COC131 Data Mining - Association Analysis

Martin D. Sykora m.d.sykora@lboro.ac.uk
Tutorial 06, Friday 27th March 2009

Association analysis is concerned with discovering interesting correlations or other relationships between variables in large databases. We are interested into relationships between features themselves, rather than features and class as in the standard classification problem setting. Hence searching for association patterns is no different from classification except that instead of predicting just the class, we try to predict arbitrary attributes or attribute combinations.

1. Fire up Weka (Waikako Environment for Knowledge Analysis) software, launch the explorer window and select the "Preprocess" tab. Open the weather.nominal data-set ("weather.nominal.arff", this should be in the ./data/ directory of the Weka install).

2. Often we are in search of discovering association rules showing attribute-value conditions that occur frequently together in a given set of data, such as; *buys(X, "computer") & buys(X, "scanner")* $\Longrightarrow$ *buys (X, "printer") [support = 2%, confidence = 60%].* Where confidence and support are measures of rule interestingness. A support of 2% means that 2% of all transactions under analysis show that computer, scanner and printer are purchased together. A confidence of 60% means that 60% of the customers who purchased a computer and a scanner also bought a printer. We are interested into association rules that apply to a reasonably large number of instances and have a reasonably high accuaracy on the instances to which they apply.
   Weka has three build-in association rule learners. These are, "Apriori", "Predictive Apriori" and "Tertius", however they are not capable of handling numeric data. Therefore in this exericse we use weather data.

   (a) Select the "Associate" tab to get into the Association rule mining perspective of Weka. Under "Associator" select and run each of the following "Apriori", "Predictive Apriori" and "Tertius".

Briefly inspect the output produced by each Associator and try to interpret its meaning.

(b) In association rule mining the number of possible association rules can be very large even with tiny datasets, hence it is in our best intrest to reduce the count of rules found, to only the most interesting ones. This is usually achieved by setting minimum thresholds on support and confidence values. Still in the "Associate" view, select the "Apriori" algorithm again, click on the textbox next to the "Choose" button and try, in turn, different values for the following parameters "lowerBoundMinSupport" *(min threshold for support)*, "minMetric" (min threshold for confidence). As you change these parameter values what do you notice about the rules that are found by the associator? *Note that the parameter "numRules" limits the maximum number of rules that the associator looks for, you can try changing this value.*

(c) This time run the Apriory algorithm with the "outputItemSets" parameter set to true. You will notice that the algorithm now also outputs a list of "Generated sets of large itemsets:" at different levels. If you have the module's Data Mining book by Witten & Frank with you, then you can compare and contrast the Apriori associator's output with the association rules on pages 114-116 *(I will have a couple copies circulating in the lab during the session, just ask me for one)*. I also strongly recommend to read through chapter 4.5 in your own time, while playing with the weather data in Weka, this chapter gives a nice & easy introduction to association rules. Notice in particular how the item sets and association rules compare with Weka and tables 4.10-4.11 in the book.

(d) Compare the association rules output from Apriori and Tertius (you can do this by navigating through the already build associator models in the "Result list" on the right side of the screen). Make sure that the Apriory algorithm shows at least 20 rules. Think about how the association rules generated by the two different methods compare to each other?

Something to always remember with association rules, is that they should not be used for prediction directly, that is without further analysis or domain knowledge, as they do not necessarily indicate causality. They are however a very helpful starting point for further exploration and for building a better understanding of our data.

3. As you should certainly know by this point, in order to identify associa-

tions between parameters a correlation matrix and scatter plot matrix can be very usefull. In order to remind yourself of this it might be helpfull to look back to tutorials 2, 3 or 5.

4. Linear regression can be very usefull in association analysis of numerical values, in fact regression analysis is a powerfull approach to modelling the relationship between a dependent and independent variables. Simple regression is when we predict from one independent variable and multiple regression is when we predict from more than one independent variables. The model we attempt to fit is a linear one which is, very simply, drawing a line through the data. Of all the lines that can possibly be drawn through the data, we are looking for the one that best fits the data. In fact, we look to find a line that best satisfies

$$\gamma = \beta_0 + \beta_1 x + \epsilon$$

So a most accurate model is that which yields a best fit line to the data in question, we are looking for minimal sum of squared deviations between actual and fitted values, this is called method of least squares. So now that we have briefly reminded ourselves of the very basics of regression lets directly move onto an example in Weka.

   (a) In Weka go back to the "Preprocess" tab. Open the iris data-set ("iris.arff", this should be in the ./data/ directory of the Weka install).

   (b) In the "Attributes" section (bottom left of the screen) select the "class" feature and click "Remove". We need to do this, as simple linear regression cannot deal with non numeric values.

   (c) Next select the "Classify" tab to get into the Classification perspective of Weka, and choose "LinearRegression" (under "functions").

   (d) Clicking on the textbox next to the "Choose" button brings up the parameter editor window. Click on the "More" button to get information about the parameters. Make sure that "attributeSelectionMethod" is set to "No attribute selection" and "eliminateColinearAttributes" is set to "False".

   (e) Finally make sure that you select the parameter "petalwidth" in the dropdown box just under the "Test Options". Hit Start to run the regression. Inspect the results, in particular pay attention to the Linear Regression Model formula returned, and the coefficients

and intercept of the straight line equation. As this is a numeric prediction/regression problem, accuracy is measured with Root Mean Squared Error, Mean Absolute Error and the likes. As most of you will have clearly noticed, you can repeat this process for regressing the other features in turn, and compare how well the different features can be predicted.