TDDD41/732A75: Association Analysis -2

Goals

- Appreciate the importance of the distance metric used within the clustering algorithm, e.g. the clusters found by the algorithm may not correspond to the clustering that is most natural for us human beings.

Procedure

- **Dataset**

  In this exercise, you will also work with a well-known dataset called the monk1 dataset. The dataset is available here . This is an artificial dataset with 124 instances, each described by 6 discrete attributes and a binary class attribute. No more information about the data is given though you may need it to solve the problem below. Search for it ! We recommend you to check the UCI Machine Learning Repository. You will see that there are 3 versions of the so-called monk problem: We are using the first version in this exercise.

- **Clusters may not correspond to classes**

  First, cluster the data with different algorithms and number of clusters. Use the Clusters to class evaluation model to see whether the clustering algorithm is able to discover the class division existing in the data. Hopefully, you won't be able to discover it. In this exercise, we want you to answer to the following question: Why can the clustering algorithms not find a clustering that matches the class division in the database? Association analysis may help you to find the answer. Proceed as follows. Use association analysis to find a set of rules that are able to accurately predict the class label from the rest of the attributes. We recommend you to use a minimum support of 0.05 and a maximum number of rules of 19. Note that the class attribute is binary, so it suffices to find rules that accurately predict class 1, i.e. an instance is assigned to class 0 if it is not assigned to class 1. Try to find as few rules predicting class 1 as possible, i.e. try to remove redundant rules. Hopefully, you will be able to perfectly describe class 1 with only 4 rules. Now, try to answer the question above. Finally, would you say that the clustering algorithms fail or perform poorly for the monk1 dataset? Why or why not?

Submission

Submit a report on how you used association analysis and the results you obtained as well as answers to all questions in the text.