GRR

získávání znalostí v geografických datech
Autoři

Knowledge Discovery Group
Faculty of Informatics
Masaryk Univerzity
Brno, Czech Republic

GRR - Popis systému - cíle

- systém pro dolování v geografických datech
- snadno dostupný pro běžné uživatele
- snadno rozšiřitelný
- co nejméně závislý na použitém geografickém informačním systému

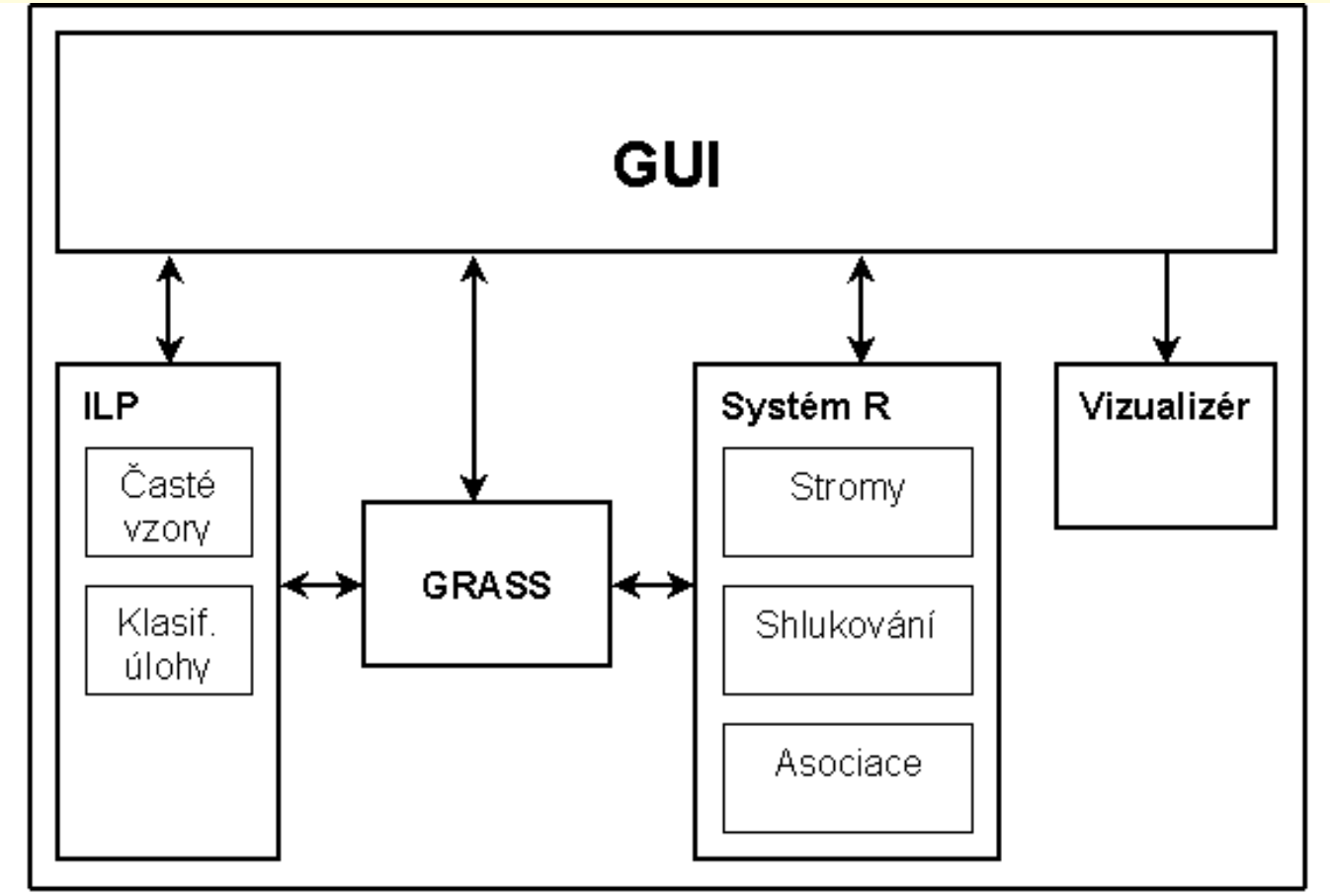
GRR - Popis systému

GRR - Systém pro dolování v geografickém informačním systému GRASS

- Struktura systému GRR
  - grafické uživatelské rozhraní
  - rozhraní pro komunikaci se systémem R
  - rozhraní pro komunikaci s vlastním geografickým informačním systémem
  - vizualizér - výsledné znalosti jsou vyjádřeny v jazyku PMML. Systém umožňuje jejich vizualizaci.

GRR - Popis systému

- Struktura systému GRR



GRR - Popis systému

- Grafické rozhraní

GRR - implementace

- implementován v jazyku Perl (verze 5.6.1)
- pro tvorbu grafického uživatelského rozhraní bylo využito Tk (verze 8.0)
- vytvořen a testován v operačním systému Linux RedHat 7.2.
- testován se systémem R verze 1.6.1.
- pro komunikaci mezi systémem GRASS a R byl použit v systému R přídavný modul GRASS (verze 0.1-11) Rogera Bivanda
- Jako vizualizátor byl použit PMML vizualizer [Wetts02] implementovaný v jazyce java

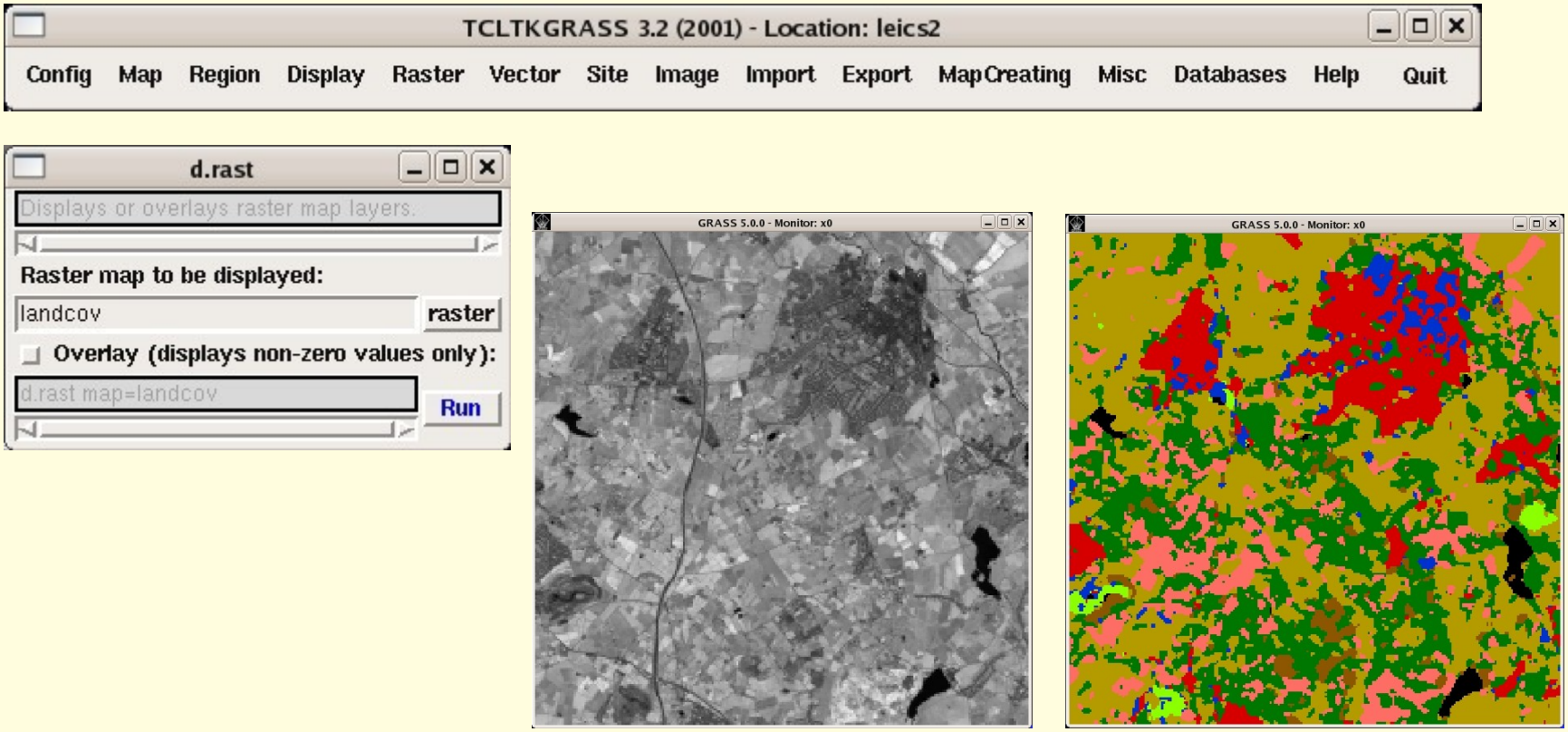
GRR - GRASS

- GRR je vystavěn nad geografickém informačním systému GRASS
- je v podstatě nádstavbou
- GRR pracuje nad daty, které máme otevřeny v systému GRASS
- dále využívá možností zobrazení vrstev v GRASSu
- GRR je odladěn nad verzí systému GRASS 5.0.0.

GRASS - Geographical Resources Analysis Support systém

- geografický informační systém (GIS)
- vyvíjený od roku 1982
- "open-source" - free software
  (v ramci projektu GNU)
- pracuje pod operačními systémy UNIX
  (na různých platformách)
  s grafickým uživatelským rozhraním X-Windows
- existuje i verze pro Windows NT/2000/XP
  (s použitím systému cygwin)
- http://grass.itc.it

- Grafické rozhraní



GRR - data

- Geografická data, která máme k dispozici pro demonstrování práce systému GRR, pokrývají severozápadní část Leicestershire v Anglii
- Jde o rastrová data pokrývající výřez velikosti 12km x 12km
- Část této oblasti je spíše nížinatá (sever a východ) a část pokrývá vysočina.

GRR - data

- V oblasti jsou:
  - dvě větší města
    - Loughborough
    - Shepshe
  - hustá síť silnic různých kategorií
  - Železnice
  - řeka (Soar)

GRR – data – jednotlivé vrstvy

- crash
  - vrstva, ve které je označeno místo nehody na dálnici
- contours
  - zde jsou zaznamenány vrstevnice procházející touto oblastí
- image
  - jde o černobílý satelitní snímek dané oblasti
- landcov
  - v této vrstvě jsou barevně odlišeny oblasti podle pokryvu resp. využití
- plant
  - zde je vyznačena čistírna odpadních vod

GRR – data – jednotlivé vrstvy

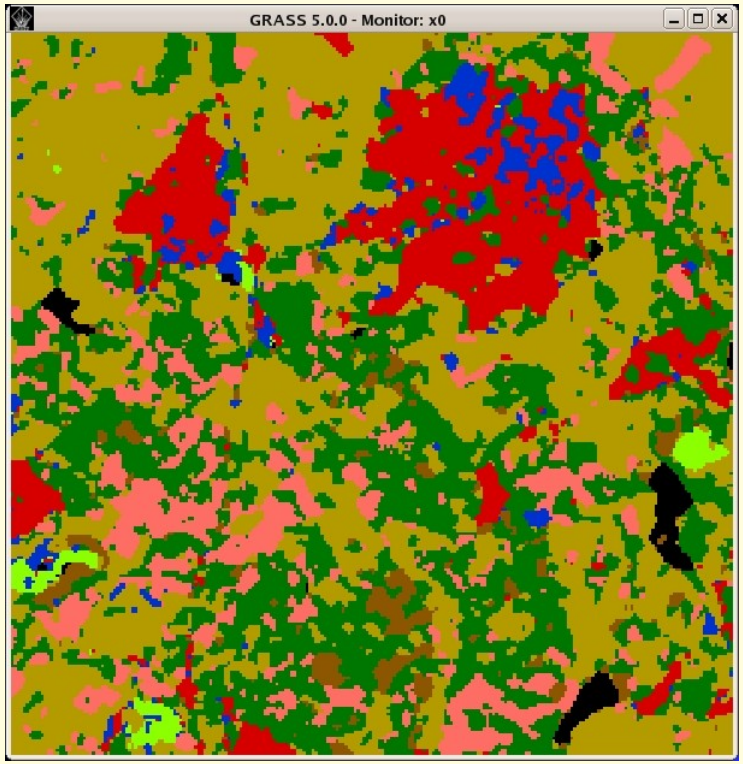
- popln
  - v této vrstvě jsou vyznačeny oblasti, které jsou zalidněny
- rail
  - vrstva, kde jsou vyznačeny železnice
- roads
  - v této vrstvě je zaznačena síť silnic, které jsou klasifikované podle typu: dálnice, plánované, silnice tř. A, silnice tř. B, silnice tř. C
- segment
  - zde je zaznamenán úsek řeky, který je znečištěn
- source
  - v této vrstvě je vyznačen zdroj znečištění

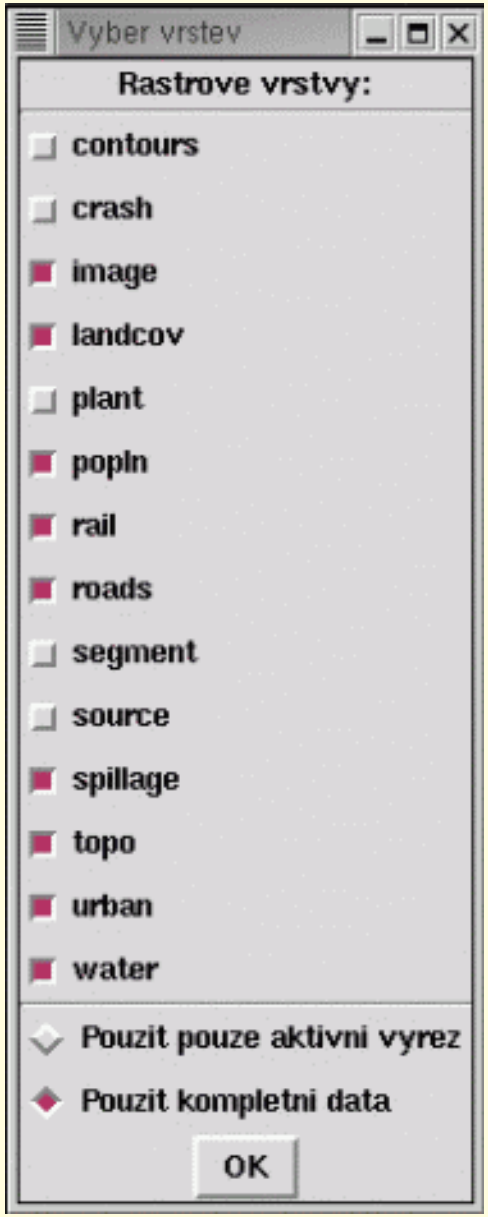
GRR – data – jednotlivé vrstvy

- spillage
  - zde je vyznačena míra rizika chemického znečištění
- topo
  - zde je pro každou buňku určena nadmořská výška (jde o digitální model)
- urban
  - v této vrstvě jsou vyznačena důležitá městská území
- water
  - zde jsou vyznačeny řeky a vodní nádrže

GRR – data – jednotlivé vrstvy

- Import vrstev do systemu GRR
- Zobrazení vrstev image a landcov v systemu Grass





GRR – data – vrstva landcov

- barevně odlišeny oblasti podle pokryvu resp. využití:
  - průmysl. obytná oblast
  - lom/důl
  - zalesněná oblast
  - orná půda
  - pastvina
  - křovinatá oblast
  - vodní plocha

GRR - R

- http://www.r-project.org
- systém pro analýzu dat a grafickou prezentaci výsledku
- volně šířitelný software v rámci projektu GNU
- pro operační systémy:
  - UNIX
  - Windows 9x/NT/2000
  - Mac OS
- Součástí systému je programovací jazyk s překladačem a systém knihoven

R

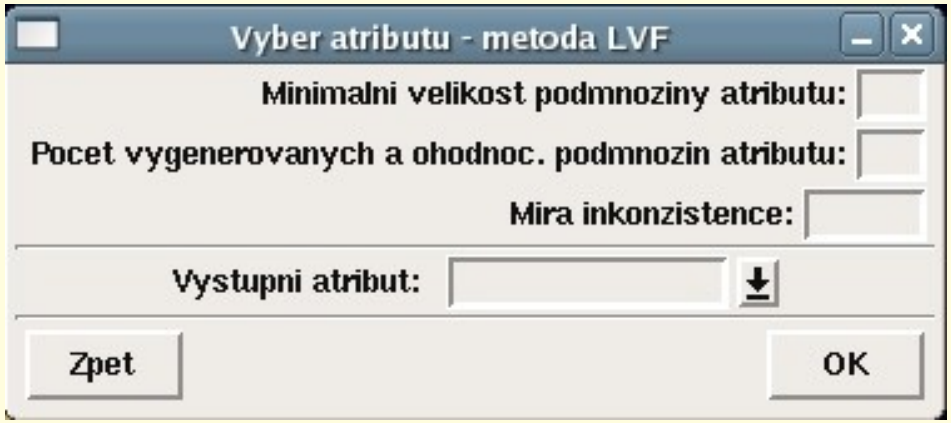
- poskytuje širokou škálu metod
  - předzpracování dat
  - statistických technik
  - algoritmů strojového učení
  - např:
    - Klasifikace
    - Shlukováni
    - lineární a nelineární modelováni
    - asociační pravidla
- Umožňuje též výsledky vizualizovat.

GRR - Metody předzpracování

- Vzorkování
  - výběr náhodného vzorku zadané velikosti
- Metody pro výběr atributů
  - LVF
- Konstrukce nových atributů
  - analýza hlavních komponent
  - hledání častých vzorů

LVF

- metoda pro výběr relevantních atributů pro klasifikaci
- implementovaná v jazyce C Filipem Procházkou
- uživatel může na základě výsledků zůžit výběr vstupních dat pro další práci



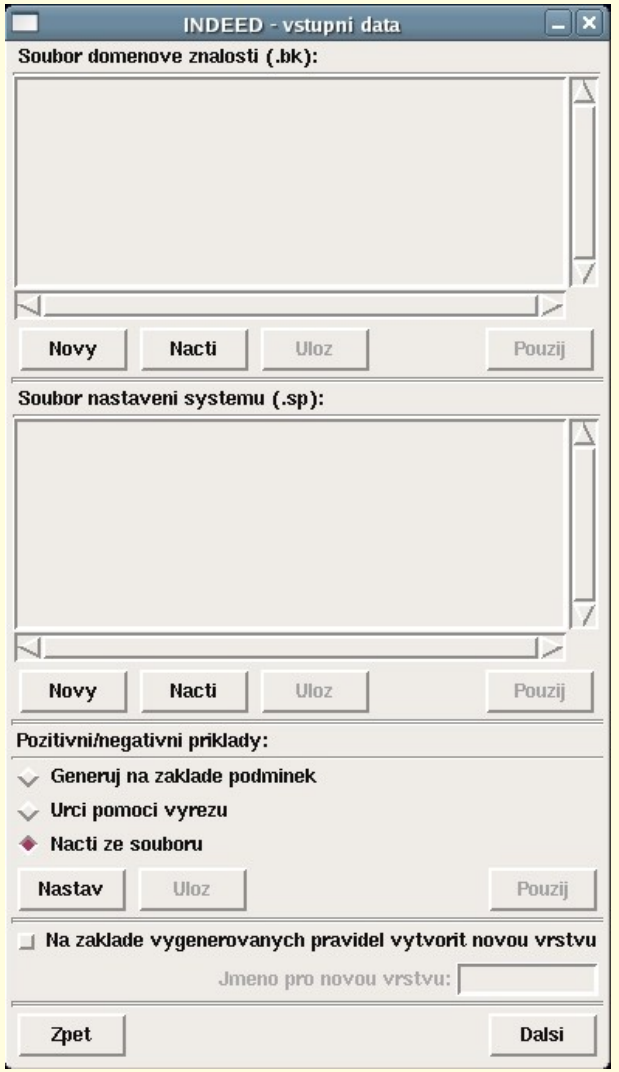
GRR -Metody dolování

- klasifikace dat pomocí rozhodovacích stromů
- metody pro hledání shluků
  - aglomerativní i hierarchická metoda
- asociační pravidla
- ILP
  - klasifikační pravidla, časté vzory a asociační pravidla
- weka

- V rámci GRR byl využit na dolování klasifikačních pravidel
- Základní postup výpočtu systému:

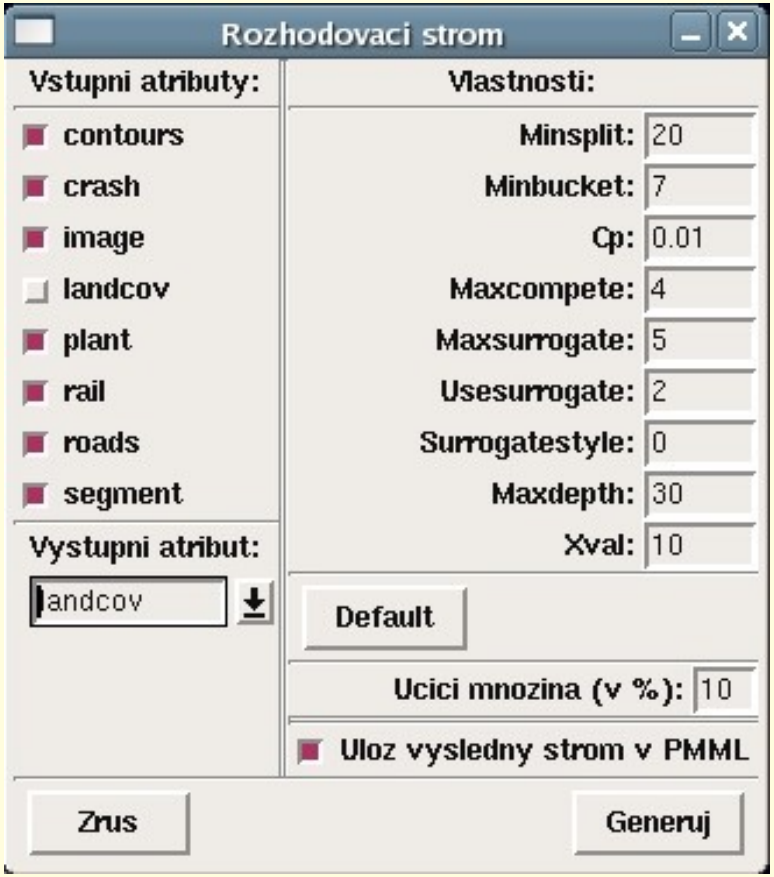
1. *Vyber příklad, který má být generalizován. Pokud žádný neexistuje, skonči, jinak pokračuj dalším krokem.*
2. *Zkonstruuj nejspecifičtější klauzuli, která pokrývá vybraný příklad a splňuje daná jazyková omezení. Obvykle jde o jednoznačnou klauzuli s mnoha literály, označovanou jako tzv. „bottom clause" (spodní klauzule). Tomuto kroku se někdy říká saturace.*
3. *Najdi obecnější klauzuli než je spodní klauzule. Toto je provedeno prohledáním některých podmnožin literálů, které mají ve spodní klauzuli „nejlepší" skóre. Tento krok se někdy označuje jako redukční krok.*
4. *Odstraň redundantní. Klauzule s nejlepším skóre je přidána do aktuální teorie a všechny příklady, které se staly redundantními, se odstraní.*
5. *Vrať se na krok 1.*

- Odvozuje klasifikační pravidla v podobě Hornových klauzulí
- Napsán v jazyce Prolog
- Vyvinut na FI MU



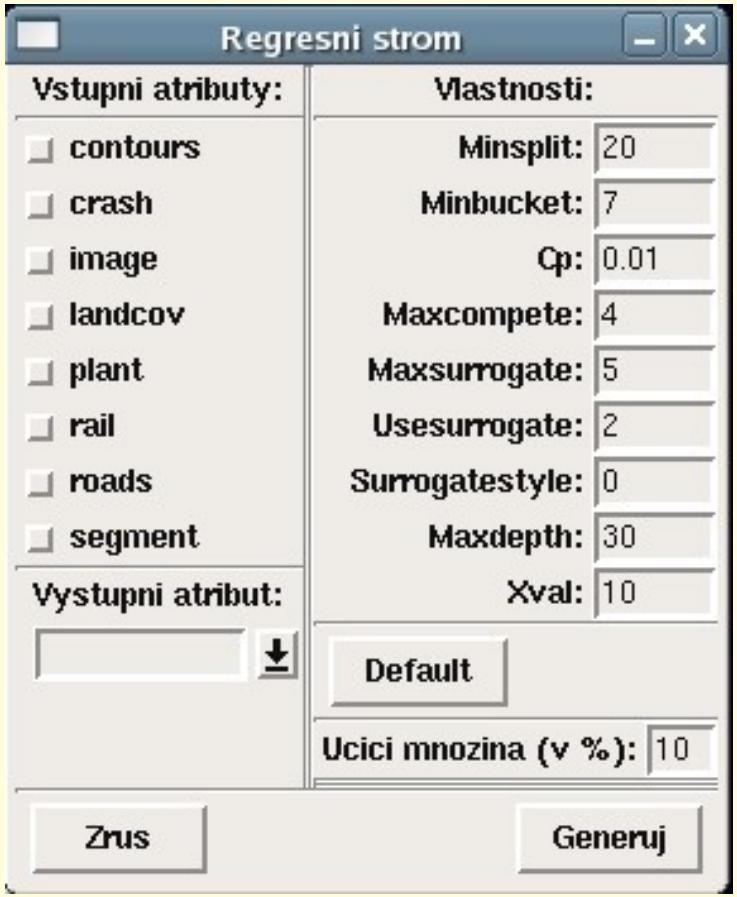
Rozhodovací strom

- cílem je pro vstupní atributy a výstupní atribut (diskrétní) nalézt, model pomocí kterého můžeme klasifikovat data
- Implementováno pomocí externí knihovny rpart systému R

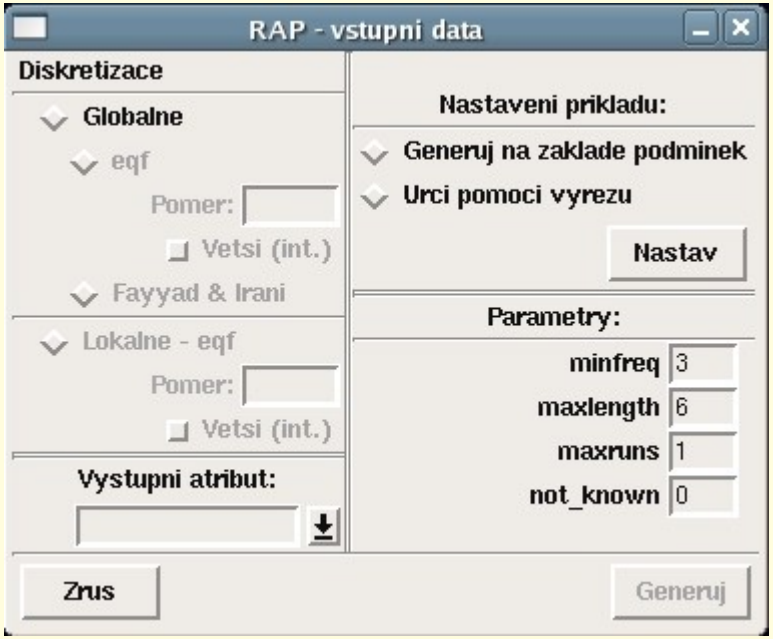


Regresní strom

- narozdíl od rozhodovacího stromu má výstupní atribut spojitý
- také využívá knihovnu rpart systému R



- Nástroj na hledání častých vzorů v datech
- Napsán v jazyce Prolog
- Vyvinut na FI MU
- Algoritmus výpočtu:



1: **dokud** vzor Q není maximální **dělej**
2: vytvoř všechna možná rozšíření *AllNewQ* vzoru Q
3: vyber *NewQ* z rozšíření *AllNewQ*
4: ověř, zda *NewQ* není mezi známými (již nalezenými) vzory
5: ověř, zda *NewQ* není mezi nefrekventovanými vzory
6: spočítej frekvenci *NewQ*
7: za Q vezmi *NewQ*

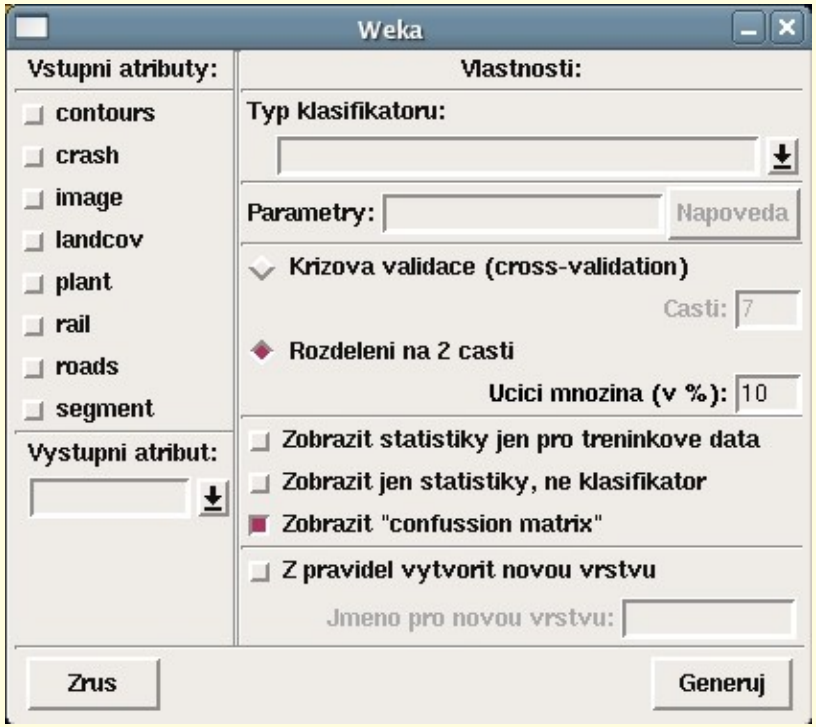
Clara

- principem algoritmu je nalézt shluky podobných záznamů
- implementována v externí knihovně cluster systému R

Autoclass

- automatický klasifikační systém, založený na metodách učení bez učitele
- k určování tříd používá Bayesovský přístup

- Systém vyvinutý na University of Waikato na Novém Zélandu
- souhrn knihoven v Javě
- implementované nástroje umožňují:
  - předzpracování dat
  - klasifikaci
  - regresi
  - shlukování
  - hledání asociačních pravidel

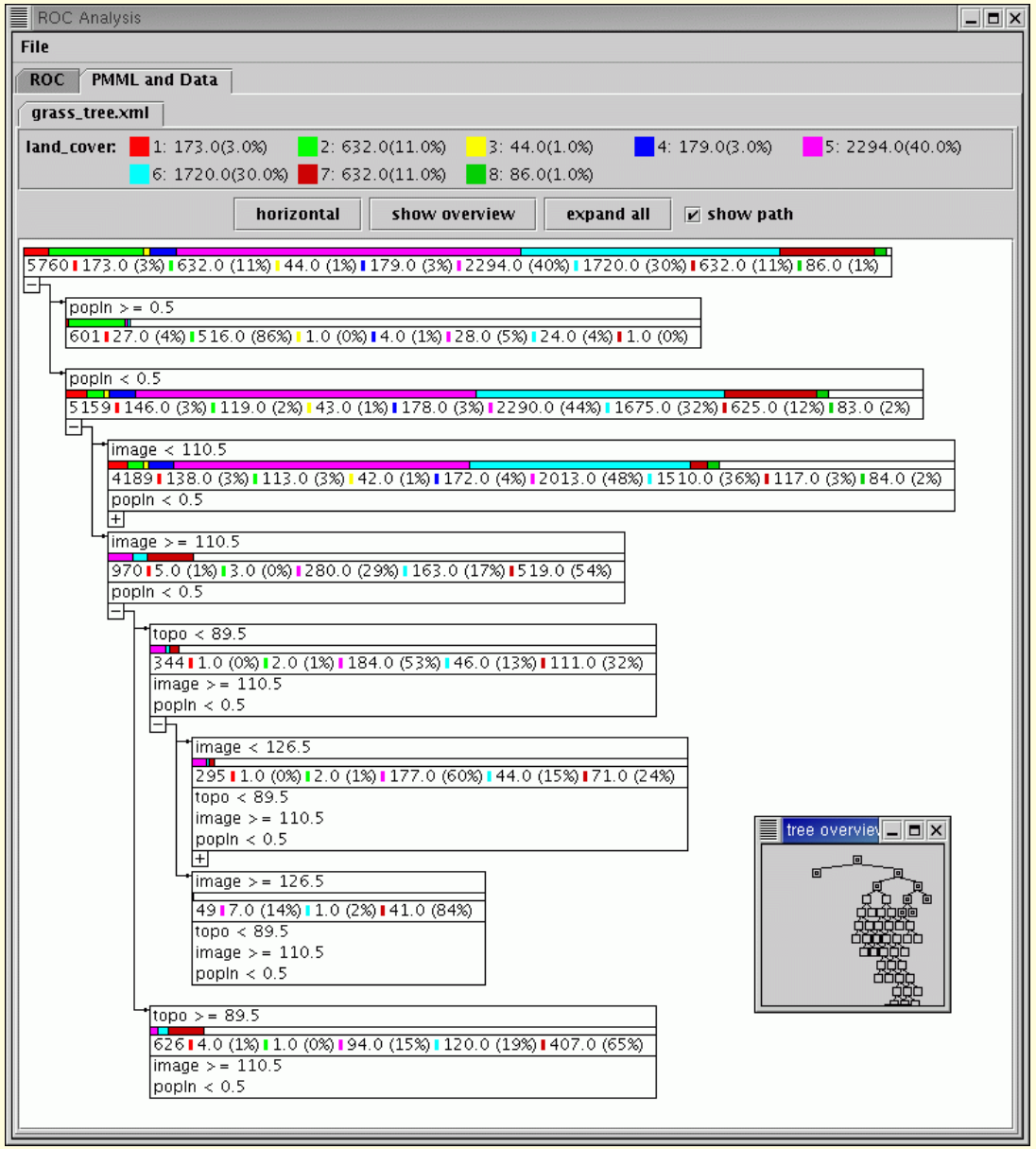


GRR - Vizualizace PMML

- PMML - Predictive Modeling Markup Language
- standard pro reprezentaci výsledků data mining
- http://www.dmg.org/pmml-v1-1.htm
- podporován jak nekomerčními systémy pro vyhledávání znalostí v databázích (Knowledge Discovery Support Engines, KDDSE) tak komerčními firmami jako je IBM. Oracle, SAS a SPSS
- výsledky dolování lze vizualizovat pomocí PMML-Vizualizéru http://soleunet.ijs.si/website/other/pmml.html

GRR - Vizualizace PMML

- Vizualizace rohodovacího stromu



- Více na: http://www.fi.muni.cz/kd/projects/grr/