**Data Mining for Business Dr. Kamesam**

**SPSS Workshop ---Association**

Prof. Kamesam (pkamesam@fordham.edu)

Graduate Assistants: Maxwell Li

**Objectives**

• View the resulting rules by browsing the model;

• Explain the meaning of rule confidence, support, rule support and lift;

• Sort the rules based on different criteria; and

• Create a rule set and use this to identify those records whose conditions are related to a selected conclusion.

**Data**

In this lesson we perform a market basket analysis of a dataset containing shopping information, ***Shopping.txt*** from **Blackboard.** The file contains fields that indicate whether or not a customer, during a single visit, purchased a particular product. Thus each record represents a store visit in which at least one product was purchased. The file also contains basic demographics, such as gender and age group.

**Introduction**

When people buy cigarettes do they tend to buy chocolate or beer? If people have high cholesterol do they also tend to have high blood pressure? If people buy car insurance do they also buy house insurance?

Answers to such questions can form the basis of brand positioning, advertising and even direct marketing. But how do we find whether associations such as these exist, and how can we begin to search for them when our databases have tens of thousands of records and many fields?

Association detection algorithms provide rules describing which values of fields typically occur together. They can therefore be used as an approach to this area of data understanding.

When PASW Modeler produces a set of association rules it also gives measures indicating the frequency and strength of the association for each rule. These measures are referred to as rule support and rule confidence and are given in the format:



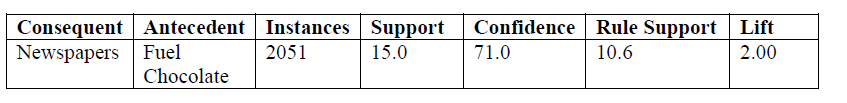
**"Instances"** is the number of records in the dataset that match the antecedents.

**Support** is the percentage of records that match the antecedents.

**Confidence** is the percentage of all records matching the antecedents that also match the consequent.

**Rule support** is the percentage of records that match the entire rule (both the antecedents and consequent).

**Lift** is the expected return using a model or rule. In this context it is the ratio of the rule confidence to the overall percentage occurrence of the consequent in the data.

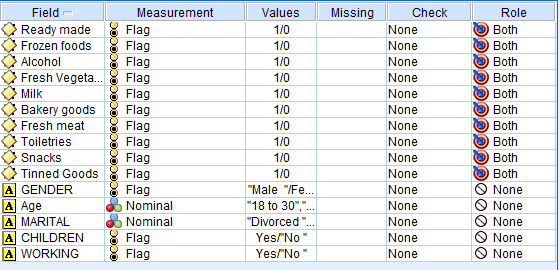
Therefore the full format of a rule will appear as: 

In this example, 15% of the customers (2051 individuals) bought fuel and chocolate. Of these, 71% also bought newspapers. 10.6% of the customers bought fuel, chocolate, and newspapers. The lift value of 2.00 indicates that those who purchase fuel and chocolate are twice as likely to buy newspapers as is the overall sample (71.0% versus 35.5%—not shown in the table).

**Steps:**

1. Click **File…New Stream** → Place a **Var.File** Node → Import data of ***Shopping.txt***

2. Place a Type node right to the shopping.txt and connect → Double-click **Type node** to define the measurement of fields → change the first 10 fields (from Ready made to Tinned Goods) to **Flag** → Change the **Role** setting for all **product fields (Ready made** to **Tinned goods)** to **Both** (select all product fields, right-click and choose **Set Role…Both** from the context menu) →Change the **Role** setting for the **demographic fields** to **None** (select the fields, right-click and choose **Set Role…None** from the context menu) (shown below)

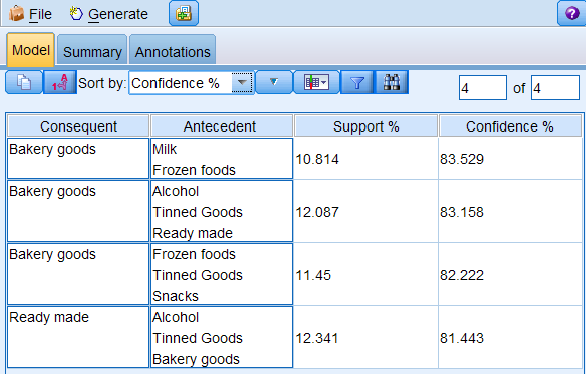


**Figure 1**

3. Click Ok to close Type Node → Place a Table Node above the Type node and connect → Double-click the Table node to get a rough understanding of the data.

4. Click **OK** to close the **Table** node → Place the **Apriori** node from the Modeling palette to the right of the **Type** node → Connect the **Type** node to the **Apriori** node →Double-click the **Apriori** node named **10 fields**

5.Click **Run** → An Apriori generated model node will appear in the Models palette → Double-click this node to read the results

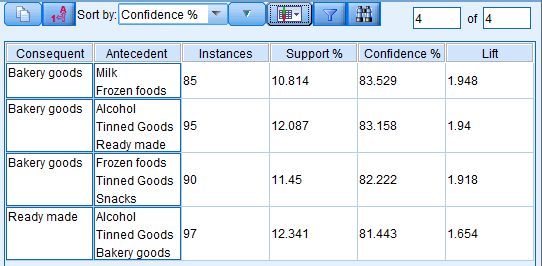
**Interpreting the results **

**Figure 2**

6.The Apriori algorithm has found only four association rules. By default, in addition to the consequent and antecedent, only the *Support %* and *Confidence %* are displayed. The rules are initially sorted by *Confidence %*. The first rule tells us that on 10.81% of the records (shopping visits), milk and frozen foods were purchased. Of this group, 83.53% also bought bakery goods. (as shown above)

7. To see the other available measures:

Click the **Show/hide** criteria menu button →**Click Instances** →Click the **Show/hide** criteria menu button again →**Click Lift**

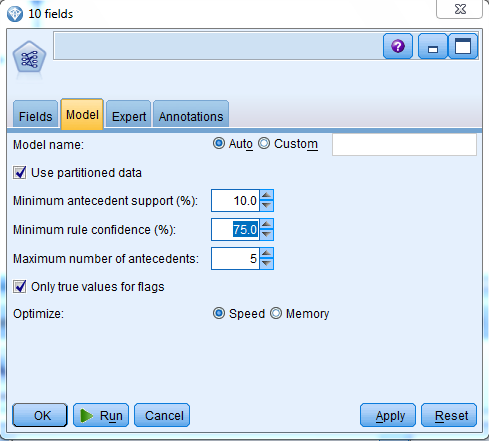
****

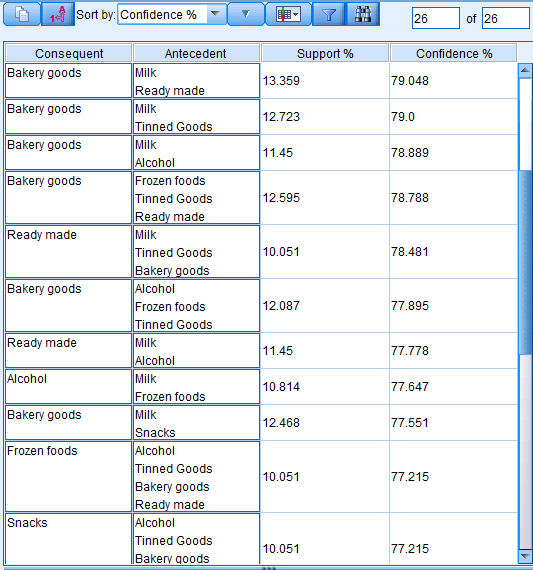
**Figure 3**

We now see that milk and frozen foods were purchased on 85 shopping trips. There are 786 records in the file, so this is (85/786)\*100, or 10.81% of the total number of shopping trips (*Support %*). (as shown above)

8. We will now investigate whether, by dropping the confidence of the rules to 75%, we can obtain a larger number of associations.

Click **File…Close** to close the Apriori Association Rules Browser window →Double-click the **Apriori** modeling node → Enter **75** in the **Minimum rule confidence:** text box (type or use the spin control) →Click **Run** → Double-click the **Apriori node** (As shown below)



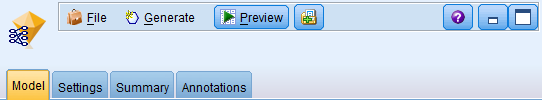


**Figure 4**

This is a far richer set of associations, and 26 rules are produced in all. The challenge is now to examine the rules for those that might be useful in the context of your business question or goal.

**Note: We will take Alcohol as an example to show following steps so make sure you have finished Step 8. (The associations about Alcohol has been included)**

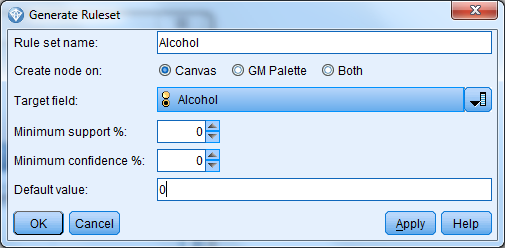
9. A rule set can be created for a selected consequent by **selecting the Rule Set option** under the Generate menu of the Association Rules browser window. This will generate a new model which, when placed in a data stream, will create a field indicating whether a rule in the rule set applies to the record, along with its confidence value. Note that more than a single rule may apply to a record and, by default, the confidence value is based on the first rule whose conditions match the record. We will create a rule set for the Alcohol field.



**Figure 5**

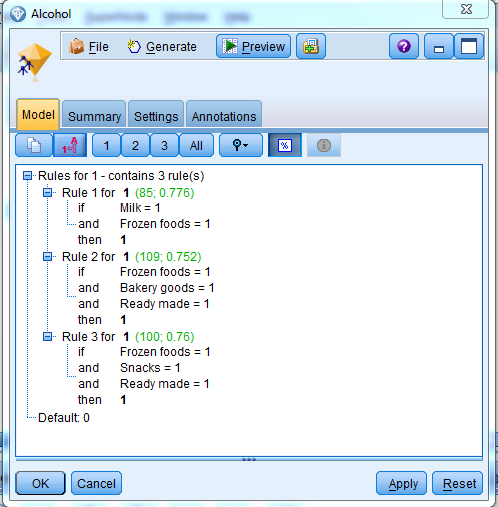
Click **Generate…Rule Set** Type **Alcohol** in the **Rule set name** text box → Select **Alcohol** in the **Target Field** list →Type **0** in the **Default value**: text box →Click **OK** → Click **File…Close** to close the Rule browser window (as shown below)

**\*\*Default value.** Allows you to specify a default value for the target field that is assigned to scored records for which no rule fires. For example, if the default value is 0, the records for which rules applied will be marked as 1, records for which rules didn’t apply will be marked as 0. We will show this in table later.



**Figure 6**

10. A generated Apriori Rule Set node named *Alcohol* will appear in the **upper left corner** of the Stream Canvas. →Double-click the generated **Apriori Rule Set** node named **Alcohol** →Click the **All** button →Click the **Show or hide instance and confidence figures** button  (As shown below)

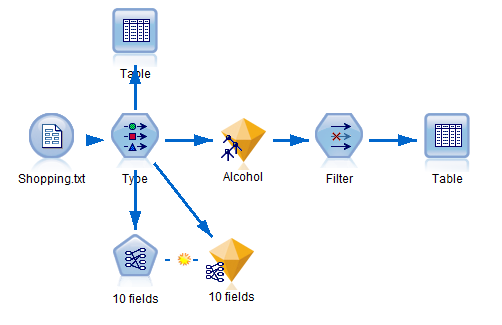


**Figure 7**

In this example the rule set contains three rules whose consequent is buying Alcohol. The first associates the purchase of milk and frozen foods. The second associates frozen foods, bakery goods and ready-made meals. The third associates frozen foods, snacks and ready-made meals.

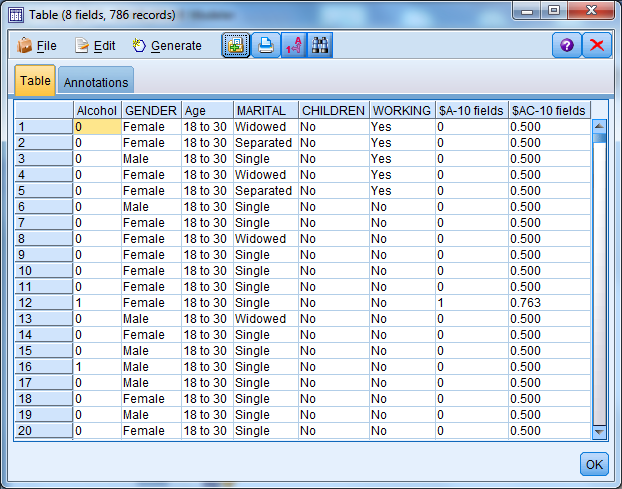
11. We can now pass the data through this node and generate a field indicating whether or not a record conforms to the conditions of any of the three rules.

Click **OK** to close the Rule Set window →Drag the **Apriori Rule Set** node named **Alcohol** to the right of the **Type** node →and Connect them→Place a **Filter** node from the Field Ops palette to the right of the **Apriori Rule Set** → Connect them→ Edit the **Filter** node and **select** the fields **Alcohol and demographic fields** (GENDER, Age, MARITAL…etc.), and then click **OK** →Place a **Table** node from the Output palette to the right of the **Filter** node →Connect the **Filter** node to the new **Table** node. (As shown below)



**Figure 8**

11. Right-click the new **Table** node, then click **Run**



**Figure 9**

Two new fields appear in the data table. The first, *$A-Alcohol*, is 0 unless one of the three rules in the rule set applies to the record, in which case it has a value of 1. The second field, *$AC-Alcohol*, represents the confidence figure for the rules decision. Notice that when the conditions of the rules do not apply to a record, its confidence value is 0.5.

We see that Rule 3 applied for record 12, since the confidence value is .76 (see the illustration under **step 10** The use of a rule set allows you to identify which customers are associated with a particular rule. From here various types of other analysis can be done (e.g., are those customers who match this rule more likely to be male or female, younger or older, etc.)