732A75 Advanced Data Mining laboratory 3 report

Yuki Washio (timwa902) and Nicolas Taba (nicta839)

March 24, 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 MONK dataset. This dataset presents 432 data points assosciated to 7 categorical attributes for the purposes of classification. [use link that is in masters bookmarks]. The MONK1 dataset is used to try to classify (class 0 or class 1) using the following constraints: a1 = a2 and a5 = 1, where a1, a2, a5 are attributes of the data points.

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

The dataset presents categorical data. We should use hierarchical clustering methods to create our clusters and obtain the desired 2 clusters.

We are going to use the complete linkage that uses euclidean distance as a metric.

(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

The algorithm does not perform well and guesses incorrectly about half of the time. This is not a satisfactory result. [exploration of the data by looking at a particular attribute].

Find a way to improve the classification by adding an outlier cluster.

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

3 Assosciation analysis

look at the rules of the problem that we want to solve and find them in there

4 Discussion