidth.* We will not do any preprocessing on this data and straight-away proceed to model building. Accuracy of each method is calculated by subtracting the incorrectly clustered instances percentage from 100.

DBSCAN

The DBSCAN clustering method is applied on the iris data set by selecting the cluster mode "classes to clusters evaluation" which evaluates clusters with respect to a class. By default, the method identifies only one cluster from the data set.

DBSCAN clustering results:
Clustered DataObjects: 150
Number of attributes: 4
Epsilon: 0.9; minPoints: 6
Number of generated clusters: 1

```
(  0.) 5.1,3.5,1.4,0.2                  -->  0
(  1.) 4.9,3,1.4,0.2                    -->  0
…
(148.) 6.2,3.4,5.4,2.3                  -->  0
(149.) 5.9,3,5.1,1.8                    -->  0

=== Model and evaluation on training set ===
Clustered Instances
0      150 (100%)

  0  <-- assigned to cluster
 50 | Iris-setosa
 50 | Iris-versicolor
 50 | Iris-virginica

Cluster 0 <-- Iris-setosa
Incorrectly clustered instances :        100.0    66.6667 %
```

As the data set contains 50 samples of each iris plant species, parameters of **Epsilon: 0.45** and **minPoints: 49** are used to get the best clustering with the DBSCAN method. Now, the method can identify two clusters with all the samples of iris-setosa in one cluster and the other two species in the other cluster. The accuracy increases from 33.33% to 66.66%

DBSCAN clustering results:
Clustered DataObjects: 150
Number of attributes: 4
Epsilon: 0.494; minPoints: 50
Distance-type:
Number of generated clusters: 2

```
(  0.) 5.1,3.5,1.4,0.2                  -->  1
(  1.) 4.9,3,1.4,0.2                    -->  1
(  2.) 4.7,3.2,1.3,0.2                  -->  1
…
(147.) 6.5,3,5.2,2                      -->  0
(148.) 6.2,3.4,5.4,2.3                  -->  0
(149.) 5.9,3,5.1,1.8                    -->  0

=== Model and evaluation on training set ===
Clustered Instances
0       50 ( 33%)
1      100 ( 67%)

  0  1  <-- assigned to cluster
 50  0 | Iris-setosa
  0 50 | Iris-versicolor
  0 50 | Iris-virginica

Cluster 0 <-- Iris-setosa
Cluster 1 <-- Iris-versicolor
Incorrectly clustered instances :         50.0      33.3333 %
```

SimpleKMeans

The SimpleKMeans clustering method is applied on the iris data set by selecting the cluster mode "classes to clusters evaluation" which evaluates clusters with respect to a class. By default, the method identifies two clusters from the data set.

KMeans clustering results:
Number of iterations: 7
Within cluster sum of squared errors: 12.143688281579722

=== Model and evaluation on training set ===
Clustered Instances
0 100 (67%)
1 50 (33%)

0 1 <-- assigned to cluster
0 50 | Iris-setosa
50 0 | Iris-versicolor
50 0 | Iris-virginica

Cluster 0 <-- Iris-versicolor
Cluster 1 <-- Iris-setosa
Incorrectly clustered instances : 50.0 33.33 %

The **k - numClusters** is set to **3** to find three clusters. The accuracy changes from 66% to 88.6%.

KMeans clustering results:
Number of iterations: 6
Within cluster sum of squared errors: 6.998114004826762

=== Model and evaluation on training set ===
Clustered Instances
0 61 (41%)
1 50 (33%)
2 39 (26%)

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
Incorrectly clustered instances : 17.0 11.33 %

HierarchicalClusterer

The HierarchicalClusterer clustering method is applied on the iris data set by selecting the cluster mode “classes to clusters evaluation” which evaluates clusters with respect to a class. By default, the method identifies two clusters from the data set.

HierarchicalClusterer clustering results:
Time taken to build the models (full training data) is **0.01 to 0.02 seconds**

The accuracy table with different types of linkage techniques:

linkType	numClusters = 2	numClusters = 3
SINGLE	66.67%	66%
COMPLETE	55,33%	88%
AVERAGE	66.66%	88.67%

Results for numClusters = 3 and linkType = AVERAGE:
=== Model and evaluation on training set ===
Clustered Instances
0 50 (33%)
1 67 (45%)
2 33 (22%)

0 1 2 <-- assigned to cluster
50 0 0 | Iris-setosa
0 50 0 | Iris-versicolor
0 17 33 | Iris-virginica

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

Incorrectly clustered instances : 17.0 11.3333 %

EM

The Expectation Maximization clustering method is applied on the iris data set by selecting the cluster mode “classes to clusters evaluation” which evaluates clusters with respect to a class. By default, the method identifies five clusters but if the numClusters is set to 3, the incorrectly clustered instances are 9.33% which gives an accuracy of 90.66%.

EM clustering results:
Number of clusters: 3
Number of iterations performed: 10
Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===
Clustered Instances
0 64 (43%)
1 50 (33%)
2 36 (24%)

Log likelihood: -2.055

```
  0  1  2  <-- assigned to cluster
  0 50  0 | Iris-setosa
 50  0  0 | Iris-versicolor
 14  0 36 | Iris-virginica
```

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

Incorrectly clustered instances : 14.0 9.33 %

Canopy

The Canopy clustering method accurately measured the number of clusters without any change in parameters and gave an accuracy of (100% - 14.667%) 85.33% in the results.

Canopy clustering
Number of canopies (cluster centers) found: 3
T2 radius: 0.717
T1 radius: 0.896

Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===
Clustered Instances
0 72 (48%)
1 50 (33%)
2 28 (19%)

```
  0  1  2  <-- assigned to cluster
  0 50  0 | Iris-setosa
 50  0  0 | Iris-versicolor
 22  0 28 | Iris-virginica
```

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

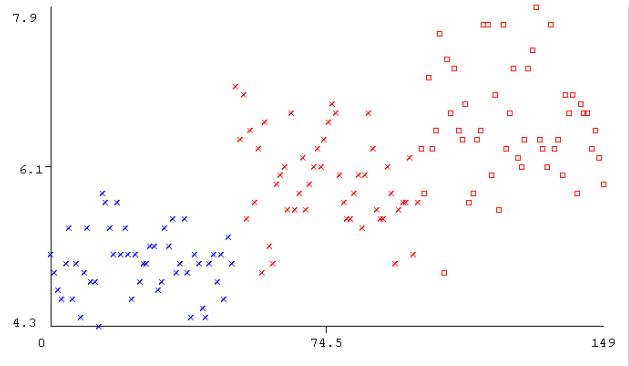
Incorrectly clustered instances : 22.0 14.667 %

RESULTS

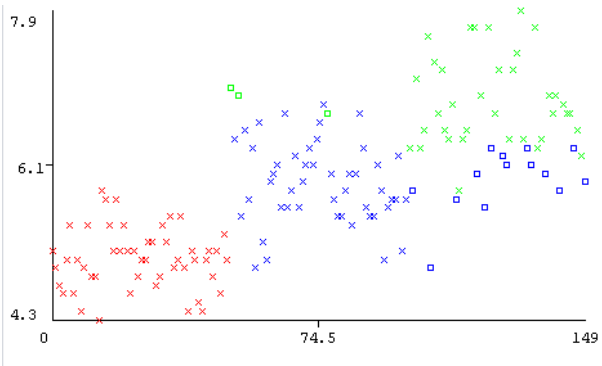
Method	Number of iterations	Sum of squared error	Time taken to build model(sec)	Accuracy(%)
DBSCAN			0	66.66
KMeans	6	6.998	0	88.67
HierarchicalClusterer			0.01	88.67
EM	10		0.01	90.66
Canopy			0	85.33

VISUALIZATIONS

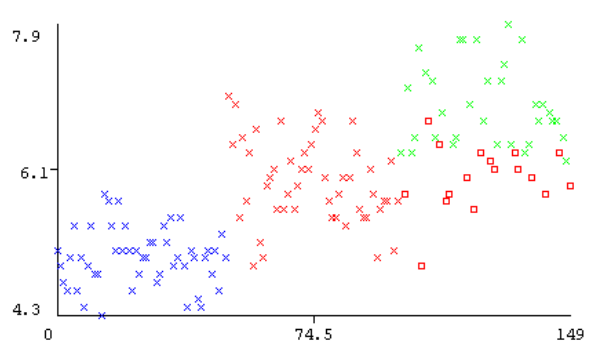
x axis- instance number, y axis- sepallength



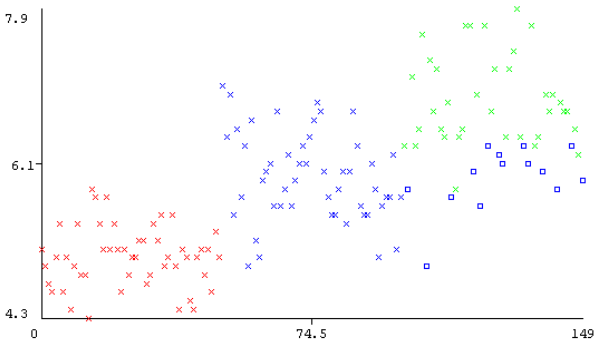
DBSCAN



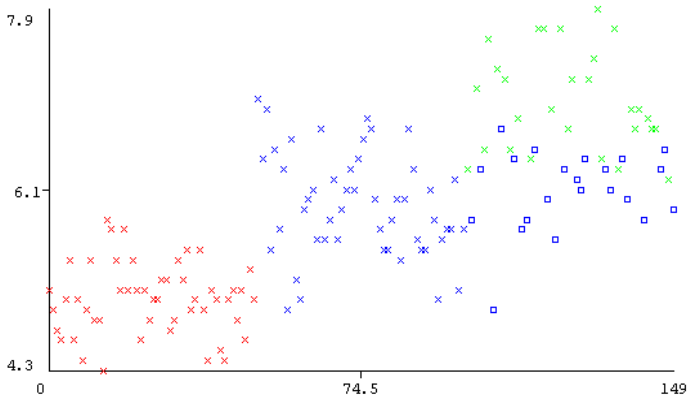
KMeans



HierarchicalClusterer



EM



Canopy

CONCLUSIONS

From the results and visualizations, we can conclude that one of the flower species is linearly separable from the other two, but the other two are not linearly separable from each other. Only the Canopy clustering method accurately measured the number of clusters. From the result table, KMeans clustering algorithm is the simplest clustering method as compared to other methods and its performance is better than Hierarchical Clustering algorithm. HierarchalClusterer method take more time than KMeans. DBSCAN is not suitable for data having very huge variations in density and hierarchical clustering algorithm is more susceptible to noisy data. The EM method takes more time to build a cluster as compared to KMeans, HierarchicalClusterer, and DBSCAN. That is why KMeans and DBSCAN methods are better than the EM method. DBSCAN has a high log likelihood value. Hence KMeans method is the best method because it is faster to build models.