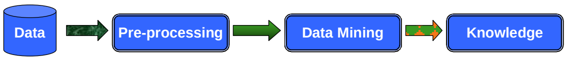
Data Mining and Statistics Within the Health Services

Tutorial for Weka

a data mining tool

Dr. Wenjia Wang
School of Computing Sciences
University of East Anglia



Content

1. Introduction to Weka
2. Data Mining Functions and Tools
3. Data Format
4. Hands-on Demos
   4.1 Weka Explorer
   - Classification
   - Attribute(feature) Selection
   4.2 Weka Experimenter
   4.3 Weka KnowledgeFlow
5. Summary

1. Introduction to WEKA

- A collection of open source of many data mining and machine learning algorithms, including
  – pre-processing on data
  – Classification:
  – clustering
  – association rule extraction
- Created by researchers at the University of Waikato in New Zealand
- Java based (also open source).

Weka Main Features

- 49 data preprocessing tools
- 76 classification/regression algorithms
- 8 clustering algorithms
- 15 attribute/subset evaluators + 10 search algorithms for feature selection.
- 3 algorithms for finding association rules
- 3 graphical user interfaces
  – "The Explorer" (exploratory data analysis)
  – "The Experimenter" (experimental environment)
  – "The KnowledgeFlow" (new process model inspired interface)

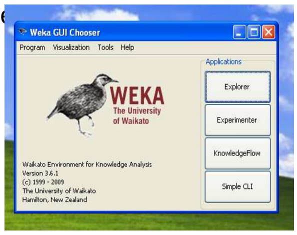
Weka: Download and Installation

- Download Weka (the stable version) from
  http://www.cs.waikato.ac.nz/ml/weka/
  – Choose a self-extracting executable (including Java VM)

  – (If you are interested in modifying/extending weka there is a developer version that includes the source code)
- After download is completed, run the self-extracting file to install Weka, and use the default set-ups.

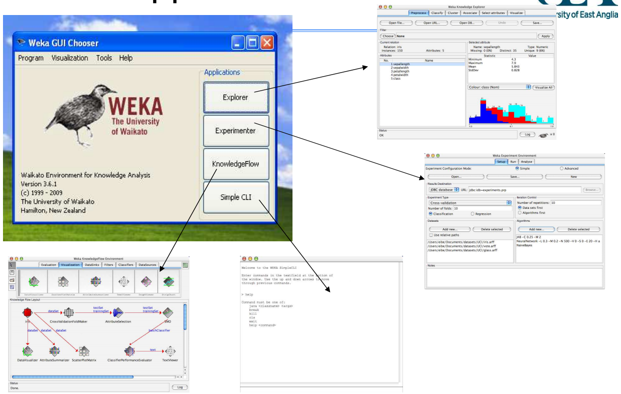
Start the Weka

- From windows desktop,
  – click "Start", choose "All programs",
  – Choose "Weka 3.6" to start Weka
  – Then the first interface window appears:
    Weka **GUI Chooser**.



Waikato Environment for Knowledge Analysis
Version 3.6.1
(c) 1999 - 2009
The University of Waikato
Hamilton, New Zealand

WEKA Application Interfaces



```
Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.

> help

Command must be one of:
    java <classname> <args>
    break
    kill
    cls
    exit
    help <command>
```

Weka Application Interfaces

- Explorer
  - preprocessing, attribute selection, learning, visualiation
- Experimenter
  - testing and evaluating machine learning algorithms
- Knowledge Flow
  - visual design of KDD process
  - Explorer
- Simple Command-line
  - A simple interface for typing commands

2. Weka Functions and Tools

- Preprocessing Filters
- Attribute selection
- Classification/Regression
- Clustering
- Association discovery
- Visualization

Load data file and Preprocessing

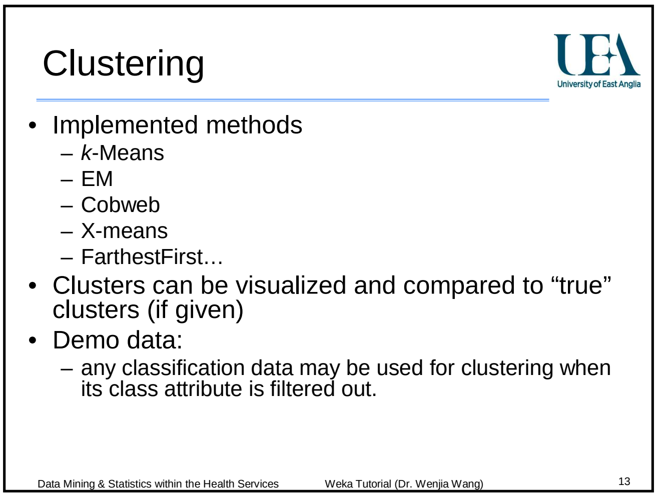
- Load data file in formats: ARFF, CSV, C4.5, binary
- Import from URL or SQL database (using JDBC)
- Preprocessing filters
  - Adding/removing attributes
  - Attribute value substitution
  - Discretization
  - Time series filters (delta, shift)
  - Sampling, randomization
  - Missing value management
  - Normalization and other numeric transformations

Feature Selection

- Very flexible: arbitrary combination of search and evaluation methods
- Search methods
  - best-first
  - genetic
  - ranking ...
- Evaluation measures
  - ReliefF
  - information gain
  - gain ratio
- Demo data: weather_nominal.arff

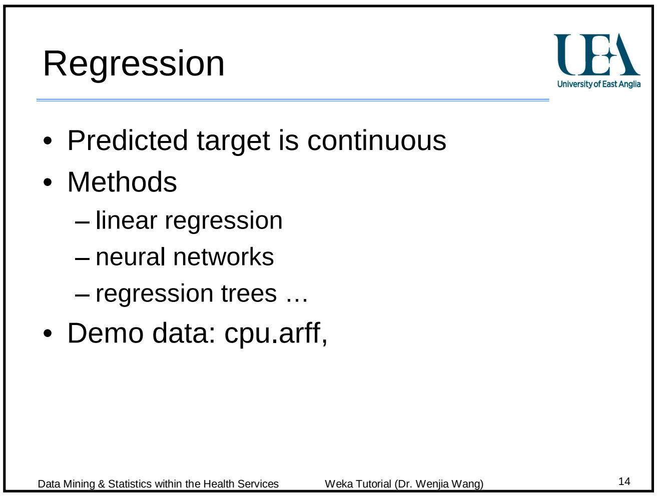
Classification

- Predicted target must be categorical
- Implemented methods
  - decision trees(J48, etc.) and rules
  - Naïve Bayes
  - neural networks
  - instance-based classifiers …
- Evaluation methods
  - test data set
  - crossvalidation
- Demo data: iris, contact lenses, labor, soybeans, etc.



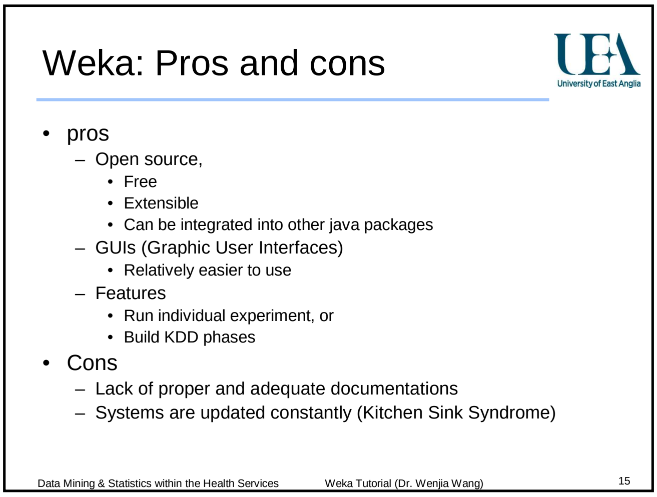
Clustering

- Implemented methods
  - *k*-Means
  - EM
  - Cobweb
  - X-means
  - FarthestFirst…
- Clusters can be visualized and compared to “true” clusters (if given)
- Demo data:
  - any classification data may be used for clustering when its class attribute is filtered out.



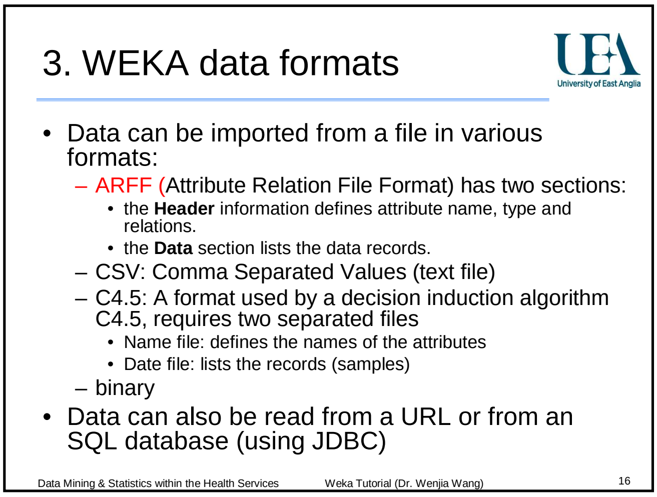
Regression

- Predicted target is continuous
- Methods
  - linear regression
  - neural networks
  - regression trees …
- Demo data: cpu.arff,



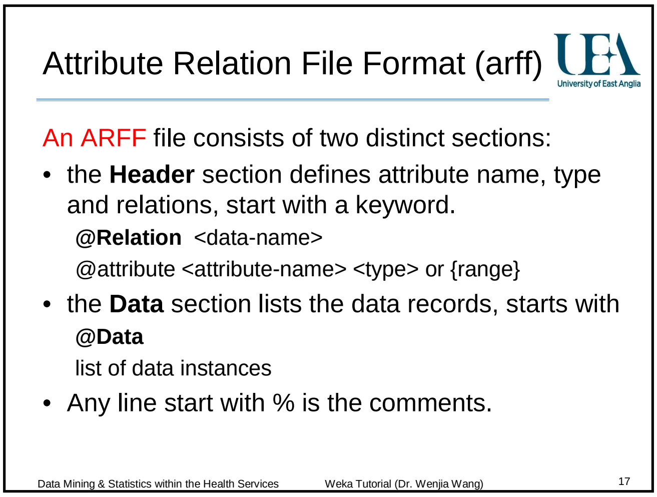
Weka: Pros and cons

- pros
  - Open source,
    - Free
    - Extensible
    - Can be integrated into other java packages
  - GUIs (Graphic User Interfaces)
    - Relatively easier to use
  - Features
    - Run individual experiment, or
    - Build KDD phases
- Cons
  - Lack of proper and adequate documentations
  - Systems are updated constantly (Kitchen Sink Syndrome)



3. WEKA data formats

- Data can be imported from a file in various formats:
  - ARFF (Attribute Relation File Format) has two sections:
    - the **Header** information defines attribute name, type and relations.
    - the **Data** section lists the data records.
  - CSV: Comma Separated Values (text file)
  - C4.5: A format used by a decision induction algorithm C4.5, requires two separated files
    - Name file: defines the names of the attributes
    - Date file: lists the records (samples)
  - binary
- Data can also be read from a URL or from an SQL database (using JDBC)



Attribute Relation File Format (arff)

An ARFF file consists of two distinct sections:

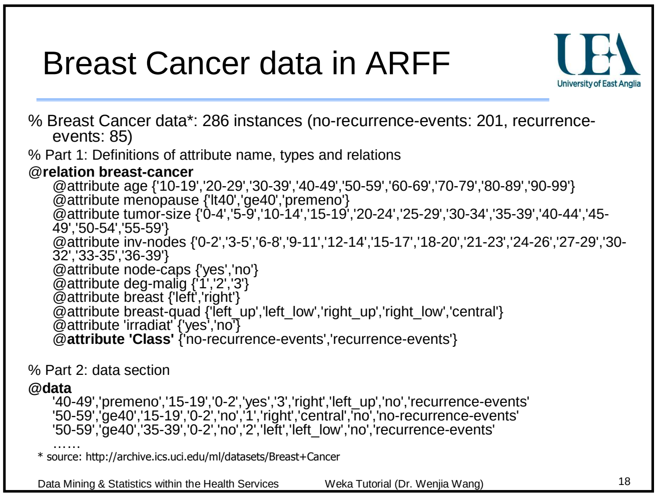
- the **Header** section defines attribute name, type and relations, start with a keyword.

  @Relation <data-name>

  @attribute <attribute-name> <type> or {range}
- the **Data** section lists the data records, starts with

  @Data

  list of data instances
- Any line start with % is the comments.



Breast Cancer data in ARFF

```
% Breast Cancer data*: 286 instances (no-recurrence-events: 201, recurrence-
    events: 85)
% Part 1: Definitions of attribute name, types and relations
@relation breast-cancer
    @attribute age {'10-19','20-29','30-39','40-49','50-59','60-69','70-79','80-89','90-99'}
    @attribute menopause {'lt40','ge40','premeno'}
    @attribute tumor-size {'0-4','5-9','10-14','15-19','20-24','25-29','30-34','35-39','40-44','45-
    49','50-54','55-59'}
    @attribute inv-nodes {'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'}
    @attribute node-caps {'yes','no'}
    @attribute deg-malig {'1','2','3'}
    @attribute breast {'left','right'}
    @attribute breast-quad {'left_up','left_low','right_up','right_low','central'}
    @attribute 'irradiat' {'yes','no'}
    @attribute 'Class' {'no-recurrence-events','recurrence-events'}

% Part 2: data section
@data
    '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'
    ……
```

* source: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer

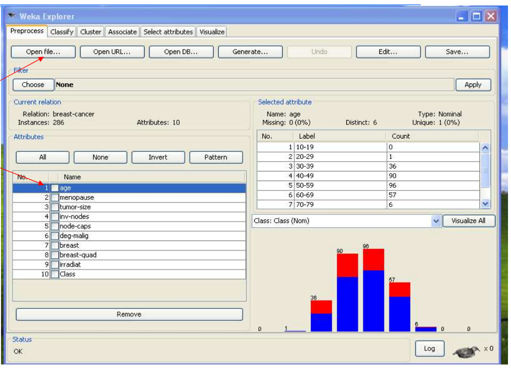
4.1 WEKA Explorer

- Click the Explorer on Weka GUI Chooser
- On the Explorer window,
  - click button “Open File” to open a data file from
    - the folder where your data files stored.
      e.g. Breast Cancer data: breast_cancer.arff

    Or (if you don’t have this data set),
    - the data folder provided by the weka package:
      e.g. C:\Program Files\Weka-3-6\data
      using “iris.arff” or “weather_nominal.arff”

Weka Explorer: open data file

- Open Breast Cancer data
- Click an attribute, e.g. age, then its distribution will be displayed in a histogram.



Weka Explorer

Filter: None

Current relation
Relation: breast-cancer
Instances: 286 Attributes: 10

Attributes

No.	Name
1	age
2	menopause
3	tumor-size
4	inv-nodes
5	node-caps
6	deg-malig
7	breast
8	breast-quad
9	irradiat
10	Class

Selected attribute
Name: age Type: Nominal
Missing: 0 (0%) Distinct: 6 Unique: 1 (0%)

No.	Label	Count
1	10-19	0
2	20-29	1
3	30-39	36
4	40-49	90
5	50-59	96
6	60-69	57
7	70-79	6

Class: Class (Nom)

Status
OK
x 0

Weka Explorer: training classifiers

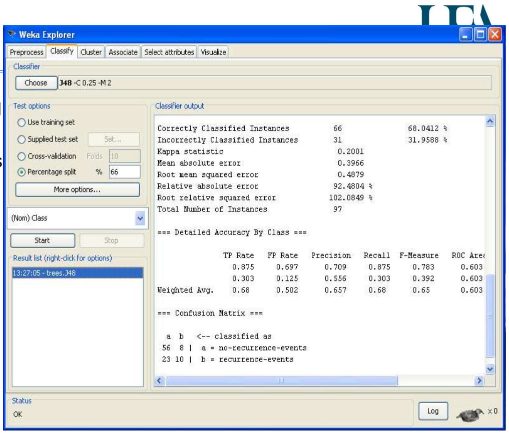
After loaded a data file, click “Classify”

- Choose a classifier,
  - Under “Classifier”: click “choose”, then a drop-down menu appears,
  - Click “trees” and select “J48” – a decision tree algorithm
- Select a test option
  - Select “percentage split”
    - with default ratio 66% for training and 34% for testing
- Click “Start” to train and test the classifier.
  - The training and testing information will be displayed in classifier output window.

Results

- Testing results:
- 97 cases used in test.

Correct:
66 (68%)
Wrong:
31 (32%)



Weka Explorer

Classifier: J48 -C 0.25 -M 2

Test options
○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 66

(Nom) Class

Result list (right-click for options)
13:27:05 - trees.J48

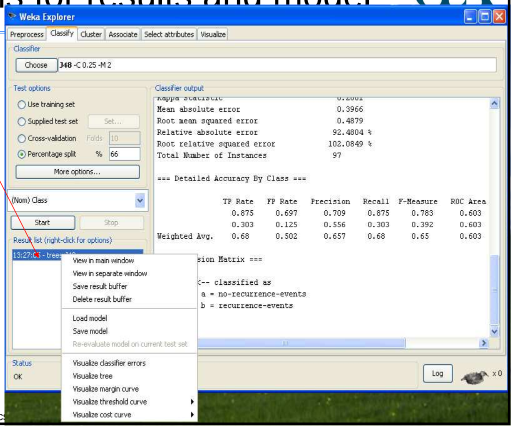
Classifier output

```
Correctly Classified Instances          66               68.0412 %
Incorrectly Classified Instances        31               31.9588 %
Kappa statistic                          0.2001
Mean absolute error                      0.3966
Root mean squared error                  0.4879
Relative absolute error                 92.4804 %
Root relative squared error            102.0849 %
Total Number of Instances               97

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
                 0.875     0.697     0.709      0.875     0.783      0.603
                 0.303     0.125     0.556      0.303     0.392      0.603
Weighted Avg.    0.68      0.502     0.657      0.68      0.65       0.603

=== Confusion Matrix ===

  a  b   <-- classified as
 56  8 |  a = no-recurrence-events
 23 10 |  b = recurrence-events
```

Status
OK
x 0

Options for results and model

- Point to result list window, and right click mouse.
- A menu will pop out to show all the options available about the model.



Weka Explorer

Classifier: J48 -C 0.25 -M 2

Test options
○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 66

(Nom) Class

Result list (right-click for options)
13:27:05 - trees.J48

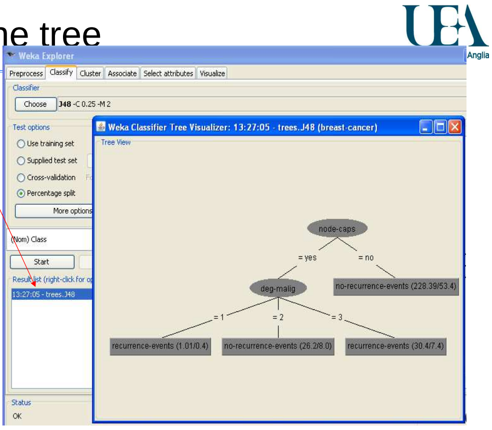
Classifier output

```
Kappa statistic                          0.2001
Mean absolute error                      0.3966
Root mean squared error                  0.4879
Relative absolute error                 92.4804 %
Root relative squared error            102.0849 %
Total Number of Instances               97

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
                 0.875     0.697     0.709      0.875     0.783      0.603
                 0.303     0.125     0.556      0.303     0.392      0.603
Weighted Avg.    0.68      0.502     0.657      0.68      0.65       0.603

sion Matrix ===

<-- classified as
 a = no-recurrence-events
 b = recurrence-events
```

Status
OK
x 0

View the tree

- Point to result list window, and right click mouse,
- Choose “**visualize tree**”, then the tree will be displayed in another window.



Weka Explorer

Weka Classifier Tree Visualizer: 13:27:05 - trees.J48 (breast-cancer)

Classifier: J48 -C 0.25 -M 2

Test options
○ Use training set
○ Supplied test set
○ Cross-validation
● Percentage split

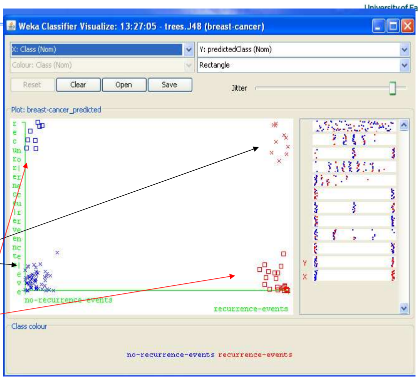
(Nom) Class

Result list (right-click for options)
13:27:05 - trees.J48

Status
OK

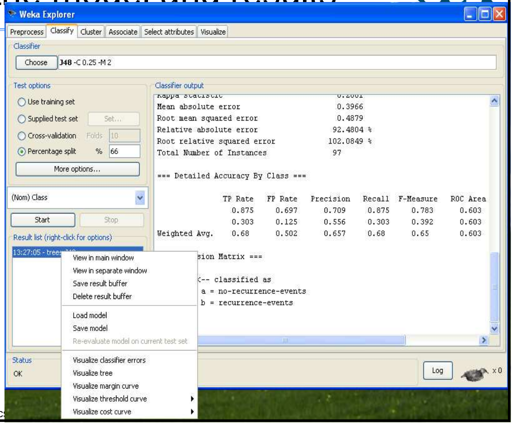
View classifier errors

- right click the result list,
- Choose "**visualize classifier error**", then a new window will be popped out to display the classifier's error.
  - Correctly predicted cases
  - Wrong cases



Save the model and results

- Right click on the result list
- Choose "save model" and "save result buffer" to save the classifier and the results to the disk folder.



Classifier: J48 -C 0.25 -M 2

○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 66

```
Kappa statistic                          0.2001
Mean absolute error                      0.3966
Root mean squared error                  0.4879
Relative absolute error                 92.4804 %
Root relative squared error            102.0849 %
Total Number of Instances               97

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
                 0.875     0.697      0.709     0.875     0.783      0.603
                 0.303     0.125      0.556     0.303     0.392      0.603
Weighted Avg.    0.68      0.502      0.657     0.68      0.65       0.603

=== Confusion Matrix ===

  <-- classified as
   a = no-recurrence-events
   b = recurrence-events
```

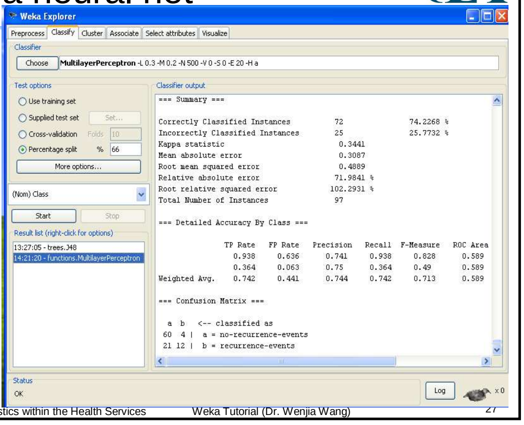
Train a neural net

Click "Choose" to select another function,

e.g. "Multilayer Perceptron" - a type of neural net.

Then click "Start" to train and test it. (note: the training may take much longer time.)

The results seem better than the tree classifier.



Classifier: MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

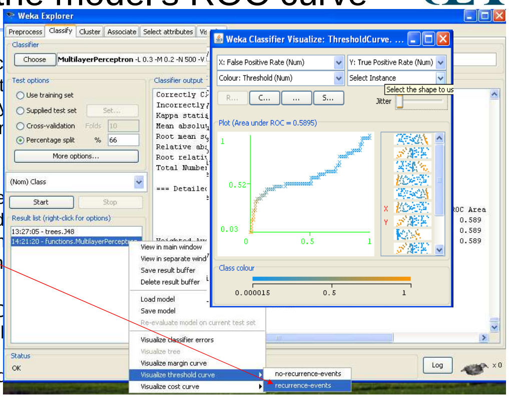
○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 66

Result list: 13:27:05 - trees.J48
14:21:20 - functions.MultilayerPerceptron

```
=== Summary ===

Correctly Classified Instances          72               74.2268 %
Incorrectly Classified Instances        25               25.7732 %
Kappa statistic                          0.3441
Mean absolute error                      0.3087
Root mean squared error                  0.4889
Relative absolute error                 71.9841 %
Root relative squared error            102.2931 %
Total Number of Instances               97

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
                 0.938     0.636      0.741     0.938     0.828      0.589
                 0.364     0.063      0.75      0.364     0.49       0.589
Weighted Avg.    0.742     0.441      0.744     0.742     0.713      0.589

=== Confusion Matrix ===

  a  b   <-- classified as
 60  4 |  a = no-recurrence-events
 21 12 |  b = recurrence-events
```

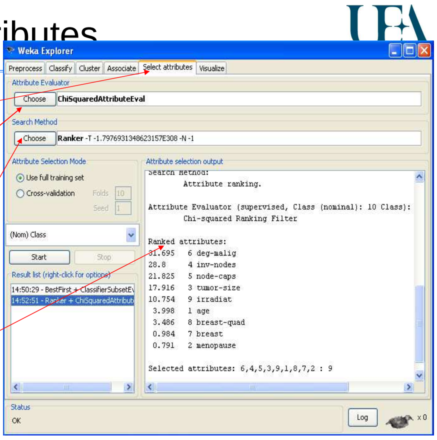
View the model's ROC curve

- Right clic the result "Multiplay rPerceptr n"
- Choose "visualize threshold curve" an "recurren events";
- The ROC curve will be displayed



Select Attributes

- Click "Select Attributes"
- Choose an "attribute evaluator"
  - e.g. chiSquare
- Choose a "Search Method"
- Then click "Start"
- The selected attributes are listed.



Attribute Evaluator: ChiSquaredAttributeEval

Search Method: Ranker -T -1.7976931348623157E308 -N -1

● Use full training set
○ Cross-validation Folds 10 Seed 1

Result list: 14:50:29 - BestFirst + ClassifierSubsetEv
14:52:51 - Ranker + ChiSquaredAttribut

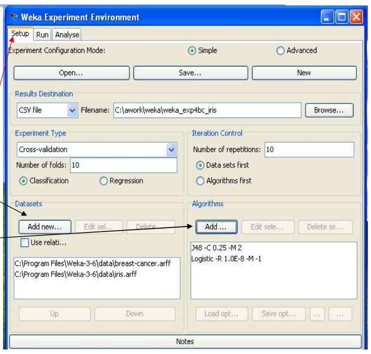
```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 10 Class):
        Chi-squared Ranking Filter

Ranked attributes:
31.695    6 deg-malig
28.8      4 inv-nodes
21.825    5 node-caps
17.916    3 tumor-size
10.754    9 irradiat
 3.998    1 age
 3.486    8 breast-quad
 0.984    7 breast
 0.791    2 menopause

Selected attributes: 6,4,5,3,9,1,8,7,2 : 9
```

4.2 Weka Experimenter

- you can use Experimenter to carry out experiments for multiple data sets using multiple methods,

e.g. classifying

- two data sets
  - Breast cancer
  - Iris
- Using two methods
  - Decision Tree: J48
  - Logistic
- The experiment is "Setup" as shown in the screenshot.
- Then click "Run"



Experiment Configuration Mode: ● Simple ○ Advanced

Results Destination: CSV file Filename: C:\awork\weka\weka_exp4bc_iris

Experiment Type: Cross-validation
Number of folds: 10
● Classification ○ Regression

Iteration Control: Number of repetitions: 10
● Data sets first
○ Algorithms first

Datasets: □ Use relati...
C:\Program Files\Weka-3-6\data\breast-cancer.arff
C:\Program Files\Weka-3-6\data\iris.arff

Algorithms: J48 -C 0.25 -M 2
Logistic -R 1.0E-8 -M -1

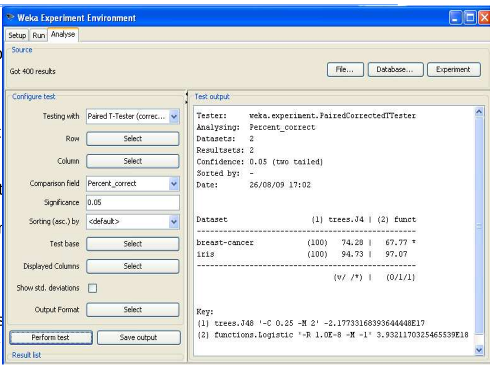
Analysis of the results

- Click “analysis” to analyse the results,

E.g.

paired t-test significance

- Click “Experiment
- Configure test: choosin appropriate test and parameters
- Click “Perform tes and the test results are listed.



Got 400 results

Testing with: Paired T-Tester (correc...

Comparison field: Percent_correct

Significance: 0.05

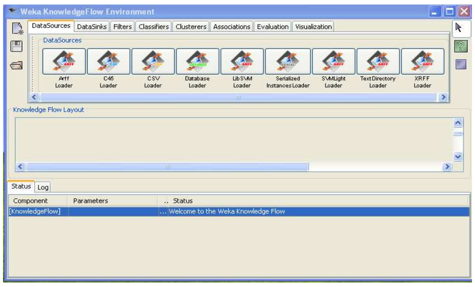
Sorting (asc.) by: <default>

Show std. deviations ☐

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  Percent_correct
Datasets:   2
Resultsets: 2
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       26/08/09 17:02


Dataset                    (1) trees.J4 | (2) funct
---------------------------------------------------
breast-cancer             (100)   74.28 |   67.77 *
iris                      (100)   94.73 |   97.07
---------------------------------------------------
                                (v/ /*) |   (0/1/1)


Key:
(1) trees.J48 '-C 0.25 -M 2' -2.17733168393644448E17
(2) functions.Logistic '-R 1.0E-8 -M -1' 3.932117032546553 9E18
```

4.3 KnowledgeFlow

- Click KnowledgeFlow on Weka GUI Chooser
- A new window opened for buidling KDD process.



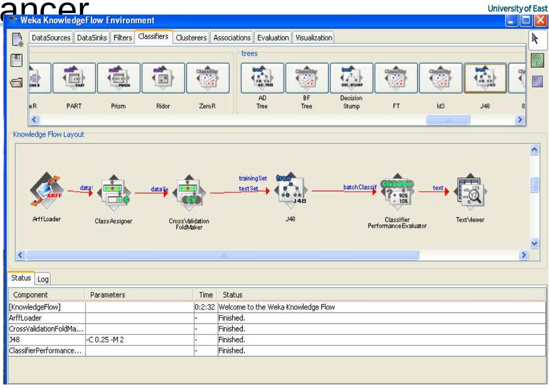
Component	Parameters	.. Status
[KnowledgeFlow]		... Welcome to the Weka Knowledge Flow

Steps for building a KDD process

Major steps for building a process

1. Adding required nodes
   1) Add nodes
   2) Add a data source node from “DataSources”
      1) Right click to configure it with a data set
   3) Add a classAssigner node from “Evaluation” and a CrossValidationFoldmaker node
   4) Add a classifier, e.g. J48, from Classifiers
   5) Add a classiferPerformanceEvaluator node from “Evaluation”
   6) Add a text viewer from “Visualisation”
2. Connect the nodes
   – Right click “DataSource” node and choose DataSet, then connect it to the ClassAssigner node,
   – do the same or similar for connecting between the other nodes.
3. Run the process (using the default setups for each node)
   – Right click DataSource node and choose “Start loading”, the process should run and “Status” window should indicate if the run is correct and completed.
4. View the results:
   – If the run is correctly completed, right click “Text Viewer” node and choose “Show results”, then another window pops out to show the results.

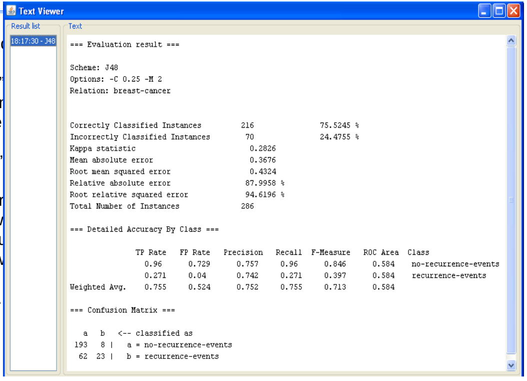
A KDD process for Breast Cancer



Component	Parameters	Time	Status
[KnowledgeFlow]		0:2:32	Welcome to the Weka Knowledge Flow
ArffLoader		-	Finished.
CrossValidationFoldMa...		-	Finished.
J48	-C 0.25 -M 2	-	Finished.
ClassifierPerformance...		-	Finished.

Results of the KDD process

- right cli “Text Viewer” node ar choose “Show results” then another window pops ou to show the results.



```
=== Evaluation result ===

Scheme: J48
Options: -C 0.25 -M 2
Relation: breast-cancer


Correctly Classified Instances         216               75.5245 %
Incorrectly Classified Instances        70               24.4755 %
Kappa statistic                          0.2826
Mean absolute error                      0.3676
Root mean squared error                  0.4324
Relative absolute error                 87.9958 %
Root relative squared error             94.6196 %
Total Number of Instances              286

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.96      0.729      0.757     0.96      0.846      0.584    no-recurrence-events
                 0.271     0.04       0.742     0.271     0.397      0.584    recurrence-events
Weighted Avg.    0.755     0.524      0.752     0.755     0.713      0.584

=== Confusion Matrix ===

   a   b   <-- classified as
 193   8 |   a = no-recurrence-events
  62  23 |   b = recurrence-events
```

5. Weka Tutorial Summary

Weka is open source data mining software that offers

- Some GUI interfaces for data mining
  – Explorer
  – Experimenter
  – KnowledgeFlow
- Many functions and tools that include
  – Methods for **classification:**
    decision trees, rule learners, naive Bayes, decision tables, locally weighted regression, SVMs, instance-based learners, logistic regression, multi-layer perceptron
  – methods for **regression/prediction:**
    linear regression, model tree generators, locally weighted regression, instance-based learners, decision tables, multi-layer perceptron
  – **Ensemble schemes**
    - Bagging, boosting, stacking, RandomFrest
  – Methods for **clustering**:
    - K-means, EM and Cobweb
  – Methods for feature selection