Lab 3 – 732A61

Emil K Svensson emisv463

Linköping University 2017-02-24

# Clustering

## I tried three clustering techniques with two distance measures. The EM-, K-Means- and Hierarchical-clustering algorithm was used along with the Euclidean and Manhattan-distance. The best result was from the Hierarchical-clustering with Wards-measure, this algorithm had a misclassification-rate of about 38 % which is good compared to the other algorithms that had misclassification-rate around 45- 50 % which is close to guessing the class-belonging.

The algorithms struggle to classify correctly due to several factors. First, since the different class labels are spread out in a large space since the data is categorical the algorithms will have problems trying to minimize distances. This becomes a larger issue when trying to find two clusters in a data set where there is a larger number of natural clusters than specified. Lastly a clustering algorithm has no proper objective function to find correct class labels, it sole purpose is to minimize the total squared with in error of each cluster not to minimize the misclassification rate.

# Association Analysis

In this task these following rules were selected to predict the 62 observations from the 19 rules generated.

1. attribute#5=1 29 ==> class=1 29 conf:(1)

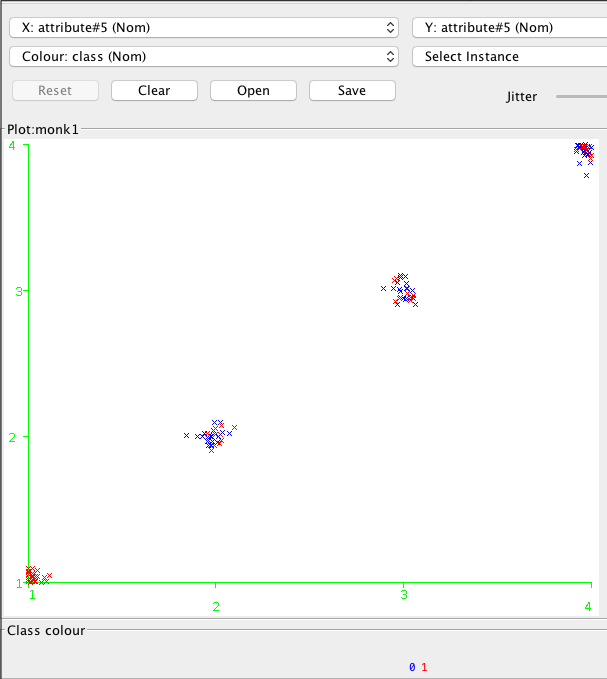
5. attribute#1=2 attribute#2=2 15 ==> class=1 15 conf:(1)

14. attribute#1=1 attribute#2=1 9 ==> class=1 9 conf:(1)

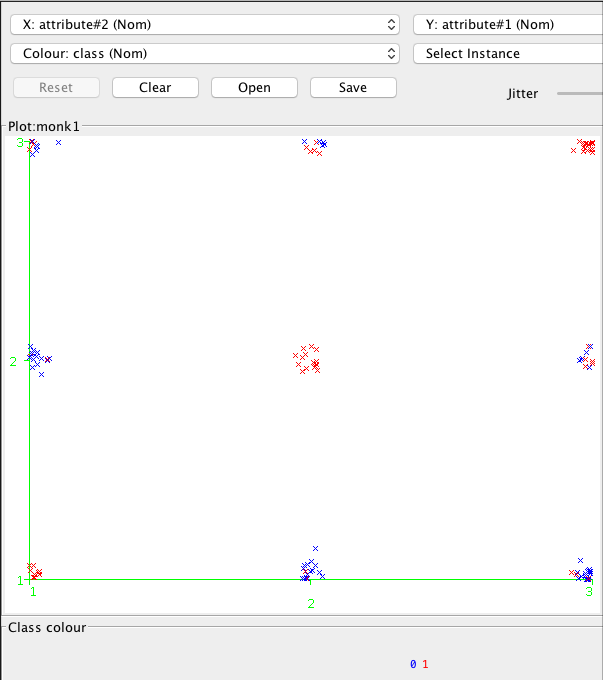
18. attribute#1=3 attribute#2=3 attribute#3=1 9 ==> class=1 9 conf:(1)

The process of selecting rules was a bit tedious since when a rule is selected it is not allowed to be in any other rule selected. So for instance when rule 1 was selected no more rules with attribute#5=1 can be selected. This because rules that have the same conditions might be an intersect or a subset of the other rule and there is no way to tell which of the observations were selected. This makes the greedy approach a bad approach (i.e. selecting the highest available support/confidence each time) since we probably won’t find the optimal set of rules as we did this time.

But instead of thinking about this as superset or subset is easier to show and select rules using visualization-tools.



So basically what we are doing with rule 1. is selecting the bottom left cluster (the red) in the graph above.



With rule 5. 14. And 18. We are selecting (identifying) the red clusters in the anti-diagonal of the graph.

This demonstrates the power of association analysis where we can see and select clearly separable clusters formed in smaller natural cluster that a clustering-algorithm can’t find. Again this is since they don’t have target for the clustering and just focuses on its objective function minimizing the distances with in the clusters.