Association Analysis -2

Yusur Al-Mter (yusal621), Roshni Sundaramurthy(rossu809)

March 5, 2019

Clustering:

The performed algorithms chosen to cluster the data are:

1) Simple K-Means.
2) The EM Algorithm.
3) The Hierarchical Clustering algorithm.

The figures below, illustrates the above mentioned algorithms performance, respectively.

Simple K-Means:

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

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

Clustered Instances

0       77 ( 62%)
1       47 ( 38%)


Class attribute: class
Classes to Clusters:

  0  1  <-- assigned to cluster
 40 22 | 0
 37 25 | 1

Cluster 0 <-- 0
Cluster 1 <-- 1

Incorrectly clustered instances :       59.0     47.5806 %
```

The EM Algorithm:

```
Time taken to build model (full training data) : 0.03 seconds

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

Clustered Instances

0       63 ( 51%)
1       61 ( 49%)


Log likelihood: -6.00373


Class attribute: class
Classes to Clusters:

  0  1  <-- assigned to cluster
 34 28 | 0
 29 33 | 1

Cluster 0 <-- 0
Cluster 1 <-- 1

Incorrectly clustered instances :       57.0     45.9677 %
```

The Hierarchical Clustering algorithm:

```
Time taken to build model (full training data) : 0.08 seconds

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

Clustered Instances

0      123 ( 99%)
1        1 (  1%)


Class attribute: class
Classes to Clusters:

  0  1  <-- assigned to cluster
 62  0 | 0
 61  1 | 1

Cluster 0 <-- 0
Cluster 1 <-- 1

Incorrectly clustered instances :       61.0     49.1935 %
```

Why can the clustering algorithms not find a clustering that matches the class division in the database?

Applying different clustering techniques such as K-means, EM Algorithm and the Hierarchical Clustering algorithm. The EM algorithm performed better than the other corresponding algorithms according to the misclassification rate equal to 45.9677%, but it's not necessarily good as it seems to be random guessing. The figures mentioned above, clearly shows that the clustering algorithms couldn't find a clustering that matches the class division in the database, since the clustering algorithm main procedure is trying to minimize the intra-cluster distances between the data points and maximize the inter-cluster distances, and also the type of the data set attributes plays a major role in computing the distances since it's sensitive for the different

distance measurements which leads in experiencing some constrains while trying to minimize/maximize the distances.

Association Analysis:

Using the association analysis and in particular the *Apriori Algorithm* and according to the structure of the association analysis which doesn't require for the data points of the same class to be close to each other, instead, it will find a set of rules that are able to accurately predict the class label from the rest of the attributes, and by setting the minimum support to 0.05 and a maximum number of rules of 19, and by removing the redundant rules, the rules where the antecedent is a super set of the antecedent of another rule, we selected rules predicting class 1 shown in the below figure, where the instance is assigned to class 0 if it is not assigned to class 1.

```
 1. attribute#5=1 29 ==> class=1 29    <conf:(1)> lift:(2) lev:(0.12) [14] conv:(14.5)

2. attribute#1=3 attribute#2=3 17 ==> class=1 17    <conf:(1)> lift:(2) lev:(0.07) [8] conv:(8.5)

5. attribute#1=2 attribute#2=2 15 ==> class=1 15    <conf:(1)> lift:(2) lev:(0.06) [7] conv:(7.5)

14. attribute#1=1 attribute#2=1 9 ==> class=1 9    <conf:(1)> lift:(2) lev:(0.04) [4] conv:(4.5)
```

By using Association analysis, we notice a significant performance and better results than for the clustering algorithm, we are able to find the clusters in the data and combining them to get a decent separable classes. Whereas the clustering algorithms were unsuccessful to perform well and produce the desired results.

Would you say that the clustering algorithms fail or perform poorly for the monk1 dataset? Why or why not?

The clustering algorithm fails with the monk1 data set for many reasons, mainly because of the inappropriate data pre-processing, hence, when using Simple k-Means algorithm the distance metrics used for clustering whether *Euclidean* or *Manhattan* were unable to perform correctly and find a proper separation between the data points.