

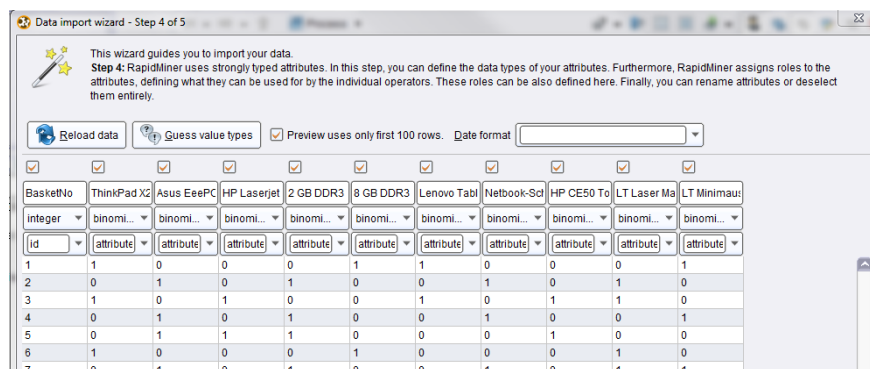
Data Mining – FSS 2021

Exercise 8: Association Analysis

8.1. 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.

Solution: Use the “Import Excel” function of RapidMiner and select the first tab of the excel document. Annotate the first column (Step 3 of 5) with name. Setup the types as described in step 4 of 5 and import the dataset.



2. Mine frequent item sets from the data set using the FP-Growth operator (find min number of itemsets = false, support = 0.2, positive value = 1). Which items are usually bought together with the laptop, the netbook and the printer?

Solution: Design a process with the retrieve operator and the FP-Growth operator. Go to the FP-Growth Tab in the result view and filter the item sets according to ThinkPad X220, HP Laserjet and Asus EeePC.

Conclusion: Items bought with

- Asus EeePC: 2GB DDR3 RAM (0.5), LT Minimaus (0.4), Netbook-Schutzhülle (0.4)
- HP Laserjet P2055: HP CE50 Toner (0.3)
- ThinkPad X220: Lenovo Tablet Sleeve (0.3)

Only the outstanding items are lined up above. Support is in brackets.

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?

Solution: Add the Create Association Rules operator in the process from 6.1.2. Use the table view of the AssociationRules tab in the result view to overview the created rules. You can filter by conclusion to get a better overview about the learned rules. Use a min confidence of 0.7.

Conclusion: The algorithm creates several different rules for the 3 mentioned products. All the rules have a lift value higher than 1 which means there are positively correlated – in this case, if one combination of products (left side – premises) is bought also the combination of products (right side – conclusions) are bought.

From the rules we can learn the following (support, confidence, lift):

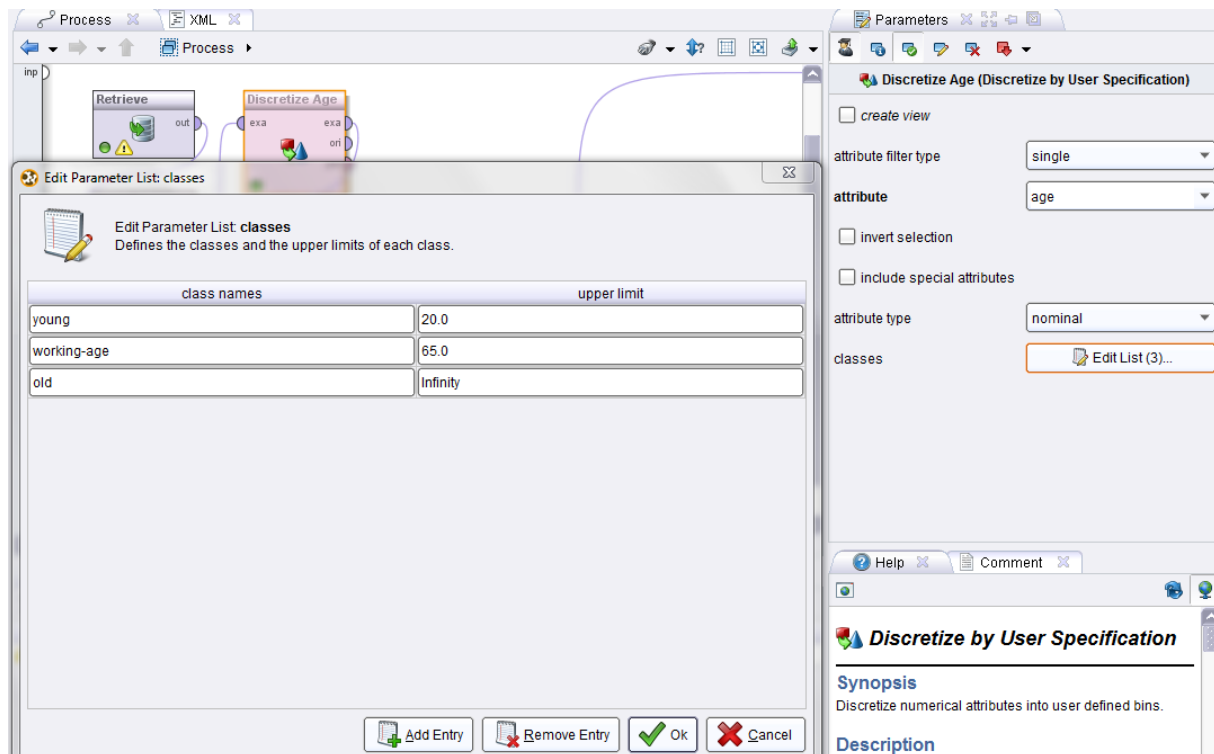
- If the 2 GB DDR3 RAM extensions is bought also the Asus EeePC is bought (0.5, 1, 1.667)
- If the Netbook-Schutzhülle is bought also the Asus EeePC is bought (0.4, 1, 1.667)
- If the 2 GB DDR3 RAM extension is bought 80% of all people also buy the Netbook-Schutzhülle (0.4, 0.8, 2)
- If the Netbook-Schutzhülle is bought also the 2 GB DDR3 RAM is bought (0.4, 1, 2)

From those relationships all other combinations can be calculated.

8.2. 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?

Solution: Build up the process as described in the task. Use selection attribute operator to select all parameters needed. Have a look at the data ranges of the 2 attributes: age and hours-per-week to find some fitting bins. Use the Discretize by User Specification operator, input and name the ranges in the operator's settings.



A possible setup for the ranges is:

- Age: ≤ 20 , $21 < x < 65$ and > 65
- Hours-per-week: ≤ 20 , $20 < x \leq 45$, > 45

Convert all attributes to binominal attributes using the “Nominal to Binominal” operator.

Conclusion: We come up with 96 attributes.

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?

Conclusion: Using the Data View on the FrequentItemSets Tab in the result view we can learn that people who are earning less than 50K are most likely from the United-States and/or within the working age and/or white and/or work 20-45 hours per week and/or are from the Private sector (workclass=private). At this point we do not know about the confidence of these combinations but we know how often they appear together.

Result Overview					
FrequentItemSets (FP-Growth)					
ExampleSet (Nominal to Binominal)					
	No. of Sets: 77	Size	Support	Item 1	Item 2
Total Max. Size: 6	2	0.687		native-country = United-States	class = <=50K
	2	0.668		age = working-age	class = <=50K
	2	0.643		race = White	class = <=50K
Min. Size: 2	2	0.555		class = <=50K	workclass = Private
Max. Size: 7	2	0.556		class = <=50K	hours-per-week = normal
Contains Item:	2	0.479		class = <=50K	sex = Male
class = <=50K	2	0.281		class = <=50K	education = HS-grad
	2	0.291		class = <=50K	sex = Female
Update View	3	0.591		native-country = United-States	age = working-age
	3	0.589		native-country = United-States	race = White
	3	0.486		native-country = United-States	class = <=50K
	3	0.492		native-country = United-States	class = <=50K
	3	0.424		native-country = United-States	class = <=50K
	3	0.258		native-country = United-States	class = <=50K
	3	0.263		native-country = United-States	class = <=50K
	3	0.552		age = working-age	race = White
	3	0.485		age = working-age	class = <=50K
	3	0.496		age = working-age	class = <=50K

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.

Solution: As we use the nominal to binominal operator we have a lot of attributes after this transformation, e.g. for education we have 16. We could try to find an aggregation for some of such nominal attributes. To reduce the chance to lose information, first have a look on the number of instances which will be included in your new created groups so that you are not creating new groups where one characteristics include over 90% of all items in the dataset. Use the Map operator.

Conclusion: Possible aggregation for education could be

- School (Preschool, 1st-12th)
- College (Some-college)
- HS-Grad (HS-Grad)
- Other-Grad (Assoc-acdm, Prof-school, Bachelors, Masters, Doctorate, Assoc-voc)

Using this aggregation, and the aggregation of US and Non-US for the *native-country* attribute (use “add default mapping” to avoid manually specifying mappings for Non-US countries) we came up with 45 attributes and 196 item sets.

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?

Conclusion: Set the support down to 0.1, and you will get a sufficient number of item sets. But as you can see they convey very similar meaning as the sets you get for “less than 50K”. When looking at the data you can see that all attributes in the discovered item sets are really strong represented in the data set (see the Statistics view of the ExampleSet tab, check the value distributions).

8.3. 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.

Solution: Add the Create association Rules operator into your process and select <=50K within the filtering of the result view tab.

Conclusion: Some of the rules which are calculated and are interesting (based on support, confidence and lift) are:

- Almost all women earn less than 50K
- If you have only a HS-Grad you will not earn more than 50K

When tuning the parameter (confidence 0.2 [Create Association Rules] and support 0.1 [FP-Growth] we can find out that the rules with the highest confidence according to conclusion class = >50K are related to sex = Male, race = White and native-country = United-States.

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?

Solution: Replace the sample operator with a Filter Sample operator.

Conclusion: Using a 0.04 min support value and a 0.5 min confidence value the following rules may be interesting:

- Being Male with Prof-specialty leads to an income over 50K for immigrants