TUTORIAL

WEKA:
The Waikato Environment for Knowledge Analysis

CONTENTS

TABLES

FIGURES

Introduction

What is Machine Learning?

One of the central problems of the information age is dealing with the enormous amount of raw information that is available. Machine learning has the potential to sift through this mass of information and convert it into knowledge that people can use. In general, machine learning enables a computer to automatically analyse a large (or small) body of data and decide what information is most relevant. This crystallised information can then be used to help people make decisions faster and more accurately.

WEKA

WEKA, the Waikato Environment for Knowledge Analysis, is an experimental software "workbench" incorporating several standard machine learning techniques. With it, a specialist in a particular field is able to use machine learning to derive knowledge from databases that are far too large to be analysed by hand.

The WEKA workbench acts as an interface to machine learning schemes and datasets. The workbench also provides access to two tools for analysing data — the Machine Learning Workbench and the Experiment Editor.

The Machine Learning Workbench

The Machine Learning Workbench is used to produce rules and decision trees based on the current dataset. These rules and decision trees are able to be used in application to the domain being studied. The Machine Learning Workbench is provided for domain experts, such as agriculturalists or horticulturalists, who are interested in learning more about their data. The Machine Learning Workbench also enables machine learning experts to learn how to analyse datasets more efficiently and effectively on different machine learning schemes.

The Experiment Editor

The Experiment Editor is an analysis tool provided for machine learning experts who want to discover the classification accuracy and speed of schemes on different datasets; and who also want to create new schemes and analyse them through the Experiment Editor for performance measures. The Experiment Editor is an analysis tool used to evaluate ways for new and existing schemes to perform more effectively.

The Attribute Editor

A third tool, the Attribute Editor, is also provided on the Workbench. The Attribute Editor adds another dimension to data analysis, by changing the existing dataset in some way. This may involve creating a new attribute or selecting a subset of the existing data. The Attribute Editor adds another dimension to data analysis.

Machine Learning Schemes

The WEKA Workbench provides access to a number of different schemes; this number is continuously growing, giving the user more ways to analyse their data.

There are currently 11 machine learning schemes, on the WEKA 2.1 Workbench. Six of these — 1R, T2, C4.5. InductH, M5 and FOIL can be used by either a domain or machine learning expert to produce useful rules or decision trees. The rules or decision trees can be used as the basis for an expert system, or purely as one of many ways of analysing a dataset.

All of the schemes can be used for analysis with the Experiment Editor; where classification accuracy and efficiency, rule and decision tree sizes, and the different algorithm performances can be continuously tested by many different datasets.

The schemes used in this tutorial are:

- 1R, which generates a very simple rule that chooses just one attribute as the criterion for the current decision being made;
- T2, which calculates decision trees with one or two levels and is guaranteed to find a close to optimal 2-level decision tree of a given complexity;
- C4.5, which generates both rules and decision trees (the decision trees generally have more than two levels) and can both be read as ‘if-then......else-if......else’ statements culminating in a decision;
- InductH, which produces either independent rules (read with an OR between each given rule) or tree-like rules (with a two-way (binary) test at each node);
- FOIL, which produces more complex rules which deal with relationships between different attributes, as well as each individual attribute.

Each machine learning scheme takes the data in a datafile and ‘runs it through the scheme’. The datafile is required to be in a specific format — Attribute Relation File Format (ARFF). ARFF specifies the relation name, and each attribute with its type, and the data in the file. An example of the ARFF file format is in Appendix 2, where the golf and breast cancer datasets are listed.

Sometimes it is not sufficient to use only one datafile to test a scheme’s accuracy. In this case, a training dataset, is used to build rules or decision trees based on the dataset’s information, and a testing dataset is used to test the rules or decision trees for initial accuracy.

Using this Tutorial

This tutorial is designed as a guide or walk-through to analysing a dataset through the WEKA workbench. The tutorial is divided into four sections: Setting the Experimental Goals; Pre-analysis; Machine Learning Workbench Analysis; and The Experiment Editor.

Two example datasets are used in this tutorial.

1. Golf, smaller and less complex, contains weather information for a golfer and is used as the main example dataset throughout the tutorial
2. Cancer, larger and more complex, is based on breast cancer information (and also includes access to a test dataset) and its analysis is included in the appendices as a more complex analysis example.

Using the Mouse or Keyboard

The predominant input device is the mouse. The keyboard can also be used for selecting options from menus. Short-cut keys are given next to the menu item. For example, a user may open a training dataset by entering:

Ctrl+T

Typographic Conventions

The following typographic conventions are used throughout this tutorial.

When a control command, such as a menu command, text button, radio button, menu name or check box, has a name which you are asked to access, its name is given in **bold** text.

Text commands are given in `Courier` font.

Window and dialog box names are in regular text, with an initial capital letter; for example, the Open Training Dataset dialog box.

File names, directories, and record files appear as UPPERCASE text.

The User Manual

A user manual is also provided with this tutorial. The user manual should be used as a support to this tutorial. It explains in greater depth many of the tools for analysing a dataset using the WEKA workbench.

Tutorial Version

This tutorial and its user manual are written for WEKA 2.1 (30 August, 1996).

Setting the Experimental Goals

You will perform the following analysis processes when you analyse the golf dataset on the WEKA workbench.

Process 1:	Experimental design
Process 2:	Pre-analysis
Process 3:	WEKA analysis
Process 4:	Result interpretation

Experimental Design

Experimental design has two parts: defining the problem in machine learning terms and defining the objectives of the experiment.

Defining the Problem in Machine Learning Terms

This is similar to defining the problem for statistical analysis. The following problem definitions provide sufficient information about what the golf machine learning experiment should achieve.

Produce rules or a decision tree about which days a golfer may play golf.[1]

Defining the Objectives of the Experiment

The objectives for the golf machine learning experiment should be based on the type of output required which the user can apply to the domain. The output is determined by the type of scheme used and the options that are selected for the schemes.

1R produces rules, testing their performance through analysis. This scheme will generate a simple rule stating which attribute has the most effect for deciding whether golf should be played or not. 1Rw is chosen as it trains the entire dataset (the golf dataset has only 14 instances), and reports how accurate it has been.

T2 produces both rules and a decision tree. As the golf dataset is simple (it has few instances and attributes). T1 can be selected to indicate the depth of the tree required. This will produce a 1R like classifier.

C4.5 builds on 1Rw and T1, and produces both rules and a decision tree. The expected decision tree will probably be small in structure (the dataset is simple). The rules and trees are usually easy to read and effective in making decisions based on the data in the dataset.

[1] Further goals may be added.

Generally, good results can be produced by leaving all other options as their default values. Time complexity measures will not be relevant due to the size of the dataset (it will be expected to be analysed quickly), so internal evaluation will be used.

InductH produces rules (read as if-then-else statements). Selecting DNF will produce a list of independent rules — where you may apply either one or another of the rules. All of the rules will cover the entire dataset.

FOIL also produces rules. These rules are more complex in structure that either 1Rw, T1, C4.5 or InductH. FOIL produces rules with relationships between attributes.

Relevant goals and objectives for the golf analysis experiment would be:

Using the attribute ‘class’ as the class, produce rules or a decision tree about which days a golfer may play golf from the following schemes.

SCHEME	CONFIGURATION
1R	with 1Rw options
T2	with T1 options
C4.5	with tree and rule output, internal evaluation, for analysis on the Machine Learning Workbench, and default settings for analysis with the Experiment Editor.[2]
InductH	with default options
FOIL	with default options

[2] Output from the Experiment Editor is not graphical; therefore, it will not be possible to output a decision tree or rules.

Pre-analysis

Pre-analysis consists of:

- investigating the original golf dataset for missing values, inconsistent values, inaccurate values;
- identifying basic distributions between one, two or more of its attributes; and
- converting the golf dataset into the required machine learning format.

Investigation of the original golf dataset involves looking for:

Missing Values

These are either:

a) unknown values, replace with a '?'
b) unrecorded values, replace with a '?'
c) zero values, replace with either 0 for an integer value or 0.0 for a real value.[3]

OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	CLASS
sunny	?	85	FALSE	'Don't Play'
sunny	80	?	TRUE	'Don't Play'
overcast	?	78	?	'Play'
rain	70	?	FALSE	'Play'
rain	68	80	FALSE	'Play'

Table 1: Golf Dataset with Missing Values

The class value is used to identify the decision made about a particular instance. If the an instance's class value contains a missing value, then the instance should be removed from the dataset.

Inconsistent Values

Inconsistent values usually occur in symbolic variables where spelling may vary; for instance, in Table 2 the attribute 'outlook' has the value 'sunny' spelt in two different ways.

OUTLOOK
suny
sunny
overcast
ran
rain

Table 2: Outlook with Inconsistent Values

[3] Integer values do not have decimal place values; real values can have decimal place values (both integer and real valued attributes are known as 'continuous').

Inaccurate Values

Inaccurate values (or outlying values) can be found by graphing one variable at a time. An inaccurate value, will significantly deviate from the pattern formed from the remaining values. However, sometimes inaccurate values are hard to find, and domain knowledge is required to know if the value is incorrect.

For instance, the attribute 'temperature' is graphed below. Looking at the non-broken line on the graph, there seem to be no obvious outlying or inaccurate values.

However, if the non-broken line was replaced with the values represented by the dotted line, it would be more obvious that an outlying value had occurred as it is significantly different from the remaining values.

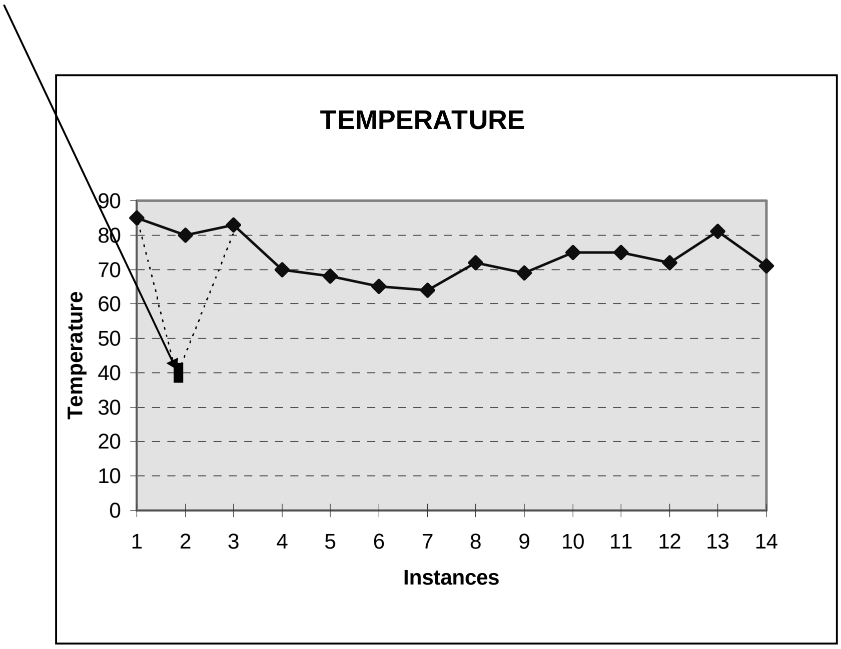


Figure 1: Temperature Distribution

Identifying Basic Distributions

Each attribute within a dataset can be graphed to view the general distribution of the data. Distributions of the data give information that can be used later for creating new attributes or a new selection of the existing data. For instance, the attribute Outlook is graphed by a bar chart, which gives a general distribution of its values.[4]

[4] 'Outlook' is a symbolic (discrete) variable. Graphing this attribute involved counting the occurrence of each value and entering this into another table, then graphing this. Graphing a numerical (continuous) variable involves sorting the dataset on that variable and graphing the values for each instance (an example of this is the graph of the 'Temperature' variable.

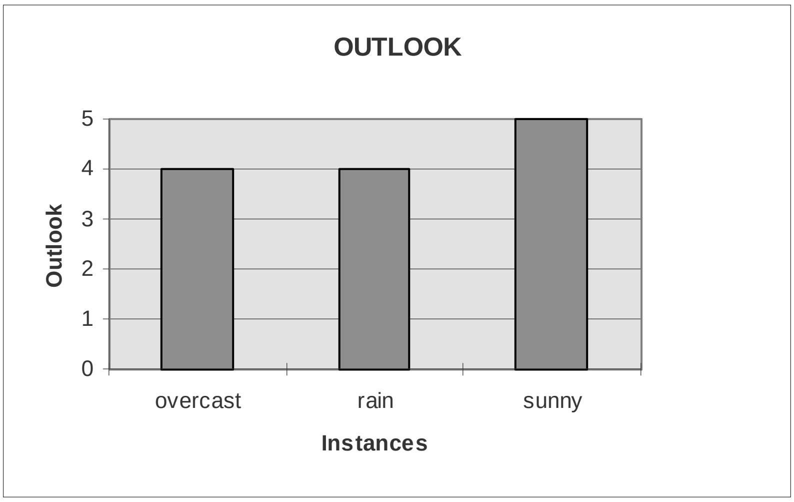


Figure 2: Outlook Distribution

Other basic pre-analysis features such as discovering relationships can also be graphed.

Converting the Dataset to Machine Learning Format

Each datafile is required to ARFF format — which specifies the relation name, and each attribute with its type, and the data in the file.

Converting your dataset to an ARFF may involve six stages. These include loading your data onto a PC or Macintosh spreadsheet package, saving the file in comma separated form (.csv), then loading it into your machine with WEKA installed.

Looking at Your Original Dataset

WEKA assumes the dataset is in a vertical format with field titles across the top and data beneath.

OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	CLASS
sunny	85	85	FALSE	'Don’t Play'
sunny	80	90	TRUE	'Don’t Play'
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	'Don’t Play'

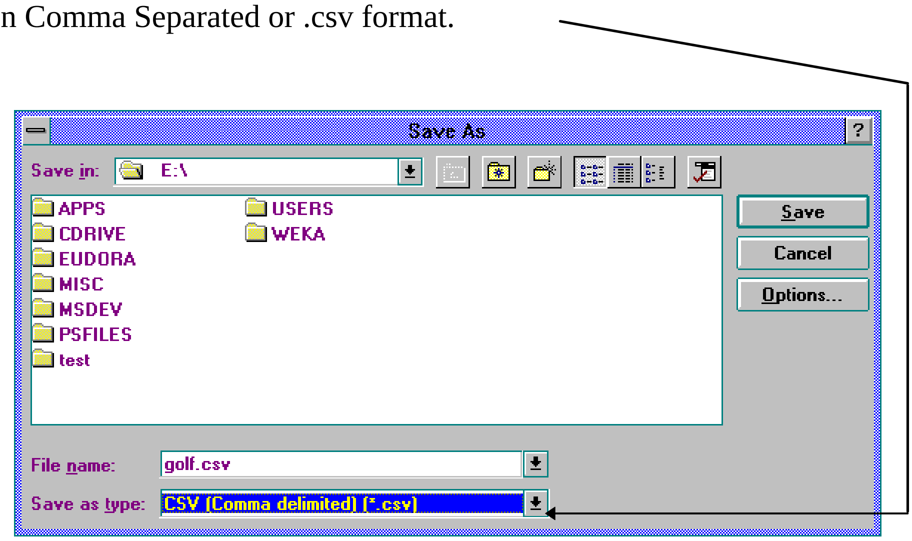
Table 3: Golf Dataset

If you do not currently have your data loaded into a spreadsheet, load it now.[5]

[5] If you have more than one row of title information then the converter will treat the additional title rows as data.

The Conversion Process

1. Using the **'Save As'** option on your spreadsheet, save your dataset in Comma Separated or .csv format.



Save in: E:\

File name: golf.csv

Save as type: CSV (Comma delimited) (*.csv)

Figure 3: MS-Word 7 — Saving golf.csv

2. FTP or copy your code to the machine with WEKA installed

3. Run the csvtoarff program as follows:

```
csvtoarff filename.csv > filename.arff
```

for example:

```
csvtoarff golf.csv > golf.arff
```

4. This will classify each column as enumerated (discrete or 'string') or as a real or integer type.[6] Check the filename.arff is a valid ARFF file:

```
arffinfo < filename.arff
```

for example:

```
arffinfo < golf.arff
```

For the golf dataset, the following result should be produced:[7]

[6] If an enumerated column has integer values, the user can force the column to be enumerated using the -e switch: `csvtoarff -e <n> dataset.csv > dataset.arff`. This would force the 'Class' column to be an enumerated type.

[7] If any warnings are displayed, edit the file and attempt to correct the problem; for example, change a real value to an integer, or an integer to an enumerated (symbolic) value, etc...

```
      relation_name: golf
      relation_attributes: 5
      relation_instances: 14
```

5. Your file can now be loaded onto the WEKA workbench.[8]

[8] If you have only one dataset, and would like to split it into a training and testing set, use the arffsplit program: `arffsplit -p66 < filename.arff train66.arff test34.arff`. This will randomly divide the filename.arff into two files, with 66% in the training dataset and 34% in the test dataset.

WEKA Analysis

A review of your goals and objectives will indicate what WEKA tools (Machine Learning workbench, Attribute Editor or Experiment Editor) you will be using to analyse your dataset:

- if you want to output either rules or decision trees then use the Machine Learning Workbench;
- if you want to alter your existing dataset in some way, then use the Attribute Editor; or
- if you want to learn how fast and accurate a number of schemes are over a number of datasets then use the Experiment Editor.

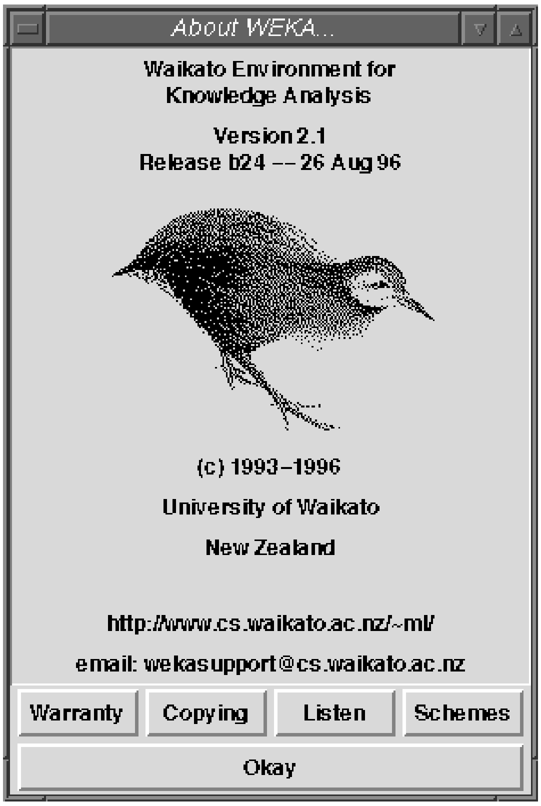
Starting WEKA 2.1

Before running any analysis experiment you need to open WEKA.

To do this on an X-terminal type:[9]

```
weka2.1
```

After a short time the Machine Learning Workbench window will appear. In front of it the About WEKA... window will also appear; which gives you options for learning more about the workbench.



Waikato Environment for
Knowledge Analysis

Version 2.1
Release b24 -- 26 Aug 96

(c) 1993-1996

University of Waikato

New Zealand

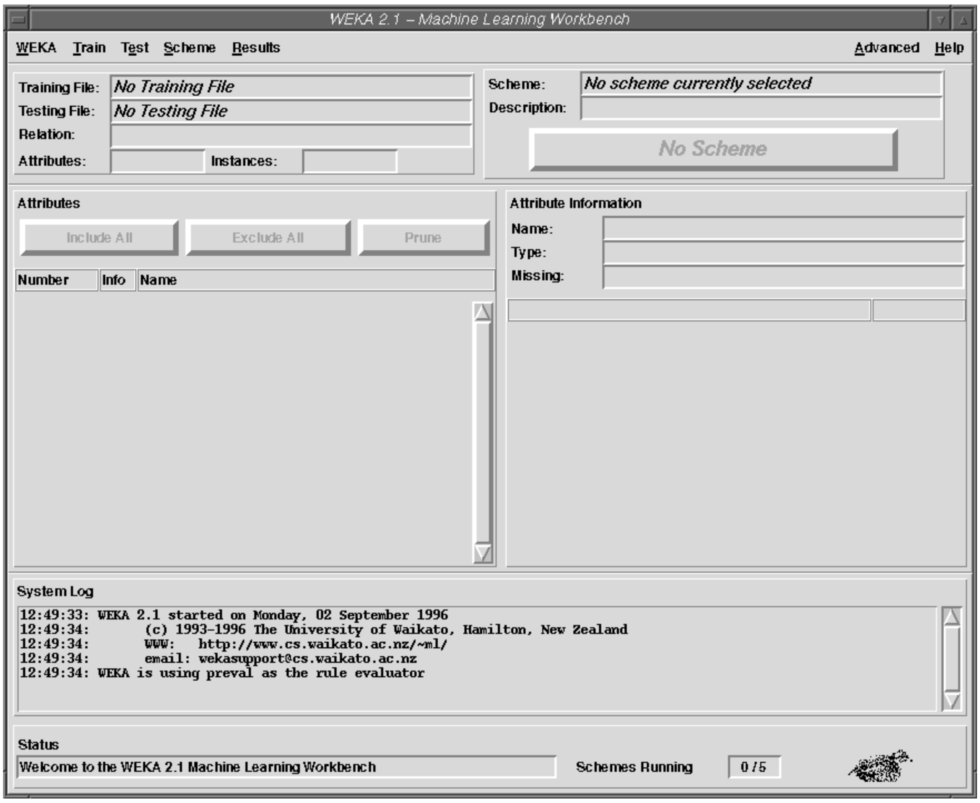
http://www.cs.waikato.ac.nz/~ml/

email: wekasupport@cs.waikato.ac.nz

Figure 4. About WEKA

[9] You may also have to set your display variable before you start WEKA. For example on a machine called pluto, type: `setenv DISPLAY pluto:0.0` or `export DISPLAY=pluto:0.0` (if the machine has a different name then change the name pluto to your machine name).

Select **Okay** on the About WEKA window. You will now be viewing the Machine Learning Workbench window only.



Training File: No Training File
Testing File: No Testing File
Relation:
Attributes:
Instances:
Scheme: No scheme currently selected
Description:
Name:
Type:
Missing:

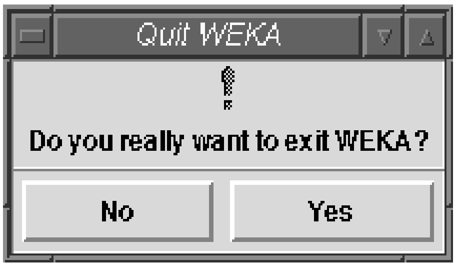
```
12:49:33: WEKA 2.1 started on Monday, 02 September 1996
12:49:34:       (c) 1993-1996 The University of Waikato, Hamilton, New Zealand
12:49:34:       WWW:   http://www.cs.waikato.ac.nz/~ml/
12:49:34:       email: wekasupport@cs.waikato.ac.nz
12:49:34: WEKA is using preval as the rule evaluator
```

Welcome to the WEKA 2.1 Machine Learning Workbench
Schemes Running 0 / 5

Figure 5. Machine Learning Workbench

Exiting WEKA 2.1

You may exit WEKA at any time during your analysis. To do this select the **Quit** option from the WEKA pull down menu.



Do you really want to exit WEKA?

Figure 6. Quit WEKA

The Machine Learning Workbench

The Machine Learning Workbench is provided for domain experts, such as agriculturalists or horticulturalists, who are interested in learning more about their data. The Machine Learning Workbench also enables machine learning experts to learn how to analyse datasets more efficiently and effectively on different machine learning schemes.

By running the golf dataset through each of the schemes individually, the Machine Learning Workbench will be used to produce :

- rules and decision trees about whether to play golf or not (for both a domain expert and machine learning expert);
- PREVAL (PROLOG) evaluation, including confusion matrices, on scheme accuracy and efficiency (for a machine learning expert).

The rules, decision trees, and PROLOG output will be based on the data in the golf dataset. Once each of the schemes has finished running the results from each schemes analysis can be viewed and interpreted.

Loading a Dataset

To open the golf.arff training dataset, select **Open Training Dataset...** from the Train pull down menu.[10]

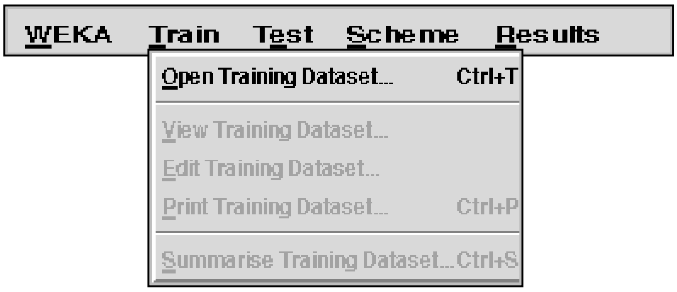
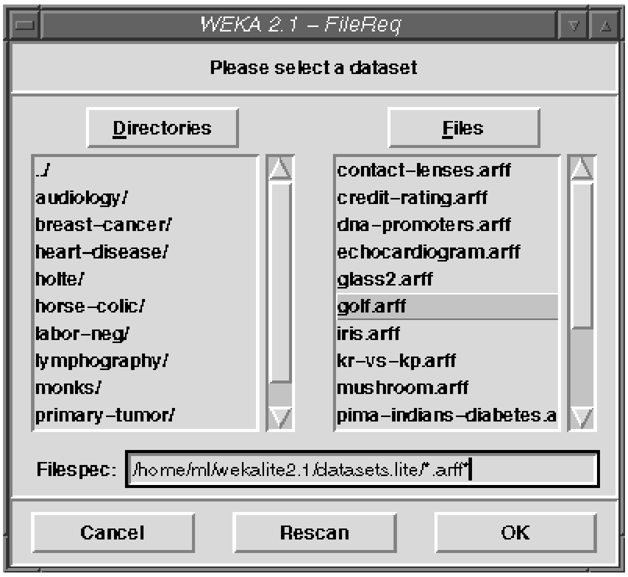

WEKA
Train
Test
Scheme
Results
Open Training Dataset... Ctrl+T
View Training Dataset...
Edit Training Dataset...
Print Training Dataset... Ctrl+P
Summarise Training Dataset... Ctrl+S

Figure 7. Open Train menu — Open Training Dataset...

[10] The training dataset is used to build the basic rules or decision tree based on the data (for further details, go to page 6).

A window appears requesting you to select a datafile.



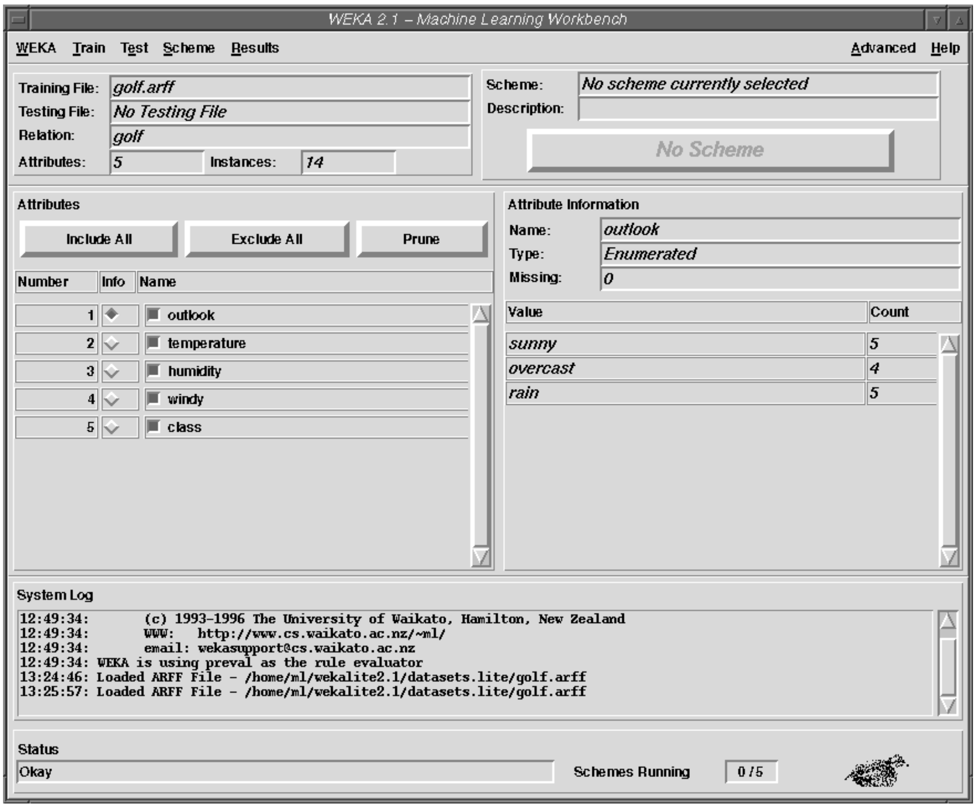
Filespec: /home/ml/wekalite2.1/datasets.lite/*.arff*

Figure 8. Please select a dataset

You should be in the directory:

```
$WEKAHOME/datasets.lite/...
```

Choose the **golf.arff** datafile, listed under Files. Select**OK**.



Training File: *golf.arff*
Testing File: *No Testing File*
Relation: *golf*
Attributes: *5* Instances: *14*

Scheme: *No scheme currently selected*
Description:

Number	Info	Name
1	◆	outlook
2	◇	temperature
3	◇	humidity
4	◇	windy
5	◇	class

Name: *outlook*
Type: *Enumerated*
Missing: *0*

Value	Count
sunny	*5*
overcast	*4*
rain	*5*

```
12:49:34:       (c) 1993-1996 The University of Waikato, Hamilton, New Zealand
12:49:34:       WWW:   http://www.cs.waikato.ac.nz/~ml/
12:49:34:       email: wekasupport@cs.waikato.ac.nz
12:49:34: WEKA is using preval as the rule evaluator
13:24:46: Loaded ARFF File - /home/ml/wekalite2.1/datasets.lite/golf.arff
13:25:57: Loaded ARFF File - /home/ml/wekalite2.1/datasets.lite/golf.arff
```

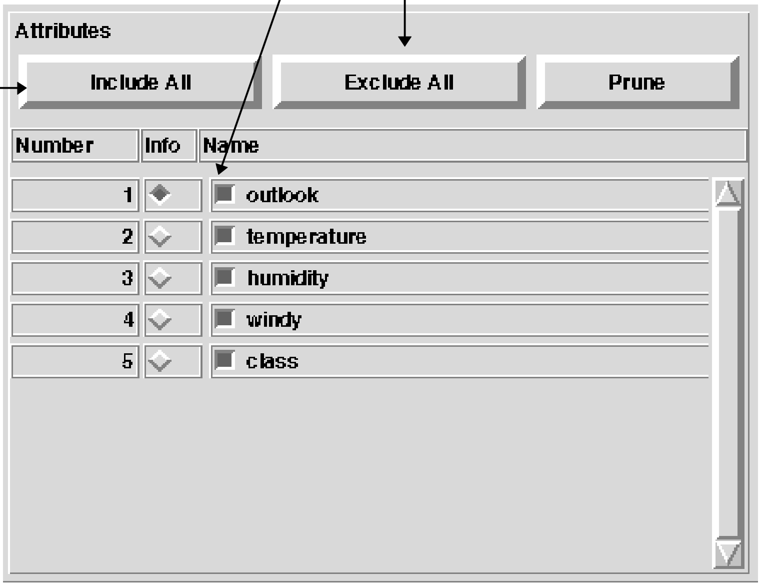
Status: Okay
Schemes Running: 0 / 5

Figure 9. Machine Learning Workbench with golf dataset

The Machine Learning Workbench now shows information about the golf dataset. This includes the file name, number of attributes and instances, and each attribute in the dataset.

Selecting the Attributes

You can select the golf dataset's attributes which you want to include in your analysis experiment. You can select or deselect and attribute by clicking on the red filled box beside the attribute name.



Number	Info	Name
1	◆	■ outlook
2	◇	■ temperature
3	◇	■ humidity
4	◇	■ windy
5	◇	■ class

Figure 10. Golf attributes.

Selecting a subset of the existing attributes means that the data in the dataset will be restricted to only those that you include. The results from any analysis on a subset of the dataset will usually vary in some way in comparison to analysing the dataset with all of its attributes. However, for the golf dataset, all its attributes will be included. To include all of the attributes click on the**Include All** button.

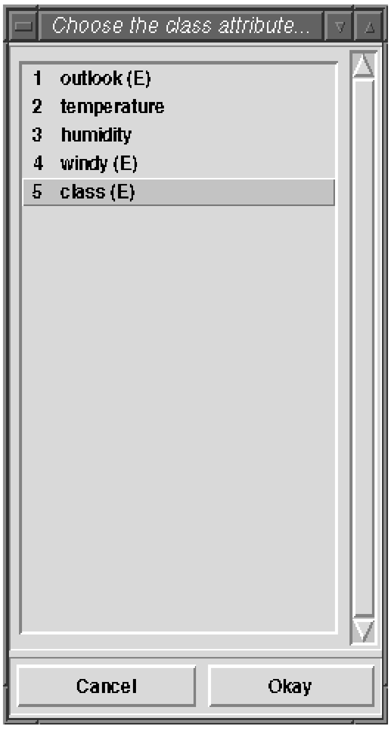
(To see how this option actually works, select the **Exclude All** button; notice how each of the boxes beside the attribute name are no longer filled. Re-select all of the attributes using the Include All button).

Choosing a Scheme

From the WEKA Scheme pull down menu select **1R**. (This is the very first scheme in the list). The Machine Learning Workbench window now shows the Run Scheme button as Run 1R. Select this.

Figure 11. Run Scheme Button

The Choose the class attribute... window appears. In this case the default class value (5 class(E)) is highlighted. Select **Okay**.



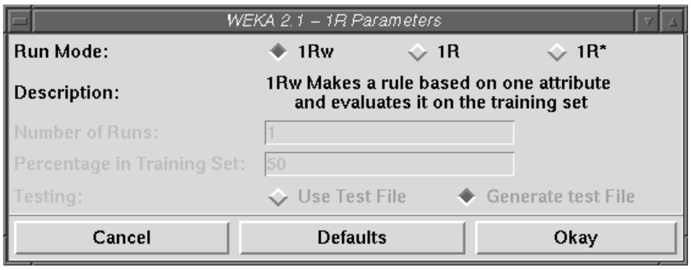
Choose the class attribute...

1 outlook (E)
2 temperature
3 humidity
4 windy (E)
5 class (E)

Figure 12. Choose the class attribute...

Configuring the Scheme

The 1R Parameters window will then appear. Select **1Rw** from the Run Mode options and then **Okay**.



WEKA 2.1 – 1R Parameters

Run Mode: ◆ 1Rw ◇ 1R ◇ 1R*

Description: 1Rw Makes a rule based on one attribute and evaluates it on the training set

Number of Runs: 1

Percentage in Training Set: 50

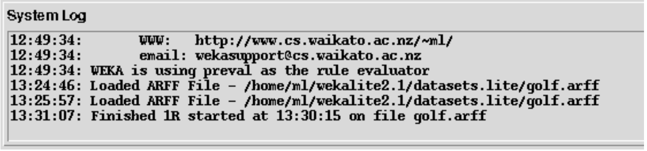
Testing: ◇ Use Test File ◆ Generate test File

Figure 13. 1R Parameters

Running the Scheme

WEKA will now run the golf dataset through the 1Rw scheme. If you look at the lower right side of the Machine Learning Workbench window you will see the WEKA figure moving. This indicates that the scheme is currently analysing the golf dataset.

When the scheme has finished executing a message will appear in the System Log.

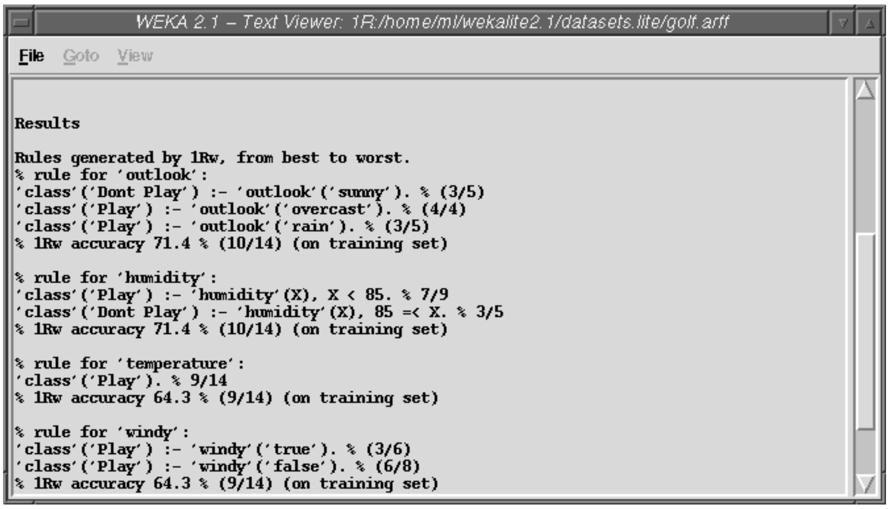


```
12:49:34:       WWW:   http://www.cs.waikato.ac.nz/~ml/
12:49:34:       email: wekasupport@cs.waikato.ac.nz
12:49:34: WEKA is using preval as the rule evaluator
13:24:46: Loaded ARFF File - /home/ml/wekalite2.1/datasets.lite/golf.arff
13:25:57: Loaded ARFF File - /home/ml/wekalite2.1/datasets.lite/golf.arff
13:31:07: Finished 1R started at 13:30:15 on file golf.arff
```

Figure 14. System Log

Viewing the Results

To view the 1Rw results, from the Results pull down menu select the italicised option **1R(golf.arff).** The Text Viewer window appears.


WEKA 2.1 – Text Viewer: 1R:/home/ml/wekalite2.1/datasets.lite/golf.arff
File
Goto
View

```
Results

Rules generated by 1Rw, from best to worst.
% rule for 'outlook':
'class'('Dont Play') :- 'outlook'('sunny'). % (3/5)
'class'('Play') :- 'outlook'('overcast'). % (4/4)
'class'('Play') :- 'outlook'('rain'). % (3/5)
% 1Rw accuracy 71.4 % (10/14) (on training set)

% rule for 'humidity':
'class'('Play') :- 'humidity'(X), X < 85. % 7/9
'class'('Dont Play') :- 'humidity'(X), 85 =< X. % 3/5
% 1Rw accuracy 71.4 % (10/14) (on training set)

% rule for 'temperature':
'class'('Play'). % 9/14
% 1Rw accuracy 64.3 % (9/14) (on training set)

% rule for 'windy':
'class'('Play') :- 'windy'('true'). % (3/6)
'class'('Play') :- 'windy'('false'). % (6/8)
% 1Rw accuracy 64.3 % (9/14) (on training set)
```

Figure 15. 1R results

Analysing the Machine Learning Workbench Scheme Results

The results from the Machine Learning Workbench need to be analysed from the point of view of the domain. Analysing these results involves answering questions concerning not only the data, but the domain, and also the machine learning environment used. These questions include (there are many more):

- are the results accurate?;
- can the results be applied to the domain?;
- can better results be gained from the machine learning schemes by using different scheme configurations or by creating new attributes or a subset of the data

At the top of each scheme's output, dataset and scheme information is given. This is in the form of:

```
<Scheme> - <Day, Date Month Year>
===============================================
Filename:         <pathname>
Relation:         <relation name>
Instances:        <number of instances>
Class Attribute:  <name of class attribute>
Attributes Used:  <numbers relating to attributes>
                  <attribute names>
No. of Attributes:<number of attributes>
Test File:        <None needed  / None
                   specified / filename>
```

Figure 16. Scheme Header Information

The output for each scheme analysis on the Machine Learning Workbench for the golf dataset is:

```
<Scheme> - <Day, Date Month Year>
================================================
Filename:
      /home/ml/wekalite2.1/datasets.lite/golf.arff
Relation:               golf
Instances:              14
Class Attribute:        class
Attributes Used:        1,2,3,4,5
                        'outlook'
                        'temperature'
                        'humidity'
                        'windy'
                        'class'
No. of Attributes:      5
Test File:              None needed/None specified
```

Figure 17. Scheme Header Information for the Golf Dataset

Each scheme requires separate analysis, as each scheme's results are different. The schemes' output will be analysed for the most effective results, these will be shown and explained.

1R Results

1R produces rules, testing their performance through analysis. It has generated a simple rule stating which attribute(s) are the most effective for deciding whether golf should be played or not.

The rules generated by 1Rw are listed from best to worst, with the best being listed first. These are measured in terms of predictive ability. For the golf dataset 1Rw has predicted two attributes as being the highest ranking at 71.4% accuracy. These are outlook and humidity. Temperature and windy have 64.3% accuracy.

```
Rules generated by 1Rw, from best to worst.
% rule for 'outlook':
'class'('Dont Play') :- 'outlook'('sunny'). % (3/5)
'class'('Play') :- 'outlook'('overcast'). % (4/4)
'class'('Play') :- 'outlook'('rain'). % (3/5)
% 1Rw accuracy 71.4 % (10/14) (on training set)

% rule for 'humidity':
'class'('Play') :- 'humidity'(X), X < 85. % 7/9
'class'('Dont Play') :- 'humidity'(X), 85 =< X. % 3/5
% 1Rw accuracy 71.4 % (10/14) (on training set)

% rule for 'temperature':
'class'('Play'). % 9/14
% 1Rw accuracy 64.3 % (9/14) (on training set)

% rule for 'windy':
'class'('Play') :- 'windy'('true'). % (3/6)
'class'('Play') :- 'windy'('false'). % (6/8)
% 1Rw accuracy 64.3 % (9/14) (on training set)
```

Figure 18. Results from 1Rw Analysis

Analysing 1R Results

These rules are read as 'if-then-else-if' statements. For example, the first rule reads as:

> If the outlook is sunny then don't play golf, otherwise (else) if the outlook is either overcast or rainy then play golf.

Comparing the first rule predicate with the original data (below), you can see that there were three instances with the 'outlook' attribute value as 'sunny' that also had 'Don't Play' as the class value.

OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	CLASS
sunny	85	85	FALSE	'Dont Play'
sunny	80	90	TRUE	'Dont Play'
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	'Dont Play'
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	'Dont Play'

Figure 19. Data from Golf Dataset

There were in fact five instances with ‘sunny’ as the ‘outlook’ attribute value. The remaining two had ‘Play’ as the class value.

sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75	70	TRUE	Play

Figure 20. Two Remaining ‘outlook’ = ‘sunny’ Instances in the Golf Dataset

The next term states that all four instances with the ‘outlook’ attribute having the value ‘overcast’ have a class value as ‘Play’.

```
'class'('Play') :- 'outlook'('overcast'). % (4/4)
```

Figure 21. 1Rw’s First Rule’s Second Term

Finally, for the third term, of the five possible instances with ‘rain’ as the value for the outlook attribute, three of these had the class value ‘Play’; the other two have the class ‘Dont Play’.

```
'class'('Play') :- 'outlook'('rain'). % (3/5)
```

Figure 22. 1Rw’s First Rule’s Third Term

The outlook attribute was accurate 71.4% of the time. This means that for every instance in the dataset (14 in total), only four were incorrectly classified (10/14).

From the results it seems that both ‘outlook’ and ‘humidity’ can both be used separately to decide whether to play golf or not. This is because they both have the highest accuracy rate at 71.4%.

1R PREVAL Evaluation

1Rw does not output any PREVAL evaluation.

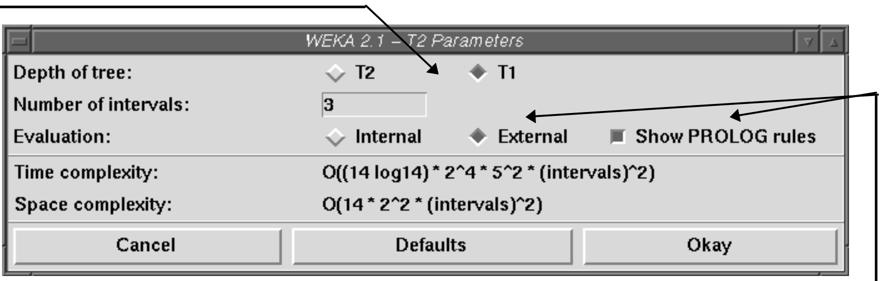
To exit the results window and return to the Machine Learning Workbench select **Close** from the File pull down menu.

Other Schemes

Follow similar instructions for selecting and running each of the following schemes; when each has finished view and analyse the results:

T2
From the Scheme pull down menu, choose **T2**. Select the **Run T2** button and choose the class attribute.

The T2 Parameters window now appears.

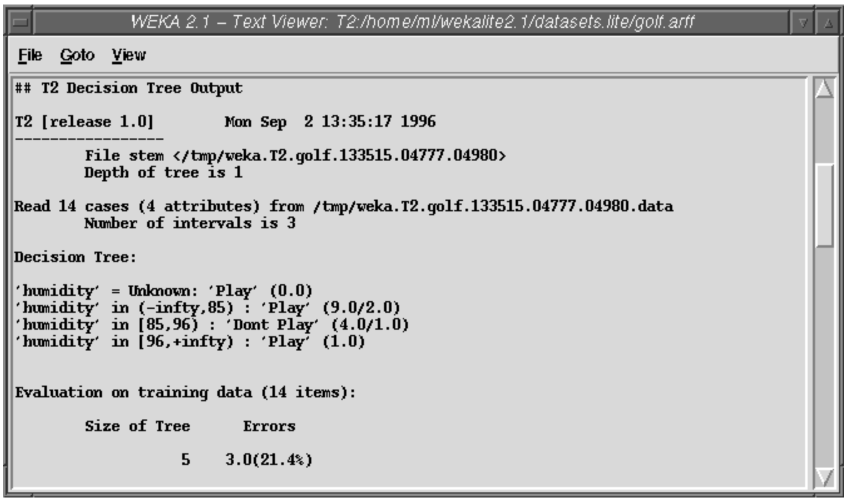


Depth of tree: ◇ T2 ◆ T1
Number of intervals: 3
Evaluation: ◇ Internal ◆ External ■ Show PROLOG rules
Time complexity: O((14 log14) * 2^4 * 5^2 * (intervals)^2)
Space complexity: O(14 * 2^2 * (intervals)^2)

Figure 23. T2 Parameters

Configure T2, choosing Depth of Tree as **T1**, leaving Evaluation as both External and Show PROLOG. Select **Okay**.

Viewing T1 Results
To view the T1 analysis results, from the Results pull down menu select the italicised option **T2(golf.arff).** The Text Viewer window then appears.


WEKA 2.1 – Text Viewer: T2:/home/ml/wekalite2.1/datasets.lite/golf.arff
File
Goto
View

```
## T2 Decision Tree Output

T2 [release 1.0]        Mon Sep  2 13:35:17 1996
-----------------
        File stem </tmp/weka.T2.golf.133515.04777.04980>
        Depth of tree is 1

Read 14 cases (4 attributes) from /tmp/weka.T2.golf.133515.04777.04980.data
        Number of intervals is 3

Decision Tree:

'humidity' = Unknown: 'Play' (0.0)
'humidity' in (-infty,85) : 'Play' (9.0/2.0)
'humidity' in [85,96) : 'Dont Play' (4.0/1.0)
'humidity' in [96,+infty) : 'Play' (1.0)


Evaluation on training data (14 items):

        Size of Tree      Errors

                   5    3.0(21.4%)
```

Figure 24. T1 results

You can also view a graphical representation of the T1 decision tree. From the View pull down menu within the Text Viewer window, select **Unpruned Tree...**[11]

[11] (Pruned Tree can not be selected as no pruned tree has been output from the T1 scheme).

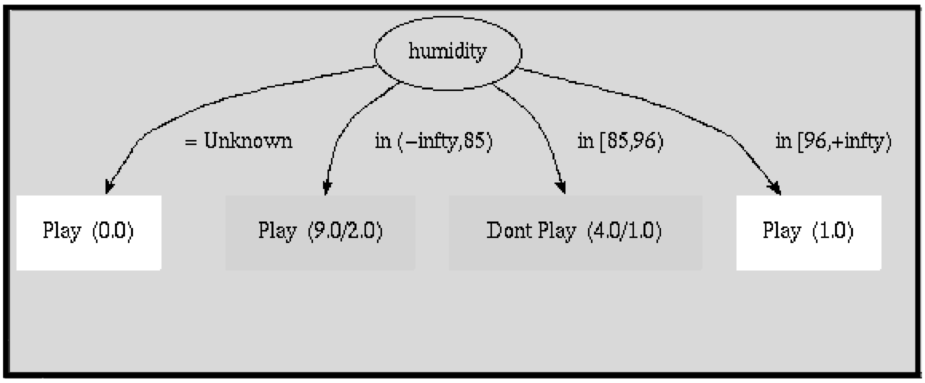


Figure 25. T1 tree output

To return to the 1R Text Viewer window, select **Close** from the File menu.

T1 Results
T1 produces a 1-level decision tree and PROLOG rules derived from the decision tree.

The T1 results are given in a decision tree format, and can be read as 'if-then....if-else...' statements.

Analysing T1 Results
T1 analysis defines 'humidity' as the single most significant attribute for deciding to play golf or not. This is similar to 1R results, where 'humidity' had the highest accuracy rate (equal with 'outlook').

Each branch of the decision tree is derived from values for the 'humidity' attribute; where 65 is the lowest value and 96 is the highest value.

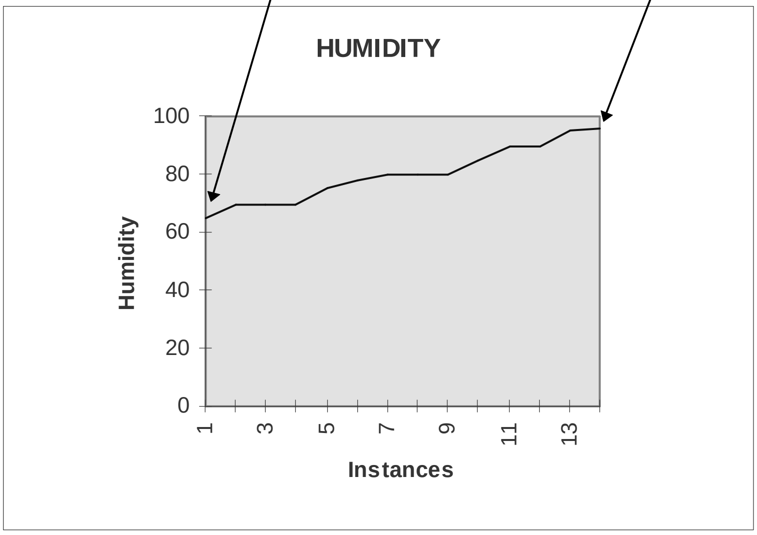


Figure 26. Humidity Distribution

```
'humidity' = Unknown: 'Play' (0.0)
'humidity' in (-infty,85) : 'Play' (9.0/2.0)
'humidity' in [85,96) : 'Dont Play' (4.0/1.0)
'humidity' in [96,+infty) : 'Play' (1.0)


Evaluation on training data (14 items):

   Size of Tree      Errors

              5    3.0(21.4%)
```

Figure 27. T1 Tree Output

The decision tree rules can be read as:

> if humidity is unknown then Play golf (the zero value in brackets indicate that this rule is never used),
>
> else if humidity is between negative infinity and 85, then play golf (the number in brackets indicate that this rule was used nine times, but incorrectly classified two instances),
>
> else if humidity is between 85 and 96 then don't play golf (this rules was used four times but incorrectly classified one instance),
>
> else if humidity is between 96 and positive infinity then play golf (this rule was used once correctly).

At the bottom of the tree, evaluation on the classification of the training data is given. The number of nodes (size of tree) in the decision tree is listed, along with the number of incorrectly classified instances (errors), and the incorrectly classification percentage.[12]

```
Evaluation on training data (14 items):

   Size of Tree      Errors

              5    3.0(21.4%)
```

Figure 28. T1 Tree Evaluation

[12] A incorrectly classified instance is where an instance was classified with the wrong class value.

The size of the tree (number of nodes) is derived from the number or rules, and the attribute that the rules concern. This is easier to view when you look at the graphical version of the decision tree (page 29). You will see that it has five nodes.

T1 PREVAL Evaluation

For T1 PREVAL outputs its own PROLOG rules based on the data in the dataset. The PROLOG rules for the golf dataset are:

```
## Prolog Rules made from T2 Decision Tree
% T2 WEKA Prolog Rules: T2 Tree


% Rule 1 - Length 1
’class’(’Play’) :-
      ( ’humidity’(?) ).

% Rule 2 - Length 1
’class’(’Play’) :-
      ( ’humidity’(X_2), X_2 < 85 ).

% Rule 3 - Length 2
’class’(’Dont Play’) :-
      ( ’humidity’(X_3), X_3 >= 85, X_3 < 96 ).

% Rule 4 - Length 1
’class’(’Play’) :-
      ( ’humidity’(X_4), X_4 >= 96 ).
```

Figure 29. PROLOG Rules from T1 Analysis

These rules are derived from the rules built by the T1 decision. There are four rules, each is output in PROLOG format.

A classification table is also produced with PREVAL evaluation. This is in the form of a confusion matrix, describing the number of instances which were correctly and incorrectly classified. The confusion matrix shows that eight instances with a class value of ‘Play’ were correctly classified, and one instance was incorrectly classified.

```
Classification Table:
   (a)  (b)   <- classified as
  ---- ----
     8    1   (a):  class ’Play’
     2    3   (b):  class ’Dont Play’
```

Figure 30. Classification Table of T1 Analysis

It also shows that three instances with a class value of ‘Don’t Play’ were correctly classified, and two were incorrectly classified.

Another table is output from PREVAL evaluation. This is a Rule Usage table which indicates how many instances were either correctly, incorrectly, or multiply classified for each rule. Also, what the decision was for each rule.

```
Rule Usage:
         Correct Incorrect Multiple
Rule   1:   0       0         0     -> 'Play'
Rule   2:   7       2         0     -> 'Play'
Rule   3:   3       1         0     -> 'Dont Play'
Rule   4:   1       0         0     -> 'Play'
              ----------------------
Total Hits  11      3         0
```

Figure 31. T1 Rule Usage Analysis

As you can see, the second rule correctly classified seven instances and incorrectly classified two instances.

Finally, a summary is given, which states the number and percentage for the correctly, incorrectly, unclassified and multiply classified instances; and includes a category for unknown results (there were none for the golf dataset).

number percentage

```
Summary:
Correctly Classified Instances      11  78.5714 %
Incorrectly Classified Instances     3  21.4286 %
UnClassified Instances               0   0.0000 %
Multiply Classified Instances        0   0.0000 %
Unknown results:                     0   0.0000 %
```

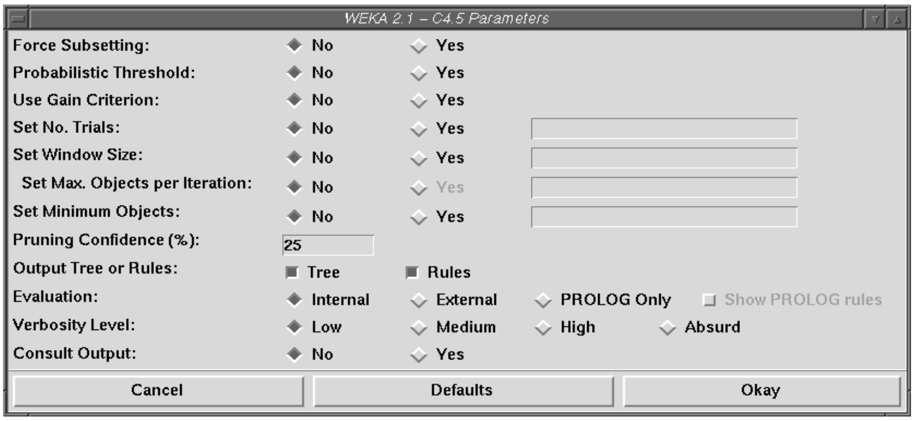
Figure 32. T1 Summary Statistics

You should be in the T2 Text Viewer window. To return to the Machine Learning Workbench window, select **Close** from the T2 Text Viewer window File menu.

C4.5

From the Scheme menu, choose **C4.5**. Select the **Run C4.5** button and choose the class attribute.

The C4.5 Parameters window now appears. Configure C4.5 selecting Output Tree or Rules as **Tree** and **Rules**, and change Evaluation to **Internal** (otherwise you will be unable to output both tree and rule). Select **Okay**.

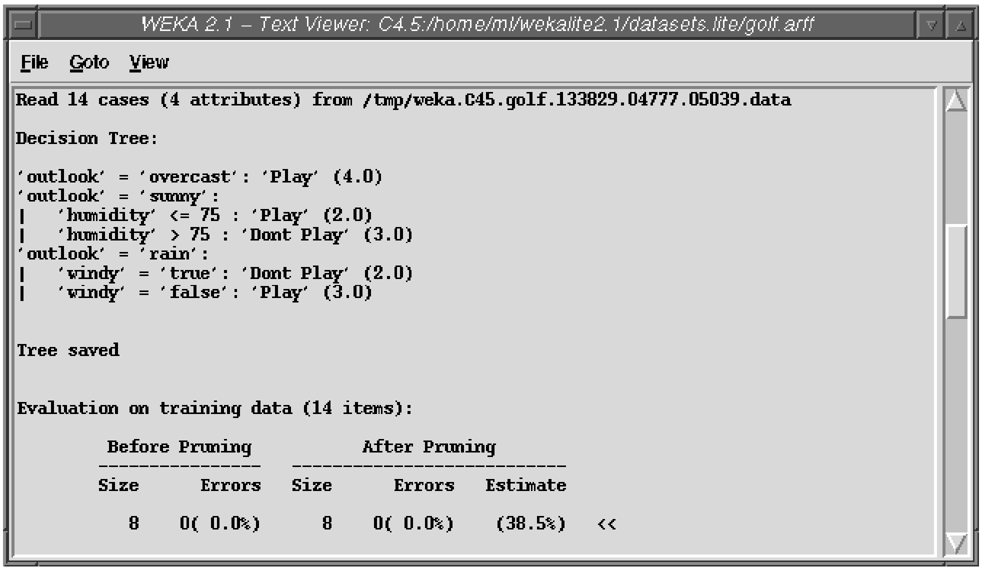


Force Subsetting: ● No ○ Yes
Probabilistic Threshold: ● No ○ Yes
Use Gain Criterion: ● No ○ Yes
Set No. Trials: ● No ○ Yes
Set Window Size: ● No ○ Yes
Set Max. Objects per Iteration: ● No ○ Yes
Set Minimum Objects: ● No ○ Yes
Pruning Confidence (%): 25
Output Tree or Rules: ☒ Tree ☒ Rules
Evaluation: ● Internal ○ External ○ PROLOG Only □ Show PROLOG rules
Verbosity Level: ● Low ○ Medium ○ High ○ Absurd
Consult Output: ● No ○ Yes

Figure 33. C4.5 Parameters

Viewing C4.5 Results

To view the results of the C4.5 analysis, from the Results menu select the italicised option **C4.5(golf.arff).** The Text Viewer window then appears.


WEKA 2.1 – Text Viewer: C4.5:/home/ml/wekalite2.1/datasets.lite/golf.arff
File
Goto
View

```
Read 14 cases (4 attributes) from /tmp/weka.C45.golf.133829.04777.05039.data

Decision Tree:

'outlook' = 'overcast': 'Play' (4.0)
'outlook' = 'sunny':
|   'humidity' <= 75 : 'Play' (2.0)
|   'humidity' > 75 : 'Dont Play' (3.0)
'outlook' = 'rain':
|   'windy' = 'true': 'Dont Play' (2.0)
|   'windy' = 'false': 'Play' (3.0)


Tree saved


Evaluation on training data (14 items):

         Before Pruning          After Pruning
        ----------------   ---------------------------
        Size      Errors   Size      Errors   Estimate

           8    0( 0.0%)      8    0( 0.0%)    (38.5%)   <<
```

Figure 34. C4.5 results

You can also view a graphical representation of the C4.5 decision tree. From the View menu within your Text Viewer window, select **Unpruned Tree**... (Pruned Tree can not be selected as no pruned tree has been output from the C4.5 scheme).

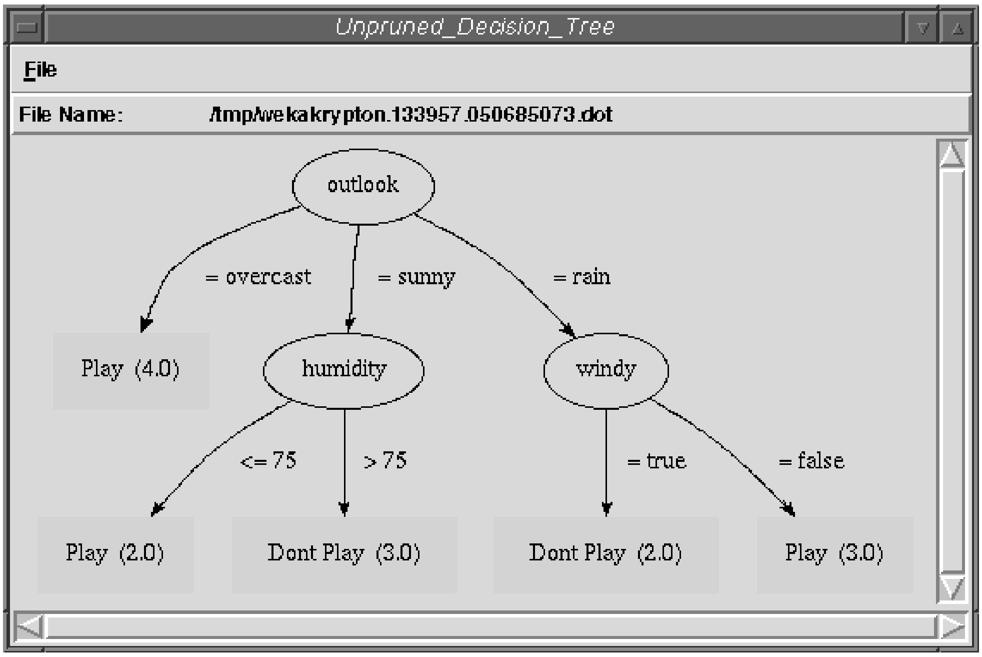


Figure 35. C4.5 Unpruned Decision Tree

To return to the C4.5 Text Viewer window, select **Close** from the File menu.

C4.5 Results

C4.5 results builds on the results from 1Rw and T1. It produces a decision tree first, then final rules which are derived from the decision tree rules.

The C4.5 rules can be read as 'if-then...else-if', 'AND' and 'OR' rules; where each rule that is listed one after another represents an 'OR' statement, each rule that is indented is an 'AND' statement.

Analysing C4.5 Results

Only one decision tree has been built by C4.5 for the golf dataset. The reason for this is that the golf dataset is rather simple, ie. it contains few instances and attributes. If the dataset was larger then both an unpruned and a pruned decision tree would have been output.

The C4.5 decision tree for the golf dataset is small in structure with eight nodes.

```
'outlook' = 'overcast': 'Play' (4.0)
'outlook' = 'sunny':
|   'humidity' <= 75 : 'Play' (2.0)
|   'humidity' > 75 : 'Dont Play' (3.0)
'outlook' = 'rain':
|   'windy' = 'true': 'Dont Play' (2.0)
|   'windy' = 'false': 'Play' (3.0)


Tree saved


Evaluation on training data (14 items):

Before Pruning           After Pruning
----------------   ---------------------------
Size      Errors   Size      Errors   Estimate

   8    0( 0.0%)      8    0( 0.0%)    (38.5%)
```

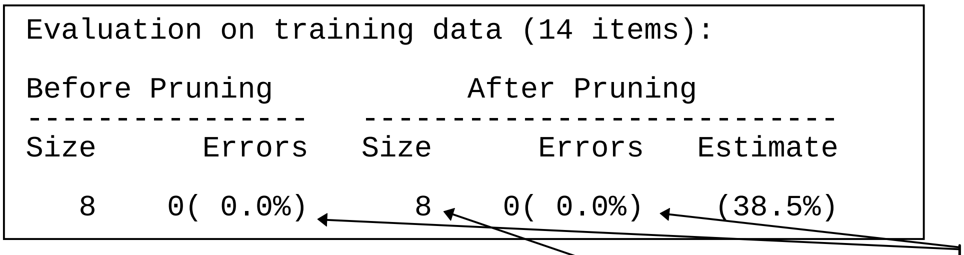
Figure 36. C4.5 Decision Tree and Training Data Evaluation

The rules read as:

> if outlook is overcast then play golf, OR
>
> if outlook is sunny AND humidity is less than 76 then play golf, OR
>
> if outlook is sunny AND humidity is greater than 75 then don’t play golf, OR
>
> if outlook is for rain AND windy is true then don’t play golf, OR
>
> if outlook is for rain AND windy is false then play golf.

The numbers in brackets after each decision has been reached indicates how often a rule is used to classify an instance, which is similar to both 1R and T1.

At the bottom of the decision tree, evaluation on the classification of the training data is given. This includes evaluation of the unpruned decision tree and evaluation of the pruned decision tree. As only one tree was built for the golf dataset, both the pruned and unpruned decision tree results refer to the same tree.



```
Evaluation on training data (14 items):

Before Pruning           After Pruning
----------------   ---------------------------
Size      Errors   Size      Errors   Estimate

   8    0( 0.0%)      8    0( 0.0%)    (38.5%)
```

Figure 37. C4.5 Evaluation on Training Data

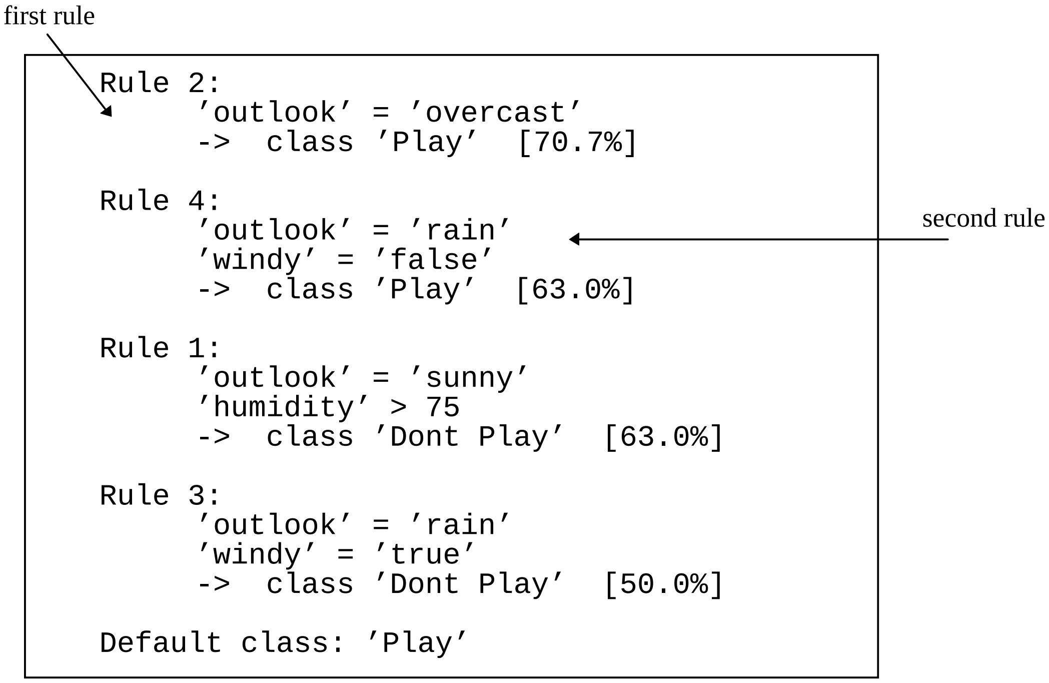
Both the pruned and unpruned evaluation give the size (number of nodes) of the tree and the number of errors and the incorrectly classified percentage.

For the golf dataset the output tree has eight nodes (look at the graphical version, Figure...), and no instances have been incorrectly classified. This is a good result. Ideally, trees with less than ten nodes and with a 90% to 100% accuracy indicate that the results will be easy to work with.

The final rules are based on the decisions from the tree. The rules are listed in order of classification accuracy. The C4.5 rules can be read as ‘if-then...else-if’ rules. For instance, look at the graphical view of the decision tree (Figure....), and starting from the top — take the branch with the highest classification accuracy (the left-most branch). The rule this depicts would read:

if ‘outlook’ is overcast then play golf.

This is, in fact, correct as the first rule states this (see the next page).



first rule

second rule

```
Rule 2:
     ’outlook’ = ’overcast’
     ->  class ’Play’  [70.7%]

Rule 4:
     ’outlook’ = ’rain’
     ’windy’ = ’false’
     ->  class ’Play’  [63.0%]

Rule 1:
     ’outlook’ = ’sunny’
     ’humidity’ > 75
     ->  class ’Dont Play’  [63.0%]

Rule 3:
     ’outlook’ = ’rain’
     ’windy’ = ’true’
     ->  class ’Dont Play’  [50.0%]

Default class: ’Play’
```

Figure 38. C4.5 rules

The first rule is used 70.7% of the time for classifying instances (or four times if viewed from the decision tree).

The second highest classification is:

if ‘outlook’ is for ‘rain’ AND ‘windy’ is ‘false’ then play golf

This is the right most branch, and has classified three instances of the 14 in the dataset.

Each of the rules can be read like this, or alternatively, you can use to the decision tree to reach a particular decision. If an instance, is given in the future which cannot be classified by any of the rules, then by default it will be classified as ‘Play’. The default class is the class value which occurs most often.

C4.5 PREVAL Evaluation

For C4.5, with internal evaluation, PREVAL outputs a table listing information about each rule and a confusion matrix. The C4.5 PROLOG table for the golf dataset are:

```
Evaluation on training data (14 items):

Rule  Size  Error  Used  Wrong    Advantage
----  ----  -----  ----  -----    ---------
   2     1  29.3%     4  0 (0.0%) 0 (0|0) 'Play'
   4     2  37.0%     3  0 (0.0%) 0 (0|0) 'Play'
   1     2  37.0%     3  0 (0.0%) 3 (3|0) 'Dont Play'
   3     2  50.0%     2  0 (0.0%) 2 (2|0) 'Dont Play'
```

Figure 39. C4.5 Evaluation on Training Data

This table lists each rule in the order that it was given in the C4.5 output. It also includes how many terms are in each rule (the rule’s size), the percentage of error each rule will have for future classifications given the same parameters, the number of times each rule was used, the number of incorrect classifications for each rule, and what advantage can be gained from using this rule. Lastly, each rules classification is given.

For example, the first rule, is actually rule 2. It has one term, with 70.7% accuracy in determining the classification. The first rule was used four times, and correctly classified all four of these instances. It’s class is ‘Play’.

Finally, a confusion matrix is given. This depicts how many instances were correctly and incorrectly classified. For the golf dataset, all instances were correctly classified, so no incorrectly classified instances are given.

```
Tested 14, errors 0 (0.0%)   <<


        (a)  (b)  <-classified as
       ---- ----
          9        (a): class 'Play'
               5   (b): class 'Dont Play'
```

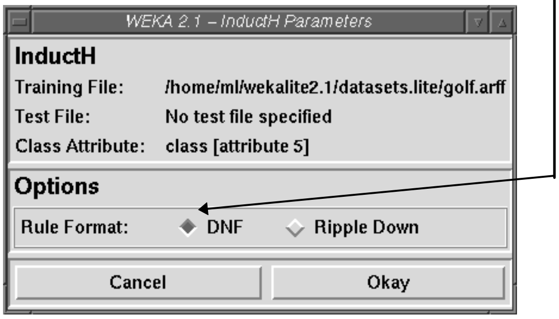
Figure 40. C4.5 Confusion Matrix

To return to the Machine Learning Workbench, select **Close** from the File menu.

InductH

From the Scheme menu on the Machine Learning Workbench, choose **InductH**. Select the **Run InductH** button and choose the class attribute.

The InductH Parameters window now appears. Configure InductH selecting Rule Format as **DNF**. (DNF is the default option for InductH). Select **Okay**.



InductH

Training File: /home/ml/wekalite2.1/datasets.lite/golf.arff

Test File: No test file specified

Class Attribute: class [attribute 5]

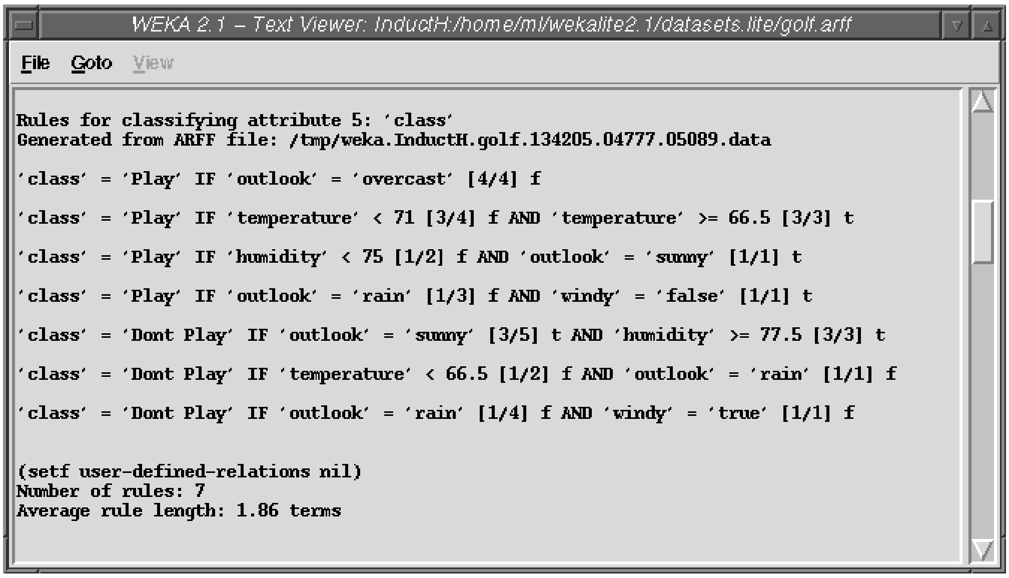
Options

Rule Format: ● DNF ○ Ripple Down

Figure 41. InductH Parameters

Viewing InductH Results

When InductH has finished analysing the golf dataset, the results can be viewed by selecting the italicised option **InductH(golf.arff)** from the Results menu. The Text Viewer window then appears.


WEKA 2.1 – Text Viewer: InductH:/home/ml/wekalite2.1/datasets.lite/golf.arff
File
Goto
View

```
Rules for classifying attribute 5: 'class'
Generated from ARFF file: /tmp/weka.InductH.golf.134205.04777.05089.data

'class' = 'Play' IF 'outlook' = 'overcast' [4/4] f

'class' = 'Play' IF 'temperature' < 71 [3/4] f AND 'temperature' >= 66.5 [3/3] t

'class' = 'Play' IF 'humidity' < 75 [1/2] f AND 'outlook' = 'sunny' [1/1] t

'class' = 'Play' IF 'outlook' = 'rain' [1/3] f AND 'windy' = 'false' [1/1] t

'class' = 'Dont Play' IF 'outlook' = 'sunny' [3/5] t AND 'humidity' >= 77.5 [3/3] t

'class' = 'Dont Play' IF 'temperature' < 66.5 [1/2] f AND 'outlook' = 'rain' [1/1] f

'class' = 'Dont Play' IF 'outlook' = 'rain' [1/4] f AND 'windy' = 'true' [1/1] f


(setf user-defined-relations nil)
Number of rules: 7
Average rule length: 1.86 terms
```

Figure 42. InductH results

InductH Results

InductH produces rules (read as if-then...else-if statements) where at least one of the rules will classify an instance. These are in the form of:

```
<class> = <class value> IF <attribute> = <attribute
                 value>...AND/IF..
```

where the class equals the class value if the data meets the rule conditions.

Analysing InductH Results

```
’class’ = ’Play’ IF ’outlook’ = ’overcast’ [4/4] f

’class’ = ’Play’ IF ’temperature’ < 71 [3/4] f
      AND ’temperature’ >= 66.5 [3/3] t

’class’ = ’Play’ IF ’humidity’ < 75 [1/2] f AND
      ’outlook’ = ’sunny’ [1/1] t

’class’ = ’Play’ IF ’outlook’ = ’rain’ [1/3] f
      AND ’windy’ = ’false’ [1/1] t

’class’ = ’Dont Play’ IF ’outlook’ = ’sunny’
      [3/5] t AND ’humidity’ >= 77.5 [3/3] t

’class’ = ’Dont Play’ IF ’temperature’ < 66.5
      [1/2] f AND ’outlook’ = ’rain’ [1/1] f

’class’ = ’Dont Play’ IF ’outlook’ = ’rain’
      [1/4] f AND ’windy’ = ’true’ [1/1] f
```

Figure 43. InductH Rules for the Golf Dataset

For example, the rules are read as:

play golf if the outlook is overcast,

else play golf if the temperature is between 66.5 and 70.9,

else play golf if the humidity is less than 75 and the outlook is sunny,

else play golf if outlook is for rain and windy is false

else don’t play golf if the outlook is sunny and the humidity is greater than 77.4,

else don’t play golf if the temperature is less than 66.5 and the outlook is for rain,

else don’t play golf if the outlook is for rain and windy is true.

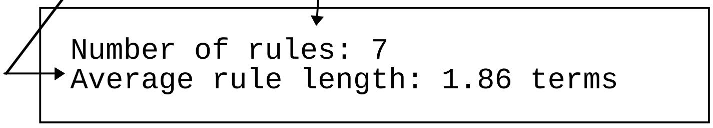
Since Play is the most frequent class value, its rules are listed first. Every rule (except the first one) is made up of conjunctive (an AND) terms (sub-rules). Each term is measured for accuracy. The accuracy is given by the numbers in brackets after each term, for example,

```
[<n>,<n>]
```

If the rule is made from conjunctive terms, the first set of bracketed numbers relates to how many instances the first term classifies. Each term is listed in this way, until the last predicate of a conjunctive rule is listed. Each predicate alters the number of instances the whole rule covers.

The first rule classifies only four instances, where the outlook is overcast and the class value is ‘Play’ (there are only four instances with outlook as overcast). The second rule is harder to interpret. There are four instances where ‘temperature’ is less than 71, but only three of these have the class value ‘Play’. The second predicate of this rule, takes these three instances and applies its rule to them. In this case, each has a temperature of more than 66.4. Therefore, three instances are entirely valid for the second rule. Each of the rules can be read like this.

For the entire rule set, each of the predicates (terms) is counted (this gives 13) and divided by the number of rules (7). This gives an average rule length of 1.86 terms.



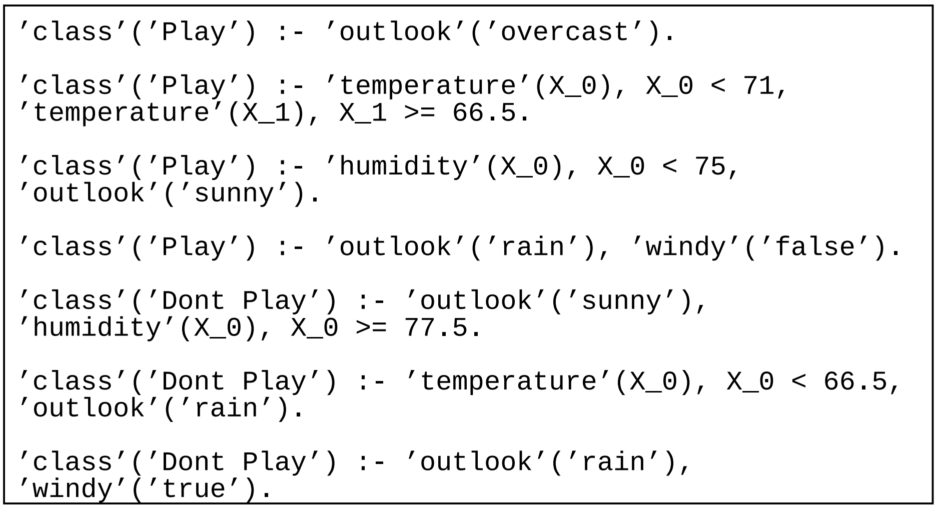
```
Number of rules: 7
Average rule length: 1.86 terms
```

Figure 44. Number of Rules and Averages

InductH PREVAL Evaluation

For InductH, PREVAL outputs PROLOG rules based on InductH rules, output earlier; a summary of the PROLOG rule classifications; a confusion matrix; and summary statistics.

The PROLOG rules based on the InductH rules are listed below. These are the same rules except they are listed in a PROLOG format.



```
’class’(’Play’) :- ’outlook’(’overcast’).

’class’(’Play’) :- ’temperature’(X_0), X_0 < 71,
’temperature’(X_1), X_1 >= 66.5.

’class’(’Play’) :- ’humidity’(X_0), X_0 < 75,
’outlook’(’sunny’).

’class’(’Play’) :- ’outlook’(’rain’), ’windy’(’false’).

’class’(’Dont Play’) :- ’outlook’(’sunny’),
’humidity’(X_0), X_0 >= 77.5.

’class’(’Dont Play’) :- ’temperature’(X_0), X_0 < 66.5,
’outlook’(’rain’).

’class’(’Dont Play’) :- ’outlook’(’rain’),
’windy’(’true’).
```

Figure 45. InductH PROLOG Rules

Below the PROLOG rules, a summary of the rule classifications is given. This lists each rule, its usage, the number and percentage that it correctly classified, and the class of each rule.

```
Summary of prolog rule classifications
--------------------------------------
Rule   Usage            Correct            Class
  1   22.2 %      4    100.0 %     4     Play
  2   16.7 %      3    100.0 %     3     Play
  3   11.1 %      2    100.0 %     2     Play
  4   16.7 %      3    100.0 %     3     Play
  5   16.7 %      3    100.0 %     3     Dont Play
  6    5.6 %      1    100.0 %     1     Dont Play
  7   11.1 %      2    100.0 %     2     Dont Play
```

Figure 46. Summary of InductH PROLOG Rules

For instance, the first rule is Rule 1, which was used 22.2% percent of the time to classify four rules which were all classified correctly with the class value of ‘Play’.

The confusion matrix, similar to the matrices for both T1 and C4.5, indicates how many instances were correctly or incorrectly classified. For the golf dataset, all instances were correctly classified.

```
Confusion Matrix
----------------
        (1)   (2) <- classified as
          9        (1) Play
                5  (2) Dont Play
```

Figure 47. InductH Confusion Matrix

Finally, the classification statistics are listed which indicate the number of instances which were correctly, incorrectly, multiply or not classified

For the golf dataset, with InductH evaluation, every instance was correctly classified.

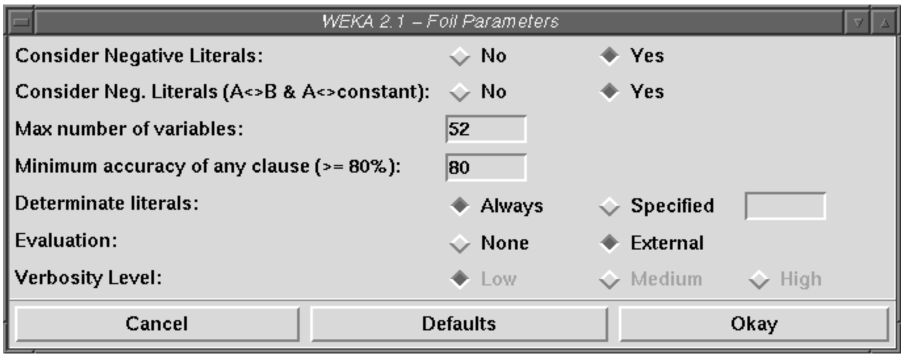
```
Correctly Classified Instances      14   100.0000 %
Incorrectly Classified Instances     0     0.0000 %
UnClassified Instances               0     0.0000 %
Multiply Classified Instances        0     0.0000 %
```

To return to the Machine Learning Workbench window, select **Close** from the InductH results window File menu.

FOIL

From the Scheme menu, choose **FOIL**. Select the **Run FOIL** button and choose the class attribute.

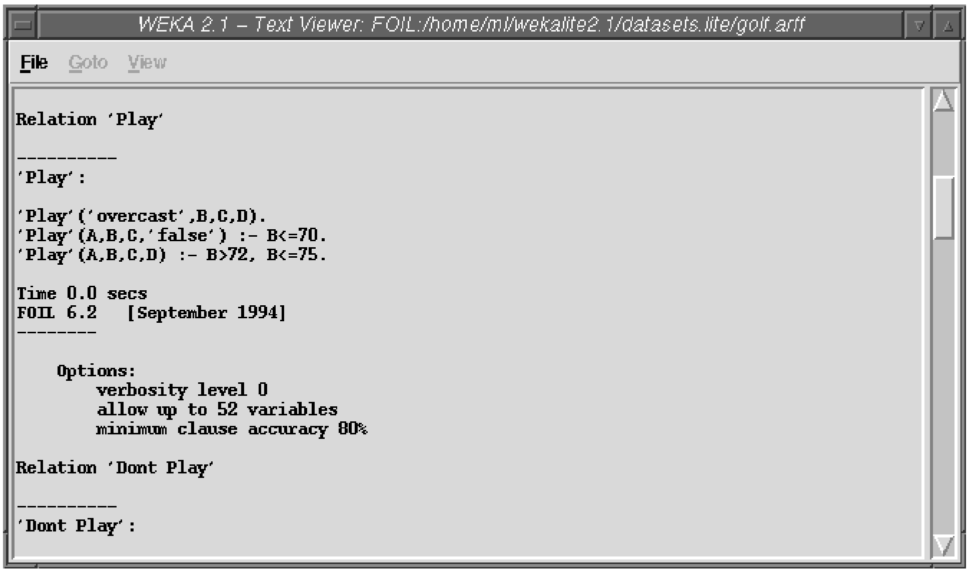
The FOIL Parameters window now appears. Select **Okay**.



Consider Negative Literals: ◇ No ◆ Yes
Consider Neg. Literals (A<>B & A<>constant): ◇ No ◆ Yes
Max number of variables: 52
Minimum accuracy of any clause (>= 80%): 80
Determinate literals: ◆ Always ◇ Specified
Evaluation: ◇ None ◆ External
Verbosity Level: ◆ Low ◇ Medium ◇ High

Figure 48. FOIL Parameters

To view the results, from the Results menu select the italicised option **FOIL(golf.arff).** The Text Viewer window will then appear.


WEKA 2.1 – Text Viewer: FOIL:/home/ml/wekalite2.1/datasets.lite/golf.arff
File
Goto
View

```
Relation 'Play'

----------
'Play':

'Play'('overcast',B,C,D).
'Play'(A,B,C,'false') :- B<=70.
'Play'(A,B,C,D) :- B>72, B<=75.

Time 0.0 secs
FOIL 6.2   [September 1994]
--------

    Options:
        verbosity level 0
        allow up to 52 variables
        minimum clause accuracy 80%

Relation 'Dont Play'

----------
'Dont Play':
```

Figure 49. FOIL results

FOIL Analysis

Like other machine learning schemes, FOIL also produces rules, in the form of 'if-then...else-if' statements. These rules are complex in structure, and involve replacing unimportant attribute values with alphabetic figures. More important symbolic attribute values are left in the rule clause; and more important numerical (real or integer) attribute values are defined after the clause.

Analysing FOIL Results

FOIL has divided the dataset by each of the class values. The class value with the highest occurrence — ‘Play’ — is evaluated first. The rules given for ‘Play’ are:

```
’Play’(’overcast’,B,C,D).
’Play’(A,B,C,’false’) :- B<=70.
’Play’(A,B,C,D) :- B>72, B<=75.
```

Figure 50. Rules for the ‘Play’ Class Value

This first rule indicates that the class is 'Play' if the first attribute — that is, 'outlook' — is 'overcast', regardless of the values of the other attributes. The second rule indicates that the class is 'Play' if the fourth attribute, 'windy', is false, provided that the second, 'temperature', is less than or equal to 70. The third rule indicates that the class is 'Play' regardless of the other attributes provided that 'temperature' is between 72 and 75.

The rules given for ‘Don’t Play’ are:

```
’Dont Play’(’sunny’,B,C,D) :- B>75.
’Dont Play’(’rain’,B,C,’true’).
’Dont Play’(’sunny’,B,C,D) :- C>90.
```

Figure 51. Rules for the ‘Don’t Play’ Class Value

The first rule indicates that the class is 'Don't Play' if the 'outlook' is 'sunny', provided that the 'temperature' is greater than 75. The second rule indicates that the class is 'Don't Play' if the 'outlook' is 'rain' and 'windy' is false. The third indicates that the class is 'Don't Play' if the 'outlook' is 'sunny', provided that the 'humidity' exceeds 90.

FOIL PREVAL Evaluation

For FOIL, PREVAL outputs PROLOG rules based on FOIL rules, output earlier; a summary of the PROLOG rule classifications; a confusion matrix; and summary statistics.

The PROLOG rules based on the FOIL rules are listed on the following page. These are the same rules except they are listed in a PROLOG format.

```
% Rule 1 - Length 1
'class'('Play') :- ( 'outlook'('overcast') ).

% Rule 2 - Length 2
'class'('Play') :- ( 'temperature'(X_2), X_2 =< 70 ),
                   ( 'windy'('false') ).

% Rule 3 - Length 2
'class'('Play') :- ( 'temperature'(X_4), X_4 > 72, X_4
      =< 75 ).

% Rule 4 - Length 2
'class'('Dont Play') :- ( 'outlook'('sunny') ),
      ( 'temperature'(X_6), X_6 > 75 ).

% Rule 5 - Length 2
'class'('Dont Play') :- ( 'outlook'('rain') ),
      ( 'windy'('true') ).

% Rule 6 - Length 2
'class'('Dont Play') :- ( 'humidity'(X_9), X_9 > 90 ),
      ( 'outlook'('sunny') ).
```

Figure 52. FOIL PROLOG Rules

Below the PROLOG rules, a summary of the rule classifications is given. This lists each rule, its usage, the number and percentage that it correctly classified, and the class of each rule.

```
Summary of prolog rule classifications
--------------------------------------
Rule         Usage          Correct      Class
  1    28.6 %      4    100.0 %     4     Play
  2    21.4 %      3    100.0 %     3     Play
  3    14.3 %      2    100.0 %     2     Play
  4    14.3 %      2    100.0 %     2     Dont Play
  5    14.3 %      2    100.0 %     2     Dont Play
  6     7.1 %      1    100.0 %     1     Dont Play
```

Figure 53. Summary of FOIL PROLOG Rule Classifications

For instance, the first rule is Rule 1, which was used 28.6% percent of the time to classify four rules which were all classified correctly with the class value of ‘Play’.

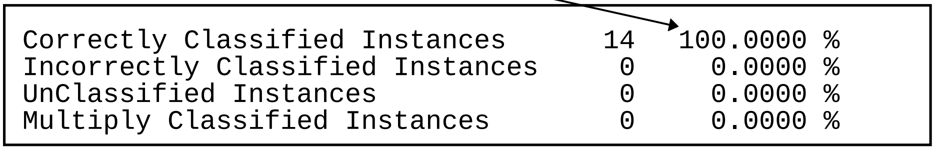
The confusion matrix, similar to the matrix for InductH, indicates how many instances were correctly or incorrectly classified. For the golf dataset, all instances were classified correctly.

```
Confusion Matrix
----------------
        (1)   (2) <- classified as
          9        (1) Play
                5 (2) Dont Play
```

Figure 54. FOIL Confusion Matrix

Finally, the classification statistics are listed which indicate the number of instances which were correctly, incorrectly, multiply or not classified

For the golf dataset, with FOIL evaluation, every instance was correctly classified.



```
Correctly Classified Instances      14   100.0000 %
Incorrectly Classified Instances     0     0.0000 %
UnClassified Instances               0     0.0000 %
Multiply Classified Instances        0     0.0000 %
```

Figure 55. Summary of FOIL rules

To return to the Machine Learning Workbench window, select **Close** from the FOIL results window File menu.

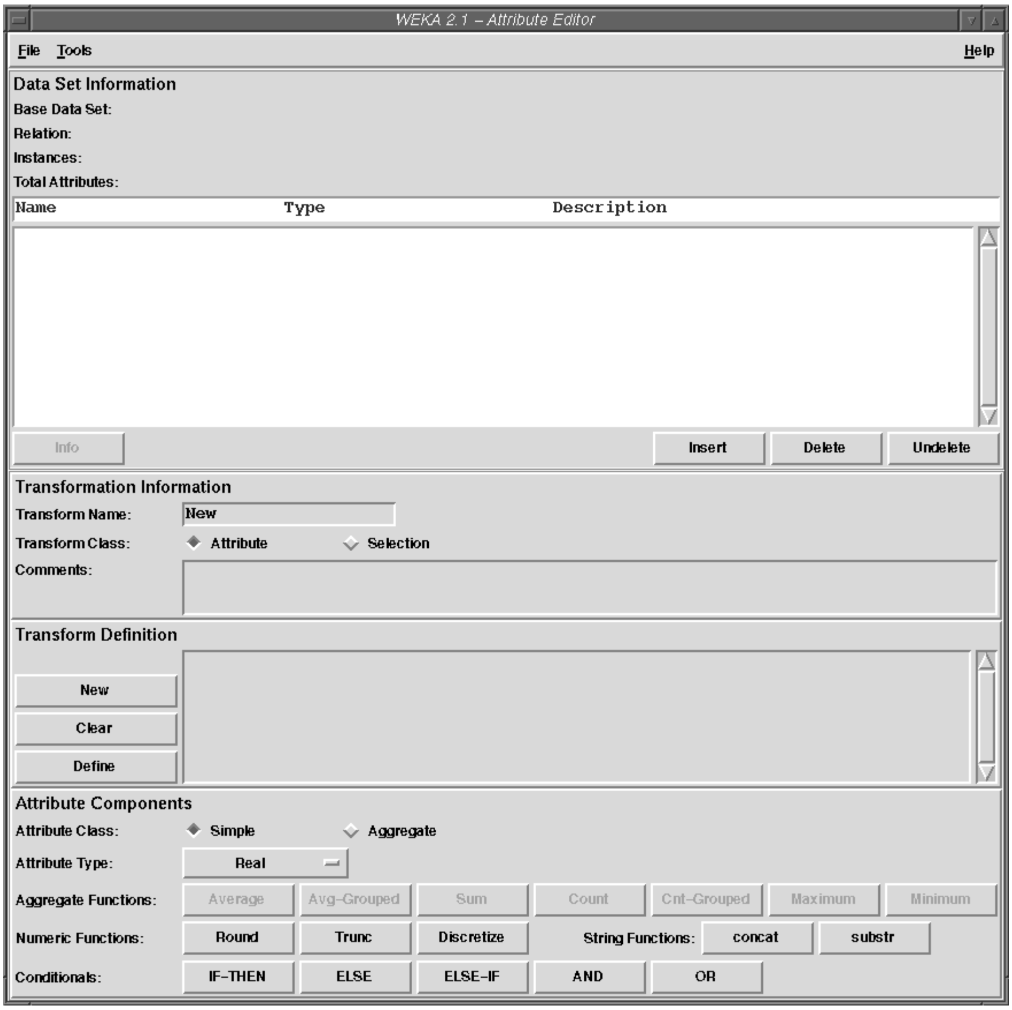
THE ATTRIBUTE EDITOR

The Attribute Editor is used to create new datasets based on existing datasets. This involves either creating a new attribute or selecting a subset of the existing data. Creating a new attribute or selection of the existing data can substantially alter the results that you will get from the scheme analysis. Once the new dataset has been created, it can be analysed on the Machine Learning Workbench.

Starting the Attribute Editor

To start the Attribute Editor, select **Attribute Editor** from the WEKA menu.

The WEKA Attribute Editor window appears on top of the Machine Learning Workbench window.



Transform Name: New

Transform Class: ◆ Attribute ◇ Selection

Attribute Class: ◆ Simple ◇ Aggregate

Attribute Type: Real

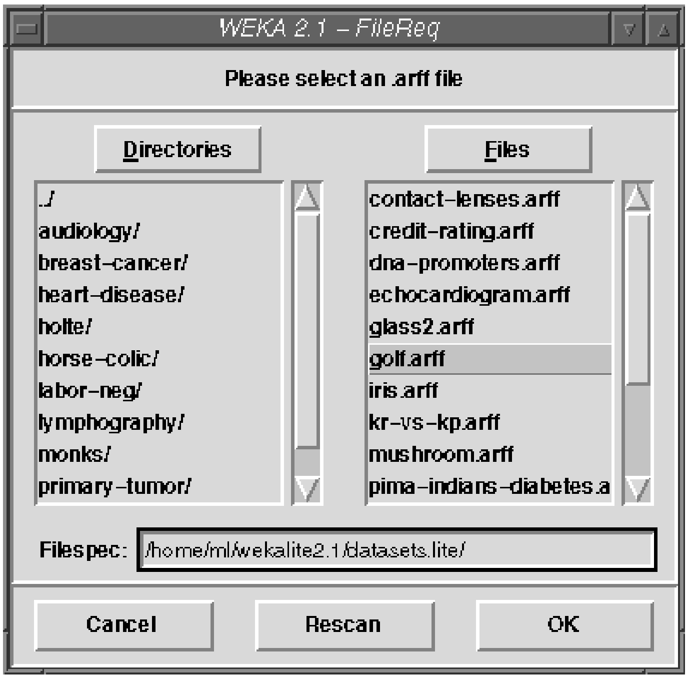
Figure 56. WEKA Attribute Editor

Quitting the Attribute Editor

You may quit from the Attribute Editor, returning to the Machine Learning Workbench window. Do this by selecting **Quit** from the Attribute Editor's File menu.

Opening a Dataset

To open the golf dataset, select **Open Dataset...** from the File menu. A window appears requesting you to select a datafile.

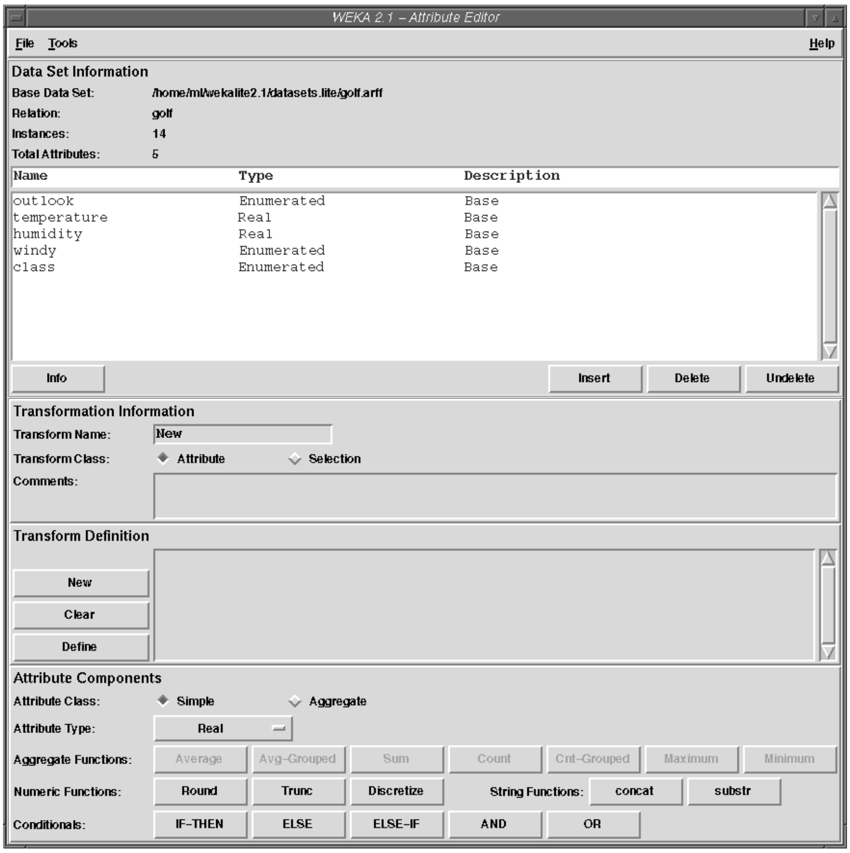


Filespec: /home/ml/wekalite2.1/datasets.lite/

Figure 57. Please select a dataset

Select the **golf.arff** dataset. And press **OK**.

The golf dataset will now be loaded on to the Attribute Editor. You are able to view the attributes in the dataset and access further information about each attribute.



Data Set Information

Base Data Set: /home/ml/wekalite2.1/datasets.lite/golf.arff
Relation: golf
Instances: 14
Total Attributes: 5

Name	Type	Description
outlook	Enumerated	Base
temperature	Real	Base
humidity	Real	Base
windy	Enumerated	Base
class	Enumerated	Base

Transformation Information

Transform Name: New
Transform Class: ● Attribute ○ Selection
Comments:

Transform Definition

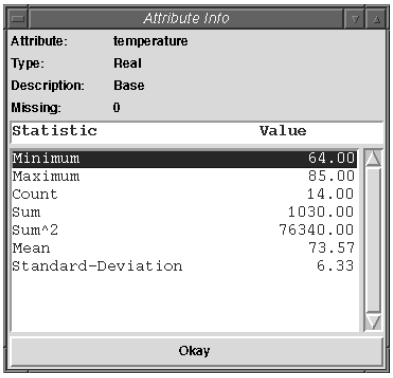
Attribute Components

Attribute Class: ● Simple ○ Aggregate
Attribute Type: Real
Aggregate Functions:
Numeric Functions:
String Functions:
Conditionals:

Figure 58. Attribute Editor — Golf dataset

Viewing the Attribute Information

To view the ‘temperature’ attribute information, select **temperature** from the attribute list, then the **Info** button seen below the attributes.



Attribute: temperature
Type: Real
Description: Base
Missing: 0

Statistic	Value
Minimum	64.00
Maximum	85.00
Count	14.00
Sum	1030.00
Sum^2	76340.00
Mean	73.57
Standard-Deviation	6.33

Figure 59. Info window
for the temperature attribute

As the temperature is a real attribute, the information that you can view is the number of instances (count) the values cover; the minimum and maximum values for the instances; the mean and standard deviation; and a total of all of the values plus the sum squared of all of the values. Select **Okay**.

Creating a New Attribute

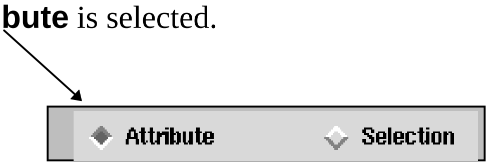
Creating a new attribute adds new dimensionality to your dataset analysis. New attributes can only be created based on existing data in the dataset.[13] Often this means creating a new symbolic (enumerated) attribute based on existing continuous (numerical) attributes. This gives less actual attribute values giving smaller and sometimes more accurate results. Most class values are created in this way.

For the golf dataset, a new binary (two valued) attribute based on the existing 'temperature' attribute can be created. Two temperature categories could be 'cold' and 'hot'; where 'cold' is defined by 'temperature' being less than 75 and 'hot' is greater than 74. The number 75 is chosen as the value to split on as it is half-way between the 'temperature' minimum value (64) and the 'temperature' maximum value (85).

To create a new attribute select the **New** button, which is under the Transform Definition label. Inside the Transformation Information > Transformation Name box enter:

```
New_temperature
```

This gives the new attribute a name. Make sure that the Transform Class — **Attribute** is selected.

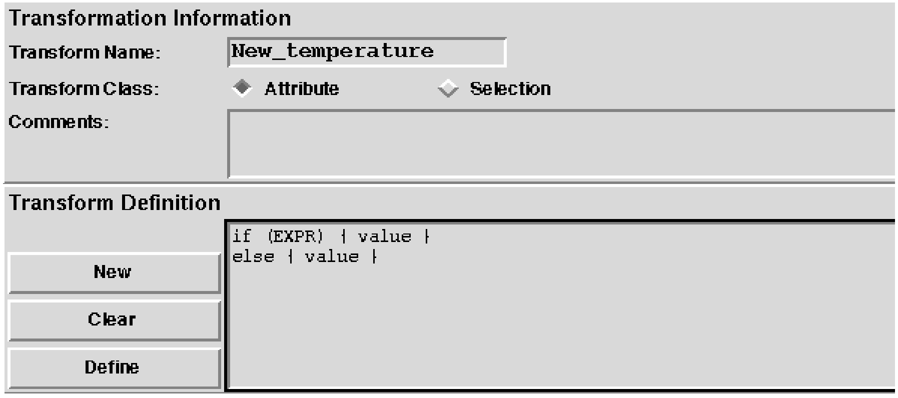


◆ Attribute ◇ Selection

Figure 60. Transformation Class Attribute selection

Click inside the Transform Definition window. From the Attribute Components options select **'IF-THEN'**; 'if (EXPR) { value }' is entered into the Transform Definition window, and the cursor is automatically put on the next line. From the Attribute Components options select **'ELSE'**; the expression 'else { value }' is added to the Transform Definition window. The Transformation Definition is pictured with the new parameters on the next page.

[13] New attributes can be created from more than one existing attribute; these can be a combination of both discrete and continuous attributes.



Transformation Information

Transform Name: New_temperature

Transform Class: ● Attribute ◇ Selection

Comments:

Transform Definition

```
if (EXPR) { value }
else { value }
```

Figure 61. Transformation Definition Window
with new attribute parameters

The expression:

```
    if(EXPR) {value}
    else {value}
```

defines the basis for the New_temperature attribute creation expression.

In the Transform Definition window select and delete ‘**(EXPR)**’. Leaving your cursor between the two brackets, select and double-click the attribute **‘temperature’** from the attributes list. You have selected ‘temperature’ as the basis New_temperature attribute. The attribute name ‘temperature’ is automatically entered between the two brackets. After ‘temperature’ type in:

```
    < 75.0
```

Move the cursor to the first { value }. Select the word **‘value’**, and enter:

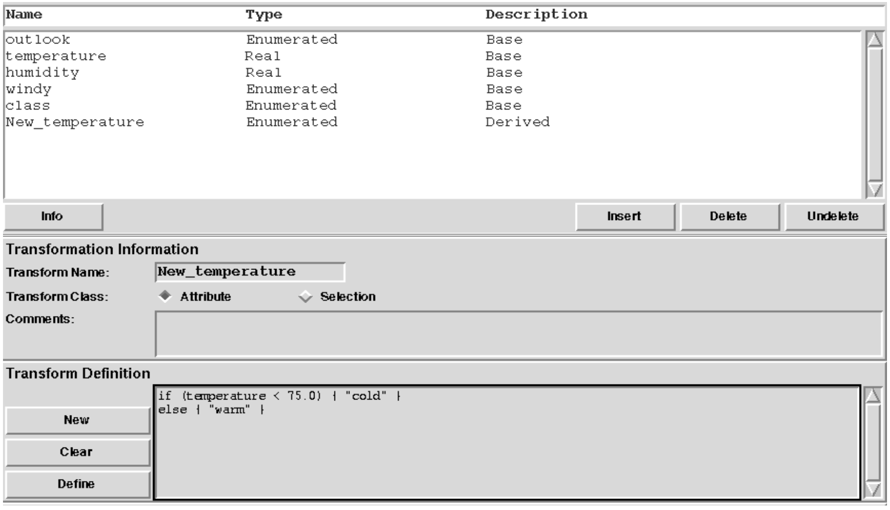
```
    “cold”
```

(make sure that you enter this value in double quotes). That defines your first attribute value. Move your cursor to the second line, and select the word **‘value’** on it, and enter:

```
    “warm”
```

(double quotes are required here as well). Now your second attribute value has been defined as well.

The **Attribute Type** shows Real as its current value. New_temperature needs to be defined as a discrete (enumerated) attribute as it contains symbolic values. Select the **Attribute Type** button and choose **Enumerated** from the given list. Then select the button **Define** under the Transform Definition label. Notice that the New_temperature attribute is now listed in the attributes list.



Name	Type	Description
outlook	Enumerated	Base
temperature	Real	Base
humidity	Real	Base
windy	Enumerated	Base
class	Enumerated	Base
New_temperature	Enumerated	Derived

Transformation Information

Transform Name: New_temperature

Transform Class: ● Attribute ○ Selection

Comments:

Transform Definition

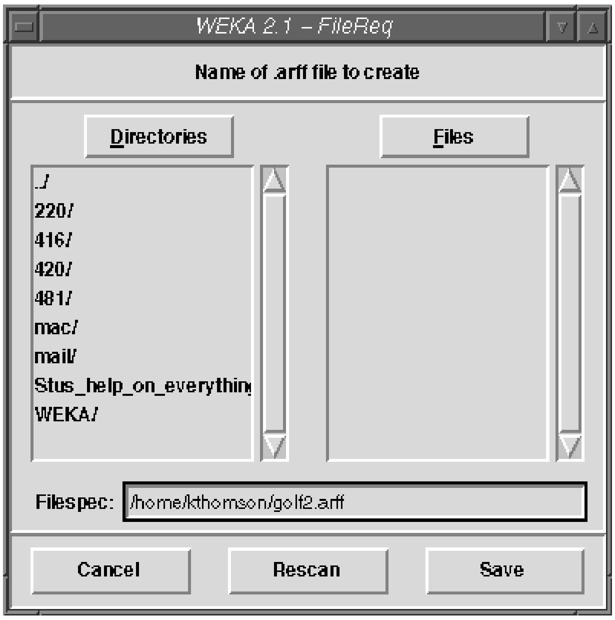
```
if (temperature < 75.0) | "cold" |
else | "warm" |
```

Figure 62. Transform Definition Window after new attribute has been defined

From the File menu, select **Save Dataset...** A window appears requesting you to give a name for your new dataset. Enter your new file name as:

```
golf2.arff
```

and select **Save**.



WEKA 2.1 – FileReq

Name of .arff file to create

Filespec: /home/kthomson/golf2.arff

Figure 63. Save the name of the new ARFF datafile

Another window appears, asking if you want to save the Perl script for later use. As this is unlikely select **No**.

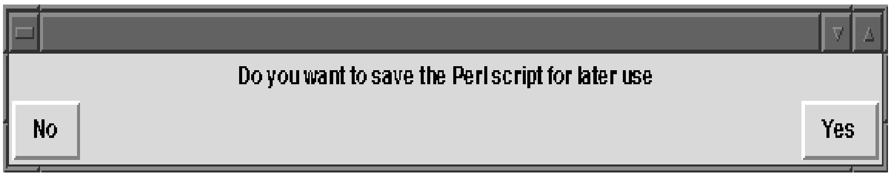
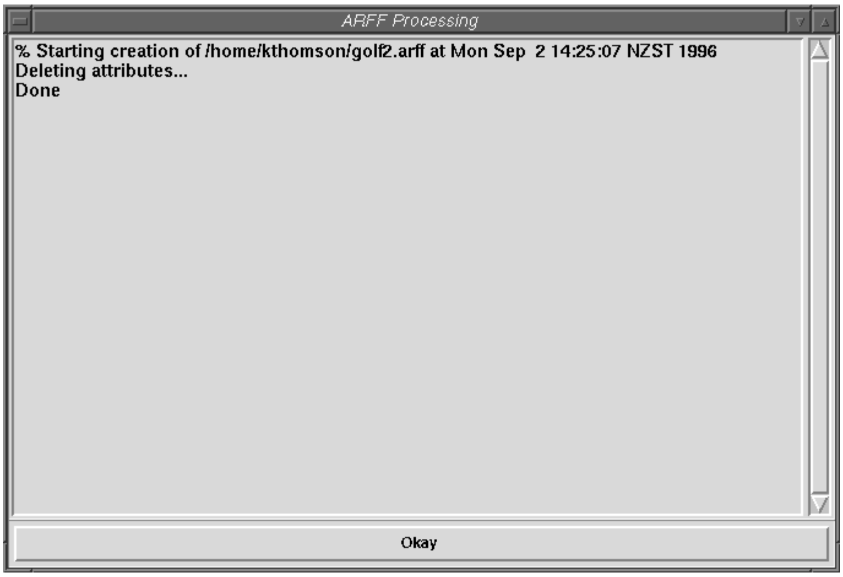

Do you want to save the Perl script for later use

Figure 64. Save Perl script for later use

The 'ARFF Processing' window appears, which indicates WEKA is analysing the golf2 datafile. When the scheme analysis has been completed, this window will display Done.



```
% Starting creation of /home/kthomson/golf2.arff at Mon Sep  2 14:25:07 NZST 1996
Deleting attributes...
Done
```

Figure 65. ARFF Processing window

Select **Okay**.

Finally another window appears asking if you want to load your new file. Select **Yes**.

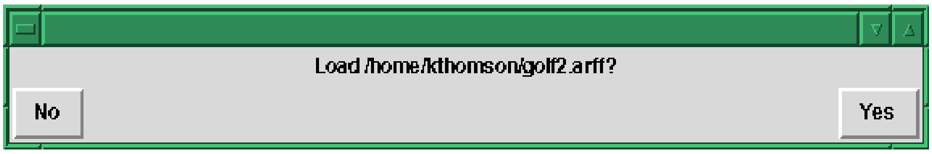
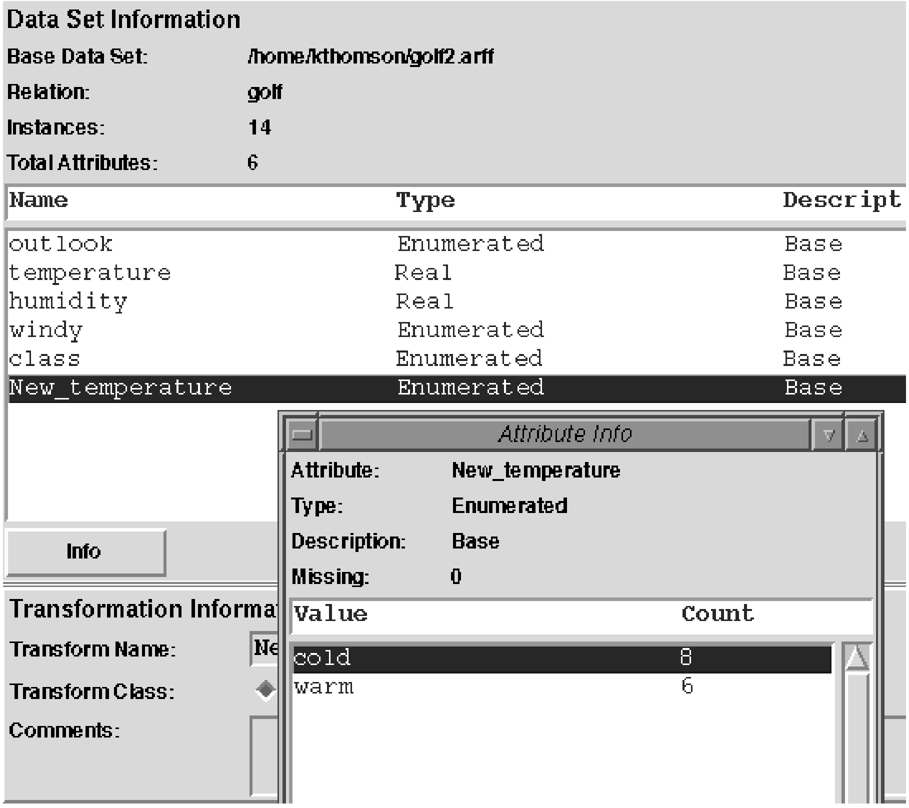

Load /home/kthomson/golf2.arff?

Figure 66.   Load new ARFF file

Your new file will now have been loaded into the attribute editor.

You may view details about your newly created attribute by selecting **New_temperature** and then the **Info** button seen under the attributes list.



Data Set Information

Base Data Set:	/home/kthomson/golf2.arff
Relation:	golf
Instances:	14
Total Attributes:	6

Name	Type	Descript
outlook	Enumerated	Base
temperature	Real	Base
humidity	Real	Base
windy	Enumerated	Base
class	Enumerated	Base
New_temperature	Enumerated	Base

Attribute Info

Attribute:	New_temperature
Type:	Enumerated
Description:	Base
Missing:	0

Value	Count
cold	8
warm	6

Figure 67. Window with new attribute and information.

Creating a Selection

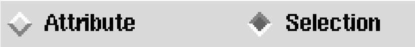
Although you may select exactly which attributes you want to include in your dataset when you are analysing it on the Machine Learning Workbench, the Attribute Editor allows you to create a new subset of your existing dataset based on a particular attribute's value.

For example, you will be creating a subset of the dataset based on the 'windy' attribute. By selecting only the data where 'windy' is 'false', this effectively removes all instances from your new dataset where 'windy' does not equal 'false'. Like creating a new attribute this means that any new analysis will generate different results than those from your original dataset.

Select the **New** button, which is under the Transform Definition label. Inside the Transformation Information > Transformation Name box enter

```
Not_Windy
```

This gives your new selection a name. Make sure that the Transform Class — **Selection** is selected.



◇ Attribute ◆ Selection

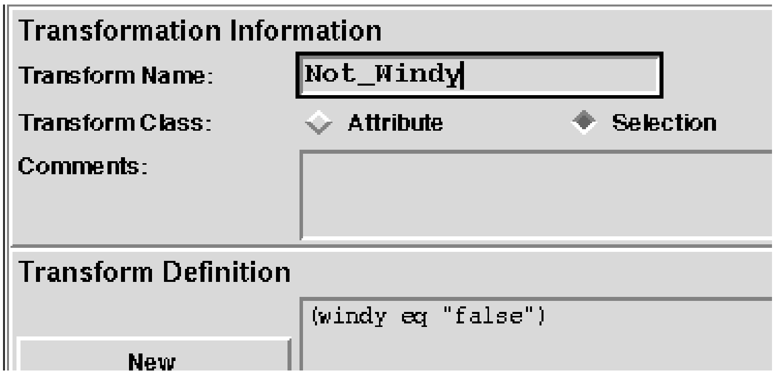
Figure 68. Transformation Class Selection

Click inside the Transform Definition window and enter the following arguments:

```
(windy eq “false”)
```

This expression defines the basis for the Not_Windy selection creation expression.

Then select the button **Define** under the Transform Definition label. Notice that your new selection is now listed in the attributes list.



Transformation Information

Transform Name: Not_Windy

Transform Class: ◇ Attribute ◆ Selection

Comments:

Transform Definition

(windy eq "false")

Figure 69. Transform Definition Window
with new selection definition

From the File menu, select **Save Dataset...** A window appears requesting you to give a name for your new dataset. Enter your new file name as:

```
golf3.arff
```

and select **Save**.

Another window appears, asking if you want to save the Perl script for later use. As this is unlikely, select **No**.

The ARFF Processing window appears, which indicates WEKA is analysing the golf3 datafile. When the analysis has finished, this window will display Done.

Select **Okay**.

Finally another window appears asking if you want to load your new file. Select **Yes**.

Your new file will be loaded into the attribute editor.

Notice that there are now eight instances in the dataset.

Exit the Attribute Editor. Both golf2 and golf3 datasets can be analysed in the same way as the original golf datafile by the Machine Learning Workbench.[14][15]

[14] When analysing golf2 on the Machine Learning the old temperature attribute needs to be removed from the existing dataset.

[15] For further Attribute Editor components that can be used for defining a new attribute reference the WEKA: Machine Learning Workbench User Manual

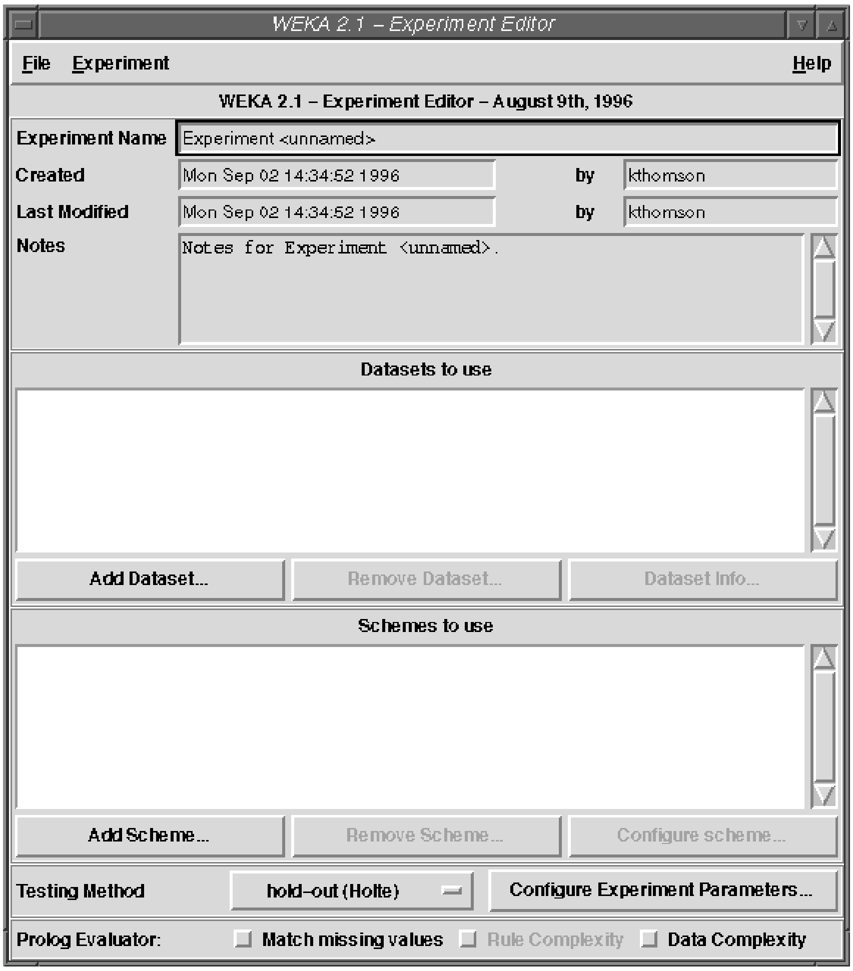
The Experiment Editor

The Experiment Editor is an analysis tool provided for machine learning experts who want to measure how accurate and fast a scheme is when analysing a number of different datasets. The Experiment Editor is used for running a number of different datasets through a number of different schemes.

Starting the Experiment Editor

To start the Experiment Editor, select **Experiment Editor** from the WEKA menu.

The WEKA Experiment Editor window appears, on top of the Machine Learning Workbench window.



WEKA 2.1 – Experiment Editor – August 9th, 1996

Experiment Name: Experiment <unnamed>

Created: Mon Sep 02 14:34:52 1996 by kthomson

Last Modified: Mon Sep 02 14:34:52 1996 by kthomson

Notes: Notes for Experiment <unnamed>.

Testing Method: hold-out (Holte)

Prolog Evaluator: ☐ Match missing values ☐ Rule Complexity ☐ Data Complexity

Figure 70. WEKA Experiment Editor

Quitting the Experiment Editor

You may quit from the Experiment Editor, returning to the Machine Learning Workbench window, by selecting **Quit** from the Experiment Editor’s File menu.

Creating a New Experiment

To create a new experiment, select **New Experiment** from the Experiment Editor's File menu. This automatically sets the environment for you to enter your experiment name and definition, and to enter your dataset(s) and schemes.

Experiment Name

You will notice that the cursor has been placed in the Experiment Name box for you. You are required to give an experiment a name to identify the experiment from others.

This experiment is based on the golf dataset. In the box enter a name for your golf experiment, an appropriate name is:

```
Golf
```

This is the name of your new golf experiment.

Created and Last Modified

Two labels are listed below the Experiment Editor — Created and Last Modified. You can not enter any information in these fields — they will be automatically updated for you. To the right of these fields are 'by' fields which automatically show your user name. This cannot be updated either.

The created fields specify the date and time that your experiment was created, and later, if you modify the experiment the last modified field will be updated with similar information.

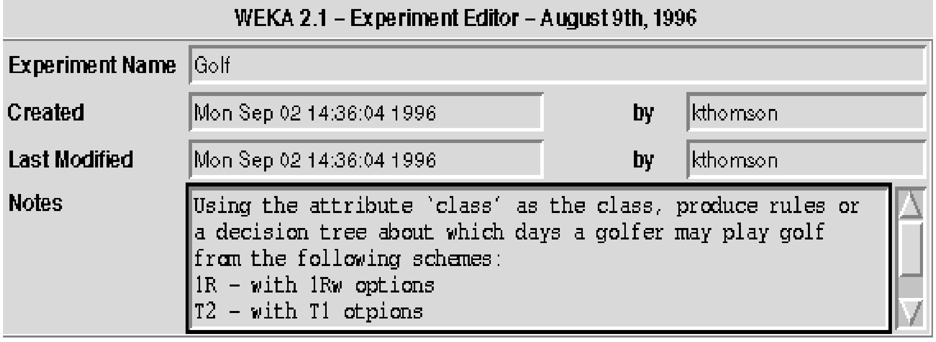
Notes

Below the Created and Last Modified labels is an area for entering experimental notes. This enables you to keep track of not only what was done in the experiment but also why they were performed. Recall the goals and objectives defined earlier? Enter the objectives given for the golf experiment in the Notes box (these are listed on the following page).

```
Using the attribute 'class' as the class,
produce rules or a decision tree about which
days a golfer may play golf from the following
schemes.
```

SCHEME	CONFIGURATION
1R	with 1Rw options
T2	with T1 options
C4.5	with tree and rule output, internal evaluation, for analysis on the Machine Learning Workbench, and default settings for analysis with the Experiment Editor.[16]
InductH	with default options
FOIL	with default options

The Experiment Editor window should now look like Figure 70, which shows the experiment name, and the created and last modified, 'by' and notes fields with golf information.



WEKA 2.1 – Experiment Editor – August 9th, 1996

Experiment Name: Golf

Created: Mon Sep 02 14:36:04 1996 by: kthomson

Last Modified: Mon Sep 02 14:36:04 1996 by: kthomson

Notes:

```
Using the attribute 'class' as the class, produce rules or
a decision tree about which days a golfer may play golf
from the following schemes:
1R - with 1Rw options
T2 - with T1 otpions
```

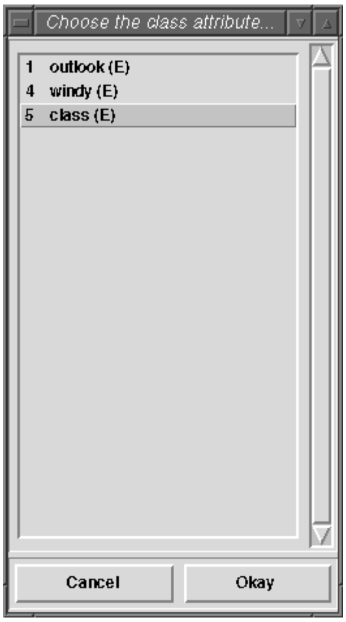
Figure 71. Experiment Editor with Golf Experiment Information

Datasets to Use

You can select any number of datasets to analyse through the Experiment Editor. However, for the golf dataset you will be using one dataset. To add this, select the **Add Dataset...** option and the Load an ARFF file window will appear. From the files column select **golf.arff**, then **OK**.

Another window, Choose the class attribute... appears. In it is a list of the enumerated variables. Select the golf dataset's class attribute; this is represented by **5 class(E).** Like analysing this dataset on the Machine Learning Workbench, the class attribute is the attribute you will be basing your decisions upon.

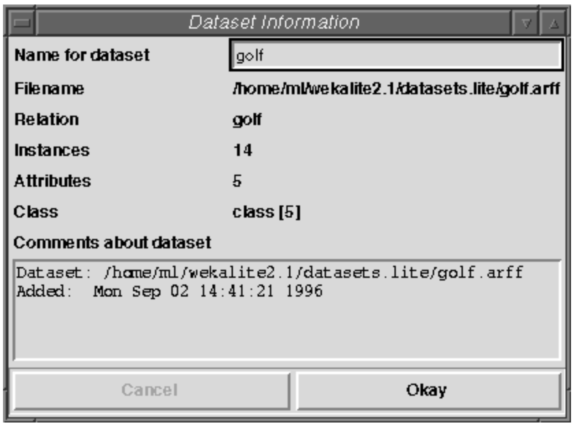
[16] Output from the Experiment Editor is not graphical; therefore, it will not be possible to output a decision tree or rules.



1 outlook (E)
4 windy (E)
5 class (E)

Figure 72. Experiment Editor —
Choose the class attribute

The Dataset Information window appears. The information here includes the dataset name; path and file name; relation name; total number of instances and attributes; the class value; and comments about the dataset just selected, which in this instance includes the path and file name of the golf.arff file and the data the golf datafile was added to the current experiment.



Name for dataset	golf
Filename	/home/ml/wekalite2.1/datasets.lite/golf.arff
Relation	golf
Instances	14
Attributes	5
Class	class [5]

Comments about dataset

```
Dataset: /home/ml/wekalite2.1/datasets.lite/golf.arff
Added:  Mon Sep 02 14:41:21 1996
```

Figure 73. Experiment Editor —
Dataset Information

Select **Okay** (notice that you are unable to highlight or select Cancel). Now look at the updated Experiment Editor window. The name of the golf dataset has been added to the Datasets to use box. If you would like to view the golf dataset information again, select the dataset then the **Dataset Info...** button.

Figure 74. Dataset Info... button

You are able to either Cancel or Update this information. In this case, as you have no updated any information, select**Cancel**.

Schemes to Use

You can select any number of schemes to analyse the golf dataset with on the Experiment Editor. The schemes you will be using have been previously defined within Chapter 1: Experiment Objectives and are also listed in the notes box for the current experiment.

These need to be added. From the Experiment Editor select the **Add Scheme** option. Another window, Choose a ML scheme..., appears. Select each of the schemes required in turn.

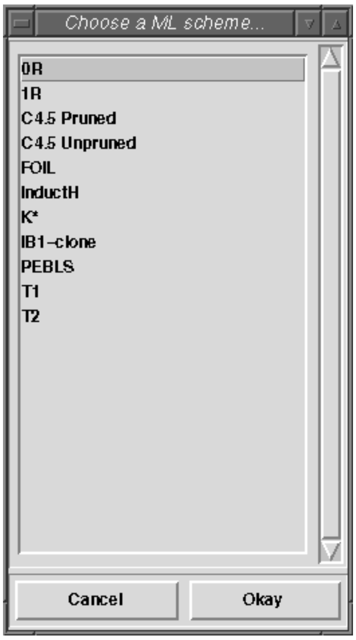
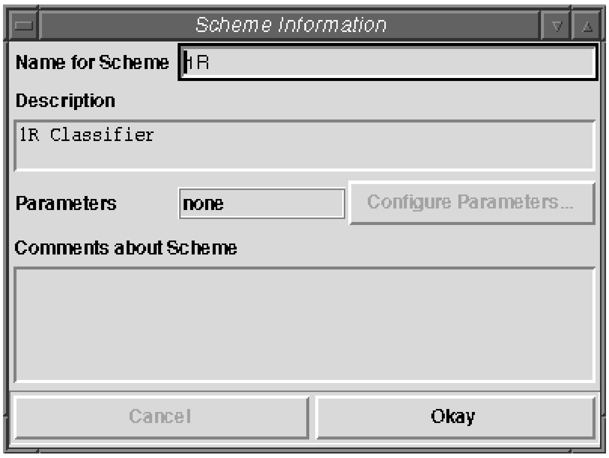


Figure 75. Experiment Editor — Choose a ML scheme...

Do this by selecting:

1R and then **Okay**.



Scheme Information

Name for Scheme: 1R

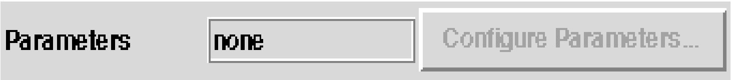
Description: 1R Classifier

Parameters: none

Comments about Scheme:

Figure 76. Experiment Editor —
1R Scheme Information

The scheme window appears. It describes what scheme has just been selected and any special parameters that will be used when it will analyse the golf dataset. You want to run 1R with 1Rw parameters; 1Rw is not the default configuration and will need to be selected. To configure a scheme used in the Experiment Editor select the **Configure Parameters...** option.



Parameters: none

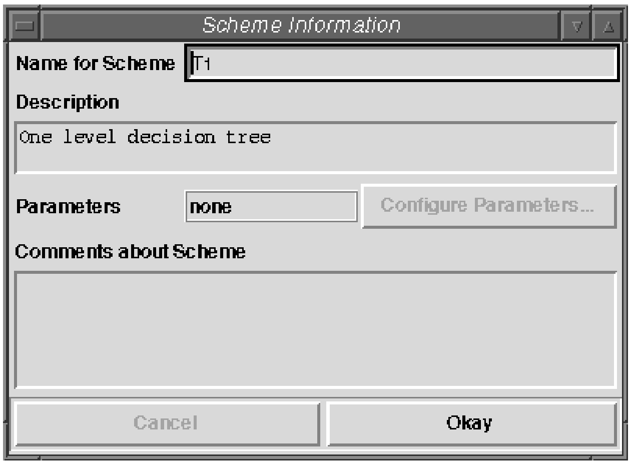
Figure 77. Schemes to Use options — Configure Parameters...is unable to be selected

However, on this configuration window you are unable to select the Configuration Parameters... option (it is viewed in grey) and configure any of 1R's parameters; specifically, you are unable to configure 1R to 1Rw options. Also notice that you are unable to select cancel. In this case, once the scheme has been added to the current experiment it will need to be deleted.

Press **Okay** on Scheme Information window. The scheme, 1R is now listed under Schemes to Use. However, you don't want 1R with these options, so delete the scheme from the experiment by selecting **1R** in the Schemes to Use list, then **Remove Scheme...** 1R has now been deleted from the Schemes to Use list.

Figure 78. Remove Scheme... button

T1, then select **Okay**.



Name for Scheme T1

Description

One level decision tree

Parameters none

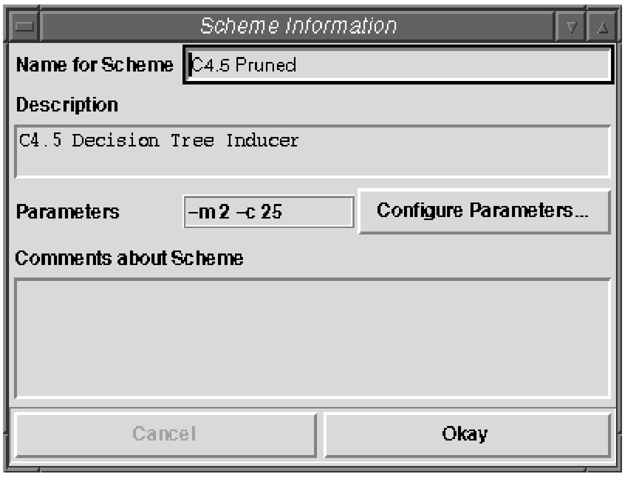
Figure 79. Experiment Editor —
T1 Scheme Information

The scheme information window appears. It describes the scheme that has just been selected and any special parameters that will be used when it will analyse the golf dataset.

The parameters shown are the required T1 parameters specified earlier. Select **Okay** from the Scheme information window. T1 has now been added as a scheme.

C4.5 Pruned, then select **Okay**

The scheme information window appears. It describes the C4.5 scheme and any special parameters that will be used when it analyses the golf dataset.



Name for Scheme C4.5 Pruned

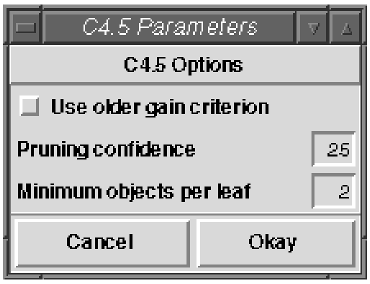
Description

C4.5 Decision Tree Inducer

Parameters -m 2 -c 25

Figure 80. Experiment Editor — C4.5 Scheme Information

The C4.5 Scheme Information window allows configuration of the C4.5 parameters. You can do this by selecting the **Configure Parameters** option.



C4.5 Options

□ Use older gain criterion

Pruning confidence: 25

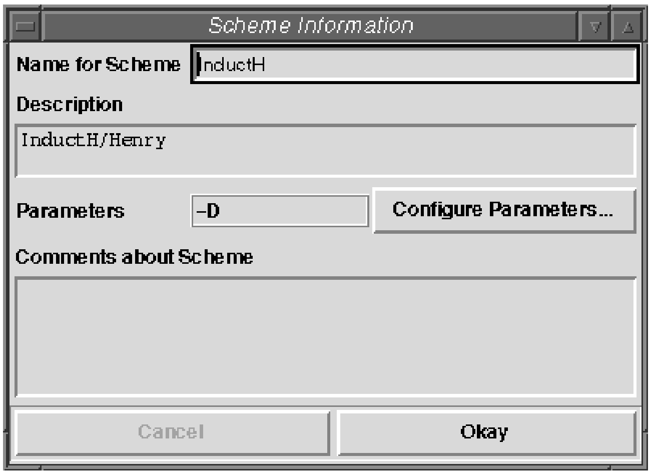
Minimum objects per leaf: 2

Figure 81. Experiment Editor —
C4.5 Scheme Information

The options shown are the required C4.5 parameters. In the Experimental Goals the C4.5 parameters were required to be the default options. On the C4.5 Parameters window, select **Cancel** as the parameters have not been altered in any way.

You will now be back at the C4.5 Pruned Scheme Information window. Select **Okay**.

InductH — then **Okay**.



Name for Scheme: InductH

Description: InductH/Henry

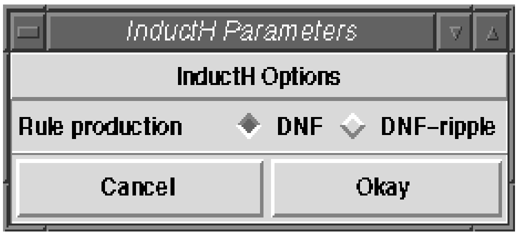
Parameters: -D

Comments about Scheme

Figure 82. Experiment Editor —
InductH Scheme Information

The InductH Scheme Information window allows configuration of the InductH parameters. You can do this by selecting the **Configure Parameters** option.

When you are inside the InductH Parameters window select **DNF** as the default option for InductH, then **Okay**.



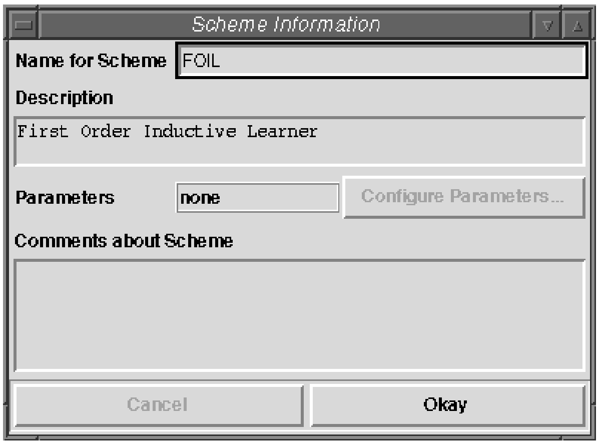
Rule production ◆ DNF ◇ DNF-ripple

Figure 83. InductH Configuration window

These are the required parameters (default settings).

Select **Okay**.

FOIL, and then **Okay**.



Name for Scheme: FOIL

Description: First Order Inductive Learner

Parameters: none

Comments about Scheme

Figure 84. Experiment Editor — FOIL Scheme Information

Although you are unable to configure FOIL further, the parameters given are the default settings; in this case, the required parameters.

Select **Okay**.

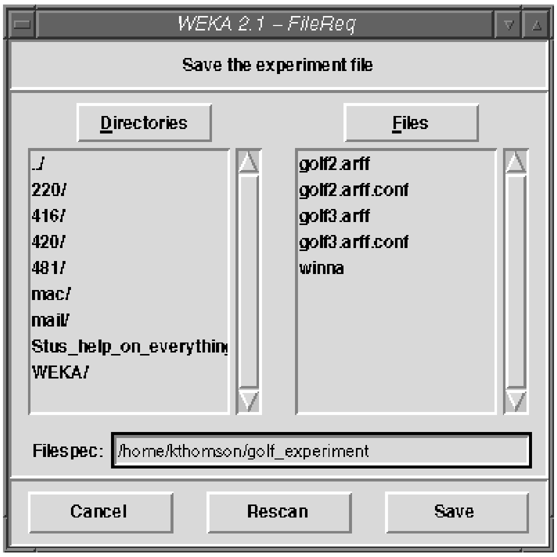
Saving the Experiment

You need to save your new experiment so that you can reference it later if required. From the File menu select **Save Experiment**. A Save the Experiment window will appear, allowing you to select the path and file name in which you wish to save the experiment.

Choose your directory and give the experiment a name:

```
golf_experiment
```

and select **Save**.



Save the experiment file

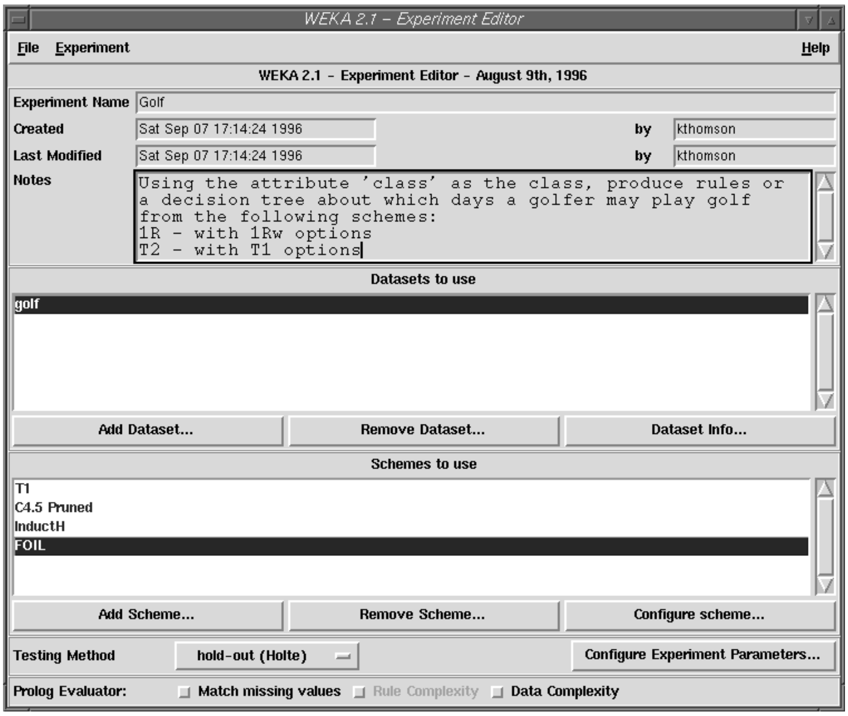
../
220/
416/
420/
481/
mac/
mail/
Stus_help_on_everythin
WEKA/

golf2.arff
golf2.arff.conf
golf3.arff
golf3.arff.conf
winna

Filespec: /home/kthomson/golf_experiment

Figure 85. Experiment Editor —
Save the Experiment File

Before running the experiment, the final Experiment Editor window will look like this:



WEKA 2.1 – Experiment Editor – August 9th, 1996

Experiment Name: Golf
Created: Sat Sep 07 17:14:24 1996 by kthomson
Last Modified: Sat Sep 07 17:14:24 1996 by kthomson
Notes:

```
Using the attribute 'class' as the class, produce rules or
a decision tree about which days a golfer may play golf
from the following schemes:
1R - with 1Rw options
T2 - with T1 options
```

Datasets to use

golf

Schemes to use

T1
C4.5 Pruned
InductH
FOIL

Testing Method: hold-out (Holte)

Prolog Evaluator: □ Match missing values □ Rule Complexity □ Data Complexity

Figure 86. Experiment Editor with Golf Information

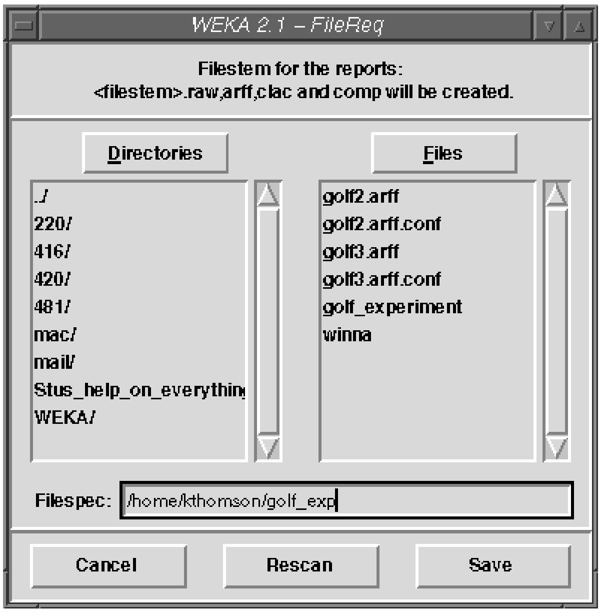
Running the Experiment

From the Experiment menu, select **Run Experiment...** A window will appear asking you to give a file name for the Golf reports. Select the required directory and enter:

```
run_golf
```

as the file name. This file name will be used as a ‘filestem’ for files that WEKA will generate to describe the results of your analysis experiment. Files with extensions .raw, .arff, .clac, and .comp will be created for this ‘filestem’. These files are for:

- <filestem.raw> - raw experimental results
- <filestem.arff> - raw results in ARFF format
- <filestem.clac> - classification accuracy report
- <filestem.comp> - PREVAL complexity report.



Filestem for the reports:
<filestem>.raw,arff,clac and comp will be created.

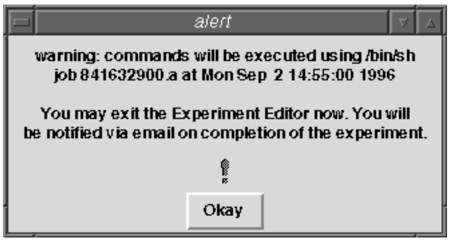
```
../
220/
416/
420/
481/
mac/
mail/
Stus_help_on_everythin
WEKA/
```

```
golf2.arff
golf2.arff.conf
golf3.arff
golf3.arff.conf
golf_experiment
winna
```

Filespec: /home/kthomson/golf_exp

Figure 87. Experiment Editor —
Filestem for the Reports

When you have saved the run_golf file, another window appears telling you that the results of your Golf experiment will be posted to you via electronic mail.



```
warning: commands will be executed using /bin/sh
job 841632900.a at Mon Sep  2 14:55:00 1996
```

You may exit the Experiment Editor now. You will
be notified via email on completion of the experiment.

Figure 88. Experiment Editor —
alert

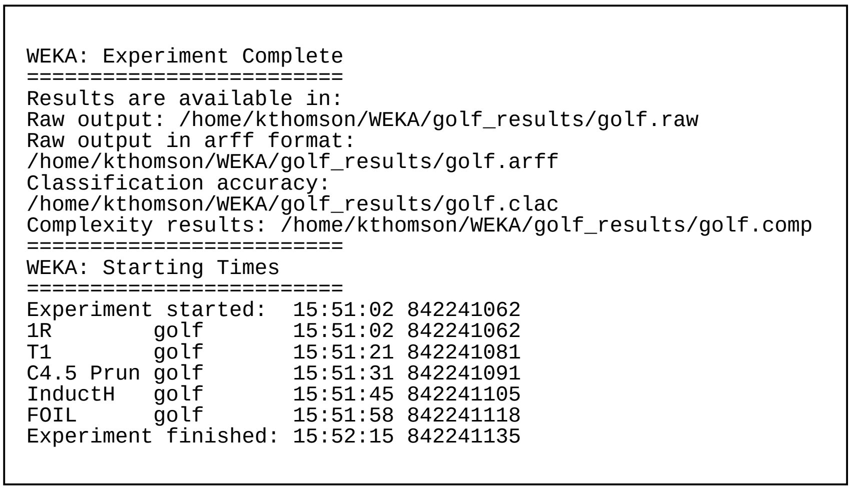
Select **Okay**. This returns you to the Experiment Editor.

Quit from the Experiment Editor and return to the Machine Learning Workbench window.

The results from the experiment are listed below.

Analysing the Experiment Editor Results

The results from the Experiment Editor can be viewed from within a text editor. First though, a message indicating that the schemes have finished analysing the datasets will be sent to you.



```
WEKA: Experiment Complete
=========================
Results are available in:
Raw output: /home/kthomson/WEKA/golf_results/golf.raw
Raw output in arff format:
/home/kthomson/WEKA/golf_results/golf.arff
Classification accuracy:
/home/kthomson/WEKA/golf_results/golf.clac
Complexity results: /home/kthomson/WEKA/golf_results/golf.comp
=========================
WEKA: Starting Times
=========================
Experiment started:  15:51:02 842241062
1R        golf       15:51:02 842241062
T1        golf       15:51:21 842241081
C4.5 Prun golf       15:51:31 842241091
InductH   golf       15:51:45 842241105
FOIL      golf       15:51:58 842241118
Experiment finished: 15:52:15 842241135
```

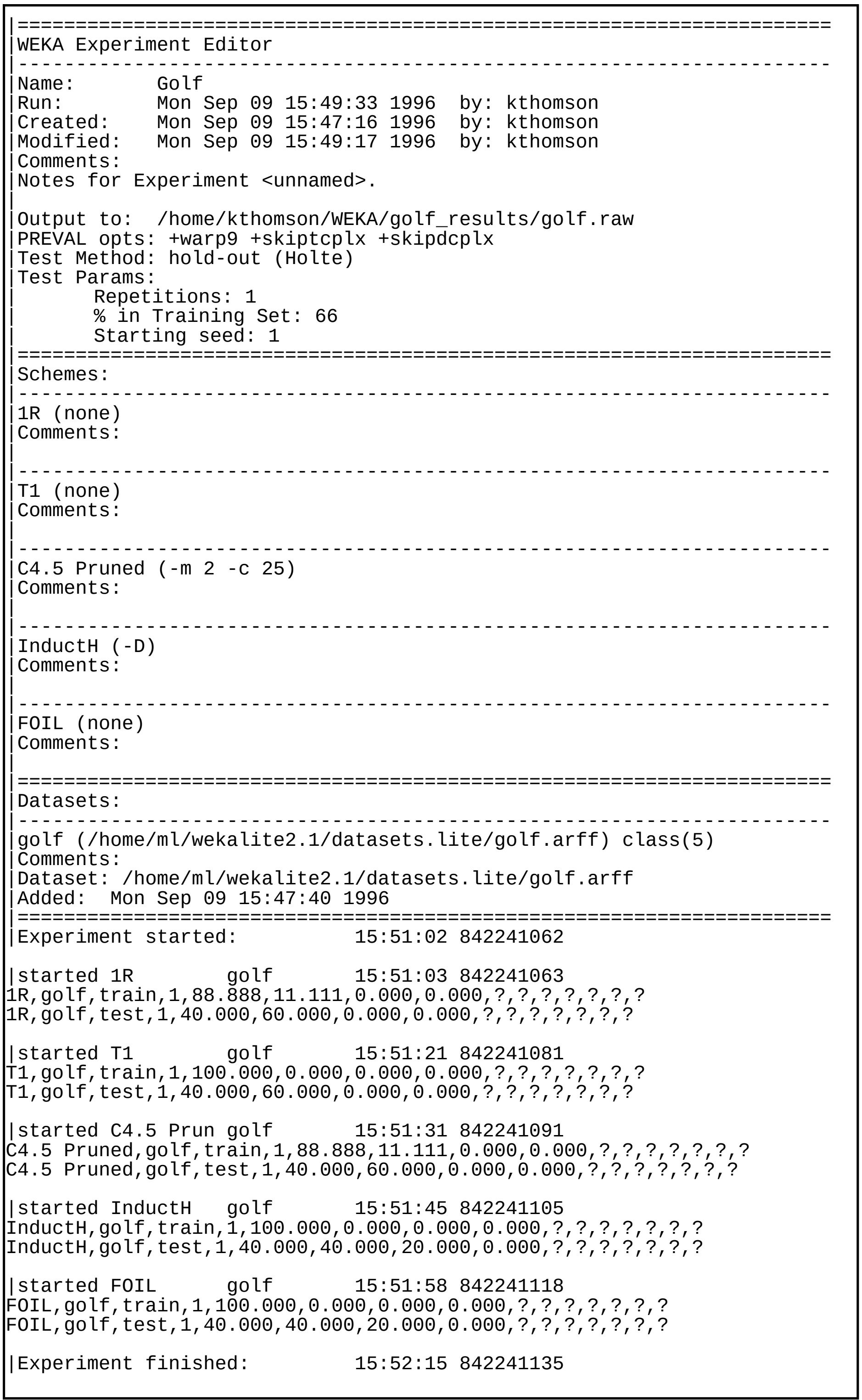
Figure 89. Experiment Complete Message

Once this has occurred you may view your results.

The results are mailed in four different files:
golf_exp.raw containing raw experimental results;
golf_exp.arff containing the raw results in ARFF format;
golf_exp.clac which is a classification accuracy report; and
golf_exp.comp which is a PREVAL complexity report.

The golf_exp.raw File

The .raw file outputs raw experimental results. It has the following output:



```
|=====================================================================
|WEKA Experiment Editor
|---------------------------------------------------------------------
|Name:       Golf
|Run:        Mon Sep 09 15:49:33 1996  by: kthomson
|Created:    Mon Sep 09 15:47:16 1996  by: kthomson
|Modified:   Mon Sep 09 15:49:17 1996  by: kthomson
|Comments:
|Notes for Experiment <unnamed>.
|
|Output to:  /home/kthomson/WEKA/golf_results/golf.raw
|PREVAL opts: +warp9 +skiptcplx +skipdcplx
|Test Method: hold-out (Holte)
|Test Params:
|       Repetitions: 1
|       % in Training Set: 66
|       Starting seed: 1
|=====================================================================
|Schemes:
|---------------------------------------------------------------------
|1R (none)
|Comments:
|
|---------------------------------------------------------------------
|T1 (none)
|Comments:
|
|---------------------------------------------------------------------
|C4.5 Pruned (-m 2 -c 25)
|Comments:
|
|---------------------------------------------------------------------
|InductH (-D)
|Comments:
|
|---------------------------------------------------------------------
|FOIL (none)
|Comments:
|
|=====================================================================
|Datasets:
|---------------------------------------------------------------------
|golf (/home/ml/wekalite2.1/datasets.lite/golf.arff) class(5)
|Comments:
|Dataset: /home/ml/wekalite2.1/datasets.lite/golf.arff
|Added:  Mon Sep 09 15:47:40 1996
|=====================================================================
|Experiment started:          15:51:02 842241062

|started 1R        golf       15:51:03 842241063
1R,golf,train,1,88.888,11.111,0.000,0.000,?,?,?,?,?,?,?
1R,golf,test,1,40.000,60.000,0.000,0.000,?,?,?,?,?,?,?

|started T1        golf       15:51:21 842241081
T1,golf,train,1,100.000,0.000,0.000,0.000,?,?,?,?,?,?,?
T1,golf,test,1,40.000,60.000,0.000,0.000,?,?,?,?,?,?,?

|started C4.5 Prun golf       15:51:31 842241091
C4.5 Pruned,golf,train,1,88.888,11.111,0.000,0.000,?,?,?,?,?,?,?
C4.5 Pruned,golf,test,1,40.000,60.000,0.000,0.000,?,?,?,?,?,?,?

|started InductH   golf       15:51:45 842241105
InductH,golf,train,1,100.000,0.000,0.000,0.000,?,?,?,?,?,?,?
InductH,golf,test,1,40.000,40.000,20.000,0.000,?,?,?,?,?,?,?

|started FOIL      golf       15:51:58 842241118
FOIL,golf,train,1,100.000,0.000,0.000,0.000,?,?,?,?,?,?,?
FOIL,golf,test,1,40.000,40.000,20.000,0.000,?,?,?,?,?,?,?

|Experiment finished:         15:52:15 842241135
```

Figure 90. .raw file output

Here, experimental information is shown, such as:

- the name of the experiment — Golf
- the date and by whom the experiment was created modified and run

```
|Name:       Golf
|Run:        Mon Sep 09 15:49:33 1996  by: kthomson
|Created:    Mon Sep 09 15:47:16 1996  by: kthomson
|Modified:   Mon Sep 09 15:49:17 1996  by: kthomson
```

Figure 91. .raw Experiment Information

- any comments about the experiment —
  - the output path and file name
  - the PREVAL evaluation options used
  - the test method used; and
  - the test parameters

```
|Output to:  /home/kthomson/WEKA/golf_results/golf.raw
|PREVAL opts: +warp9 +skiptcplx +skipdcplx
|Test Method: hold-out (Holte)
|Test Params:
|      Repetitions: 1
|      % in Training Set: 66
|      Starting seed: 1
```

Figure 92. .raw Experiment Comments

- the schemes that were run and the parameters used, for example InductH was set to DNF, it has -D as its parameter. None of the schemes have any comments.

```
|Schemes:
|----------------------------------------------------------------------
|1R (none)
|Comments:
|
|----------------------------------------------------------------------
|T1 (none)
|Comments:
|
|----------------------------------------------------------------------
|C4.5 Pruned (-m 2 -c 25)
|Comments:
|
|----------------------------------------------------------------------
|InductH (-D)
|Comments:
|
|----------------------------------------------------------------------
|FOIL (none)
|Comments:
|
|======================================================================
```

Figure 93. .raw Scheme Information

- the datasets used; this includes:
  - the name of each dataset
  - the path to the dataset
  - the dataset’s class
  - any comments about the dataset
  - and the date the dataset was added to the experiment

```
|Datasets:
|----------------------------------------------------------------------
|golf (/home/ml/wekalite2.1/datasets.lite/golf.arff) class(5)
|Comments:
|Dataset: /home/ml/wekalite2.1/datasets.lite/golf.arff
|Added:  Mon Sep 09 15:47:40 1996
|======================================================================
```

Figure 94. .raw Dataset Information

- the experiment times
  - when the experiment started
  - the time each scheme finished analysing the golf dataset

```
|Experiment started:          15:51:02 842241062
|started 1R        golf       15:51:03 842241063
|started T1        golf       15:51:21 842241081
|started C4.5 Prun golf       15:51:31 842241091
|started InductH   golf       15:51:45 842241105
|started FOIL      golf       15:51:58 842241118
```

Figure 95. .raw Experiment Times Information

- the experimental information about each scheme analysing the golf dataset; this includes:
  - the mode (testing or training); and
  - each scheme run with the same information given for the golf_exp.arff file.

```
|Experiment started:          15:51:02 842241062

|started 1R        golf       15:51:03 842241063
1R,golf,train,1,88.888,11.111,0.000,0.000,?,?,?,?,?,?,?
1R,golf,test,1,40.000,60.000,0.000,0.000,?,?,?,?,?,?,?

|started T1        golf       15:51:21 842241081
T1,golf,train,1,100.000,0.000,0.000,0.000,?,?,?,?,?,?,?
T1,golf,test,1,40.000,60.000,0.000,0.000,?,?,?,?,?,?,?

|started C4.5 Prun golf       15:51:31 842241091
C4.5 Pruned,golf,train,1,88.888,11.111,0.000,0.000,?,?,?,?,?,?,?
C4.5 Pruned,golf,test,1,40.000,60.000,0.000,0.000,?,?,?,?,?,?,?

|started InductH   golf       15:51:45 842241105
InductH,golf,train,1,100.000,0.000,0.000,0.000,?,?,?,?,?,?,?
InductH,golf,test,1,40.000,40.000,20.000,0.000,?,?,?,?,?,?,?

|started FOIL      golf       15:51:58 842241118
FOIL,golf,train,1,100.000,0.000,0.000,0.000,?,?,?,?,?,?,?
FOIL,golf,test,1,40.000,40.000,20.000,0.000,?,?,?,?,?,?,?

|Experiment finished:         15:52:15 842241135
```

Figure 96. .raw Experiment Information

The golf_exp.arff File

The ARFF file is different from the original golf.arff file containing golf data. This ARFF file contains the raw results from golf_exp.raw in ARFF format. However, it has a similar format; containing the relation name, the different attributes, and the data within the golf_exp.arff dataset.

The golf_exp.arff file has the following output:

```
@relation ’Golf’
@attribute Scheme          { ’1R’,’T1’,’C4.5 Pruned’,’InductH’,’FOIL’ }
@attribute Dataset         { ’golf’ }
@attribute Mode            { test, train }
@attribute Run { 1 }
@attribute Pct_Correct       REAL
@attribute Pct_Incorrect     REAL
@attribute Pct_Unclassified  REAL
@attribute Pct_Multiple      REAL
@attribute R                 REAL
@attribute C2                REAL
@attribute D2                REAL
@attribute T2                REAL
@attribute C3                REAL
@attribute D3                REAL
@attribute T3                REAL
@data
’1R’,’golf’,train,1,88.888,11.111,0.000,0.000,?,?,?,?,?,?,?
’1R’,’golf’,test,1,40.000,60.000,0.000,0.000,?,?,?,?,?,?,?
’T1’,’golf’,train,1,100.000,0.000,0.000,0.000,?,?,?,?,?,?,?
’T1’,’golf’,test,1,40.000,60.000,0.000,0.000,?,?,?,?,?,?,?
’C4.5 Pruned’,’golf’,train,1,88.888,11.111,0.000,0.000,?,?,?,?,?,?,?
’C4.5 Pruned’,’golf’,test,1,40.000,60.000,0.000,0.000,?,?,?,?,?,?,?
’InductH’,’golf’,train,1,100.000,0.000,0.000,0.000,?,?,?,?,?,?,?
’InductH’,’golf’,test,1,40.000,40.000,20.000,0.000,?,?,?,?,?,?,?
’FOIL’,’golf’,train,1,100.000,0.000,0.000,0.000,?,?,?,?,?,?,?
’FOIL’,’golf’,test,1,40.000,40.000,20.000,0.000,?,?,?,?,?,?,?
```

Figure 97. Output from golf_exp.arff

The information held in the golf_exp.arff dataset includes:

- which schemes were run on the golf.arff dataset, these were 1R, T1, C4.5 Pruned, InductH, and FOIL;
- the golf dataset name, which is given at all times as ‘golf’;
- what mode the golf dataset was analysed in — the golf dataset was analysed in both training and testing modes; and
- how many runs were made for the golf experiment — this was one.

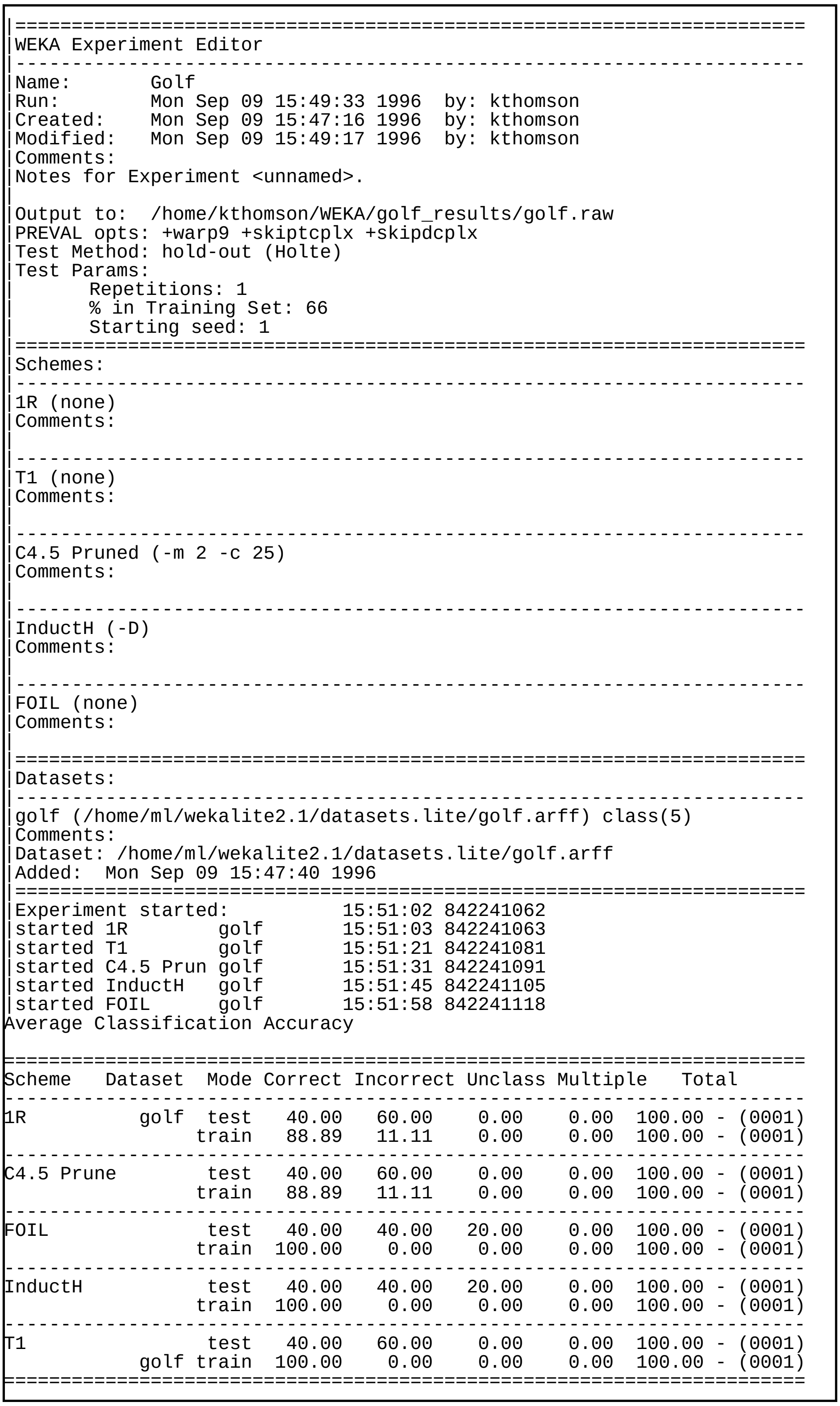
For each scheme, information is given for:

- the percentage of the instances correctly classified;
- the percentage of the instances incorrectly classified;
- the percentage of the instances not classified; and
- the percentage of the instances multiply classified.

The attributes R, C2, D2, T2, C3, D3, and T3 are PREVAL measures.

The golf_exp.clac File

The .clac file outputs a classification accuracy report, and has the following output:

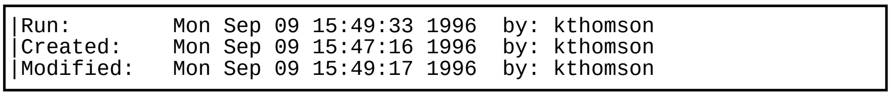


```
|======================================================================
|WEKA Experiment Editor
|----------------------------------------------------------------------
|Name:       Golf
|Run:        Mon Sep 09 15:49:33 1996  by: kthomson
|Created:    Mon Sep 09 15:47:16 1996  by: kthomson
|Modified:   Mon Sep 09 15:49:17 1996  by: kthomson
|Comments:
|Notes for Experiment <unnamed>.
|
|Output to:  /home/kthomson/WEKA/golf_results/golf.raw
|PREVAL opts: +warp9 +skiptcplx +skipdcplx
|Test Method: hold-out (Holte)
|Test Params:
|       Repetitions: 1
|       % in Training Set: 66
|       Starting seed: 1
|======================================================================
|Schemes:
|----------------------------------------------------------------------
|1R (none)
|Comments:
|
|----------------------------------------------------------------------
|T1 (none)
|Comments:
|
|----------------------------------------------------------------------
|C4.5 Pruned (-m 2 -c 25)
|Comments:
|
|----------------------------------------------------------------------
|InductH (-D)
|Comments:
|
|----------------------------------------------------------------------
|FOIL (none)
|Comments:
|
|======================================================================
|Datasets:
|----------------------------------------------------------------------
|golf (/home/ml/wekalite2.1/datasets.lite/golf.arff) class(5)
|Comments:
|Dataset: /home/ml/wekalite2.1/datasets.lite/golf.arff
|Added:  Mon Sep 09 15:47:40 1996
|======================================================================
|Experiment started:          15:51:02 842241062
|started 1R        golf       15:51:03 842241063
|started T1        golf       15:51:21 842241081
|started C4.5 Prun golf       15:51:31 842241091
|started InductH   golf       15:51:45 842241105
|started FOIL      golf       15:51:58 842241118
Average Classification Accuracy

=======================================================================
Scheme   Dataset  Mode Correct Incorrect Unclass Multiple   Total
-----------------------------------------------------------------------
1R          golf  test   40.00   60.00    0.00    0.00  100.00 - (0001)
                 train   88.89   11.11    0.00    0.00  100.00 - (0001)
-----------------------------------------------------------------------
C4.5 Prune        test   40.00   60.00    0.00    0.00  100.00 - (0001)
                 train   88.89   11.11    0.00    0.00  100.00 - (0001)
-----------------------------------------------------------------------
FOIL              test   40.00   40.00   20.00    0.00  100.00 - (0001)
                 train  100.00    0.00    0.00    0.00  100.00 - (0001)
-----------------------------------------------------------------------
InductH           test   40.00   40.00   20.00    0.00  100.00 - (0001)
                 train  100.00    0.00    0.00    0.00  100.00 - (0001)
-----------------------------------------------------------------------
T1                test   40.00   60.00    0.00    0.00  100.00 - (0001)
            golf train  100.00    0.00    0.00    0.00  100.00 - (0001)
=======================================================================
```

Figure 98. .clac file output

Here, experimental information is shown, such as:

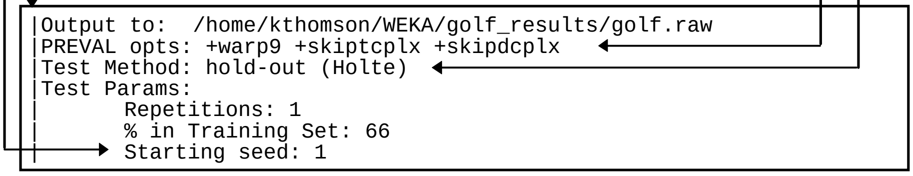
- the name of the experiment — Golf
- the date and by whom the experiment was created modified and run



```
|Run:         Mon Sep 09 15:49:33 1996  by: kthomson
|Created:     Mon Sep 09 15:47:16 1996  by: kthomson
|Modified:    Mon Sep 09 15:49:17 1996  by: kthomson
```

Figure 99. .clac Experiment Information

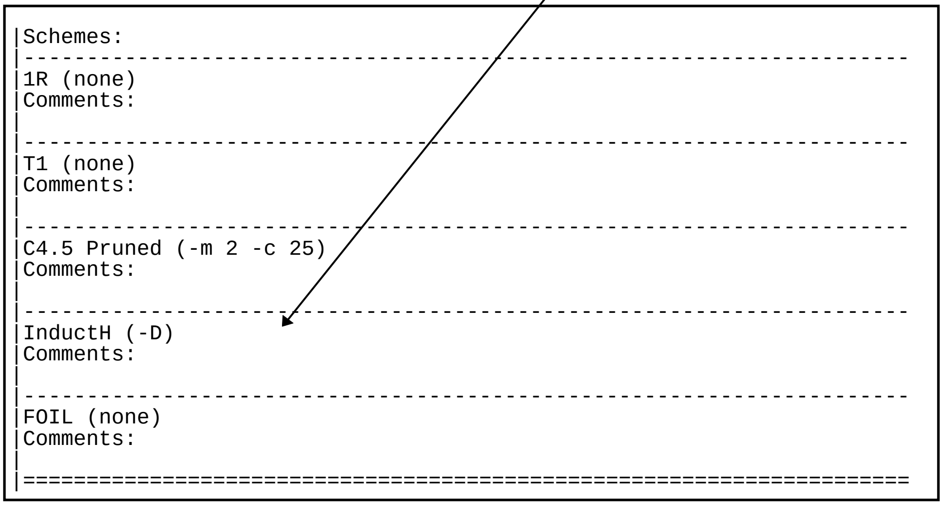
- any comments about the experiment
  - the output path and file name
  - the PREVAL evaluation options used
  - the test method used; and
  - the test parameters



```
|Output to:  /home/kthomson/WEKA/golf_results/golf.raw
|PREVAL opts: +warp9 +skiptcplx +skipdcplx
|Test Method: hold-out (Holte)
|Test Params:
|      Repetitions: 1
|      % in Training Set: 66
|      Starting seed: 1
```

Figure 100. .clac Experiment Comments

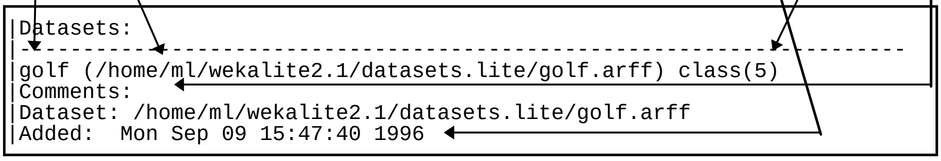
- the schemes that were run and the parameters used, for example InductH was set to DNF, it has -D as its parameter. None of the schemes have any comments.



```
|Schemes:
|----------------------------------------------------------------------
|1R (none)
|Comments:
|
|----------------------------------------------------------------------
|T1 (none)
|Comments:
|
|----------------------------------------------------------------------
|C4.5 Pruned (-m 2 -c 25)
|Comments:
|
|----------------------------------------------------------------------
|InductH (-D)
|Comments:
|
|----------------------------------------------------------------------
|FOIL (none)
|Comments:
|
|======================================================================
```

Figure 101. .clac Scheme Information

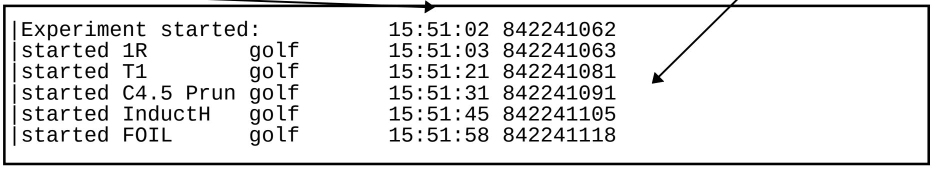
- the datasets used; this includes:
  - the name of each dataset
  - the path to the dataset
  - the dataset’s class
  - any comments about the dataset
  - and the date the dataset was added to the experiment



```
|Datasets:
|----------------------------------------------------------------------
|golf (/home/ml/wekalite2.1/datasets.lite/golf.arff) class(5)
|Comments:
|Dataset: /home/ml/wekalite2.1/datasets.lite/golf.arff
|Added:  Mon Sep 09 15:47:40 1996
```

Figure 102. .clac Dataset Information

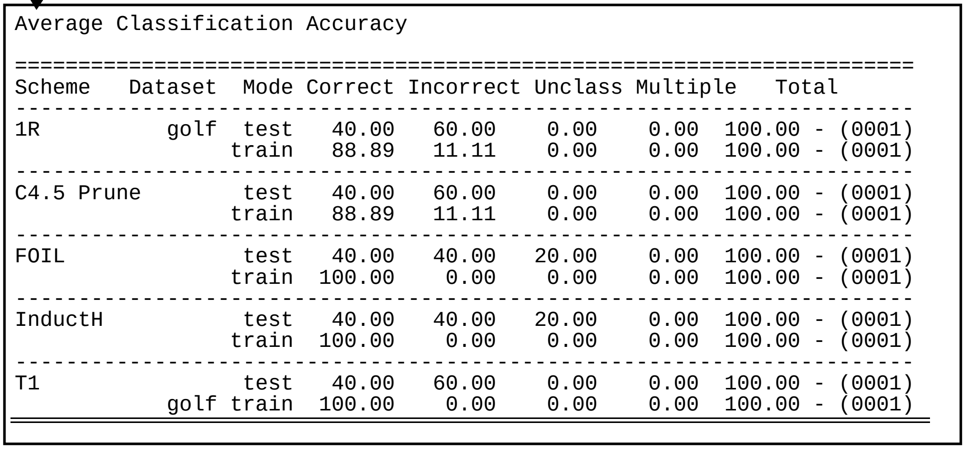
- the experiment times
  - when the experiment started
  - the time each scheme finished analysing the golf dataset



```
|Experiment started:          15:51:02 842241062
|started 1R        golf       15:51:03 842241063
|started T1        golf       15:51:21 842241081
|started C4.5 Prun golf       15:51:31 842241091
|started InductH   golf       15:51:45 842241105
|started FOIL      golf       15:51:58 842241118
```

Figure 103. .clac Experiment Times

- the average classification accuracy of each scheme analysing the golf dataset; this includes:
  - the mode (testing or training);
  - the percentage of instances correctly classified;
  - the percentage of instances incorrectly classified;
  - the percentage of instances unclassified;
  - the percentage of instances multiply classified; and
  - the total classifications made by each scheme.

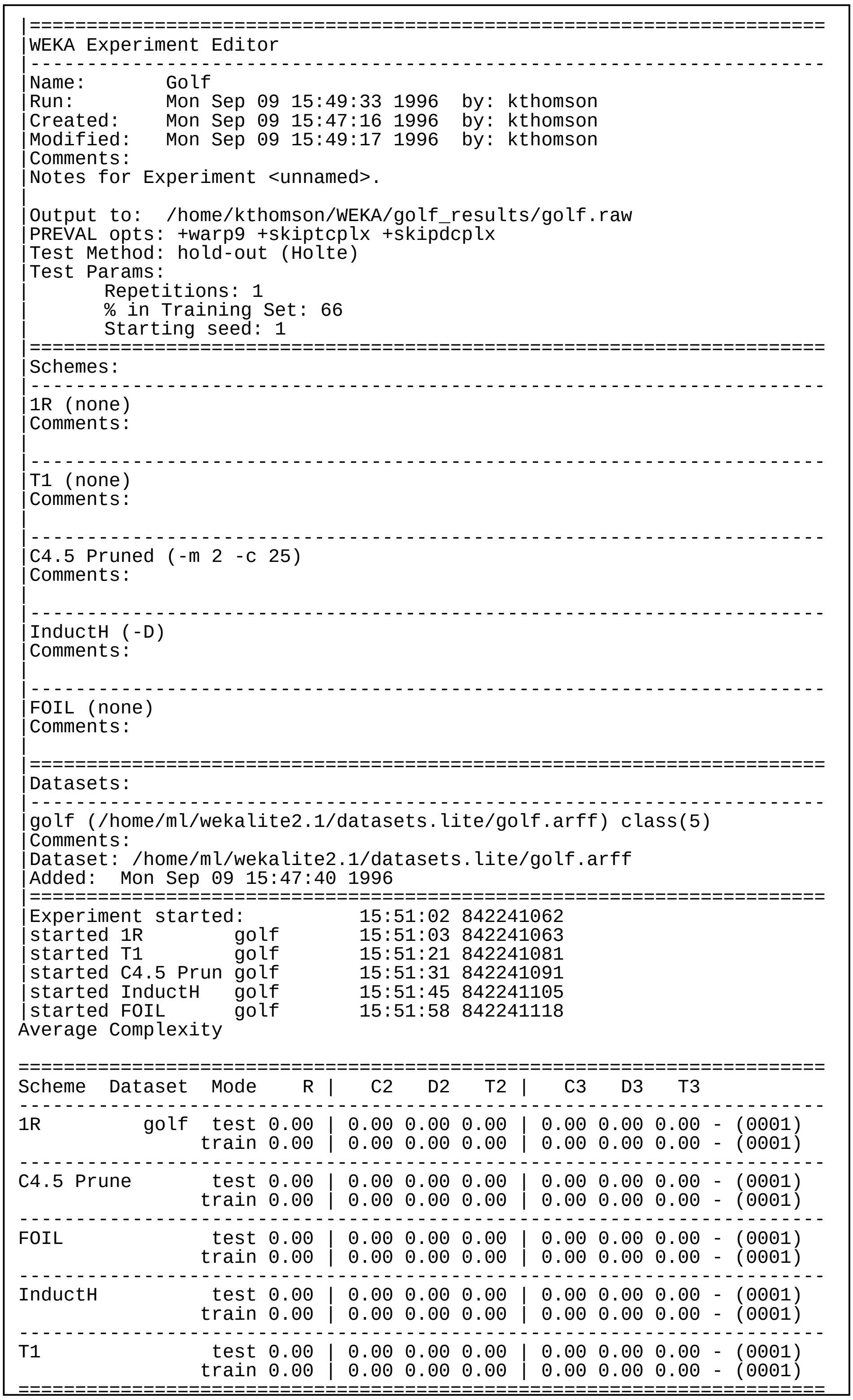


```
Average Classification Accuracy

=======================================================================
Scheme   Dataset  Mode Correct Incorrect Unclass Multiple   Total
-----------------------------------------------------------------------
1R          golf  test   40.00   60.00    0.00    0.00  100.00 - (0001)
                 train   88.89   11.11    0.00    0.00  100.00 - (0001)
-----------------------------------------------------------------------
C4.5 Prune        test   40.00   60.00    0.00    0.00  100.00 - (0001)
                 train   88.89   11.11    0.00    0.00  100.00 - (0001)
-----------------------------------------------------------------------
FOIL              test   40.00   40.00   20.00    0.00  100.00 - (0001)
                 train  100.00    0.00    0.00    0.00  100.00 - (0001)
-----------------------------------------------------------------------
InductH           test   40.00   40.00   20.00    0.00  100.00 - (0001)
                 train  100.00    0.00    0.00    0.00  100.00 - (0001)
-----------------------------------------------------------------------
T1                test   40.00   60.00    0.00    0.00  100.00 - (0001)
            golf train  100.00    0.00    0.00    0.00  100.00 - (0001)
```

Figure 104. .clac Average Classification Accuracy Information

The golf_exp.comp File

The .comp file outputs a PREVAL complexity report; its output is:



```
|======================================================================
|WEKA Experiment Editor
|----------------------------------------------------------------------
|Name:       Golf
|Run:        Mon Sep 09 15:49:33 1996  by: kthomson
|Created:    Mon Sep 09 15:47:16 1996  by: kthomson
|Modified:   Mon Sep 09 15:49:17 1996  by: kthomson
|Comments:
|Notes for Experiment <unnamed>.
|
|Output to:  /home/kthomson/WEKA/golf_results/golf.raw
|PREVAL opts: +warp9 +skiptcplx +skipdcplx
|Test Method: hold-out (Holte)
|Test Params:
|       Repetitions: 1
|       % in Training Set: 66
|       Starting seed: 1
|======================================================================
|Schemes:
|----------------------------------------------------------------------
|1R (none)
|Comments:
|
|----------------------------------------------------------------------
|T1 (none)
|Comments:
|
|----------------------------------------------------------------------
|C4.5 Pruned (-m 2 -c 25)
|Comments:
|
|----------------------------------------------------------------------
|InductH (-D)
|Comments:
|
|----------------------------------------------------------------------
|FOIL (none)
|Comments:
|
|======================================================================
|Datasets:
|----------------------------------------------------------------------
|golf (/home/ml/wekalite2.1/datasets.lite/golf.arff) class(5)
|Comments:
|Dataset: /home/ml/wekalite2.1/datasets.lite/golf.arff
|Added:  Mon Sep 09 15:47:40 1996
|======================================================================
|Experiment started:          15:51:02 842241062
|started 1R        golf       15:51:03 842241063
|started T1        golf       15:51:21 842241081
|started C4.5 Prun golf       15:51:31 842241091
|started InductH   golf       15:51:45 842241105
|started FOIL      golf       15:51:58 842241118
Average Complexity

=======================================================================
Scheme  Dataset  Mode    R |   C2   D2   T2 |   C3   D3   T3
-----------------------------------------------------------------------
1R         golf  test 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 - (0001)
                train 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 - (0001)
-----------------------------------------------------------------------
C4.5 Prune       test 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 - (0001)
                train 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 - (0001)
-----------------------------------------------------------------------
FOIL             test 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 - (0001)
                train 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 - (0001)
-----------------------------------------------------------------------
InductH          test 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 - (0001)
                train 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 - (0001)
-----------------------------------------------------------------------
T1               test 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 - (0001)
                train 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 - (0001)
=======================================================================
```

Figure 105. .comp file output

Here, experimental information is shown, such as:

- the name of the experiment — Golf
- the date and by whom the experiment was created modified and run

```
|Name:        Golf
|Run:         Mon Sep 09 15:49:33 1996  by: kthomson
|Created:     Mon Sep 09 15:47:16 1996  by: kthomson
|Modified:    Mon Sep 09 15:49:17 1996  by: kthomson
```

Figure 106. .clac Experiment Information

- any comments about the experiment
  - the output path and file name
  - the PREVAL evaluation options used
  - the test method used; and
  - the test parameters

```
|Output to:  /home/kthomson/WEKA/golf_results/golf.raw
|PREVAL opts: +warp9 +skiptcplx +skipdcplx
|Test Method: hold-out (Holte)
|Test Params:
|       Repetitions: 1
|       % in Training Set: 66
|       Starting seed: 1
```

Figure 107. .clac Experiment Comments

- the schemes that were run and the parameters used, for example InductH was set to DNF, it has -D as its parameter. None of the schemes have any comments.

```
|Schemes:
|----------------------------------------------------------------------
|1R (none)
|Comments:
|
|----------------------------------------------------------------------
|T1 (none)
|Comments:
|
|----------------------------------------------------------------------
|C4.5 Pruned (-m 2 -c 25)
|Comments:
|
|----------------------------------------------------------------------
|InductH (-D)
|Comments:
|
|----------------------------------------------------------------------
|FOIL (none)
|Comments:
|
|======================================================================
```

Figure 108. .clac Scheme Information

- the datasets used; this includes:
    - the name of each dataset
    - the path to the dataset
    - the dataset's class
    - any comments about the dataset
    - and the date the dataset was added to the experiment

```
|Datasets:
|-------------------------------------------------------------------------
|golf (/home/ml/wekalite2.1/datasets.lite/golf.arff) class(5)
|Comments:
|Dataset: /home/ml/wekalite2.1/datasets.lite/golf.arff
|Added:  Mon Sep 09 15:47:40 1996
|=========================================================================
```

Figure 109. .clac Dataset Information

- the experiment times
    - when the experiment started
    - the time each scheme finished analysing the golf dataset

```
|Experiment started:          15:51:02 842241062
|started 1R        golf       15:51:03 842241063
|started T1        golf       15:51:21 842241081
|started C4.5 Prun golf       15:51:31 842241091
|started InductH   golf       15:51:45 842241105
|started FOIL      golf       15:51:58 842241118
```

Figure 110. .clac Experiment Times

- the PREVAL complexity measures for each scheme analysing the golf dataset; this includes:
    - the mode (testing or training); and

```
=======================================================================
Scheme  Dataset  Mode    R |   C2   D2   T2 |   C3   D3   T3
-----------------------------------------------------------------------
1R         golf  test 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 - (0001)
                train 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 - (0001)
-----------------------------------------------------------------------
C4.5 Prune       test 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 - (0001)
                train 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 - (0001)
-----------------------------------------------------------------------
FOIL             test 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 - (0001)
                train 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 - (0001)
-----------------------------------------------------------------------
InductH          test 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 - (0001)
                train 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 - (0001)
-----------------------------------------------------------------------
T1               test 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 - (0001)
                train 0.00 | 0.00 0.00 0.00 | 0.00 0.00 0.00 - (0001)
=======================================================================
```

Figure 111. .clac PREVAL Complexity Measures

Appendices

APPENDIX 1: WEKA EXPERIMENT WITH TRAINING AND TESTING DATASETS

Here is another example of data analysis performed on the WEKA workbench. This analysis has been included to show how both a training and a testing dataset can also be used in analysis.

The breast cancer file, introduced at the beginning of the Tutorial is used as the example datafile. The original breast cancer file has already been divided into separate training and testing datasets.[17] Both datasets have been saved as ARFF files and can be found in:

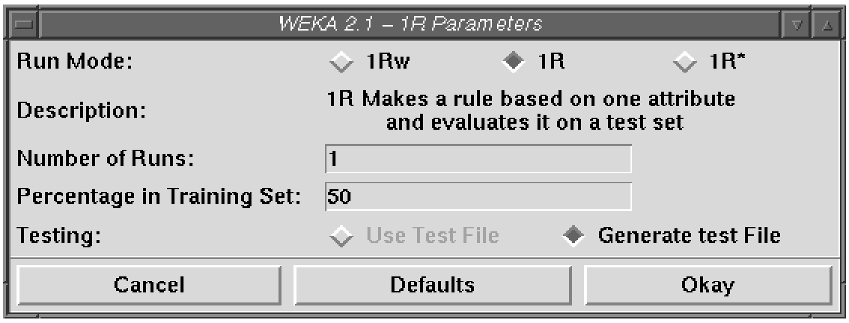
```
$WEKAHOME/datasets.lite/breast-cancer
```

Load the breast-cancer training dataset:

```
breast-cancer-data.arff
```

From the scheme menu choose each of the following schemes with the class attribute as class **10 Class (E)**:

1R — Run the scheme with 1R options



Run Mode: ◇ 1Rw ◆ 1R ◇ 1R*

Description: 1R Makes a rule based on one attribute and evaluates it on a test set

Number of Runs: 1

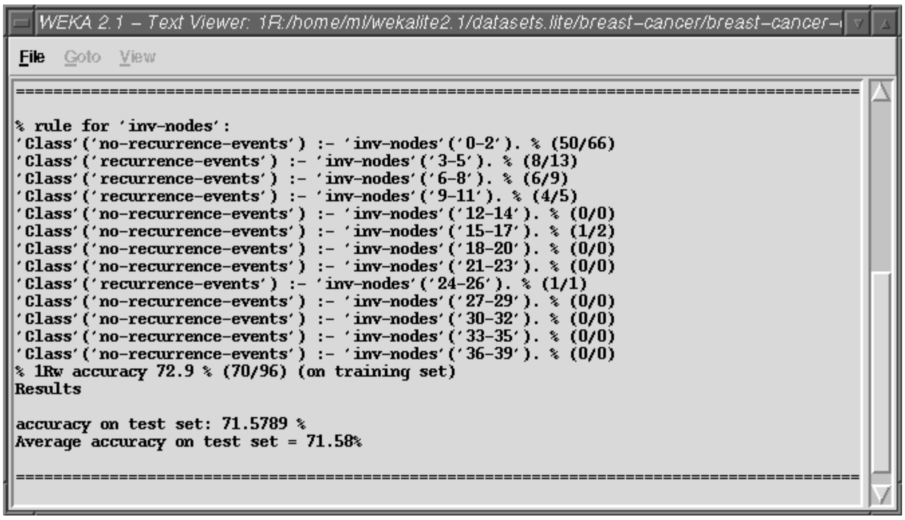
Percentage in Training Set: 50

Testing: ◇ Use Test File ◆ Generate test File

Figure 112. 1R Parameters

and view the results

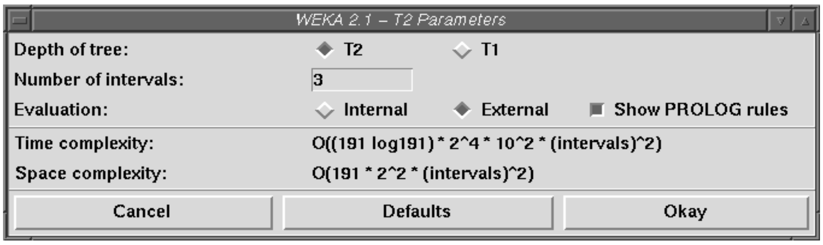
[17] Training and testing datasets are described in the Introduction.


WEKA 2.1 – Text Viewer: 1R:/home/ml/wekalite2.1/datasets.lite/breast-cancer/breast-cancer-
File Goto View

```
================================================================================
% rule for 'inv-nodes':
'Class'('no-recurrence-events') :- 'inv-nodes'('0-2'). % (50/66)
'Class'('recurrence-events') :- 'inv-nodes'('3-5'). % (8/13)
'Class'('recurrence-events') :- 'inv-nodes'('6-8'). % (6/9)
'Class'('recurrence-events') :- 'inv-nodes'('9-11'). % (4/5)
'Class'('no-recurrence-events') :- 'inv-nodes'('12-14'). % (0/0)
'Class'('no-recurrence-events') :- 'inv-nodes'('15-17'). % (1/2)
'Class'('no-recurrence-events') :- 'inv-nodes'('18-20'). % (0/0)
'Class'('no-recurrence-events') :- 'inv-nodes'('21-23'). % (0/0)
'Class'('recurrence-events') :- 'inv-nodes'('24-26'). % (1/1)
'Class'('no-recurrence-events') :- 'inv-nodes'('27-29'). % (0/0)
'Class'('no-recurrence-events') :- 'inv-nodes'('30-32'). % (0/0)
'Class'('no-recurrence-events') :- 'inv-nodes'('33-35'). % (0/0)
'Class'('no-recurrence-events') :- 'inv-nodes'('36-39'). % (0/0)
% 1Rw accuracy 72.9 % (70/96) (on training set)
Results

accuracy on test set: 71.5789 %
Average accuracy on test set = 71.58%

================================================================================
```

Figure 113. 1R Results

T2 — Run the T2 scheme with T2 options



Depth of tree: ◆ T2 ◇ T1

Number of intervals: 3

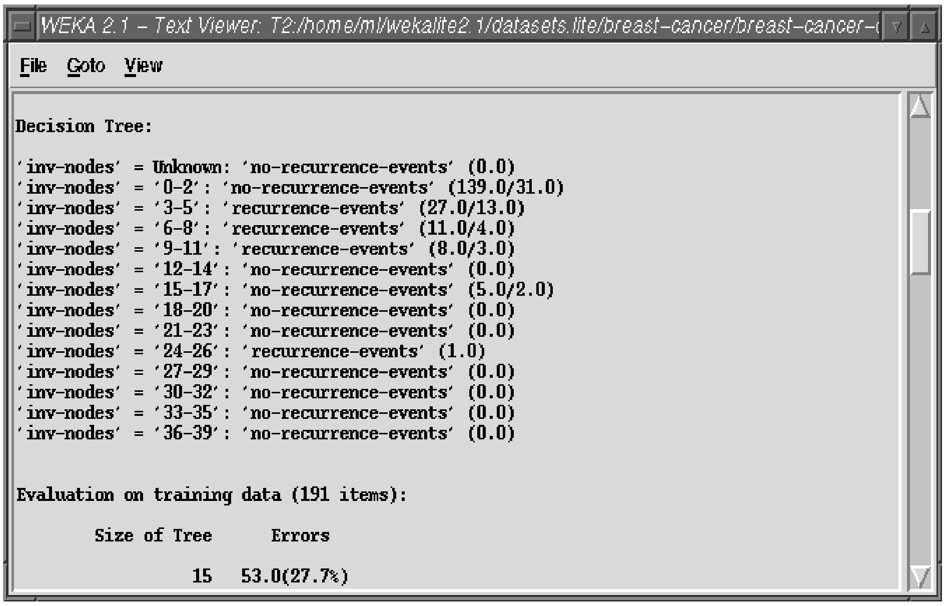
Evaluation: ◇ Internal ◆ External ■ Show PROLOG rules

Time complexity: O((191 log191) * 2^4 * 10^2 * (intervals)^2)

Space complexity: O(191 * 2^2 * (intervals)^2)

Figure 114. T2 Parameters

and view the results

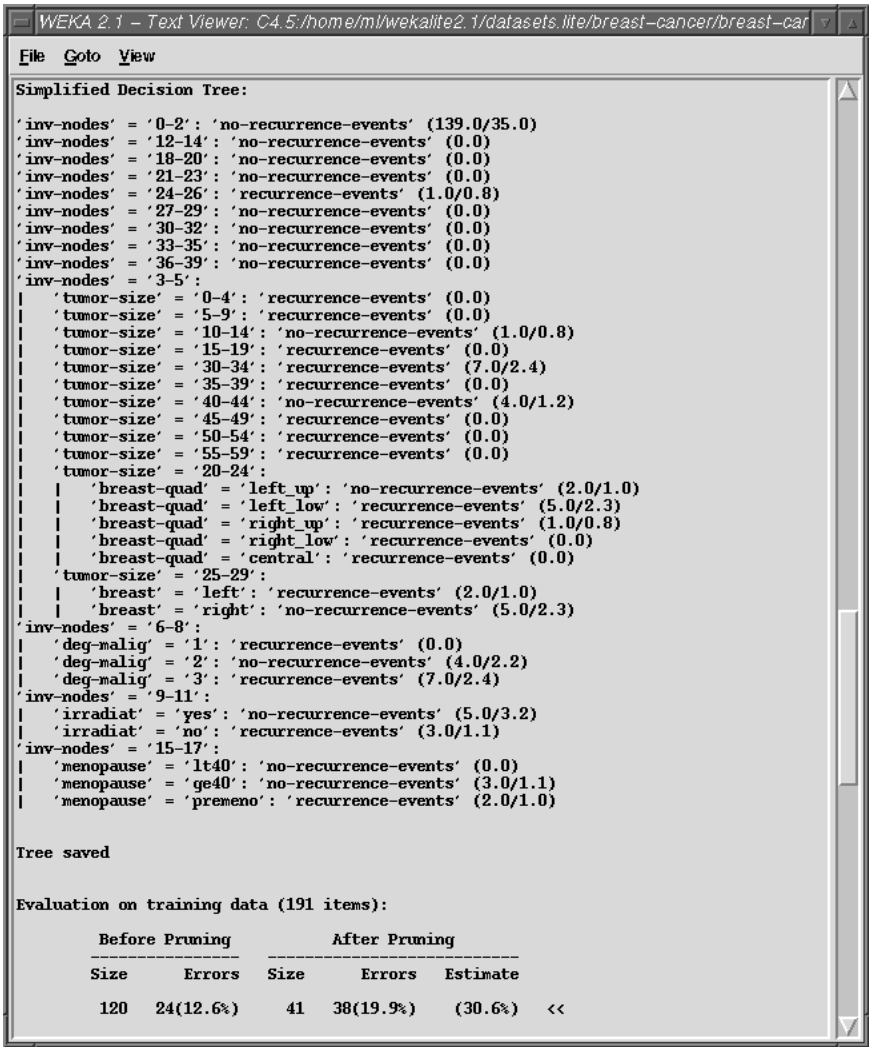

WEKA 2.1 – Text Viewer: T2:/home/ml/wekalite2.1/datasets.lite/breast-cancer/breast-cancer-(
File Goto View

```
Decision Tree:

'inv-nodes' = Unknown: 'no-recurrence-events' (0.0)
'inv-nodes' = '0-2': 'no-recurrence-events' (139.0/31.0)
'inv-nodes' = '3-5': 'recurrence-events' (27.0/13.0)
'inv-nodes' = '6-8': 'recurrence-events' (11.0/4.0)
'inv-nodes' = '9-11': 'recurrence-events' (8.0/3.0)
'inv-nodes' = '12-14': 'no-recurrence-events' (0.0)
'inv-nodes' = '15-17': 'no-recurrence-events' (5.0/2.0)
'inv-nodes' = '18-20': 'no-recurrence-events' (0.0)
'inv-nodes' = '21-23': 'no-recurrence-events' (0.0)
'inv-nodes' = '24-26': 'recurrence-events' (1.0)
'inv-nodes' = '27-29': 'no-recurrence-events' (0.0)
'inv-nodes' = '30-32': 'no-recurrence-events' (0.0)
'inv-nodes' = '33-35': 'no-recurrence-events' (0.0)
'inv-nodes' = '36-39': 'no-recurrence-events' (0.0)


Evaluation on training data (191 items):

        Size of Tree      Errors

                 15   53.0(27.7%)
```

Figure 115. T2 Decision Tree Results

If you want to see a graphical representation of the decision tree then choose **Unpruned Decision Tree...** from the View menu.

C4.5 — Run this scheme with Tree and Internal output, and internal evaluation. View the results.

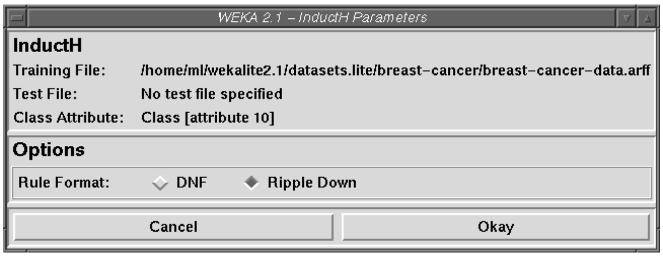


WEKA 2.1 – Text Viewer: C4.5:/home/ml/wekalite2.1/datasets.lite/breast-cancer/breast-car
File
Goto
View

```
Simplified Decision Tree:

'inv-nodes' = '0-2': 'no-recurrence-events' (139.0/35.0)
'inv-nodes' = '12-14': 'no-recurrence-events' (0.0)
'inv-nodes' = '18-20': 'no-recurrence-events' (0.0)
'inv-nodes' = '21-23': 'no-recurrence-events' (0.0)
'inv-nodes' = '24-26': 'recurrence-events' (1.0/0.8)
'inv-nodes' = '27-29': 'no-recurrence-events' (0.0)
'inv-nodes' = '30-32': 'no-recurrence-events' (0.0)
'inv-nodes' = '33-35': 'no-recurrence-events' (0.0)
'inv-nodes' = '36-39': 'no-recurrence-events' (0.0)
'inv-nodes' = '3-5':
|   'tumor-size' = '0-4': 'recurrence-events' (0.0)
|   'tumor-size' = '5-9': 'recurrence-events' (0.0)
|   'tumor-size' = '10-14': 'no-recurrence-events' (1.0/0.8)
|   'tumor-size' = '15-19': 'recurrence-events' (0.0)
|   'tumor-size' = '30-34': 'recurrence-events' (7.0/2.4)
|   'tumor-size' = '35-39': 'recurrence-events' (0.0)
|   'tumor-size' = '40-44': 'no-recurrence-events' (4.0/1.2)
|   'tumor-size' = '45-49': 'recurrence-events' (0.0)
|   'tumor-size' = '50-54': 'recurrence-events' (0.0)
|   'tumor-size' = '55-59': 'recurrence-events' (0.0)
|   'tumor-size' = '20-24':
|   |   'breast-quad' = 'left_up': 'no-recurrence-events' (2.0/1.0)
|   |   'breast-quad' = 'left_low': 'recurrence-events' (5.0/2.3)
|   |   'breast-quad' = 'right_up': 'recurrence-events' (1.0/0.8)
|   |   'breast-quad' = 'right_low': 'recurrence-events' (0.0)
|   |   'breast-quad' = 'central': 'recurrence-events' (0.0)
|   'tumor-size' = '25-29':
|   |   'breast' = 'left': 'recurrence-events' (2.0/1.0)
|   |   'breast' = 'right': 'no-recurrence-events' (5.0/2.3)
'inv-nodes' = '6-8':
|   'deg-malig' = '1': 'recurrence-events' (0.0)
|   'deg-malig' = '2': 'no-recurrence-events' (4.0/2.2)
|   'deg-malig' = '3': 'recurrence-events' (7.0/2.4)
'inv-nodes' = '9-11':
|   'irradiat' = 'yes': 'no-recurrence-events' (5.0/3.2)
|   'irradiat' = 'no': 'recurrence-events' (3.0/1.1)
'inv-nodes' = '15-17':
|   'menopause' = 'lt40': 'no-recurrence-events' (0.0)
|   'menopause' = 'ge40': 'no-recurrence-events' (3.0/1.1)
|   'menopause' = 'premeno': 'recurrence-events' (2.0/1.0)


Tree saved


Evaluation on training data (191 items):

         Before Pruning           After Pruning
        ----------------   ---------------------------
        Size      Errors   Size      Errors   Estimate

         120   24(12.6%)     41   38(19.9%)    (30.6%)   <<
```

Figure 116. C4.5 Decision Tree

InductH — Run this scheme with Rippledown Rules (this takes about three minutes).



InductH

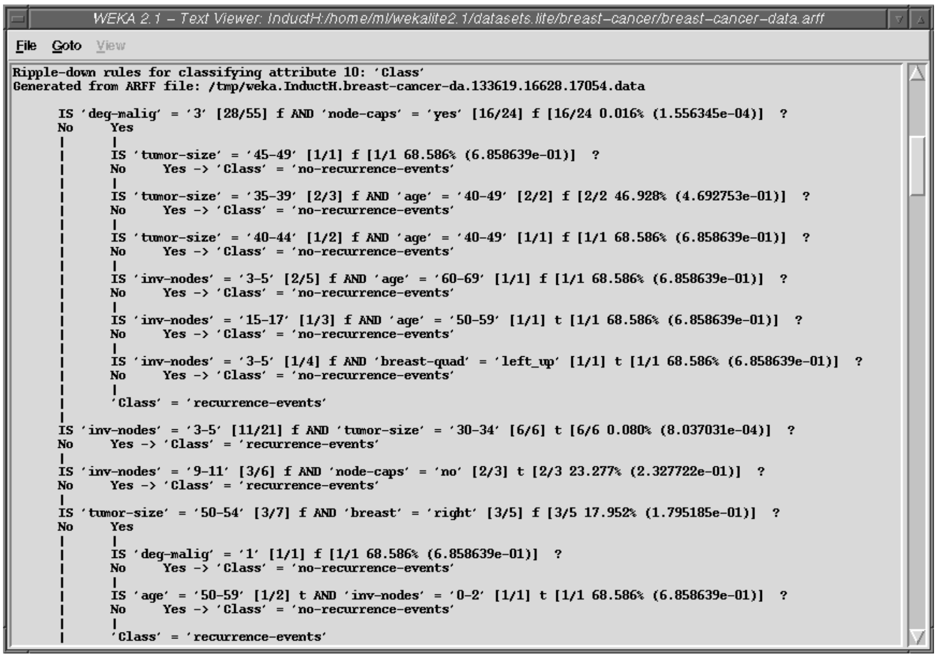
Training File: /home/ml/wekalite2.1/datasets.lite/breast-cancer/breast-cancer-data.arff
Test File: No test file specified
Class Attribute: Class [attribute 10]

Options

Rule Format: ◇ DNF ◆ Ripple Down

Figure 117. InductH Parameters

and view the results.

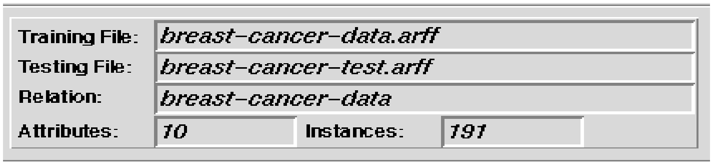

WEKA 2.1 – Text Viewer: InductH:/home/ml/wekalite2.1/datasets.lite/breast-cancer/breast-cancer-data.arff
File
Goto
View

```
Ripple-down rules for classifying attribute 10: 'Class'
Generated from ARFF file: /tmp/weka.InductH.breast-cancer-da.133619.16628.17054.data

      IS 'deg-malig' = '3' [28/55] f AND 'node-caps' = 'yes' [16/24] f [16/24 0.016% (1.556345e-04)]  ?
      No     Yes
      |      |
      |      IS 'tumor-size' = '45-49' [1/1] f [1/1 68.586% (6.858639e-01)]  ?
      |      No     Yes -> 'Class' = 'no-recurrence-events'
      |      |
      |      IS 'tumor-size' = '35-39' [2/3] f AND 'age' = '40-49' [2/2] f [2/2 46.928% (4.692753e-01)]  ?
      |      No     Yes -> 'Class' = 'no-recurrence-events'
      |      |
      |      IS 'tumor-size' = '40-44' [1/2] f AND 'age' = '40-49' [1/1] f [1/1 68.586% (6.858639e-01)]  ?
      |      No     Yes -> 'Class' = 'no-recurrence-events'
      |      |
      |      IS 'inv-nodes' = '3-5' [2/5] f AND 'age' = '60-69' [1/1] f [1/1 68.586% (6.858639e-01)]  ?
      |      No     Yes -> 'Class' = 'no-recurrence-events'
      |      |
      |      IS 'inv-nodes' = '15-17' [1/3] f AND 'age' = '50-59' [1/1] t [1/1 68.586% (6.858639e-01)]  ?
      |      No     Yes -> 'Class' = 'no-recurrence-events'
      |      |
      |      IS 'inv-nodes' = '3-5' [1/4] f AND 'breast-quad' = 'left_up' [1/1] t [1/1 68.586% (6.858639e-01)]  ?
      |      No     Yes -> 'Class' = 'no-recurrence-events'
      |      |
      |      'Class' = 'recurrence-events'
      |
      IS 'inv-nodes' = '3-5' [11/21] f AND 'tumor-size' = '30-34' [6/6] t [6/6 0.080% (8.037031e-04)]  ?
      No     Yes -> 'Class' = 'recurrence-events'
      |
      IS 'inv-nodes' = '9-11' [3/6] f AND 'node-caps' = 'no' [2/3] t [2/3 23.277% (2.327722e-01)]  ?
      No     Yes -> 'Class' = 'recurrence-events'
      |
      IS 'tumor-size' = '50-54' [3/7] f AND 'breast' = 'right' [3/5] f [3/5 17.952% (1.795185e-01)]  ?
      No     Yes
      |      |
      |      IS 'deg-malig' = '1' [1/1] f [1/1 68.586% (6.858639e-01)]  ?
      |      No     Yes -> 'Class' = 'no-recurrence-events'
      |      |
      |      IS 'age' = '50-59' [1/2] t AND 'inv-nodes' = '0-2' [1/1] t [1/1 68.586% (6.858639e-01)]  ?
      |      No     Yes -> 'Class' = 'no-recurrence-events'
      |      |
      |      'Class' = 'recurrence-events'
```

Figure 118. InductH Results

Load the breast-cancer testing dataset:

`breast-cancer-test.arff`

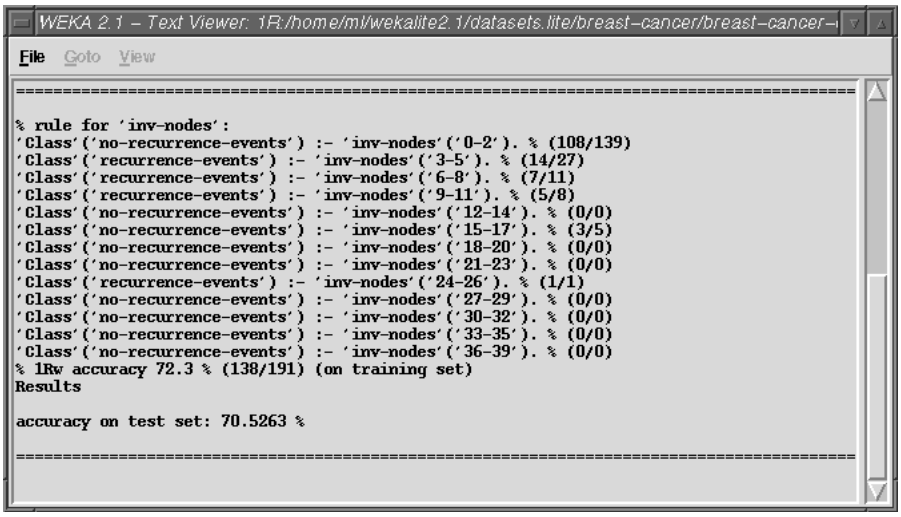


Training File: *breast–cancer–data.arff*
Testing File: *breast–cancer–test.arff*
Relation: *breast–cancer–data*
Attributes: *10* Instances: *191*

Figure 119. Data and Test Files loaded on the Experiment Editor

From the scheme menu choose each of the following schemes with the class attribute as class **10 Class (E)**:

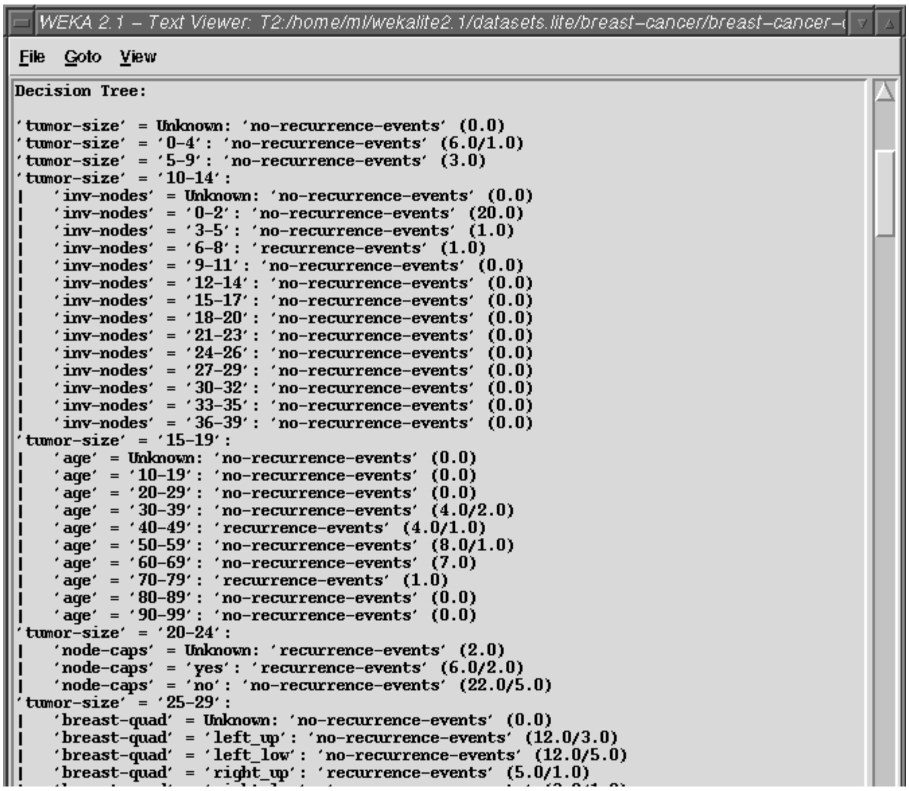
1R — Run the scheme with 1R options; view the results and compare these with the training dataset 1R results.


WEKA 2.1 – Text Viewer: 1R:/home/ml/wekalite2.1/datasets.lite/breast-cancer/breast-cancer-
File
Goto
View

```
==========================================================================================

% rule for 'inv-nodes':
'Class'('no-recurrence-events') :- 'inv-nodes'('0-2'). % (108/139)
'Class'('recurrence-events') :- 'inv-nodes'('3-5'). % (14/27)
'Class'('recurrence-events') :- 'inv-nodes'('6-8'). % (7/11)
'Class'('recurrence-events') :- 'inv-nodes'('9-11'). % (5/8)
'Class'('no-recurrence-events') :- 'inv-nodes'('12-14'). % (0/0)
'Class'('no-recurrence-events') :- 'inv-nodes'('15-17'). % (3/5)
'Class'('no-recurrence-events') :- 'inv-nodes'('18-20'). % (0/0)
'Class'('no-recurrence-events') :- 'inv-nodes'('21-23'). % (0/0)
'Class'('recurrence-events') :- 'inv-nodes'('24-26'). % (1/1)
'Class'('no-recurrence-events') :- 'inv-nodes'('27-29'). % (0/0)
'Class'('no-recurrence-events') :- 'inv-nodes'('30-32'). % (0/0)
'Class'('no-recurrence-events') :- 'inv-nodes'('33-35'). % (0/0)
'Class'('no-recurrence-events') :- 'inv-nodes'('36-39'). % (0/0)
% 1Rw accuracy 72.3 % (138/191) (on training set)
Results

accuracy on test set: 70.5263 %

==========================================================================================
```

Figure 120. 1R Test Results

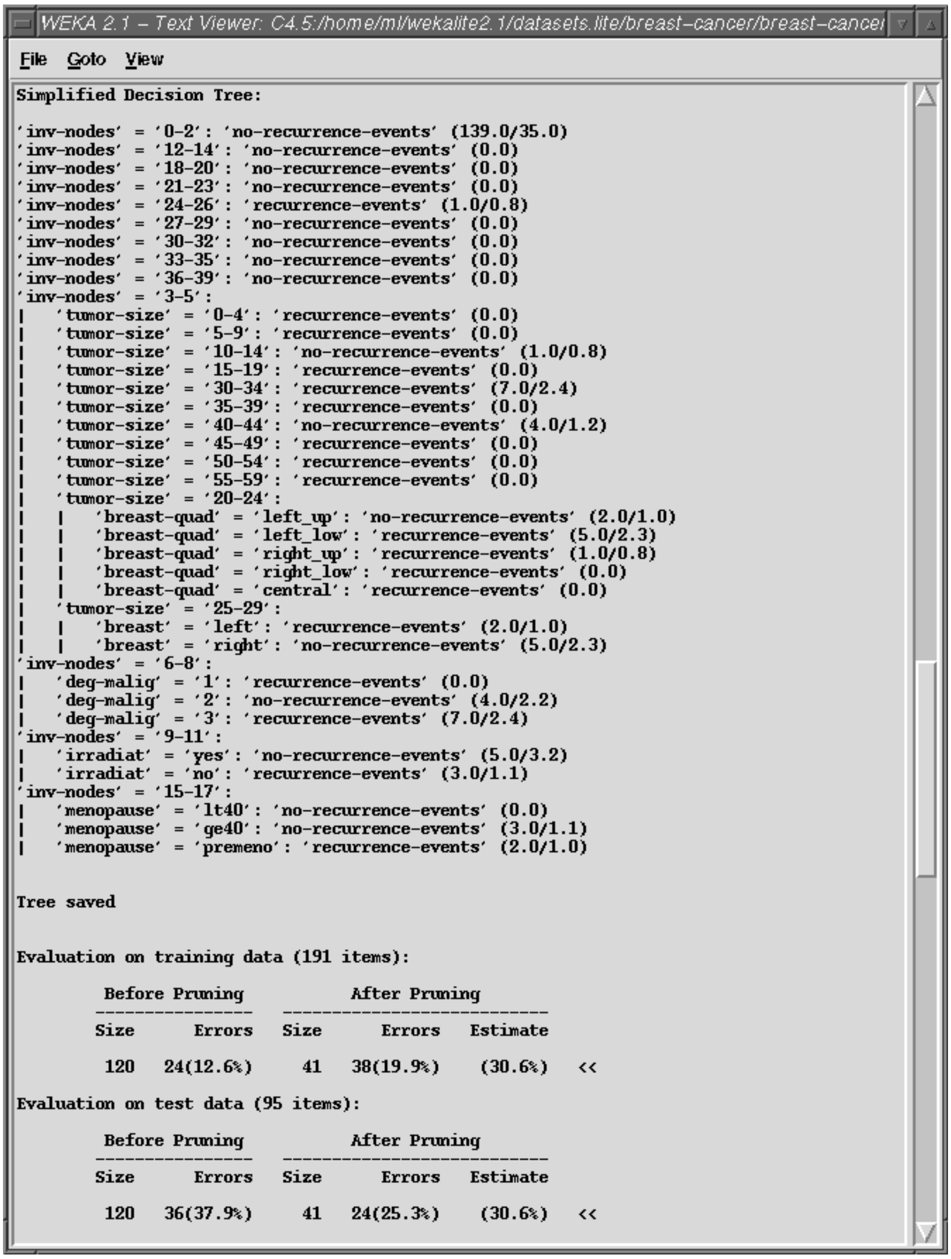
T2 — Run the scheme with T2 options; view the results and compare these with the training dataset results.


WEKA 2.1 – Text Viewer: T2:/home/ml/wekalite2.1/datasets.lite/breast-cancer/breast-cancer-
File
Goto
View

```
Decision Tree:

'tumor-size' = Unknown: 'no-recurrence-events' (0.0)
'tumor-size' = '0-4': 'no-recurrence-events' (6.0/1.0)
'tumor-size' = '5-9': 'no-recurrence-events' (3.0)
'tumor-size' = '10-14':
|   'inv-nodes' = Unknown: 'no-recurrence-events' (0.0)
|   'inv-nodes' = '0-2': 'no-recurrence-events' (20.0)
|   'inv-nodes' = '3-5': 'no-recurrence-events' (1.0)
|   'inv-nodes' = '6-8': 'recurrence-events' (1.0)
|   'inv-nodes' = '9-11': 'no-recurrence-events' (0.0)
|   'inv-nodes' = '12-14': 'no-recurrence-events' (0.0)
|   'inv-nodes' = '15-17': 'no-recurrence-events' (0.0)
|   'inv-nodes' = '18-20': 'no-recurrence-events' (0.0)
|   'inv-nodes' = '21-23': 'no-recurrence-events' (0.0)
|   'inv-nodes' = '24-26': 'no-recurrence-events' (0.0)
|   'inv-nodes' = '27-29': 'no-recurrence-events' (0.0)
|   'inv-nodes' = '30-32': 'no-recurrence-events' (0.0)
|   'inv-nodes' = '33-35': 'no-recurrence-events' (0.0)
|   'inv-nodes' = '36-39': 'no-recurrence-events' (0.0)
'tumor-size' = '15-19':
|   'age' = Unknown: 'no-recurrence-events' (0.0)
|   'age' = '10-19': 'no-recurrence-events' (0.0)
|   'age' = '20-29': 'no-recurrence-events' (0.0)
|   'age' = '30-39': 'no-recurrence-events' (4.0/2.0)
|   'age' = '40-49': 'recurrence-events' (4.0/1.0)
|   'age' = '50-59': 'no-recurrence-events' (8.0/1.0)
|   'age' = '60-69': 'no-recurrence-events' (7.0)
|   'age' = '70-79': 'recurrence-events' (1.0)
|   'age' = '80-89': 'no-recurrence-events' (0.0)
|   'age' = '90-99': 'no-recurrence-events' (0.0)
'tumor-size' = '20-24':
|   'node-caps' = Unknown: 'recurrence-events' (2.0)
|   'node-caps' = 'yes': 'recurrence-events' (6.0/2.0)
|   'node-caps' = 'no': 'no-recurrence-events' (22.0/5.0)
'tumor-size' = '25-29':
|   'breast-quad' = Unknown: 'no-recurrence-events' (0.0)
|   'breast-quad' = 'left_up': 'no-recurrence-events' (12.0/3.0)
|   'breast-quad' = 'left_low': 'no-recurrence-events' (12.0/5.0)
|   'breast-quad' = 'right_up': 'recurrence-events' (5.0/1.0)
```

Figure 121. T2 Test Results

C4.5 — Run the scheme with Tree and Rule output and internal evaluation; view the results and compare these with the training dataset C4.5 results.

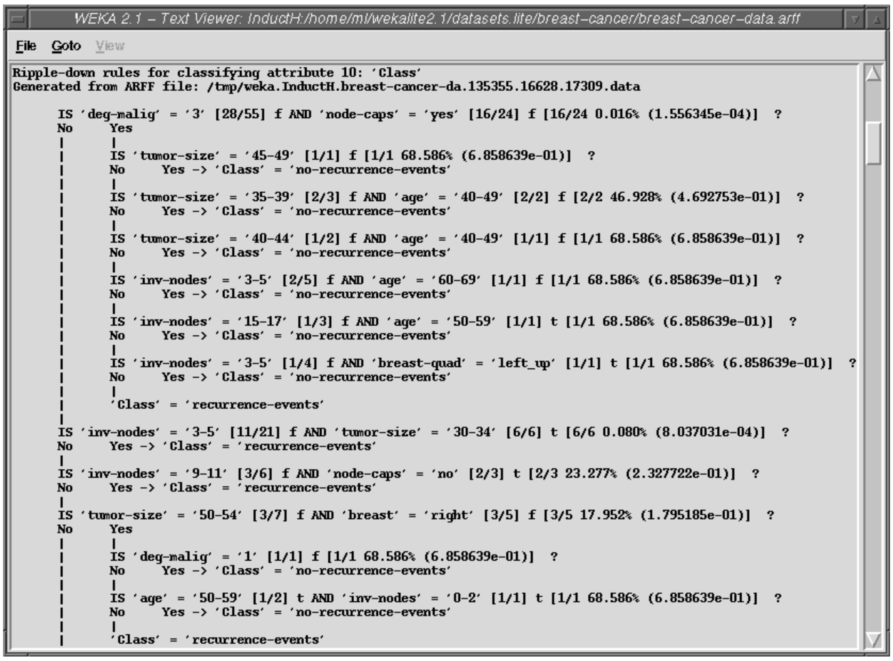


WEKA 2.1 – Text Viewer: C4.5:/home/ml/wekalite2.1/datasets.lite/breast–cancer/breast–cancer
File Goto View

```
Simplified Decision Tree:

'inv-nodes' = '0-2': 'no-recurrence-events' (139.0/35.0)
'inv-nodes' = '12-14': 'no-recurrence-events' (0.0)
'inv-nodes' = '18-20': 'no-recurrence-events' (0.0)
'inv-nodes' = '21-23': 'no-recurrence-events' (0.0)
'inv-nodes' = '24-26': 'recurrence-events' (1.0/0.8)
'inv-nodes' = '27-29': 'no-recurrence-events' (0.0)
'inv-nodes' = '30-32': 'no-recurrence-events' (0.0)
'inv-nodes' = '33-35': 'no-recurrence-events' (0.0)
'inv-nodes' = '36-39': 'no-recurrence-events' (0.0)
'inv-nodes' = '3-5':
|   'tumor-size' = '0-4': 'recurrence-events' (0.0)
|   'tumor-size' = '5-9': 'recurrence-events' (0.0)
|   'tumor-size' = '10-14': 'no-recurrence-events' (1.0/0.8)
|   'tumor-size' = '15-19': 'recurrence-events' (0.0)
|   'tumor-size' = '30-34': 'recurrence-events' (7.0/2.4)
|   'tumor-size' = '35-39': 'recurrence-events' (0.0)
|   'tumor-size' = '40-44': 'no-recurrence-events' (4.0/1.2)
|   'tumor-size' = '45-49': 'recurrence-events' (0.0)
|   'tumor-size' = '50-54': 'recurrence-events' (0.0)
|   'tumor-size' = '55-59': 'recurrence-events' (0.0)
|   'tumor-size' = '20-24':
|   |   'breast-quad' = 'left_up': 'no-recurrence-events' (2.0/1.0)
|   |   'breast-quad' = 'left_low': 'recurrence-events' (5.0/2.3)
|   |   'breast-quad' = 'right_up': 'recurrence-events' (1.0/0.8)
|   |   'breast-quad' = 'right_low': 'recurrence-events' (0.0)
|   |   'breast-quad' = 'central': 'recurrence-events' (0.0)
|   'tumor-size' = '25-29':
|   |   'breast' = 'left': 'recurrence-events' (2.0/1.0)
|   |   'breast' = 'right': 'no-recurrence-events' (5.0/2.3)
'inv-nodes' = '6-8':
|   'deg-malig' = '1': 'recurrence-events' (0.0)
|   'deg-malig' = '2': 'no-recurrence-events' (4.0/2.2)
|   'deg-malig' = '3': 'recurrence-events' (7.0/2.4)
'inv-nodes' = '9-11':
|   'irradiat' = 'yes': 'no-recurrence-events' (5.0/3.2)
|   'irradiat' = 'no': 'recurrence-events' (3.0/1.1)
'inv-nodes' = '15-17':
|   'menopause' = 'lt40': 'no-recurrence-events' (0.0)
|   'menopause' = 'ge40': 'no-recurrence-events' (3.0/1.1)
|   'menopause' = 'premeno': 'recurrence-events' (2.0/1.0)


Tree saved


Evaluation on training data (191 items):

         Before Pruning           After Pruning
        ----------------   ---------------------------
        Size      Errors   Size      Errors   Estimate

         120   24(12.6%)     41   38(19.9%)    (30.6%)   <<

Evaluation on test data (95 items):

         Before Pruning           After Pruning
        ----------------   ---------------------------
        Size      Errors   Size      Errors   Estimate

         120   36(37.9%)     41   24(25.3%)    (30.6%)   <<
```

Figure 122. C4.5 Decision Tree Test Results

InductH — Run the scheme with Rippledown options; view the results and compare these with the training dataset results.


WEKA 2.1 – Text Viewer: InductH:/home/ml/wekalite2.1/datasets.lite/breast-cancer/breast-cancer-data.arff
File
Goto
View

```
Ripple-down rules for classifying attribute 10: 'Class'
Generated from ARFF file: /tmp/weka.InductH.breast-cancer-da.135355.16628.17309.data

      IS 'deg-malig' = '3' [28/55] f AND 'node-caps' = 'yes' [16/24] f [16/24 0.016% (1.556345e-04)]  ?
      No     Yes
      |      |
      |      IS 'tumor-size' = '45-49' [1/1] f [1/1 68.586% (6.858639e-01)]  ?
      |      No     Yes -> 'Class' = 'no-recurrence-events'
      |      |
      |      IS 'tumor-size' = '35-39' [2/3] f AND 'age' = '40-49' [2/2] f [2/2 46.928% (4.692753e-01)]  ?
      |      No     Yes -> 'Class' = 'no-recurrence-events'
      |      |
      |      IS 'tumor-size' = '40-44' [1/2] f AND 'age' = '40-49' [1/1] f [1/1 68.586% (6.858639e-01)]  ?
      |      No     Yes -> 'Class' = 'no-recurrence-events'
      |      |
      |      IS 'inv-nodes' = '3-5' [2/5] f AND 'age' = '60-69' [1/1] f [1/1 68.586% (6.858639e-01)]  ?
      |      No     Yes -> 'Class' = 'no-recurrence-events'
      |      |
      |      IS 'inv-nodes' = '15-17' [1/3] f AND 'age' = '50-59' [1/1] t [1/1 68.586% (6.858639e-01)]  ?
      |      No     Yes -> 'Class' = 'no-recurrence-events'
      |      |
      |      IS 'inv-nodes' = '3-5' [1/4] f AND 'breast-quad' = 'left_up' [1/1] t [1/1 68.586% (6.858639e-01)]  ?
      |      No     Yes -> 'Class' = 'no-recurrence-events'
      |      |
      |      'Class' = 'recurrence-events'
      |
      IS 'inv-nodes' = '3-5' [11/21] f AND 'tumor-size' = '30-34' [6/6] t [6/6 0.080% (8.037031e-04)]  ?
      No     Yes -> 'Class' = 'recurrence-events'
      |
      IS 'inv-nodes' = '9-11' [3/6] f AND 'node-caps' = 'no' [2/3] t [2/3 23.277% (2.327722e-01)]  ?
      No     Yes -> 'Class' = 'recurrence-events'
      |
      IS 'tumor-size' = '50-54' [3/7] f AND 'breast' = 'right' [3/5] f [3/5 17.952% (1.795185e-01)]  ?
      No     Yes
      |      |
      |      IS 'deg-malig' = '1' [1/1] f [1/1 68.586% (6.858639e-01)]  ?
      |      No     Yes -> 'Class' = 'no-recurrence-events'
      |      |
      |      IS 'age' = '50-59' [1/2] t AND 'inv-nodes' = '0-2' [1/1] t [1/1 68.586% (6.858639e-01)]  ?
      |      No     Yes -> 'Class' = 'no-recurrence-events'
      |      |
      |      'Class' = 'recurrence-events'
```

Figure 123. InductH Test Results

Open the Experiment Editor. Create a new experiment and enter the experiment name and notes for the experiment.

The notes should be similar to those written for the golf experiment. The files that will be used will be both the breast cancer data and test datasets. The schemes that will be used are FOIL, K*, and PEBLS; each with default parameters.

Enter this information. Save the experiment and run the experiment. When this has finished view your results.

APPENDIX 2: GOLF AND BREAST CANCER ARFF FILES

Both the golf and breast cancer ARFF files are given below.

Golf

A simple dataset which uses weather information to decided whether to play golf or not.

Dataset information:
Number of Instances: 14
Number of Attributes: 4 + class attribute
Missing values: none.

Attribute Information:

ATTRIBUTE	TYPE	VALUES
outlook:	enumerated	sunny, overcast, rain
temperature:	real with range	[0.0,100]
humidity:	real	
windy:	enumerated	true, false
class:	enumerated	Play, ’Dont Play’

Table 4. Golf Dataset Attribute Information

Class Distribution:
The class for this dataset is called ‘class’.

Play:	9 (64.3%)
Dont Play':	5 (35.7%)

Note : The distribution is given as the number of instances with the class value with in the entire dataset (ie. out of 14 instances)

Data:

```
sunny, 85, 85, FALSE, ’Dont Play’
sunny, 80, 90, TRUE, ’Dont Play’
overcast, 83, 78, FALSE, Play
rain, 70, 96, FALSE, Play
rain, 68, 80, FALSE, Play
rain, 65, 70, TRUE, ’Dont Play’
overcast, 64, 65, TRUE, Play
sunny, 72, 95, FALSE, ’Dont Play’
sunny, 69, 70, FALSE, Play
rain, 75, 80, FALSE, Play
sunny, 75, 70, TRUE, Play
overcast, 72, 90, TRUE, Play
overcast, 81, 75, FALSE, Play
rain, 71, 80, TRUE, ’Dont Play’
```

Sources:

"C4.5: Programs for Machine Learning", Morgan Kaufmann, Oct 1992

Note: This is NOT a UCI dataset - SRG 7 Nov 1994

Breast Cancer

This is one of three domains provided by the Oncology Institute that has repeatedly appeared in the machine learning literature. (See also lymphography and primary-tumor.)

This data set includes 201 instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are linear and some are nominal.

Dataset information:

Number of Instances: 286.

201 instances are of one class, 85 are of another class.

Number of Attributes: nine + class attribute.

Some attributes are linear and some are nominal.

Missing values: there are nine missing values.

Eight values are missing from attribute node-caps, and one value is missing from the attribute breast-quad.

Attribute Information:

ATTRIBUTE	TYPE	VALUES
age:	enumerated	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
menopause:	enumerated	lt40, ge40, premeno
tumor-size:	enumerated	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59
inv-nodes:	enumerated	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26,27-29, 30-32, 33-35, 36-39
node-caps:	enumerated (binary)	yes, no
deg_malig:	enumerated	1, 2, 3
breast:	enumerated	left, right
breast-quad:	enumerated	left-up, left-low, right-up,right-low, central
irradiat:	enumerated	yes, no
Class:	enumerated	no-recurrence-events, recurrence-events

Table 5. Breast-cancer Dataset Attribute Information

Class Distribution:

The class for this dataset is called ‘Class’.

No-recurrence-events:	201
recurrence-events:	85

Note : The distribution is given as the number of instances with the class value with in the entire dataset (ie. out of 286 instances)

Data:

The data has been divided into training and testing datasets.

TRAINING FILE:

```
40-49, premeno, 15-19, 0-2, yes, 3, right, left_up, no, recurrence-events
50-59, ge40, 15-19, 0-2, no, 1, right, central, no, no-recurrence-events
50-59, ge40, 35-39, 0-2, no, 2, left, left_low, no, recurrence-events
40-49, premeno, 35-39, 0-2, yes, 3, right, left_low, yes, no-recurrence-events
40-49, premeno, 30-34, 3-5, yes, 2, left, right_up, no, recurrence-events
50-59, premeno, 25-29, 3-5, no, 2, right, left_up, yes, no-recurrence-events
50-59, ge40, 40-44, 0-2, no, 3, left, left_up, no, no-recurrence-events
40-49, premeno, 10-14, 0-2, no, 2, left, left_up, no, no-recurrence-events
40-49, premeno, 0-4, 0-2, no, 2, right, right_low, no, no-recurrence-events
40-49, ge40, 40-44, 15-17, yes, 2, right, left_up, yes, no-recurrence-events
50-59, premeno, 25-29, 0-2, no, 2, left, left_low, no, no-recurrence-events
60-69, ge40, 15-19, 0-2, no, 2, right, left_up, no, no-recurrence-events
50-59, ge40, 30-34, 0-2, no, 1, right, central, no, no-recurrence-events
50-59, ge40, 25-29, 0-2, no, 2, right, left_up, no, no-recurrence-events
40-49, premeno, 25-29, 0-2, no, 2, left, left_low, yes, recurrence-events
30-39, premeno, 20-24, 0-2, no, 3, left, central, no, no-recurrence-events
50-59, premeno, 10-14, 3-5, no, 1, right, left_up, no, no-recurrence-events
60-69, ge40, 15-19, 0-2, no, 2, right, left_up, no, no-recurrence-events
50-59, premeno, 40-44, 0-2, no, 2, left, left_up, no, no-recurrence-events
50-59, ge40, 20-24, 0-2, no, 3, left, left_up, no, no-recurrence-events
50-59, lt40, 20-24, 0-2, ?, 1, left, left_low, no, recurrence-events
60-69, ge40, 40-44, 3-5, no, 2, right, left_up, yes, no-recurrence-events
50-59, ge40, 15-19, 0-2, no, 2, right, left_low, no, no-recurrence-events
40-49, premeno, 10-14, 0-2, no, 1, right, left_up, no, no-recurrence-events
30-39, premeno, 15-19, 6-8, yes, 3, left, left_low, yes, recurrence-events
50-59, ge40, 20-24, 3-5, yes, 2, right, left_up, no, no-recurrence-events
50-59, ge40, 10-14, 0-2, no, 2, right, left_low, no, no-recurrence-events
40-49, premeno, 10-14, 0-2, no, 1, right, left_up, no, no-recurrence-events
60-69, ge40, 30-34, 3-5, yes, 3, left, left_low, no, no-recurrence-events
40-49, premeno, 15-19, 15-17, yes, 3, left, left_low, no, recurrence-events
60-69, ge40, 30-34, 0-2, no, 3, right, central, no, recurrence-events
60-69, ge40, 25-29, 3-5, ?, 1, right, left_low, yes, no-recurrence-events
50-59, ge40, 25-29, 0-2, no, 3, left, right_up, no, no-recurrence-events
50-59, ge40, 20-24, 0-2, no, 3, right, left_up, no, no-recurrence-events
40-49, premeno, 30-34, 0-2, no, 1, left, left_low, yes, recurrence-events
30-39, premeno, 15-19, 0-2, no, 1, left, left_low, no, no-recurrence-events
40-49, premeno, 10-14, 0-2, no, 2, right, left_up, no, no-recurrence-events
60-69, ge40, 45-49, 6-8, yes, 3, left, central, no, no-recurrence-events
40-49, ge40, 20-24, 0-2, no, 3, left, left_low, no, no-recurrence-events
40-49, premeno, 10-14, 0-2, no, 1, right, right_low, no, no-recurrence-events
30-39, premeno, 35-39, 0-2, no, 3, left, left_low, no, recurrence-events
40-49, premeno, 35-39, 9-11, yes, 2, right, right_up, yes, no-recurrence-events
60-69, ge40, 25-29, 0-2, no, 2, right, left_low, no, no-recurrence-events
50-59, ge40, 20-24, 3-5, yes, 3, right, right_up, no, recurrence-events
30-39, premeno, 15-19, 0-2, no, 1, left, left_low, no, no-recurrence-events
50-59, premeno, 30-34, 0-2, no, 3, left, right_up, no, recurrence-events
60-69, ge40, 10-14, 0-2, no, 2, right, left_up, yes, no-recurrence-events
40-49, premeno, 35-39, 0-2, yes, 3, right, left_up, yes, no-recurrence-events
50-59, premeno, 50-54, 0-2, yes, 2, right, left_up, yes, no-recurrence-events
50-59, ge40, 40-44, 0-2, no, 3, right, left_up, no, no-recurrence-events
```

```
70-79, ge40, 15-19, 9-11, ?, 1, left, left_low, yes, recurrence-events
50-59, lt40, 30-34, 0-2, no, 3, right, left_up, no, no-recurrence-events
40-49, premeno, 0-4, 0-2, no, 3, left, central, no, no-recurrence-events
70-79, ge40, 40-44, 0-2, no, 1, right, right_up, no, no-recurrence-events
40-49, premeno, 25-29, 0-2, ?, 2, left, right_low, yes, no-recurrence-events
50-59, ge40, 25-29, 15-17, yes, 3, right, left_up, no, no-recurrence-events
50-59, premeno, 20-24, 0-2, no, 1, left, left_low, no, no-recurrence-events
50-59, ge40, 35-39, 15-17, no, 3, left, left_low, no, no-recurrence-events
50-59, ge40, 50-54, 0-2, no, 1, right, right_up, no, no-recurrence-events
30-39, premeno, 0-4, 0-2, no, 2, right, central, no, recurrence-events
50-59, ge40, 40-44, 6-8, yes, 3, left, left_low, yes, recurrence-events
40-49, premeno, 30-34, 0-2, no, 2, right, right_up, yes, no-recurrence-events
40-49, ge40, 20-24, 0-2, no, 3, left, left_up, no, no-recurrence-events
40-49, premeno, 30-34, 15-17, yes, 3, left, left_low, no, recurrence-events
40-49, ge40, 20-24, 0-2, no, 2, right, left_up, no, recurrence-events
50-59, ge40, 15-19, 0-2, no, 1, right, central, no, no-recurrence-events
30-39, premeno, 25-29, 0-2, no, 2, right, left_low, no, no-recurrence-events
60-69, ge40, 15-19, 0-2, no, 2, left, left_low, no, no-recurrence-events
50-59, premeno, 50-54, 9-11, yes, 2, right, left_up, no, recurrence-events
30-39, premeno, 10-14, 0-2, no, 1, right, left_low, no, no-recurrence-events
50-59, premeno, 25-29, 3-5, yes, 3, left, left_low, yes, recurrence-events
60-69, ge40, 25-29, 3-5, ?, 1, right, left_up, yes, no-recurrence-events
60-69, ge40, 10-14, 0-2, no, 1, right, left_low, no, no-recurrence-events
50-59, ge40, 30-34, 6-8, yes, 3, left, right_low, no, recurrence-events
30-39, premeno, 25-29, 6-8, yes, 3, left, right_low, yes, recurrence-events
50-59, ge40, 10-14, 0-2, no, 1, left, left_low, no, no-recurrence-events
50-59, premeno, 15-19, 0-2, no, 1, left, left_low, no, no-recurrence-events
40-49, premeno, 25-29, 0-2, no, 2, right, central, no, no-recurrence-events
40-49, premeno, 25-29, 0-2, no, 3, left, right_up, no, recurrence-events
60-69, ge40, 30-34, 6-8, yes, 2, right, right_up, no, no-recurrence-events
50-59, lt40, 15-19, 0-2, no, 2, left, left_low, no, no-recurrence-events
40-49, premeno, 25-29, 0-2, no, 2, right, left_low, no, no-recurrence-events
40-49, premeno, 30-34, 0-2, no, 1, right, left_up, no, no-recurrence-events
60-69, ge40, 15-19, 0-2, no, 2, left, left_up, yes, no-recurrence-events
30-39, premeno, 0-4, 0-2, no, 2, right, central, no, no-recurrence-events
50-59, ge40, 35-39, 0-2, no, 3, left, left_up, no, no-recurrence-events
40-49, premeno, 40-44, 0-2, no, 1, right, left_up, no, no-recurrence-events
30-39, premeno, 25-29, 6-8, yes, 2, right, left_up, yes, no-recurrence-events
50-59, ge40, 20-24, 0-2, no, 1, right, left_low, no, no-recurrence-events
50-59, ge40, 30-34, 0-2, no, 1, left, left_up, no, no-recurrence-events
60-69, ge40, 20-24, 0-2, no, 1, right, left_up, no, recurrence-events
30-39, premeno, 30-34, 3-5, no, 3, right, left_up, yes, recurrence-events
50-59, lt40, 20-24, 0-2, ?, 1, left, left_up, no, recurrence-events
50-59, premeno, 10-14, 0-2, no, 2, right, left_up, no, no-recurrence-events
50-59, ge40, 20-24, 0-2, no, 2, right, left_up, no, no-recurrence-events
40-49, premeno, 45-49, 0-2, no, 2, left, left_low, yes, no-recurrence-events
30-39, premeno, 40-44, 0-2, no, 1, left, left_up, no, recurrence-events
50-59, premeno, 10-14, 0-2, no, 1, left, left_low, no, no-recurrence-events
60-69, ge40, 30-34, 0-2, no, 3, right, left_up, yes, recurrence-events
40-49, premeno, 35-39, 0-2, no, 1, right, left_up, no, recurrence-events
40-49, premeno, 20-24, 3-5, yes, 2, left, left_low, yes, recurrence-events
50-59, premeno, 15-19, 0-2, no, 2, left, left_low, no, recurrence-events
50-59, ge40, 30-34, 0-2, no, 3, right, left_low, no, no-recurrence-events
60-69, ge40, 20-24, 0-2, no, 2, left, left_up, no, no-recurrence-events
40-49, premeno, 20-24, 0-2, no, 1, left, right_low, no, no-recurrence-events
60-69, ge40, 30-34, 3-5, yes, 2, left, central, yes, recurrence-events
60-69, ge40, 20-24, 3-5, no, 2, left, left_low, yes, recurrence-events
50-59, premeno, 25-29, 0-2, no, 2, left, right_up, no, recurrence-events
```

```
50-59, ge40, 30-34, 0-2, no, 1, right, right_up, no, no-recurrence-events
40-49, premeno, 20-24, 0-2, no, 2, left, right_low, no, no-recurrence-events
60-69, ge40, 15-19, 0-2, no, 1, right, left_up, no, no-recurrence-events
60-69, ge40, 30-34, 0-2, no, 2, left, left_low, yes, no-recurrence-events
30-39, premeno, 30-34, 0-2, no, 2, left, left_up, no, no-recurrence-events
30-39, premeno, 40-44, 3-5, no, 3, right, right_up, yes, no-recurrence-events
60-69, ge40, 5-9, 0-2, no, 1, left, central, no, no-recurrence-events
60-69, ge40, 10-14, 0-2, no, 1, left, left_up, no, no-recurrence-events
40-49, premeno, 30-34, 6-8, yes, 3, right, left_up, no, recurrence-events
60-69, ge40, 10-14, 0-2, no, 1, left, left_up, no, no-recurrence-events
40-49, premeno, 35-39, 9-11, yes, 2, right, left_up, yes, no-recurrence-events
40-49, premeno, 20-24, 0-2, no, 1, right, left_low, no, no-recurrence-events
40-49, premeno, 30-34, 0-2, yes, 3, right, right_up, no, recurrence-events
50-59, premeno, 25-29, 0-2, yes, 2, left, left_up, no, no-recurrence-events
40-49, premeno, 15-19, 0-2, no, 2, left, left_low, no, no-recurrence-events
30-39, premeno, 35-39, 9-11, yes, 3, left, left_low, no, recurrence-events
30-39, premeno, 10-14, 0-2, no, 2, left, right_low, no, no-recurrence-events
50-59, ge40, 30-34, 0-2, no, 1, right, left_low, no, no-recurrence-events
60-69, ge40, 30-34, 0-2, no, 2, left, left_up, no, no-recurrence-events
60-69, ge40, 25-29, 0-2, no, 2, left, left_low, no, no-recurrence-events
40-49, premeno, 15-19, 0-2, no, 2, left, left_up, no, recurrence-events
60-69, ge40, 15-19, 0-2, no, 2, right, left_low, no, no-recurrence-events
40-49, premeno, 30-34, 0-2, no, 2, left, right_low, no, no-recurrence-events
20-29, premeno, 35-39, 0-2, no, 2, right, right_up, no, no-recurrence-events
40-49, premeno, 30-34, 0-2, no, 3, right, right_up, no, recurrence-events
40-49, premeno, 25-29, 0-2, no, 2, right, left_low, no, recurrence-events
30-39, premeno, 30-34, 0-2, no, 3, left, left_low, no, no-recurrence-events
30-39, premeno, 15-19, 0-2, no, 1, right, left_low, no, recurrence-events
50-59, ge40, 0-4, 0-2, no, 1, right, central, no, no-recurrence-events
50-59, ge40, 0-4, 0-2, no, 1, left, left_low, no, no-recurrence-events
60-69, ge40, 50-54, 0-2, no, 3, right, left_up, no, recurrence-events
50-59, premeno, 30-34, 0-2, no, 1, left, central, no, no-recurrence-events
60-69, ge40, 20-24, 24-26, yes, 3, left, left_low, yes, recurrence-events
40-49, premeno, 25-29, 0-2, no, 2, left, left_up, no, no-recurrence-events
40-49, premeno, 30-34, 3-5, no, 2, right, left_up, no, recurrence-events
50-59, premeno, 20-24, 3-5, yes, 2, left, left_low, no, no-recurrence-events
50-59, ge40, 15-19, 0-2, yes, 2, left, central, yes, no-recurrence-events
50-59, premeno, 10-14, 0-2, no, 3, left, left_low, no, no-recurrence-events
30-39, premeno, 30-34, 9-11, no, 2, right, left_up, yes, recurrence-events
60-69, ge40, 10-14, 0-2, no, 1, left, left_low, no, no-recurrence-events
40-49, premeno, 40-44, 0-2, no, 2, right, left_low, no, no-recurrence-events
50-59, ge40, 30-34, 9-11, ?, 3, left, left_up, yes, no-recurrence-events
40-49, premeno, 50-54, 0-2, no, 2, right, left_low, yes, recurrence-events
50-59, ge40, 15-19, 0-2, no, 2, right, right_up, no, no-recurrence-events
50-59, ge40, 40-44, 3-5, yes, 2, left, left_low, no, no-recurrence-events
30-39, premeno, 25-29, 3-5, yes, 3, left, left_low, yes, recurrence-events
60-69, ge40, 10-14, 0-2, no, 2, left, left_low, no, no-recurrence-events
60-69, lt40, 10-14, 0-2, no, 1, left, right_up, no, no-recurrence-events
30-39, premeno, 30-34, 0-2, no, 2, left, left_up, no, recurrence-events
30-39, premeno, 20-24, 3-5, yes, 2, left, left_low, no, recurrence-events
50-59, ge40, 10-14, 0-2, no, 1, right, left_up, no, no-recurrence-events
60-69, ge40, 25-29, 0-2, no, 3, right, left_up, no, no-recurrence-events
50-59, ge40, 25-29, 3-5, yes, 3, right, left_up, no, no-recurrence-events
40-49, premeno, 30-34, 6-8, no, 2, left, left_up, no, no-recurrence-events
60-69, ge40, 50-54, 0-2, no, 2, left, left_low, no, no-recurrence-events
50-59, premeno, 30-34, 0-2, no, 3, left, left_low, no, no-recurrence-events
40-49, ge40, 20-24, 3-5, no, 3, right, left_low, yes, recurrence-events
50-59, ge40, 30-34, 6-8, yes, 2, left, right_low, yes, recurrence-events
```

```
60-69, ge40, 25-29, 3-5, no, 2, right, right_up, no, recurrence-events
40-49, premeno, 20-24, 0-2, no, 2, left, central, no, no-recurrence-events
40-49, premeno, 20-24, 0-2, no, 2, left, left_up, no, no-recurrence-events
40-49, premeno, 50-54, 0-2, no, 2, left, left_low, no, no-recurrence-events
50-59, ge40, 20-24, 0-2, no, 2, right, central, no, recurrence-events
50-59, ge40, 30-34, 3-5, no, 3, right, left_up, no, recurrence-events
40-49, ge40, 25-29, 0-2, no, 2, left, left_low, no, no-recurrence-events
50-59, premeno, 25-29, 0-2, no, 1, right, left_up, no, recurrence-events
40-49, premeno, 40-44, 3-5, yes, 3, right, left_up, yes, no-recurrence-events
40-49, premeno, 20-24, 0-2, no, 2, right, left_up, no, no-recurrence-events
40-49, premeno, 20-24, 3-5, no, 2, right, left_up, no, no-recurrence-events
40-49, premeno, 25-29, 9-11, yes, 3, right, left_up, no, recurrence-events
40-49, premeno, 25-29, 0-2, no, 2, right, left_low, no, recurrence-events
40-49, premeno, 20-24, 0-2, no, 1, right, right_up, no, no-recurrence-events
30-39, premeno, 40-44, 0-2, no, 2, right, right_up, no, no-recurrence-events
60-69, ge40, 10-14, 6-8, yes, 3, left, left_up, yes, recurrence-events
40-49, premeno, 35-39, 0-2, no, 1, left, left_low, no, no-recurrence-events
50-59, ge40, 30-34, 3-5, no, 3, left, left_low, no, recurrence-events
40-49, premeno, 5-9, 0-2, no, 1, left, left_low, yes, no-recurrence-events
60-69, ge40, 15-19, 0-2, no, 1, left, right_low, no, no-recurrence-events
40-49, premeno, 30-34, 0-2, no, 3, right, right_up, no, no-recurrence-events
40-49, premeno, 25-29, 0-2, no, 3, left, left_up, no, recurrence-events
50-59, ge40, 5-9, 0-2, no, 2, right, right_up, no, no-recurrence-events
50-59, premeno, 25-29, 0-2, no, 2, right, right_low, no, no-recurrence-events
50-59, premeno, 25-29, 0-2, no, 2, left, right_up, no, recurrence-events
```

TESTING FILE:

```
40-49, premeno, 10-14, 0-2, no, 2, left, left_low, yes, no-recurrence-events
60-69, ge40, 35-39, 6-8, yes, 3, left, left_low, no, recurrence-events
60-69, ge40, 50-54, 0-2, no, 2, right, left_up, yes, no-recurrence-events
40-49, premeno, 25-29, 0-2, no, 2, right, left_up, no, no-recurrence-events
30-39, premeno, 20-24, 3-5, no, 2, right, central, no, no-recurrence-events
30-39, premeno, 30-34, 0-2, no, 1, right, left_up, no, recurrence-events
60-69, lt40, 30-34, 0-2, no, 1, left, left_low, no, no-recurrence-events
40-49, premeno, 15-19, 12-14, no, 3, right, right_low, yes, no-recurrence-events
60-69, ge40, 20-24, 0-2, no, 3, right, left_low, no, recurrence-events
30-39, premeno, 5-9, 0-2, no, 2, left, right_low, no, no-recurrence-events
40-49, premeno, 30-34, 0-2, no, 3, left, left_up, no, no-recurrence-events
60-69, ge40, 30-34, 0-2, no, 3, left, left_low, no, no-recurrence-events
40-49, premeno, 25-29, 0-2, no, 1, right, right_low, no, no-recurrence-events
40-49, premeno, 25-29, 0-2, no, 1, left, right_low, no, no-recurrence-events
60-69, ge40, 40-44, 3-5, yes, 3, right, left_low, no, recurrence-events
50-59, ge40, 25-29, 0-2, no, 2, left, left_low, no, no-recurrence-events
50-59, premeno, 30-34, 0-2, no, 3, right, left_up, yes, recurrence-events
40-49, ge40, 30-34, 3-5, no, 3, left, left_low, no, recurrence-events
40-49, premeno, 25-29, 0-2, no, 1, right, left_low, yes, no-recurrence-events
40-49, ge40, 25-29, 12-14, yes, 3, left, right_low, yes, recurrence-events
40-49, premeno, 40-44, 0-2, no, 1, left, left_low, no, recurrence-events
40-49, premeno, 20-24, 0-2, no, 2, left, left_low, no, no-recurrence-events
50-59, ge40, 25-29, 0-2, no, 1, left, right_low, no, no-recurrence-events
40-49, premeno, 20-24, 0-2, no, 2, right, left_up, no, no-recurrence-events
70-79, ge40, 40-44, 0-2, no, 1, right, left_up, no, no-recurrence-events
60-69, ge40, 25-29, 0-2, no, 3, left, left_up, no, recurrence-events
50-59, premeno, 25-29, 0-2, no, 2, left, left_low, no, no-recurrence-events
60-69, ge40, 45-49, 0-2, no, 1, right, right_up, yes, recurrence-events
50-59, ge40, 20-24, 0-2, yes, 2, right, left_up, no, no-recurrence-events
```

```
50-59, ge40, 25-29, 0-2, no, 1, left, left_low, no, no-recurrence-events
50-59, ge40, 20-24, 0-2, no, 3, left, left_up, no, no-recurrence-events
40-49, premeno, 20-24, 3-5, no, 2, right, left_low, no, no-recurrence-events
50-59, ge40, 35-39, 0-2, no, 2, left, left_up, no, no-recurrence-events
30-39, premeno, 20-24, 0-2, no, 3, left, left_up, yes, recurrence-events
60-69, ge40, 30-34, 0-2, no, 1, right, left_up, no, no-recurrence-events
60-69, ge40, 25-29, 0-2, no, 3, right, left_low, no, no-recurrence-events
40-49, ge40, 30-34, 0-2, no, 2, left, left_up, yes, no-recurrence-events
30-39, premeno, 25-29, 0-2, no, 2, left, left_low, no, no-recurrence-events
40-49, premeno, 20-24, 0-2, no, 2, left, left_low, no, recurrence-events
30-39, premeno, 20-24, 0-2, no, 2, left, right_low, no, no-recurrence-events
40-49, premeno, 10-14, 0-2, no, 2, right, left_low, no, no-recurrence-events
50-59, premeno, 15-19, 0-2, no, 2, right, right_low, no, no-recurrence-events
50-59, premeno, 25-29, 0-2, no, 1, right, left_up, no, no-recurrence-events
60-69, ge40, 20-24, 0-2, no, 2, right, left_up, no, no-recurrence-events
60-69, ge40, 40-44, 0-2, no, 2, right, left_low, no, recurrence-events
30-39, lt40, 15-19, 0-2, no, 3, right, left_up, no, no-recurrence-events
40-49, premeno, 30-34, 12-14, yes, 3, left, left_up, yes, recurrence-events
60-69, ge40, 30-34, 0-2, yes, 2, right, right_up, yes, recurrence-events
50-59, ge40, 40-44, 6-8, yes, 3, left, left_low, yes, recurrence-events
50-59, ge40, 30-34, 0-2, no, 3, left, ?, no, recurrence-events
70-79, ge40, 10-14, 0-2, no, 2, left, central, no, no-recurrence-events
30-39, premeno, 40-44, 0-2, no, 2, left, left_low, yes, no-recurrence-events
40-49, premeno, 30-34, 0-2, no, 2, right, right_low, no, no-recurrence-events
40-49, premeno, 30-34, 0-2, no, 1, left, left_low, no, no-recurrence-events
60-69, ge40, 15-19, 0-2, no, 2, left, left_low, no, no-recurrence-events
40-49, premeno, 10-14, 0-2, no, 2, left, left_low, no, no-recurrence-events
60-69, ge40, 20-24, 0-2, no, 1, left, left_low, no, no-recurrence-events
50-59, ge40, 10-14, 0-2, no, 1, left, left_up, no, no-recurrence-events
50-59, premeno, 25-29, 0-2, no, 1, left, left_low, no, no-recurrence-events
50-59, ge40, 30-34, 9-11, yes, 3, left, right_low, yes, recurrence-events
50-59, ge40, 10-14, 0-2, no, 2, left, left_low, no, no-recurrence-events
40-49, premeno, 30-34, 0-2, no, 1, left, right_up, no, no-recurrence-events
70-79, ge40, 0-4, 0-2, no, 1, left, right_low, no, no-recurrence-events
40-49, premeno, 25-29, 0-2, no, 3, right, left_up, yes, no-recurrence-events
50-59, premeno, 25-29, 0-2, no, 3, right, left_low, yes, recurrence-events
50-59, ge40, 40-44, 0-2, no, 2, left, left_low, no, no-recurrence-events
60-69, ge40, 25-29, 0-2, no, 3, left, right_low, yes, recurrence-events
40-49, premeno, 30-34, 3-5, yes, 2, right, left_low, no, no-recurrence-events
50-59, ge40, 20-24, 0-2, no, 2, left, left_up, no, recurrence-events
70-79, ge40, 20-24, 0-2, no, 3, left, left_up, no, no-recurrence-events
30-39, premeno, 25-29, 0-2, no, 1, left, central, no, no-recurrence-events
60-69, ge40, 30-34, 0-2, no, 2, left, left_low, no, no-recurrence-events
40-49, premeno, 20-24, 3-5, yes, 2, right, right_up, yes, recurrence-events
50-59, ge40, 30-34, 9-11, ?, 3, left, left_low, yes, no-recurrence-events
50-59, ge40, 0-4, 0-2, no, 2, left, central, no, no-recurrence-events
40-49, premeno, 20-24, 0-2, no, 3, right, left_low, yes, no-recurrence-events
30-39, premeno, 35-39, 0-2, no, 3, left, left_low, no, recurrence-events
60-69, ge40, 30-34, 0-2, no, 1, left, left_up, no, no-recurrence-events
60-69, ge40, 20-24, 0-2, no, 1, left, left_low, no, no-recurrence-events
50-59, ge40, 25-29, 6-8, no, 3, left, left_low, yes, recurrence-events
50-59, premeno, 35-39, 15-17, yes, 3, right, right_up, no, recurrence-events
30-39, premeno, 20-24, 3-5, yes, 2, right, left_up, yes, no-recurrence-events
40-49, premeno, 20-24, 6-8, no, 2, right, left_low, yes, no-recurrence-events
50-59, ge40, 35-39, 0-2, no, 3, left, left_low, no, no-recurrence-events
50-59, premeno, 35-39, 0-2, no, 2, right, left_up, no, no-recurrence-events
40-49, premeno, 25-29, 0-2, no, 2, left, left_up, yes, no-recurrence-events
40-49, premeno, 35-39, 0-2, no, 2, right, right_up, no, no-recurrence-events
```

50-59, premeno, 30-34, 3-5, yes, 2, left, left_low, yes, no-recurrence-events
40-49, premeno, 20-24, 0-2, no, 2, right, right_up, no, no-recurrence-events
60-69, ge40, 15-19, 0-2, no, 3, right, left_up, yes, no-recurrence-events
50-59, ge40, 30-34, 6-8, yes, 2, left, left_low, no, no-recurrence-events
50-59, premeno, 25-29, 3-5, yes, 2, left, left_low, yes, no-recurrence-events
30-39, premeno, 30-34, 6-8, yes, 2, right, right_up, no, no-recurrence-events
50-59, premeno, 15-19, 0-2, no, 2, right, left_low, no, no-recurrence-events
50-59, ge40, 40-44, 0-2, no, 3, left, right_up, no, no-recurrence-events

Citation Request:

This breast cancer domain was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. Thanks go to M. Zwitter and M. Soklic for providing the data.

Sources:

Matjaz Zwitter & Milan Soklic (physicians)
Institute of Oncology
University Medical Center
Ljubljana, Yugoslavia

Donors:

Ming Tan and Jeff Schlimmer (Jeffrey.Schlimmer@a.gp.cs.cmu.edu)
Date: 11 July 1988