WEKA
Werkzeuge

Werkzeug

Zeug zum Werken

- Zeug
- Dinge, Kram, Arbeitsgeräte

- Werken
- tätig sein, schaffen

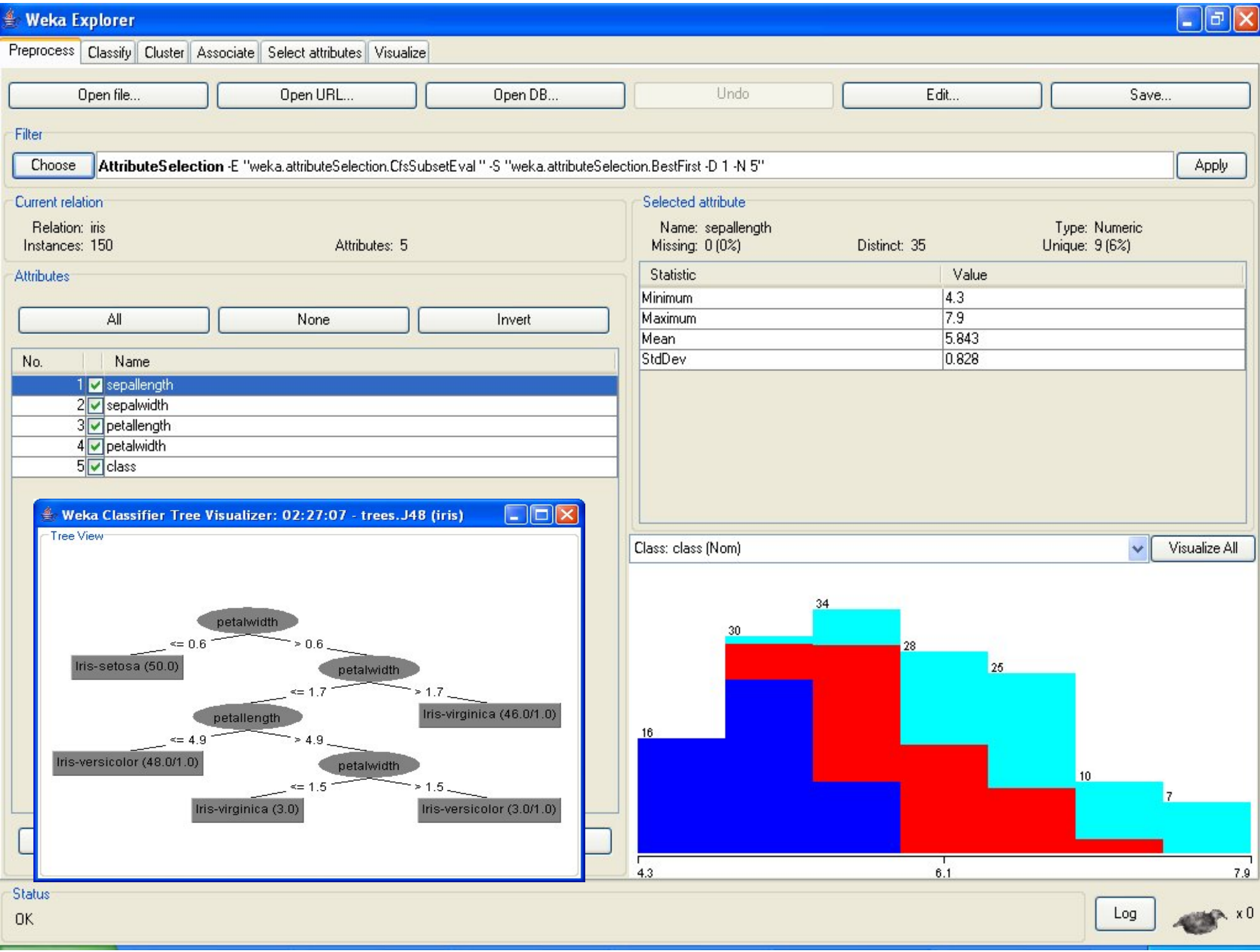
d.h. Dinge die zum arbeiten benutzt werden

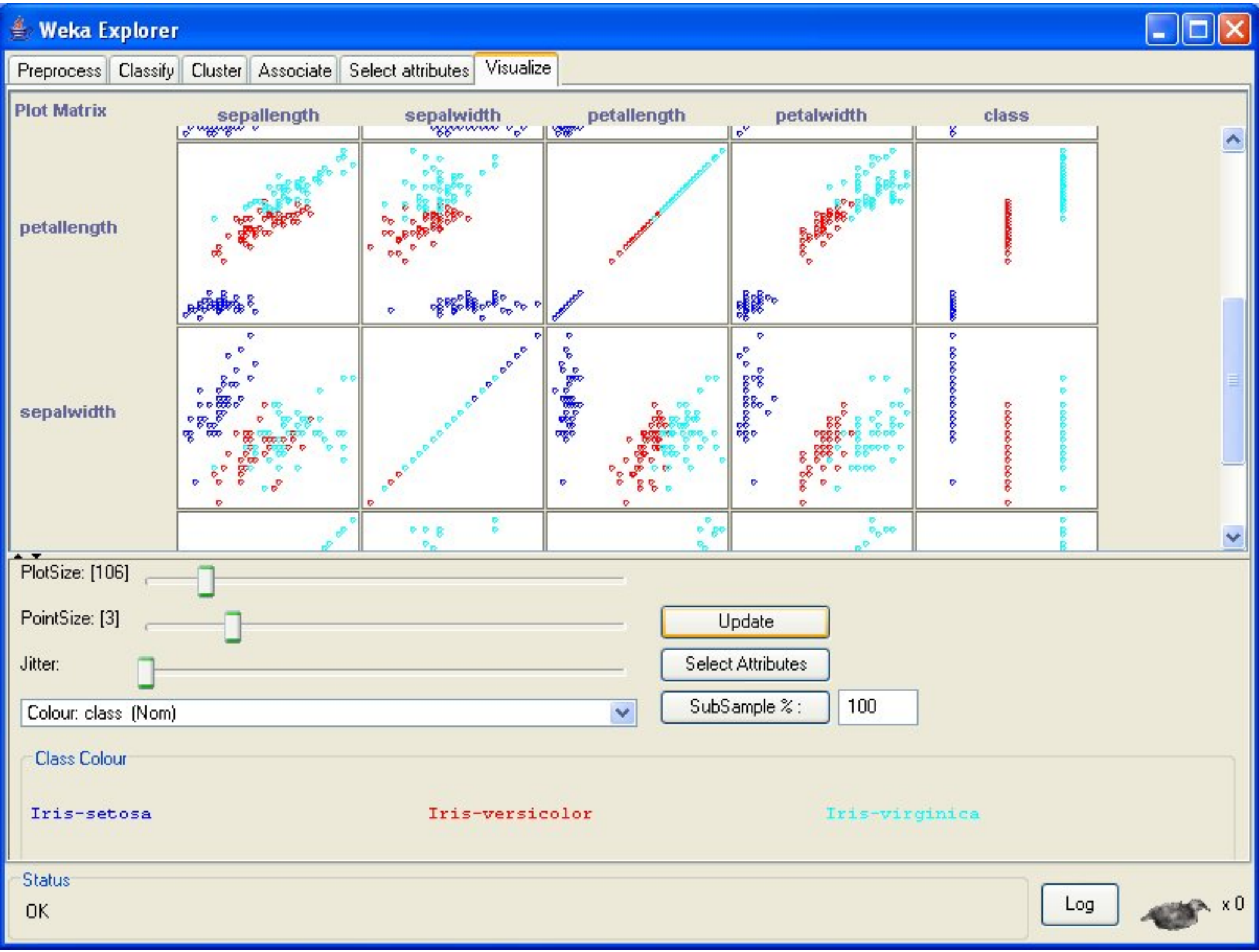
WEKA

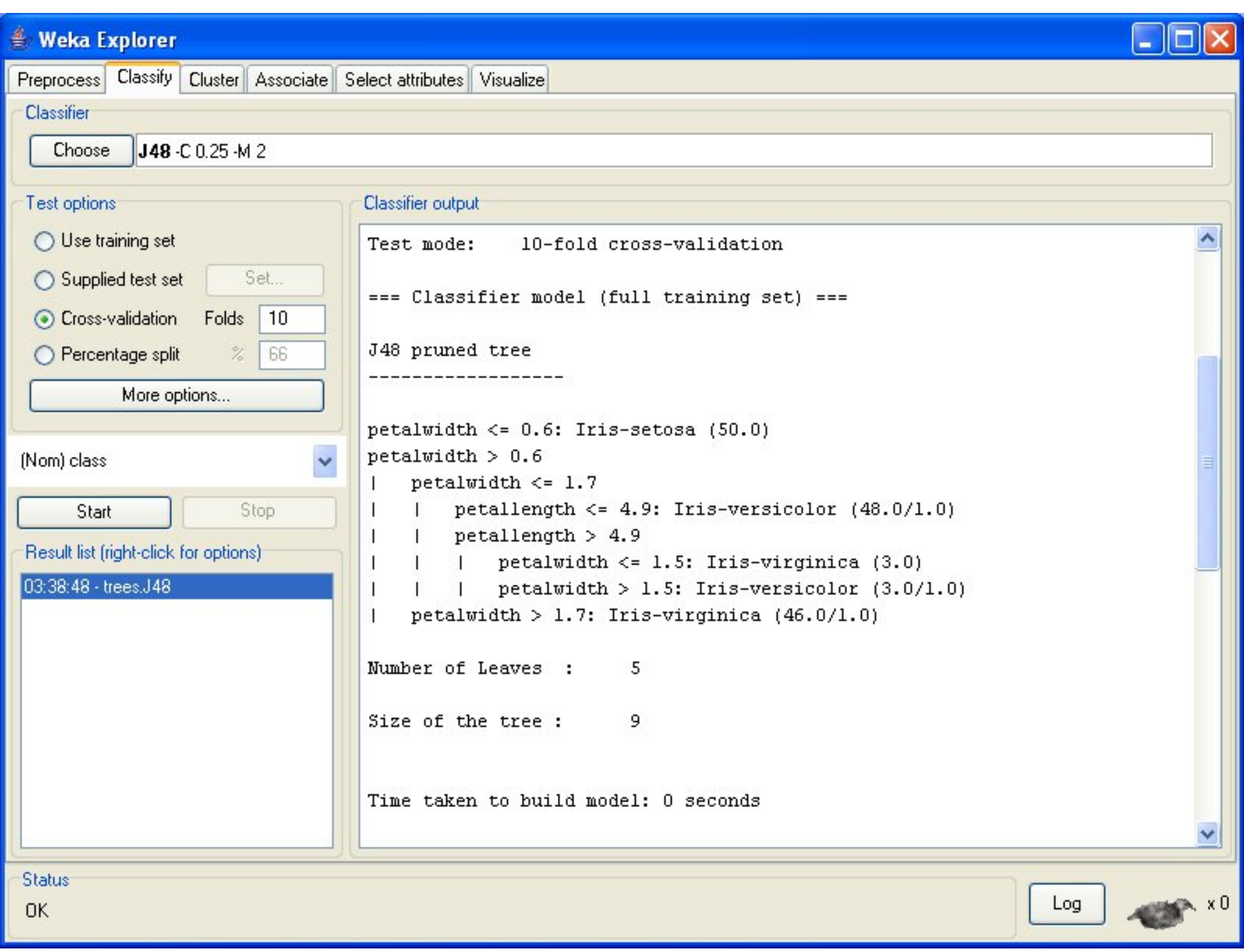
- Waikato Environment for Knowledge Analysis
- Werkzeug zum Data-Mining
- GNU-Lizenz
- geschrieben in JAVA

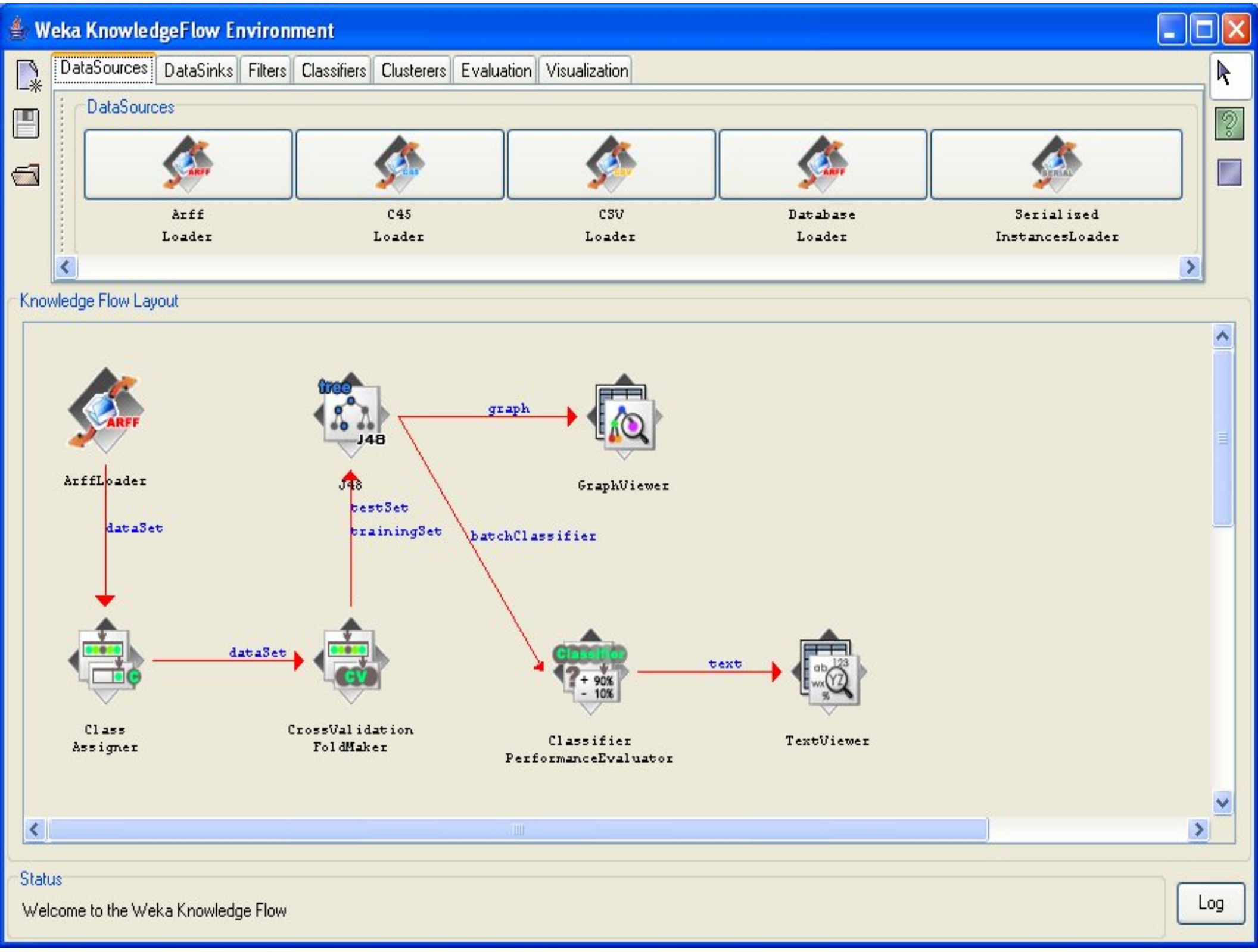
Userinterfaces

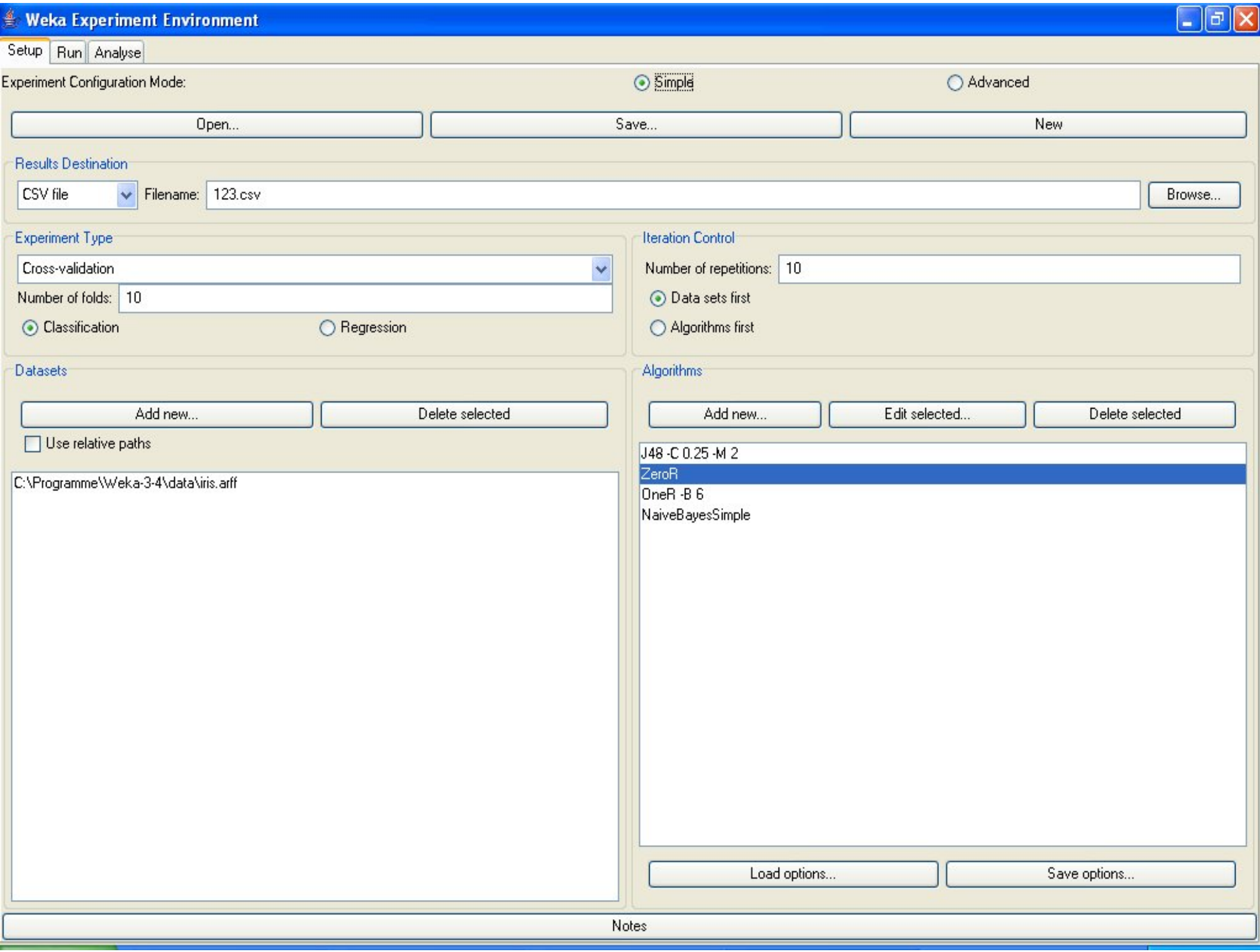
- Explorer
- Experimenter
- Knowledge Flow
- Simple Command Line


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Filter
Choose
AttributeSelection -E "weka.attributeSelection.CfsSubsetEval " -S "weka.attributeSelection.BestFirst -D 1 -N 5"
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Selected attribute
Name: sepallength
Missing: 0 (0%)
Distinct: 35
Type: Numeric
Unique: 9 (6%)
Statistic
Value
Minimum
4.3
Maximum
7.9
Mean
5.843
StdDev
0.828
Attributes
All
None
Invert
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Weka Classifier Tree Visualizer: 02:27:07 - trees.J48 (iris)
Tree View
petalwidth
<= 0.6
> 0.6
Iris-setosa (50.0)
petalwidth
<= 1.7
> 1.7
petallength
Iris-virginica (46.0/1.0)
<= 4.9
> 4.9
Iris-versicolor (48.0/1.0)
petalwidth
<= 1.5
> 1.5
Iris-virginica (3.0)
Iris-versicolor (3.0/1.0)
Class: class (Nom)
Visualize All
34
30
28
25
16
10
7
4.3
6.1
7.9
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
sepallength
sepalwidth
petallength
petalwidth
class
petallength
sepalwidth
PlotSize: [106]
PointSize: [3]
Jitter:
Colour: class (Nom)
Update
Select Attributes
SubSample % :
100
Class Colour
Iris-setosa
Iris-versicolor
Iris-virginica
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
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
Start
Stop
Result list (right-click for options)
03:38:48 - trees.J48
Classifier output
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
----
petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
| petalwidth <= 1.7
| | petallength <= 4.9: Iris-versicolor (48.0/1.0)
| | petallength > 4.9
| | | petalwidth <= 1.5: Iris-virginica (3.0)
| | | petalwidth > 1.5: Iris-versicolor (3.0/1.0)
| petalwidth > 1.7: Iris-virginica (46.0/1.0)
Number of Leaves : 5
Size of the tree : 9
Time taken to build model: 0 seconds
Status
OK
Log
x 0


Weka KnowledgeFlow Environment
DataSources
DataSinks
Filters
Classifiers
Clusterers
Evaluation
Visualization
DataSources
Arff Loader
C45 Loader
CSV Loader
Database Loader
Serialized InstancesLoader
Knowledge Flow Layout
ArffLoader
J48
GraphViewer
graph
dataSet
testSet
trainingSet
batchClassifier
Class Assigner
dataSet
CrossValidation FoldMaker
Classifier PerformanceEvaluator
text
TextViewer
Status
Welcome to the Weka Knowledge Flow
Log


Weka Experiment Environment
Setup
Run
Analyse
Experiment Configuration Mode:
Simple
Advanced
Open...
Save...
New
Results Destination
CSV file
Filename:
123.csv
Browse...
Experiment Type
Cross-validation
Number of folds:
10
Classification
Regression
Iteration Control
Number of repetitions:
10
Data sets first
Algorithms first
Datasets
Add new...
Delete selected
Use relative paths
C:\Programme\Weka-3-4\data\iris.arff
Algorithms
Add new...
Edit selected...
Delete selected
J48 -C 0.25 -M 2
ZeroR
OneR -B 6
NaiveBayesSimple
Load options...
Save options...
Notes

Vergleich Interfaces

- Explorer: schnell, einfach
- Knowledge Flow: Anschaulich, sequenzielles laden der Daten
- Experimenter: Arbeiten im Cluster

Datenvorbereitung

- Qualität der Daten
- Entsprechendes Format
- Auswahl der betrachteten Instanzen und Eigenschaften

Datenqualität

- Aggregationsgrad
- Ausreißer
- Fehler

ARFF

- Attribut relation file format
- Orientierung an relationalen Datenbanken
  - Benennung einer Relation
  - Benennung der Attribute, Definierung der Datentypen der Attribute
  - Aufzählung der Instanzen

Lernen

- Das „Konzept“ wird gesucht, beschrieben
- Möglich z.B mit: Klassifikation, Clustering

Entscheidungsbäume

- Erstellung durch Divide&Conquer
- Müssen normalerweise gekappt werden
- Umwandlung in Klassifizierungsregeln
- Umgang mit Datenlücken

Entscheidungsbäume Datenlücken

- Ersetzen durch Default-Wert
- Teilen der Datensätze
- Ignorieren der Datensätze mit fehlenden Werten

Entscheidungsbäume Kappung

- Postpruning / Prepruning
- Subtreereplacement: es wird beim Blatt begonnen, an jedem Knoten geprüft ob geschnitten werden soll
- Subtreeraising

Clustering

- K-means- Methode
- Inkrementelles Verfahren
- Wahrscheinlichkeitsbasiertes clustering
- EM-Algorithmus

k-means

- Setzen von k Punkten
- Zuordnung aller Datenpunkte zum nächsten der k Punkte
- Neusetzung der k Punkte am Centroid der zu ihnen gehörigen
- Wenn k unbekannt MDL mit Kosten für jeden neuen Cluster anwenden

Inkrementelles clustering

- Die Instanzen werden inkrementell als Blätter in einen Baum eingefügt
- Eine große Neustrukturierung wird durchgeführt in Abhängigkeit von der category utility
- Category utility ist ein Maß für die Qualität der Cluster im Subbaum

Wahrscheinlichkeitsbasiertes clustering

- Jedem Cluster wird eine Verteilung und eine Wahrscheinlichkeit unterstellt
- Ergibt Wahrscheinlichkeit für eine Instanz zu einem Cluster zu gehören
- Bleibt Unbekanntheit der Verteilungen

EM-Algorithmus

- Kombination aus k-mean und Wahrscheinlichkeitsbasiertem Clustering
- Initial werden die Parameter für W. Geraten
- Berechnung der Clusteranteile
- Wiederholung