

ARMADA



*Association
Rule
Miner
And
Deduction
Analysis*

User Manual

By James Malone

Contents

Introduction To <i>ARMADA</i>	Page 2
System Requirements	Page 2
Installing And Running The Software	Page 3
Familiarisation With The Program	Page 4
Getting Started	Page 8
Selecting A File To Mine	Page 8
Selecting Mining Criteria	Page 8
Using Rule Goal Builder	Page 10
Sampling the Data Set	Page 11
Beginning Mining	Page 12
The Mining Results	Page 13
Analysing The Rules	Page 13
Graphical Analysis	Page 14
Printing a Graphical Display	Page 16
Saving A Results File	Page 16
Opening A Results File	Page 16
Troubleshooting	Page 18
Hot-Key Summary	Page 19
Index	Page 20

ARMADA Support:

Email: James.Malone@Sunderland.ac.uk

©2003 *ARMADA* - Copyright James Malone

Introduction To *ARMADA*

ARMADA is a Data Mining tool that extracts Association Rules from numerical data files using a variety of selectable techniques and criteria. The program integrates several mining methods which allow the efficient extraction of rules, while allowing the thoroughness of the mine to be specified at the users discretion.

The name ARMADA stands for Association Rule Miner And Deduction Analysis. The program was designed as a tool to assist in the analysis of both the knowledge extracted and the deduction processes by which such a task is undertaken. However, the program can also be used as a straightforward Data Mining tool for the efficient extraction of Association Rules.

The actual knowledge extracted is presented in the form of easy-to-understand rules, while the details of the process, such as time taken and file size considered, are conveniently summarised in the 'Mining Report' section. These mining results can also be saved and opened for analysis in a 'dmr' (data mining results) file.

The program also allows the results to be displayed through various graphical representations, such as bar charts and line graphs. Such graphics can often help to summarise the knowledge being analysed by providing a concise conceptualisation of the data under scrutiny. A facility to print such graphics is also included.

Although the type of numerical data which ARMADA could be used to mine are virtually endless, common examples of data sets include;

- POS (Point of Sale) Transaction data
- Medical databases
- Census data
- Statistical data
- Lottery Results (not guaranteed to provide winning lines!)

System Requirements

The following are the minimum system requirements to run ARMADA;

- MATLAB[®] version 5.x or greater*
- 16MB memory
- 200 hard Disk Space (for installation of program)

Recommended system to assist with efficient mining;

- 32MB memory (or greater)
- 133MHz processor (or faster)

Installing And Running The Software

To install the software simply copy over the files from the disk to the directory marked 'MATLAB/Bin'. The program can then be ran by loading MATLAB and entering ARMADA at the command line prompt.

Familiarisation With The Program

The ARMADA system consists of two main parts:

The ARMADA Criteria Window. This is the initial, pre-mining part of the program which deals with specifying the criteria by which the mining process is going to be undertaken.

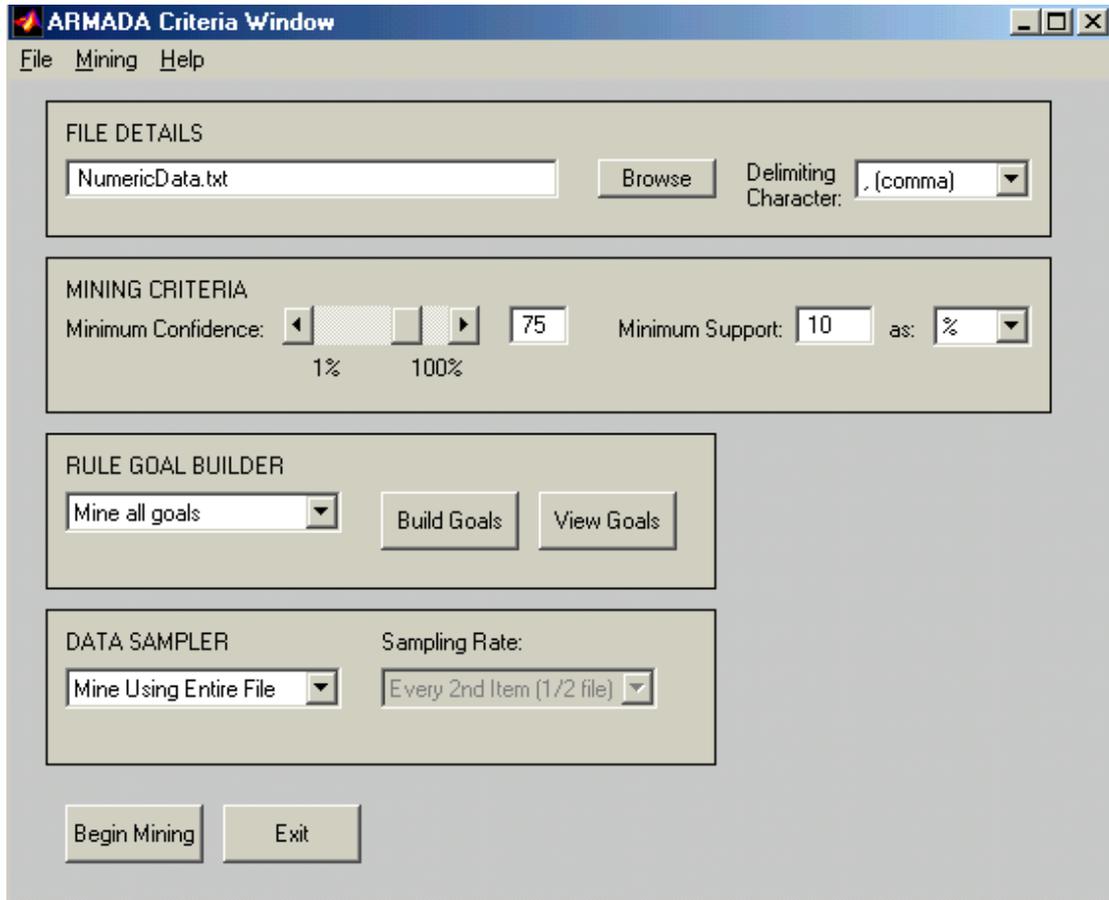


Figure 1. The ARMADA Criteria Window

This window can be broken down into four further parts:

The *File Details* section. This deals with the selection of the file related criteria, such as the file and path name and the delimiting character which indicates the character that separates one numeric item from the next within the file.

The *Mining Criteria* section. This deals with the specifying of two important attributes used to evaluate Association Rules – that of Minimum Confidence and Minimum Support.

The *Rule Goal Builder* section. This allows the creation and viewing of goals by which rules are mined.

The *Data Sampler* section. This section specifies the thoroughness by which the mining is undertaken, allowing the data set to be analysed in full, as a specified sample or as both for analysis purposes.

The bottom two buttons are used to begin the mining or to exit the program.

The menu bar along the top provides further options to open a results file, create a new mining criteria screen and open the help screen.

The Mining Results Window. This is the post-mining part of the program which displays the Association Rules that have been extracted and a report down the right hand side.

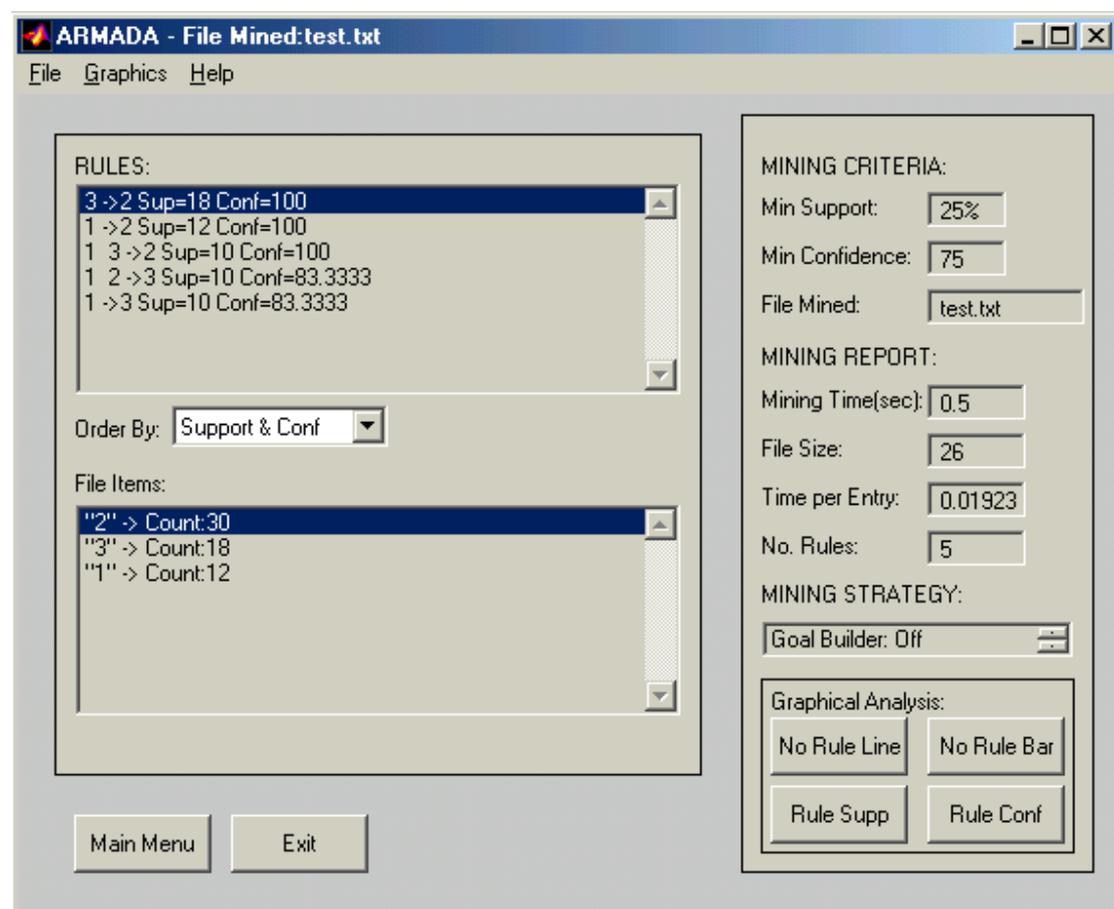


Figure 2. The Mining Results Window

Again, this can be further broken down into six parts:

The *Rules* section. This box displays all of the rules that have been mined using the specified criteria. The rules appear in the format of;

LHS *Item(s)* -> RHS *Item(s)* Sup = *number* Conf = *number*

The LHS (Left Hand Side), or antecedent, item(s) appear to the left of the '->' symbol. Multiple items are separated with a space. Similarly, the RHS (Right Hand Side), or consequent, items appear to the right of the '->' symbol. The support of the rule is represented as a numeric value after the 'Sup =' part. The confidence of the rule is represented as a numeric value after the 'Conf =' part.

The rules box can be sorted for displaying purposes by either support then confidence value of each rule or by the LHS size of the rules (1 LHS part rules at top).

The *File Items* section. This box displays all of the file items that are above the specified minimum support in descending order, i.e. the item with greatest support at the top of the list.

The *Mining Criteria* section. This displays the mining criteria that were specified.

The *Mining Report* section. This displays a report on the mining to summarise the process undertaken. The report includes important factors which can be of benefit when analysing the rules.

The *Mining Strategy* section. This displays the strategy that was specified. Specifically, it displays whether or not the Goal Builder was used, the Sampler was used and what Sample Rate was specified (if applicable).

The *Graphical Analysis* section. This allows various graphical summaries of the rules extracted and their attributes to be displayed by clicking on the relevant button.

The buttons to the bottom-left of the Window allow the main criteria window to be displayed or the program to be exited.

The menu bar along the top provides further options to save the results file being displayed, open a results file and open the help screen.

A further variation of the Mining Results Window is the **Analysis Results Window**. This is similar to the Mining Results Window in almost every way except that it displays two sets of rules, file items and mining reports. This window is displayed when the analysis mode is selected from the 'Data Sampler' options.

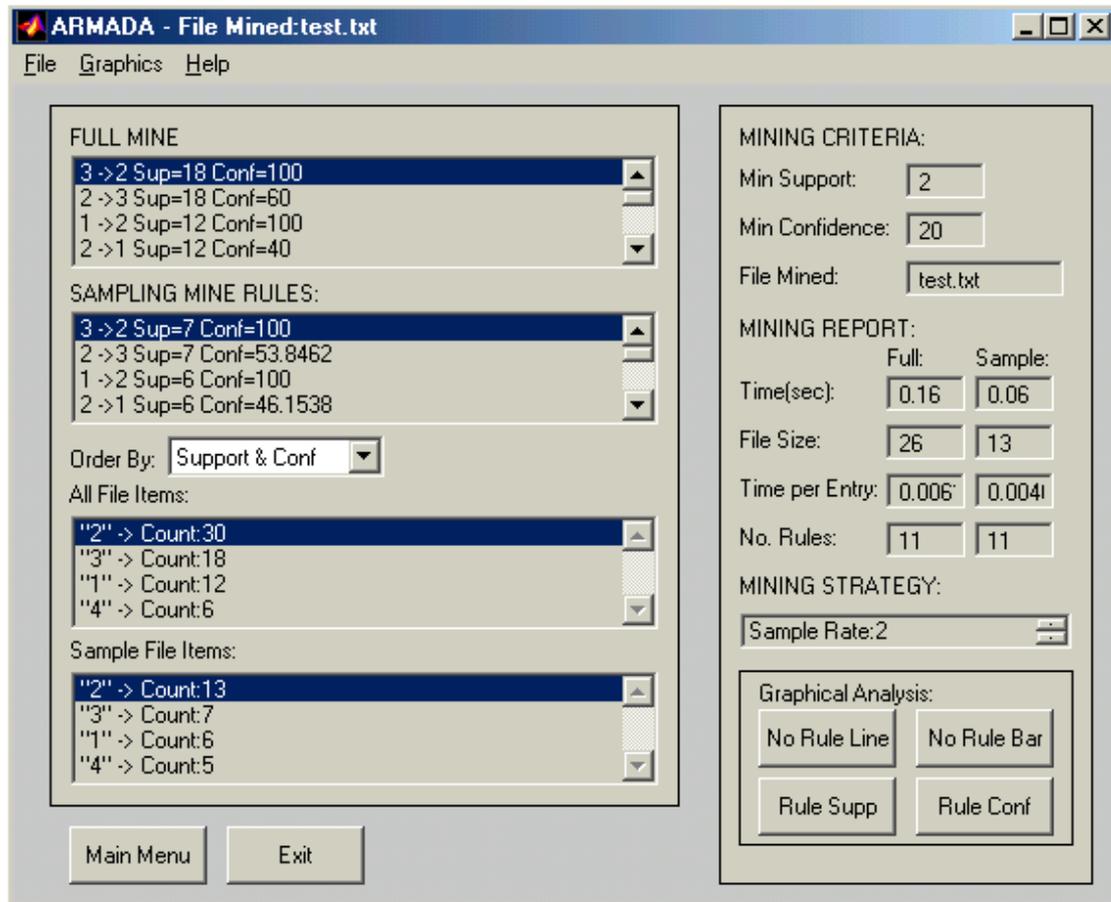


Figure 3. The Analysis Results Window

Getting Started

Selecting criteria to perform mining by is a task that may, at first, appear daunting. There are questions that arise when making the selections before mining which can not be awarded concrete answers. For example, what level of confidence is going to provide a 'useful' set of Association Rules, would the results be just as effective if the Data Sampler was used rather than a full mine, and so on.

Here lies the oxymoron that is Association Rule mining. The value of discovering specific and accurate knowledge from such data mining, is in the unknown quantities of the data set being mined. The simple answer to some of these questions, and more, is that there is no simple answer, at least before mining is undertaken.

For this reason, this section is meant as a guide to assist in the process of selecting criteria by which to perform mining, not as a set of concrete rules that must be followed every time in order to produce effective results.

Selecting A File To Mine

One part of the program that does adhere to specific guides is the selection of the file to perform mining upon. The file must contain numerical data, separated by one of five *delimiting characters*. A delimiting character simply represents a break between one item and the next. The five characters are;

- a comma (,)
- a semi-colon (;)
- a colon (:)
- a full stop (.)
- a space ()

To specify a file, the file name and path can be entered in the white box under the file details heading. NOTE: If a path is not entered, then the MATLAB default working path is used. The file can also be selected from a standard file open dialogue, which allows the traversal of the system's directories, by selecting the 'Browse' button.

Selecting Mining Criteria

One of the most difficult decisions that must be taken is the selection of the mining criteria, specifically, the two attributes that fall into this category – minimum Support and minimum Confidence. There are no hard and fast rules to selecting suitable values for either, however there are some pieces of information that can help when making a decision.

Firstly, the definition of what Support and Confidence are must be understood. **Support** is the number of times the items in a rule appear together in a single entry within the entire set. **Confidence** is the number of times that the LHS of a rule leading to the RHS is true within the data set.

So, if all the items in a rule appeared together 5 times in a data set with only 10 entries, then the support is 5 or 50%. If the LHS of a particular rule led to the RHS in 4 out of those 5 occurrence mentioned, then the confidence is 80%.

The next piece of information to have in mind is that, the lower the values for each of these two criteria, the more rules will be extracted. Therefore, the most rules will be extracted when the values for each are set to 1 (which is the lowest value permitted). Conversely, the higher the values of the two criteria, the smaller the number of rules that will be extracted. Therefore, the least rules will be extracted when the values are set to 100%.

In practice, the rule of thumb is that to extract all possible rules (sometimes called brute force mining) set the values to the lowest permitted. To extract only rules that apply to every entry in a data set, make the values for the criteria 100%. The latter, however, will usually extract no rules at all.

Using this theory, it would not be unreasonable to suppose that, by setting each criterion to 50%, exactly half of the total number of possible rules would be extracted. However, because of the nature of Association Rule Mining, this does not follow. This is because the rules being mined are dependant upon the items within the data set; the criteria are relative to relationships between items within the data not the overall results. 50% support means items that appear together in 50% of each entry in the data set, not 50% of the total amount of rules that can be extracted.

Although there is no single solution to extracting rules, using the above information, the following can be concluded and hence used as guidelines;

1. If a broad range of rules is required, a low minimum criteria should be selected
2. If a small number of highly occurring rules is required then keep criteria high.

One alternative approach is to begin mining with a very high criteria and, if the rules are not sufficient in number, repeat the process for a slightly lower threshold. This will allow mining to be performed until the number of rules are sufficient to an individual's needs. Of course, the pay-off here is that the mining process must be repeated until this goal is achieved which could be time-consuming.

Using Rule Goal Builder

One strategy that features within ARMADA is that of Rule Goal Builder. This allows rules to be mined which contain specific goals that are determined before the mining process is undertaken. This strategy can help act as a filter when examining rules and is particularly useful when the interest is in the relationships help for a specific item or set of items.

Defining goals is a relatively simple process. By selecting the 'Build Goals' button on the Mining Criteria Window, the Rule Builder Window is now displayed (figure 4).

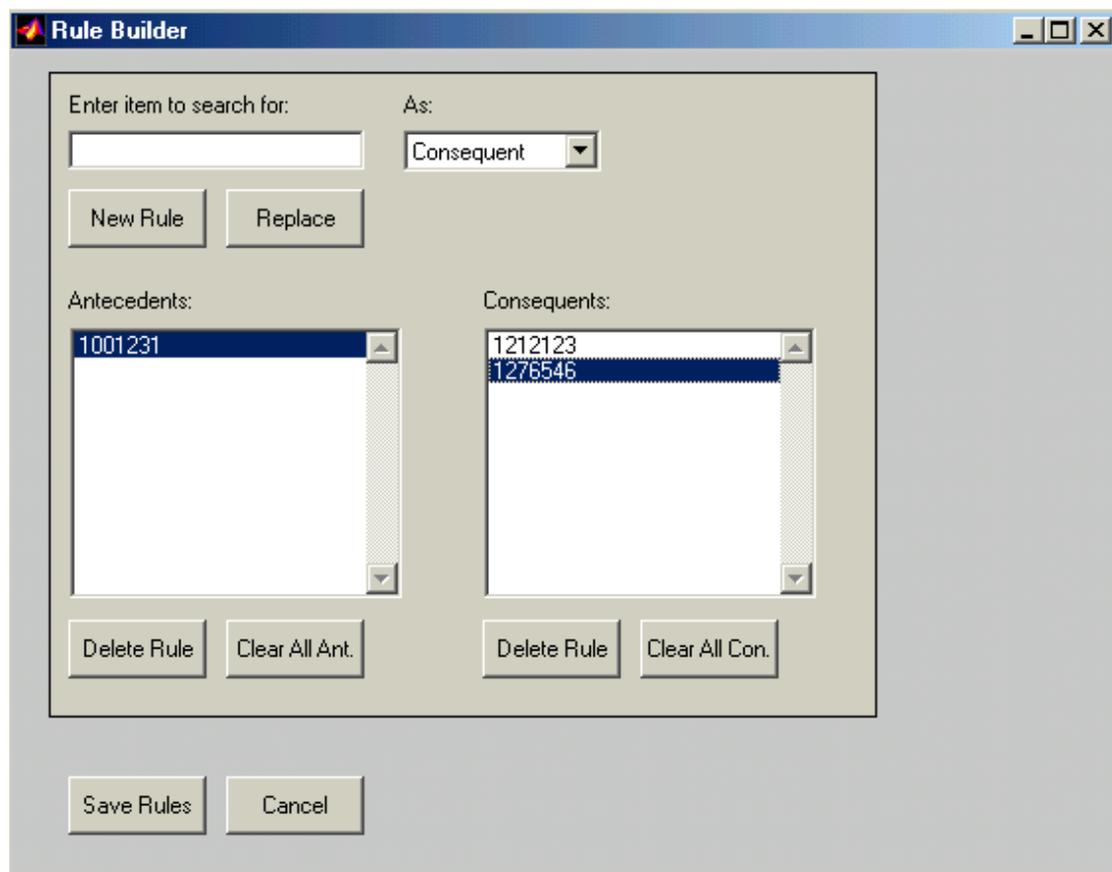


Figure 4. Rule Builder Window

Goals can be entered in the box below 'Enter item to search for:' and added to either the Antecedent box or the Consequent box, depending upon the selection made in the drop-down menu below the 'As:'. An item can be added to the list by selecting the 'New Rule' button or replace an existing item in the list by selecting the 'Replace' button. When replacing a goal, the currently selected item in the relevant goal box will be replaced, indicated by the blue bar. In the above example (Figure 4), the only item in the Antecedent box is selected. The second item down is the item selected in the Consequent box.

To remove an item, select the goal from the box and click on the 'Delete Rule' button below the relevant box. To remove all the rules in a particular box select the appropriate 'Clear All' button.

Once goals have been built they can be saved by pressing 'Save Rules' which stores them in memory so they can be utilised if selected when mining begins. Alternatively, any changes that have been made since the window was opened can be disregarded by selecting the 'Cancel' button.



Figure 5. Selecting Goal Builder

Once rules have been defined, they can be used to target the Association Rules that are mined. This is achieved by selecting the option 'Mine using built goals' from the drop-down box shown above (Figure 5).

NOTE: the default option for Rule Goal Builder is set to 'Mine all goals', which does not exclusively consider those goals that may have been defined. A common error made when mining is undertaken is that of defining a set of goals, but not changing the drop-down menu to 'Mine using built goals'. The lesson to learn here is *be aware of the option selected before beginning mining*.

Sampling the Data Set

The final strategy that can be used to undertake mining is that of *data sampling*. This option allows the specification of just how thorough the mine being performed is, with regards to how many entries within a data set that are looked at during the process.

A 'normal' mine for Association Rules is generally considered as extremely thorough – analysing the entire data set. However, when constraints play a role, for example time, thoroughness is not always the primary concern. The technique of data sampling is one which takes such constraints into consideration and allows a reduction of the thoroughness of the mine, to decrease the time that the mining takes.

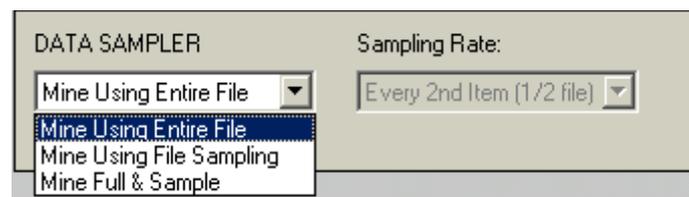


Figure 6. Selecting The Data Sampler

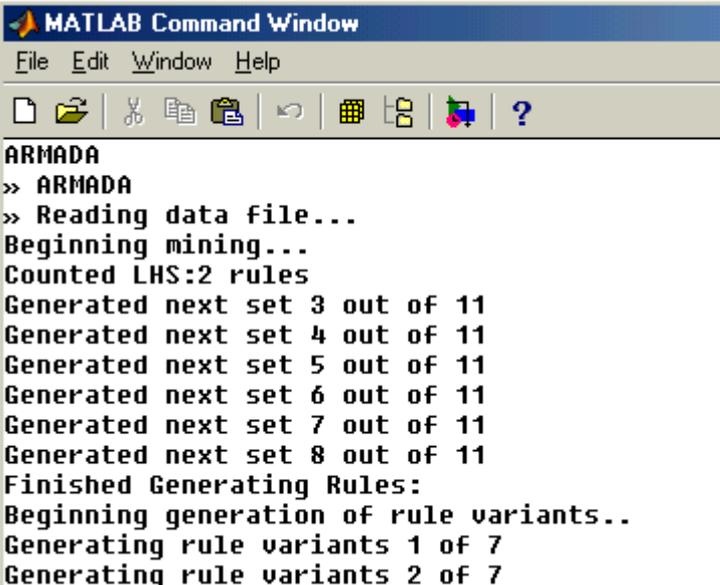
The options available in Data Sampler are three-fold. The data sample can be set to either 'Mine Using Entire File', which mines all of the data set, 'Mine Using File Sampling', which mines a sample of the data set (specified by 'Sampling Rate' drop-down) or 'Mine Full & Sample' which performs the mine in analysis mode. The first two options are relatively self-explanatory, however the third may not be so obvious.

The 'Mine Full & Sample' option performs a mine using the full data set – but also mines using the sample set. The rules extracted are then presented in the 'Analysis Results Window' (see page 6) which displays both of these results and allows a comparative analysis to be performed. A particularly useful feature of this method is the graphical displays which plots the values for both sets of results. This is an invaluable aid when comparing the two strategies and helps to provide some of the value to ARMADA as an analysis tool as well as a Data Mining program.

Beginning Mining

Once all the criteria have been selected, mining can then begin. To start the process simply click on the 'Begin Mining' button or select Mining->Begin Mining from the menu bar. Alternatively, the hot-key 'ctrl+B' can be used.

If the data file can not be read because it is invalid in some way or any of the criteria entered are invalid, then an error message will be displayed. Otherwise, the process of mining begins. The stage of the mining process that ARMADA is currently at can be obtained by viewing the MATLAB Command Window screen (Figure 7). This can help to give some idea of where the program is at and therefore how long it may have until completion with regards to the number of phases remaining.



```
MATLAB Command Window
File Edit Window Help
ARMADA
>> ARMADA
>> Reading data file...
Beginning mining...
Counted LHS:2 rules
Generated next set 3 out of 11
Generated next set 4 out of 11
Generated next set 5 out of 11
Generated next set 6 out of 11
Generated next set 7 out of 11
Generated next set 8 out of 11
Finished Generating Rules:
Beginning generation of rule variants..
Generating rule variants 1 of 7
Generating rule variants 2 of 7
```

Figure 7. MATLAB Command Window

The Mining Results

Arguably the most important phase of Data Mining is the analysis and subsequent conclusions made from the results. This section aims to assist in the understanding of the mining results by describing how they are represented and how interpret the graphical displays.

Analysing The Rules

The rules are displayed in a straightforward manner which is relatively self-explanatory. The rules displayed in the Rule Box appear in the format of;

LHS *Item(s)* -> RHS *Item(s)* Sup = *number* Conf = *number*

An example of such a rule could look like;

1234 2345 3456 -> 4567 Sup=10 Conf=70

This would translate to mean that the items '1234', '2345' and '3456' lead to '4567' with support of 10 and confidence of 70.

The displayed rules can also be sorted in order of either their support and confidence (highest top of the list) and by the six of the LHS part of the rule (with one part LHS rules top of the list).

The file items list is also provided to give some insight into the data set that has been mined. These items are ordered by their support (highest top of the list).

The criteria that were specified are also displayed in the window. Minimum support is displayed as a number alone if it was specified as such, or, if it was specified as a percentage of the data set, then a '%' sign follows the value. Minimum confidence is always expressed as a percentage.

The mining report section displays;

- the total time that mining took
- the file size that was analysed for mining
- the time per entry, which is the time taken to mine each rule
- the number of rules that have been extracted from the mining

The mining strategy box summarises the strategies that were selected to perform mining by.

Graphical Analysis

Another useful feature of ARMADA is the ability to summarise the knowledge extracted graphically. These graphics come in four parts:

Number of Rules Line Graph. This plots a graph of the size of the LHS of the rules against the no of rules extracted. This helps to show the proportions of rule numbers for varying rule sizes. (Figure 8)

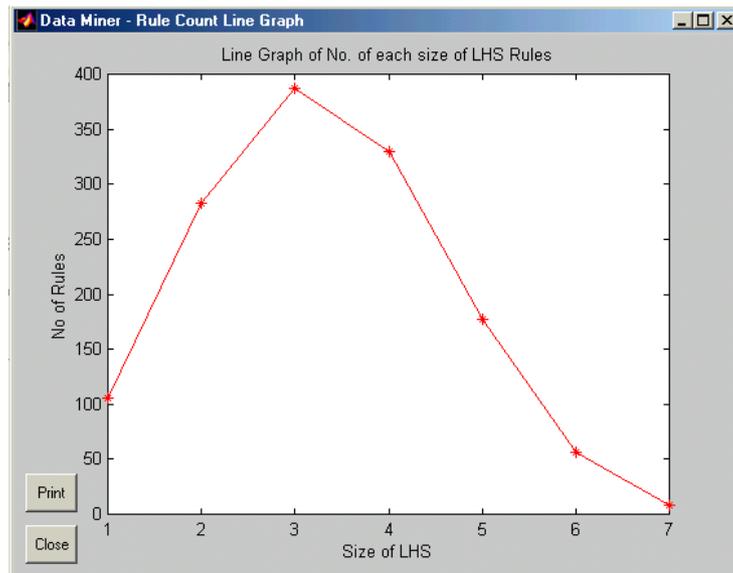


Figure 8. No of Rules Line Graph

Number of Rules Bar Chart. This plots a bar chart of the size of the LHS of the rules against the no of rules extracted. This helps to show the proportions of rule numbers for varying rule sizes. (Figure 9)

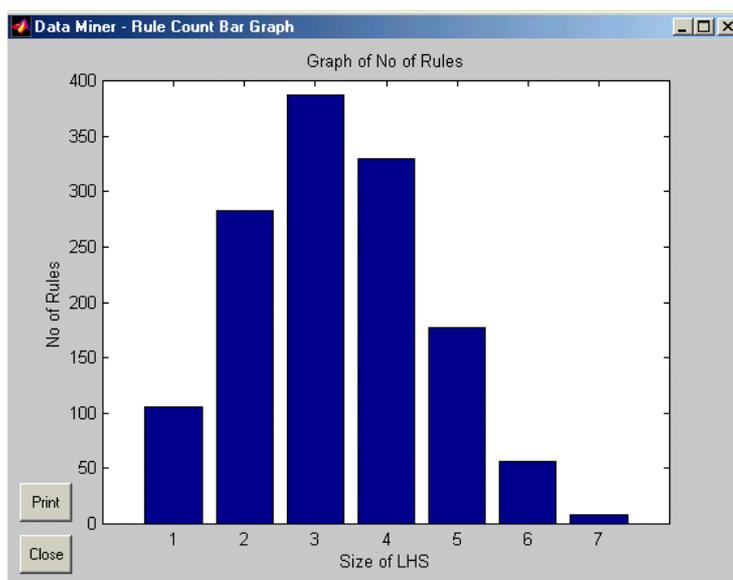


Figure 9. No of Rules Bar Chart

Rule Support Line Graph. This plots a line graph of the support of the rules against the number of rules, starting from the highest support as the left most value down to the lowest at the right most value. This graph often describes a 'waterfall' effect when analysing the rules as the support decreases or a straight horizontal line if the support is constant throughout. (Figure 10)

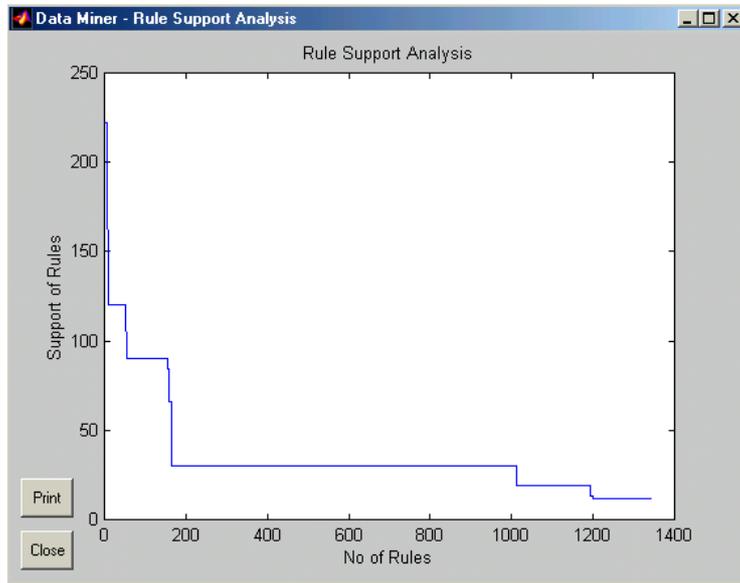


Figure 10. Rule Support Line Graph

Rule Confidence Line Graph. This plots a line graph of the confidence of the rules against the number of rules, starting from the highest as the left most value down to the lowest as the right most value. Again, this graph often describes a 'waterfall' effect when analysing the support of the rule or a straight line if the confidence is constant throughout. (Figure 11)

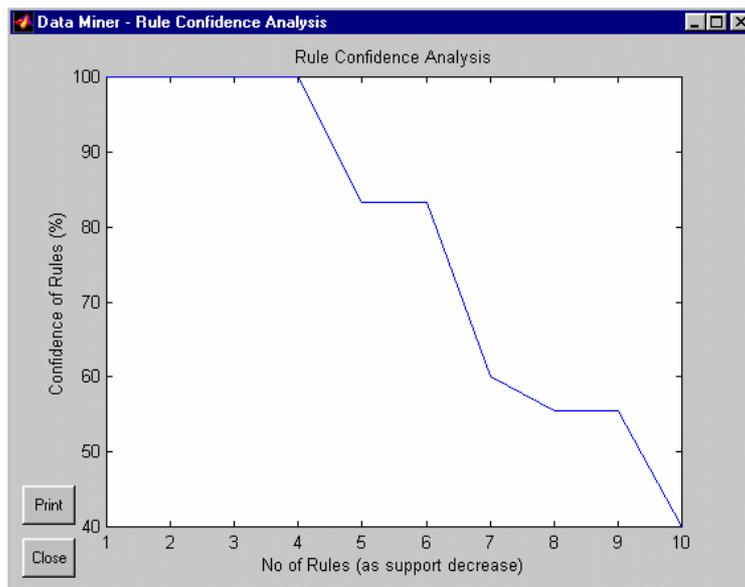


Figure 11. Rule Confidence Line Graph

Printing A Graphical Display

Any of the graphical displays can be printed. To print a graph simply open up the required display and click on the 'Print' button. This will print out the current graphic using the current default printer and it's default settings. Therefore, if a coloured print-out is required make sure the current default settings specify this.

Saving A Results File

Once mining has finished and the results are being displayed they can, if required, be stored for later analysis. ARMADA results files are stored in files with extension '.dmr' (Data Mining Results) and can be given any name that is valid within the current operating system.

To save a file:

1. Select File->Save from the menu bar or use hot-key 'ctrl+s'. This brings up the save file dialogue.
2. Enter a file name in the 'File Name' box. NOTE: there is no need to enter .dmr after the file name as ARMADA will automatically do this.
3. Click on the 'save' button. If a file name is enter for which a file already exists the option to cancel or to proceed and overwrite the current contents of this file will be given. Alternatively, a file can be selected from the displayed folder to save the results to, however this will overwrite the current contents of the selected file.

Opening A Results File

To open a mining results (dmr) file:

1. Select File->Open from the menu bar or use hot-key 'ctrl+o'. This brings up the open file dialgoue.
2. Enter a file name in the 'File Name' box or select a file from the contents of the folder being displayed in the window.
3. Click on the 'open' button. If the file name is invalid or is as of an invalid format then an error message is displayed.
4. If the file is valid, the results file will now be displayed in a Results Window.

NOTE: When a results file is opened, the current mining results being displayed are lost. Therefore, if these results are required for later analysis ensure they have been saved first, before opening a new file.

Troubleshooting

This section contains a list of common of common problems with the possible reason(s) for the problem occurring and the action to take to remedy them.

Problem	Possible Reason	Action
Program won't run after installation.	Installation has not been completed successfully.	Re-install program files.
	MATLAB version is not compatible with ARAMADA.	Install MATLAB version 5 or above.
	Command to run the program is being entered incorrectly.	Check that the command being entered is 'ARMADA'. Check that it is in uppercase (some versions of MATLAB are case-sensitive).
	System does not meet minimum requirements.	Check minimum requirements section in manual. Upgrading system to minimum requirements is only solution if this is the case.
Program won't begin mining – displays an error message.	File has not been specified.	Enter a file name in file details box.
	Minimum support is invalid.	Enter a numeric value ≥ 1 if as a No., or in range $\geq 1 \leq 100$ if as %.
	File does not exist.	Make sure file specified is one which exists.
	File is invalid.	Check that the correct file has been specified. Check that delimiting character is the correct one for the file specified. Check that data is numeric values only.
	No goals have been defined and option is set to mine using built goals.	Switch off goal builder or define some goals.
Using full mine - no rules were extracted from mining.	Minimum criteria are too high.	Decrease minimum support and/or decrease minimum confidence.

Problem	Possible Reason	Action
Using Goal builder mine - no rules were extracted from mining.	Minimum criteria are too high.	Decrease minimum support and/or decrease minimum confidence.
	Goals specified are not contained within data set.	Check that goals built are items within the data set.
Mining process is taking long time to complete.	Minimum criteria are too low.	Increase minimum support and/or increase minimum confidence.
	File being mined is large.	Try reducing entries looked at by using data sampler.
Too many rules are being extracted.	Minimum criteria are too low.	Increase minimum support and/or increase minimum confidence.
Results file will not open.	File entered is invalid.	Reselect a valid '.dmr' file.
	File specified is corrupt.	Usually no cure for this problem. Some diagnostic tools such as 'Disk Doctor' <i>may</i> be able to rescue the data.
	File is in 'resource locked'.	Free file from other user's resources.
Graphics will not print to printer.	Printer is switched off.	Switch printer on. Check power cable is connected correctly.
	Printer is not connected.	Check printer is connected correctly.
	Default printer is not installed.	Check a default printer is installed. Install driver updates from manufacturers.
Graphics will not print in colour.	Default settings are set to black only.	Change default settings to print in colour.
Operating system crashes when running program.	Memory resources are exhausted.	Check memory meets minimum requirements. Close down other programs running in background. Upgrade memory.
	Operating system problem.	Re-install operating system. Contact operating system manufacturer.
	Virus infection.	Run virus checker.
	ARMADA program files are corrupted.	Re-install ARMADA program. Contact ARMADA support.

Hot-Keys Summary

Action	Hot-Key
New mine screen	Ctrl + n
Open results file	Ctrl + o
Save results to file	Ctrl + s
Exit program	Ctrl + x
Begin Mining	Ctrl + b
Help Contents	Ctrl + h

Index

A

Analysing the rules; 13
Analysis mode, mining in; 12
Analysis Results Window; 6

B

Beginning mining; 12

C

Confidence, definition; 9
Confidence line graph; 15
Criteria, selecting; 8
Criteria Window; 4

D

Delimiting character; 8
DMR file; 16

F

File details; 8
File, valid formats; 8
Full mine; 12

G

Getting started; 8
Goal builder; 10
Graphical analysis; 14

H

Hot-Key summary; 19

I

Installing the software; 3

L

LHS; 13

M

MATLAB Command Window; 12
MATLAB requirements; 2
Memory requirements; 2
Mining criteria section; 8

N

No. of Rules Line Graph; 14
No. of Rules Bar Chart; 14

O

Opening a results file; 16

P

Printing a graphic; 16

R

Results Window; 5
RHS; 13
Rule Builder Window; 10
Running the software; 3

S

Sampling the data set; 11
Saving a results file; 16
Support, definition; 9
Support Line Graph; 15
Support, contact details; 1
System requirements; 2

T

Troubleshooting guide; 17

U

Uses, examples of; 2