

WEKA

Waikato Environment for Knowledge Analysis

Attribute Selection

Prof. Pietro Ducange



Attribute Selection (1)

- Attribute selection involves searching through all possible combinations of attributes in the data to find which subset of attributes works best for prediction.
- Attribute selection consists basically of two different types of algorithms:
 - **evaluator** – determines the merit of single attributes or subsets of attributes
 - **search algorithm** – the search heuristic

Attribute Selection (2)

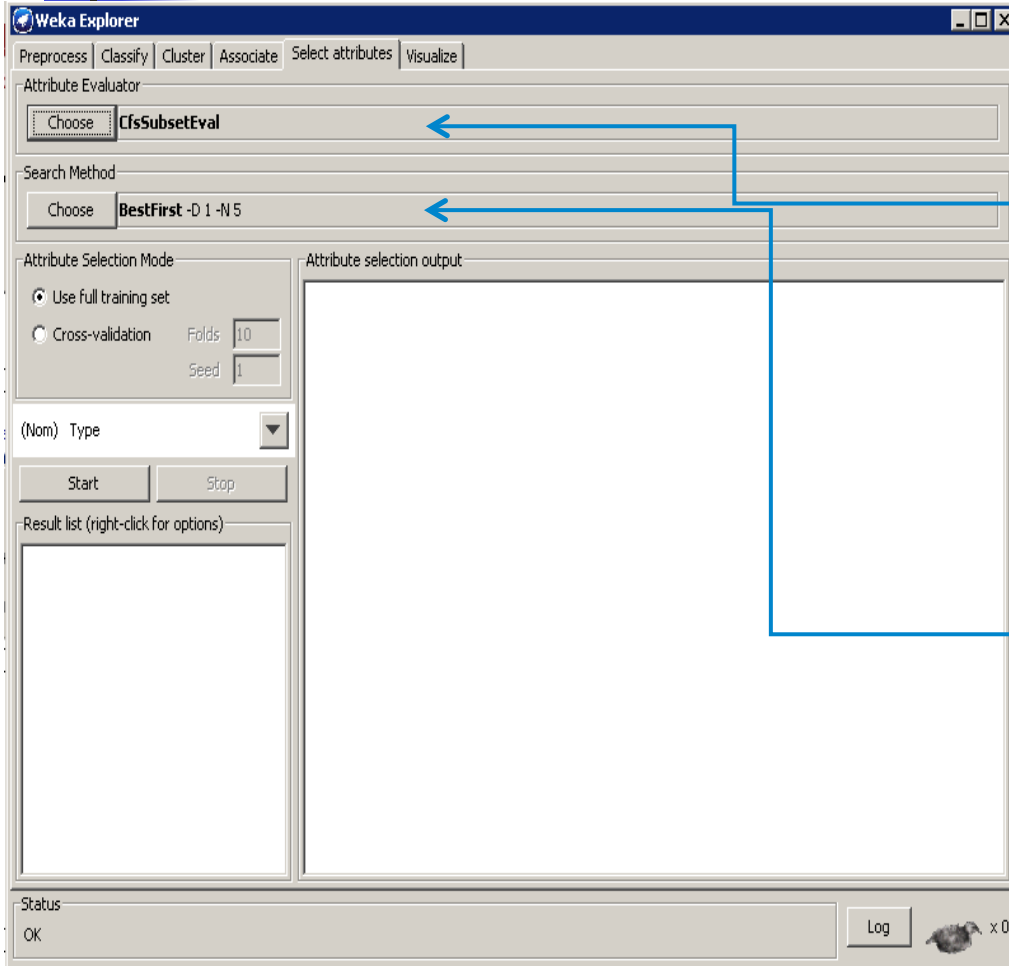
The image displays two screenshots of the Weka Explorer software interface, illustrating the process of selecting attribute selection methods.

Left Screenshot: The 'Attribute Evaluator' window is open, showing a tree view of available methods under the 'weka' folder. The 'attributeSelection' sub-folder is expanded, and 'CfsSubsetEval' is selected. Other methods listed include ChiSquaredAttributeEval, ClassifierSubsetEval, ConsistencySubsetEval, CostSensitiveAttributeEval, CostSensitiveSubsetEval, FilteredAttributeEval, FilteredSubsetEval, GainRatioAttributeEval, InfoGainAttributeEval, LatentSemanticAnalysis, OneRAttributeEval, PrincipalComponents, ReliefFAttributeEval, SVMAttributeEval, SymmetricalUncertAttributeEval, and WrapperSubsetEval. Buttons for 'Filter...', 'Remove filter', and 'Close' are visible at the bottom of the list.

Right Screenshot: The 'Search Method' window is open, showing a tree view of search methods under the 'weka' folder. The 'attributeSelection' sub-folder is expanded, and 'BestFirst' is selected. Other methods listed include ExhaustiveSearch, GeneticSearch, GreedyStepwise, LinearForwardSelection, RaceSearch, RandomSearch, Ranker, RankSearch, ScatterSearchV1, and SubsetSizeForwardSelection. A 'Close' button is visible at the bottom of the list.

Both screenshots show the 'Attribute Evaluator' window with a 'Choose' button and the selected method name ('CfsSubsetEval' in the left window and 'BestFirst' in the right window) displayed in the main area.

Attribute Selection: First Example (1)



Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them

Searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility. Forward or Backward Search can be selected

Attribute Selection: First Example (2)

=== Attribute Selection on all input data ===

Search Method:

Best first.

Start set: no attributes

Search direction: forward

Stale search after 5 node expansions

Total number of subsets evaluated: 12

Merit of best subset found: 0.887

Attribute Subset Evaluator (supervised, Class (nominal): 5 class):

CFS Subset Evaluator

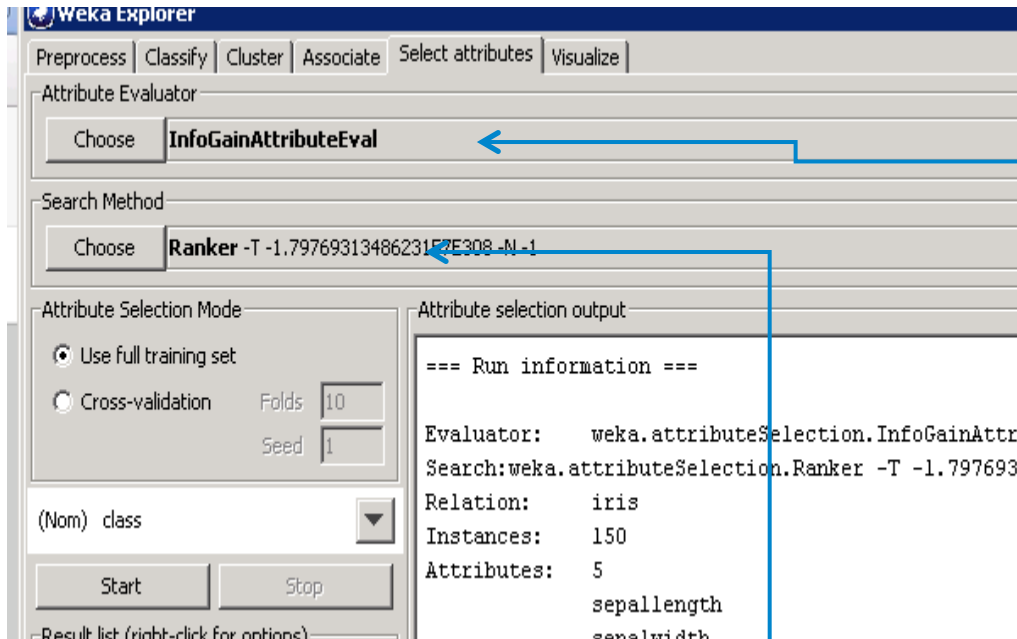
Including locally predictive attributes

Selected attributes: 3,4 : 2

petallength

petalwidth

Attribute Selection: Second Example (1)



Evaluates the worth of an attribute by measuring the information gain with respect to the class

Ranks attributes by their individual evaluations. A selection threshold can be fixed

Attribute Selection: Second Example (2)

=== Attribute Selection on all input data ===

Search Method:

Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 5 class):
Information Gain Ranking Filter

Ranked attributes:

1.418 3 petallength

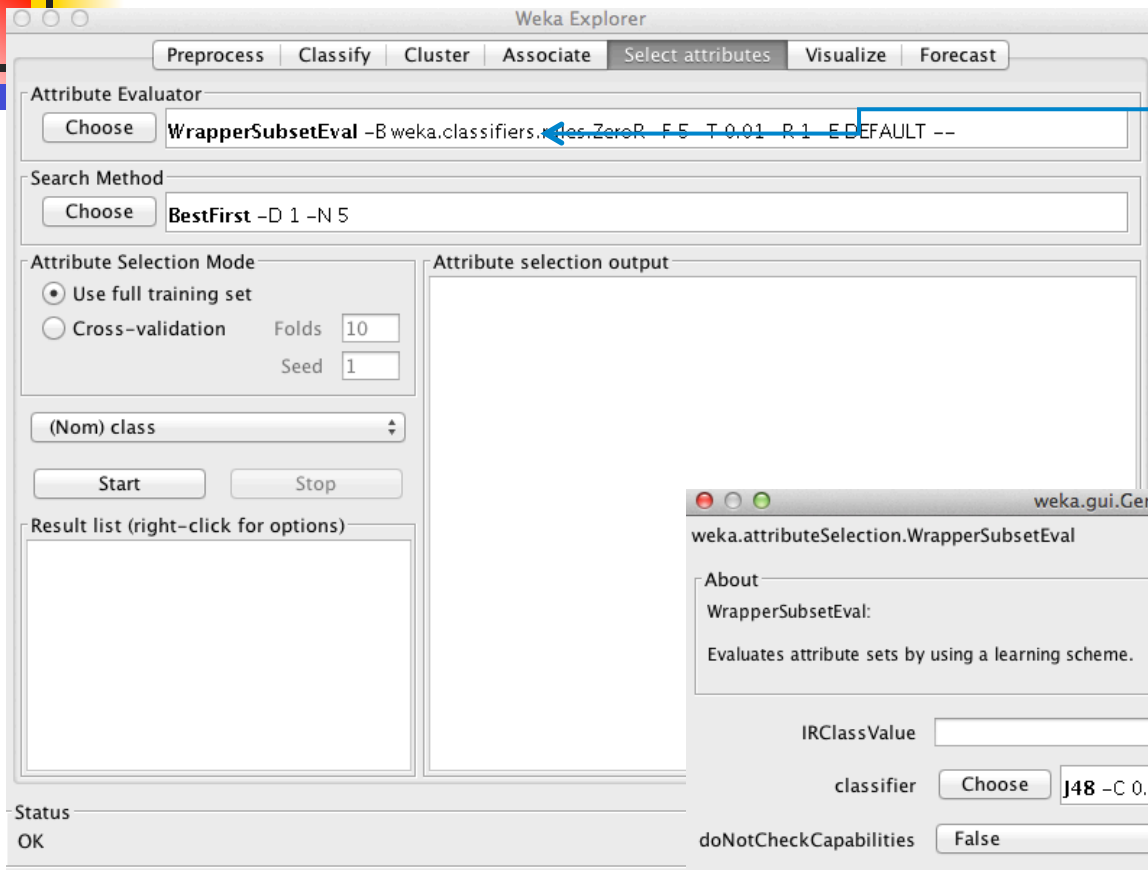
1.378 4 petalwidth

0.698 1 sepallength

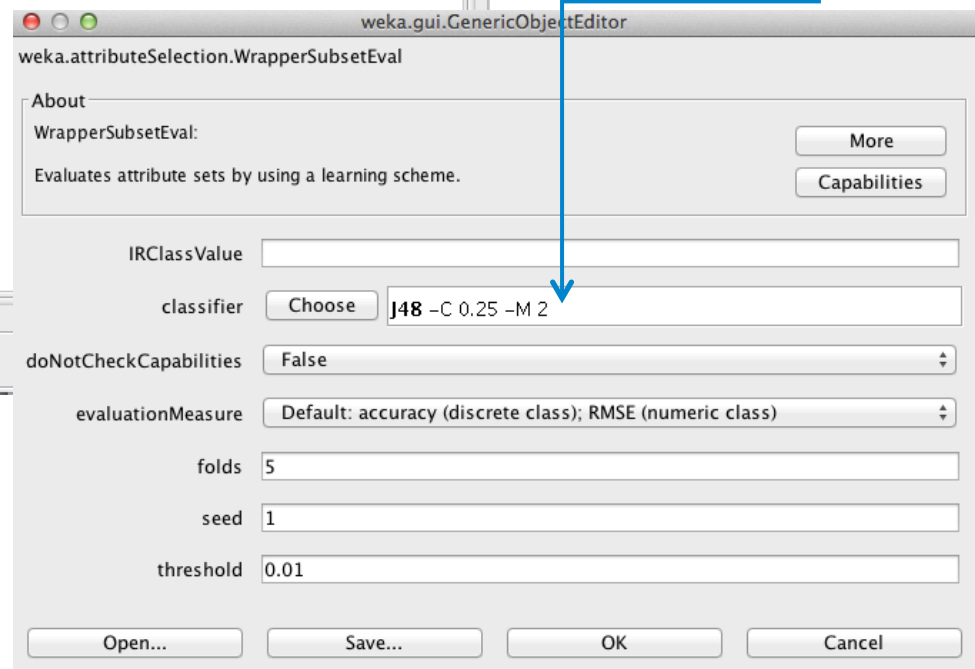
0.376 2 sepalwidth

Selected attributes: 3,4,1,2 : 4

Attribute Selection: Wrapper Method (2)



Evaluates the worth of a set of attributes by using a specific classifier



Attribute Selection: Wrapper Method (2)

=== Attribute Selection on all input data ===

Search Method:

Best first.

Start set: no attributes

Search direction: forward

Stale search after 5 node expansions

Total number of subsets evaluated: 11

Merit of best subset found: 0.947

Attribute Subset Evaluator (supervised, Class (nominal): 5 class):

Wrapper Subset Evaluator

Learning scheme: weka.classifiers.trees.J48

Scheme options: -C 0.25 -M 2

Subset evaluation: classification accuracy

Number of folds for accuracy estimation: 5

Selected attributes: 4 : 1

petalwidth

Attribute Selection as a Filter

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter

weka

- filters
 - AllFilter
 - MultiFilter
 - supervised
 - attribute
 - AddClassification
 - AttributeSelection**
 - ClassOrder
 - Discretize
 - NominalToBinary
 - PLSFilter
 - instance
 - unsupervised

AttributeSelection -E "weka.attributeSelection.CfsSubsetEval" -S "weka.attributeSelection.Be: Apply

Selected attribute

Name: sepal.length Type: Numeric
 Missing: 0 (0%) Distinct: 35 Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: class (Nom) Visualize All

16 30 34 28 25 10 7

4.3 6.1 7.9

Filter... Remove filter Close

Log x 0



Attribute Selection as a Filter (setting parameters)

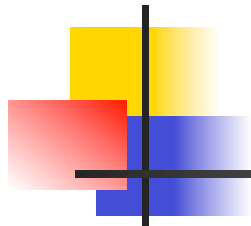
The screenshot displays the Weka Explorer interface with the 'Filter' tab active. The filter selected is 'AttributeSelection' with the command: `-E "weka.attributeSelection.CfsSubsetEval" -S "weka.attributeSelection.BestFirst -D 1 -N 5"`. The current relation is 'iris' with 150 instances. The attributes list includes: 1 sepallength, 2 sepalwidth, 3 petallength, 4 petalwidth, and 5 class.

Two dialog boxes are open for configuration:

- weka.filters.supervised.attribute.AttributeSelection**:
 - About: A supervised attribute filter that can be used to select attributes.
 - evaluator: `CfsSubsetEval`
 - search: `BestFirst -D 1 -N 5`
- weka.attributeSelection.CfsSubsetEval**:
 - About: CfsSubsetEval : Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.
 - locallyPredictive: `True`
 - missingSeparate: `False`



Attribute Selection as a Filter (results)



The screenshot shows the Weka software interface with the 'Attribute Selection' filter applied. The 'Current relation' is 'iris-weka.filters.supervised.attribute.AttributeSelect...' with 150 instances and 3 attributes. The 'Attributes' list shows 'petalength', 'petalwidth', and 'class'. The 'Selected attribute' is 'petalength', which is numeric with 43 distinct values and 10 unique values (7%). A histogram shows the distribution of 'petalength' values across three classes: class 1 (blue), class 2 (red), and class 3 (cyan). The histogram bars are labeled with their respective counts: 50 for class 1, 34 for class 2, and 47 for class 3. The x-axis represents 'petalength' values from 1 to 6.9, and the y-axis represents the count of instances.

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Now we can switch to the classify module and perform a cross validation analysis....

...Is it a correct way to act?



Classification and Attribute Selection (1)

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier: **AttributeSelectedClassifier** -E "weka.attributeSelection.CfsSubsetEval" -S "weka.attributeSelection.BestFirst -D 1 -N 5" -W weka.classifiers.trees.J48 -- -C 0.25 -M 2

Test options:

- Use training set
- Supplied test set (Set...)
- Cross-validation (Folds: 10)
- Percentage split (%: 66)

More options...

(Nom) class: [Dropdown]

Start | Stop

Result list (right-click for options):

- 12:28:51 - lazy.IBk
- 12:29:08 - lazy.IBk
- 12:31:38 - bayes.NaiveBayes**

Classifier output:

```

sepalwidth
mean
std.
weight
precision

petalwidth
mean
std.
weight
precision

petalwidth
mean
std.
weight
precision

Time taken to build model: 0 seconds
    
```

Dimensionality of training and test data is reduced by attribute selection before being passed on to a classifier

weka.gui.GenericObjectEditor

weka.classifiers.meta.AttributeSelectedClassifier

About

Dimensionality of training and test data is reduced by attribute selection before being passed on to a classifier.

More | Capabilities

classifier: Choose **J48 -C 0.25 -M 2**

debug: False

evaluator: Choose **CfsSubsetEval**

search: Choose **BestFirst -D 1 -N 5**

Open... | Save... | OK | Cancel



Classification and Attribute Selection (2)

Selected attributes: 3,4 : 2
petallength
petalwidth

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances	49	96.0784 %
Incorrectly Classified Instances	2	3.9216 %
Kappa statistic	0.9408	
Mean absolute error	0.0396	
Root mean squared error	0.1579	
Relative absolute error	8.8979 %	
Root relative squared error	33.4091 %	
Total Number of Instances	51	

Command Line Attribute Selection: An Example

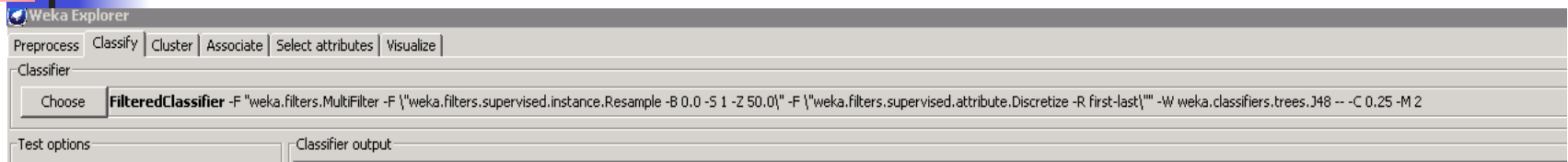
- **Generating new training and test files**

```
java weka.filters.supervised.attribute.AttributeSelection \  
-E "weka.attributeSelection.CfsSubsetEval " \  
-S "weka.attributeSelection.BestFirst -D 1 -N 5" \  
-b \  
-i <Training.arff> \  
-o <TrainingSel.arff> \  
-r <Test.arff> \  
-s <TestSel.arff>
```

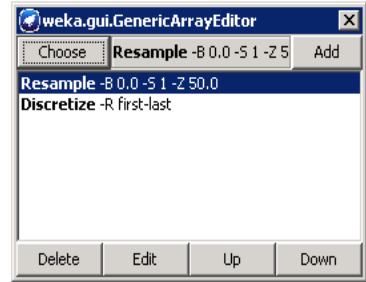
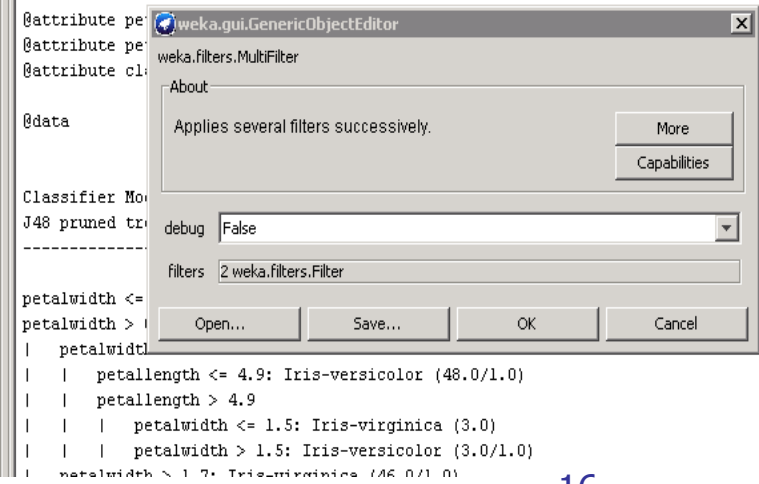
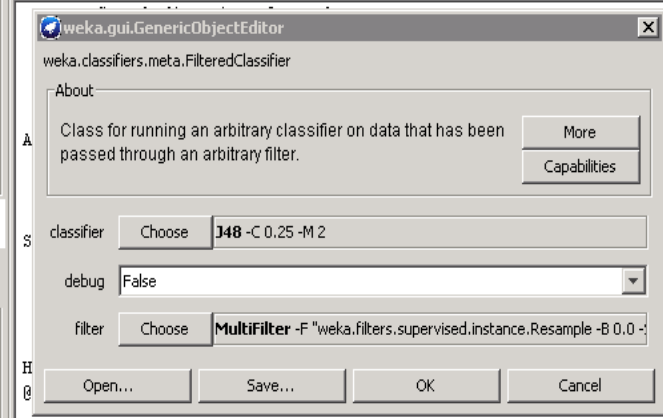
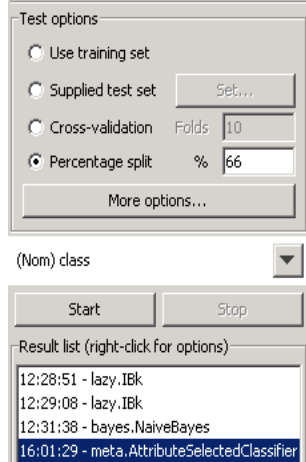
- **Generating and testing a classifier**

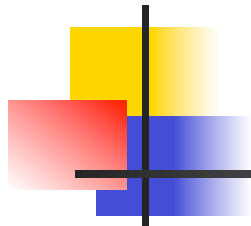
```
java weka.classifiers.trees.J48 -t TrainingSel.arff -T TestSel.arff
```

Classification and Filtering Data (1)



The structure of the filter is based exclusively on the training data and test instances will be processed by the filter without changing their structure





Exercise II

Perform the classification by using the three different meta classifiers (select a classification algorithm and three different attribute selection methods) and the Pima Diabets dataset with a 5-fold cross validation.

Which is the best attribute selection method?

