732A75 Advanced Data Mining laboratory 3 report

Yuki Washio and Nicolas Taba

March 4, 2021

1 Introduction

The aim of this laboratory exercise is to test the limits of the clustering algorithm using the distance metric and understand that the clsutering algorithms can sometimes fail to provide clusters that correspond to classes that we (humans) may intuit better.

In order to illustrate this, we are using the MONK1 dataset. [Present the dataset here]. [FIND REFERENCE]

We will first start an exploratory analysis and clustering using appropriate clustering methods for our problem using the distance as a metric and trying to improve on it as much as possible before turning to assosciation analysis.

2 Clustering

In this section of the laboratory, we know the make up of this dataset and want to familiarize ourselves with the assosciation analysis tool of Weka. We begin by discretizing the dataset in 3 bins of data since continuous attributes cannot be processed by the software. We perform the simple Kmeans algorithm on the data like in the previous laboratory and crosstabulate the clusters with the class labels.

(a) data separated into 3 bins according to petal-width

(b) kmeans clustering using k=3

Figure 1: clustering using Kmeans with k=3 and n=3 bins

Here we see that a few elements of versicolor and virginica are wrongly clustered and account for a 68% error in the clustering. We now perform the apriori algorithm on this clustering to try and ascertain if the rules used to make this clusters have high confidence.

We use the apriori algorithm to find rules that have the clusters as consequent.

Figure 2: Apriori rules for k=3 and n=3 bins.

3 Assosciation analysis

4 Discussion