Practical Data Mining

COMP-321B

Tutorial 6: Association Rules

Gabi Schmidberger
Mark Hall

September 11, 2006

1 Introduction

This tutorial is about association rule learners. Association rule learners are started in the **Associate** panel, which is introduced in the next chapter. Only one association rules learner is used:

Apriori (weka.associations.Apriori).

1.1 Introduction to the data sets used

This tutorial on association rule learning uses the following three data sets (which can be found in /home/ml/321/Tutorial6/).

vote.arff This data set contains information about how each of the U.S. House of Representatives Congressmen voted on the 16 key notes. One instance represents the voting history of one congressmen and her/his party affiliation. (For classification the party affiliation was used as class attribute.)

weather.nominal.arff This dataset was already used in Tutorial 1. It is a very small data set with only nominal attributes.

supermarket.arff This data set describes the shopping habits of supermarket customers. Most of the attributes stand for one particular item group. The value is 't' if the customer had bought an item out of a item range and missing otherwise. There is one instance per customer. The data set contains no class attribute, as this is not required for learning association rules.

2 The Associate Panel

The Associate Panel looks very similar to the Classify Panel. It is basically the same just the Test options box and the class selection field are missing. Both are not relevant for association rules. Association rules don't generally give one attribute a special position as classification does with the class attribute. The testing panel is not needed since association rule learning is mostly seen as a mainly exploratory data mining task. This means there is no strong emphasis on precise evaluation.

2.1 The output of the association rules learner Apriori

Load the data set 'vote.arff' and change into the Associate Panel. Select 'Apriori' as associator. After pressing start Apriori starts to build its model and writes its output into the output field. The first part of the output ('Run information') describes the option that have been set and the data set used.

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     vote
Instances:    435
Attributes:   17
              handicapped-infants
              water-project-cost-sharing
              adoption-of-the-budget-resolution
              physician-fee-freeze
              el-salvador-aid
              religious-groups-in-schools
              anti-satellite-test-ban
              aid-to-nicaraguan-contras
              mx-missile
              immigration
              synfuels-corporation-cutback
              education-spending
              superfund-right-to-sue
              crime
              duty-free-exports
              export-administration-act-south-africa
              Class
```

The next part of the output is information about how the rules were generated. Since you have not learned how the rules are build we want to ignore that part. But it should look like this:

```
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.45 (196 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 11

Generated sets of large itemsets:

Size of set of large itemsets L(1): 20

Size of set of large itemsets L(2): 17

Size of set of large itemsets L(3): 6

Size of set of large itemsets L(4): 1
```

The rules that have been generated are listed at the end of the output. Mostly not all rules are shown. By default the 10 most valuable one's according to their confidence level are shown. Each rule consists of some attribute values on a left hand side of the arrow, the arrow sign and the right hand side list of attribute values. Right of the arrow sign are the predicted attribute values.

Taking rule 1 as example, the rule means if the value for 'adoption-of-the-budget-resolution' is 'y' and the value for 'physician-fee-freeze' is 'n' then the value for the attribute 'class' should be predicted as 'democrat' (Note that NOT in all rules the predicted value is a value of the class attribute.) This prediction has a certain support and confidence value.

The number before the arrow sign is the number of instances the rule applies to. The number after the arrow sign is the number of instances predicted correctly. The number in brackets after 'conf:' is the confidence of the rule.

3 Using Apriori

If not yet loaded, load the data set 'vote.arff'.

Task A1: Run Apriori using the default settings of the options. The last 10 lines of the output below 'Best rules found' are the 10 best rules generated. The confidence of rule 10 is 0.96. How was this confidence value computed? Write down the proportion as division.

Task A2: How many instances is the support of rule 8?

Task A3: What does it mean 'rule applies to a certain number of instances'? Explain using rule number 7 as example. (Hint: You can check the numbers on the Preprocess panel.)

Task A4: What does it mean 'number of instances predicted correctly'? Explain using rule number 9 as example.

Task A5: Study the description of the parameters for Apriori by pressing the 'More' button in the window that allows you to change the options for 'Apriori'. Try to change the number of rules listed in the output. Do you think the number generated rules can be more than 100. If yes, why is that?

Task A6: What does 'best rules' mean? What criterion is used to decide what the best rules are?

Task A7: Which rule tells you how likely it is that if a congressman has not voted for aid to El Salvador, that he has also voted for aid to Nicaraguan contras?

Task A8: The 10 best rules contain rules that have 'Class=democrat' in their right hand side. Does this say something about the voting habits of the democratic congressmen?

4 The weather data set again

Load the data set 'weather.nominal.arff'. This is a small data set which we have already used in the first tutorial.

Task B1: Consider the rule:

```
temperature=hot ==> humidity=normal.
```

What is the support of this rule? How many instances apply to this rule and what is the confidence value? (Open the 'Viewer' window in the Preprocess panel to answer this question.)

Task B2: Consider the rule:

```
temperature=hot humidity=high ==> windy=TRUE.
```

What is the support of this rule? How many instances apply to this rule and what is the confidence value? Further write down the instance numbers of the instances that support the rule and the number of the instances that apply to this rule.

Task B3: Can a rule have tests on two (or more) attributes on its right hand side like in the example below:

```
outlook=sunny temperature=cool
==> humidity=normal play=yes
```

5 Make association rules for the supermarket data set

Load the data set 'supermarket.arff'.

Association rules are primarily aimed to support exploratory data analysis. Use Apriori to generate rules and use them to say something about the shopping habits of supermarket customers. Generate about 30 rules.

It could also be interesting to generate rules with one particular attribute on the right hand side. These can be generated by setting the first option to 'true' and the second option to the attribute index value (**Attribute indexes for this option start with** 0 **not with** 1) you want to see at the right hand side of the rules.

Task C1: Study some generated rules and describe one observation YOU think to have made about supermarket customer purchasing habits. Also note down the relevant rules for this observation.

Task C2: Describe a second observation YOU think to have made about supermarket customer purchasing habits. Also note down the relevant rules for this observation.

Task C3: Do the observations you made in Tasks C1 and C2 suggest any courses of action for the supermarket manager? If so, what might they be?

6 Answers

Answer A1: Confidence of rule 10, conf: 0.96 = ... divided by .....

Answer A2: Support for rule 8 is .....

Answer A3:

Answer A4:

Answer A5:

Answer A6:

Answer A7:

Answer A8:

Answer B1:

```
temperature=hot ==> humidity=normal.
```

Support ..... No. instances applying to this rule .....

Confidence value .....

Answer B2:

```
 temperature=hot humidity=high ==> windy=TRUE.
```

Support ..... No. instances applying to this rule .....

Confidence value .....

List of the indexes that support .....

List of the indexes that apply to .....

Answer B3:

Answer C1:

Answer C2:

Answer C3: