UNIVERSITY OF MANNHEIM

Data and Web Science Group
Prof. Dr. Christian Bizer
B6 – B1.15
68159 Mannheim

Data Mining – FSS 2020

Exercise 6: Association Analysis

6.1. Analyzing the Shopping Basket Data Set

1. The Shopping Basket data set is provided as an Excel file on the website. Load the data set into RapidMiner. In the import dialog, specify that the BasketNo attribute has the role id and that all other attributes have the type binominal.
2. Mine frequent item sets from the data set using the FP-Growth operator (support = 0.2, positive value = 1). Which items are usually bought together with the laptop, the netbook and the printer?
3. Create association rules from the frequent item sets. What do the rules tell you about the relationship between Asus EeePC netbooks, 2 GB DDR3 RAM extensions and Netbook Schutzhüllen? What do the lift values tell you about the interestingness of the rules?

6.2. Finding Frequent Pattern in the Adult Data Set

1. Import the Adult-tweaked data set into RapidMiner. The Adult-tweaked data set is provided on the website as an ARFF file.
2. Prepare the data set for Frequent Pattern Mining by: 1. reducing the size of the data set to 5000 examples using sampling; 2. removing the attributes fnlwgt, education-num, capital-gain, capital-loss, marital-status and relationship; 3. discretizing the attributes age and hours-per-week into three user defined ranges (think about ranges that could make sense for the attributes), 4. converting all attributes into binominal attributes. How many attributes does the resulting data set have?
3. Apply the FP-Growth operator to find the frequent item sets that have a support above 0.2. What can you learn from these item sets about the people how earn less than 50K a year?
4. Given the large number of examples and the low min-support threshold, the number of frequent item sets containing *education* attribute is surprisingly low. Moreover, only one value for this attribute is present in the resulting frequent item sets. Why is this the case? How could you aggregate the data to change this without losing too much information? Also look at the *native-country* attribute, and think of a possible aggregation for it.
5. Use the FP-Growth *must contain* parameter to restrict pattern to the ones containing “class = >50K” and lower the support so that a decent number of item sets is discovered. What can you learn from these item sets about the people who earn more than 50K a year?

6.3. Mining Association Rules from the Adult Data Set

1. Use the Create Association Rules operator to create association rules from the frequent item set resulting from exercise 6.2.3. Which rules do you consider interesting? Consider both = >50K and <50K classes.
2. Now we want to focus on the relationship of the occupation, education and income of immigrants: instead of sampling the data set, filter the data set with an invert filter: native-country = United-States. Which rules do you consider interesting?