TDDD41 Data Mining - Clustering and Association Analysis

Lab 3 - Group 4 Report

Lawrence Thanakumar Rajappa (lawra776)
Kyriakos Domanos (kyrdo817)

March 7, 2020

Association Analysis-2

Data Preparation

In this lab assignment, we used Monk1 dataset which contains 124 samples of data, 7 attributes and target variable is binary. As a starting process, we checked whether the data contains continuous format, but all attributes were discrete.

Clustering

Now, we tried to apply ***SimpleKmeans*** clustering algorithm with the following specifications;

* **Seed value** : 10
* **No. of Clusters** : 2

and igonred ***class*** attribute and selected Classes to clusters evaluation to crosstabulate the clustering. We have got 47.5806% incorrectly clustered instances. When we tried with ***MakeDensity BasedClusterer*** with the default parameters, we got 45.9677% which was not a good improvement.

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

When visualizing the results of cluster, we could see that there is overlapping of data points and moreover, there is not well-defined boundary, this makes the separation tedious. In order to have a proper separation, similar data points should lie close to each other, this would make the jobs of clustering algorithm easy. In order to have a proper clustering, we could either use ***overlapping clustering*** technique or preprocess the data by adding some more attributes to the data which could increase the chances of separation.

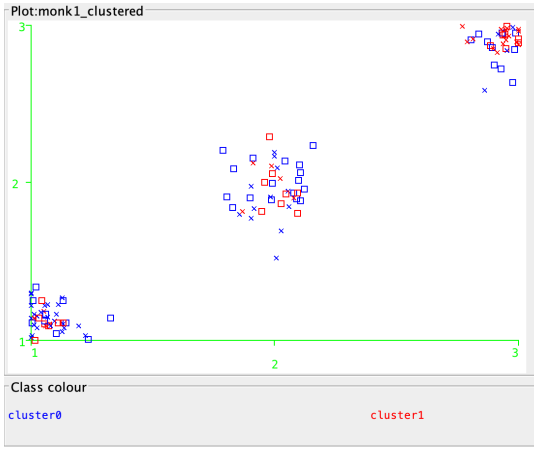

Figure 1: Overlapping data points

Would you say that it goes poorly for monk1? Why or why not?

From the data, we could conclude that the results of clustering algorithms were bad are because of characteristics of the data. As said earlier, if we provide the preprocessed data to the algorithms, we could get a better result and performance. But there is another way to handle this, i.e. we need to create a new distance function which can be incorporated into the existing algorithms to create a separation for this type of data, but on the other hand, it requires lot of expertise and knowledge for a domain for the data related to that domain.

Association Analysis

In order to perform association analysis on the dataset, it is not a thumb rule that similar points should lie close to each other, only rules are required to classify the points based on the pattern in the data.

We then performed the ***addCluster*** in the preprocess section and applied ***Apriori Algorithm***, and got the following association rules table which is given below;

Association rule(s)	**Cluster**	**Occurrence**	**Confidence**
attribute#5=1	1	29	1
attribute#1=3 attribute#2=3	1	17	1
attribute#5=1 attribute#6=2	1	13	1
attribute#1=2 attribute#2=2	1	15	1

We ignore some rules in the association analysis result where antecedent is a super set of antecedent of another rule which are not useful for our analysis. Having lower confidence yields more information for super set antecedent and we should not discard them. So, in this case the above rule of ignoring association rules where antecendent is a super set of antecendent of another rule applies, but it is not a general rule, it can be applied when the rules have confidence = 1.