732A61
Lab 3

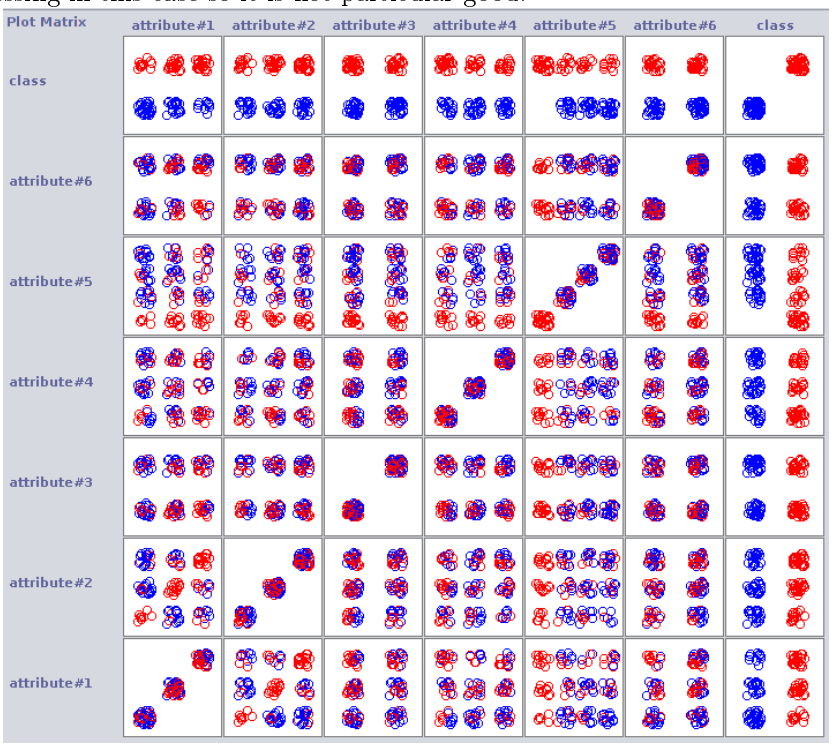
Rasmus Holm

March 1, 2017

1 Clustering

From the plot below we can see that the two classes overlap in all the variables and therefore it is impossible to separate the two classes perfectly in the input space for the methods I tried.

I tried k-means, density based k-means, hierarhical, and EM with EM performing the best with a misclassification rate of 42.7% which is close to random guessing in this case so it is not particular good.



A major flaw with these tests are that clustering algorithms are not designed to classify observations so the objective functions are completely different from what a classification algorithm would have. Clustering algorithms tries to minimize the intra-cluster distances and maximize the inter-cluster distances and categorical variables are particular sensitive for distance measures like the Euclidean distance since they are either the same or not the same.

2 Association Analysis

I used the Apriori algorithm with minimum support of 0.05 and 19 rules in total. Below are the rules I choose.

2.1 Rules

- attribute#5=1 29 ==> class=1 29 <conf:(1)>
- attribute#1=3 attribute#2=3 17 ==> class=1 17 <conf:(1)>
- attribute#1=2 attribute#2=2 15 ==> class=1 15 <conf:(1)>
- attribute#1=1 attribute#2=1 9 ==> class=1 9 <conf:(1)>

I choose these particular rules since those correspond to those I found in the documentation about the data set.

The rules correspond to the following plots:

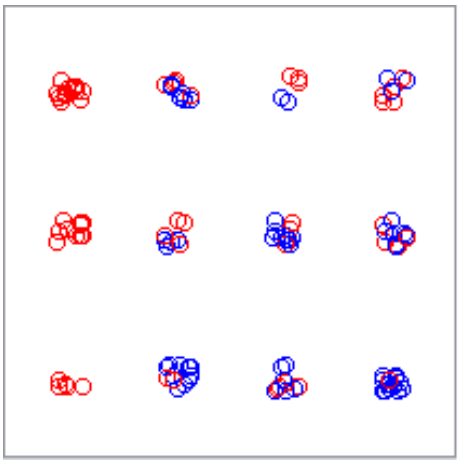


Figure 1: X-axis: Attribute 5, Y-axis: Attribute 2 The red clusters to the left.

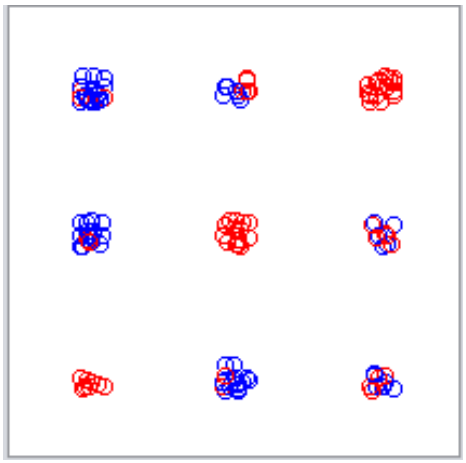


Figure 2: X-axis: Attribute 1, Y-axis: Attribute 2
The red clusters in the anti-diagonal

2.2 Analysis

Association analysis have an advantage here of being able to find smaller clusters within the data and then aggregate those together to perfectly distinguish the classes in this particular case. The clustering algorithms do not possess this ability and therefore cannot perform as good.