Association Analysis (2) - the importance of the distance metric used within the clustering algorithm

Nahid Farazmand (nahfa911) and Anubhav Dikshit (anudi287)

March 1, 2019

Dataset

Monk dataset contains 124 instance and each instance has 6 attributes and belongs to one of 2 classes. Here you can see 10 first instances of the dataset.

```
##    att1 att2 att3 att4 att5 att6 class
## 1     1    1    1    1    3    1     1
## 2     1    1    1    1    3    2     1
## 3     1    1    1    3    2    1     1
## 4     1    1    1    3    3    2     1
## 5     1    1    2    1    2    1     1
## 6     1    1    2    1    2    2     1
## 7     1    1    2    2    3    1     1
## 8     1    1    2    2    4    1     1
## 9     1    1    2    3    1    2     1
## 10    1    2    1    1    1    2     1
```

Clustering

Here we want to investigate whether or not the clustering algorithms can find the classes which has been identified before. For this purpose, we First, cluster the data with different algorithms and number of clusters by Use of the Clusters to class evaluation model to see whether the clustering algorithm is able to discover the class division existing in the data.

Here, we used 4 different cases:

```
1. 2-Mean algorithm, seed = 10

2. 2-Mean algorithm, seed = 10

3. Density-Based Cluster with minStdDev = 1.0E-6 and 2-Mean

4. Density-Based Cluster with minStdDev = 1.0E-6 and 4-Mean
```

The confusion matrices for these cases can bee seen in the picture bellow:

Confusion Matrix - Classes vs Clusters

2-Mean algorithm

```
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 %
```

4-Mean algorithm

```
Classes to Clusters:

  0  1  2  3  <-- assigned to cluster
 29 17 11  5 | 0
 21 19 13  9 | 1

Cluster 0 <-- 0
Cluster 1 <-- 1
Cluster 2 <-- No class
Cluster 3 <-- No class

Incorrectly clustered instances :       76.0     61.2903 %
```

Density-Based Cluster with 2-Mean

```
Classes to Clusters:

  0  1  <-- assigned to cluster
 44 18 | 0
 39 23 | 1

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

Incorrectly clustered instances :       57.0     45.9677 %
```

Density-Based Cluster with 4-Mean

```
Classes to Clusters:

  0  1  2  3  <-- assigned to cluster
 28 17 12  5 | 0
 23 18 11 10 | 1

Cluster 0 <-- 0
Cluster 1 <-- 1
Cluster 2 <-- No class
Cluster 3 <-- No class

Incorrectly clustered instances :       78.0     62.9032 %
```

It can be clearly seen that the algorithms could not discover the class division existing in the data!

Now we want to find out why the clustering algorithms were not successful, for this purpose we got help from Association analysis. By Use of association analysis we will try to find a set of rules that are able to accurately predict the class label from the rest of the attributes. Actually, we want to mine the classes to find out how we can decribe the classes by looking at the attributes then we can find out how was the performance of our clustering algorithms.

Association analysis

For association analysis we used density-Based Cluster with minStdDev = 1.0E-6 and 2-Mean and we added the cluster attribute to the dataset.Then we implement the association analysis with minimum support of 0.05 and a maximum number of rules of 19 and Tried to find as few rules predicting class 1 as possible. The rules are as bellow:

```
Apriori
=======

Minimum support: 0.05 (6 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 19

Best rules found:

1. attribute#1=1 attribute#5=1 6 ==> class=1 6    <conf:(1)>
2. attribute#1=2 attribute#5=1 attribute#6=1 6 ==> class=1 6    <conf:(1)>
3. attribute#1=2 attribute#2=2 attribute#3=2 6 ==> class=1 6    <conf:(1)>
4. attribute#1=2 attribute#2=2 attribute#4=3 6 ==> class=1 6    <conf:(1)>
```

By looking at the rules we can obviousely see that all founded rules have minimum support (not more than that) this can be one of the reasons why the clustering algorithm could not find the classes. On the other hand, the more important problem is that the attributes are categorical (not numerical) and here the only way

that we can estimate the proximity between instances were euclidean and Manhattan distances which can be used for numerical data. This is the most important reason why the clustering algorithm could not discover the class division existing in the data. We would say that with the euclidean and Manhattan formlas we cannot obviousely distinguish between these rules and other combination of attributes in different instances.