CHAPTER FIVE:
ASSOCIATION RULES

CONTEXT AND PERSPECTIVE

Roger is a city manager for a medium-sized, but steadily growing, city. The city has limited resources, and like most municipalities, there are more needs than there are resources. He feels like the citizens in the community are fairly active in various community organizations, and believes that he may be able to get a number of groups to work together to meet some of the needs in the community. He knows there are churches, social clubs, hobby enthusiasts and other types of groups in the community. What he doesn't know is if there are connections between the groups that might enable natural collaborations between two or more groups that could work together on projects around town. He decides that before he can begin asking community organizations to begin working together and to accept responsibility for projects, he needs to find out if there are any existing associations between the different types of groups in the area.

LEARNING OBJECTIVES

After completing the reading and exercises in this chapter, you should be able to:

- Explain what association rules are, how they are found and the benefits of using them.
- Recognize the necessary format for data in order to create association rules.
- Develop an association rule model in RapidMiner.
- Interpret the rules generated by an association rule model and explain their significance, if any.

ORGANIZATIONAL UNDERSTANDING

Roger's goal is to identify and then try to take advantage of existing connections in his local community to get some work done that will benefit the entire community. He knows of many of

the organizations in town, has contact information for them and is even involved in some of them himself. His family is involved in an even broader group of organizations, so he understands on a personal level the diversity of groups and their interests. Because people he and his family knows are involved in other groups around town, he is aware in a more general sense of many different types of organizations, their interests, objectives and potential contributions. He knows that to start, his main concern is finding types of organizations that seem to be connected with one another. Identifying individuals to work with at each church, social club or political organization will be overwhelming without first categorizing the organizations into groups and looking for associations between the groups. Only once he's checked for existing connections will he feel ready to begin contacting people and asking them to use their cross-organizational contacts and take on project ownership. His first need is to find where such associations exist.

DATA UNDERSTANDING

In order to answer his question, Roger has enlisted our help in creating an **association rules** data mining model. Association rules are a data mining methodology that seeks to find frequent connections between attributes in a data set. Association rules are very common when doing shopping basket analysis. Marketers and vendors in many sectors use this data mining approach to try to find which products are most frequently purchased together. If you have ever purchased items on an e-Commerce retail site like Amazon.com, you have probably seen the fruits of association rule data mining. These are most commonly found in the recommendations sections of such web sites. You might notice that when you search for a smartphone, recommendations for screen protectors, protective cases, and other accessories such as charging cords or data cables are often recommended to you. The items being recommended are identified by mining for items that previous customers bought in conjunction with the item you search for. In other words, those items are found to be *associated* with the item you are looking for, and that association is so frequent in the web site's data set, that the *association* might be considered a *rule*. Thus is born the name of this data mining approach: "association rules". While association rules are most common in shopping basket analysis, this modeling technique can be applied to a broad range of questions. We will help Roger by creating an association rule model to try to find linkages across types of community organizations.

Working together, we using Roger's knowledge of the local community to create a short survey which we will administer online via a web site. In order to ensure a measure of data integrity and to try to protect against possible abuse, our web survey is password protected. Each organization invited to participate in the survey is given a unique password. The leader of that organization is asked to share the password with his or her membership and to encourage participation in the survey. Community members are given a month to respond, and each time an individual logs on complete the survey, the password used is recorded so that we can determine how many people from each organization responded. After the month ends, we have a data set comprised of the following attributes:

- **Elapsed_Time**: This is the amount of time each respondent spent completing our survey. It is expressed in decimal minutes (e.g. 4.5 in this attribute would be four minutes, thirty seconds).
- **Time_in_Community**: This question on the survey asked the person if they have lived in the area for 0-2 years, 3-9 years, or 10+ years; and is recorded in the data set as Short, Medium, or Long respectively.
- **Gender**: The survey respondent's gender.
- **Working**: A yes/no column indicating whether or not the respondent currently has a paid job.
- **Age**: The survey respondent's age in years.
- **Family:** A yes/no column indicating whether or not the respondent is currently a member of a family-oriented community organization, such as Big Brothers/Big Sisters, childrens' recreation or sports leagues, genealogy groups, etc.
- **Hobbies:** A yes/no column indicating whether or not the respondent is currently a member of a hobby-oriented community organization, such as amateur radio, outdoor recreation, motorcycle or bicycle riding, etc.
- **Social_Club:** A yes/no column indicating whether or not the respondent is currently a member of a community social organization, such as Rotary International, Lion's Club, etc.
- **Political:** A yes/no column indicating whether or not the respondent is currently a member of a political organization with regular meetings in the community, such as a political party, a grass-roots action group, a lobbying effort, etc.

- **Professional:** A yes/no column indicating whether or not the respondent is currently a member of a professional organization with local chapter meetings, such as a chapter of a law or medical society, a small business owner's group, etc.
- **Religious:** A yes/no column indicating whether or not the respondent is currently a member of a church in the community.
- **Support_Group:** A yes/no column indicating whether or not the respondent is currently a member of a support-oriented community organization, such as Alcoholics Anonymous, an anger management group, etc.

In order to preserve a level of personal privacy, individual respondents' names were not collected through the survey, and no respondent was asked to give personally identifiable information when responding.

DATA PREPARATION

A CSV data set for this chapter's exercise is available for download at the book's companion web site (https://sites.google.com/site/dataminingforthemasses/). If you wish to follow along with the exercise, go ahead and download the Chapter05DataSet.csv file now and save it into your RapidMiner data folder. Then, complete the following steps to prepare the data set for association rule mining:

1) Import the Chapter 5 CSV data set into your RapidMiner data repository. Save it with the name Chapter5. If you need a refresher on how to bring this data set into your RapidMiner repository, refer to steps 7 through 14 of the Hands On Exercise in Chapter 3. The steps will be the same, with the exception of which file you select to import. Import all attributes, and accept the default data types. This is the same process as was done in Chapter 4, so hopefully by now, you are getting comfortable with the steps to import data into RapidMiner.

2) Drag your Chapter5 data set into a new process window in RapidMiner, and run the model in order to inspect the data. When running the model, if prompted, save the process as Chapter5_Process, as shown in Figure 5-1.

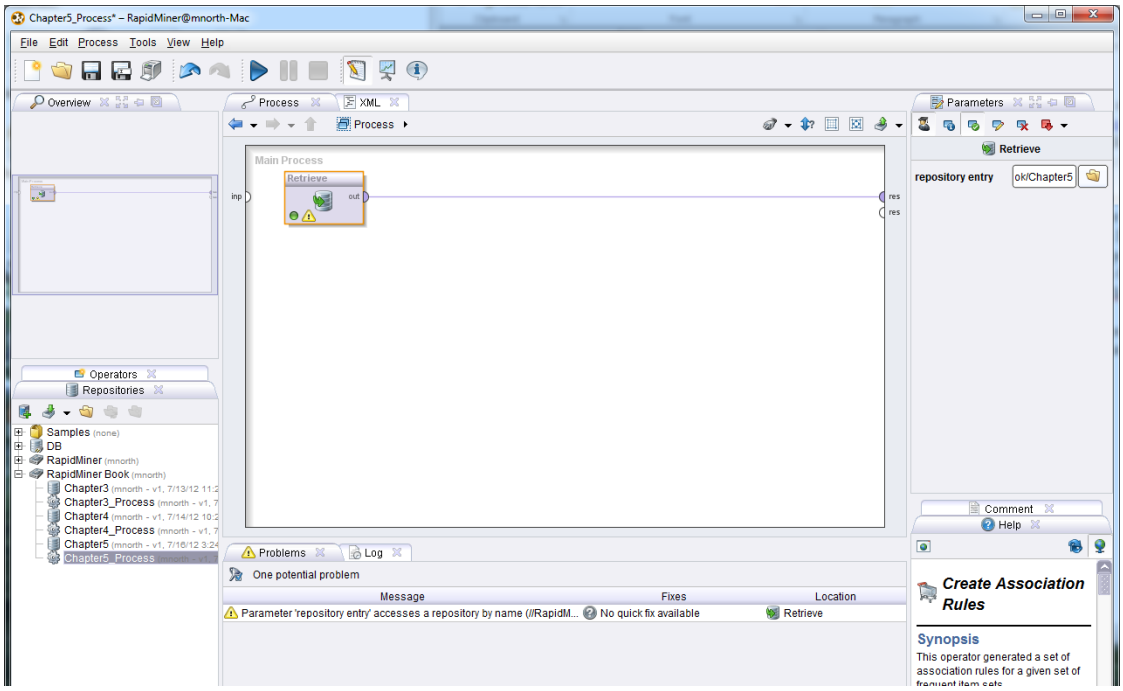
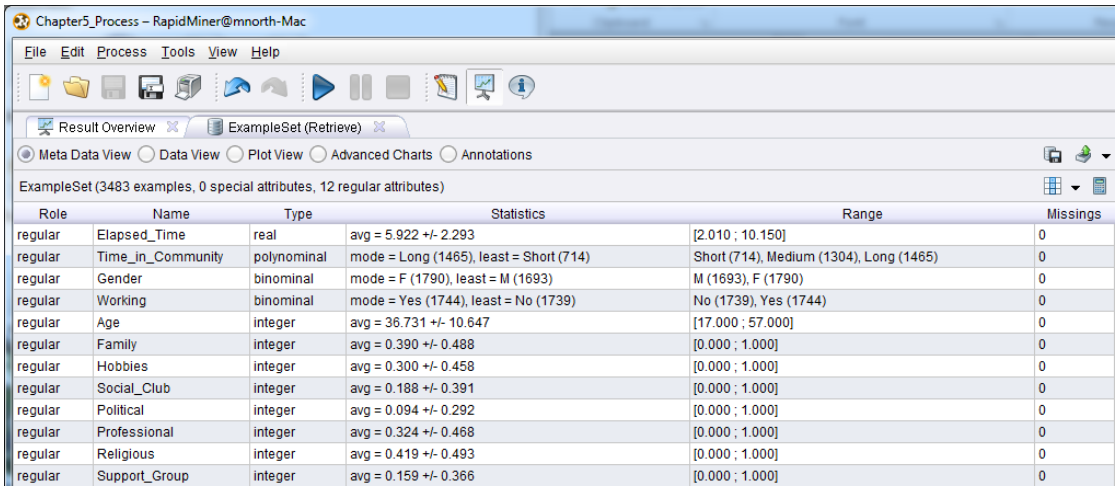


Figure 5-1. Adding the data for the Chapter 5 example model.

3) In results perspective, look first at Meta Data view (Figure 5-2). Note that we do not have any missing values among any of the 12 attributes across 3,483 observations. In examining the statistics, we do not see any inconsistent data. For numeric data types, RapidMiner has given us the **average** (avg), or **mean**, for each attribute, as well the **standard deviation** for each attribute. Standard deviations are measurements of how dispersed or varied the values in an attribute are, and so can be used to watch for inconsistent data. A good rule of thumb is that any value that is smaller than two standard deviations below the mean (or arithmetic average), or two standard deviations above the mean, is a statistical outlier. For example, in the Age attribute in Figure 5-2, the average age is 36.731, while the standard deviation is 10.647. Two standard deviations above the mean would be 58.025 (36.731+(2*10.647)), and two standard deviations below the mean would be 15.437 (36.731-(2*10.647)). If we look at the Range column in Figure 5-2, we can see that the Age attribute has a range of 17 to 57, so all of our observations fall within two standard deviations of the mean. We find no inconsistent data in this attribute. This won't always be the case, so a data miner should always be watchful for such indications of inconsistent data. It's important to realize also that while two standard deviations is a guideline, it's not a hard-and-fast rule. Data miners should be thoughtful about why some observations may be legitimate and yet far from the mean, or why some values that fall within two standard deviations of the mean should still be scrutinized. One other item should be noted as we

examine Figure 5-2: the yes/no attributes about whether or not a person was a member of various types of community organizations was recorded as a 0 or 1 and those attributes were imported as 'integer' data types. The association rule operators we'll be using in RapidMiner require attributes to be of 'binominal' data type, so we still have some data preparation yet to do.



Role	Name	Type	Statistics	Range	Missings
regular	Elapsed_Time	real	avg = 5.922 +/- 2.293	[2.010 ; 10.150]	0
regular	Time_in_Community	polynominal	mode = Long (1465), least = Short (714)	Short (714), Medium (1304), Long (1465)	0
regular	Gender	binominal	mode = F (1790), least = M (1693)	M (1693), F (1790)	0
regular	Working	binominal	mode = Yes (1744), least = No (1739)	No (1739), Yes (1744)	0
regular	Age	integer	avg = 36.731 +/- 10.647	[17.000 ; 57.000]	0
regular	Family	integer	avg = 0.390 +/- 0.488	[0.000 ; 1.000]	0
regular	Hobbies	integer	avg = 0.300 +/- 0.458	[0.000 ; 1.000]	0
regular	Social_Club	integer	avg = 0.188 +/- 0.391	[0.000 ; 1.000]	0
regular	Political	integer	avg = 0.094 +/- 0.292	[0.000 ; 1.000]	0
regular	Professional	integer	avg = 0.324 +/- 0.468	[0.000 ; 1.000]	0
regular	Religious	integer	avg = 0.419 +/- 0.493	[0.000 ; 1.000]	0
regular	Support_Group	integer	avg = 0.159 +/- 0.366	[0.000 ; 1.000]	0

Figure 5-2. Meta data of our community group involvement survey.

4) Switch back to design perspective. We have a fairly good understanding of our objectives and our data, but we know that some additional preparation is needed. First off, we need to reduce the number of attributes in our data set. The elapsed time each person took to complete the survey isn't necessarily interesting in the context of our current question, which is whether or not there are existing connections between types of organizations in our community, and if so, where those linkages exist. In order to reduce our data set to only those attributes related to our question, add a Select Attributes operator to your stream (as was demonstrated in Chapter 3), and select the following attributes for inclusion, as illustrated in Figure 5-3: Family, Hobbies, Social_Club, Political, Professional, Religious, Support_Group. Once you have these attributes selected, click OK to return to your main process.

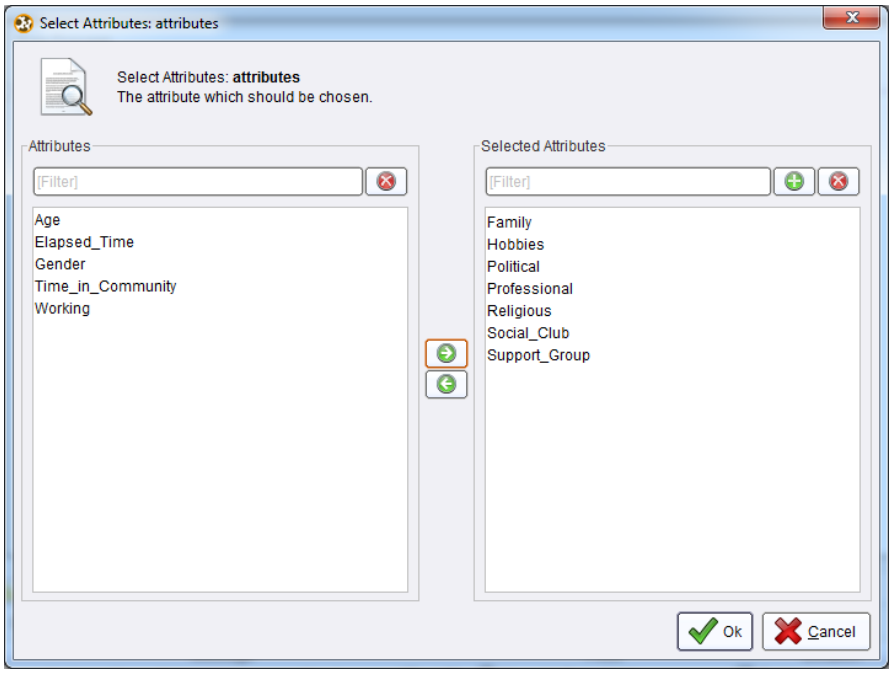


Figure 5-3. Selection of attributes to include in the association rules model.

5) One other step is needed in our data preparation. This is to change the data types of our selected attributes from integer to binominal. As previously mentioned, the association rules operators need this data type in order to function properly. In the search box on the Operators tab in design view, type 'Numerical to' (without the single quotes) to locate the operators that will change attributes with a numeric data type to some other data type. The one we will use is Numerical to Binominal. Drag this operator into your stream.

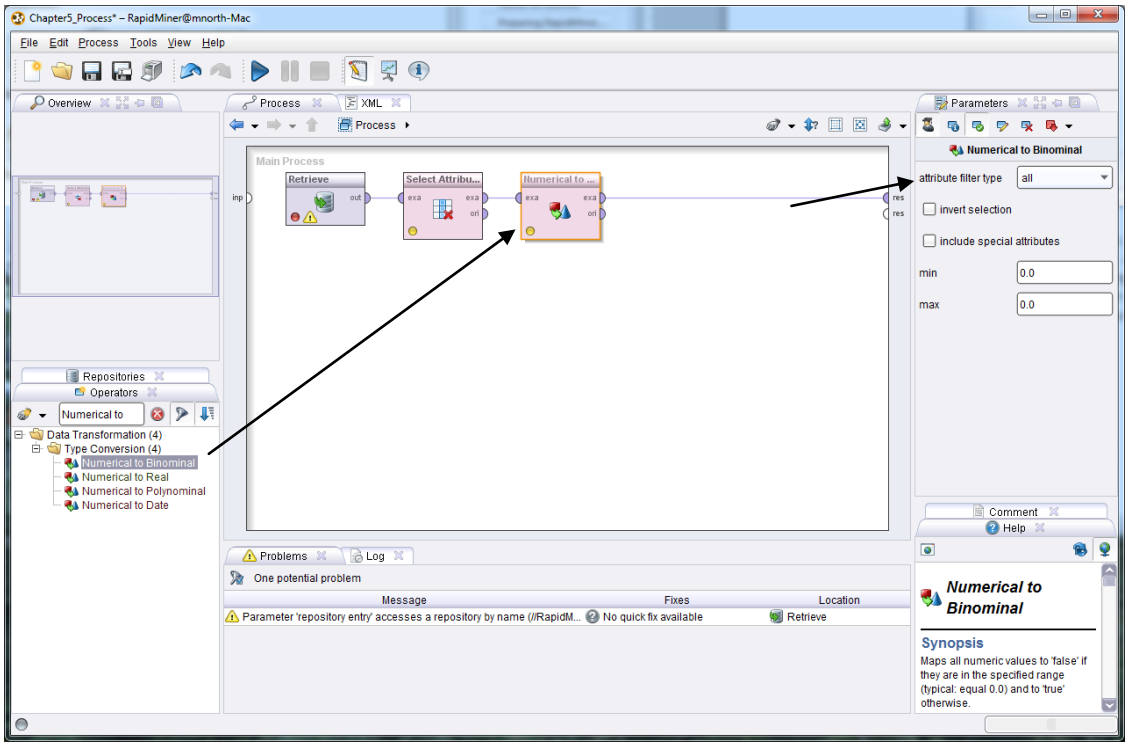
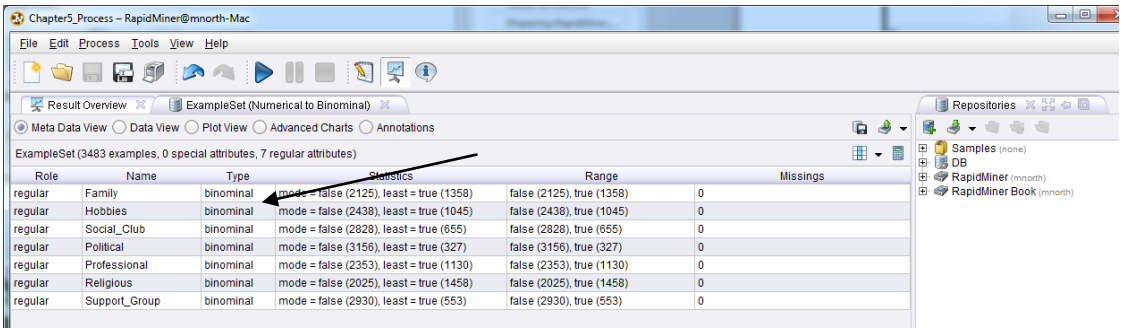


Figure 5-4. Adding a data type converstion operator to a data mining model.

6) For our purposes, all attributes which remain after application of the Select Attributes operator need to be converted from numeric to binominal, so as the black arrow indicates in Figure 5-4, we will convert 'all' from the former data type to the latter. We could convert a subset or a single attribute, by selecting one of those options in the attribute filter type dropdown menu. We have done this in the past, but in this example, we can accept the default and covert all attributes at once. You should also observe that within RapidMiner, the data type **binominal** is used instead of **binomial**, a term many data analysts are more used to. There is an important distinction. *Binomial* means one of two numbers (usually 0 and 1), so the basic underlying data type is still numeric. *Binominal* on the other hand, means one of two values which may be numeric *or* character based. Click the play button to run your model and see how this conversion has taken place in our data set. In results perspective, you should see the transformation, as depicted in Figure 5-5.



Role	Name	Type	Statistics	Range	Missings
regular	Family	binominal	mode = false (2125), least = true (1358)	false (2125), true (1358)	0
regular	Hobbies	binominal	mode = false (2438), least = true (1045)	false (2438), true (1045)	0
regular	Social_Club	binominal	mode = false (2828), least = true (655)	false (2828), true (655)	0
regular	Political	binominal	mode = false (3156), least = true (327)	false (3156), true (327)	0
regular	Professional	binominal	mode = false (2353), least = true (1130)	false (2353), true (1130)	0
regular	Religious	binominal	mode = false (2025), least = true (1458)	false (2025), true (1458)	0
regular	Support_Group	binominal	mode = false (2930), least = true (553)	false (2930), true (553)	0

Figure 5-5. The results of a data type transformation.

7) For each attribute in our data set, the values of 1 or 0 that existed in our source data set now are reflected as either 'true' or 'false'. Our data preparation phase is now complete and we are ready for…

MODELING

8) Switch back to design perspective. We will use two specific operators in order to generate our association rule data mining model. Understand that there are many other operators offered in RapidMiner that can be used in association rule models. At the outset, we established that this book is not a RapidMiner training manual and thus, will not cover every possible operator that could be used in a given model. Thus, please do not assume that this chapter's example is demonstrating the one and only way to mine for association rules. This is one of several possible approaches, and you are encouraged to explore other operators and their functionality.

To proceed with the example, use the search field in the operators tab to look for an operator called FP-Growth. Note that you might find one called W-FPGrowth. This is simply a slightly different implementation of the FP-Growth algorithm that will look for associations in our data, so do not be confused by the two very similar names. For this chapter's example, select the operator that is just called FP-Growth. Go ahead and drag it into your stream. The FP in FP-Growth stands for **Frequency Pattern**. Frequency pattern analysis is handy for many kinds of data mining, and is a necessary component of association rule mining. Without having frequencies of attribute combinations, we cannot determine whether any of the patterns in the data occur often enough to be considered rules. Your stream should now look like Figure 5-6.

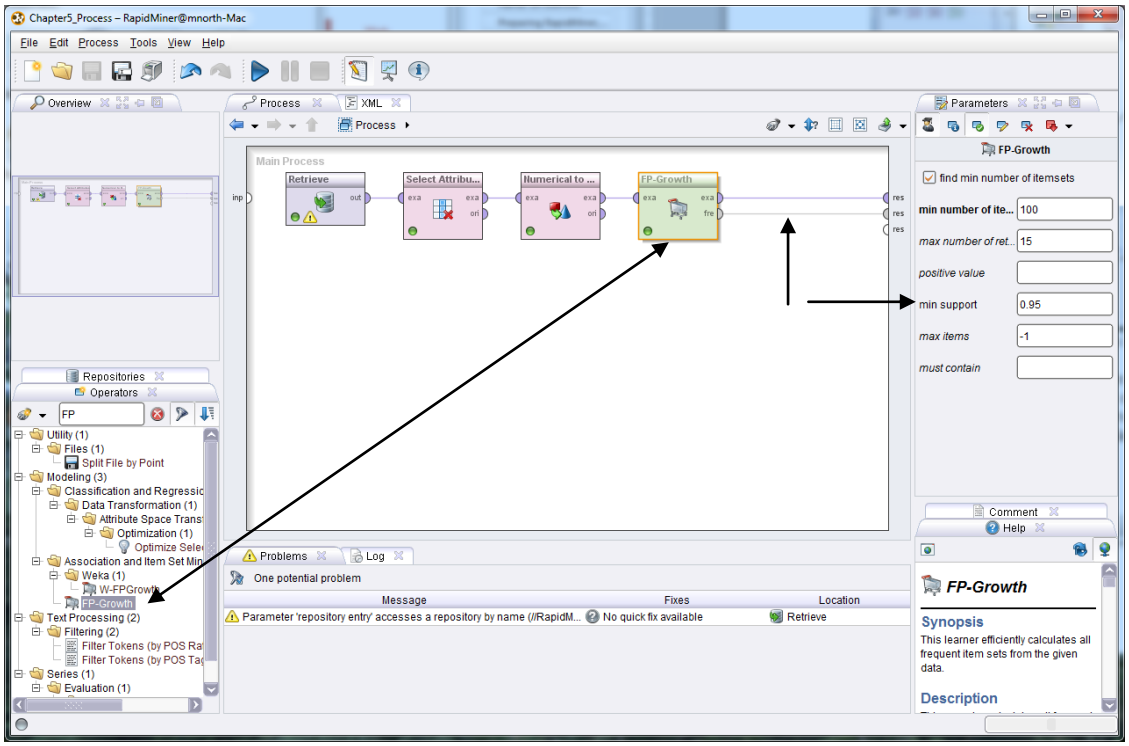


Figure 5-6. Addition of an FP-Growth operator to an association rule model.

9) Take note of the *min support* parameter on the right hand side. We will come back to this parameter during the evaluation portion of this chapter's example. Also, be sure that both your *exa* port and your *fre* port are connected to *res* ports. The *exa* port will generate a tab of your examples (your data set's observations and meta data), while the *fre* port will generate a matrix of any frequent patterns the operator might find in your data set. Run your model to switch to results perspective.

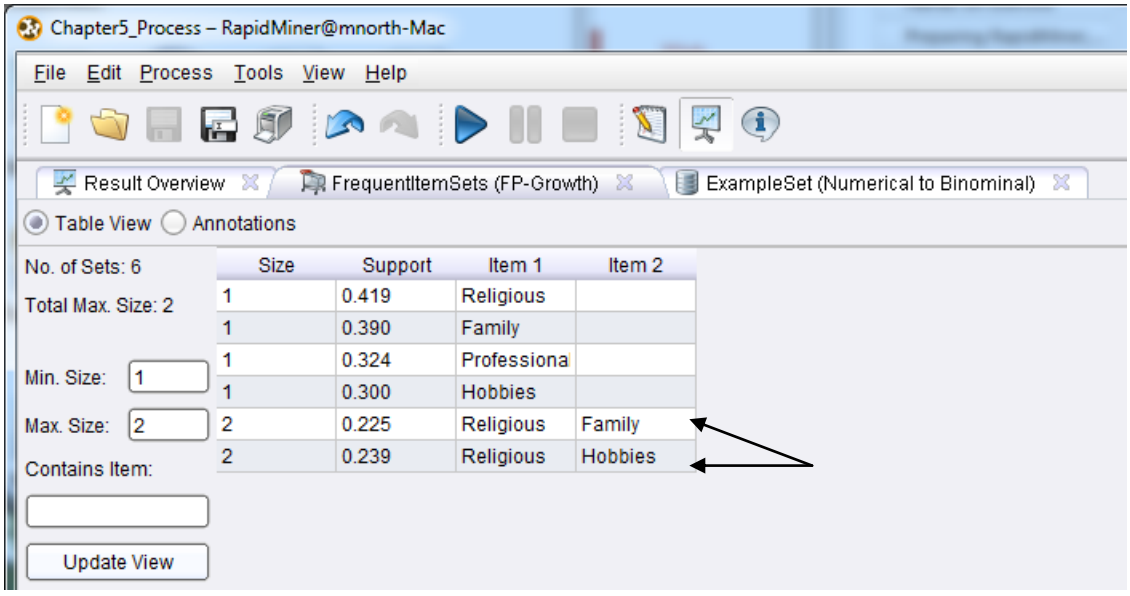


Figure 5-7. Results of an FP-Growth operator.

10) In results perspective, we see that some of our attributes appear to have some frequent patterns in them, and in fact, we begin to see that three attributes look like they might have some association with one another. The black arrows point to areas where it seems that Religious organizations might have some natural connections with Family and Hobby organizations. We can investigate this possible connection further by adding one final operator to our model. Return to design perspective, and in the operators search box, look for 'Create Association' (again, without the single quotes). Drag the Create Association Rules operator over and drop it into the spline that connects the *fre* port to the *res* port. This operator takes in frequent pattern matrix data and seeks out any patterns that occur so frequently that they could be considered rules. Your model should now look like Figure 5-8.

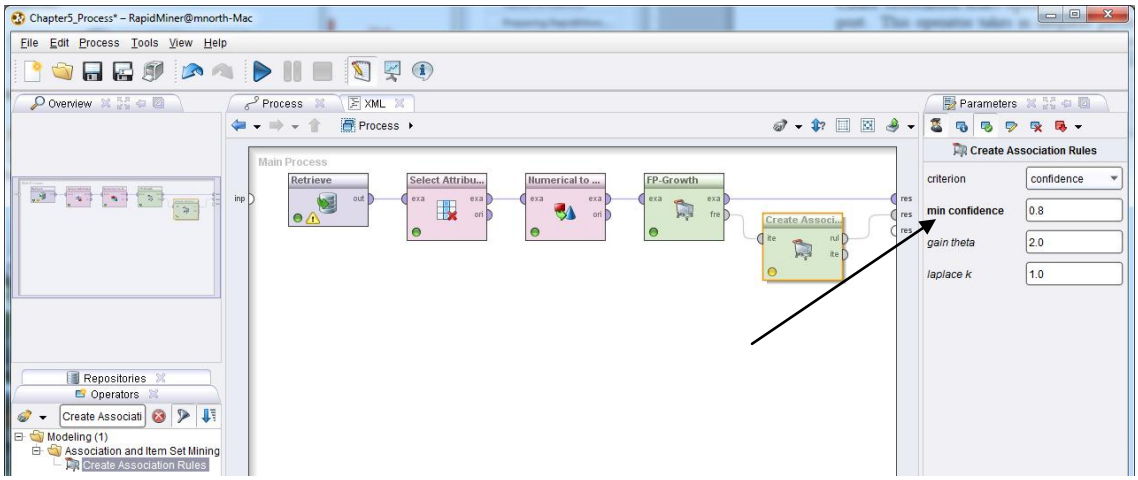


Figure 5-8. Addition of Create Association Rules operator.

11) The Create Association Rules operator can generate both a set of rules (through the *rul* port) and a set of associated items (through the *ite* port). We will simply generate rules, and for now, accept the default parameters for the Create Association Rules, though note the *min confidence* parameter, which we will address in the evaluation phase of our mining. Run your model.

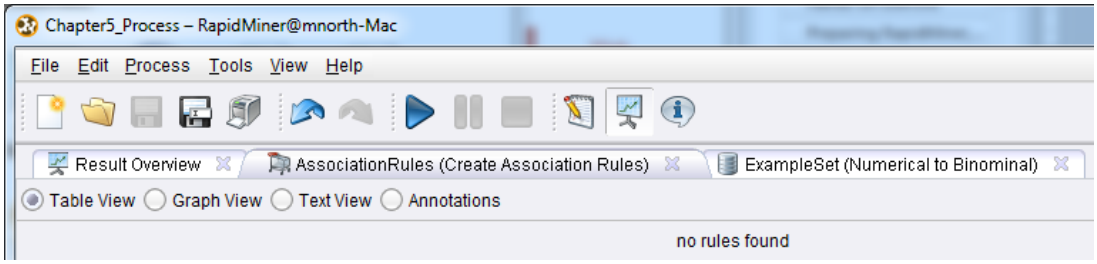


Figure 5-9. The results of our association rule model.

12) Bummer. No rules found. Did we do all that work for nothing? It seemed like we had some hope for some associations back in step 9, what happened? Remember from Chapter 1 that the CRISP-DM process is cyclical in nature, and sometimes, you have to go back and forth between steps before you will create a model that yields results. Such is the case here. We have nothing to consider here, so perhaps we need to tweak some of our model's parameters. This may be a process of trial and error, which will take us back and forth between our current CRISP-DM step of Modeling and…

EVALUATION

13) So we've evaluated our model's first run. No rules found. Not much to evaluate there, right? So let's switch back to design perspective, and take a look at those parameters we highlighted briefly in the previous steps. There are two main factors that dictate whether or not frequency patterns get translated into association rules: **Confidence percent** and **Support percent**. Confidence percent is a measure of how confident we are that when one attribute is flagged as true, the associated attribute will also be flagged as true. In the classic shopping basket analysis example, we could look at two items often associated with one another: cookies and milk. If we examined ten shopping baskets and found that cookies were purchased in four of them, and milk was purchased in seven, and that further, in three of the four instances where cookies were purchased, milk was also in those baskets, we would have a 75% confidence in the association rule: cookies → milk. This is calculated by dividing the three instances where cookies and milk coincided by the four instances where they *could have* coincided (3/4 = .75, or 75%). The rule cookies → milk had a chance to occur four times, but it only occurred three, so our confidence in this rule is not absolute.

Now consider the reciprocal of the rule: milk → cookies. Milk was found in seven of our ten hypothetical baskets, while cookies were found in four. We know that the coincidence, or frequency of connection between these two products is three. So our confidence in milk → cookies falls to only 43% (3/7 = .429, or 43%). Milk had a chance to be found with cookies seven times, but it was only found with them three times, so our confidence in milk → cookies is a good bit lower than our confidence in cookies → milk. If a person

comes to the store with the intention of buying cookies, we are more confident that they will also buy milk than if their intentions were reversed. This concept is referred to in association rule mining as **Premise → Conclusion**. Premises are sometimes also referred to as **antecedents**, while conclusions are sometimes referred to as **consequents**. For each pairing, the confidence percentages will differ based on which attribute is the premise and which the conclusion. When associations between three or more attributes are found, for example, cookies, crackers → milk, the confidence percentages are calculated based on the two attributes being found with the third. This can become complicated to do manually, so it is nice to have RapidMiner to find these combinations and run the calculations for us!

The support percent is an easier measure to calculate. This is simply the number of times that the rule *did* occur, divided by the number of observations in the data set. The number of items in the data set is the absolute number of times the association *could have* occurred, since every customer could have purchased cookies and milk together in their shopping basket. The fact is, they didn't, and such a phenomenon would be highly unlikely in any analysis. Possible, but unlikely. We know that in our hypothetical example, cookies and milk were found together in three out of ten shopping baskets, so our support percentage for this association is 30% (3/10 = .3, or 30%). There is no reciprocal for support percentages since this metric is simply the number of times the association did occur over the number of times it could have occurred in the data set.

So now that we understand these two pivotal parameters in association rule mining, let's make a parameter modification and see if we find any association rules in our data. You should be in design perspective again, but if not, switch back now. Click on your Create Association Rules operator and change the *min confidence* parameter to .5 (see Figure 5-10). This indicates to RapidMiner that any association with at least 50% confidence should be displayed as a rule. With this as the confidence percent threshold, if we were using the hypothetical shopping baskets discussed in the previous paragraphs to explain confidence and support, cookies → milk would return as a rule because its confidence percent was 75%, while milk → cookies would not, due to that association's 43% confidence percent. Let's run our model again with the .5 confidence value and see what we get.

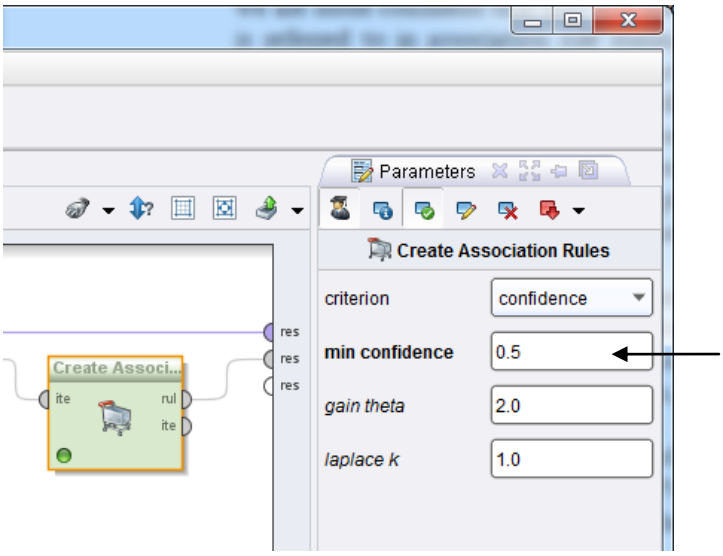


Figure 5-10. Chaning the confidence percent threshold.

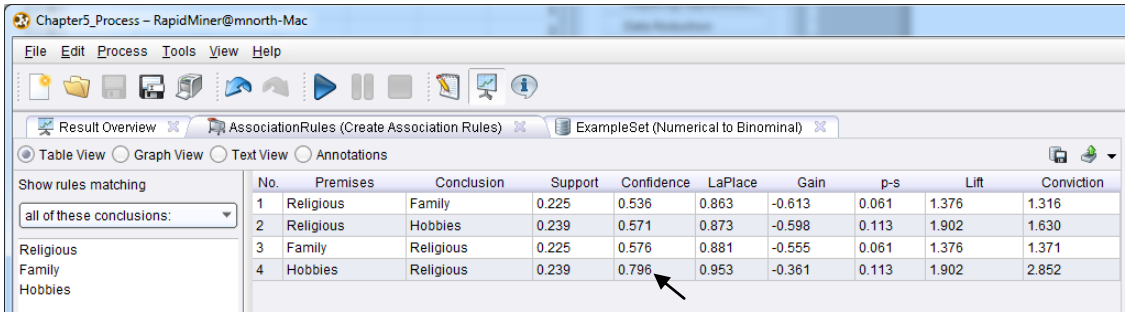


Figure 5-11. Four rules found with the 50% confidence threshold.

14) Eureka! We have found rules, and our hunch that Religious, Family and Hobby organizations are related was correct (remember Figure 5-7). Look at rule number four. It just barely missed being considered a rule with an 80% confidence threshold at 79.6%. Our other associations have lower confidence percentages, but are still quite good. We can see that for each of these four rules, more than 20% of the observations in our data set support them. Remember that since support is not reciprocal, the support percents for rules 1 and 3 are the same, as they are for rules 2 and 4. As the premises and conclusions were reversed, their confidence percentages did vary however. Had we set our confidence percent threshold at .55 (or 55% percent), rule 1 would drop out of our results, so Family → Religious would be a rule but Religious → Family would not. The other calculations to the right (LaPlace…Conviction) are additional arithmetic indicators of the strength of the rules' relationships. As you compare these values to support and confidence percents, you will see that they track fairly consistently with one another.

If you would like, you may return to design perspective and experiment. If you click on the FP-Growth operator, you can modify the *min support* value. Note that while support percent is the metric calculated and displayed by the Create Association Rules operator, the *min support* parameter in the FP-Growth actually calls for a confidence level. The default of .95 is very common in much data analysis, but you may want to lower it a bit and re-run your model to see what happens. Lowering *min support* to .5 does yield additional rules, including some with more than two attributes in the association rules. As you experiment you can see that a data miner might need to go back and forth a number of times between modeling and evaluating before moving on to…

DEPLOYMENT

We have been able to help Roger with his question. Do existing linkages between types of community groups exist? Yes, they do. We have found that the community's churches, family, and hobby organizations have some common members. It may be a bit surprising that the political and professional groups do not appear to be interconnected, but these groups may also be more specialized (e.g. a local chapter of the bar association) and thus may not have tremendous cross-organizational appeal or need. It seems that Roger will have the most luck finding groups that will collaborate on projects around town by engaging churches, hobbyists and family-related organizations. Using his contacts among local pastors and other clergy, he might ask for volunteers from their congregations to spearhead projects to clean up city parks used for youth sports (family organization association rule) or to improve a local biking trail (hobby organization association rule).

CHAPTER SUMMARY

This chapter's fictional scenario with Roger's desire to use community groups to improve his city has shown how association rule data mining can identify linkages in data that can have a practical application. In addition to learning about the process of creating association rule models in RapidMiner, we introduced a new operator that enabled us to change attributes' data types. We also used CRISP-DM's cyclical nature to understand that sometimes data mining involves some back and forth 'digging' before moving on to the next step. You learned how support and

confidence percentages are calculated and about the importance of these two metrics in identifying rules and determining their strength in a data set.

REVIEW QUESTIONS

1) What are association rules? What are they good for?

2) What are the two main metrics that are calculated in association rules and how are they calculated?

3) What data type must a data set's attributes be in order to use Frequent Pattern operators in RapidMiner?

4) How are rule results interpreted? In this chapter's example, what was our strongest rule? How do we know?

EXERCISE

In explaining support and confidence percentages in this chapter, the classic example of shopping basket analysis was used. For this exercise, you will do a shopping basket association rule analysis. Complete the following steps:

1) Using the Internet, locate a sample shopping basket data set. Search terms such as 'association rule data set' or 'shopping basket data set' will yield a number of downloadable examples. With a little effort, you will be able to find a suitable example.

2) If necessary, convert your data set to CSV format and import it into your RapidMiner repository. Give it a descriptive name and drag it into a new process window.

3) As necessary, conduct your Data Understanding and Data Preparation activities on your data set. Ensure that all of your variables have consistent data and that their data types are appropriate for the FP-Growth operator.

4) Generate association rules for your data set. Modify your confidence and support values in order to identify their most ideal levels such that you will have some interesting rules with reasonable confidence and support. Look at the other measures of rule strength such as LaPlace or Conviction.

5) Document your findings. What rules did you find? What attributes are most strongly associated with one another. Are there products that are frequently connected that surprise you? Why do you think this might be? How much did you have to test different support and confidence values before you found some association rules? Were any of your association rules good enough that you would base decisions on them? Why or why not?

Challenge Step!

6) Build a new association rule model using your same data set, but this time, use the W-FPGrowth operator. (Hints for using the W-FPGrowth operator: (1) This operator creates its own rules without help from other operators; and (2) This operator's support and confidence parameters are labeled U and C, respectively.

Exploration!

7) The Apriori algorithm is often used in data mining for associations. Search the RapidMiner Operators tree for Apriori operators and add them to your data set in a new process. Use the Help tab in RapidMiner's lower right hand corner to learn about these operators' parameters and functions (be sure you have the operator selected in your main process window in order to see its help content).