

WEKA software



Bc. Václav Souhrada

v.souhrada@gmail.com



Obsah

- Rozpoznávání a klasifikace
- **Strojové učení – Machine Learning**
- Dolování dat – Data Mining
- **Co je to WEKA?**
- Historie
- **WEKA GUI**
- WEKA projekty



Rozpoznávání a klasifikace

- Význam českého slova **rozpoznávání** budeme chápat jako ekvivalent disciplíny anglicky nazývané **pattern recognition**.
- **Rozpoznávání** chápeme jako úlohu, při které zařazujeme objekty do tříd podle jejich společných vlastností tak, že objekty vzájemně si podobné zařazujeme do stejné třídy.



Rozpoznávání a klasifikace

- **Rozlišujeme:**

- **Klasifikaci** – zařazujeme do předem známého, pevného počtu tříd
 - Např. rozpoznávání znaků
- **Rozpoznávání** – zařazujeme do předem neznámého počtu tříd, které identifikujeme až během vlastního rozpoznávání



Strojové učení (Machine Learning)

- Učení
 - „...získání znalostí, nebo dovedností na základě studování, zkušeností, ...“
(OED)
- Strojové učení (Machine Learning)
 - „.... počítačový program se strojově učí, jestliže se jeho výkonnost na dané úloze zlepšuje s postupným získáváním zkušeností s ohledem na daná pravidla ...“
(Mitchell, 1997)



Strojové učení (Machine Learning)

- ML je v současné době “*jádrem*” umělé inteligence a její rozvoj je předpokladem dalšího rozvoje ML
- ML je prakticky až na název, shodné s moderní disciplínou nazývanou “*Dolování dat*” (Data Mining)
- ML ve světě:
 - Marketing – předvídaní vývoje cen akcií
 - Letecké a technické technologie
 - Robotní kopaná a další ...



Strojové učení (Machine Learning)

- Rozlišujeme:
 - Učení s učitelem (supervised learning)
 - Učení bez učitele (unsupervised learning)



Data Mining

- **Data Mining** (DM) = proces objevování vzorů (patterns) v datech
- Tento proces musí být automatický nebo polautomatický
- Objevené vzory musí být smysluplné, aby vedly k něčemu užitečnému



Data Mining

- Příslušné vzory mohou získat netriviálně skryté a potenciálně užitečné informace z dat
- **DM se používá např. v:**
 - *Komerční sféře* (rozhodnutí o rozesílání reklamních letáků)
 - *Vědeckém výzkumu* (např. při analýze genetické informace)
 - *Monitorování internetu*
 - *V "boji" s terorismem*



Co je to WEKA ?



Copyright: Martin Kramer (mkramer@wxs.nl)

Co je to WEKA ?

Základní charakteristika



- Waikato Environment for Knowledge Analysis
- “Nejmodernější” sbírka algoritmů *strojového učení* a nástrojů pro *předzpracování dat*
- WEKA je implementována v program. jazyku **Java**
- Licence - **GNU GPL**



Co je to WEKA ?

Základní charakteristika



- Snadné využití WEKA algoritmů
- Vhodná pro vytváření nových schémat strojového učení
- Efektivní implementace plug-inů



Co je to WEKA ?

Základní charakteristika



- Nástroje:
 - Předzpracování dat
 - Třídění
 - Regrese
 - Seskupení (Cluster)
 - Spojení pravidel
 - Vizualizace



Co je to WEKA ?

Základní charakteristika



- 49 nástrojů pro předzpracování dat
- 76 klasifikačních a regresních algoritmů
- 8 algoritmů pro předzpracování dat
- 3 algoritmy asociačních pravidel
- 15 ohodnocujících algoritmů
- 10 vyhledávacích algoritmů



Co je to WEKA ?

Základní charakteristika



- Tři základní GUI
 - “*The Explorer*” (analýza dat)
 - “*The Experimenter*” (experimenty)
 - “*The Knowledge Flow*” (vytváření nových schémat)
- Verze:
 - Book
 - Developer



Co je to WEKA ?

Podporované datové formáty



- ARFF (Attribute-Relation File Format)
- XRFF (eXtensible attribute-Relation File Format)
- Databáze (JDBC, MS Acces)
- URL
- CSV
- BSI, C4.5 (*.names*, *.data*) a binární soubory



Co je to WEKA ?

Podporované datové formáty



Formát ARFF



Co je to WEKA ?

Formát ARFF - charakteristika



- Skládá se ze dvou částí:
 - **Hlavička (Header)**
 - **Datová část (Data)**
- Řádek začínající znakem % se považuje za komentář
- První řádek (netýká se to komentářů) ARFF formátu **musí** začínat názvem relace ***@relation název_relace***



Co je to WEKA ?

Formát ARFF - charakteristika



- **Header**
 - Tato část obsahuje jméno relace a seznam atributů (sloupce v datotech).
 - **Jméno relace:** *@relation název_relace*
- **Deklarace atributů:** *@attribute jméno_atributu typ_atributu*
- Každý atribut je definován svým unikátním jménem
- **Datové typy:**
 - numeric
 - integer
 - real
 - string
 - date (*@attribute datum DATE "yyyy-MM-dd HH:mm:ss"*)
 - relational (pro multi-instanční atributy)
 - nominal (*@attribute pocasi {slunečno,zataženo,}*)



Co je to WEKA ?

Formát ARFF - charakteristika



- Header

Multinominální typ atributu

@relation elektrody

@attribute stroj {EEG32,EEGPRO,BRAINamp}

@attribute elektrody relational

@attribute e1 numeric

@attribute e2 numeric

@attribute e3 numeric

@attribute e4 numeric

@attribute e5 numeric

@end elektrody

@attribute normalniStav {true,false}



Co je to WEKA ?

Formát ARFF - charakteristika



- **Datová část**
- Tato část obsahuje data k příslušným atributům

Začátek datové sekce:

@data

Zápis dat:

sunny,75,70,TRUE,yes

overcast,72,90,TRUE,yes



Co je to WEKA ?

Formát ARFF



@relation weather

```
@attribute outlook {sunny, overcast,rainy}  
@attribute temperature numeric  
@attribute humidity numeric  
@attribute windy {TRUE,FALSE}  
@attribute play {yes,no}
```

@data

```
sunny,85,85, FALSE, no  
sunny,80,90, TRUE, no  
overcast,83,86, FALSE, yes  
rainy,70,96, FALSE, yes  
rainy,68,80, FALSE, yes  
rainy,65,70, TRUE, no
```



Historie



- Od r. 1993 – WEKA projekt financován vládou Nového Zélandu
(dříve financována prof. Ian Wittenem)



Historie



- **1993** – Vývoj rozhraní a infrastruktury
 - **Geoff Holmes** – vymyslel zkratku WEKA
 - **Andrew Donkin** – navrhl formát ARFF
- **1994** – *První* vydání v rámci univerz. **Waikato**
- **1996** – První veřejné vydání pod verzí 2.1
- **1997** – (červenec) **WEKA 2.2**



Historie



- **1997** – Rozhodnutí o přepsání do jaz. Java



- **Eibe Frank** – disertační práce
- Původně označena jako **JAWS** (**J**Ava **We**ka **S**ystem)

- **1998** – WEKA 2.3

- Poslední verze založená na TCL/TK systému

– WEKA 3

- 100% v **Jave**
- Poprvé verze **Developer**

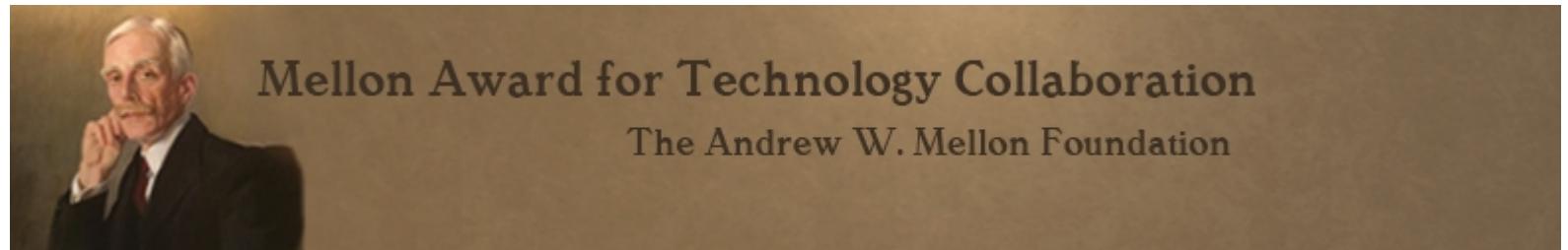


Historie



- 2008 – WEKA 3.5.8
- 2008 – Nominace na cenu – MATC

Mellon Award for Technology Collaboration



Copyright:(<http://matc.mellon.org/>)





WEKA GUI



- Posvítíme si na *strukturální vzorce*
- **Příklad** - “The weather problem” (TWP)
- TWP je malá datová sada (tiny dataset), kterou vývojáři WEKY používají jako ilustrativní příklad metod ML.



- Atributy:

- **outlook** (*počasi*)
 - *sunny* (*slunečno*)
 - *overcast* (*zataženo*)
 - *rainy* (*oblačno*)
- **temperature** (*teplota*)
 - *hot* (*vedro*)
 - *mild* (*příjemně*)
 - *cool* (*zima*)



- Atributy:

- **humidity** (*vlhkost*)
 - *high* (*slunečno*)
 - *normal* (*zataženo*)
- **windy** (vítr)
 - *TRUE* (*pravda*)
 - *FALSE* (*neprvada*)
- **play** (hra)
 - *yes* (*ano*)
 - *no* (*ne*)



WEKA GUI



Vysvětlení příkladu – tabulka

outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no



- **Výběr některých asociačních pravidel:**

- **if** temperature = cool **then** humidity = normal
- **if** humidity = normal and windy = FALSE **then** play = yes
- **if** outlook = sunny and play = no

then humidity = high

- **if** windy = FALSE and play = no

then outlook = sunny
and humidity = high



Vysvětlení příkladu – dataset v ARFF

@relation weather

```
@attribute outlook {sunny, overcast,rainy}  
@attribute temperature {hot, mild, cool}  
@attribute humidity {high, normal}  
@attribute windy {TRUE,FALSE}  
@attribute play {yes,no}
```



WEKA GUI



Vysvětlení příkladu - dataset v ARFF

@data

```
sunny,hot,high, FALSE,no  
sunny,hot,high, TRUE,no  
overcast,hot,high, FALSE, yes  
rainy,mild,high, FALSE, yes  
rainy,cool,normal, FALSE, yes  
rainy,cool,normal, TRUE, no  
overcast,cool,normal, TRUE, yes  
sunny,mild,high, FALSE, no  
sunny,cool,normal, FALSE, yes  
rainy,mild,normal, FALSE, yes  
sunny,mild,normal, TRUE, yes  
overcast,mild,high, TRUE, yes  
overcast,hot,normal, FALSE, yes  
rainy,mild,high, TRUE, no
```



WEKA GUI



- **The Explorer**
- **The Experimenter**
- **The Knowledge Flow**
- Simple CLI
- **Boundary Visualizer**
- Plot



WEKA GUI



- ROC
- Tree Visualizer
- Graph Visualizer
- Arff Viewer
- Sql Viewer
- Neuron Network





WEKA GUI

WEKA 2.1 – Machine Learning Workbench

WEKA Train Test Scheme Results Advanced Help

Training File: *golf.arff*
Testing File: *No Testing File*
Relation: *golf*
Attributes: 5 Instances: 14
Scheme: No scheme currently selected
Description:
No Scheme

Attributes

Include All Exclude All Prune

Number	Info	Name
1	◆	outlook
2	◆	temperature
3	◆	humidity
4	◆	windy
5	◆	class

Attribute Information

Name: *outlook*
Type: *Enumerated*
Missing: 0

Value	Count
sunny	5
overcast	4
rain	5

System Log

```
12:49:34: (c) 1993-1996 The University of Waikato, Hamilton, New Zealand
12:49:34: www: http://www.cs.waikato.ac.nz/~ml/
12:49:34: email: wekasupport@cs.waikato.ac.nz
12:49:34: WEKA is using preval as the rule evaluator
13:24:46: Loaded ARFF File - /home/ml/wekalite2.1/datasets.lite/golf.arff
13:25:57: Loaded ARFF File - /home/ml/wekalite2.1/datasets.lite/golf.arff
```

Status Okay Schemes Running 0 / 5

WEKA GUI



Weka 3.5.6 - Explorer

Program Applications Tools Visualization Windows Help

Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter Choose **None** Apply

Current relation
Relation: weather Instances: 14 Attributes: 5

Attributes
All None Invert Pattern

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

Class: play (Nom) Visualize All

Status OK Log x 0

WEKA GUI



The Explorer



WEKA GUI

The Explorer



Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize |

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter
Choose **None** Apply

Current relation
Relation: weather Instances: 14 Attributes: 5

Attributes
All None Invert Pattern

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

Class: play (Nom) Visualize All

5 4 5

Remove

Status OK Log x 0

A screenshot of the WEKA Explorer interface. The window title is 'Explorer'. The menu bar includes 'Preprocess', 'Classify', 'Cluster', 'Associate', 'Select attributes', and 'Visualize'. Below the menu is a toolbar with buttons for 'Open file...', 'Open URL...', 'Open DB...', 'Generate...', 'Undo', 'Edit...', and 'Save...'. A 'Filter' section allows choosing a filter ('None') and applying it. The 'Current relation' section shows 'Relation: weather', 'Instances: 14', and 'Attributes: 5'. The 'Attributes' section lists attributes: 'outlook', 'temperature', 'humidity', 'windy', and 'play'. The 'Selected attribute' section shows 'outlook' as nominal with 3 distinct values and 0 unique values. A table displays the count of each outlook value: sunny (5), overcast (4), and rainy (5). The 'Class: play (Nom)' section has a 'Visualize All' button. At the bottom, there are three colored bars (red, blue, red) labeled 5, 4, and 5 respectively. The status bar at the bottom left says 'Status OK' and the bottom right has a 'Log' button and a small bird icon with 'x 0'.

WEKA GUI

The Explorer



- Obsahuje 6 panelů:
- Preprocess panel
 - Výchozí panel
 - Nahrávání souborů
 - Filtrování
 - Prohlížení charakter. vlastností atributů



WEKA GUI

The Explorer



- **Classifier panel**
 - Konfigurace klasifikátorů
 - Klasifikace
- **Cluster panel**
 - Provádění seskupení (“clustrování”)
 - Jednotlivé clustry lze zobrazit v pop-up okénku



WEKA GUI

The Explorer



- **Associate panel**
 - “Dolování” dat
- **Select Attributes panel**
 - Konfigurace a použití kombinaci WEKA “odhadců” atributů a režim vyhledávání pro vybrání vhodných atributů z dat
- **Visualize panel**
 - Zobrazení 2D grafů



WEKA GUI



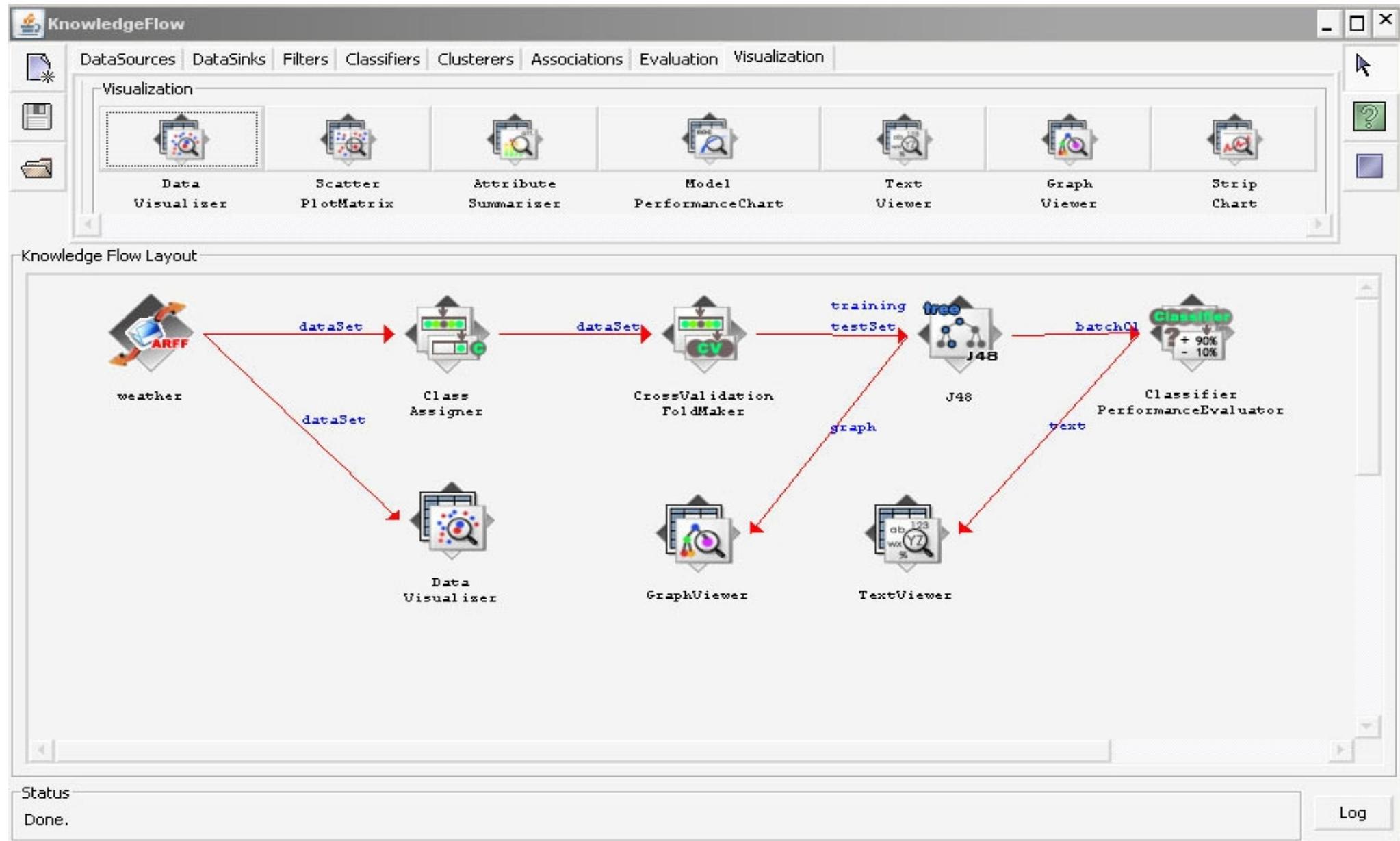
The Knowledge Flow



WEKA GUI



The Knowledge Flow



WEKA GUI

The Knowledge Flow



- Alternativa ke GUI *The Explorer* jako přední grafická část jádra algoritmů soft. WEKA
- Neustále je toto GUI ve vývoji
- Prezentuje datový tok inspirovaný WEKA rozhraním



- Uživatel si může vybrat WEKA komponenty z menu a umístit je na plátno
- Tyto komponenty mohou být na plátně vzájemně propojeny tak, aby vytvářely “tok znalostí” pro zpracování a analýzu dat
- K tomuto GUI se dnes ubírá veškerá vývojářská pozornost – **budoucnost** soft. WEKA



- Základní vlastnosti:
 - Stylové rozvržení datových toků
 - Hromadné nebo přírustkové zpracování dat
 - Zpracování vícenásobných datových proudů v dávkách nebo paralelně (každý separátní tok vykonává jeho samostatné vlákno)
 - Dokáže zobrazit více ROC křivek, což Explorer neumí



- **DataSources** – zavaděče
- **DataSinks** – uložiště
- **Filters** – filtry
- **Classifiers** – klasifikátory
- **Clusterers** – shluky, seskupení



The Knowledge Flow - Komponenty

- **Evaluation – vyhodnocení**

- ***TrainingSetMaker*** – vytvoření skupiny dat do tréninkové sady
- ***TestSetMaker*** – vytvoření skupiny dat do testovací sady
- ***CrossValidationFoldMaker*** – rozštěpení nějaké skupiny dat, tréninkové sady nebo testovací sady do drážek
- ***TrainTestSplitMaker*** – rozštěpení skupiny dat, tréninkové sady nebo testovací sady do tréninkové sady a testovací sady
- ***ClassAssigner*** – přiřazení sloupce tak, že by byl třídou pro nějakou skupinu dat, tréninkovou sadu nebo testovací sadu



The Knowledge Flow - Komponenty

- **Evaluation – vyhodnocení**

- ***ClassValuePicker*** – zvolení třídní hodnoty tak, aby byla považována za pozitivní třídu
- ***ClassifierPerformanceEvaluator*** – zhodnotit výkon skupiny trénovaných/testovaných klasifikátorů
- ***IncrementalClassifierEvaluator*** – zhodnocení výkonu přírůstkově vycvičených klasifikátorů
- ***ClustererPerformanceEvaluator*** – zhodnocení výkonu seskupené trénované/testované skupiny
- ***PredictionAppender*** - připojit klasifikátor prognózy k testovací sadě. Pro třídní problémy, může bud' přidat předpověděné třídní štítky nebo rozdělit pravděpodobnosti



The Knowledge Flow - Komponenty

● Visualization – vizualizace

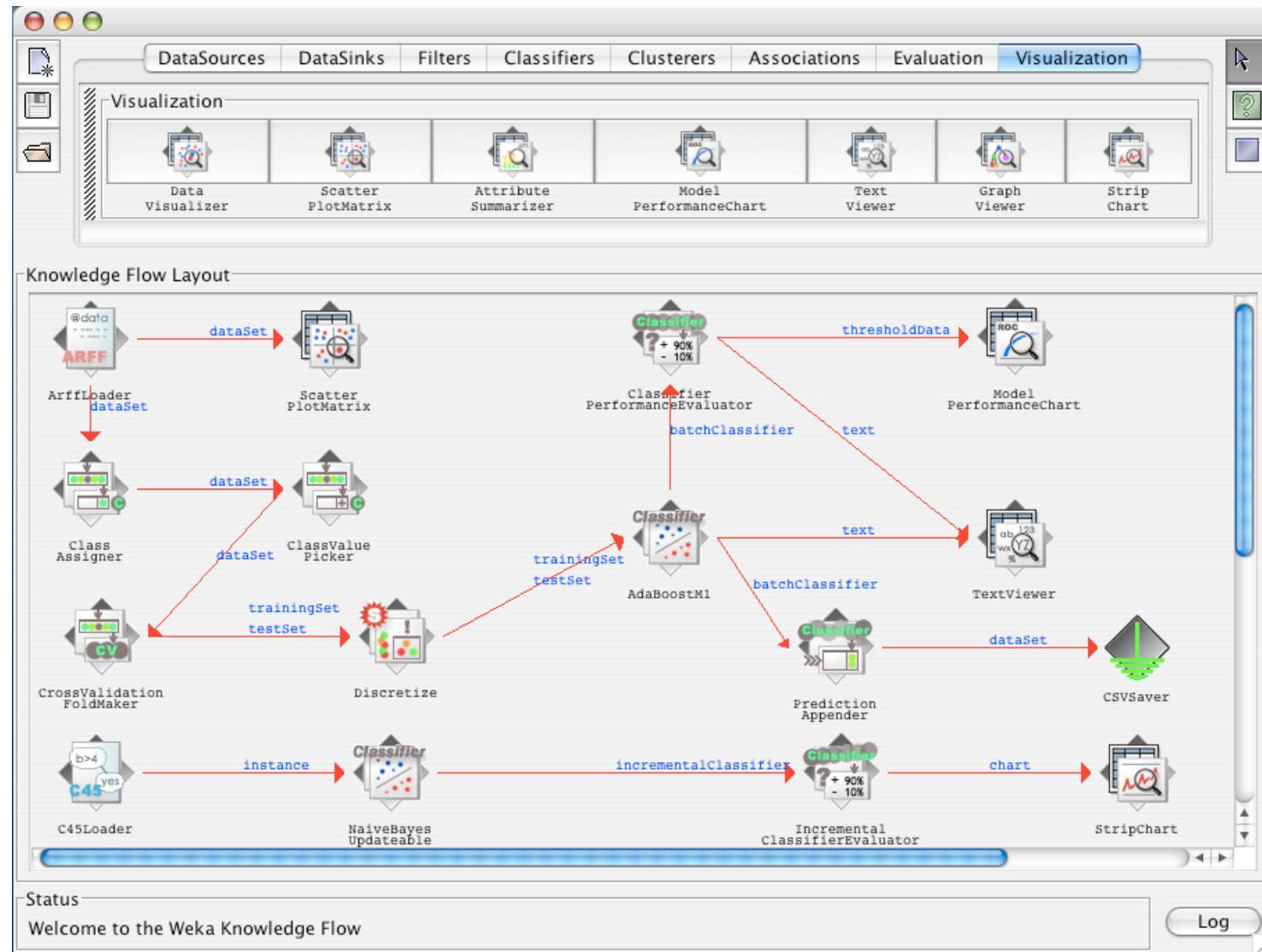
- **DataVisualizer** – komponenta, která může vyskočit jako panel pro vizualizaci dat v jednotlivě velkém 2S diagramu rozptylu(plotu)
- **ScatterPlotMatrix** – komponenta, která se může objevovat jako panel obsahující matici malého rozptylu plotu (kliknutím na malý plot vyskočí plot větší)
- **AttributeSummarizer** – komponenta, která se může objevovat jako panel obsahující matici histogramu plotů – jedna pro každý z atributů vstupních datech
- **ModelPerformanceChart** – komponenta, která se může objevovat jako panel pro vizualizaci prahové křivky (to jest ROC styl)
- **TextViewer** – komponenta pro zobrazení textových dat a může ukázat skupiny dat statistického přehledu klasifikace
- **GraphViewer** – komponenta, která se může objevovat jako panel pro vizualizaci stromu základních modelů
- **StripChart** – komponenta, která se může objevovat jako panel zobrazující rolovací *plot* dat, užívaných pro online prohlížení výkonu přírůstkových klasifikátorů



WEKA GUI

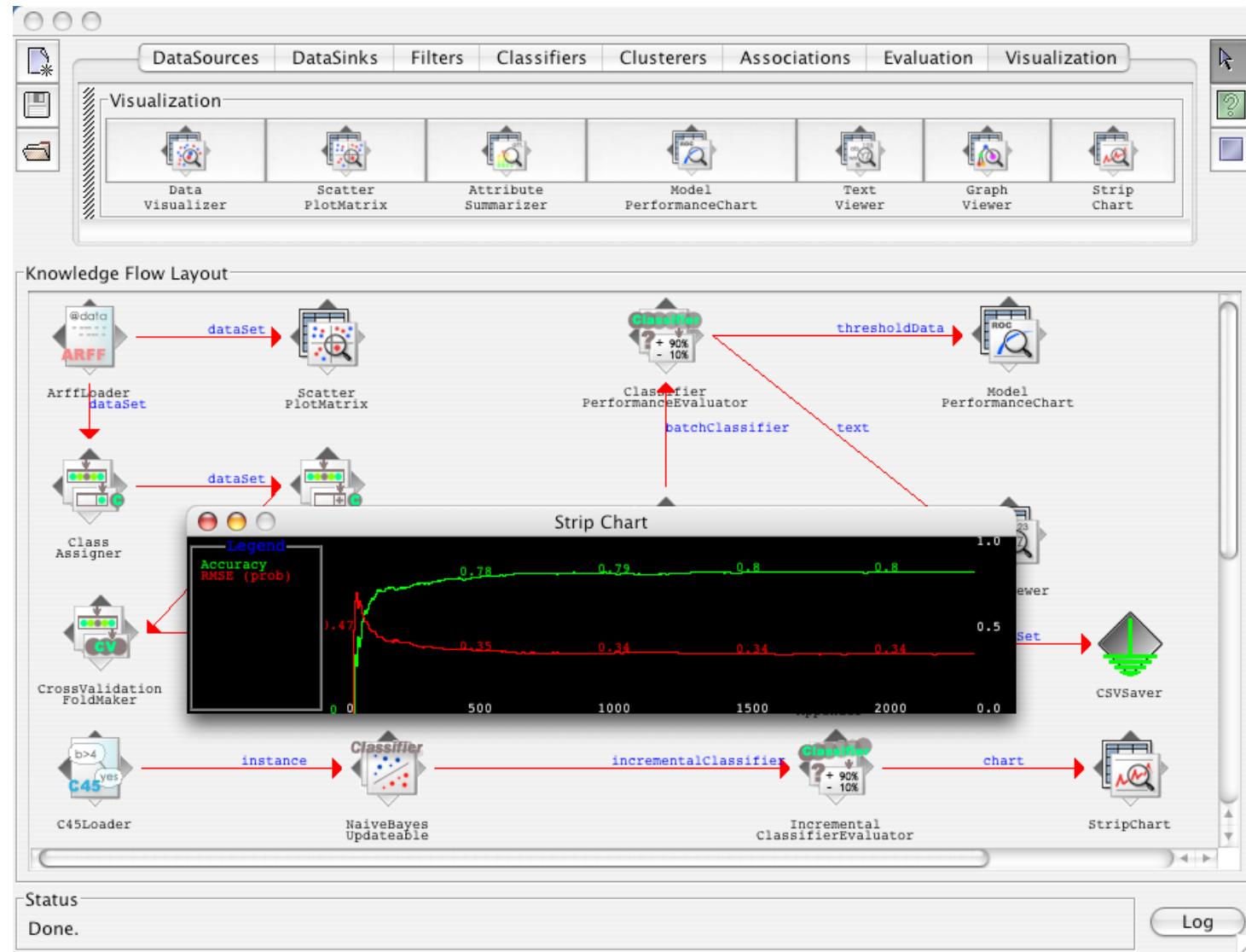


The Knowledge Flow



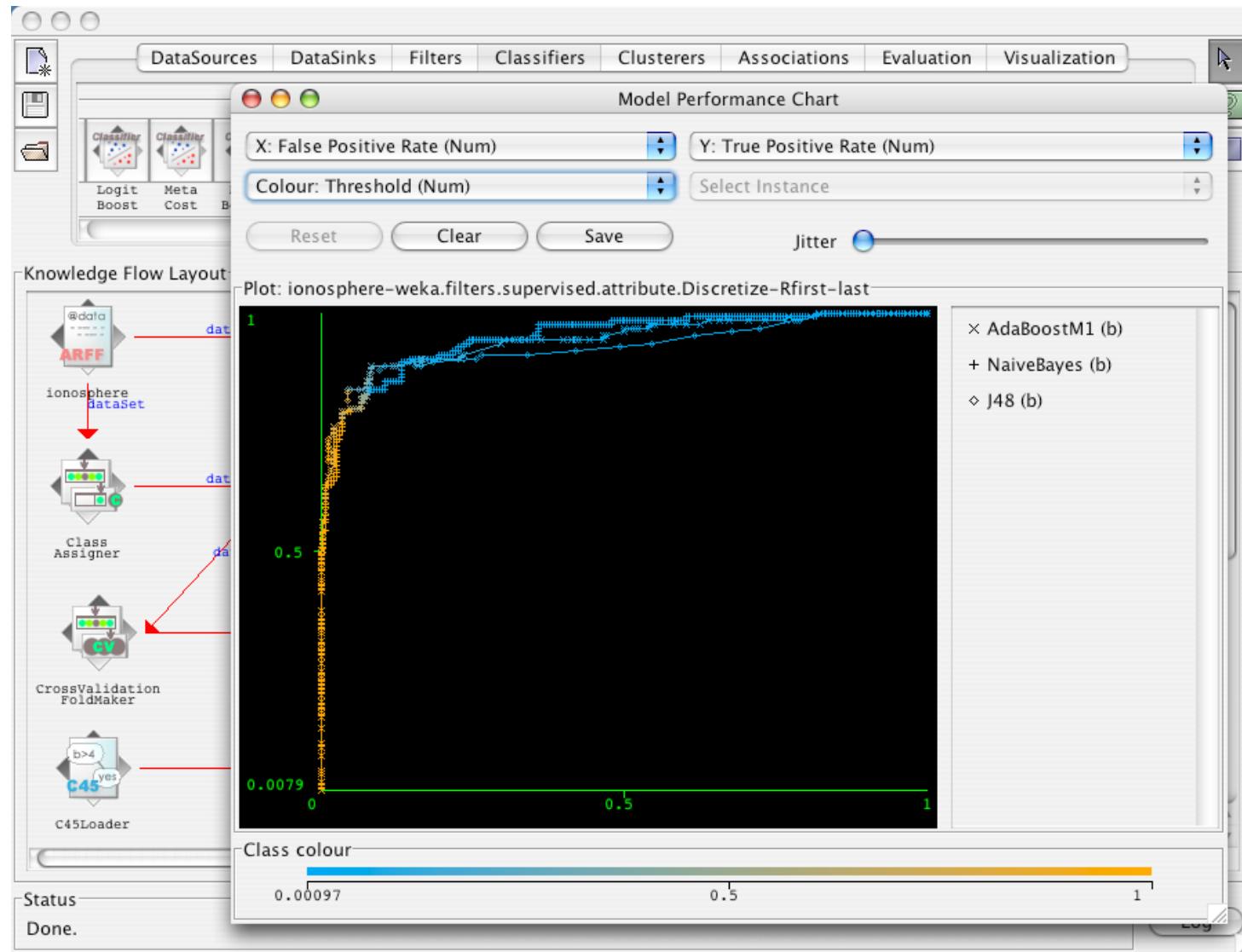
WEKA GUI

The Knowledge Flow



WEKA GUI

The Knowledge Flow



WEKA GUI



The Experimenter



WEKA GUI

The Experimenter



Weka 3.5.6 - Experimenter

Program Applications Tools Visualization Windows Help

Experimenter

Setup | Run | Analyse |

Experiment Configuration Mode: Simple Advanced

Open... Save... New

Results Destination: ARFF file

Experiment Type: Cross-validation Number of folds: Classification Regression

Iteration Control: Number of repetitions: Data sets first Algorithms first

Datasets: Add new... Edit selected... Delete selected

Algorithms: Add new... Edit selected... Delete selected

Notes: Up Down Load options... Save options... Up Down

This screenshot shows the Weka 3.5.6 Experimenter interface. The window title is 'Weka 3.5.6 - Experimenter'. The menu bar includes Program, Applications, Tools, Visualization, Windows, and Help. The main tab is 'Experimenter'. Below the tabs are 'Setup' (selected), 'Run', and 'Analyse'. The 'Experiment Configuration Mode' section has 'Simple' selected. There are three buttons: 'Open...', 'Save...', and 'New'. The 'Results Destination' section allows saving to an ARFF file with a browse button. The 'Experiment Type' section shows 'Cross-validation' selected with a dropdown arrow, and 'Number of folds' set to 5. It also includes classification and regression radio buttons. The 'Iteration Control' section has 'Number of repetitions' set to 1 and two iteration control modes: 'Data sets first' and 'Algorithms first'. The 'Datasets' section contains buttons for adding, editing, and deleting datasets, and a checkbox for 'Use relative paths'. The 'Algorithms' section contains similar buttons for managing algorithms. At the bottom, there are buttons for loading and saving options, and up/down arrows for sorting.

- Nastavení rozsáhlých experimentů
- Uvést tyto experimenty do „chodu“
- Odejít od probíhajících experimentů a po jejich dokončení se k nim opět vrátit a můžeme pak analyzovat statistiku výkonu proběhlých experimentů



- Výsledná statistika experimentu, pak může být uložena a tudíž můžeme nad námi uloženou statistikou provádět opětovně další experimenty
- Pomocí RMI jenž má WEKA implementované, lze provádět experimenty na více zařízeních
- Přístup k datům pomocí databáze (JDBC driver)



WEKA GUI



The Simple CLI



WEKA GUI

The Simple CLI



Weka 3.5.6 - SimpleCLI

Program Applications Tools Visualization Windows Help

SimpleCLI

Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of the window. Use the up and down arrows to move through previous commands.

Command completion for classnames and files is initiated with <Tab>. In order to distinguish between files and classnames, file names must be either absolute or start with '..'.

<Alt+BackSpace> is used for deleting the text in the commandline in chunks.

> help

Command must be one of:

```
java <classname> <args> [ > file]
break
kill
cls
exit
help <command>
```

|

WEKA GUI

The Simple CLI



- *Simple CLI* poskytuje přístup ke všem WEKA třídám, např. třídiče, filtry, sekupení atd, ale bez sporu s nastavením ClassPath (to jedno z ulehčených Weky od jejího vzniku).
- Nabízí jednoduché „Weka shell“ s oddělenou command-line a výstupem.
- Od verze **WEKA 3.5.6** lze provádět základní přesměrování.

java weka.classifiers.trees.J48 -t test.arff >j48.txt

- K tomu, aby WEKA vyvolávala třídy, musí být nazačatku příkaz „**java**“. Tento příkaz říká *SimpleCLI* to, jak má naložit s danou třídou a vykonat to se zadanými argumenty. Například **J48** klasifikátor může být vyvolán na iris dataset následujícím příkazem :

java weka.classifiers.trees.J48 -t c:/temp/iris.arff



WEKA projekty



WEKA projekty



WEKA GUI

BioWeka



- Projekt *BioWeka* poskytuje *framework* pro DM v rámci analýzy dat v *biologii*, *biochemii* a *bio-informatice*
- Je vyuvíjen univerzitou *Ludwig Maximilians-Universität München*
- licencí GNU GPL

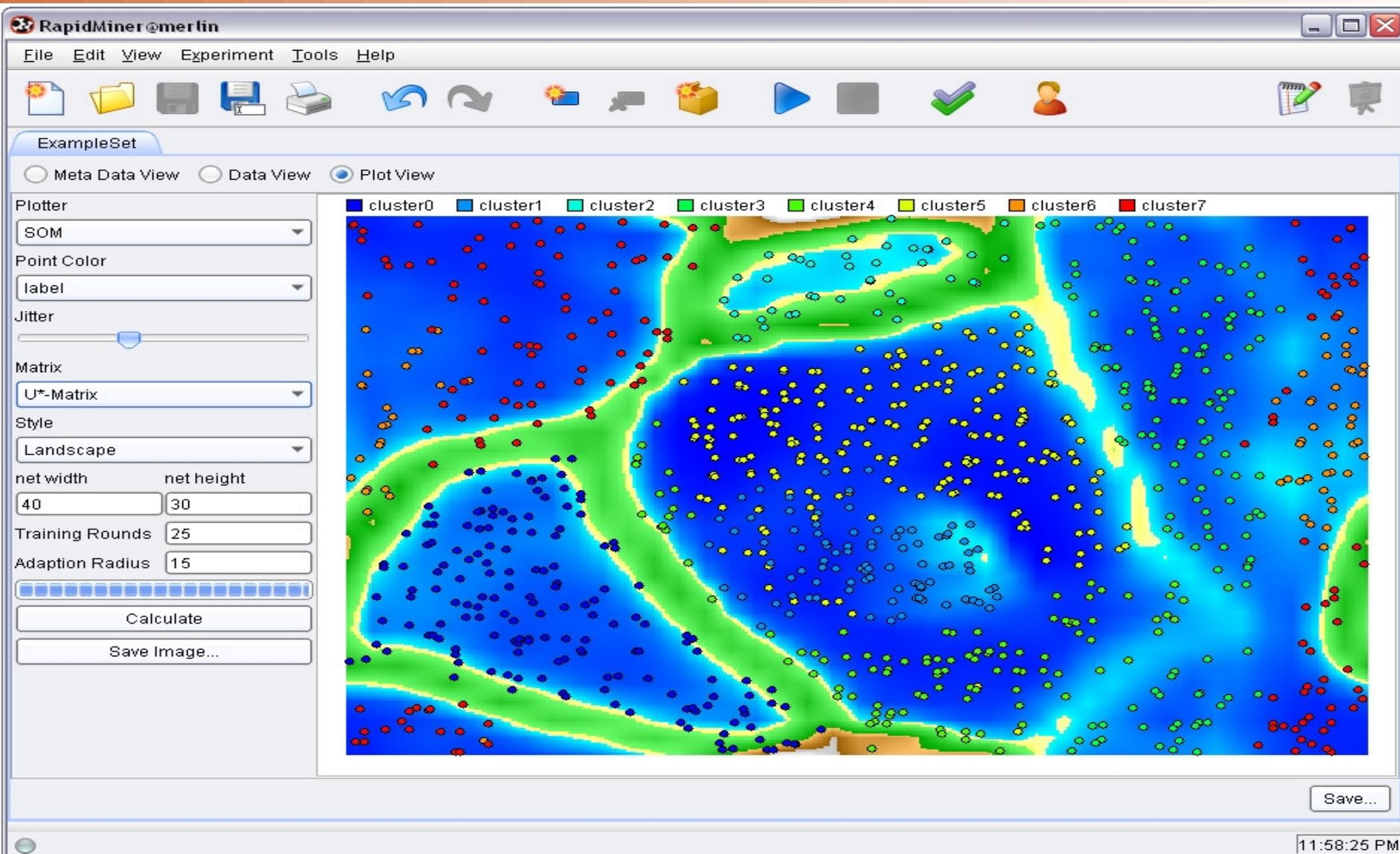


- *RapidMiner* (dříve YALE) je dle svých vývojářů předním *open source software* zabývajícím se problematikou DM. Tento projekt se může pochlubit velmi rychlými algoritmy
- Dále se chlubí velmi propracovaným GUI, zejména vizualizací dat
- *RapidMiner* však není projektem, který by nějakým způsobem WEKU rozšiřoval. Jedná se spíše o WEKU zabalenou v jiném softwaru.
- **Tři základních verze:**
 - verze s otevřeným zdrojovým kódem, distribuována pod licencí GNU GPL [4]
 - verze, která je zdarma ke stažení, ale zdrojový kód již není k dispozici
 - placená verze, která obsahuje velmi propracované GUI



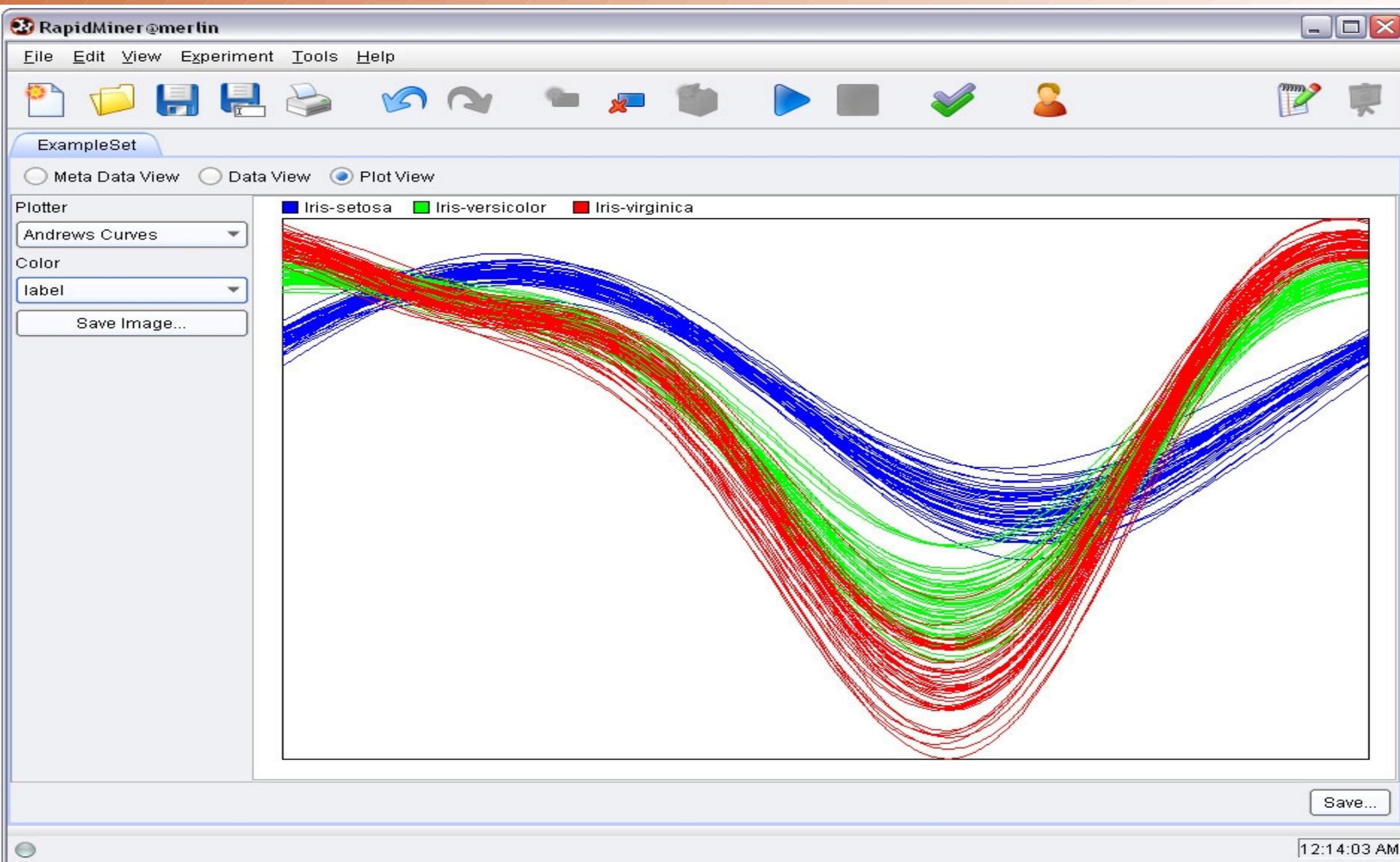
WEKA GUI

RapidMiner



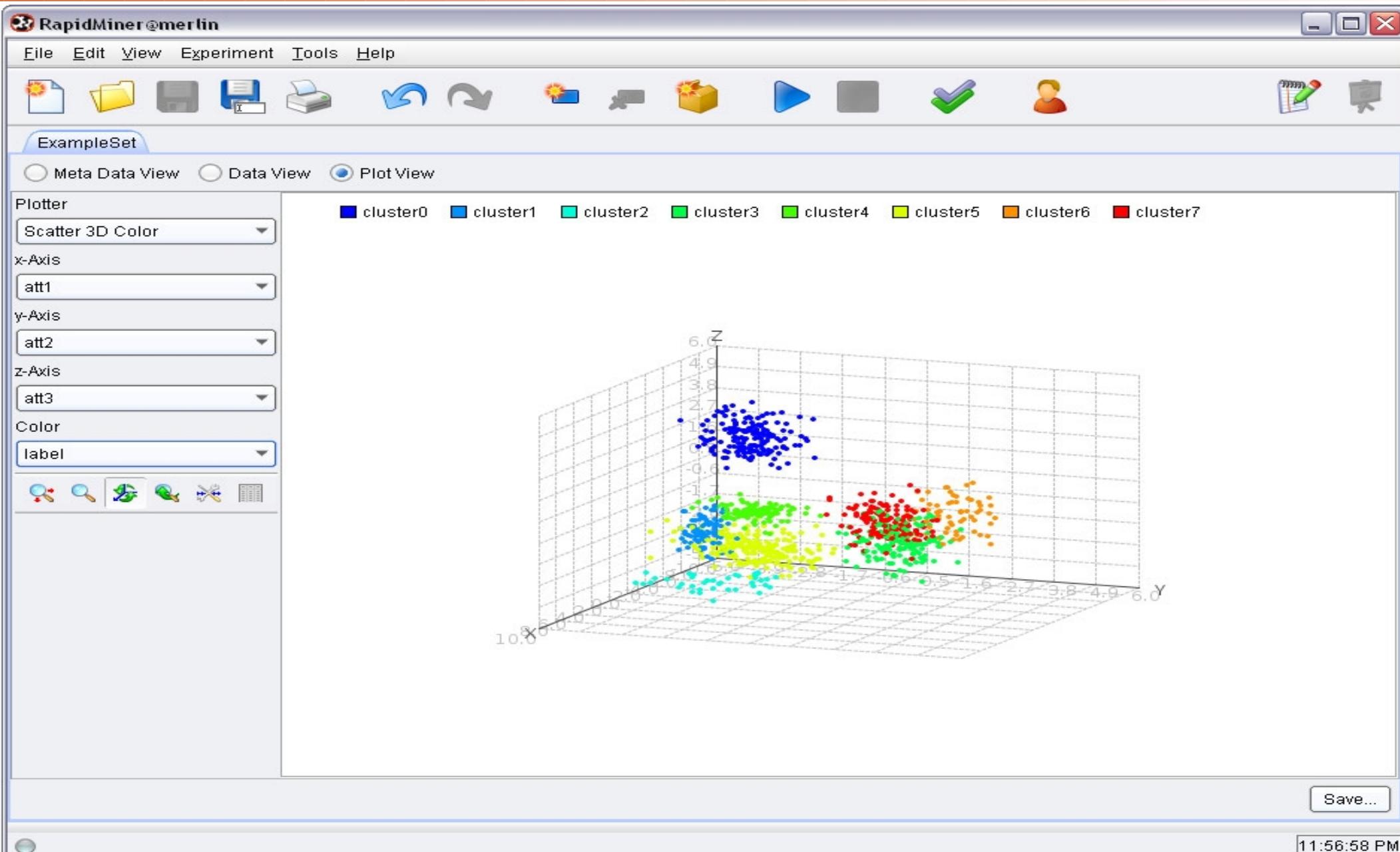
WEKA GUI

RapidMiner



WEKA GUI

RapidMiner



Už bude konec :-)



- WEKA - <http://www.cs.waikato.ac.nz/~ml/weka/>
- WekaWiki - http://weka.sourceforge.net/wekadoc/index.php/Main_Page
- BioWeka - http://bioweka.sourceforge.net/index.php/Main_Page
- RapidMiner - <http://rapid-i.com/content/blogcategory/10/69/lang,en/>
- WEKA-ERP - <http://weka-erp.origo.ethz.ch/>

