Lab 8 – Association Analysis

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.

Therefore, we are in search of association rules showing attribute-value conditions that occur frequently together in a given set of data. For example,

buys(X, “computer”) & buys(X, “scanner”) ⇒ buys (X, “printer”)

with [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 accuracy on the instances to which they apply.

1. Fire up Weka (Waikako Environment for Knowledge Analysis) software, launch the explorer window and select the “Pre-process” tab. Open the weather.nominal dataset (“weather.nominal.arff”).

2. Weka has three built-in association rule learners. These are “Apriori”, “Predictive Apriori” and “Tertius”, however they are not capable of handling numeric data. Therefore, in this exercise we use weather data.

(a) Select the “Associate” tab to get into the Association rule mining perspective of Weka. Under “Associator” choose 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 interest to reduce the 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 116, 2nd Edition or 120-121 3rd Edition.

(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 about association rules is that they should not be used for prediction directly 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.

As you should certainly know by this point, in order to identify associations between parameters a correlation matrix and scatter plot matrix can be very useful. In order to remind yourself of this it might be helpful to look back to previous labs.