

Magnum Opus 4.6.3

Tutorial Introduction

Table of contents

| | |
|-----------------------------------------------------------|----|
| 1. Preliminaries | 2 |
| 2. A simple example | 2 |
| 3. A worked example | 3 |
| 4. Attribute-value data | 7 |
| 5. A worked example using attribute-value data | 9 |
| 6. Searching by strength..... | 14 |
| 7. Searching by lift..... | 19 |
| 8. Selecting RHS elements..... | 22 |
| 9. Itemsets | 25 |
| 10. Contrast discovery | 27 |
| 11. Statistically sound association discovery..... | 31 |
| A worked example of holdout evaluation for rules | 32 |
| A worked example of holdout evaluation for itemsets | 38 |
| 12. Computation time, snapshots and anytime results..... | 41 |
| 13. Some final thoughts | 43 |

1. Preliminaries

Magnum Opus detects associations within data.

The data is imported into the system from a text file. Users typically extract data from a database into a text file for use with the system. There is considerable flexibility in the text file formats that may be employed.

The user selects settings that control a search for associations in the data. The user can choose the type of association to be found and between alternative measures of the relative value of an association. The user also specifies the maximum number of associations to be found and any further restrictions on the associations to be considered.

Within the restrictions specified by the user, **Magnum Opus** finds the associations with the highest values on the specified measure. **Magnum Opus** will only find fewer than the specified number of associations if the search is terminated by the user or there are fewer than the specified number that satisfy the user specified constraints.

The associations found are recorded in an output file and may optionally be exported to a comma separated value file suitable for input into a spreadsheet for further analysis.

2. A simple example

We start with a simple invented example of analyzing the purchasing habits of a customer of a fictitious grocery store. The customer has visited the store on ten occasions, each time buying a different selection of goods. The following item-list file records the customer's purchasing behavior. Each line represents the items bought on a single visit.

```
plums, lettuce, tomatoes
celery, confectionery
apples, carrots, tomatoes, potatoes
potatoes
confectionery
carrots
apples, oranges, lettuce, tomatoes
peaches, oranges, celery, potatoes, confectionery
oranges, lettuce, carrots, tomatoes
apples, bananas, plums, carrots, tomatoes, onions
```

These can be processed by **Magnum Opus** to find rules such as the following four.

```
apples -> tomatoes [Coverage=0.300 (3); Support=0.300 (3);
Strength=1.000; Lift=2.00; Leverage=0.1500 (1.5)]

lettuce -> tomatoes [Coverage=0.300 (3); Support=0.300 (3);
Strength=1.000; Lift=2.00; Leverage=0.1500 (1.5)]

tomatoes -> apples [Coverage=0.500 (5); Support=0.300 (3);
Strength=0.600; Lift=2.00; Leverage=0.1500 (1.5)]

tomatoes & oranges -> lettuce [Coverage=0.200 (2); Support=0.200
(2); Strength=1.000; Lift=3.33; Leverage=0.1400 (1.4)]
```

Each rule presents a list of items to the left of the arrow that are associated with the single item to the right of the arrow. Then a number of statistics are presented that describe the nature of the association. For example, the first two of these rules indicate that whenever either apples or lettuce are purchased, tomatoes are also purchased. The third and fourth rules indicate that both apples and lettuce are more likely to be purchased if tomatoes are purchased. The final rule shows that whenever both tomatoes and oranges are purchased, lettuce is also purchased. We will explore exactly how the statistics that are displayed reveal this information as we proceed.

This is a very simplistic example. In practice it would be foolish to draw strong conclusions from such limited data. Indeed, **Magnum Opus** includes facilities for assessing the strength of evidence in support of a rule, and these mechanisms would reject all the above rules as having insufficient support. This example is intended to illustrate the type of analysis that **Magnum Opus** performs, albeit, normally on much larger volumes of more complex data.

3. A worked example

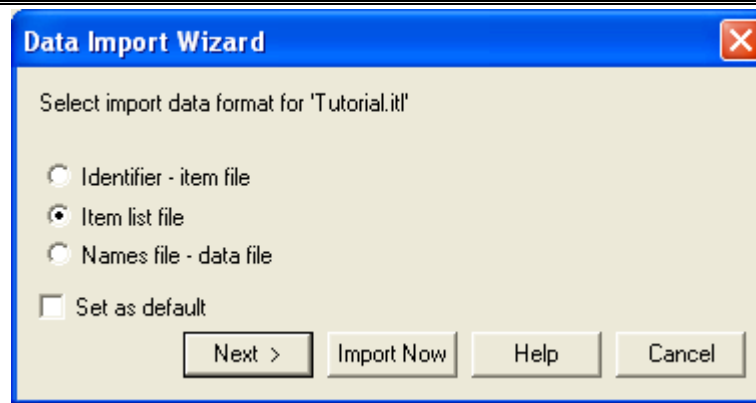
We now provide a fully worked example of an extended variant of the above scenario. The data is now extended to include all customers of the store for a given period of time, resulting in a total of 1000 transactions. The data is contained in the example file distributed with **Magnum Opus** called `tutorial.itl`.

Note, there are two versions of **Magnum Opus**. The command line version runs on Linux systems. The interactive version runs under Windows. In the following and all subsequent examples we provide both a command line for executing the example on the command line system and a step-through of the process for running it on the interactive system. We present the output from the interactive system. This may vary in minor respects from the output of the command line system.

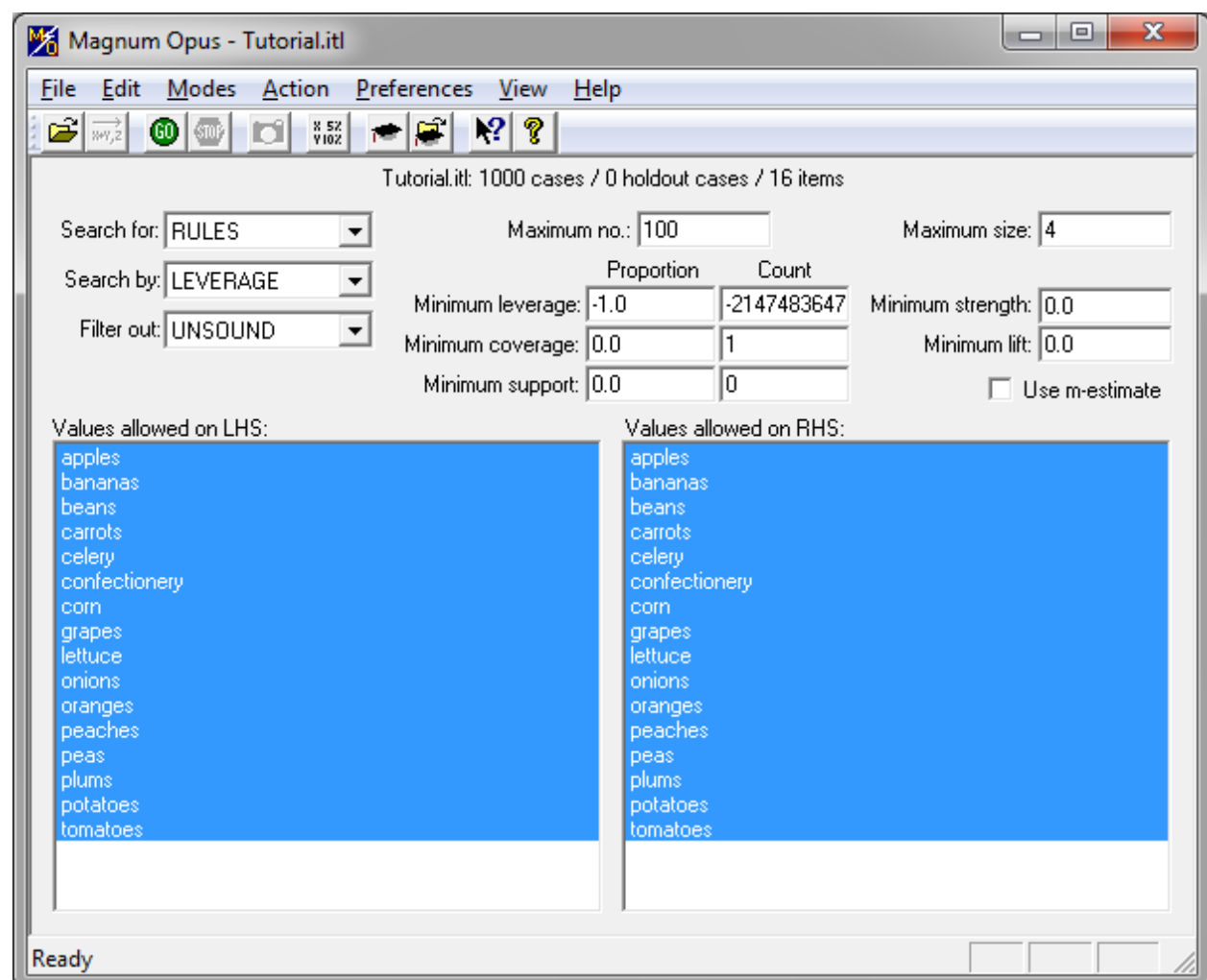
In the first example we run **Magnum Opus** with its default settings, except that we limit the number of rules produced to five only.

| |
|--------------------------------------------------------------------------------------|
| Command line: <code>mocl item-list-file=tutorial.itl maximum-results=5</code> |
|--------------------------------------------------------------------------------------|

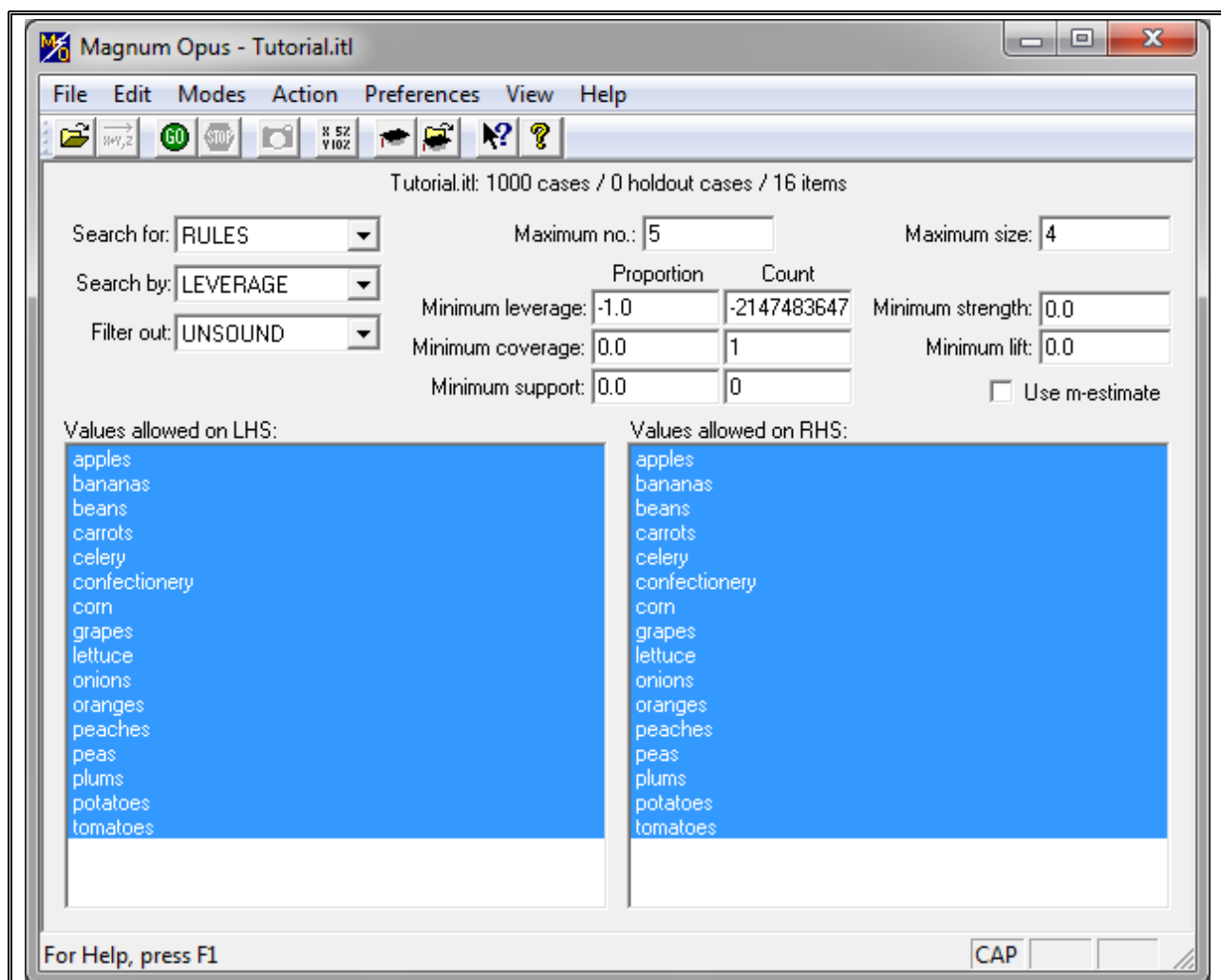
| |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Interactive system. <i>First run Magnum Opus. From the File Menu select Import Data. The system will display a dialog for selecting a file to open. If necessary, navigate to the Example Files folder within the folder into which you installed the software. Select the file tutorial.itl. The system will now display the following dialog box.</i> |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|



The system recognizes from the itl file extension that the file is probably an item list file. As this is correct and we wish to use the default data import settings, click the Import Now button. After importing the data the screen should appear as follows.



As we want to limit the number of rules to five, edit the Maximum no. edit box accordingly.



Now click the GO button to commence a search with the selected settings. A dialog will be displayed that allows you to select the file into which the results will be stored. Specify a file name and navigate to the folder in which you want the file stored. Then click on the Save button. The system will perform the search, saving the results in the specified file and then open the file for inspection.

Output:

Magnum Opus - The leader in association discovery technology.
Version 4.6.3
Copyright (c) 1999-2010 G. I. Webb & Associates Pty Ltd.

Data file: Tutorial.itl

1000 cases / 0 holdout cases / 16 items

Sun Aug 08 18:55:56 2010

Search for rules
Search by leverage
Filter out rules that are unsound.

```
Maximum number of attributes on LHS = 4
Maximum number of rules = 5
Minimum leverage = -1.0
Minimum leverage count = -2147483647
Minimum coverage = 0.0
Minimum coverage count = 1
Minimum support = 0.0
Minimum support count = 0
Minimum lift = 0.0
Minimum strength = 0.0

All values allowed on LHS

All values allowed on RHS

Search space for LHS size 1 = 120, adjusted critical value =
0.000104167
Search space for LHS size 2 = 1680, adjusted critical value =
7.44048E-006
Search space for LHS size 3 = 7280, adjusted critical value =
1.71703E-006
Search space for LHS size 4 = 21840, adjusted critical value =
5.72344E-007

Found 5 rules

tomatoes -> lettuce
[Coverage=0.263 (263); Support=0.111 (111); Strength=0.422; Lift=1.94;
Leverage=0.0539 (53.9); p=2.35E-019]

lettuce -> tomatoes
[Coverage=0.217 (217); Support=0.111 (111); Strength=0.512; Lift=1.94;
Leverage=0.0539 (53.9); p=2.35E-019]

tomatoes -> carrots
[Coverage=0.263 (263); Support=0.085 (85); Strength=0.323; Lift=1.85;
Leverage=0.0390 (39.0); p=1.83E-012]

carrots -> tomatoes
[Coverage=0.175 (175); Support=0.085 (85); Strength=0.486; Lift=1.85;
Leverage=0.0390 (39.0); p=1.83E-012]

onions -> potatoes
[Coverage=0.189 (189); Support=0.082 (82); Strength=0.434; Lift=1.53;
Leverage=0.0285 (28.5); p=5.30E-007]
```

The output file begins with a record of the settings used to produce the rules. It then states the number of rules found, followed by each of those rules. Each rule is composed of two parts. The left-hand-side (LHS) appears before the arrow and the right-hand-side (RHS) appears after the arrow. Then a number of statistics are presented that describe the relationship between the LHS and RHS.

The first rule describes an association between tomatoes and lettuce. The following measures are presented that describe the association.

| | |
|-----------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Coverage | The <i>coverage</i> of the rule is the number of cases that contain the LHS. In this data 263 cases contain tomatoes, which is 0.263 of the 1000 cases in the data. |
| Support | The <i>support</i> of the rule is the number of cases that contain both the LHS and the RHS. In this data there are 111 cases that contain both tomatoes and lettuce which represents 0.111 of the total data. |
| Strength | The <i>strength</i> is the support divided by the coverage. This represents the proportion of the cases that contain the LHS that also contain the RHS. It can be thought of as an estimate of the probability that the RHS will occur in a case if the LHS occurs. |
| Lift | The <i>lift</i> is the strength divided by the strength that would be expected if there were no relationship between the LHS and the RHS. A value of 1.0 suggests that there is no relationship between the two. Higher values suggest stronger positive relationships. Lower values suggest stronger negative relationships (the presence of the LHS reduces the likelihood of the RHS). |
| Leverage | The <i>leverage</i> is the support minus the support that would be expected if the LHS and RHS were unrelated to one another. A positive value suggests a positive relationship and a negative value suggests a negative relationship. |
| <i>p</i> | The result of a statistical evaluation of the significance of the rule. The lower this value the less likely that this rule is spurious, either because the LHS and RHS are unrelated to one another, or because one or more of the values in the LHS do not contribute to the association with the RHS. |

Congratulations! You now know enough to undertake standard association analysis with transaction data. If you are using transaction data, you might want to stop here and start analyzing your data. You might return to the tutorial after you have gained a little more familiarity with the system and are ready to master some more advanced techniques.

4. Attribute-value data

So far we have considered only data in the form of lists of items. Many data are recorded in tabular format, with columns representing *attributes* or *fields* and each row representing a distinct entity. The cells contain the values of the respective attributes or fields for the given entity. **Magnum Opus** supports such data, which must be listed in a *data file*. The columns are separated by a delimiter character such as a TAB or COMMA.

It is also necessary to specify the names and types of the attributes. This information is provided in a separate file called the *names file*. Each line of a names file starts with the name of an attribute, the first line referring to the leftmost column, the second line to the second leftmost column, and so on.

For categorical attributes, the attribute name is followed by a colon (:) and then either the keyword `categorical` or a comma separated list of the values that are allowed for the attribute.

Example:

```
Department: bakery, dairy, beverages
```

This specifies that the attribute Department can assume any one of three values `bakery`, `dairy`, or `beverages`. Any case containing any other value will be discarded and an error message generated.

Example:

```
Department: categorical
```

This specifies that the attribute Department can assume any value that appears in the data file.

For compatibility with See-5, Magnum Opus also accepts the keyword `discrete` which is treated as equivalent to `categorical`.

Numeric attributes must be divided into sub-ranges. These can be specified in the names file. Alternatively, the names file can simply identify the number of sub-ranges and Magnum Opus will select the sub-ranges for you.

For a numeric attribute with specified sub-ranges, the attribute name is followed by a list of sub-range cut points. These indicate how the numeric values for the attribute are to be subdivided into sub-ranges. Each cut point is introduced by one of the relations `<` or `<=` which is followed by the value that terminates the sub-range. If the relation is `<`, the sub-range includes all values less than the specified value. If the relation is `<=`, the sub-range includes all values less than or equal to the specified value.

Example:

```
Spend < 10 <= 100
```

This specifies that the attribute Spend has three sub-ranges, below the first cut point, between the two cut points, and above the last cut point:

```
Spend < 10
```

```
10 <= Spend <= 100
```

```
Spend > 100
```

To allow Magnum Opus to select sub-ranges, use the keyword `numeric`, followed by the number of sub-ranges required.

Example:

```
Spend: numeric 5
```

For compatibility with See-5, Magnum Opus also accepts the keyword `continuous` which is treated as `numeric 3`.

The keyword `ignore` instructs Magnum Opus to discard any data for the given attribute. This is useful for handling attributes that may appear in the data but which should not be used, such as record identifiers.

5. A worked example using attribute-value data

We now provide a worked example using the attribute-value example files distributed with **Magnum Opus**, tutorial.nam and tutorial.data. Tutorial.nam contains the following:

```
Profitability99: numeric 3
Profitability98: numeric 3
Spend99: numeric 3
Spend98: numeric 3
NoVisits99: numeric 3
NoVisits98: numeric 3
Dairy: numeric 3
Deli: numeric 3
Bakery: numeric 3
Grocery: numeric 3
SocioEconomicGroup: categorical
Promotion1: t, f
Promotion2: t, f
```

Most of these attributes are numeric. These numeric attributes have been designated `numeric 3`, indicating that they should be divided into three sub-ranges, each of which contains approximately the same number of cases. The profitability attributes represent respectively the profit made from a customer in 1999 and 1998. The spend attributes represent the total amount spent by a customer in each year. The NoVisits attributes represent the numbers of store visits in each year. The Dairy, Deli, Bakery, and Grocery attributes record the customer's total spend in each of four significant departments. The remaining three attributes are categorical. The SocioEconomicGroup attribute records an assessment of the customer's socio-economic group. The keyword `categorical` tells **Magnum Opus** to use whatever values it finds in the corresponding column in the data file. The final two attributes record whether the customer participated in each of two store promotions. The values that are allowed are listed. This allows error checking. If any other value appears in the column for the attribute an error message will be displayed.

The first line of the data file describes the first entity:

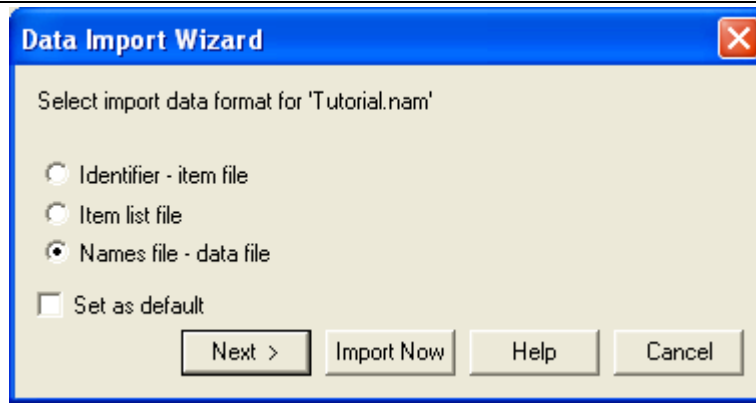
```
829, 709, 5250, 6560, 70, 82, 1074, 390, 878, 1995, C, f, f
```

This indicates that for the first entity the value of Profitability99 is 829 and so on through to the value of Promotion2 being 'f'.

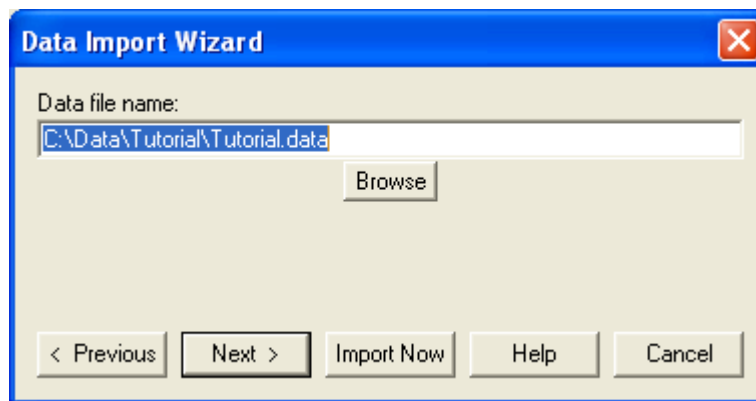
In the next example we run **Magnum Opus** on this names file and data file with its default settings, except that we limit the number of rules produced to five only.

Command line: `mocl names-file=tutorial.nam \`
`data-file=tutorial.data maximum-results=5`

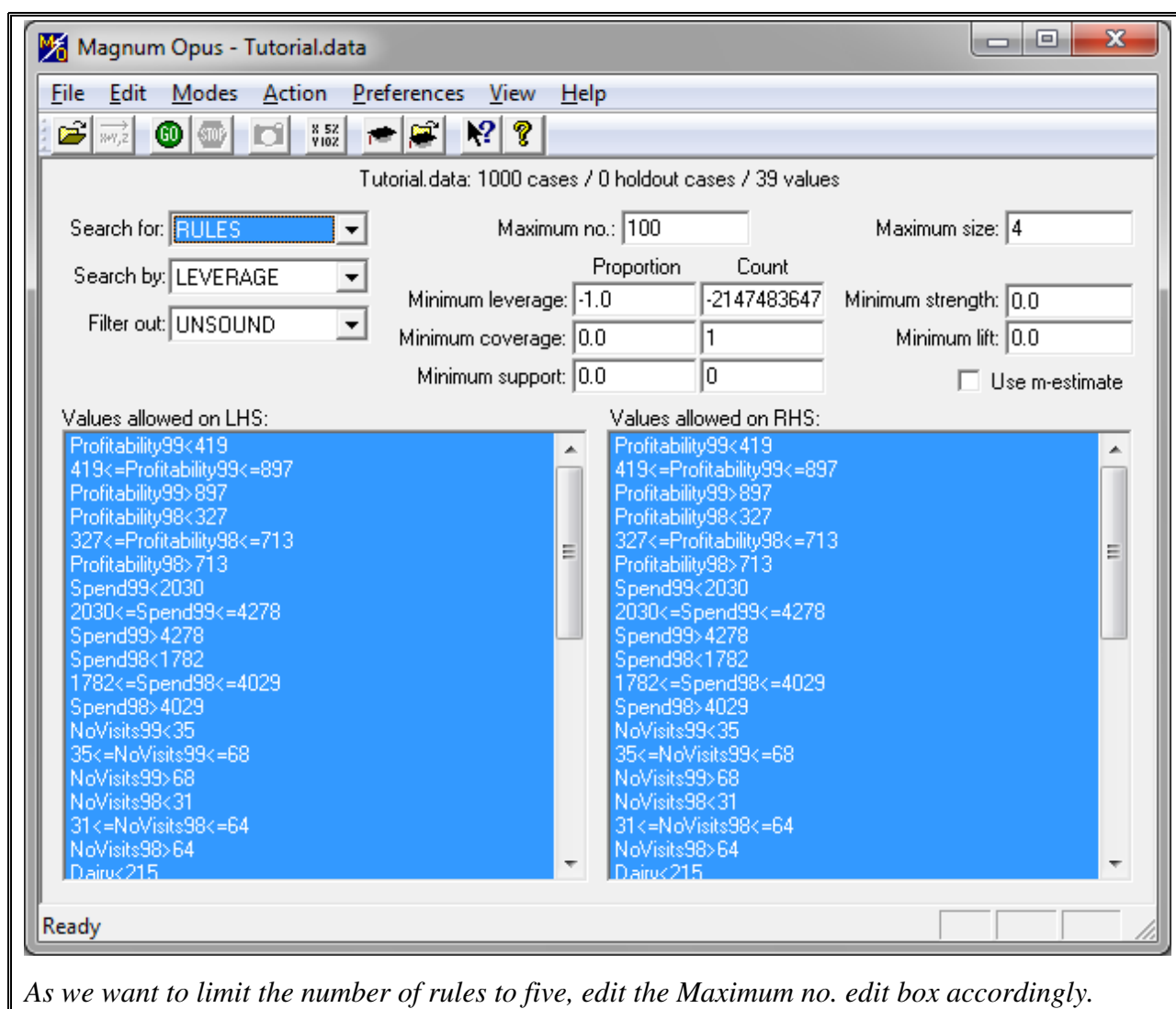
Interactive system. *First run **Magnum Opus**. From the File Menu select Import Data. The system will display a dialog for selecting a file to open. If necessary, navigate to the Example Files folder within the folder into which you installed the software. Select the file tutorial.nam. The system will now display the following dialog box.*



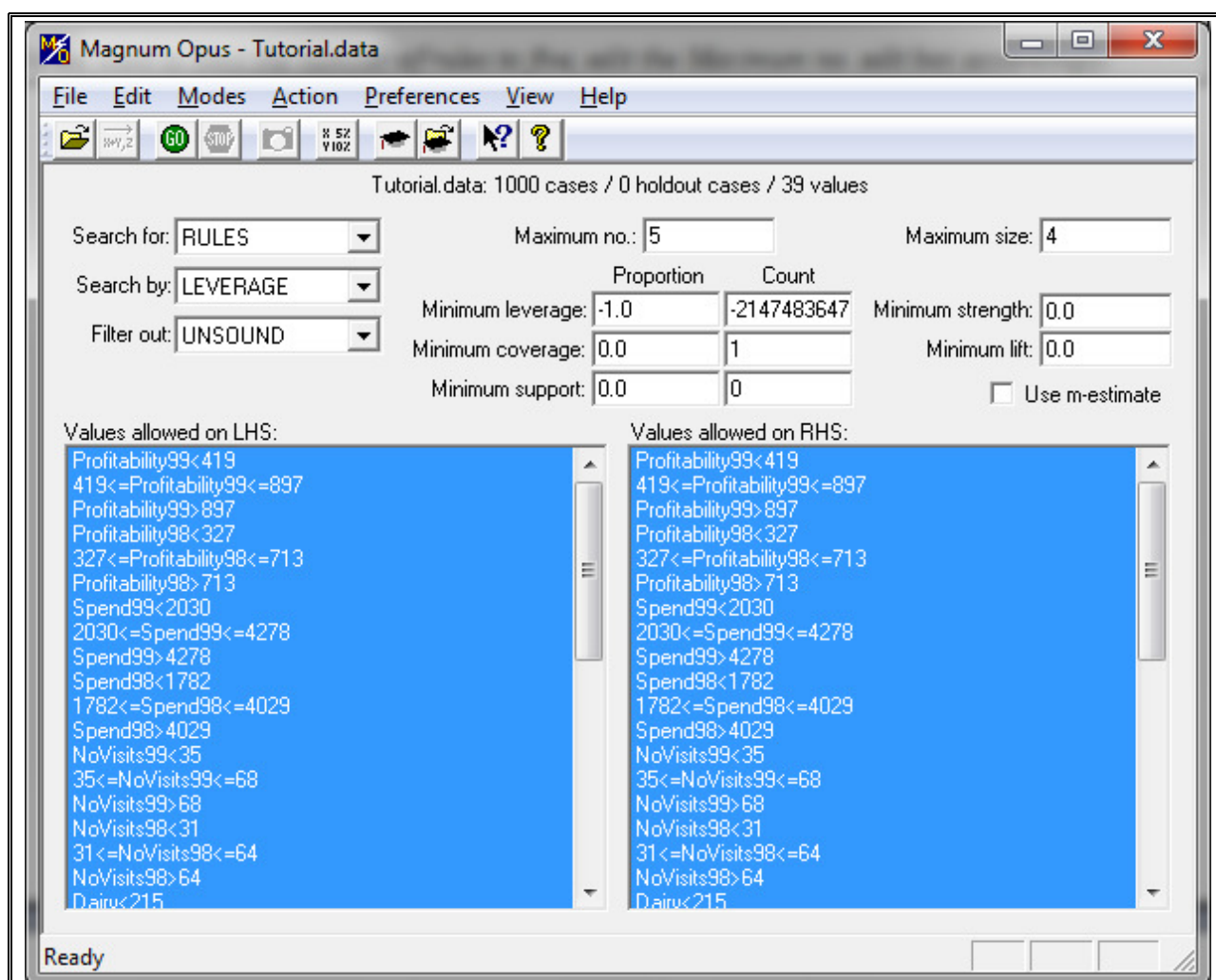
The system recognizes from the nam file extension that the file is a names file. As this is correct, we click the Next > button. The system then displays the following dialog box for selecting the data file.



As the system has defaulted to the correct file name and we wish to use the default settings, click the Import Now button. After importing the data the screen should appear as follows.



As we want to limit the number of rules to five, edit the Maximum no. edit box accordingly.



Now click the **GO** button to commence a search with the selected settings. A dialog will be displayed that allows you to select the file into which the results will be stored. Specify a file name and navigate to the folder in which you want it stored. Then click on the **Save** button. The system will perform the search, saving the results in the specified file and then open the file for inspection.

Output:

Magnum Opus - The leader in association discovery technology.
Version 4.6.3
Copyright (c) 1999-2010 G. I. Webb & Associates Pty Ltd.

Names file: Tutorial.nam
Data file: Tutorial.data

1000 cases / 0 holdout cases / 39 values

Sun Aug 08 18:52:22 2010

Search for rules
Search by leverage

Filter out rules that are unsound.

Maximum number of attributes on LHS = 4

Maximum number of rules = 5

Minimum leverage = -1.0

Minimum leverage count = -2147483647

Minimum coverage = 0.0

Minimum coverage count = 1

Minimum support = 0.0

Minimum support count = 0

Minimum lift = 0.0

Minimum strength = 0.0

All values allowed on LHS

All values allowed on RHS

Search space for LHS size 1 = 699, adjusted critical value = 1.78827E-005

Search space for LHS size 2 = 22875, adjusted critical value = 5.46448E-007

Search space for LHS size 3 = 225960, adjusted critical value = 5.53195E-008

Search space for LHS size 4 = 1.50093E+006, adjusted critical value = 8.32817E-009

Found 5 rules

Spend99<2030 -> Profitability99<419

[Coverage=0.333 (333); Support=0.302 (302); Strength=0.907; Lift=2.72; Leverage=0.1911 (191.1); p=1.66E-178]

Profitability99<419 -> Spend99<2030

[Coverage=0.333 (333); Support=0.302 (302); Strength=0.907; Lift=2.72; Leverage=0.1911 (191.1); p=1.66E-178]

Spend98<1782 -> Profitability98<327

[Coverage=0.331 (331); Support=0.295 (295); Strength=0.891; Lift=2.68; Leverage=0.1848 (184.8); p=5.12E-165]

Profitability98<327 -> Spend98<1782

[Coverage=0.333 (333); Support=0.295 (295); Strength=0.886; Lift=2.68; Leverage=0.1848 (184.8); p=5.12E-165]

NoVisits98<31 -> NoVisits99<35

[Coverage=0.325 (325); Support=0.288 (288); Strength=0.886; Lift=2.69; Leverage=0.1811 (181.1); p=1.89E-159]

As can be seen, the output is very similar to that for transaction data, except that each item consists of an attribute-value pair.

You now know enough to undertake standard association analysis with both transaction and attribute-value data. If you have not yet done so, you might want to take a break from the

tutorial and start analyzing your own data, returning to the following advanced topics when you have gained a little familiarity with using the system.

6. Searching by strength

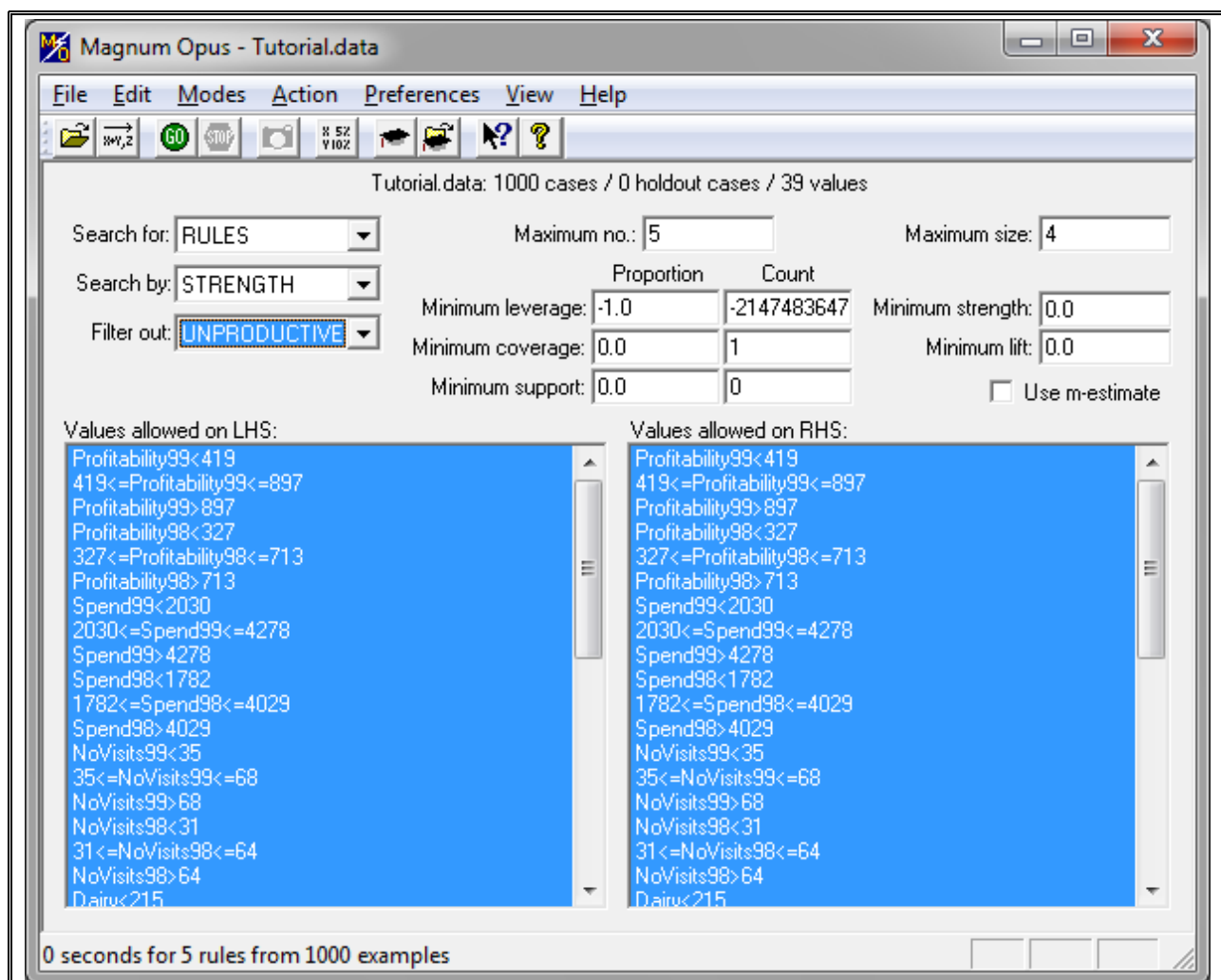
Magnum Opus has several valuable features that distinguish it from most association discovery systems. One important difference is that it allows the user to specify both how many associations to find and what measure should be used to judge how interesting an association is. Any of the measures coverage, support, strength, lift or leverage can be used for this purpose.

The first example run, above, found the five rules with the highest leverage. High leverage rules have a strong positive association between the LHS and RHS and maximize the number of times more frequently the RHS occurs in the context of the LHS than would be expected if they were not associated with one another.

The two other measures that are most frequently used are strength and lift. For our next example we will rerun the first analysis using strength as the measure by which to search. To help illustrate one of the issues involved in search by strength and lift, we will change the filter that is used to the unproductive filter. The default unsound filter discards potential associations that do not pass stringent statistical tests. The unproductive filter is much less strict.

```
Command line: mocl names-file=tutorial.nam data-file=tutorial.data \  
                 maximum-results=5 filter=unproductive \  
                 search-mode=strength
```

Interactive system. *Continuing from the previous example, select Strength in the Search by combo box and Unproductive in the Filter out combo box. The screen should appear as follows.*



Now click the GO button to commence the search. As previously, a dialog will be displayed that allows you to select the file into which the results will be stored. Specify a file name and navigate to the folder in which you want the file stored. Then click on the Save button. The system will perform the search, saving the results in the specified file and then open the file for inspection.

Output:

Magnum Opus - The leader in association discovery technology.
Version 4.6.3
Copyright (c) 1999-2010 G. I. Webb & Associates Pty Ltd.

Names file: Tutorial.nam
Data file: Tutorial.data

1000 cases / 0 holdout cases / 39 values

Mon Aug 09 21:10:49 2010

Search for rules
Search by strength

Filter out unproductive rules

Maximum number of attributes on LHS = 4

Maximum number of rules = 5

Minimum leverage = -1.0

Minimum leverage count = -2147483647

Minimum coverage = 0.0

Minimum coverage count = 1

Minimum support = 0.0

Minimum support count = 0

Minimum lift = 0.0

Minimum strength = 0.0

All values allowed on LHS

All values allowed on RHS

Found 5 rules

NoVisits98<31 & SocioEconomicGroup=B & Promotion1=t -> Spend98<1782
[Coverage=0.014 (14); Support=0.014 (14); Strength=1.000; Lift=3.02;
Leverage=0.0094 (9.4)]

NoVisits98<31 & SocioEconomicGroup=B & Promotion1=t & Promotion2=t ->
Profitability98<327
[Coverage=0.005 (5); Support=0.005 (5); Strength=1.000; Lift=3.00;
Leverage=0.0033 (3.3)]

SocioEconomicGroup=D2 & Promotion1=t & Promotion2=t -> Deli<208
[Coverage=0.005 (5); Support=0.005 (5); Strength=1.000; Lift=3.00;
Leverage=0.0033 (3.3)]

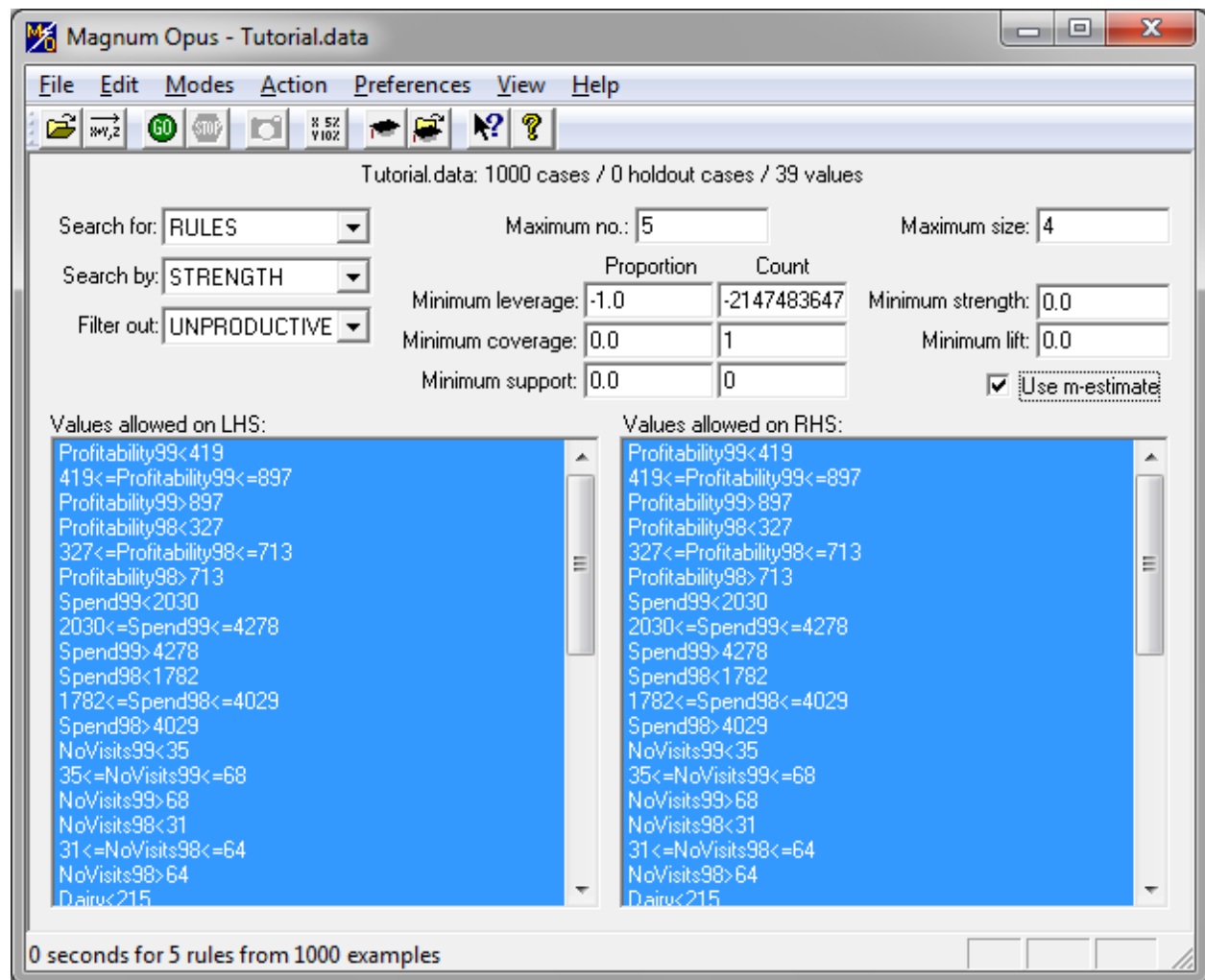
NoVisits98<31 & SocioEconomicGroup=A & Promotion1=t & Promotion2=t ->
NoVisits99<35
[Coverage=0.002 (2); Support=0.002 (2); Strength=1.000; Lift=3.04;
Leverage=0.0013 (1.3)]

NoVisits98<31 & SocioEconomicGroup=A & Promotion1=t & Promotion2=t ->
210<=Bakery<=558
[Coverage=0.002 (2); Support=0.002 (2); Strength=1.000; Lift=2.99;
Leverage=0.0013 (1.3)]

Comparing the two sets of rules, the first thing to note is that, because a statistical filter is not used, p-values are not reported. Note also that in the first set all rules have substantially higher leverage while the second have much higher strength, as these are the measures that each seeks to optimize. It is also notable that the coverage for the rules in the second set is much lower. When coverage is small, there is a substantial risk that values of strength and lift will be overestimated. To guard against this, **Magnum Opus** supports a Bayesian smoothing mechanism called the *m-estimate* that adjusts values of strength and lift to reduce this risk. For our next example we will rerun the previous analysis using this mechanism.

Command line: `mocl names-file=tutorial.nam data-file=tutorial.data \`
`maximum-results=5 filter=unproductive \`
`search-mode=strength m=2`

Interactive system. Continuing from the previous point, select the *m*-estimate check box. The screen should now appear as follows.



Now click the **GO** button to commence the search. As previously, a dialog will be displayed that allows you to select the file into which the results will be stored. Specify a file name and navigate to the folder in which you want it stored. Then click on the **Save** button. The system will perform the search, saving the results in the specified file and then open the file for inspection.

Output:

Magnum Opus - The leader in association discovery technology.
Version 4.6.3
Copyright (c) 1999-2010 G. I. Webb & Associates Pty Ltd.

Names file: Tutorial.nam
Data file: Tutorial.data

1000 cases / 0 holdout cases / 39 values

Mon Aug 09 21:22:15 2010

Search for rules
Search by strength
Filter out unproductive rules

Maximum number of attributes on LHS = 4
Maximum number of rules = 5
Minimum leverage = -1.0
Minimum leverage count = -2147483647
Minimum coverage = 0.0
Minimum coverage count = 1
Minimum support = 0.0
Minimum support count = 0
Minimum lift = 0.0
Minimum strength = 0.0

Use m-estimate, m = 2

All values allowed on LHS

All values allowed on RHS

Found 5 rules

Profitability99<419 & Dairy<215 & Bakery<210 -> Spend99<2030
[Coverage=0.210 (210); Support=0.210 (210); Strength estimate=0.994;
Lift estimate=2.98; Leverage=0.1401 (140.1)]

Profitability99<419 & NoVisits99<35 & Dairy<215 -> Spend99<2030
[Coverage=0.208 (208); Support=0.208 (208); Strength estimate=0.994;
Lift estimate=2.98; Leverage=0.1387 (138.7)]

Profitability99<419 & Bakery<210 & Grocery<873 -> Spend99<2030
[Coverage=0.207 (207); Support=0.207 (207); Strength estimate=0.994;
Lift estimate=2.98; Leverage=0.1381 (138.1)]

Profitability99<419 & Dairy<215 & Deli<208 -> Spend99<2030
[Coverage=0.204 (204); Support=0.204 (204); Strength estimate=0.994;
Lift estimate=2.98; Leverage=0.1361 (136.1)]

```
Profitability99<419 & Spend98<1782 & NoVisits99<35 & Deli<208 ->  
Spend99<2030  
[Coverage=0.192 (192); Support=0.192 (192); Strength estimate=0.993;  
Lift estimate=2.98; Leverage=0.1281 (128.1)]
```

Note first of all that the values for strength and lift are called *Strength Estimate* and *Lift Estimate* when the m-estimate is used. Also note that while a number of the same rules are discovered as previously, the estimates of their strength and lift are substantially reduced. Finally, note that the rules discovered using the m-estimate have substantially higher coverage than those previously discovered. The use of m-estimates is strongly advised when searching by strength or lift.

7. Searching by lift

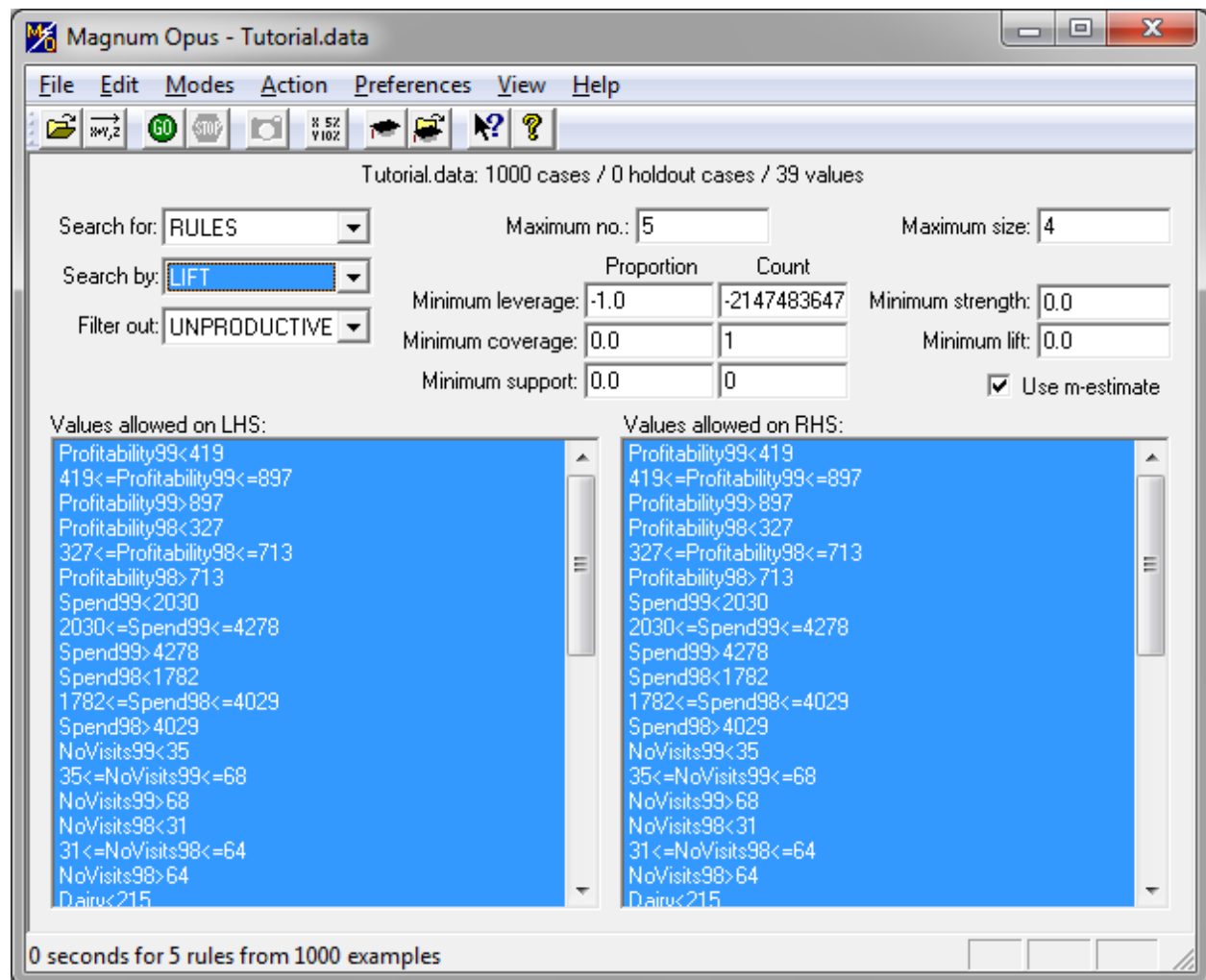
A search by strength with an m-estimate will tend to find strongly predictive rules. These are rules for which the RHS is very likely whenever the LHS occurs. However, some times rather than highly predictive rules, it is desirable to find rules that ‘beat the odds.’ For example, suppose there is a product that most people buy most of the time, such as might be the case if customers are required to purchase bags if they wish to have their purchases packed. Let us assume that 90% of customers buy bags. In this case the rule

```
confectionery -> bags [Coverage=0.336 (336); Support=0.302 (302);  
Strength=0.900; Lift=1.000; Leverage=0.0000 (-0.4)]
```

will enable us to predict with reasonable accuracy that the probability of a customer purchasing a bag if they purchase confectionery is 90%. However, such a rule may not be very useful, as it does not change our default expectation of the probability the customer will purchase a bag. Lift measures how much the rule increases the probability of the RHS relative to the default. To illustrate this, we next perform a search by lift. Note that we will use an m-estimate, as in the previous example.

```
Command line: mocl names-file=tutorial.nam data-file=tutorial.data \  
maximum-results=5 filter=unproductive \  
search-mode=lift m=2
```

Interactive system. Continuing from the previous point, select *Lift* in the Search by ComboBox. The screen should now appear as follows.



Now click the **GO** button to commence the search. As previously, a dialog will be displayed that allows you to select the file into which the results will be stored. Specify a file name and navigate to the folder in which you want it stored. Then click on the **Save** button. The system will perform the search, saving the results in the specified file and then open the file for inspection.

Output:

Magnum Opus - The leader in association discovery technology.
 Version 4.6.3
 Copyright (c) 1999-2010 G. I. Webb & Associates Pty Ltd.

Names file: Tutorial.nam
 Data file: Tutorial.data

1000 cases / 0 holdout cases / 39 values

Mon Aug 09 21:28:35 2010

```

Search for rules
Search by lift
Filter out unproductive rules

Maximum number of attributes on LHS = 4
Maximum number of rules = 5
Minimum leverage = -1.0
Minimum leverage count = -2147483647
Minimum coverage = 0.0
Minimum coverage count = 1
Minimum support = 0.0
Minimum support count = 0
Minimum lift = 0.0
Minimum strength = 0.0

Use m-estimate, m = 2

All values allowed on LHS

All values allowed on RHS

Found 5 rules

327<=Profitability98<=713 & 2030<=Spend99<=4278 & Spend98<1782 &
Bakery>558 -> SocioEconomicGroup=D2
[Coverage=0.003 (3); Support=0.003 (3); Strength estimate=0.639; Lift
estimate=6.59; Leverage=0.0027 (2.7)]

NoVisits99>68 & 208<=Deli<=556 & 210<=Bakery<=558 & Grocery<873 ->
SocioEconomicGroup=A
[Coverage=0.004 (4); Support=0.004 (4); Strength estimate=0.704; Lift
estimate=6.34; Leverage=0.0036 (3.6)]

Profitability98>713 & 1782<=Spend98<=4029 & 210<=Bakery<=558 &
Promotion2=t -> SocioEconomicGroup=D2
[Coverage=0.005 (5); Support=0.004 (4); Strength estimate=0.599; Lift
estimate=6.18; Leverage=0.0035 (3.5)]

NoVisits99>68 & NoVisits98>64 & 210<=Bakery<=558 & Grocery<873 ->
SocioEconomicGroup=A
[Coverage=0.003 (3); Support=0.003 (3); Strength estimate=0.644; Lift
estimate=5.81; Leverage=0.0027 (2.7)]

327<=Profitability98<=713 & NoVisits99>68 & Dairy<215 & Grocery<873 ->
SocioEconomicGroup=A
[Coverage=0.003 (3); Support=0.003 (3); Strength estimate=0.644; Lift
estimate=5.81; Leverage=0.0027 (2.7)]

```

Whereas the search by strength found rules with higher strength, this search finds rules with reasonable strength for items that are not frequently purchased. For example, only 9.7% of

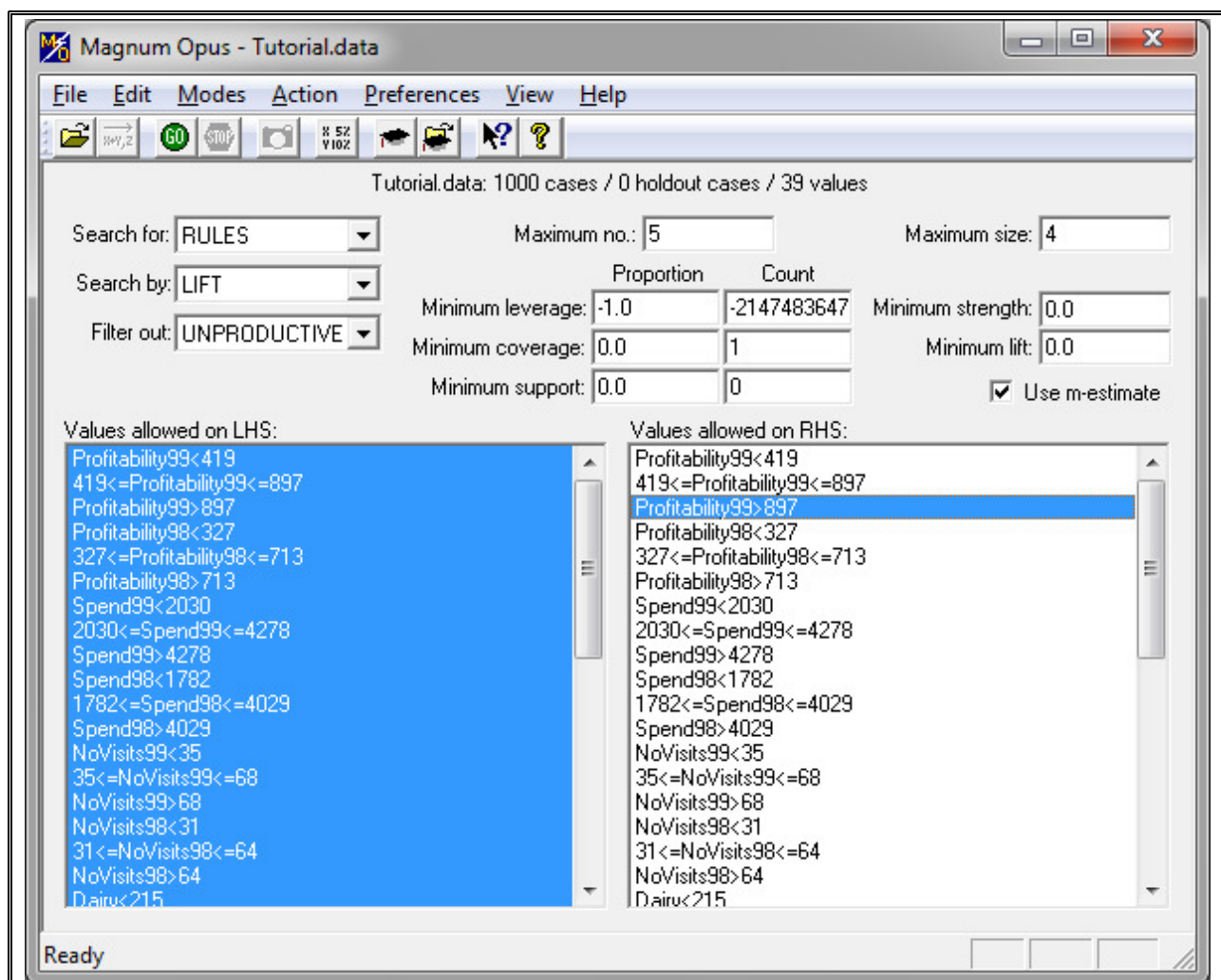
customers belong to socio-economic group D2, but all customers with $327 \leq \text{Profitability}_{98} \leq 713$ and $2030 \leq \text{Spend}_{99} \leq 4278$ and $\text{Spend}_{98} < 1782$ and $\text{Bakery} > 558$ belong to this group. While the system discounts this evidence due to the small number of examples, it is still taken as evidence of a large increase in the frequency with which such customers belong to this group.

8. Selecting RHS elements

Sometimes it will be desirable to find rules for predicting one particular outcome. For example, you might only be interested in predicting the likelihood that a customer will have high profitability. To this end, you can restrict the items that are allowed to appear on either the LHS or RHS of a rule. For the next example, we will rerun the last analysis but with the RHS restricted to $\text{Profitability}_{99} > 897$.

| |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Command line: <code>moel names-file=tutorial.nam data-file=tutorial.data \ maximum-results=5 filter=unproductive \ search-mode=lift m=2 rhs-available=Profitability99>897</code> |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

| |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Interactive system. <i>Continuing from the previous point, select $\text{profitability}_{99} > 897$ in the Values allowed on RHS selection box. The screen should now appear as follows.</i> |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|



Now click the GO button to commence the search. As previously, a dialog will be displayed that allows you to select the file into which the results will be stored. Specify a file name and navigate to the folder in which you want it stored. Then click on the Save button. The system will perform the search, saving the results in the specified file and then open the file for inspection.

Output:

Magnum Opus - The leader in association discovery technology.
Version 4.6.3
Copyright (c) 1999-2010 G. I. Webb & Associates Pty Ltd.

Names file: Tutorial.nam
Data file: Tutorial.data

1000 cases / 0 holdout cases / 39 values

Sat Aug 14 17:05:12 2010

Search for rules
Search by lift

Filter out unproductive rules

Maximum number of attributes on LHS = 4

Maximum number of rules = 5

Minimum leverage = -1.0

Minimum leverage count = -2147483647

Minimum coverage = 0.0

Minimum coverage count = 1

Minimum support = 0.0

Minimum support count = 0

Minimum lift = 0.0

Minimum strength = 0.0

Use m-estimate, m = 2

All values allowed on LHS

Values allowed on RHS:

Profitability99>897

Found 5 rules

Profitability98>713 & Spend99>4278 & Dairy>516 & Grocery>2113 ->
Profitability99>897
[Coverage=0.125 (125); Support=0.125 (125); Strength estimate=0.989;
Lift estimate=2.99; Leverage=0.0836 (83.6)]

Profitability98>713 & Dairy>516 & Bakery>558 & Grocery>2113 ->
Profitability99>897
[Coverage=0.098 (98); Support=0.098 (98); Strength estimate=0.987;
Lift estimate=2.98; Leverage=0.0656 (65.6)]

Profitability98>713 & Spend99>4278 & Deli>556 & Grocery>2113 ->
Profitability99>897
[Coverage=0.131 (131); Support=0.130 (130); Strength estimate=0.982;
Lift estimate=2.97; Leverage=0.0866 (86.6)]

Profitability98>713 & NoVisits98>64 & Deli>556 & Grocery>2113 ->
Profitability99>897
[Coverage=0.109 (109); Support=0.108 (108); Strength estimate=0.979;
Lift estimate=2.96; Leverage=0.0719 (71.9)]

Profitability98>713 & NoVisits99>68 & Deli>556 & Grocery>2113 ->
Profitability99>897
[Coverage=0.109 (109); Support=0.108 (108); Strength estimate=0.979;
Lift estimate=2.96; Leverage=0.0719 (71.9)]

Only rules with Profitability99>897 on the RHS are returned.

Sometimes some data elements represent inputs to a process and other outputs. In such circumstances it will often be useful to limit the LHS values to the inputs and the RHS values to the outputs. The rules that are discovered will then represent ways of manipulating the inputs in order to produce specific outcomes.

9. Itemsets

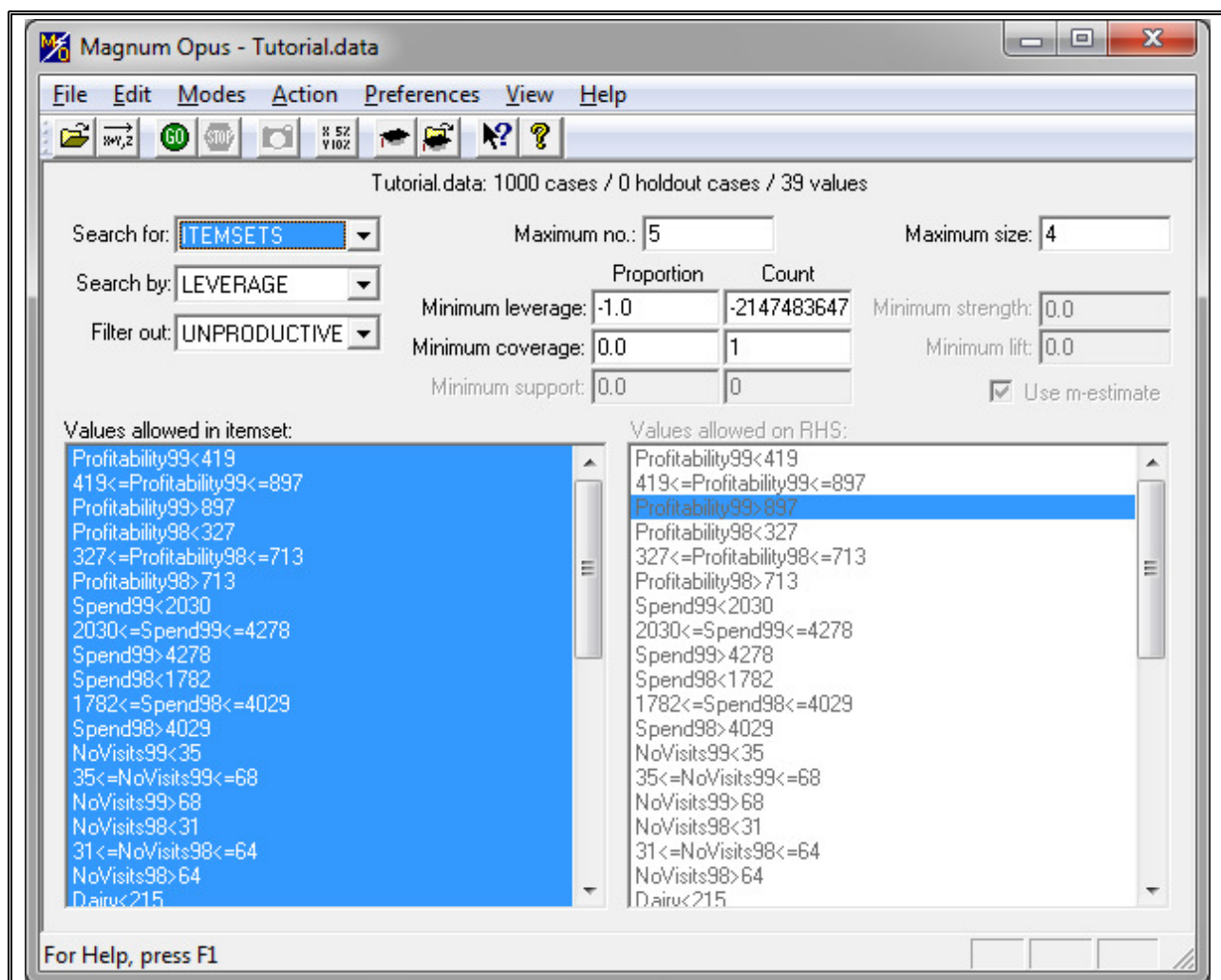
Rules are a useful way to describe interactions between elements of the data when the objective is to predict the probability of specific items in specific contexts. Sometimes, however, the primary issue is simply to identify which items occur together. In this case, presenting the interactions as rules can be distracting. For example, a single interaction between elements can result in many rules.

Itemsets are simply collections of items that appear together. The system supports two measures of the importance of an itemset, *coverage* and *leverage*. The coverage is the number of transactions or cases that contain the itemset. The leverage is the difference between this and the maximum coverage that would be expected assuming that any two subsets of the items were unrelated to one another.

The next example finds itemsets for the tutorial. data.

| |
|------------------------------------------------------------------------------------------------------------------------------------------|
| Command line: <code>mocl names-file=tutorial.nam data-file=tutorial.data \</code> <code>maximum-results=5 find-itemsets</code> |
|------------------------------------------------------------------------------------------------------------------------------------------|

| |
|--------------------------------------------------------------------------------------------------------------------------------------------------------|
| Interactive system. <i>Continuing from the previous point, select itemsets in the Search for comboBox. The screen should appear as follows.</i> |
|--------------------------------------------------------------------------------------------------------------------------------------------------------|



Now click the GO button to commence the search. As previously, a dialog will be displayed that allows you to select the file into which the results will be stored. Specify a file name and navigate to the folder in which you want it stored. Then click on the Save button. The system will perform the search, saving the results in the specified file and then open the file for inspection.

Output:

Magnum Opus - The leader in association discovery technology.
Version 4.6.3
Copyright (c) 1999-2010 G. I. Webb & Associates Pty Ltd.

Names file: Tutorial.nam
Data file: Tutorial.data

1000 cases / 0 holdout cases / 39 values

Mon Aug 09 21:40:39 2010

Search for itemsets
Search by leverage

```
Filter out unproductive itemsets

Maximum number of values in an itemset = 4
Maximum number of itemsets = 5
Minimum leverage = -1.0
Minimum leverage count = -2147483647
Minimum coverage = 0.0
Minimum coverage count = 1

All values allowed

Found 5 itemsets

Profitability99<419 & Spend99<2030
[Coverage=0.302 (302); Leverage=0.1911 (191.1)]

Profitability98<327 & Spend98<1782
[Coverage=0.295 (295); Leverage=0.1848 (184.8)]

NoVisits99<35 & NoVisits98<31
[Coverage=0.288 (288); Leverage=0.1811 (181.1)]

Profitability99>897 & Spend99>4278
[Coverage=0.287 (287); Leverage=0.1768 (176.8)]

Spend98<1782 & NoVisits98<31
[Coverage=0.277 (277); Leverage=0.1694 (169.4)]
```

Each itemset is presented as a list of the items in the set. The coverage and leverage statistics that are provided were described above.

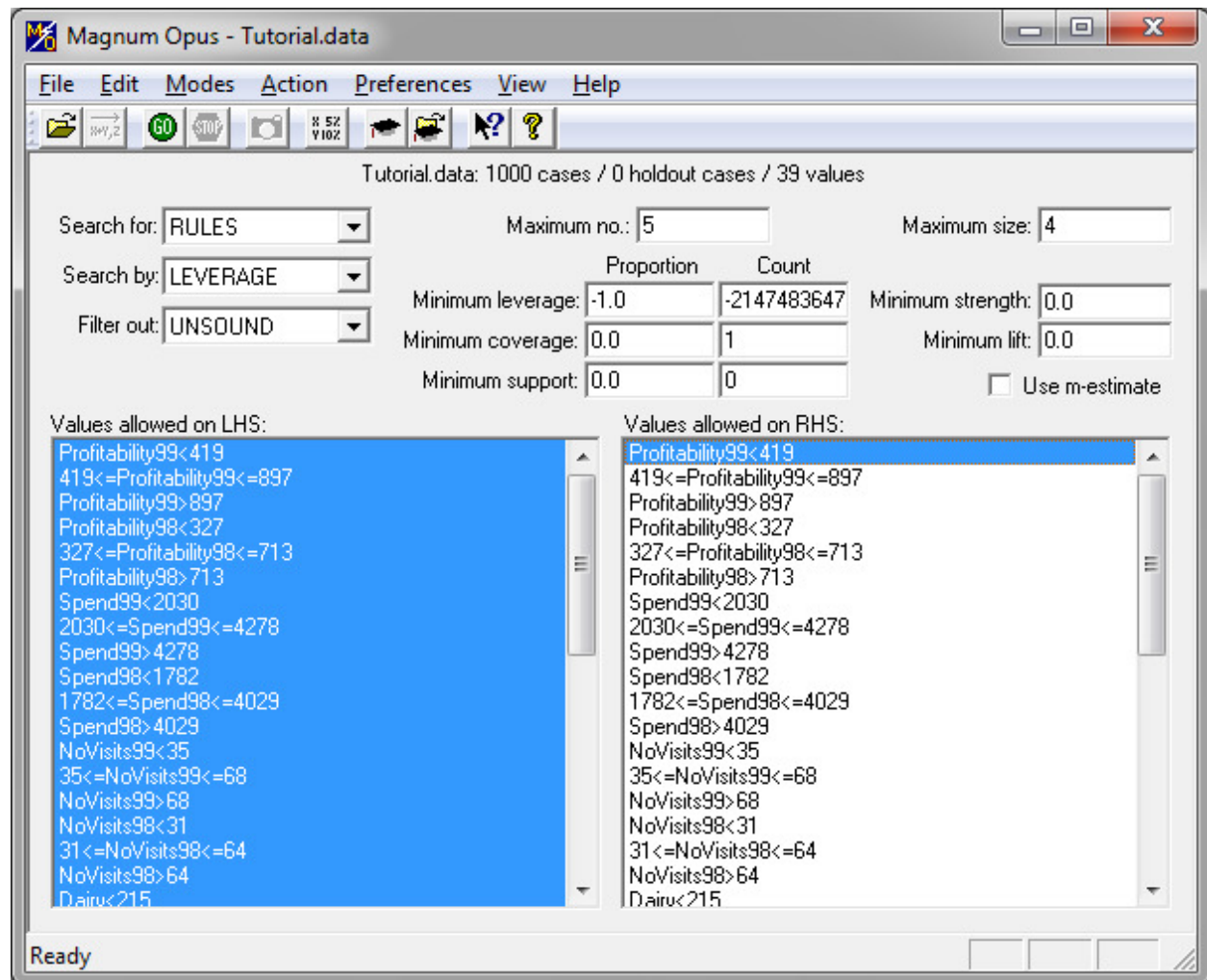
10. Contrast discovery

A common analytic task seeks to identify factors that distinguish different groups. This type of analysis is called *contrast discovery*. To perform contrast discovery it is necessary to provide each example in the data with a label identifying to which group it belongs. For attribute-value data this means providing an attribute whose values indicate group membership. For example, in the tutorial.data file, the Profitability99 attribute might be used to indicate that each example belongs to one of three groups, *low profit* (Profitability99<419), *medium profit* (419<=Profitability99<=897) or *high profit* (Profitability99>897). For transaction data it is necessary to add another item to each transaction. It is important to use a name for these labels that will not be used or mistaken for a standard item. For example, one might add items such as *profitable* and *unprofitable* to the transactions in the tutorial.itl data.

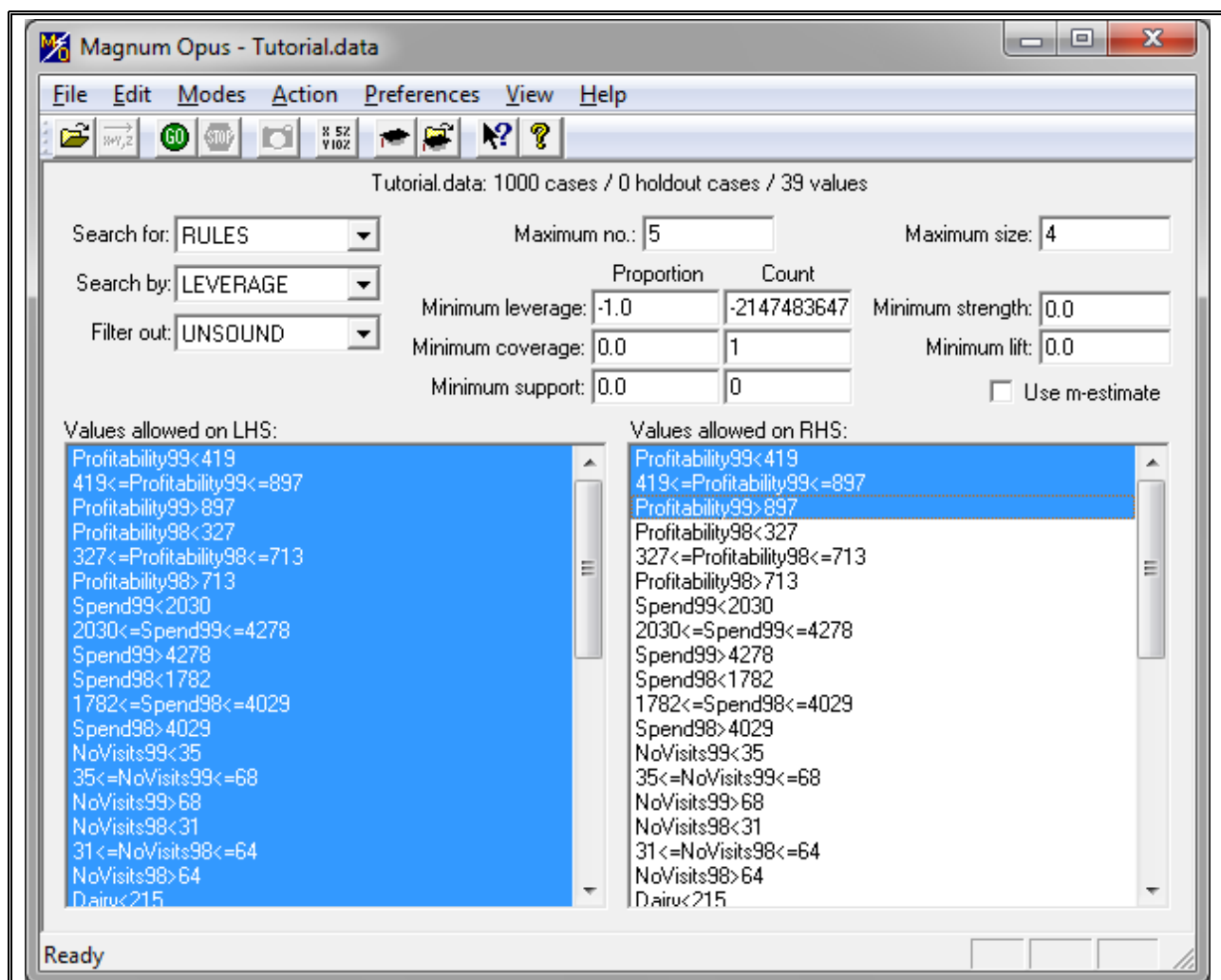
Once group labels have been added to the data, simply run **Magnum Opus** restricting the RHS values to the group labels. The next example illustrates this process using the data in the file tutorial.data, treating the Profitability99 attribute as the group variable.

Command line: mol names-file=tutorial.nam data-file=tutorial.data \
maximum-results=5 rhs-available=Profitability99

Interactive system. Continuing from the previous point, reset the Search By combo box to Leverage, restore the Filter-out combo box to Unsound and uncheck the Use m-estimate box. The screen should appear as follows.. Then, select the three values for profitability in the Values allowed on RHS edit box by first left-clicking Profitability99<419



and then, holding down the SHIFT key and left-clicking Profitability99>897.



Now click the **GO** button to commence a search with the selected settings. A dialog will be displayed that allows you to select the file into which the results will be stored. Specify a file name and navigate to the folder in which you want it stored. Then click on the **Save** button. The system will perform the search, saving the results in the specified file and then open the file for inspection.

Output:

Magnum Opus - The leader in association discovery technology.
 Version 4.6.3
 Copyright (c) 1999-2010 G. I. Webb & Associates Pty Ltd.

Names file: Tutorial.nam
 Data file: Tutorial.data

1000 cases / 0 holdout cases / 39 values

Sun Aug 08 19:21:35 2010

Search for rules
 Search by leverage

Filter out rules that are unsound.

Maximum number of attributes on LHS = 4

Maximum number of rules = 5

Minimum leverage = -1.0

Minimum leverage count = -2147483647

Minimum coverage = 0.0

Minimum coverage count = 1

Minimum support = 0.0

Minimum support count = 0

Minimum lift = 0.0

Minimum strength = 0.0

All values allowed on LHS

Values allowed on RHS:

Profitability99<419 419<=Profitability99<=897 Profitability99>897

Search space for LHS size 1 = 108, adjusted critical value = 0.000115741

Search space for LHS size 2 = 1773, adjusted critical value = 7.0502E-006

Search space for LHS size 3 = 17556, adjusted critical value = 7.12007E-007

Search space for LHS size 4 = 116802, adjusted critical value = 1.07019E-007

Found 5 rules

Spend99<2030 -> Profitability99<419

[Coverage=0.333 (333); Support=0.302 (302); Strength=0.907; Lift=2.72; Leverage=0.1911 (191.1); p=1.66E-178]

Spend99>4278 -> Profitability99>897

[Coverage=0.333 (333); Support=0.287 (287); Strength=0.862; Lift=2.60; Leverage=0.1768 (176.8); p=8.57E-149]

Spend99<2030 & Grocery<873 -> Profitability99<419

[Coverage=0.278 (278); Support=0.265 (265); Strength=0.953; Lift=2.86; Leverage=0.1724 (172.4); p=2.52E-008]

Grocery<873 -> Profitability99<419

[Coverage=0.333 (333); Support=0.277 (277); Strength=0.832; Lift=2.50; Leverage=0.1661 (166.1); p=6.14E-129]

Profitability98<327 & Spend99<2030 -> Profitability99<419

[Coverage=0.256 (256); Support=0.246 (246); Strength=0.961; Lift=2.89; Leverage=0.1608 (160.8); p=2.74E-008]

The LHS of each rule that is discovered indicates a set of factors that are more frequently associated with the RHS than with any of the other groups. For example, the first rule indicates that customers with Profitability99<419 are more likely to have a low value for Spend99 than are customers with other levels of Spend99.

11. Statistically sound association discovery

Due to the large number of potential associations that are considered during association discovery, it is inevitable that some associations will be ‘discovered’ that only appear strong by chance. **Magnum Opus** incorporates unique facilities for using statistical tests to control the risk of finding such associations. These tests are adjusted for the size of the search space and the number of associations found, as appropriate. Assuming the sample data are a random sample of the broader population about which you wish to reach conclusions, these tests ensure that the risk of ‘discovering’ a spurious association is no greater than the user-specified significance level. By default, significance levels are set to 0.05.

Magnum Opus supports two mechanisms for statistically sound association discovery. *Within-search* testing adjusts the significance level applied to statistical tests that are used while the search is being conducted. Use the *Unsound* filter to perform within-search testing. For rule discovery the unsound filter discards any rule whose strength is not significantly higher than that of any of its generalizations (rules formed by deleting elements from the LHS). For itemset discovery, the unsound filter discards itemsets that are not significantly more frequent than could be expected by assuming that any two subsets of the itemset are independent of one another.

Note that the *Insignificant* filter, also applies a statistical test, but that this test is not adjusted for the size of the search space and hence is not statistically sound. The Insignificant filter is useful for discarding rules and itemsets that are very likely to be spurious, but is likely to still accept some spurious associations.

The second mechanism is *holdout evaluation*. This requires that the data are divided into an exploratory and a holdout set. The associations are discovered from the exploratory data and tested on the holdout data. One way to do this is to have **Magnum Opus** randomly divide the data into these two sets when it is imported. You must then specify that holdout evaluation is to be performed and which statistical tests to employ.

The following holdout evaluation tests are supported for rules.

| Test | Null Hypothesis | Statistical technique |
|-----------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| Minimum Coverage | Coverage \leq Min Coverage | Binomial sign test |
| Minimum Support | Support \leq Min Support | Binomial sign test |
| Minimum Strength | Strength \leq Min Strength | Binomial sign test |
| Minimum Lift | Lift \leq Min Lift | Binomial sign test |
| Minimum Leverage | Leverage \leq Min Leverage | Binomial sign test |
| Positive correlation | Support \leq Coverage \times RHS_Coverage | Fisher exact test |
| Improvement over generalizations | Strength \leq the maximum Strength of any generalization of the current rule | Fisher exact test |
| Partial with respect to specializations | There exists another rule GLHS \rightarrow RHS in the set of best rules, that has not been rejected by holdout evaluation, that is a specialization of the current rule, and such that the LHS and RHS of the current rule are conditionally independent given the negation of | Fisher exact test |

GLHS.

The following holdout evaluation tests are supported for itemsets.

| Test | Null Hypothesis | Statistical technique |
|----------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| Minimum Coverage | Coverage \leq Min Coverage | Binomial sign test |
| Minimum Leverage | Leverage \leq Min Leverage | Binomial sign test |
| Improvement over generalizations | Coverage \leq the maximum of coverage(A) \times coverage(B) for any partition of the current itemset into two subsets A and B. | Fisher exact test |
| Self-sufficient | Coverage \leq the maximum of coverage(A) \times coverage(B) for any partition of the current itemset into two subsets A and B within the set of cases not covered by the difference between the current itemset and any of its productive supersets. | |

The positive correlation test is the default test for rules. It tests whether the leverage of the rule is greater than zero. The improvement over generalizations test is the default test for itemsets. The improvement over generalization tests are equivalent to the tests applied by the unsound filter. The Partial with respect to specializations and Self-sufficient tests check whether a specialization of a rule (a rule created by adding elements to the LHS) or the supersets of an itemset, can explain the frequency with which the itemset occurs.

For more information on statistically sound association discovery see the following worked examples and the research papers:

Webb, G.I. (2007). **Discovering Significant Patterns**. *Machine Learning* 68(1). Netherlands: Springer, pages 1-33.

Webb, G.I. (2008). **Layered Critical Values: A Powerful Direct-Adjustment Approach to Discovering Significant Patterns**. *Machine Learning* 71(2-3). Netherlands: Springer, pages 307-323 [Technical Note].

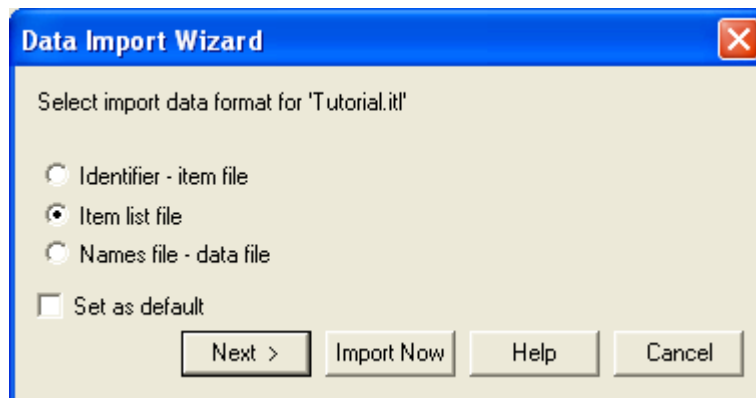
Webb, G.I. (2010). **Self-Sufficient Itemsets: An Approach to Screening Potentially Interesting Associations Between Items**. *Transactions on Knowledge Discovery from Data* 4. ACM, pages 3:1-3:20.

A worked example of holdout evaluation for rules

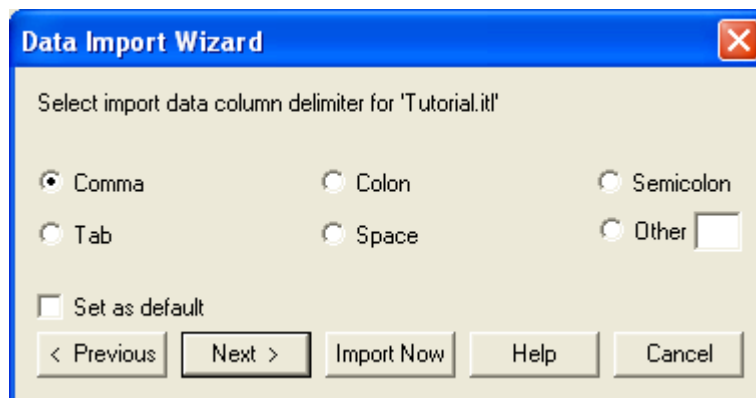
To illustrate holdout evaluation for rules, we run **Magnum Opus** on the tutorial.itl data, selecting 50% of the data for the exploratory set and the remaining 50% for the holdout set, using holdout evaluation, employing the partialness and improvement tests, searching by support, using no filtering and selecting only *tomatoes* and *potatoes* for the LHS and *lettuce* and *carrots* for the RHS.

Command line: `mol item-list-file=tutorial.itl proportion=0.5 \ out-of-sample-holdout-evaluation \ test-partialness=yes test-improvement=yes \ search-mode=support filter=none \ lhs-available=tomatoes,potatoes \ rhs-available=lettuce,carrots`

Interactive system. First run **Magnum Opus**. From the File Menu select Import Data. The system will display a dialog for selecting a file to open. If necessary, navigate to the Example Files folder within the folder into which you installed the software. Select the file tutorial.itl. The system will now display the following dialog box.



The system recognizes from the itl file extension that the file is an item list file. As this is correct, click the Next > button to go to the next screen.



This screen allows you to select the delimiter character. As the default is correct for this file, click Next > to go to the next screen.

Data Import Wizard

Enter the percentage of cases to be imported from 'Tutorial.it'

Minimum = 1.
To import all cases set percentage to 100.

Percentage:

☐ Set as default

< Previous Next > Import Now Help Cancel

This screen allows you to select how much data of the should be loaded into the exploratory set. For this example we wish to load 50%, so change the Percentage box to 50.

Data Import Wizard

Enter the percentage of cases to be imported from 'Tutorial.it'

Minimum = 1.
To import all cases set percentage to 100.

Percentage:

☐ Set as default

< Previous Next > Import Now Help Cancel

Now click Next >.

Data Import Wizard

Set hold out evaluation treatment

☐ No holdout evaluation

☒ Use out of sample data for holdout evaluation

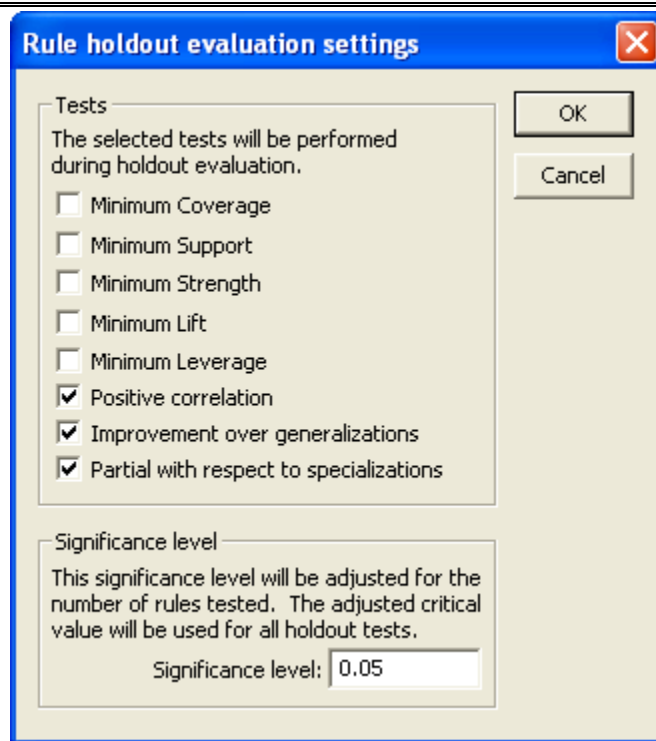
☐ Load holdout data from:

☐ Set as default

< Previous Import Data Help Cancel

The next screen allows you to select whether holdout evaluation is to be performed. If it is, you have the choice of either using the data not included in the exploratory set (the default), or of loading the data from another file. As we wish to use the default, click Import Data. This takes us to the main screen.

We want to select the holdout tests, so select Rule Evaluation Holdout Settings from the Preferences menu. This leads to a dialog that allows you to select the tests and significance level to be applied during holdout evaluation. Select Improvement over generalizations and Partial with respect to specializations.



The image shows a dialog box titled "Rule holdout evaluation settings" with a blue header bar and a red close button. The dialog is divided into two main sections. The top section, titled "Tests", contains a descriptive text and a list of seven checkboxes. The bottom section, titled "Significance level", contains a descriptive text and a text input field. On the right side of the dialog, there are two buttons: "OK" and "Cancel".

Rule holdout evaluation settings

Tests
The selected tests will be performed during holdout evaluation.

- ☐ Minimum Coverage
- ☐ Minimum Support
- ☐ Minimum Strength
- ☐ Minimum Lift
- ☐ Minimum Leverage
- ☒ Positive correlation
- ☒ Improvement over generalizations
- ☒ Partial with respect to specializations

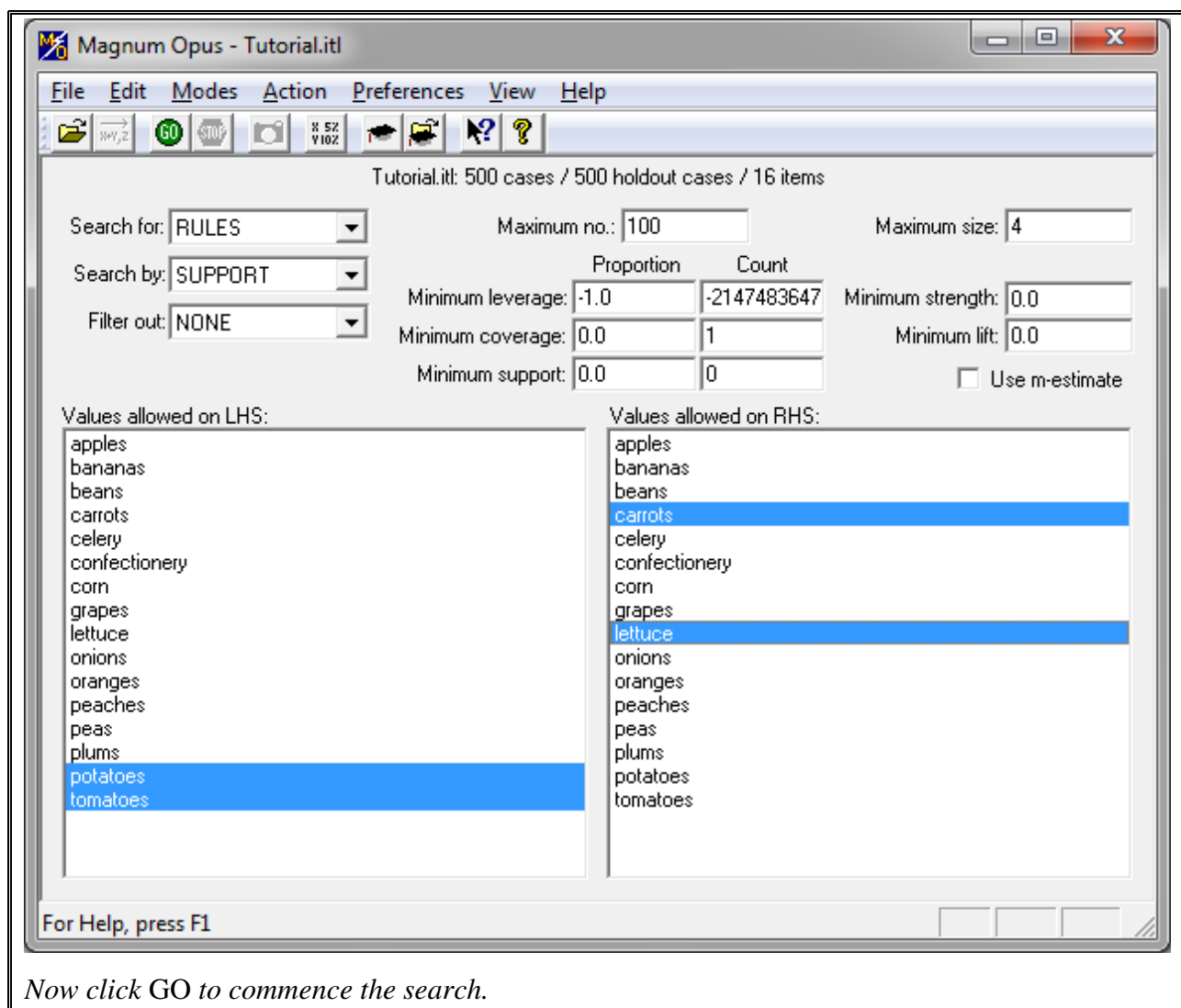
Significance level
This significance level will be adjusted for the number of rules tested. The adjusted critical value will be used for all holdout tests.

Significance level:

OK
Cancel

Then click OK to return to the main screen.

On the main screen select Search by Support and Filter out None. Then select potatoes and tomatoes for the Values allowed on the LHS and carrots and lettuce for the Values allowed on the RHS.



Now click GO to commence the search.

Output:

Magnum Opus - The leader in association discovery technology.

Version 4.6.3

Copyright (c) 1999-2010 G. I. Webb & Associates Pty Ltd.

Data file: Tutorial.itl [50% sample]

500 cases / 500 holdout cases / 16 items

Sun Aug 08 19:27:03 2010

Search for rules

Search by support

Maximum number of attributes on LHS = 4

Maximum number of rules = 100

Minimum leverage = -1.0

Minimum leverage count = -2147483647

Minimum coverage = 0.0

Minimum coverage count = 1

Minimum support = 0.0
Minimum support count = 0
Minimum lift = 0.0
Minimum strength = 0.0

Values allowed on LHS:

potatoes
tomatoes

Values allowed on RHS:

carrots
lettuce

Only 6 rules satisfy the specified constraints.

The following 2 rules passed holdout evaluation

tomatoes -> lettuce
[Coverage=0.244 (122); Support=0.106 (53); Strength=0.434; Lift=1.96;
Leverage=0.0518 (25.9)]

tomatoes -> carrots
[Coverage=0.244 (122); Support=0.080 (40); Strength=0.328; Lift=1.95;
Leverage=0.0390 (19.5)]

The following 4 rules failed holdout evaluation, adjusted critical value = 0.0125

potatoes -> lettuce
[Coverage=0.272 (136); Support=0.076 (38); Strength=0.279; Lift=1.26;
Leverage=0.0156 (7.8)]
Holdout coverage = 147, holdout support = 37, holdout strength = 0.252
Fails positive correlation, p = 0.101
Fails significant improvement with respect to DEFAULT, p = 0.101
Fails partial test with respect to tomatoes & potatoes, p = 0.952

potatoes -> carrots
[Coverage=0.272 (136); Support=0.056 (28); Strength=0.206; Lift=1.23;
Leverage=0.0103 (5.2)]
Holdout coverage = 147, holdout support = 40, holdout strength = 0.272
Fails partial test with respect to tomatoes & potatoes, p = 0.120

tomatoes & potatoes -> lettuce
[Coverage=0.072 (36); Support=0.040 (20); Strength=0.556; Lift=2.50;
Leverage=0.0240 (12.0)]
Holdout coverage = 41, holdout support = 23, holdout strength = 0.561
Fails significant improvement with respect to tomatoes, p = 0.0172

tomatoes & potatoes -> carrots
[Coverage=0.072 (36); Support=0.028 (14); Strength=0.389; Lift=2.31;
Leverage=0.0159 (8.0)]
Holdout coverage = 41, holdout support = 19, holdout strength = 0.463
Fails significant improvement with respect to tomatoes, p = 0.0166

The rules that fail holdout evaluation are listed after those that pass. The rule `tomatoes & potatoes -> lettuce` illustrates the significant improvement test. The rule `tomatoes -> lettuce` has strength 0.434. The 20 examples that provide the support for the longer rule do not provide sufficient evidence that the strength of association is truly higher than that of the shorter rule.

The rule `potatoes -> lettuce` illustrates the partialness test. The rule `tomatoes & potatoes -> lettuce` covers 36 of the 136 examples covered by the shorter rule. It also covers 20 out of the 38 examples that have both potatoes and lettuce. Once the 36 examples covered by the longer rule are removed, the remaining support is just 18 out of 100 examples. The resulting Strength (0.180) is lower than the default strength for tomatoes of (0.244). In consequence, it appears that the increased frequency of lettuce in the context of potatoes is solely due to its increased frequency when both potatoes and tomatoes are present.

A worked example of holdout evaluation for itemsets

To illustrate holdout evaluation for itemsets we continue the previous example. As before, we use **Magnum Opus** on the `tutorial.itl` data, selecting 50% of the data for the exploratory set and the remaining 50% for the holdout set and using holdout evaluation. This time, however, we search for itemsets, searching by coverage, using no filtering and selecting only carrots, lettuce, potatoes and tomatoes for the allowed items.

```
Command line: mocl item-list-file=tutorial.itl proportion=0.5 \ out-of-  
sample-holdout-evaluation \  
find-itemsets test-self-sufficient=yes \  
test-improvement=yes search-mode=coverage \ filter=none \  
items-available= carrots,lettuce,potatoes,tomatoes
```

Interactive system. *Continuing from the previous example, select ITEMSETS in the Search for box, select COVERAGE in the Search by box, and select the items carrots, lettuce, potatoes and tomatoes for the Values allowed in itemset.*

Magnum Opus - Tutorial.itl

File Edit Modes Action Preferences View Help

Tutorial.itl: 500 cases / 500 holdout cases / 16 items

Search for: ITEMSETS Maximum no.: 100 Maximum size: 4

Search by: COVERAGE

Filter out: NONE

| | Proportion | Count | |
|-------------------|------------|-------------|-----------------------------------------|
| Minimum leverage: | -1.0 | -2147483647 | Minimum strength: 0.0 |
| Minimum coverage: | 0.0 | 1 | Minimum lift: 0.0 |
| Minimum support: | 0.0 | 0 | <input type="checkbox"/> Use m-estimate |

Values allowed in itemset:

- apples
- bananas
- beans
- carrots
- celery
- confectionery
- corn
- grapes
- lettuce
- onions
- oranges
- peaches
- peas
- plums
- potatoes
- tomatoes

Values allowed on RHS:

- apples
- bananas
- beans
- carrots
- celery
- confectionery
- corn
- grapes
- lettuce
- onions
- oranges
- peaches
- peas
- plums
- potatoes
- tomatoes

For Help, press F1

Click OK to return to the main window. Now press GO to commence the search.

Output:

Magnum Opus - The leader in association discovery technology.

Version 4.6.3

Copyright (c) 1999-2010 G. I. Webb & Associates Pty Ltd.

Data file: Tutorial.itl [50% sample]

500 cases / 500 holdout cases / 16 items

Sun Aug 15 09:14:06 2010

Search for itemsets

Search by coverage

Maximum number of values in an itemset = 4

Maximum number of itemsets = 100

Minimum leverage = -1.0

Minimum leverage count = -2147483647

Minimum coverage = 0.0

Minimum coverage count = 1

Values allowed:

carrots
lettuce
potatoes
tomatoes

Only 16 itemsets satisfy the specified constraints.

The following 9 itemsets passed holdout evaluation

```
{}  
[Coverage=1.000 (500); Leverage=0.0000 (0.0)]  
  
potatoes  
[Coverage=0.272 (136); Leverage=0.0000 (0.0)]  
  
tomatoes  
[Coverage=0.244 (122); Leverage=0.0000 (0.0)]  
  
lettuce  
[Coverage=0.222 (111); Leverage=0.0000 (0.0)]  
  
carrots  
[Coverage=0.168 (84); Leverage=0.0000 (0.0)]  
  
lettuce & tomatoes  
[Coverage=0.106 (53); Leverage=0.0518 (25.9)]  
  
tomatoes & carrots  
[Coverage=0.080 (40); Leverage=0.0390 (19.5)]  
  
carrots & potatoes  
[Coverage=0.056 (28); Leverage=0.0103 (5.2)]  
  
lettuce & tomatoes & carrots  
[Coverage=0.036 (18); Leverage=0.0182 (9.1)]
```

The following 7 itemsets failed holdout evaluation, adjusted critical value = 0.00313

```
lettuce & potatoes  
[Coverage=0.076 (38); Leverage=0.0156 (7.8)]  
Holdout coverage = 37  
Fails significant improvement with respect to lettuce and potatoes, p = 0.101  
  
tomatoes & potatoes  
[Coverage=0.072 (36); Leverage=0.0056 (2.8)]  
Holdout coverage = 41  
Fails significant improvement with respect to tomatoes and potatoes, p = 0.580  
  
lettuce & carrots  
[Coverage=0.042 (21); Leverage=0.0047 (2.4)]  
Holdout coverage = 24  
Fails significant improvement with respect to lettuce and carrots, p = 0.118  
  
lettuce & tomatoes & potatoes  
[Coverage=0.040 (20); Leverage=0.0112 (5.6)]
```



```

Holdout coverage = 23
Fails significant improvement with respect to lettuce & tomatoes and potatoes,
p = 0.0498

tomatoes & carrots & potatoes
[Coverage=0.028 (14); Leverage=0.0062 (3.1)]
Holdout coverage = 19
Fails significant improvement with respect to tomatoes & carrots and potatoes,
p = 0.0381

lettuce & carrots & potatoes
[Coverage=0.016 (8); Leverage=0.0032 (1.6)]
Holdout coverage = 13
Fails significant improvement with respect to lettuce and carrots & potatoes,
p = 0.0570

lettuce & tomatoes & carrots & potatoes
[Coverage=0.016 (8); Leverage=0.0062 (3.1)]
Holdout coverage = 11
Fails significant improvement with respect to lettuce & tomatoes & carrots and
potatoes, p = 0.0204

```

The itemset, tomatoes & potatoes provides a good example of the improvement test. Tomatoes occurs in 0.282 of all holdout records and potatoes occurs in 0.294 of all holdout records. If these items were independent of each other then tomatoes & potatoes would be expected to occur in 0.083 (41.45) of all holdout records. In fact they occur in 41 holdout records, and hence do not indicate any improvement.

12. Computation time, snapshots and anytime results

Magnum Opus provides tremendous flexibility to the user. Many forms of analysis can be requested, and **Magnum Opus** always provides exact results. However, some analyses are intrinsically difficult, and hence require large amounts of computation to complete. Unfortunately, it is not possible to accurately predict in advance which analyses will take extreme lengths of time to complete and which will complete quickly.

Fortunately, however, **Magnum Opus** is an *anytime* system. While it may take a long time to produce precise results, it often finds very good results very early in the search process, and these can be inspected at any time.

When a computation is taking a long time it is often helpful to view the best results discovered so far. This allows you to both assess whether you are actually performing the correct analysis and whether the results already obtained satisfy the analytic requirement. A set of intermediate results created while computation is in progress is called a *snapshot*. The following process is used to create a snapshot.

Command line: *While the system is running, send the SIGUSR1 signal to the process. The exact command required may vary depending upon the precise operating system and command shell used. The following provides an example under bash on Linux.*

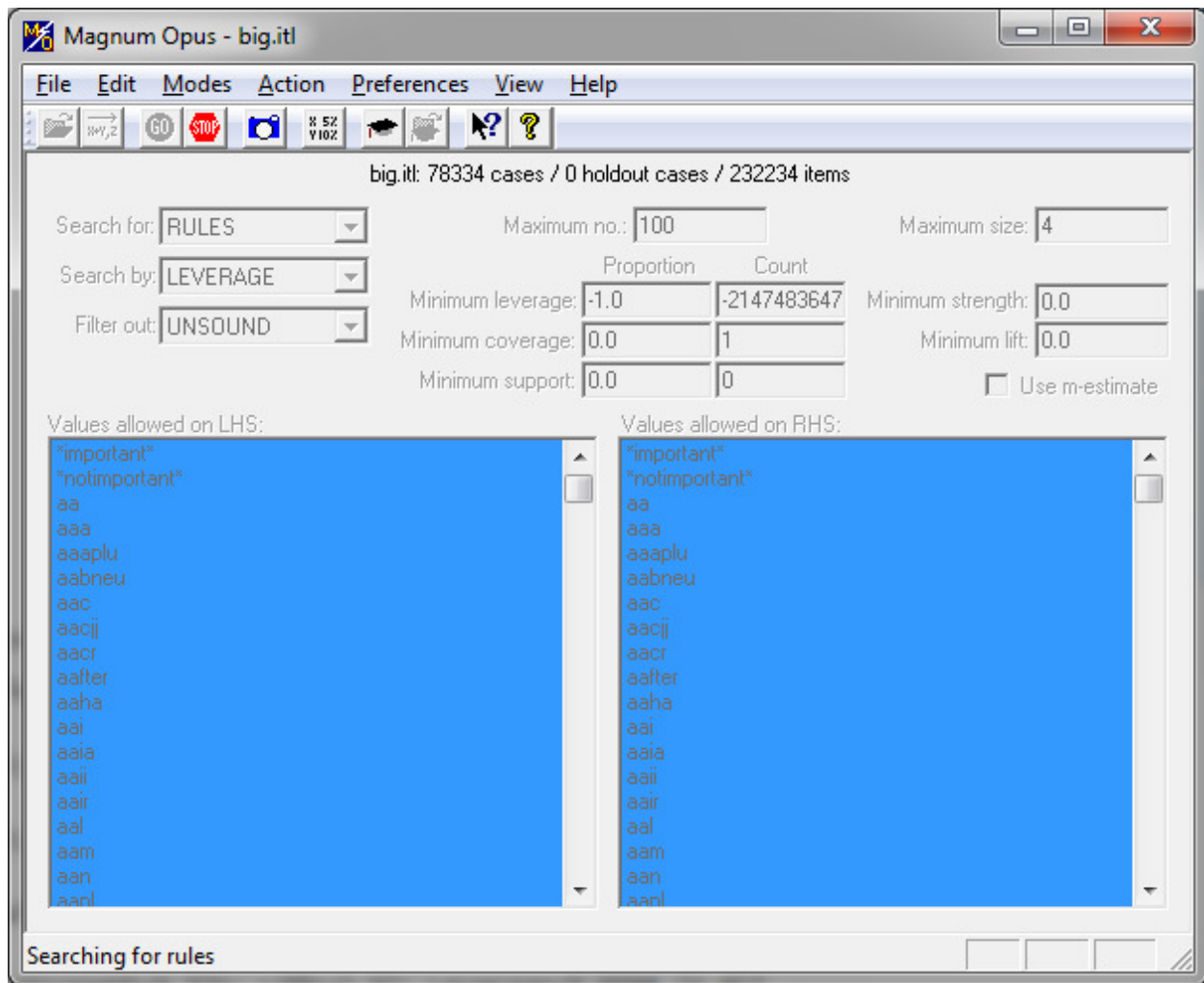
```

% mocl names-file=tutorial.nam data-file=tutorial.data > tutorial.out &
[1] 3342
% kill -SIGUSR1 3342

```

In this example the process has been run in the background and has been assigned the process ID 3342.

Interactive system. When the system is in the process of a search the screen will appear as follows:



Simply click on the blue camera icon. A dialog will appear that allows you to specify a file into which the snapshot will be saved.

In general the following actions will decrease compute time.

- Increase the minimum leverage.
- Increase the minimum coverage.
- Increase the minimum support.
- Decrease the maximum LHS length.
- Decrease the maximum number of rules to be found.
- Decrease the number of values allowed on the LHS and the RHS of rules.

Note, increasing the minimum lift or strength will only decrease compute time if use m-estimate is checked or the minimum coverage or support is set to a high value. Increasing minimum lift or strength when minimum coverage and support are both low can substantially increase compute time.

Search by lift and search by strength are both substantially faster when the m-estimate is used.

13. Some final thoughts

Magnum Opus is a powerful and flexible tool. The default settings are sufficient for many analytic tasks. However, advanced users can use the sophisticated controls to perform a wide variety of complex analyses. We recommend that new users start by using the default settings and only start using the other controls as they become familiar with the system.