Strojno učenje

Vježbe 2: Selekcija (probir) varijabli i napredno korištenje

Zašto probir značajki (engl. *Feature selection*)?

- Značajka (engl. *feature*) == **atribut**, varijabla
- Probir (engl. *selection*) == odabir, selekcija
- Dodavanje **irelevantnog atributa** može znatno smanjiti performanse modela
  - k-NN: potreban broj primjera za učenje eksponencijalno raste s brojem irelevantnih atributa
  - Naivni Bayes: nema problema, ali…
- Relevantni, ali **redundantni atributi** također mogu biti štetni:
  - vrijeme izvođenja algoritma za učenje
  - složenost modela
  - ..

Metode probira značajki

- Ugradbene (engl. *embedded*) metode
- Filter metode
- Metode “omotača” (engl. *wrapper*)

Embedded metode

- Neki algoritmi strojnog učenja inherentno uključuju probir atributa kao sastavni dio svoga procesa učenja
  - Npr. stabla odlučivanja
- Budući da su *embedded* metode sadržane u samom algoritmu učenja, više nas zanimaju metode probira atributa kao zasebni korak pretprocesiranja prije učenja

Filter metode

- Druga ideja zasniva se da se ne koristi algoritam za učenje u samom probiru već se relevantnost atributa zasniva na nekom drugom kriteriju
- Prednosti:
  - Brzina
  - Mogućnost rangiranja atributa
- Rangiranje atributa ← *attribute weighting*

Wrapper metode

- Ideja ovih metoda zasniva se na korištenju nekog algoritma učenja za ocjenu koliko su „dobri” (evaluaciju) različiti podskupovi atributa
- Iscrpnom ili heurističkom pretragom nalazi se optimalni podskup atributa koji omogućuje konstrukciju modela s najboljom kvalitetom predikcije
- Prednosti:
  - evaluiran je čitav proces učenja nad podskupnom atributa
  - može se koristili bilo koji algoritam za učenje unutar omotača
- Nedostatci:
  - Skupo i sporo: proces učenja (koji je sam po sebi glavni dio cijelog procesa) izvodi se prilikom probira mnogo puta

RapidMiner (RM)

- Probir atributa u RM-u
  - http://rapid-i.com/wiki/index.php?title=Feature_selection
  - Video predavanja:
    - http://www.youtube.com/watch?v=JlhoTAk1ow8
    - http://www.youtube.com/watch?v=qsr-SwS776I&noredirect=1
    - http://www.youtube.com/watch?v=0UdOIF3dvR8
- RM proširenje (engl. Extension) za odabir atributa
  - http://sourceforge.net/projects/rm-featselext/

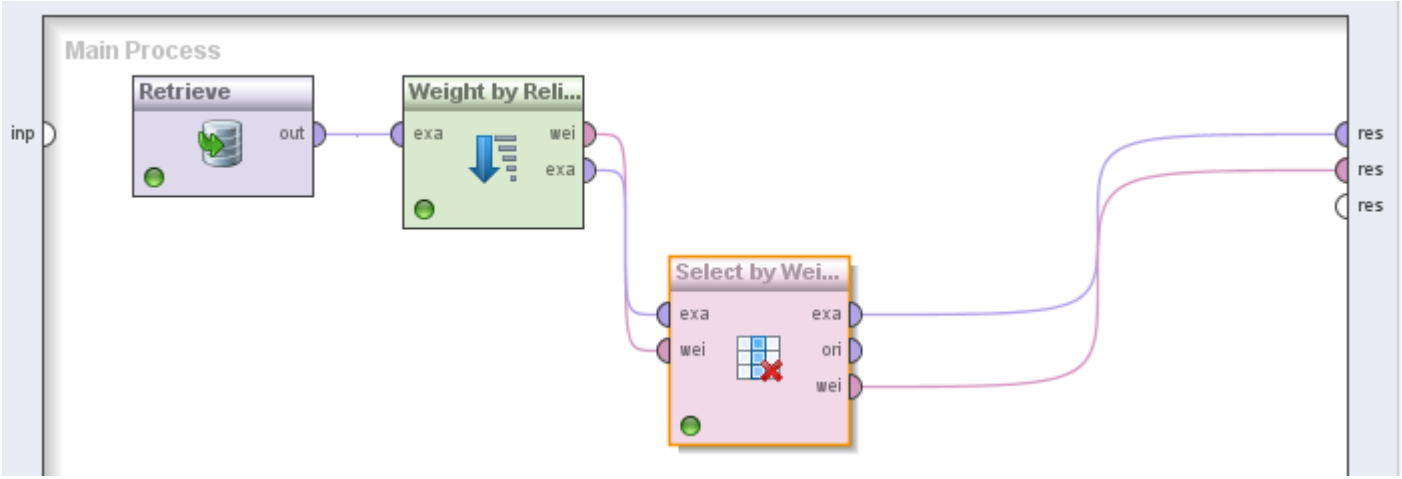
Attribute Weighting u RM

- RapidMiner pruža mnogo različitih operatora za *Attribute Weighting:*
  - *Weight by Relief*
  - *Weight by Information gain* (prisjetiti se stabla odlučivanja)
  - *Weight by Information Gain Ratio*
  - *Weight by …*
- Rezultat je skup težina atributa koje mogu biti sortirane (rangiranje atributa), vizualizirane ili npr. korištene u operatoru *Select by Weights*
- Operatori izračunaju vektor težina atributa za dani skup primjera pri čemu ga ne mijenjaju
- Kasnije atributi mogu biti filtrirani po tim težinama koristeći operator *Select by Weights* (ili skalirani koristeći operator *Scale by Weights*)

Filter metode u RM

- Filtar metode u RM → operator *Select by Weights*
  - 1. korak: Određivanje težina atributa korištenje m operatora *Weight by …*
  - 2. korak: Filtriranje atributa koje ne ispunjavaju određeni krtiterij težine korištenjem operatora *Select by Weight*
- Prednosti:
  - Brzina
- Nedostaci:
  - Nema interakcije između atributa prilikom filtriranja
  - Ne ovisi o korištenoj metodi učenja

RM primjer 1:
Attribute Weighting i Filter metoda

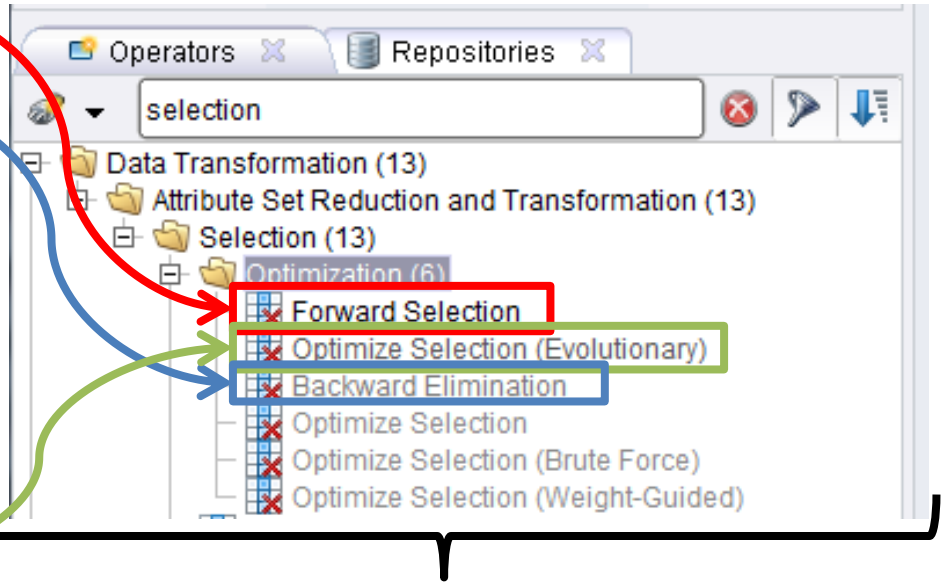


Pretraživanje podskupova atributa

- Problem: broj podskupova atributa eksponencijalno zavisi o ukupnom broju atributa -> iscrpno je neostvarivo
- Uobičajene pohlepne strategije (engl. greedy):
  - *Forward elimination*
  - *Backward selection*
- Naprednije strategije:
  - *Bidirectional search*
  - *Best-first search*
  - *Beam search*
  - *Genetic algorithms*

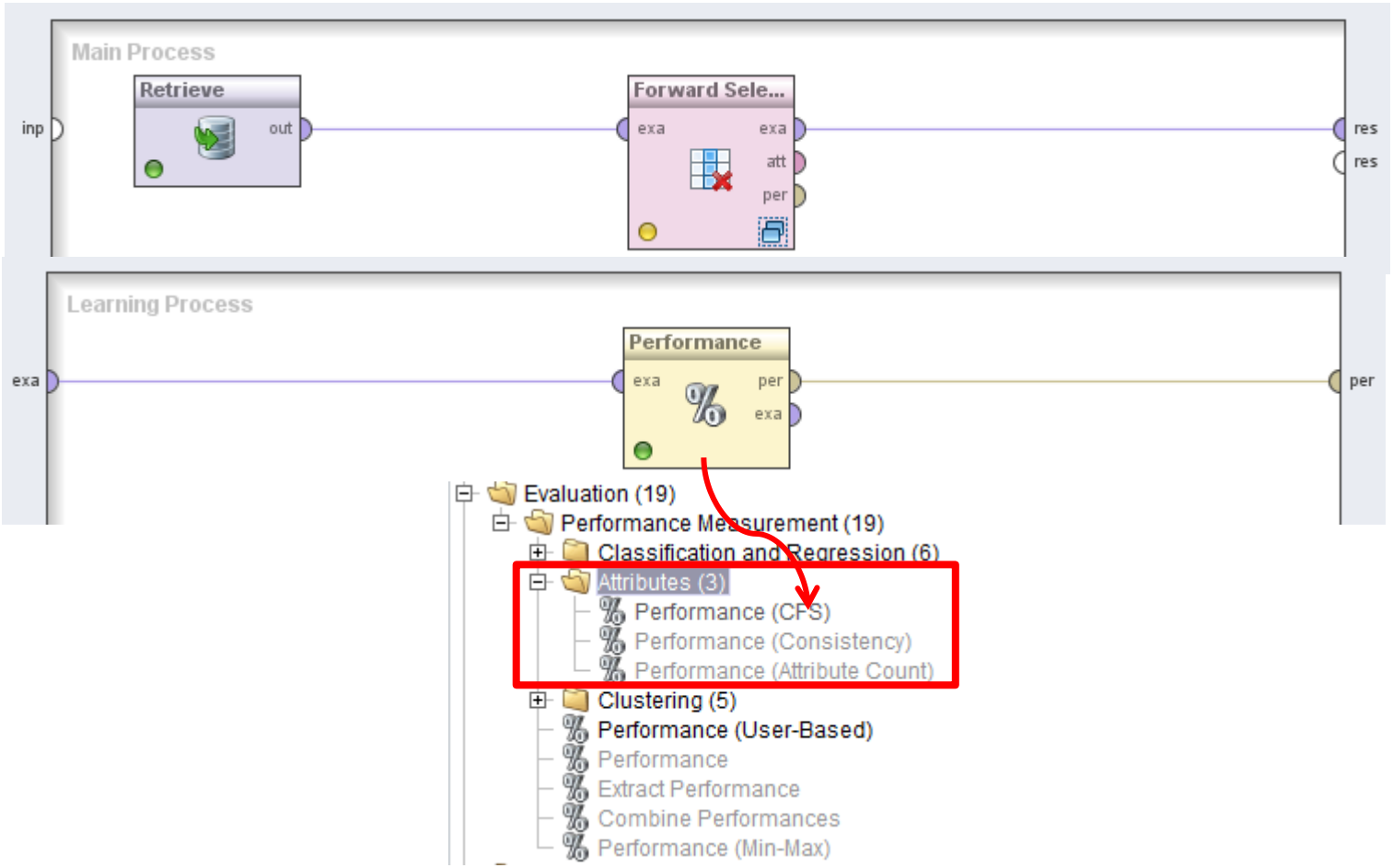
Pretraživanje podskupova atributa u RM

- Problem: broj podskupova atributa eksponencijalno zavisi o ukupnom broju atributa -> iscrpno je neostvarivo
- Uobičajene pohlepne strategije (engl. greedy):
  - *Forward elimination*
  - *Backward selection*
- Naprednije strategije:
  - *Bidirectional search*
  - *Best-first search*
  - *Beam search*
  - *Genetic algorithms*



Pretraživanje podskupa atributa u RM

RM Primjer 2: *Forward selection* kao filter

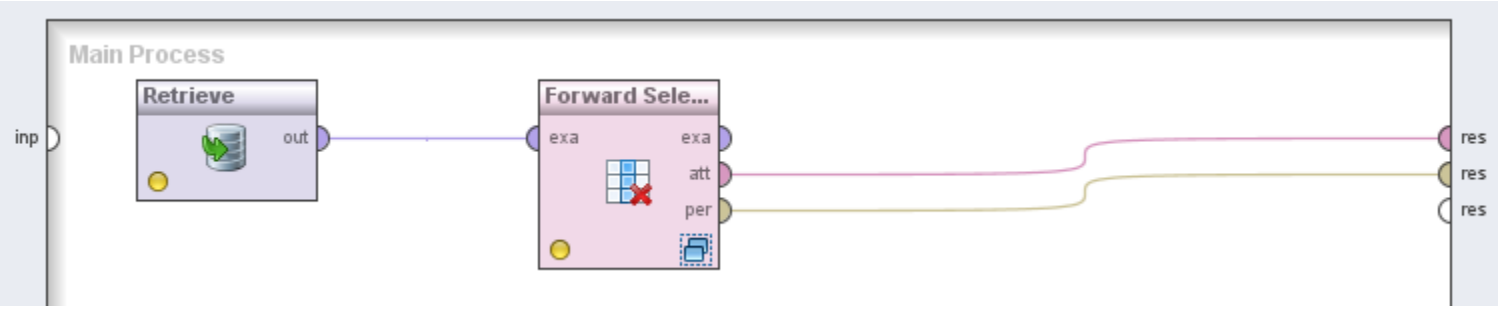


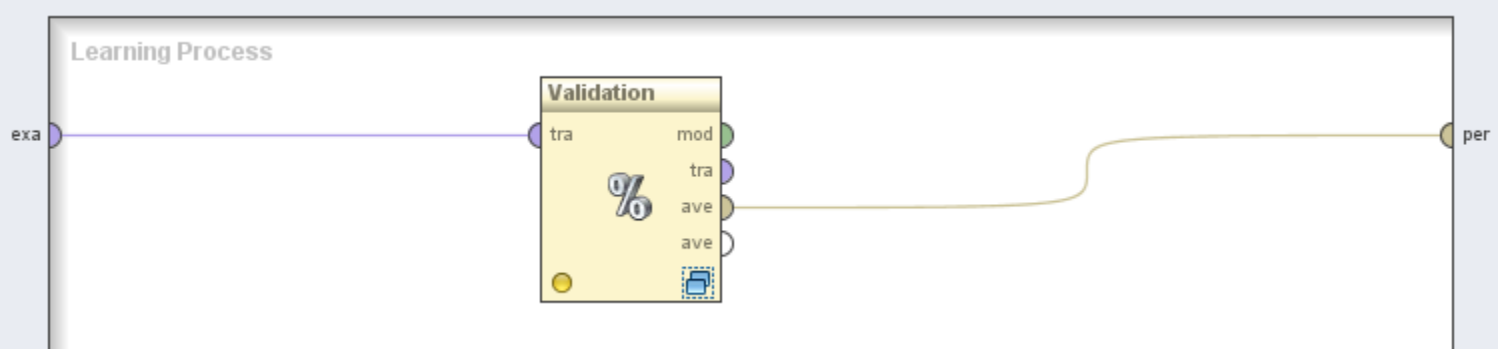
"Trick"

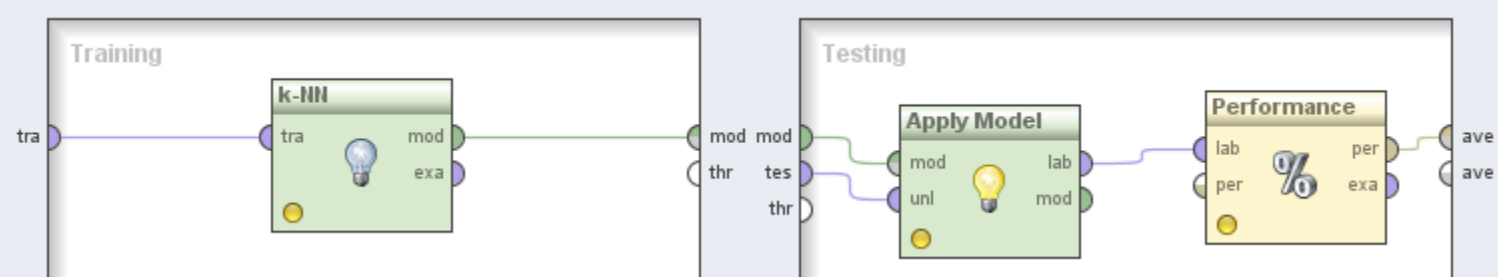
Wrapper metode u RM

- Evaluacija nekog modela učenja unutar nekog operatora za pretraživanje
  - Evaluacijski kriterij: unakrsna validacija (XV) za neki model strojnog učenja (stoga, ovisi od metodi za učenje)
- Može se koristiti bilo koji operator optimizacije pretraživanja podskupova atributa
  - Forward/Backward elimination, Evolutionary (RM impl. Genetski algoritam), …
  - strojnog učenja (stoga, ovisi od metodi za učenje)
- Efikasno samo za brze metode učenja
  - npr. Naivni Bayes

RM primjer 3: *Wrapper* sa *Forward Selection* operatorom

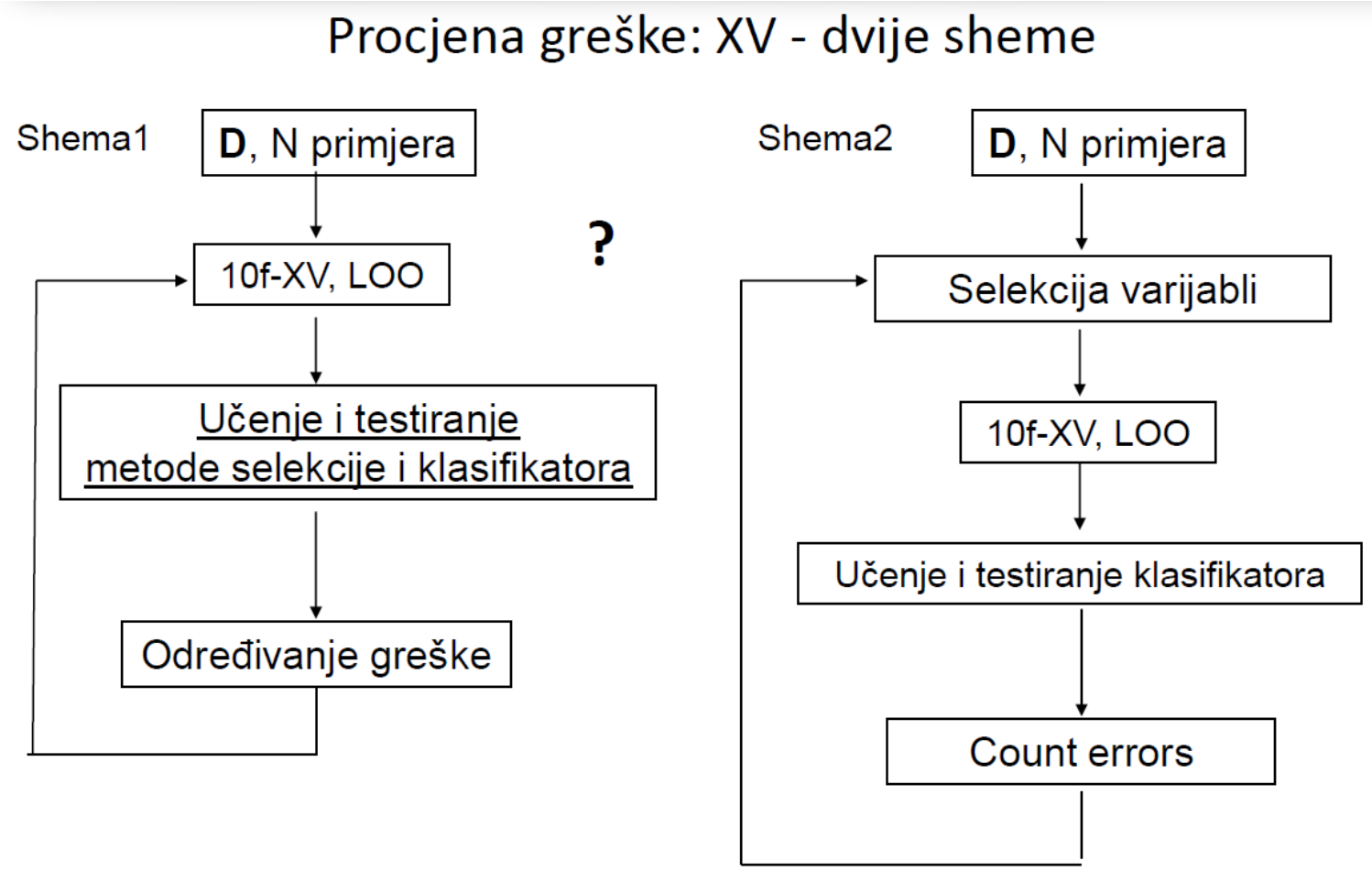






Evaluacija metoda probira atributa

Slajd s predavanja!



Evaluacija probira atributa u RM

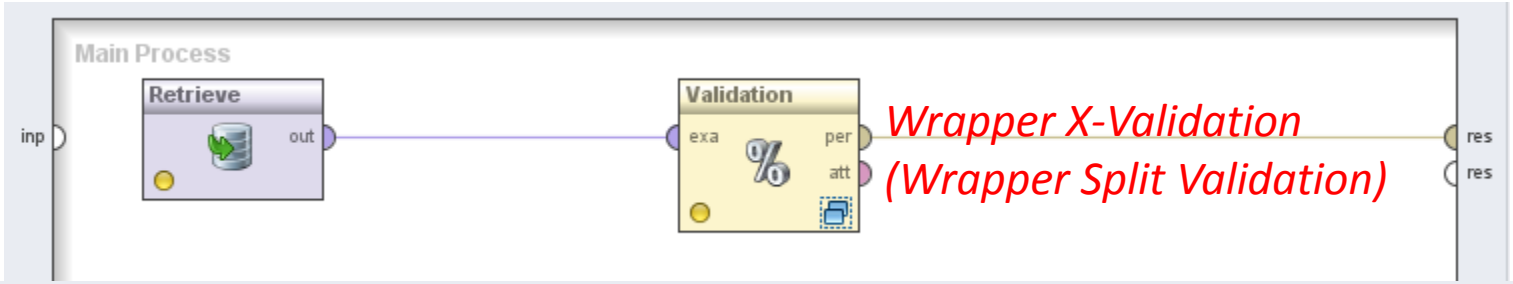
- Korištenjem ugnježđenih operatora za pretraživanje podskupova atributa
- Specijalizirani ugnježđeni operatori:
  - *Wrapper Split Validation*
  - *Wrapper-X-Validation*
- 3 podprocesa:

1. Odabir atributa (*weighting/filtering* ili *wrapper*)
2. Učenje modela
3. Evaluacija modela (na podskupu podataka za testiranje)

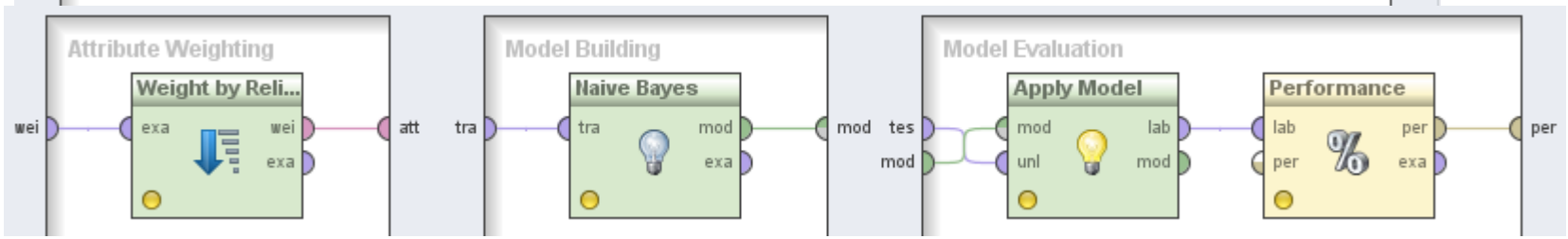
Izgradnja modela

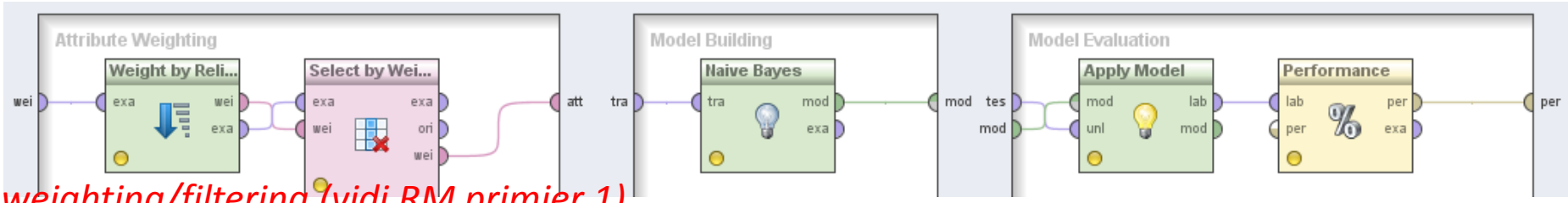
Evaluacija modela

RM primjer 4: *Wrapper X-Validation*

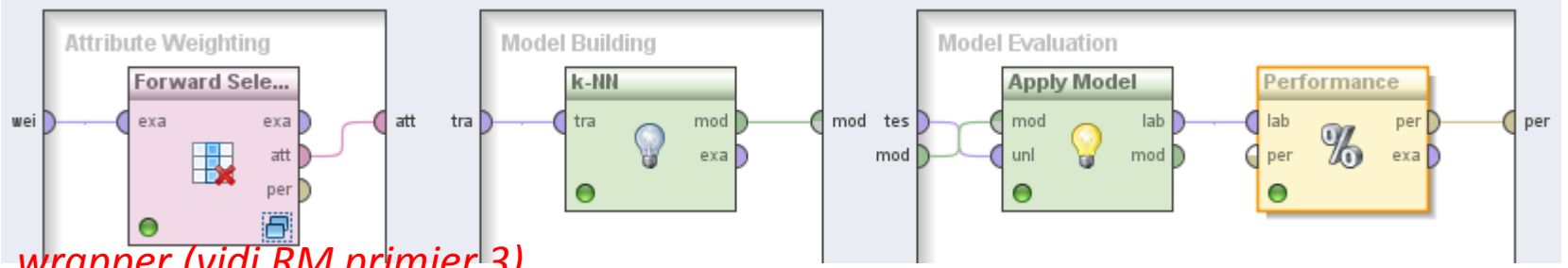


Wrapper X-Validation
(Wrapper Split Validation)





weighting/filtering (vidi RM primjer 1)



wrapper (vidi RM primjer 3)

RM Vježba 1

1. Prikažite skup primjera *"Iris"* iz *sample* repozitorija u *scatter plot*-u: Pomoću kojih atributa čovjek može najbolje i najlakše predviđati labelu?

2. Primijenite neki od operatora *Weight by …*: Koje težine oni pridružuju atributima? Je li te vrijednosti odgovaraju s vašim razmišljanjima iz 1.

3. Primijenite *wrapper* metodu: iskoristite *Forward Selection* i Backward Elimination u kombinaciji s unakrsnom validacijom i metodom učenja po vašem izboru. Koristite *breakpoins* kako bi promotrili poredak u kojem se atributi dodaju i oduzimaju.

4. Također primijenite Optimize Selection (Evolutionary) operator (koji implementira genetski algoritam).

RM Vježba 2

1. Učitajte skup primjera *“Sonar”* iz *sample* repozitorija.

2. Iskoristite genetski algoritam za pretraživanje atributa uz evaluaciju s unakrsnom validacijom modela linearne regresije koja sama interno ne probira atribute, s 10 generacija

3. Dodajte operator *Log* nakon unakrsne validacije i logirajte generaciju, trenutnu najbolju performansu i ukupnu najbolju performansu.

4. Analizirajte rezultate

Automatska optimizacija parametara u RM (ponavljanje)

- RM ima na stotine operatora
- Neki od njih mogu imati više parametara
- Kako naći najbolji odabir parametara?
- Klasični primjeri:
  - k za k-NN
  - C i epsilon za SVM

Automatska optimizacija parametara u RM (Ponavljanje)

- RM operator: *Loop Parameters*
  - Ručno odaberemo parametre i njihove pripadne vrijednosti koje želimo testirati
  - U svakoj iteraciji ugnježđeni proces u operatora koristi drugu kombinaciju parametara
  - Nedostatak je da sami trebamo ocijeniti koje vrijednosti parametara su nam dobre
- RM operator: *Optimize Parameters*
  - Operator vraća najbolji skup parametara i najbolju iteraciju
  - Definicija najboljeg skupa određena je *Performance Vectorom*
- RM operator *Log* možemo iskoristiti za određivanje kako performansa modela ovisi o parametrima

RM Vježba 3 (Ponavljanje)

1. Učitajte skup primjera *“Sonar”* iz *sample* repozitorija.

2. Nađite optimalne parametre SVM modela

3. Za skup primjera *“Iris”* iz *sample* repozitorija odredite modele koristeći 10, 20, …, 150 primjera za unakrsnu validaciju i vizualizirajte kako veličina primjera za učenje utječe na točnost – krivulja učenja (ne koristiti operator *Learning Curve*)

Optimizacija strukture RM procesa

- Osim optimizacije parametra danog modela, možemo:
  - Odabirati koje modele učenja koristiti
  - Odabrati koju metodu probira atributa koristiti
- Koristi se kombinacija operatora *Optimize (Loop) Parameters* i *Select Subprocess*
- Odabrani podprocesi postaju parametar koji se može optimirati

RM Vježba 4
(Odabir metoda za probir atributa)

1. Učitajte skup primjera *“Iris”* iz *sample* repozitorija (ili neki proizvoljni skup primjera)

2. Koju je najbolje metodu probira koristiti? Nađite je.

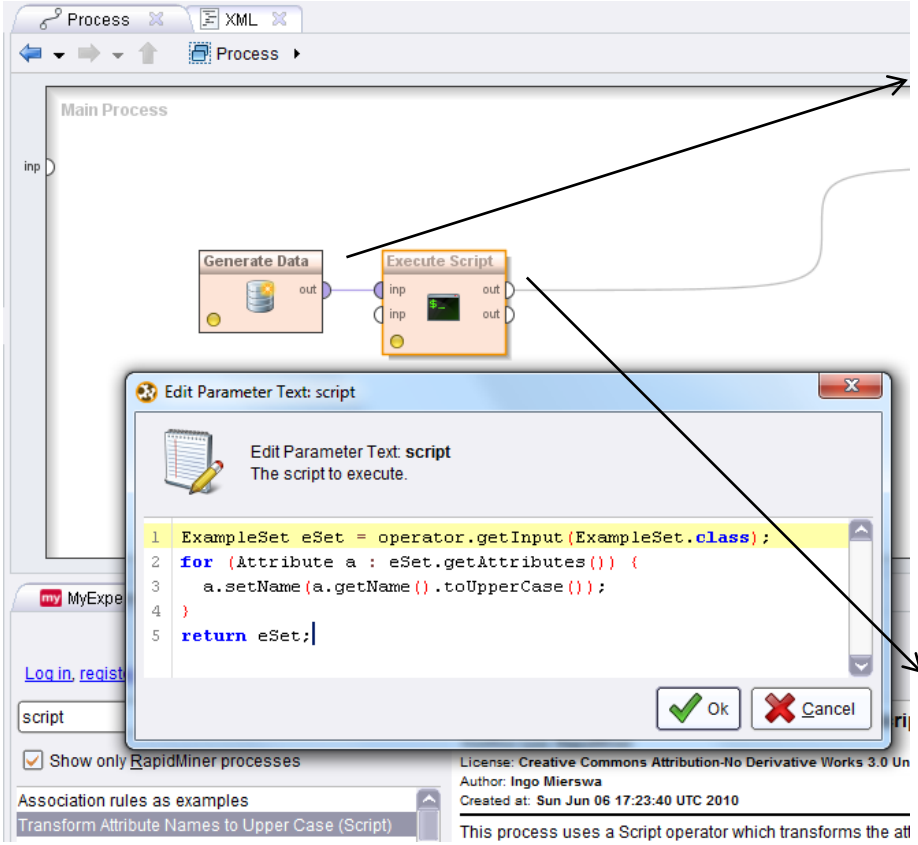
Razvoj i integracija novih metoda u RM

1. Opeator: *Execute Script*
   - *Groovy Script:* http://groovy.codehaus.org/
   - Korisno za mala brza jednostavna *custom* rješenja
2. RM ekstenzija (engl. *extension*) za R
   - Omogućuje jednostavno korištenje R operatora i skrpiti unutar RM-a
3. Pisanje novih operatora i stvaranje RM ekstenzije
   - Najefikasnije i neograničene mogućnosti
   - Preporučljivo za upotrebu u produkciji

- RM API dokumentaciju
  - http://rapid-i.com/api/rapidminer-5.1/help-doc.html
- RM izvorni kod (engl. *source code*): napisan u Javi
  - https://rapidminer.svn.sourceforge.net/svnroot/rapidminer/
  - http://rapid-i.com/content/view/25/48/lang,en/

Execute Script: Groovy

- Korisno naći primjere koristeći *myExperiment Browser*
- Potrebno poznavati API dokumentaciju RM-a




```
//Uzimanje objekta na ulazu:
ExampleSet exSet = input[0];

//Manipulacija na skupu primjera
Attributes attributes = exSet.getAttributes();
Attribute attr = attributes.get("class");
For(Example ex: exSet) {
   ex.setValue(attr, 1);
}

//Vrati rezultat:
return exSet;
```

RM ekstenzija za R

- R je široko korišteni paket za statistiku koji se primarno koristi skriptiranjem
  - http://www.r-project.org/
  - Ima puno raznih biblioteka, neke i za strojno učenje
- Uvodni video: http://rapid-i.com/content/view/239/1/
- Korisno naći primjere koristeći *myExperiment Browser*
- Napredni primjeri korištenja RM + R:
  - http://www.aphysicistinwallstreet.com/

Stvaranje novih operatora i RM ekstenzije

- Potrebno dobro poznavanje koncepata objektno orijentiranog programiranja i programskog jezika *Java*
- Početni koraci:
  - *Eclipse* IDE+SVN, JDK
  - Preuzeti izvorni kod RM-a (npr. verzija Unuk)
    - https://rapidminer.svn.sourceforge.net/svnroot/rapidminer/Unuk/
  - Preuzeti izborni kod za neku postojeću ekstenziju (npr. ValueSeries)
    - https://rapidminer.svn.sourceforge.net/svnroot/rapidminer/Plugins/ValueSeries/Unuk/
  - Proučiti osnovnu strukturu ekstenzije i pojedinih vrsta operatora
  - Proučiti *Ant build process* za ekstenziju i RM
    - stvaranje .jar datotke ekstenzije i kopiranje u /lib/plugins direktorij
  - Korištenje API dokumentacije RM-a

RapidMiner Feature Selection Extension

- Neslužbena *Custom ekstenzija za RM 5.x*
  - sadrži dodatne metode za probir atributa i algoritme za klasifikaciju dobro prilagođene za višedimenzionalne podatke
  - http://sourceforge.net/projects/rm-featselext/
- Instalacija
  - Jednostavno kopiranje datoteke "rapidminer-Feature-Selection-Extension-1.0.x.jar“ u direktorij .../RM_HOME/plugins/lib

RM Vježba 5

1. Instalirajte RM ekstenziju *"RapidMiner Feature Selection Extension"*

2. Proučite oratore za probir atributa koje ekstenzija nudi.
   - Attribute Selection - Meta-Operators:
     - Ensemble Feature Selection
     - Windowed Weighting
     - Recursive Feature Elimination (RFE)
     - Feautre Selection Stability Evaluation
   - Attribute Selection :
     - Select by SVM-RFE
     - Select by MRMR / CFS
     - Select by Feature Quantile Filter
     - Performance (MRMR)