



**Weka**

**Praktické použití**

**Antonín Pavelka**

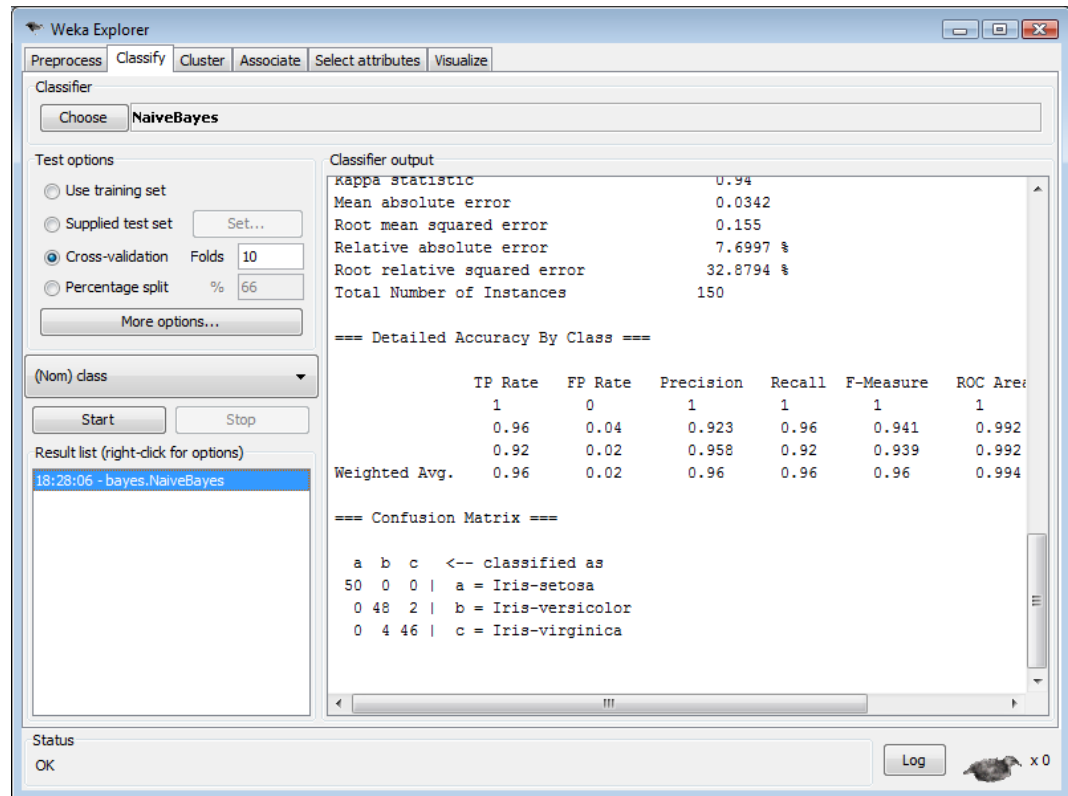
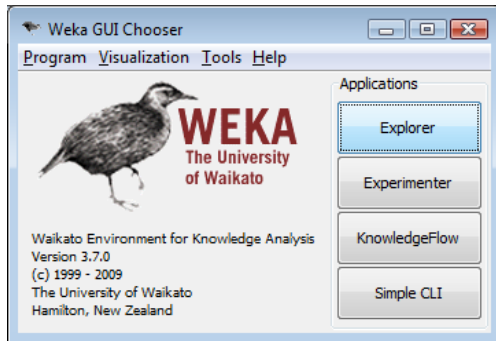
# Weka - úvod

- kolekce algoritmů strojového učení pro dolování z dat
- University of Waikato, Nový Zéland
- 1993 TCL/TK, C, Makefiles
- 1997 rozhodnutí přejít na čistou Javu
- integrována
  - RapidMiner
  - Petaho (systém business intelligence)
- GNU General Public License

# Ovládání

- spuštění  
java -Xmx1024m -jar weka.jar
- grafické rozhraní
  - Explorer – jednotlivé činnosti na kliknutí
  - Experimenter – systematické srovnání
  - Knowledge flow – činnosti jako tok
- příkazový řádek
- Java API

# Ukázka – grafické rozhraní ...



# ... příkazový řádek ...

```
java -classpath weka.jar  
    weka.classifiers.bayes.NaiveBayes  
    -t data/iris.arff
```

# ... Java API

```
Instances instances = new Instances(  
    new BufferedReader(  
        new FileReader("iris.arff")));  
instances.setClassIndex(instances.numAttributes() - 1);  
  
NaiveBayes c = new NaiveBayes();  
  
Evaluation eval = new Evaluation(instances);  
  
eval.crossValidateModel(c, instances, 10, new Random(1));  
  
System.out.println(eval.toSummaryString());  
System.out.println(eval.toMatrixString());
```

# 1. Attribute-Relation File Format (ARFF)

## ARFF soubor

```
@relation spambase
% spam, non-spam
@attribute word_freq_make real
@attribute 'char_freq_#' real
@attribute {spam, ham}
@data
0,0.64,0.64,spam
0.21,0.28,0.5,spam
0.06,0,0.71,ham
```

## Čas

```
@ATTRIBUTE timestamp DATE "yyyy-MM-dd HH:mm:ss"
@DATA "2001-04-03 12:12:12" "2001-05-03 12:59:55"
```

## Řídký formát

```
0, X, 0, Y, "class A" → {1 X, 3 Y, 4 "class A"}
0, 0, W, 0, "class B" → {2 W, 4 "class B"}
```

## Řetězce

```
@attribute LCC string
@attribute LCSH string

@data
AG5, 'Encyclopedias and dictionaries.;Twentieth century.,
```

## Chybějící hodnoty

```
4.4,?,1.5,?,Tolkien
```

# 2. Předzpracování dat

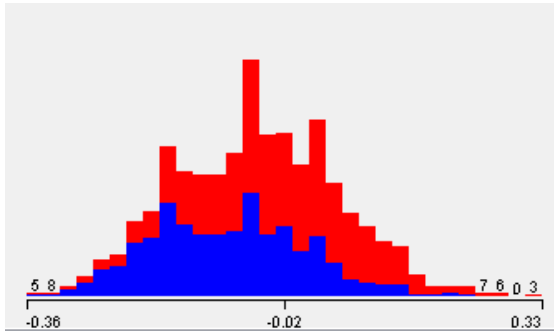
The screenshot shows the Weka Explorer interface. The 'Preprocess' tab is active. The 'Current relation' section shows 'Relation: deleterious', 'Instances: 4898', 'Attributes: 9', and 'Sum of weights: 4898'. The 'Attributes' list includes PDB, INDEX, FROM, TO, AUTOMUTE (selected), MAPP, SIFT, SNAP, and EFFECT. The 'Selected attribute' section for AUTOMUTE shows 'Name: AUTOMUTE', 'Type: Numeric', 'Missing: 0 (0%)', 'Distinct: 67', and 'Unique: 4 (0%)'. A table of statistics for AUTOMUTE is shown below:

Statistic	Value
Minimum	-0.36
Maximum	0.33
Mean	-0.053
StdDev	0.108

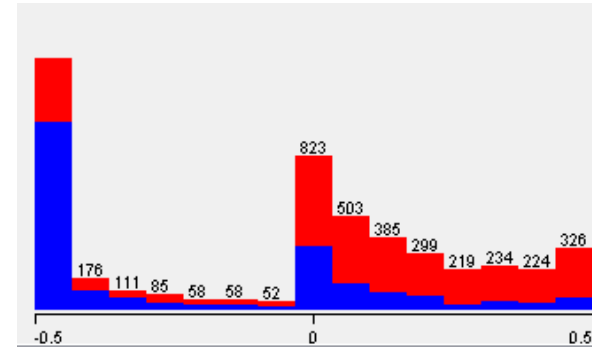
The histogram at the bottom right shows the distribution of the AUTOMUTE attribute. The x-axis ranges from -0.36 to 0.33. The distribution is bimodal, with a peak around -0.1 and another peak around 0.1. The histogram is overlaid with a blue curve and a red curve. The class is set to 'EFFECT (Nom)' and the 'Visualize All' button is visible.



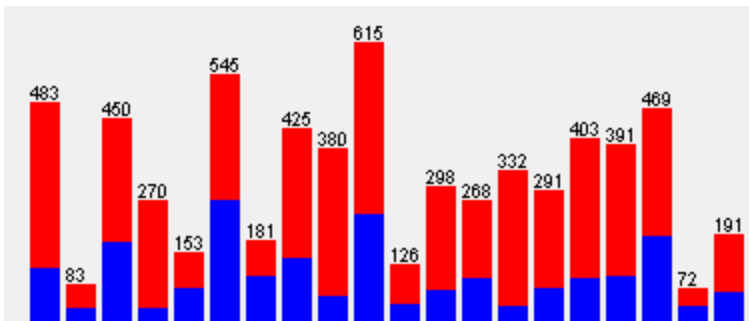
# Histogramy



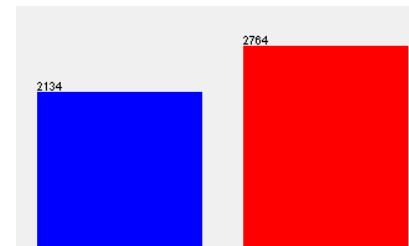
užitečný číselný atribut



podezřelý číselný atribut



20-hodnotový atribut



binární cílový atribut

# Filtry

## Unsupervised

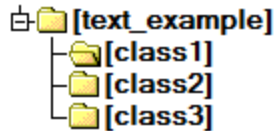
- Remove –V –R 1-5,8 (V = inverze, zachovej pouze tyto atributy)
- Discretize
  - některé algoritmy nepracují s čísly
  - urychlení
  - někdy i zvýšení přesnosti
- StringToWordVector

## Supervised

- Discretize
- AddClassification
- AttributeSelection

Multifilter – aplikuje několik filtrů po sobě

# StringToWordVector



- Dumbek's Random Stuff
- Random Stuff
- Stefan Tilkov's Random Stuff

htm  
htm  
htm

```
TextDirectoryLoader loader = new TextDirectoryLoader();  
loader.setDirectory(new File("c:/data/text_example"));  
Instances dataRaw = loader.getDataSet();
```

```
ArffSaver s1 = new ArffSaver();  
s1.setInstances(dataRaw);  
s1.setFile(new File("c:/data/text1.arff"));  
s1.writeBatch();
```



```
@attribute text string  
@attribute class {class1,class2,class3}  
  
@data  
'<html>\n\t<head>\n\t\t<title>Dumbek\'s Rand  
'<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.  
'<html>\r\n\r\n<head>\r\n<meta name="\descri  
'<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1
```

```
StringToWordVector filter = new StringToWordVector();  
filter.setInputFormat(dataRaw);  
Instances dataFiltered = Filter.useFilter(dataRaw, filter);
```

```
ArffSaver s2 = new ArffSaver();  
s2.setInstances(dataFiltered);  
s2.setFile(new File("c:/data/text2.arff"));  
s2.writeBatch();
```



```
@attribute class {class1,class2,class3}  
@attribute ago numeric  
@attribute align= numeric  
@attribute all numeric  
@attribute always numeric  
@attribute business numeric  
@attribute but numeric  
@attribute button numeric  
  
@data  
{1 1,3 1,4 1,11 1,12 1,13 1,14 1,15.....  
{10 1,34 1,37 1,49 1,50 1,53 1,99 1....  
{2 1,5 1,6 1,7 1,8 1,9 1,31 1,32 1,.....
```

```
J48 c = new J48();  
c.buildClassifier(dataFiltered);  
System.out.println("Classifier model: " + c);
```

# Tab Classify - možnosti

- textový výstup
- vizualizace klasifikátoru
- More options – Output predictions
- parametry klasifikátoru
  - SMO - buildLogisticModels

# Tab Classify - algoritmy

- rules
  - ZeroR
- bayes
  - NaiveBayes
  - Adaptive One Dependence Estimators (AODE)
- functions
  - support vector machine: SMO, SMOreg, LibSVM
  - neuronová síť: MultilayerPerceptron
- trees
  - J48, RandomForest
- meta
  - boosting, bagging
  - *FilteredClassifier, CVPParameterSelection, AttributeSelectedClassifier, CostSensitiveClassifier*

# Optimalizace parametrů

- `meta.CVParameterSelection -P "C 1 100 20" ...`

Cross-validation Parameter: '-C' ranged from 1.0 to 100.0 with 20.0 steps

Classifier Options: -C 25.0 ...

# Vážení chyb

TP Rate

0.81

0.915

- `meta.CostSensitiveClassifier`

```
% Rows Columns
```

```
2 2
```

```
% Matrix elements
```

```
0 2
```

```
1 0
```

- cena za špatně klasifikovaný P je 2x větší než za N

# Tab Select attributes

- metoda hodnocení podmnožiny atributů
  - CfsSubsetEval – prediktivní schopnost jednotlivých atributů a jejich redundance
  - ClassifierSubsetEval, WrapperSubsetEval
- nebo
- metoda hodnocení jednotlivých atributů
  - ChiSquaredAttributeEval
- prohledávací metoda
  - ExhaustiveSearch, BestFirst, GeneticSearch
- validace
  - křížová
  - filtr AttributeSelectedClassifier



# Experimenter

Weka Experiment Environment

Setup Run Analyse

Source

Got 12 results

File... Database... Experiment

Configure test

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

Row: Select

Column: Select

Comparison field: Percent\_correct ▾

Significance: 0.05

Sorting (asc.) by: <default> ▾

Test base: Select

Displayed Columns: Select

Show std. deviations:

Output Format: Select

Perform test Save output

Result list

16:39:39 - Available resultsets

16:39:41 - Percent\_correct - bayes.NaiveBayes " 5995:

Test output

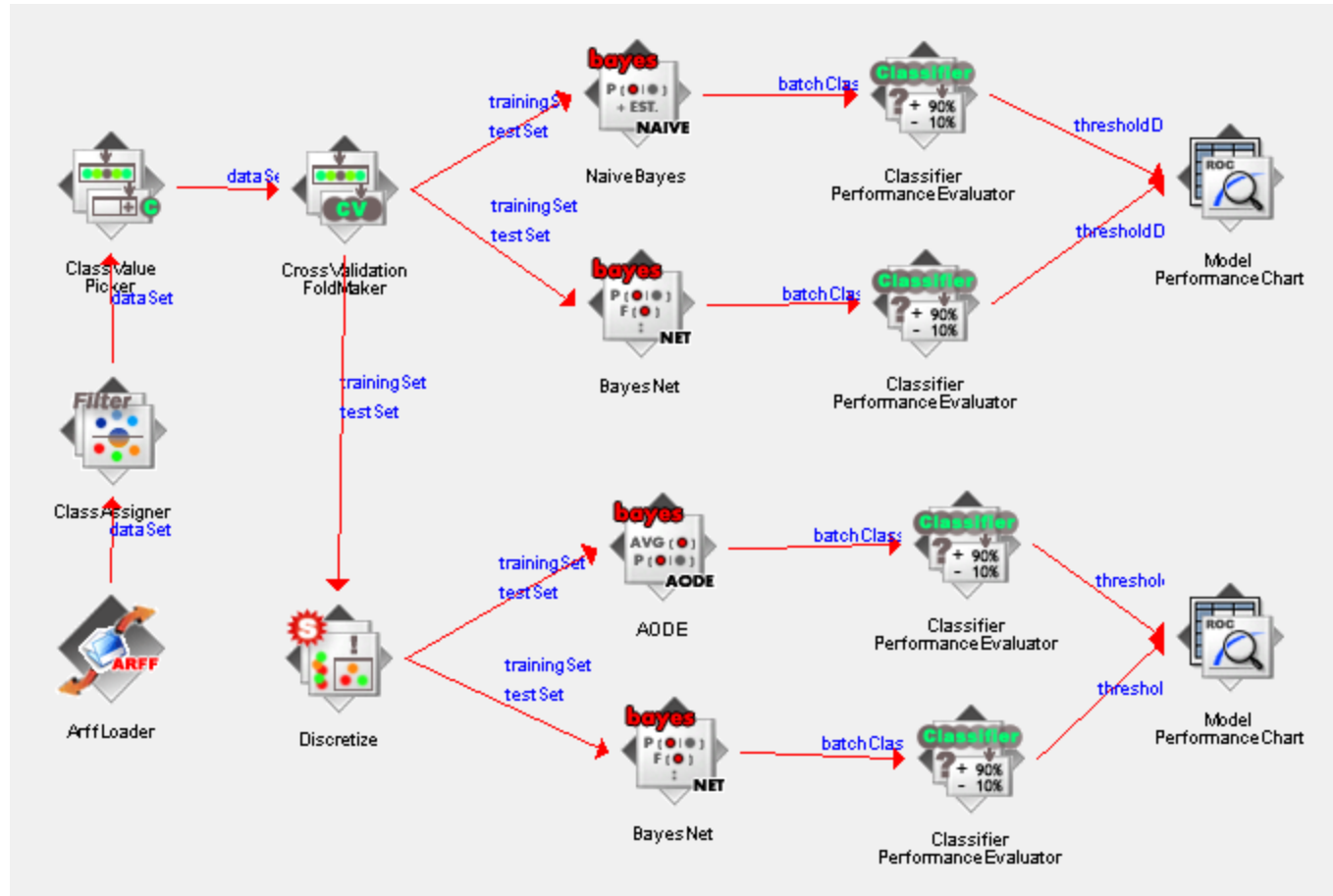
```

Tester: weka.experiment.PairedCorrectedTTester
Analysing: Percent_correct
Datasets: 1
Resultsets: 6
Confidence: 0.05 (two tailed)
Sorted by: -
Date: 1.11.09 16:39

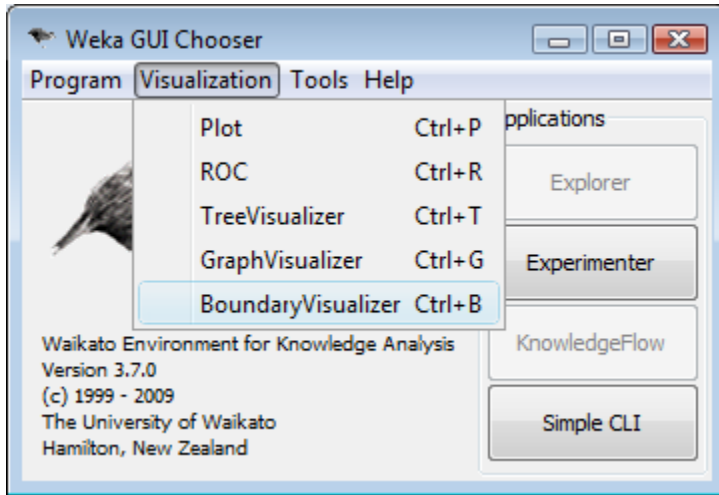
Dataset (1) bayes.Na | (2) bayes (3) bayes (4) funct (5) trees (6) trees
-----
spambase-weka.filters.sup (2) 90.18 | 90.31 93.05 v 93.87 v 92.24 93.07
-----
(v/ /*) | (0/1/0) (1/0/0) (1/0/0) (0/1/0) (0/1/0)

Key:
(1) bayes.NaiveBayes '' 5995231201785697655
(2) bayes.BayesNet '-D -Q bayes.net.search.local.K2 -- -P 1 -S BAYES -E bayes.net.estimate.S
(3) bayes.AODE '-F 1' 9197439980415113523
(4) functions.SMO '-C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K \"functions.supportVector.
(5) trees.J48 '-C 0.25 -M 2' -217733168393644444
(6) trees.RandomForest '-I 10 -K 0 -S 1' 4216839470751428698
        
```

# Knowledge Flow



# Vizualizace hranic



- jen pro nominální třídu

# Zdroje

## **Knihy**

WEKA Manual for Version 3-7-0

Data Mining: Practical Machine Learning Tools and Techniques

## **Web**

<http://www.cs.waikato.ac.nz/ml/weka/>

<http://weka.wikispaces.com/>

<http://wekadoocs.com/>

<http://www.hakank.org/weka/>