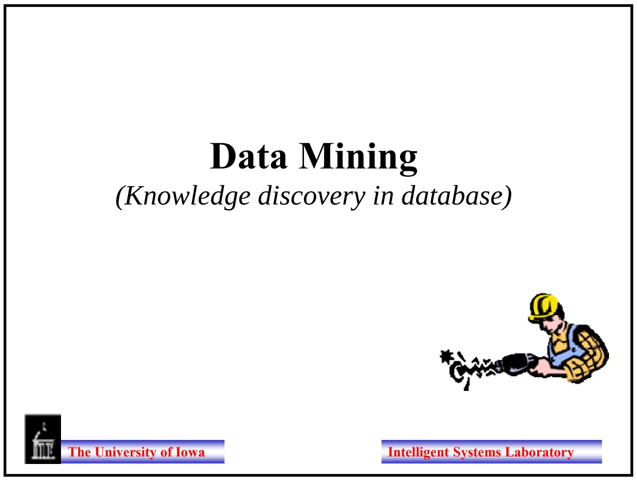
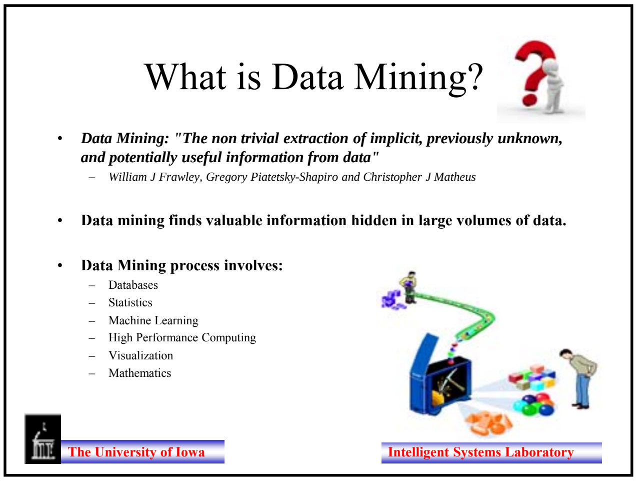


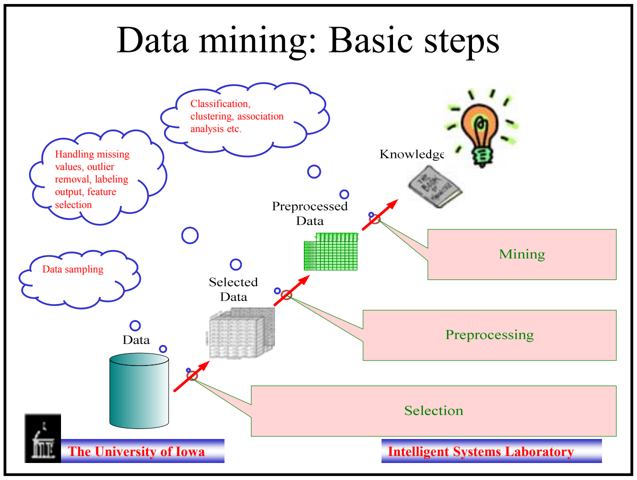
Data Mining

(Knowledge discovery in database)



What is Data Mining?

- ***Data Mining: "The non trivial extraction of implicit, previously unknown, and potentially useful information from data"***
  - *William J Frawley, Gregory Piatetsky-Shapiro and Christopher J Matheus*
- **Data mining finds valuable information hidden in large volumes of data.**
- **Data Mining process involves:**
  - Databases
  - Statistics
  - Machine Learning
  - High Performance Computing
  - Visualization
  - Mathematics



Data mining: Basic steps

Classification, clustering, association analysis etc.

Handling missing values, outlier removal, labeling output, feature selection

Data sampling

Knowledge

Preprocessed Data

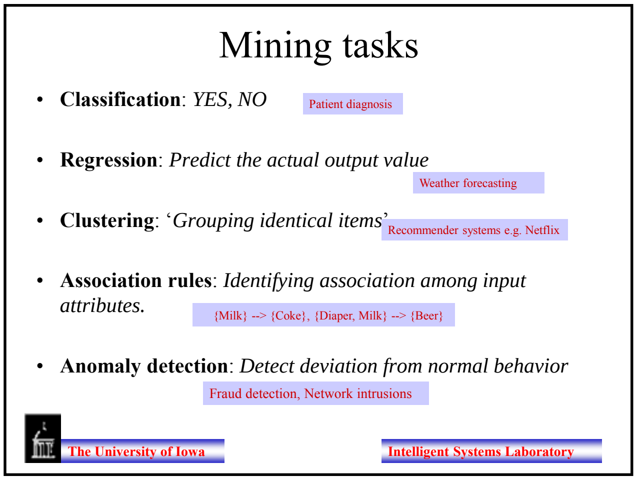
Selected Data

Data

Mining

Preprocessing

Selection



Mining tasks

- **Classification**: *YES, NO* Patient diagnosis
- **Regression**: *Predict the actual output value* Weather forecasting
- **Clustering**: *'Grouping identical items'* Recommender systems e.g. Netflix
- **Association rules**: *Identifying association among input attributes.* {Milk} --> {Coke}, {Diaper, Milk} --> {Beer}
- **Anomaly detection**: *Detect deviation from normal behavior* Fraud detection, Network intrusions

Data Mining Software

- **Enterprise-level: (US $10,000 and more)**
  - Fair Isaac, IBM, Insightful, KXEN, Oracle, SAS, and SPSS
- **Department-level: (from $1,000 to $9,999)**
  - Angoss, CART/MARS/TreeNet/Random Forests, Equbits, GhostMiner, Gornik, Mineset, MATLAB, Megaputer, Microsoft SQL Server, Statsoft Statistica, ThinkAnalytics
- **Personal-level: (from $1 to $999):**
  - Excel, See5, MATLAB
- **Free:**
  - C4.5, R, **Weka**, Xelopes

Data Mining: WEKA

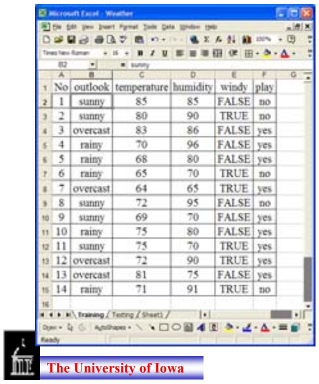
Outline

- Data preparation
- Preprocessing and “arff” files
- Filters, classifiers, and visualization
- Attribute selection
- Training and testing
- Quality measurements
- Interpretation of results

Data Mining Procedures

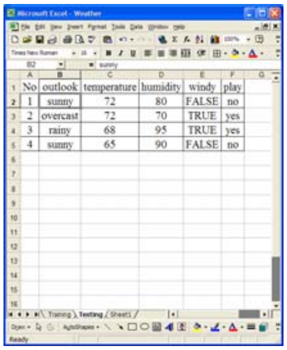
- Prepare the data into desired formats
- Preprocess the data if necessary
- Select different algorithms based on application or domain expertise
- Evaluate the results and repeat experiments again if necessary

Data Sets



Microsoft Excel - Weather
File Edit View Insert Format Tools Data Window Help
Training / Testing / Sheet1
Ready

No	outlook	temperature	humidity	windy	play
1	sunny	85	85	FALSE	no
2	sunny	80	90	TRUE	no
3	overcast	83	86	FALSE	yes
4	rainy	70	96	FALSE	yes
5	rainy	68	80	FALSE	yes
6	rainy	65	70	TRUE	no
7	overcast	64	65	TRUE	yes
8	sunny	72	95	FALSE	no
9	sunny	69	70	FALSE	yes
10	rainy	75	80	FALSE	yes
11	sunny	75	70	TRUE	yes
12	overcast	72	90	TRUE	yes
13	overcast	81	75	FALSE	yes
14	rainy	71	91	TRUE	no



Microsoft Excel - Weather
File Edit View Insert Format Tools Data Window Help
Training / Testing / Sheet1
Ready

No	outlook	temperature	humidity	windy	play
1	sunny	72	80	FALSE	no
2	overcast	72	70	TRUE	yes
3	rainy	68	95	TRUE	yes
4	sunny	65	90	FALSE	no

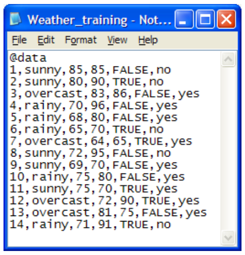
Arff file

Header of arff file

@relation	weather	
@attribute	No	real
@attribute	outlook	{sunny,overcast,rainy}
@attribute	temperature	real
@attribute	humidity	real
@attribute	windy	{TRUE,FALSE}
@attribute	play	{yes,no}

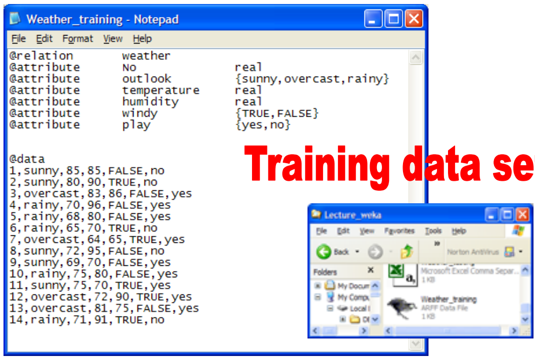
Arff file

Data in arff file



```
@data
1,sunny,85,85,FALSE,no
2,sunny,80,90,TRUE,no
3,overcast,83,86,FALSE,yes
4,rainy,70,96,FALSE,yes
5,rainy,68,80,FALSE,yes
6,rainy,65,70,TRUE,no
7,overcast,64,65,TRUE,yes
8,sunny,72,95,FALSE,no
9,sunny,69,70,FALSE,yes
10,rainy,75,80,FALSE,yes
11,sunny,75,70,TRUE,yes
12,overcast,72,90,TRUE,yes
13,overcast,81,75,FALSE,yes
14,rainy,71,91,TRUE,no
```

Arff file

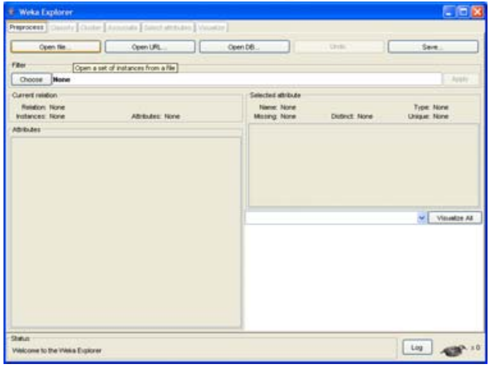


Weather_training - Notepad
File Edit Format View Help
Lecture_weka
File Edit View Favorites Tools Help
Back
Norton AntiVirus
Folders
My Docum
My Compu
Local
Microsoft Excel Comma Separ...
1 KB
Weather_training
ARFF Data File
1 KB

```
@relation       weather
@attribute      No              real
@attribute      outlook         {sunny,overcast,rainy}
@attribute      temperature     real
@attribute      humidity        real
@attribute      windy           {TRUE,FALSE}
@attribute      play            {yes,no}


@data
1,sunny,85,85,FALSE,no
2,sunny,80,90,TRUE,no
3,overcast,83,86,FALSE,yes
4,rainy,70,96,FALSE,yes
5,rainy,68,80,FALSE,yes
6,rainy,65,70,TRUE,no
7,overcast,64,65,TRUE,yes
8,sunny,72,95,FALSE,no
9,sunny,69,70,FALSE,yes
10,rainy,75,80,FALSE,yes
11,sunny,75,70,TRUE,yes
12,overcast,72,90,TRUE,yes
13,overcast,81,75,FALSE,yes
14,rainy,71,91,TRUE,no
```

Training data set

WEKA Explorer



Weka Explorer

Filter

Open a set of instances from a file

None

Current relation

Relation: None
Instances: None Attributes: None

Attributes

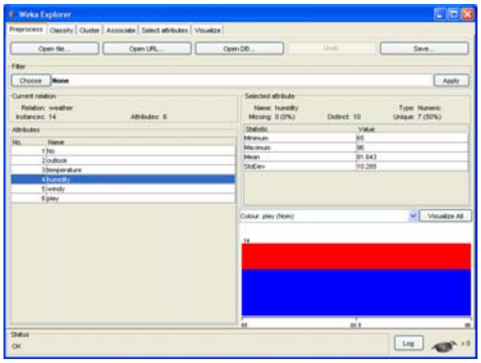
Selected attribute

Name: None Type: None
Missing: None Distinct: None Unique: None

Status

Welcome to the Weka Explorer

WEKA



Weka Explorer

Filter

None

Current relation

Relation: weather
Instances: 14 Attributes: 6

Attributes

No.	Name
1	No
2	outlook
3	temperature
4	humidity
5	windy
6	play

Selected attribute

Name: humidity Type: Numeric
Missing: 0 (0%) Distinct: 10 Unique: 7 (50%)

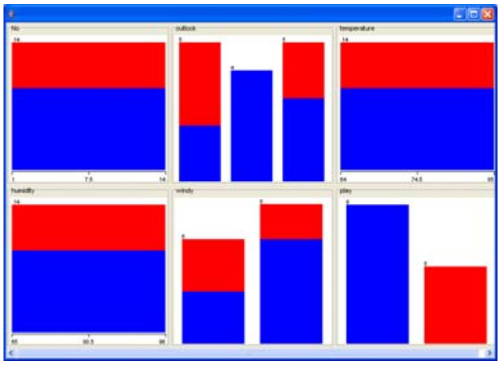
Statistic	Value
Minimum	65
Maximum	96
Mean	81.643
StdDev	10.285

Colour: play (Nom)

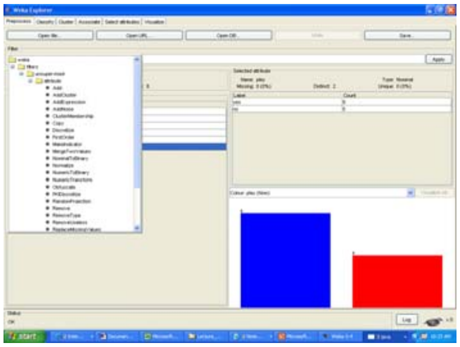
Status

OK

Parameter Distribution



Filters



Weka Explorer

Filter

- weka
  - filters
    - unsupervised
      - attribute
        - Add
        - AddCluster
        - AddExpression
        - AddNoise
        - ClusterMembership
        - Copy
        - Discretize
        - FirstOrder
        - MakeIndicator
        - MergeTwoValues
        - NominalToBinary
        - Normalize
        - NumericToBinary
        - NumericTransform
        - Obfuscate
        - PKIDiscretize
        - RandomProjection
        - Remove
        - RemoveType
        - RemoveUseless
        - ReplaceMissingValues

Selected attribute

Name: play Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

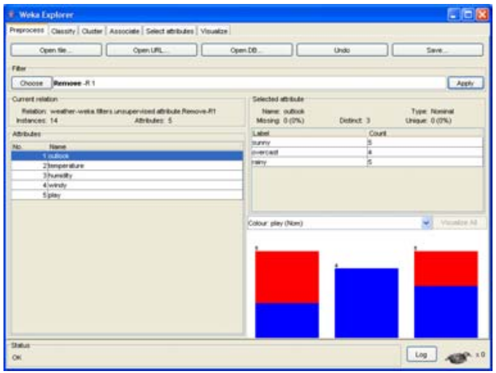
Label	Count
yes	9
no	5

Colour: play (Nom)

Status

OK

Filters



Filter: **Remove** -R 1

Current relation
Relation: weather-weka.filters.unsupervised.attribute.Remove-R1
Instances: 14
Attributes: 5

Attributes

No.	Name
1	outlook
2	temperature
3	humidity
4	windy
5	play

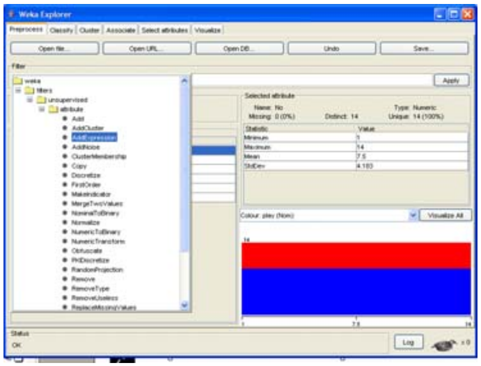
Selected attribute
Name: outlook
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)

Label	Count
sunny	5
overcast	4
rainy	5

Colour: play (Nom)

Status
OK

Filters



Filter

- weka
  - filters
    - unsupervised
      - attribute
        - Add
        - AddCluster
        - AddExpression
        - AddNoise
        - ClusterMembership
        - Copy
        - Discretize
        - FirstOrder
        - MakeIndicator
        - MergeTwoValues
        - NominalToBinary
        - Normalize
        - NumericToBinary
        - NumericTransform
        - Obfuscate
        - PKIDiscretize
        - RandomProjection
        - Remove
        - RemoveType
        - RemoveUseless
        - ReplaceMissingValues

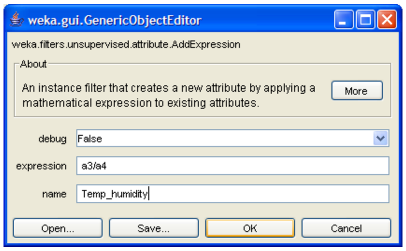
Selected attribute
Name: No
Type: Numeric
Missing: 0 (0%)
Distinct: 14
Unique: 14 (100%)

Statistic	Value
Minimum	1
Maximum	14
Mean	7.5
StdDev	4.183

Colour: play (Nom)

Status
OK

Filters



weka.filters.unsupervised.attribute.AddExpression

About

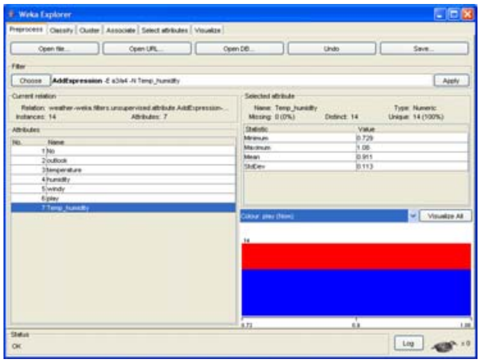
An instance filter that creates a new attribute by applying a mathematical expression to existing attributes.

debug: False

expression: a3/a4

name: Temp_humidity

Filters



Filter: **AddExpression** -E a3/a4 -N Temp_humidity

Current relation
Relation: weather-weka.filters.unsupervised.attribute.AddExpression-...
Instances: 14
Attributes: 7

Attributes

No.	Name
1	No
2	outlook
3	temperature
4	humidity
5	windy
6	play
7	Temp_humidity

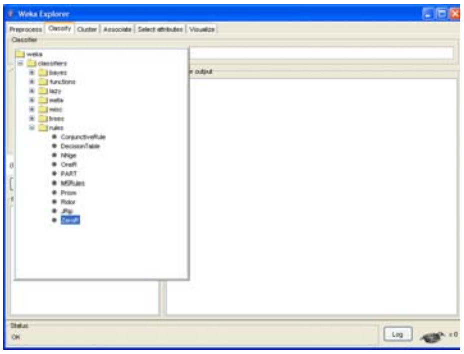
Selected attribute
Name: Temp_humidity
Type: Numeric
Missing: 0 (0%)
Distinct: 14
Unique: 14 (100%)

Statistic	Value
Minimum	0.729
Maximum	1.08
Mean	0.911
StdDev	0.113

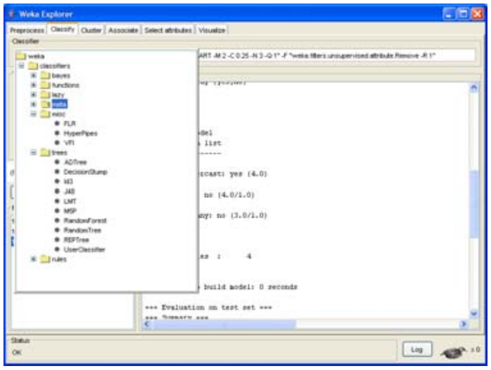
Colour: play (Nom)

Status
OK

Classifier



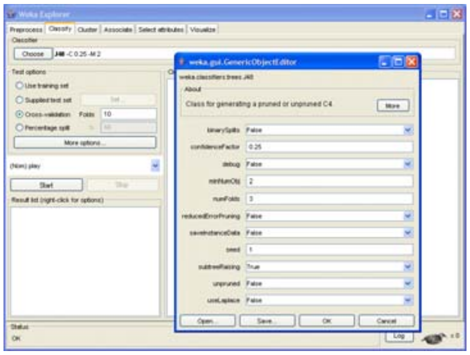
Classifiers



ART -M 2 -C 0.25 -N 3 -Q 1" -F "weka.filters.unsupervised.attribute.Remove -R 1"

```
odel
 list
------

rcast: yes (4.0)

  no (4.0/1.0)

ny: no (3.0/1.0)


es  :     4


 build model: 0 seconds

=== Evaluation on test set ===
=== Summary ===
```

Decision Tree



J48 -C 0.25 -M 2

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

More options...

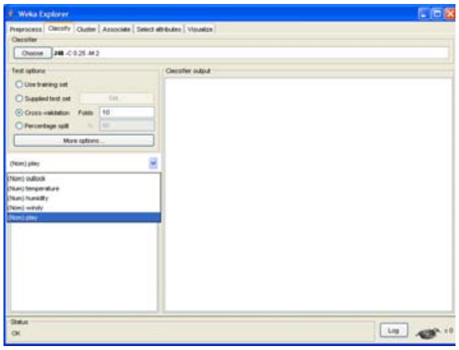
(Nom) play

weka.classifiers.trees.J48

Class for generating a pruned or unpruned C4.

binarySplits False
confidenceFactor 0.25
debug False
minNumObj 2
numFolds 3
reducedErrorPruning False
saveInstanceData False
seed 1
subtreeRaising True
unpruned False
useLaplace False

Decision Tree



J48 -C 0.25 -M 2

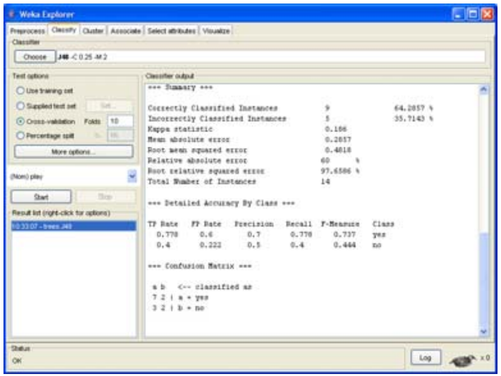
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split

More options...

(Nom) play

(Nom) outlook
(Num) temperature
(Num) humidity
(Nom) windy
(Nom) play

Decision Tree



Classifier: J48 -C 0.25 -M 2

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

(Nom) play

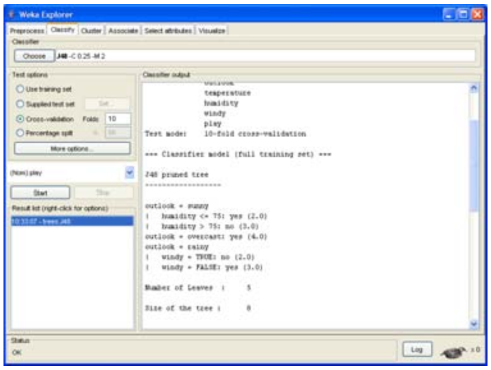
Result list (right-click for options): 10:33:07 - trees.J48

Classifier output

```
=== Summary ===

Correctly Classified Instances           9               64.2857 %
Incorrectly Classified Instances         5               35.7143 %
Kappa statistic                          0.186
Mean absolute error                      0.2857
Root mean squared error                  0.4818
Relative absolute error                 60      %
Root relative squared error             97.6586 %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.778     0.6        0.7       0.778     0.737    yes
  0.4       0.222      0.5       0.4       0.444    no

=== Confusion Matrix ===

 a b   <-- classified as
 7 2 | a = yes
 3 2 | b = no
```

Status: OK

Decision Tree



Classifier: J48 -C 0.25 -M 2

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

(Nom) play

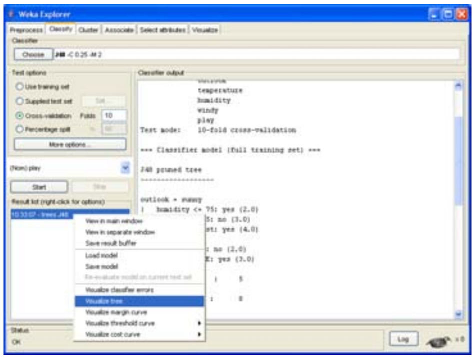
Result list (right-click for options): 10:33:07 - trees.J48

Classifier output

```
              outlook
              temperature
              humidity
              windy
              play
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

outlook = sunny
|   humidity <= 75: yes (2.0)
|   humidity > 75: no (3.0)
outlook = overcast: yes (4.0)
outlook = rainy
|   windy = TRUE: no (2.0)
|   windy = FALSE: yes (3.0)

Number of Leaves  :     5

Size of the tree :      8
```

Status: OK

Decision Tree



Classifier: J48 -C 0.25 -M 2

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

(Nom) play

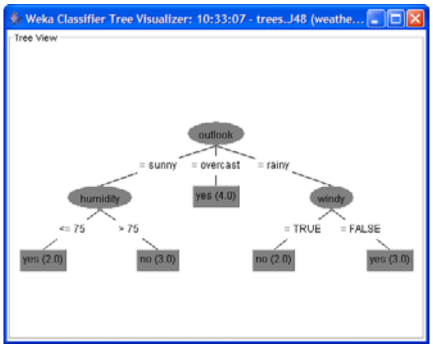
Result list (right-click for options): 10:33:07 - trees.J48

Classifier output

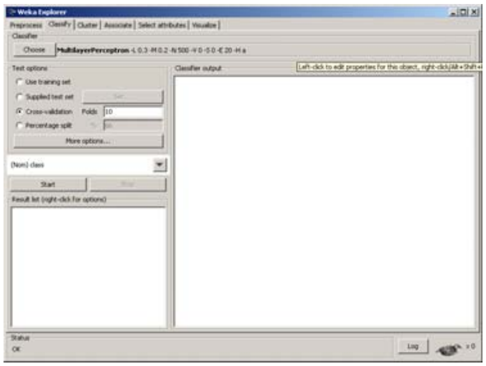
```
              outlook
              temperature
              humidity
              windy
              play
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

outlook = sunny
|   humidity <= 75: yes (2.0)
              5: no (3.0)
              st: yes (4.0)

              : no (2.0)
              E: yes (3.0)

              :     5

              :     8
```

Status: OK

Decision Tree



Neural Networks


Weka Explorer
Preprocess | Classify | Cluster | Associate | Select attributes | Visualize
Classifier
Choose
Test options
Classifier output
Start
Stop
Result list (right-click for options)
Status
Log

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

Left-click to edit properties for this object, right-click/Alt+Shift+...

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

More options...

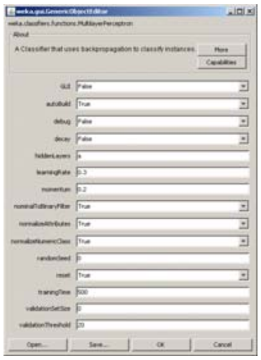
(Nom) class

Status: OK

The University of Iowa

Intelligent Systems Laboratory

Neural Networks



weka.classifiers.functions.MultilayerPerceptron

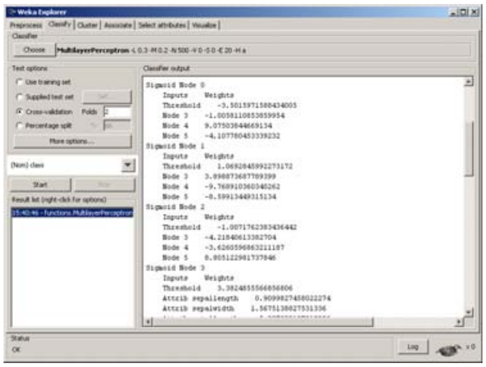
A Classifier that uses backpropagation to classify instances.

GUI: False
autoBuild: True
debug: False
decay: False
hiddenLayers: a
learningRate: 0.3
momentum: 0.2
nominalToBinaryFilter: True
normalizeAttributes: True
normalizeNumericClass: True
randomSeed: 0
reset: True
trainingTime: 500
validationSetSize: 0
validationThreshold: 20

The University of Iowa

Intelligent Systems Laboratory

Neural Networks


Weka Explorer
Preprocess | Classify | Cluster | Associate | Select attributes | Visualize
Classifier
Choose
Test options
Classifier output
Start
Stop
Result list (right-click for options)
Status
Log

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 2
○ Percentage split % 66

More options...

(Nom) class

15:40:46 - functions.MultilayerPerceptron

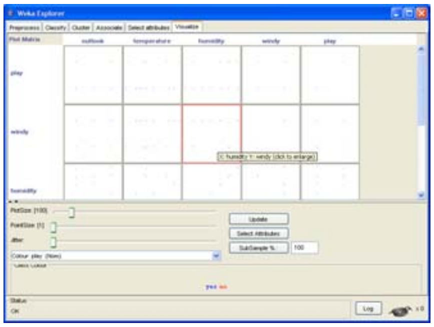
```
Sigmoid Node 0
    Inputs    Weights
    Threshold    -3.5015971588434005
    Node 3    -1.0058110853859954
    Node 4    9.07503844669134
    Node 5    -4.107780453339232
Sigmoid Node 1
    Inputs    Weights
    Threshold    1.0692645992273172
    Node 3    3.898873687789399
    Node 4    -9.768910360340262
    Node 5    -8.59913449315134
Sigmoid Node 2
    Inputs    Weights
    Threshold    -1.0071762383436442
    Node 3    -4.218406133827O4
    Node 4    -3.626059686321187
    Node 5    8.805122981737846
Sigmoid Node 3
    Inputs    Weights
    Threshold    3.382485556656806
    Attrib sepallength    0.9099827458022274
    Attrib sepalwidth    1.5675138827531336
```

Status: OK

The University of Iowa

Intelligent Systems Laboratory

Visualization



PlotSize: [100]

PointSize: [1]

Jitter

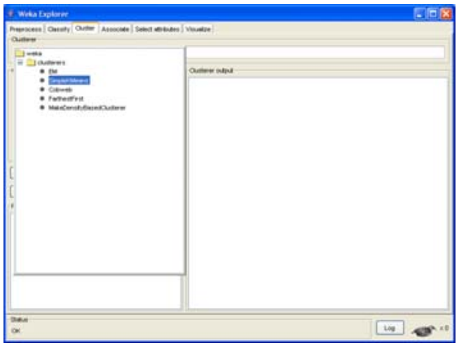
Colour: play (Nom)

SubSample %: 100

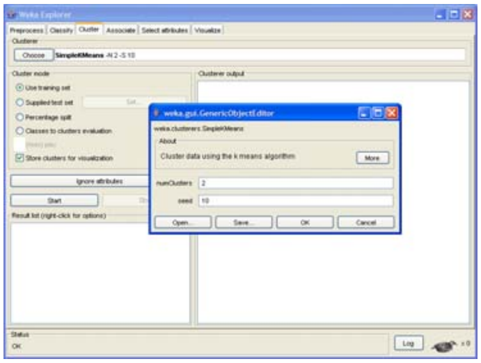
The University of Iowa

Intelligent Systems Laboratory

Clustering



Clustering



Clusterer: SimpleKMeans -N 2 -S 10

Cluster mode: ◉ Use training set

☑ Store clusters for visualization

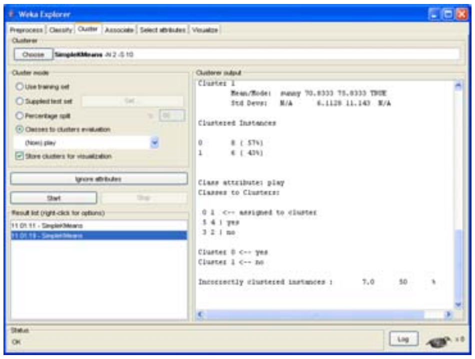
weka.clusterers.SimpleKMeans

Cluster data using the k means algorithm

numClusters: 2

seed: 10

Clustering



Clusterer: SimpleKMeans -N 2 -S 10

Cluster mode: ◉ Classes to clusters evaluation

(Nom) play

☑ Store clusters for visualization

Result list:

11:01:11 - SimpleKMeans

11:01:19 - SimpleKMeans

```
Cluster 1
        Mean/Mode:   sunny 70.8333 75.8333 TRUE
        Std Devs:    N/A      6.1128 11.143  N/A

Clustered Instances

0       8 ( 57%)
1       6 ( 43%)


Class attribute: play
Classes to Clusters:

 0 1  <-- assigned to cluster
 5 4 | yes
 3 2 | no

Cluster 0 <-- yes
Cluster 1 <-- no

Incorrectly clustered instances :       7.0      50      %
```

UCI repository

- http://archive.ics.uci.edu/ml/index.html