Weka

Praktické použití

Antonín Pavelka

Weka - úvod

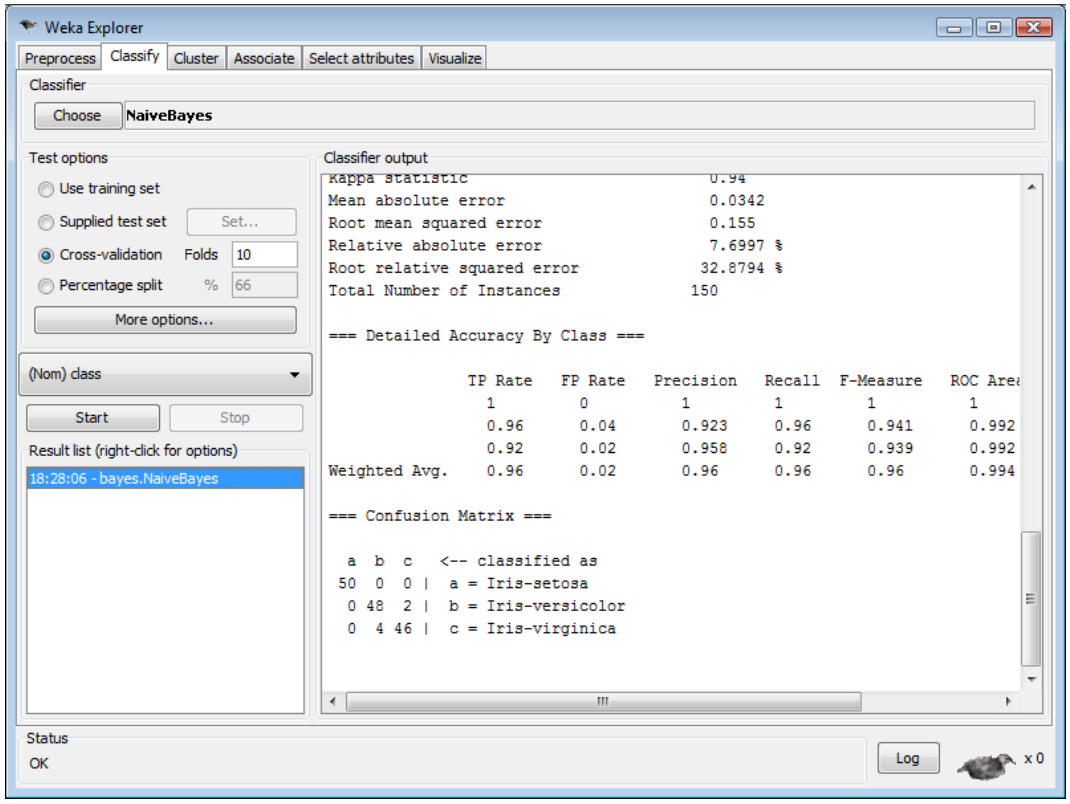
- kolekce algoritmů strojového učení pro dolování z dat
- University of Waikato, Nový Zéland
- 1993 TCL/TK, C, Makefiles
- 1997 rozhodnutí přejít na čistou Javu
- integrována
  - RapidMiner
  - Petaho (systém business intelligence)
- GNU General Public License

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

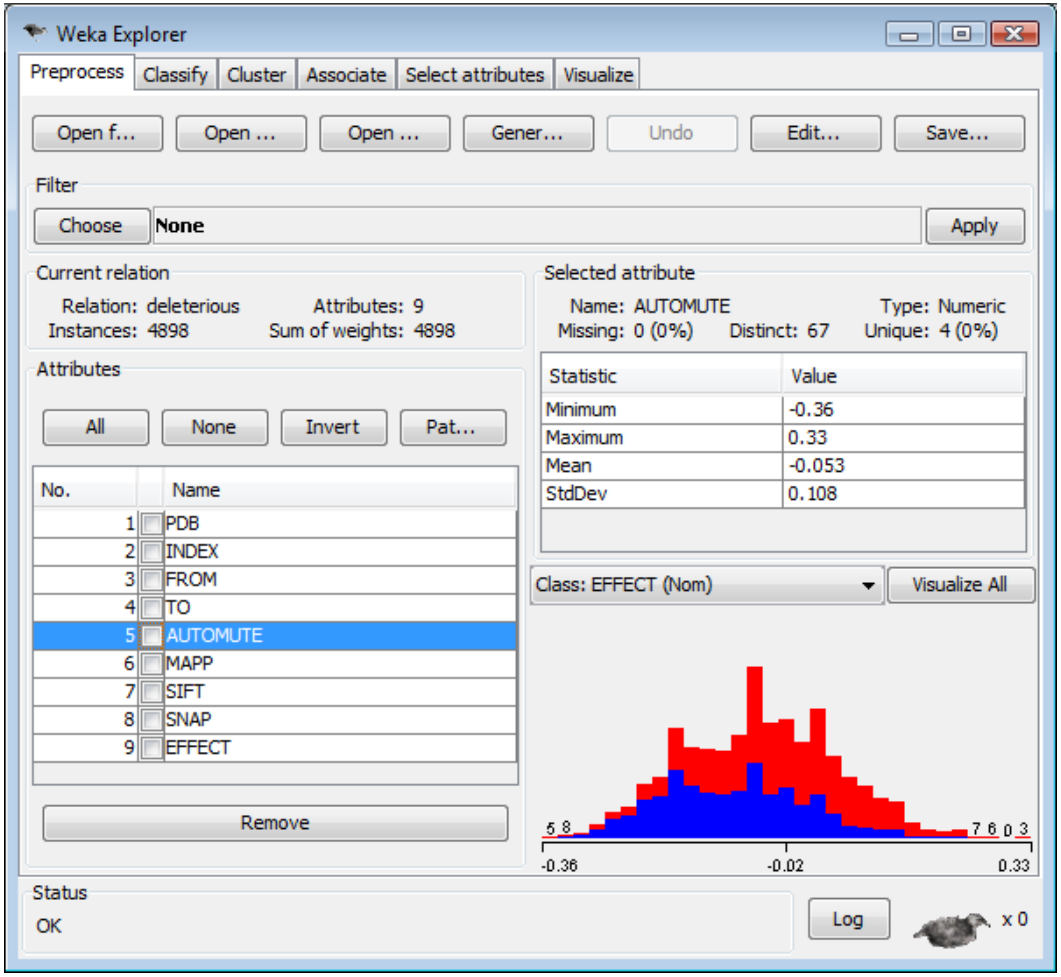
@attribute LCC string
@attribute LCSH string

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

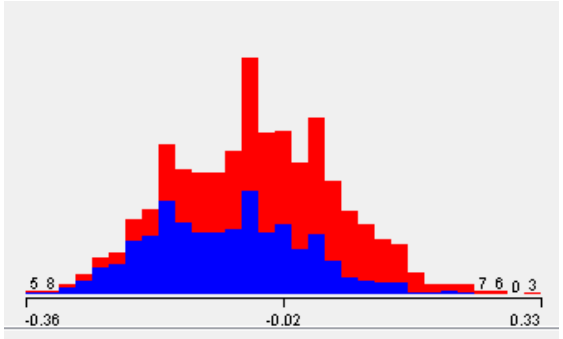
Chybějící hodnoty

4.4,?,1.5,?,Tolkien

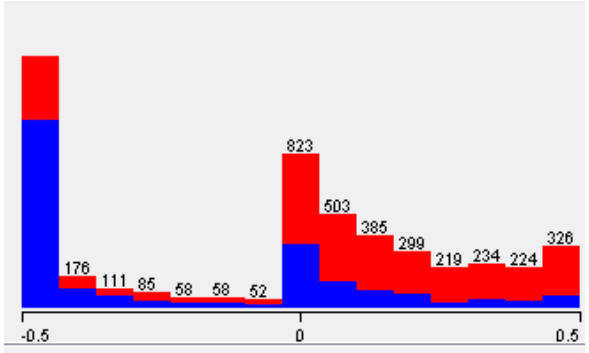
2. Předzpracování dat



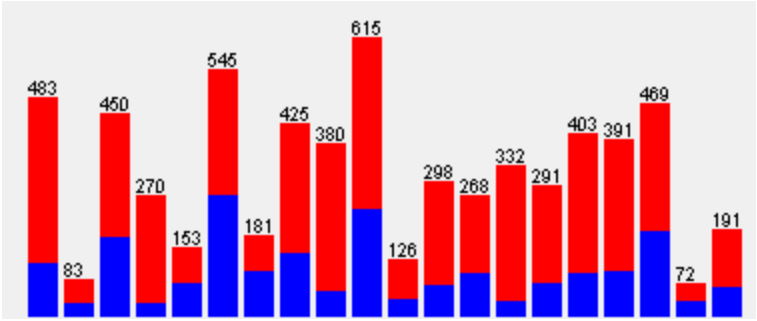
Histogramy



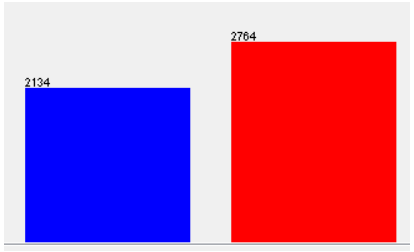
užitečný číselný atribut



podezřelý číselný atribut



20-hodnotový atribut



binární cílový atribut

Filtry

Unsupervised

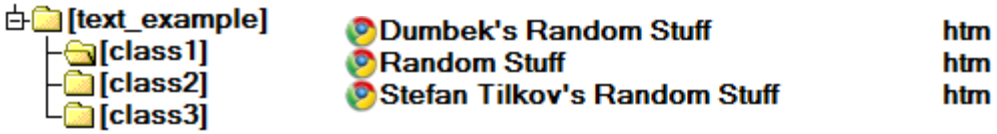
- Remove –V –R 1-5,8 (V = inverze, zachovej pouze tyto atributy)
- Discretize
  - některé algoritmy nepracují s čísly
  - urychlení
  - někdy i zvýšení přesnosti
- StringToWordVector

Supervised

- Discretize
- AddClassification
- AttributeSelection

Multifilter – aplikuje několik filtrů po sobě

StringToWordVector




```
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();

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();

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

```
@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
```

```
@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,.....
```

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, *CVParameterSelection, 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

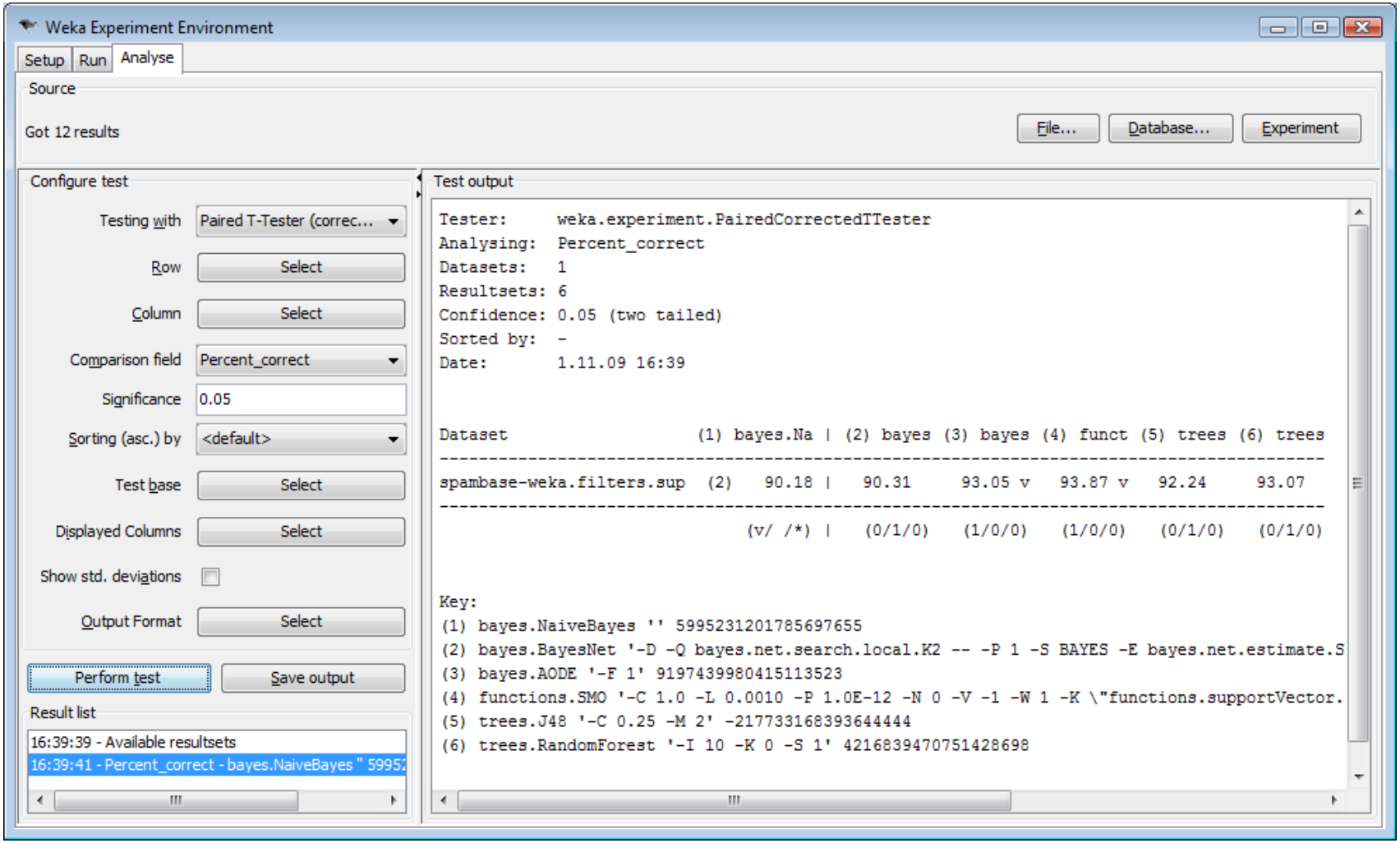
nebo

- metoda hodnocení jednotlivých atributů
  - ChiSquaredAttributeEval

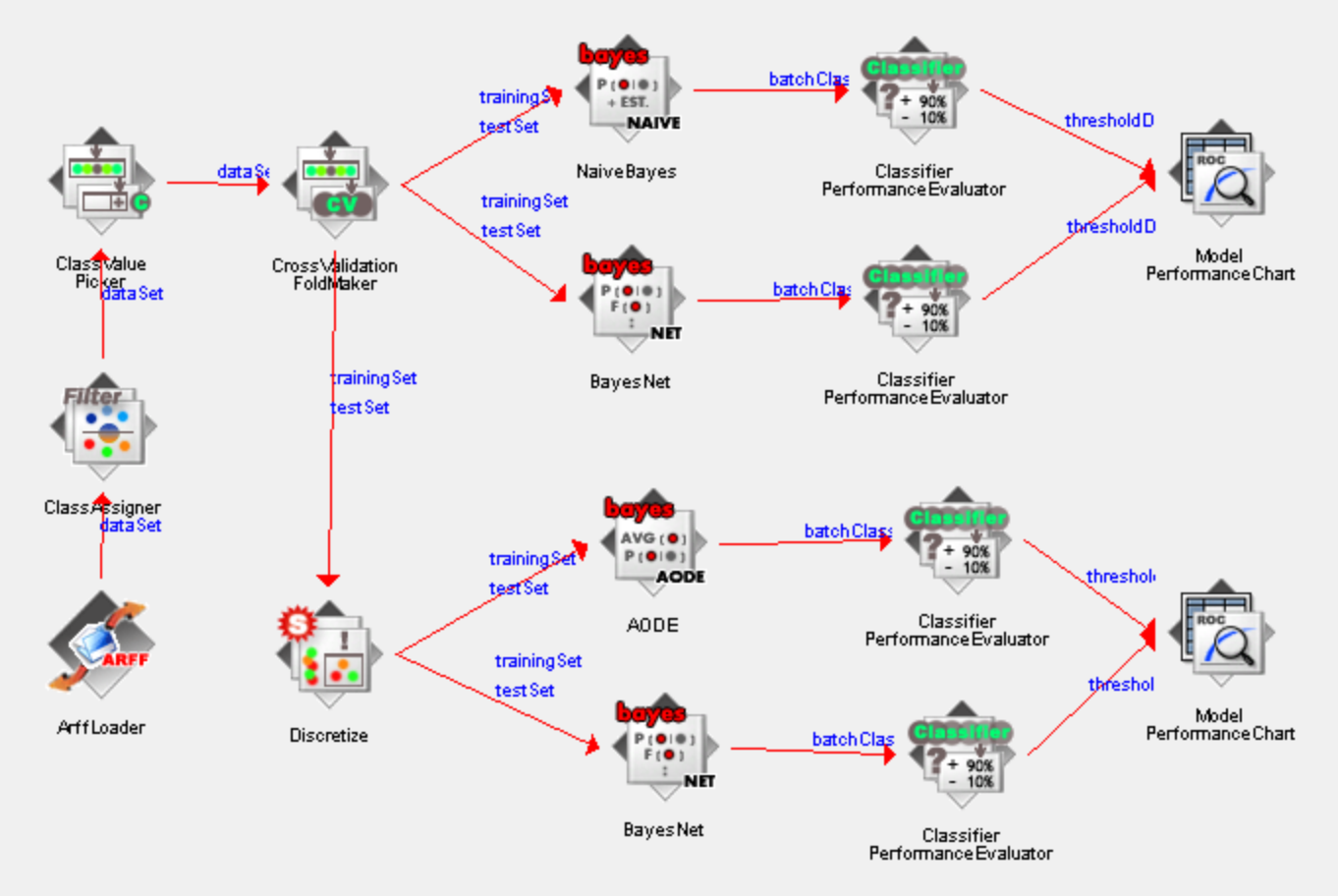
- prohledávací metoda
  - ExhaustiveSearch, BestFirst, GeneticSearch

- validace
  - křížová
  - filtr AttributeSelectedClassifier

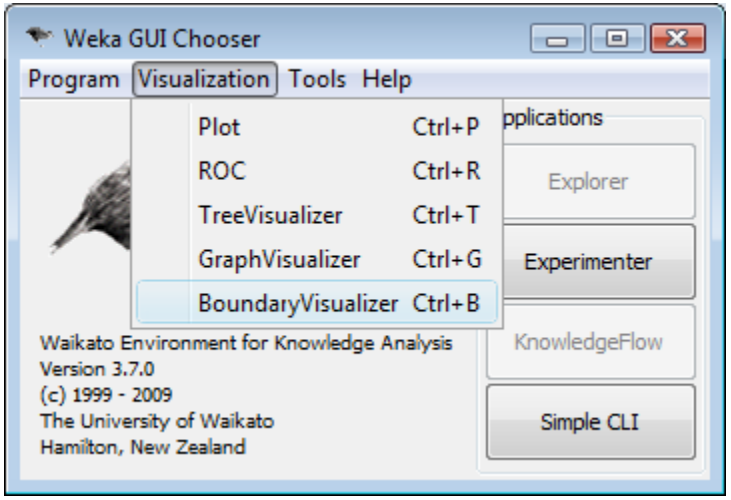
Experimenter



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://wekadocs.com/

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