02450 Project 2

Report

by Karol Dzitkowski
Marco Becattini

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Introduction to Machine Learning and Data Mining
Tue Herlau
28th November 2014

Abstract

The objective of this second report is to apply the methods you have learned in the second section of the course on „Supervised learning: Classification and regression" in order to solve both a relevant classification and regression problem for your data.

Classification

2.1 Problem

Association Mining

Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. We divide values of features into three categories: small values, medium values and high values and than we are going to do association mining for this three categories separately. So we have three different dataset and attributes are in set when they belong to that category. For instance if we consider the item set of smal values (from 1 to 4) we binarize our data by transforming the value of each attribute to 1 if it belong to the interval [1-4], and 0 otherwise. The medium values are defined from 5 to 7 while high values are defined from 8 to 16 and we binarized the data in the same way described above. For each item set we apply Apriori algorithm in order to find possible associations among the attributes.

The table below shows the best association rules found by the algorithm including its confidence percentage.

RULE	CONFIDENCE
Assiociations for HIGH values of attributes (from 8 to 16)	
xegvy → xy2br	62.96 %
xy2br → xegvy	74.62 %
yegvx → xy2br	72.00 %
xy2br → yegvx	72.64 %
yegvx → xegvy	76.39 %
xegvy → yegvx	65.03 %
Assiociations for MEDIUM values of attributes (from 4 to 7)	
x2ybr → width	67.53 %
width → x2ybr	60.37 %
width → x-box	63.67 %
x-box → width	82.51 %
high → width	80.93 %
width → high	69.55%
Assiociations for SMALL values of attributes (from 1 to 4)	
x-ege → onpix	76.01 %
onpix → x-ege	72.54 %
y-ege → onpix	78.07 %
onpix → y-ege	60.95 %
onpix → x-box	76.15 %
x-box → onpix	81.39 %
x-ege → x-box	68.88 %
x-box → x-ege	70.25 %

In the first report we computed correlation between attributes and we found out that the strongest correlation was between x-box and width (corr=0.85). Thus the results of Apriori confirms the findings from the our previous analysis ,in fact, the association for medium values of attributes x-box → width has the highest confidece.