Strojno učenje
Vježbe
WEKA 1

Tomislav Šmuc

Literatura:

Knjiga:

Ian H. Witten and Eibe Frank.

Data mining: Practical machine learning tools and techniques. 2005.

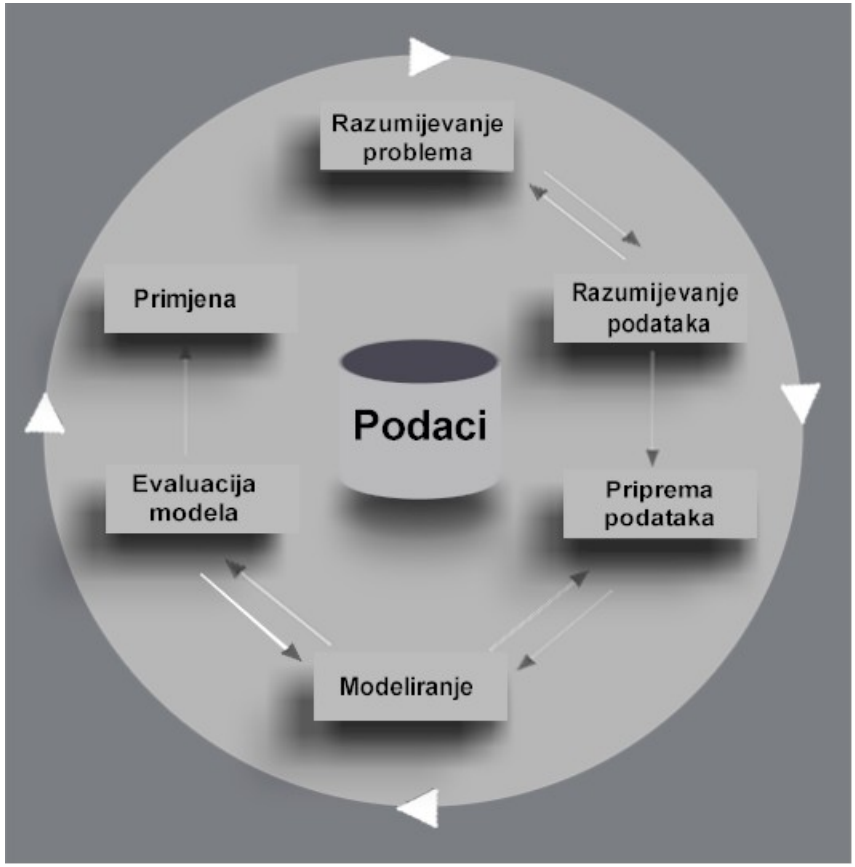
WEB – site: www.cs.waikato.ac.nz/ml/weka

WEKA

- Zbirka algoritama za data mining(JAVA, GPL)
- Gotovo svi važniji (i nevažniji) algoritmi strojnog učenja
- Source code
- puno projekata koji se nadovezuju ili koriste na WEKA-u
  - Pred-procesiranje podataka (re-sampling, filtriranje: atributi, primjeri)
  - Nadzirano, nenadzirano učenje
  - Klasifikacija, regresija, “clustering”, asocijativna pravila,.....
  - Algoritmi: Decision trees, rule learning, naiveBayes, NN, Bnets, SVM, Random Forest....
  - Meta-learning sheme
    - (boosting, bagging, stacking) i tehnike za kombiniranje više modela ili algoritama za učenje

Strojno učenje – u praksi: primjena u data mining-u

DM – kao standardni proces (CRISP-DM)



	Važnost	Trajanje
Razumijevanje problema i podataka	**80 %**	**20%**
Priprema podataka, modeliranje i evaluacija	**20 %**	**80%**

Naravno – uz pretpostavku da znate kako se to radi .....

WEKA - download

www.cs.waikato.ac.nz/ml/weka

- Requirements (Java 5.0+)
- Download (nova verzija 3.7.x - možda najbolje stable version + manual !!)
- Documentation, FAQ
- Tutorials (npr. http://maya.cs.depaul.edu/~classes/ect584/WEKA/index.html)
- Datasets (odavde možete downloadati UCI Irvine ML datasets – samo mali broj – ali već u arff formatu !)
- .....
- Related Projects .....

WEKA - format ulaznih podataka

- arff; xarff, csv, c4.5 (names/data), binary,...
  + 10tak drugih formata

- **Moj_problem.arff – format podataka**

  <= Komentari – bilo gdje: % - ispred teksta, komentari nisu ograniceni kolicinom

```
% 1. Title: moj_problem
%  Author: Tom Smuc
% 2. Sources:

@relation  moj_problem

@attribute  x1  numeric
@attribute  x2  {plavo, bijelo, “crveno”}
.....
@attribute  xn  numeric
@attribute  class  {on,off}

@data
5.1, plavo,.... ,0.2,on
4.3, bijelo,.... ,2.2,off
4.3, ?,.... ,2.2,off
```

<= **ime datoteke ili problema**

<= **prazna linija**

<= **prvi atribut - numerički**

<= **drugi atribut – kategorički**

<= **n-ti atribut – numerički**

<= **n+1-vi atribut – default – zadnji = ciljni atribut**

<= **prazna linija**

<= **odavde pa do kraja - podaci – CSV format !**

<= **? Označava “nedostajuće” podatke (en. Missing data)**

WEKA - format ulaznih podataka

- iris.arff – problem dataset

```
@RELATION iris

@ATTRIBUTE sepallength      numeric
@ATTRIBUTE sepalwidth       numeric
@ATTRIBUTE petallength      numeric
@ATTRIBUTE petalwidth       numeric
@ATTRIBUTE class            {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
.....
.....
6.3,2.5,5.0,1.9,Iris-virginica
6.5,3.0,5.2,2.0,Iris-virginica
6.2,3.4,5.4,2.3,Iris-virginica
5.9,3.0,5.1,1.8,Iris-virginica
```

WEKA - format ulaznih podataka

- iris.csv

Sepallength,sepalwidth,petallength,petalwidth,class
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
.....
.....
6.3,2.5,5.0,1.9,Iris-virginica
6.5,3.0,5.2,2.0,Iris-virginica
6.2,3.4,5.4,2.3,Iris-virginica
5.9,3.0,5.1,1.8,Iris-virginica

WEKA - format ulaznih podataka

WINDOWS

- **CLASSPATH – environment varijabla**
- **set CLASSPATH=c:\Program Files\Weka-3-5\weka.jar**

U nekom vašem direktoriju .....

%prompt> java weka.classifiers.trees.J48

- Output : help on J48 (stabla odlučivanja – C4.5)

%prompt> java weka.classifiers.lazy.IB1

- Output : help on IB1 (1-nn !)

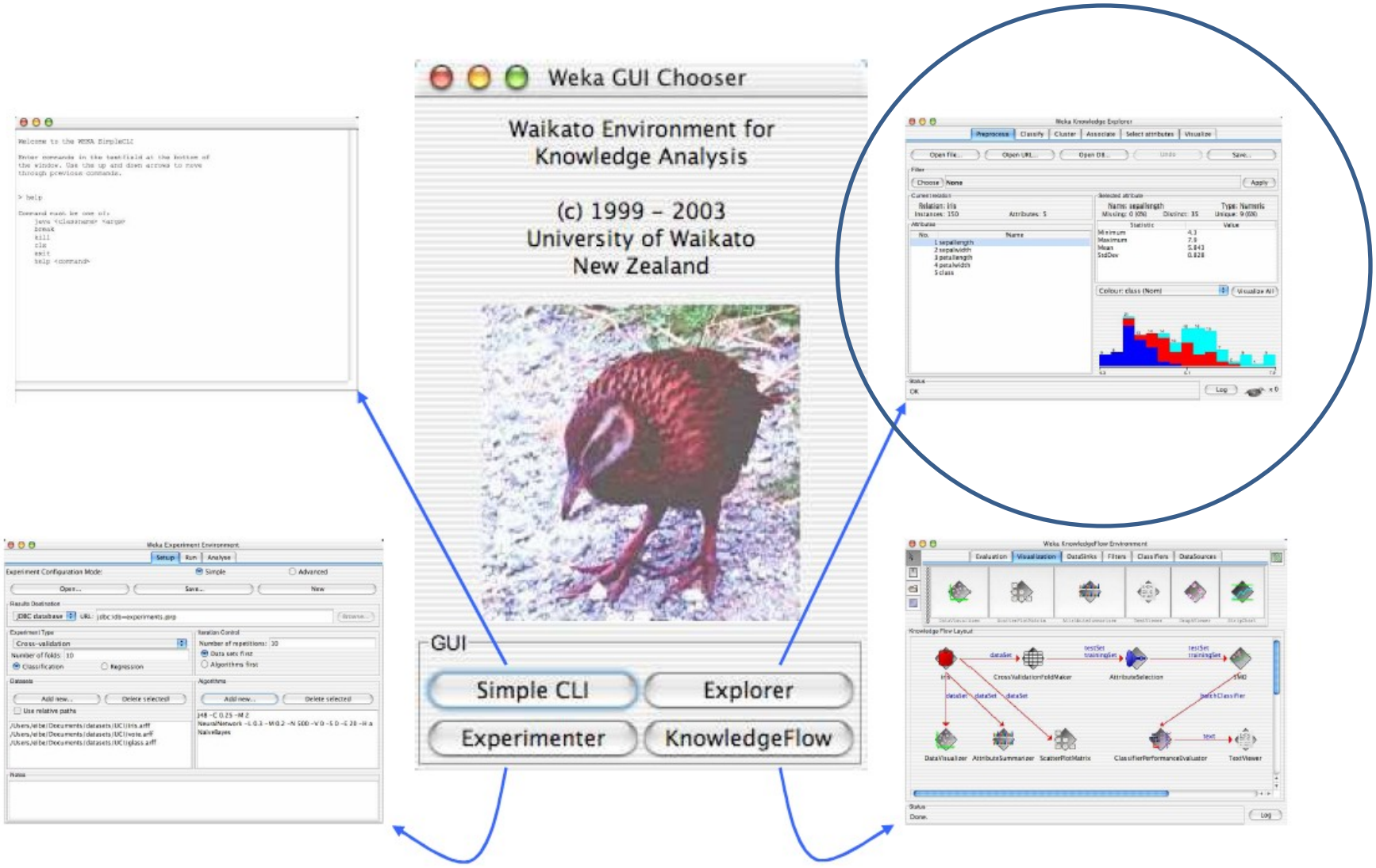
....

%prompt> java weka.classifiers.trees.J48 – t data\labor.arff

Output: summary rezultata na "training" skupu

-x 10 (opcija koja pokreće 10-fold cross validation)

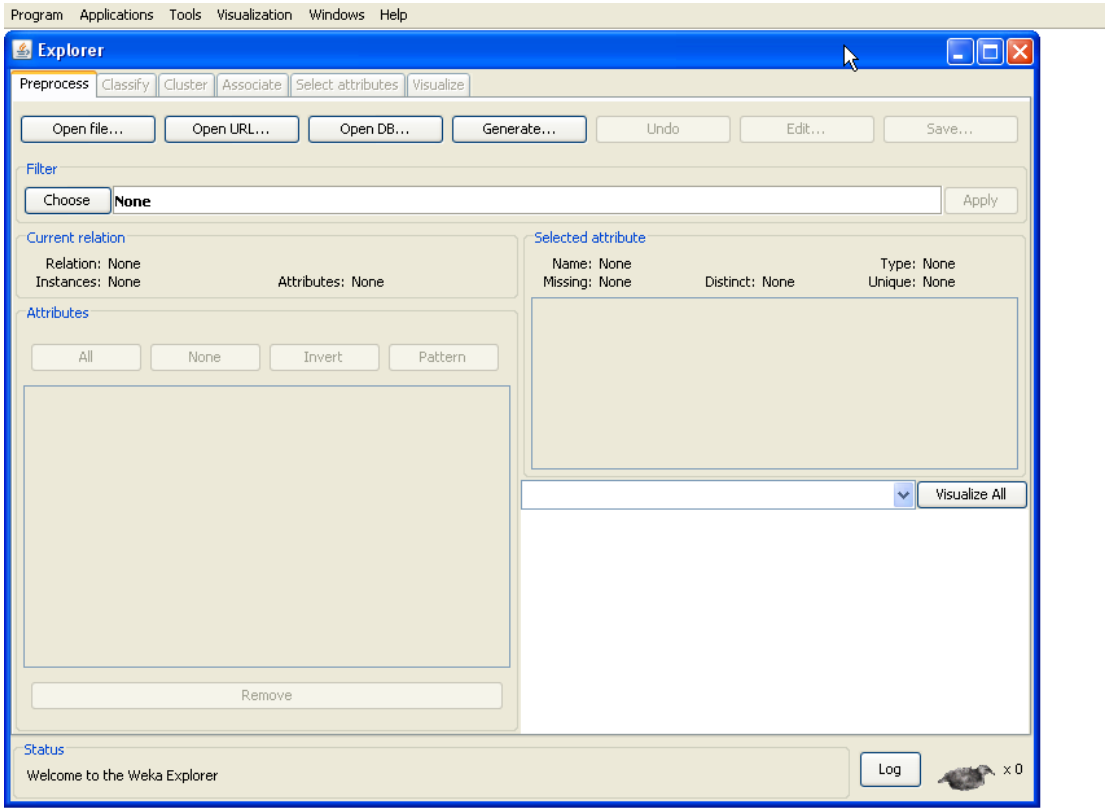
WEKA – glavni izbornik



WEKA – glavni izbornik

Izbornik

- **Explorer**
- Experimenter
- KnowledgeFlow
- SimpleCLI



Preprocessing (filteri)

- Supervised
- Unsupervised

- Za primjere (en. instances)

- za atribute (varijable)

Svrha filtera je priprema datoteke primjera za strojno učenje:

- uklanjanje/izbor/promjena dijela primjera
- uklanjanje/izbor/promjena nekih atributa/varijabli
- rješavanje problema nedostajućih vrijednosti
- ..... Puno scenarija i mogućnosti

WEKA – running – J48 = Stabla odlučivanja

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     labor-neg-data
Instances:    57
Attributes:   17
              duration
              wage-increase-first-year
              wage-increase-second-year
              wage-increase-third-year
              cost-of-living-adjustment
              working-hours
              pension
              standby-pay
              shift-differential
              education-allowance
              statutory-holidays
              vacation
              longterm-disability-assistance
              contribution-to-dental-plan
              bereavement-assistance
              contribution-to-health-plan
              class
Test mode:    10-fold cross-validation
```

WEKA – output (Decision Trees)

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

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

wage-increase-first-year <= 2.5: bad (15.27/2.27)
wage-increase-first-year > 2.5
|   statutory-holidays <= 10: bad (10.77/4.77)
|   statutory-holidays > 10: good (30.96/1.0)

Number of Leaves  : 3

Size of the tree :  5

Time taken to build model: 0 seconds
```

WEKA – output (Decision Trees)

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          42               73.6842 %
Incorrectly Classified Instances        15               26.3158 %
Kappa statistic                          0.4415
Mean absolute error                      0.3192
Root mean squared error                  0.4669
Relative absolute error                 69.7715 %
Root relative squared error             97.7888 %
Total Number of Instances               57
```

WEKA – output (Decision Trees)

```
=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.7       0.243      0.609     0.7       0.651      0.695    bad
  0.757     0.3        0.824     0.757     0.789      0.695    good

=== Confusion Matrix ===

  a  b   <-- classified as
 14  6 |  a = bad
  9 28 |  b = good
```

WEKA –za danas

Pred-procesiranje (podataka)

- 3 dataset-a (možda i još koji)
  - Iris
  - zoo
  - Waveform
- “Igra” s podacima:
  - Filteri
    - sampliranje, spremanje....
  - Vizualizacija podataka:
    - summary stats, scatter plots...

WEKA –za danas

Klasifikacija

- 3 klasifikatora (možda i koji više...)
  - Decision trees (J48)
  - Naive Bayes
  - k-nn
- Određivanje točnosti klasifikatora
  - Training error
  - Training/test
  - Cross validation
  - ROC krivulje
- Vizualizacija modela

WEKA –za zadaću

- Download/install
- Pročitati - upute/(dio knjige)
- korištenje

UCI datasets - zadatak 4+

- dostupni sa
  - ww.cs.waikato.ac.nz/ml/weka
  - www.math.hr/nastava/su

- Odabrati 4+ skupa podataka – po vlastitom nahođenju
- 4+ algoritma:
  - Naive Bayes(NB), Decision Tree (J48), k-nn (IBk)
  - MLP - MultiLayer Perceptron, (neuralna mreža)

WEKA –za zadaću

UCI datasets - zadatak 4+

- Napraviti krivulje učenja:
- veličina skupa za učenje za određivanje krivulje učenja – (10%, 20%, 40%, 80%, 100%)

- Usporediti algoritme (točnost) na svakom od ta 4+ skupa:
  - Train/test (66% split)
  - 2/5/10 fold XV
  - ROC (AUC)

- Nađite najbolji model (točnost, AUC...) !
  - (mijenjanje slobodnih parametara algoritama)

- Vizualizirajte najbolje modele (J48 npr)

WEKA –za zadaću

Napišite mali izvještaj o tome .....

ROK: Dva tjedna

Shema projekata – gruba skica

Praktični dio
Postepeno rješavanje kroz faze (uz malu pomoć:)
Za to će biti korištene vježbe (u petom mjesecu)

Faze:

a) odabir problema
b) proučavanje problema – što su podaci
c) čitanje član(a)ka - planiranje
c) razvoj rješenja (iteracije)
- prilagodba podataka (ako i koliko je potrebno)
- optimiranje modela
e) report – shema eksperimenta i rezultati

Može biti napravljeno u Rapid Miner-u, WEKA-i, R ili u alatu koji vi odaberete ili napravite!