Priprema podataka

NIKOLA MILIKIĆ

EMAIL: nikola.milikic@fon.bg.ac.rs

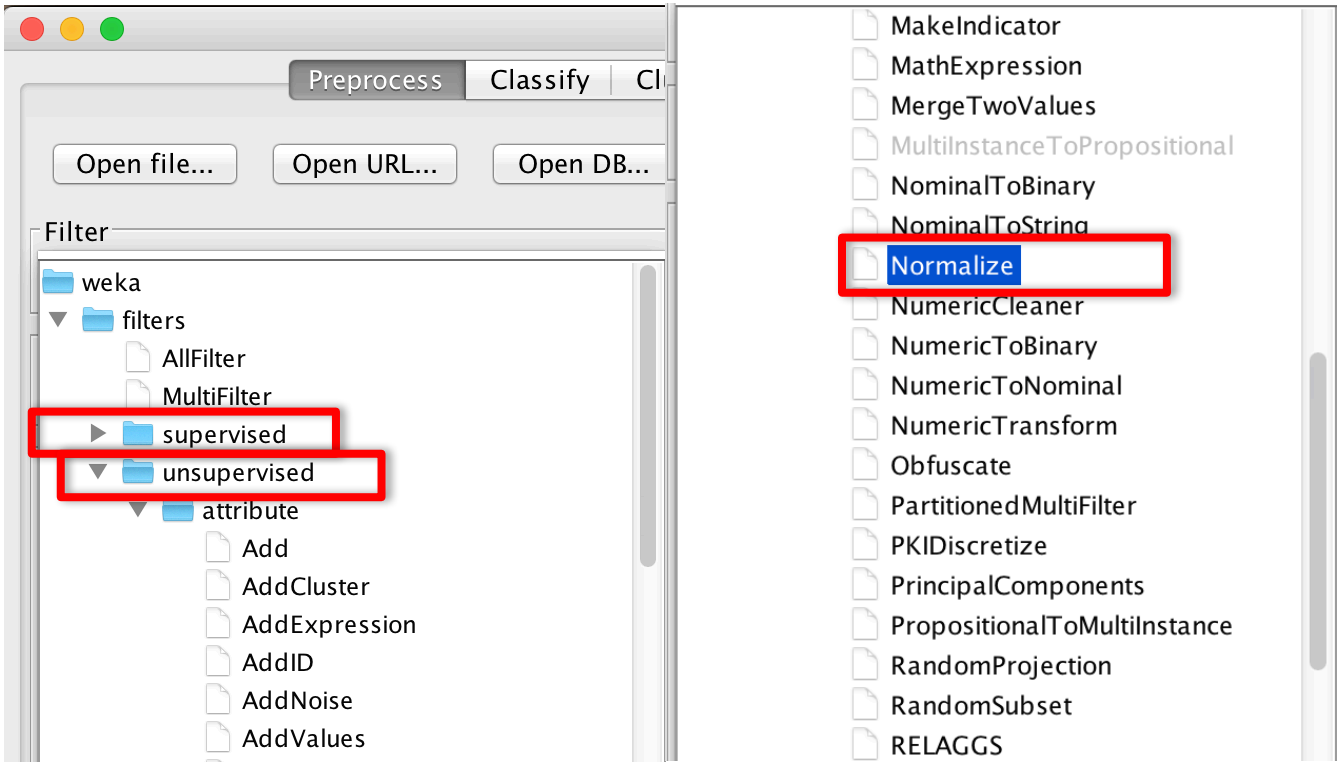
URL: http://nikola.milikic.info

Normalizacija

Normalizacija je svođenje vrednosti na neki opseg (obično 0 - 1). Koristi se kada atributi imaju razlicite merne jedinice ili opsege vrednosti.

Nedostatak je ako postoje *outliers*.

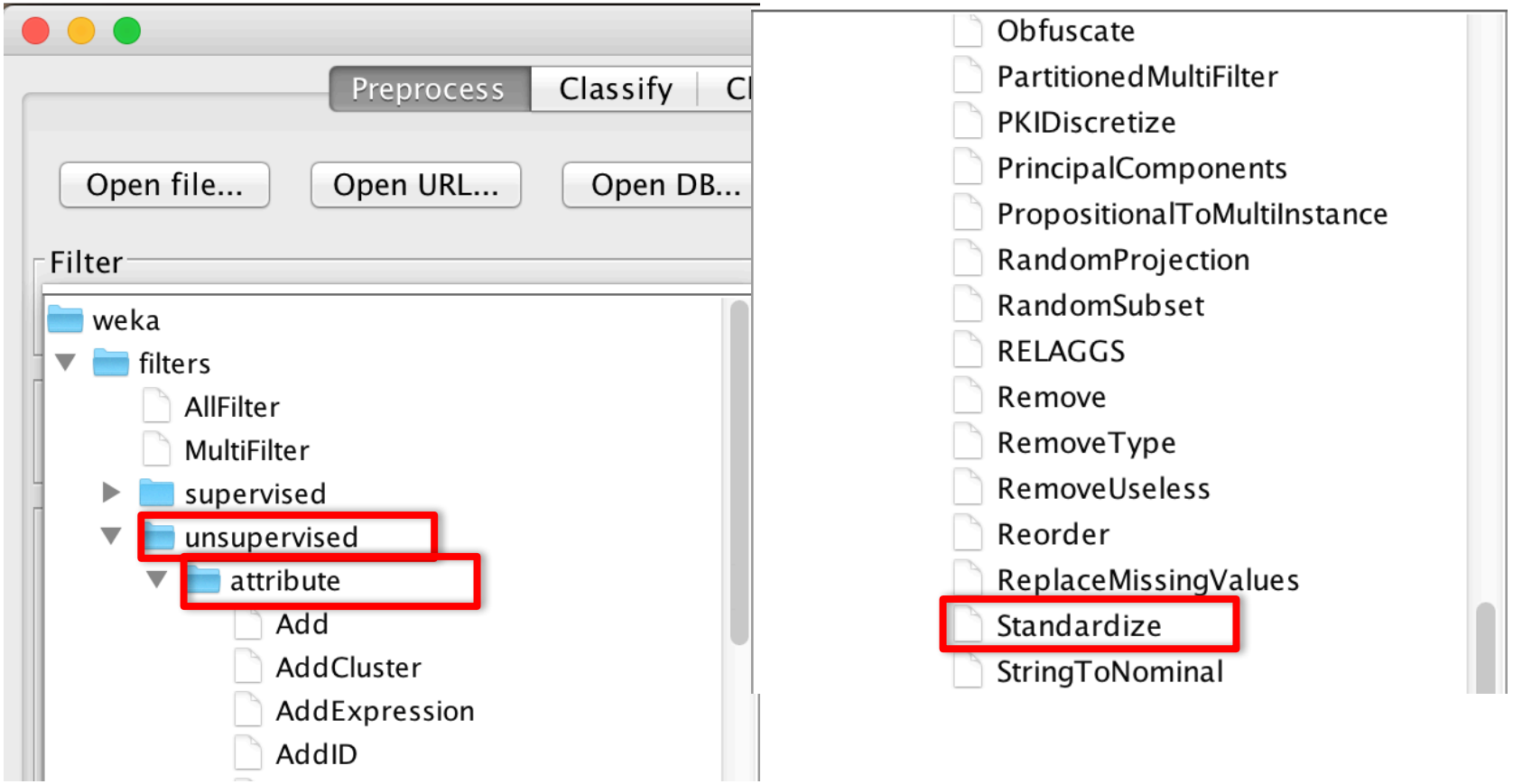
FishersIrisDataset.arff



Standardizacija

Standardizacija je svođenje srednje vrednosti na 0, a standardne devijacije na vrednost 1

FishersIrisDataset.arff



Diskretizacija atributa

Diskretizacija je proces transformacije numeričkih podataka u nominalne tako što se numeričke vrednosti smeštaju u odgovarajuće grupe kojih ima konačan broj.

Najčešći pristupi diskretizacije su:

- Nenadgledani pristupi:
  - Jednake širine opsega (Equal-width binning)
  - Jednaka pojavljivanja u opsezima (Equal-frequency binning)
- Nadgledani pristup – uzima u obzir klase

Jednake širine opsega

Jednake širine opsega (eng. Equal-width binning) deli opseg mogućih vrednosti na **N** podopsega iste širine.

širina = (maks. vrednost – min. vrednost) / N

Primer: Ako je opseg posmatranih vrednosti između 0 – 100, možemo kreirati 5 podopsega na sledeći način:

Širina = (100 – 0) / 5 = 20

Opsezi su: [0-20], (20-40], (40-60], (60-80], (80-100]

Obično se prvi i poslednji opsezi proširuju kako bi uključili vrednosti van opsega.

Jednaka pojavljivanja u opsezima

Jednaka pojavljivanja u opsezima (eng. Equal-frequency ili equal-height binning) deli opseg mogućih vrednosti na **N** podopsega gde svaki podopseg sadrži isti broj instanci.

Primer: Pretpostavimo da želimo da smestimo u 5 podopsega vrednosti:

5, 7, 12, 35, 65, 82, 84, 88, 90, 95

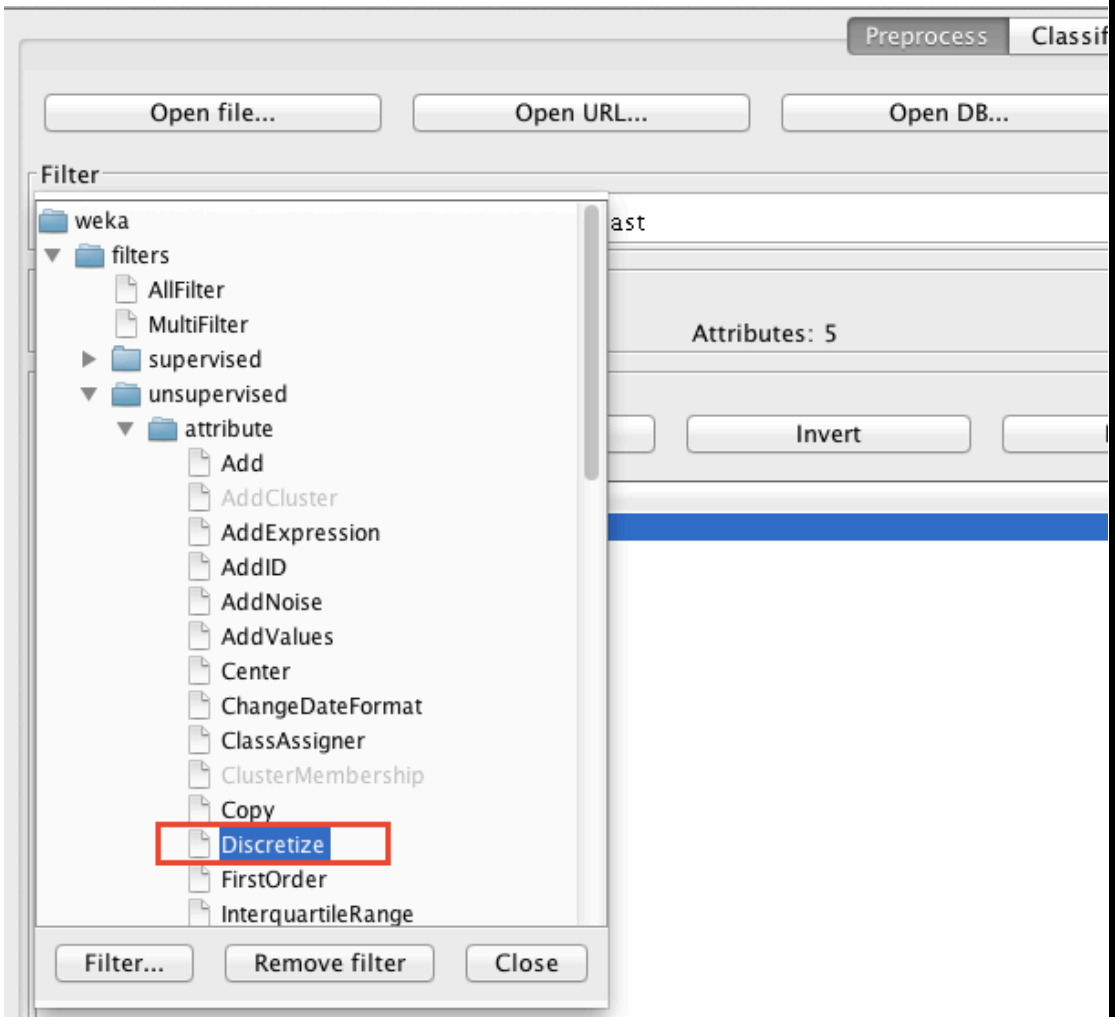
Podopsege ćemo podeliti tako što će svaki sadržati po dve instance:

5, 7, | 12, 35, | 65, 82, | 84, 88, | 90, 95

Diskretizacija u Weka-i

Atributi se diskretizuju tako što se nad njihovim vrednostima primeni odgovarajući *Filter*.

Na *Preprocess* tabu se bira opcija *Choose -> Filter* i u folderu *filters/unsupervised/attribute* se odabira filter ***Discretize***.



FishersIrisDataset.arff

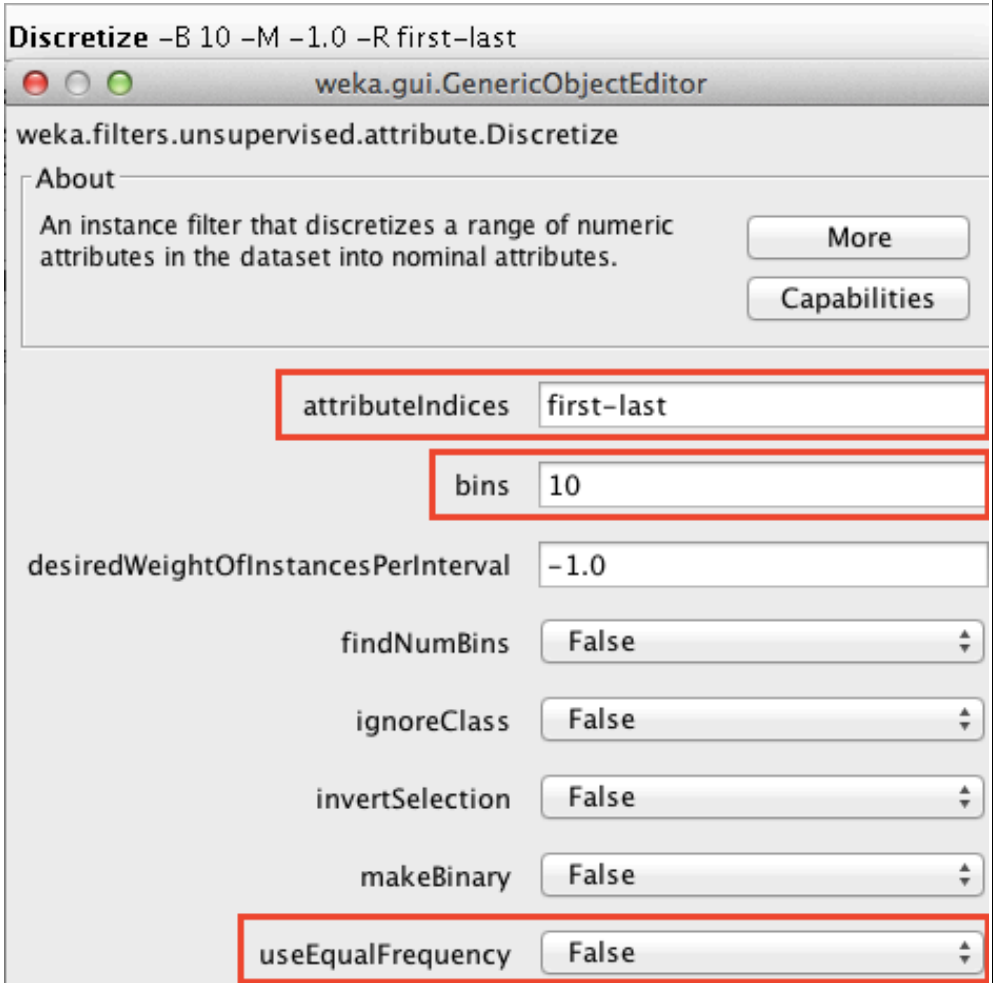
Diskretizacija u Weka-i

Po defaultu se primenjuje Diskretizacija sa jednakim širinama opsega.

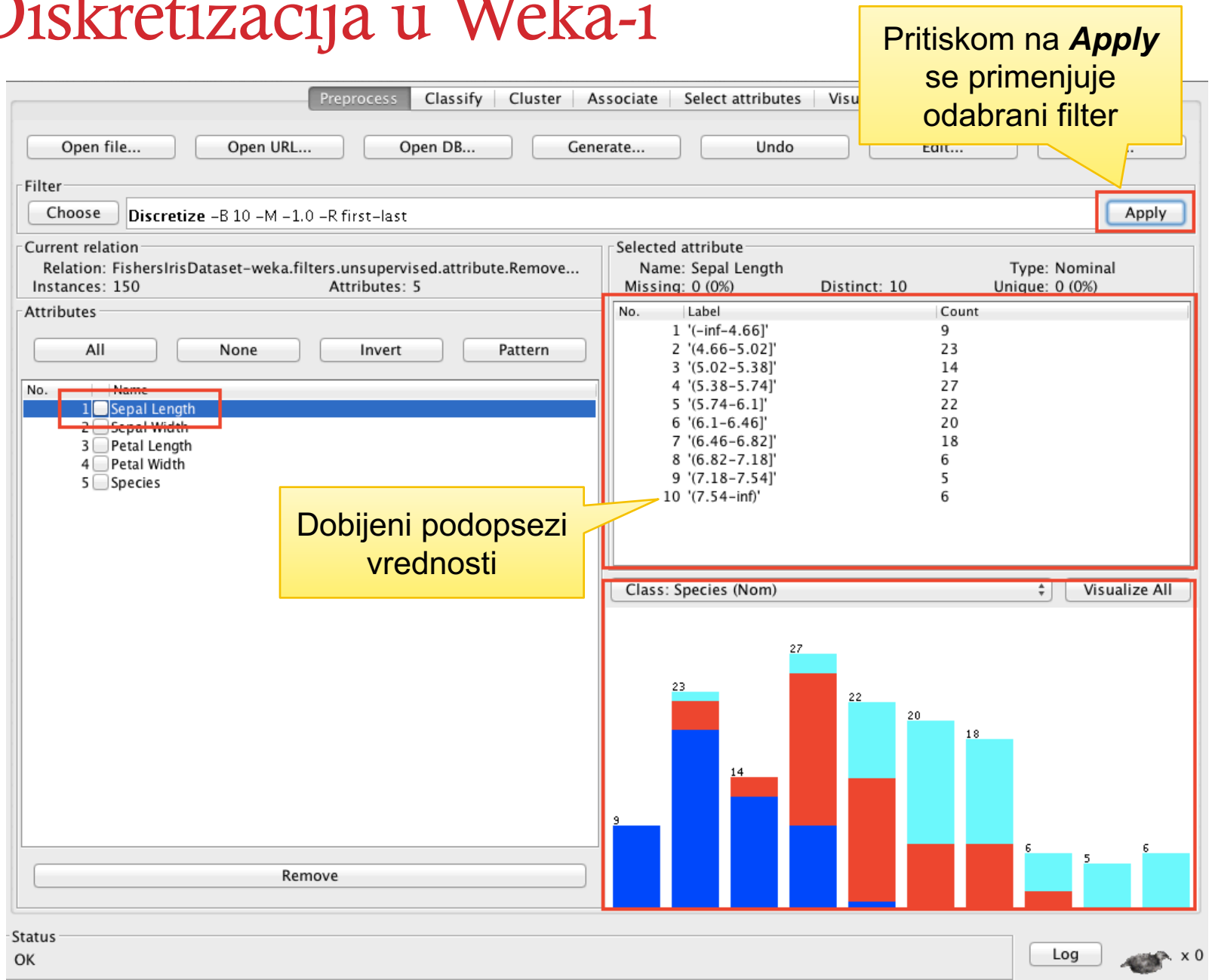
- ***attributeIndices*** - vrednost *first-last* označava da diskretizjemo sve atribute. Mogu se navesti i redni brojevi atributa

- ***bins*** - željeni broj opsega

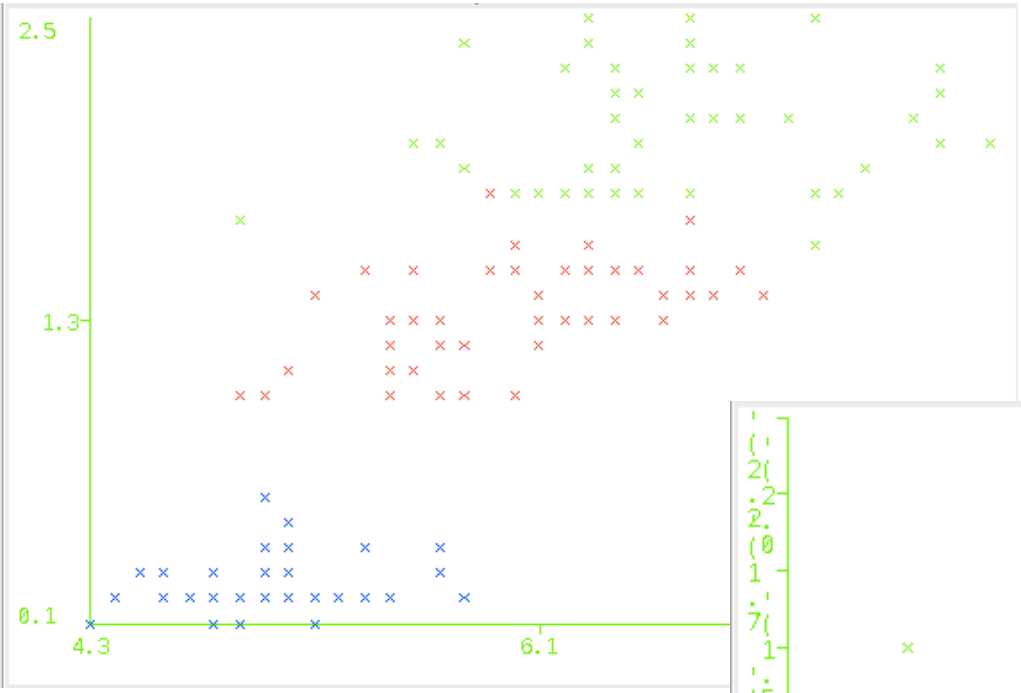
- ***useEqualFrequency*** – *true* ako se koristi diskretizacija sa jednakim pojavljivanjima u opsezima, *false* ako se koristi Diskretizacija sa jednakim širinama opsega



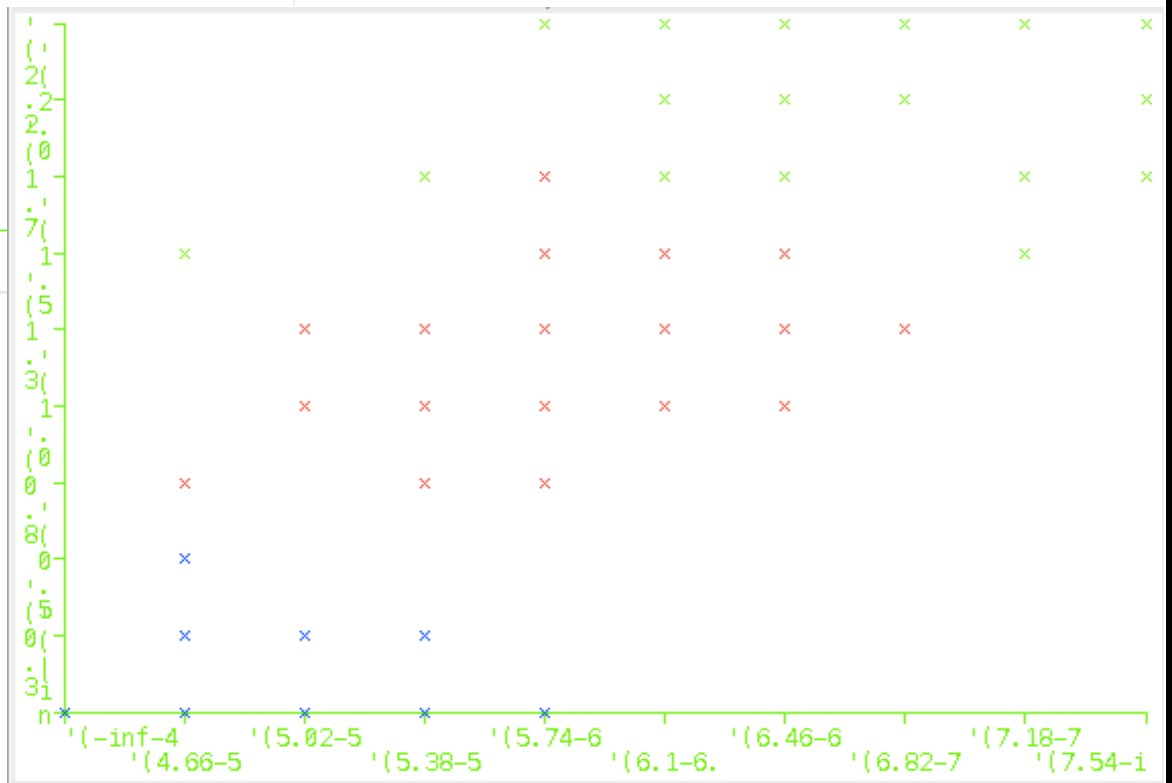
Diskretizacija u Weka-i



Podaci pre i posle diskretizacije



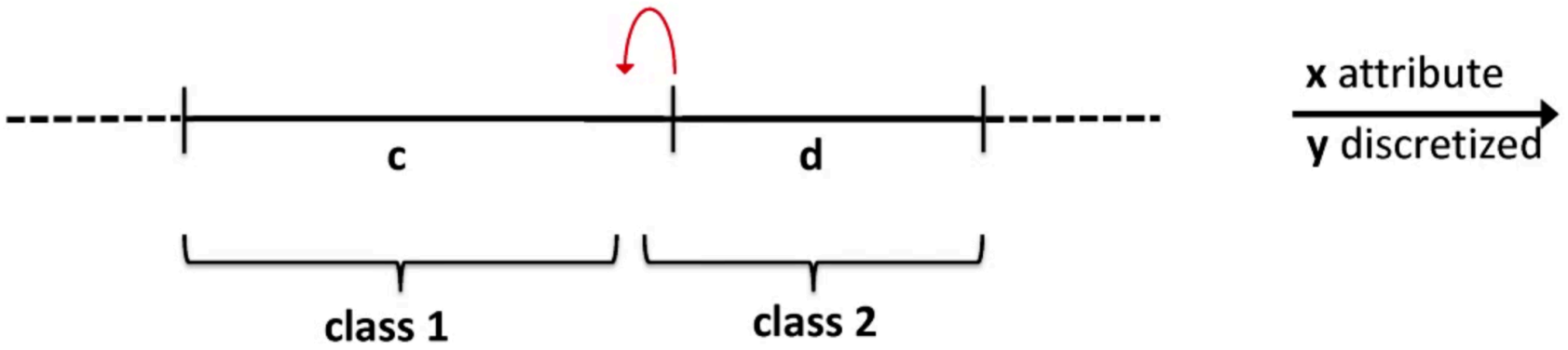
Podaci pre diskretizacije



Podaci nakon diskretizacije

Nadgledana diskretizacija

- Šta ako sve instance u jednom binu pripadaju jednoj klasi, a sve instance drugog bina pripadaju drugoj klasi osim prvog koji pripada prvoj klasi?



- Nadgledana diskretizacija uzima u obzir i klasu

Nadgledana diskretizacija

- Jedan od pristupa je koristiti entropiju
- U primeru *weather.numeric.arff*, kod atributa *temperature*

64	65	68	69	70	71	72	75	80	81	83	85
yes	no	yes	yes	yes	no	no	yes	no	yes	yes	no
						yes	yes				

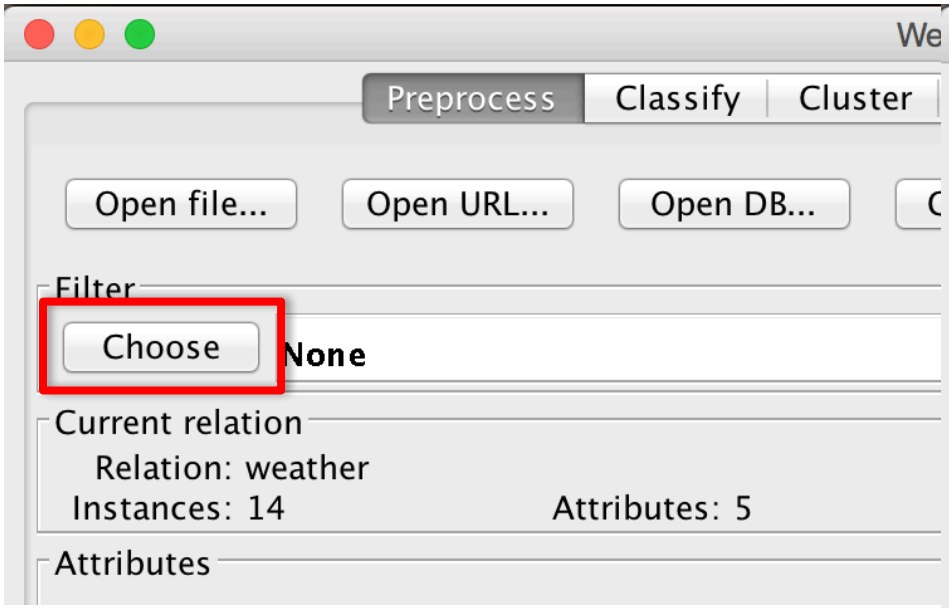
4 yes, 1 no 5 yes, 4 no
entropy = 0.934 bits

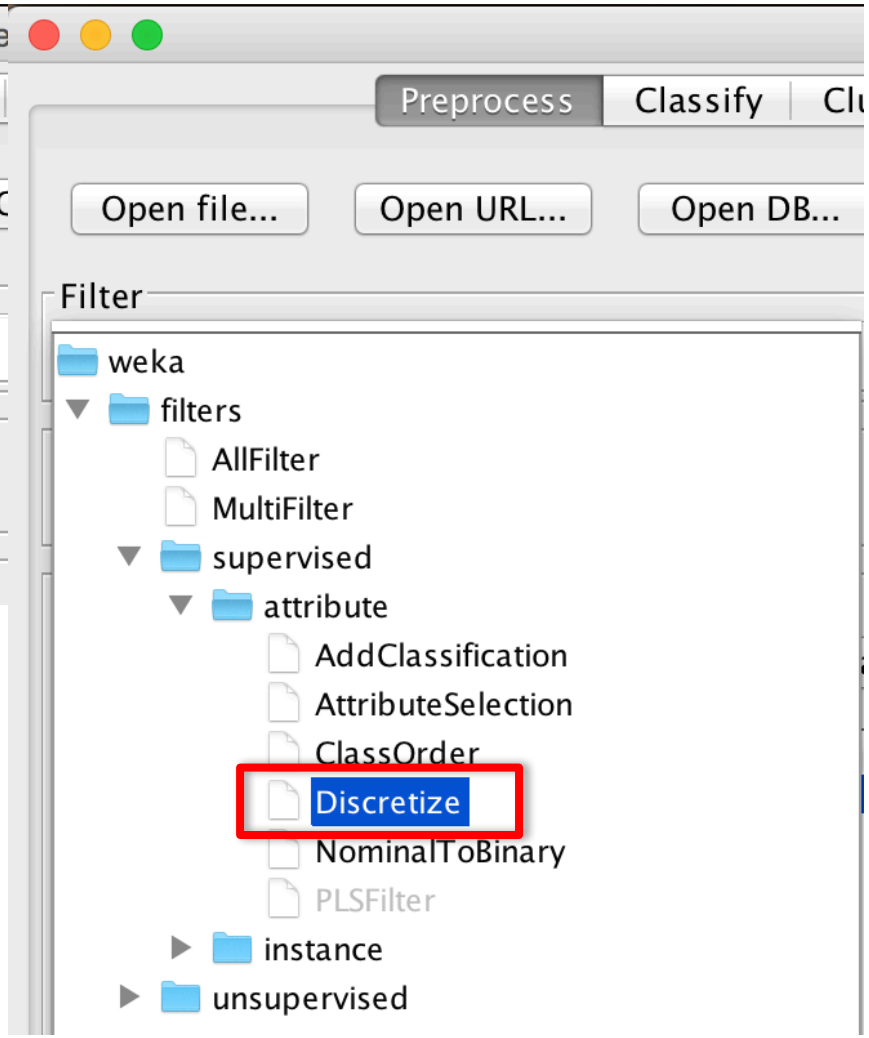
- Uzima se razdelnik sa najmanjom entropijom (najvećom informacionom dobiti)

64	65	68	69	70	71	72	75	80	81	83	85
yes	no	yes	yes	yes	no	no	yes	no	yes	yes	no
						yes	yes				

Nadgledana diskretizacija u Weka-i

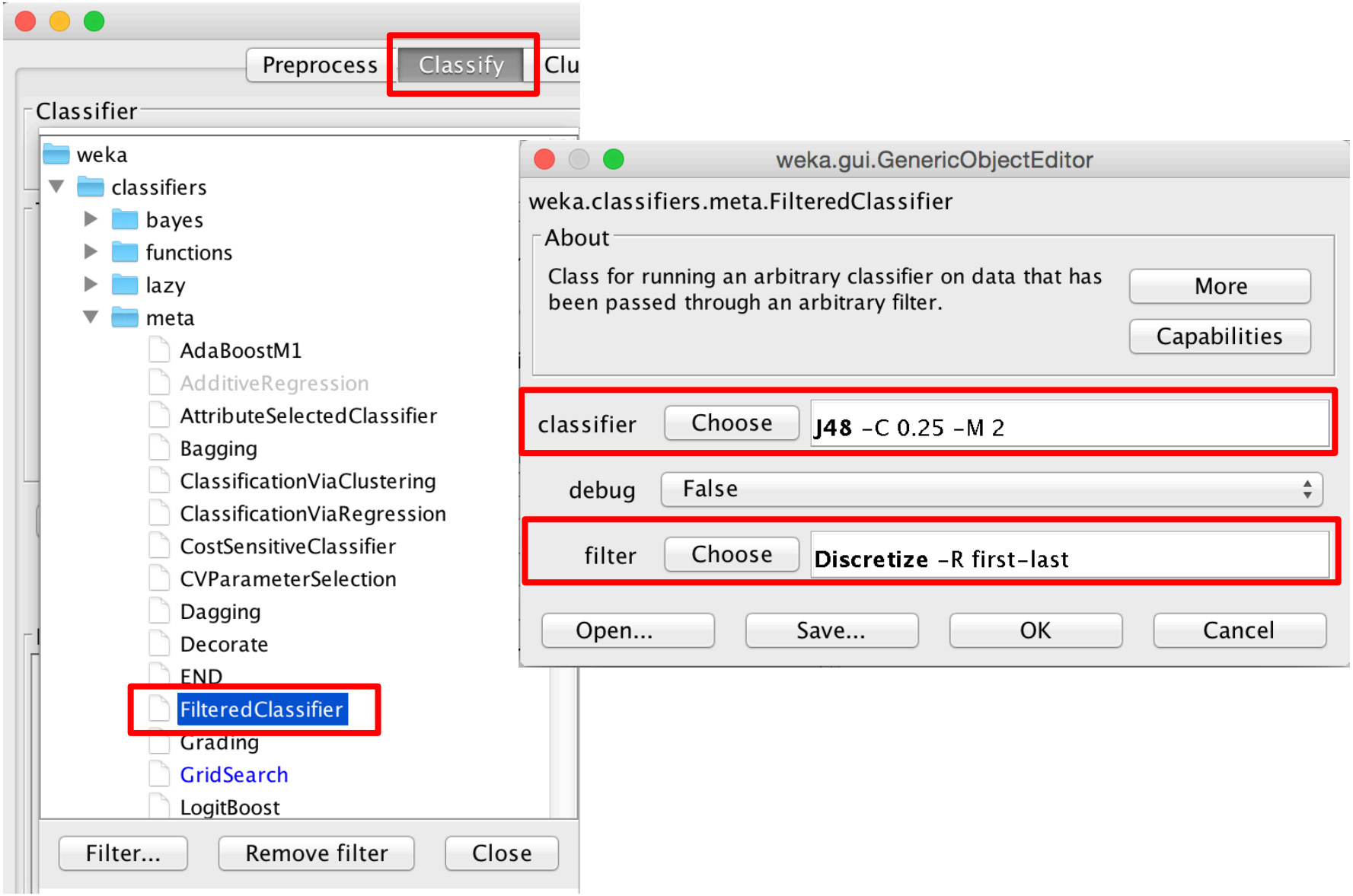
weather.numeric.arff





Problem je što prilikom nadgledane diskretizacije koristimo podatke iz celog dataset-a, pa samim tim i test podatke nad kojima ćemo posle vršiti testiranje performansi klasifikatora

meta>FilteredClassifier



Selekcija atributa

Selekcija atributa (eng. Attribute Selection ili Feature Selection) je proces odabira podskupa relevantnih atributa koji će se koristiti.

Primenjuje se u slučajevima kada se u datasetu nalaze atributi koji su redundantni ili nerelevantni.

- Redundanti atributi su oni koji ne pružaju nikakve dodatne informacije u odnosu na već selektovane atribute.
- Nerelevantni atributi su oni koji ne pružaju nikakve informacije u datom kontekstu.

Prednosti primene selekcije atributa

Suvišni atributi mogu degradirati performanse modela.

Prednosti selekcije atributa:

- Poboljšava čitljivost modela time što se model sastoji samo iz relevantnih atributa
- Kraće vreme treniranja
- Povećana generalizacija time što smanjuje mogućnosti za overfitting

Najbolji način za selekciju atributa je ručno ukoliko se dobro poznaje problem koji se rešava. I automatizovani pristupi selekcije daju dobre rezultate.

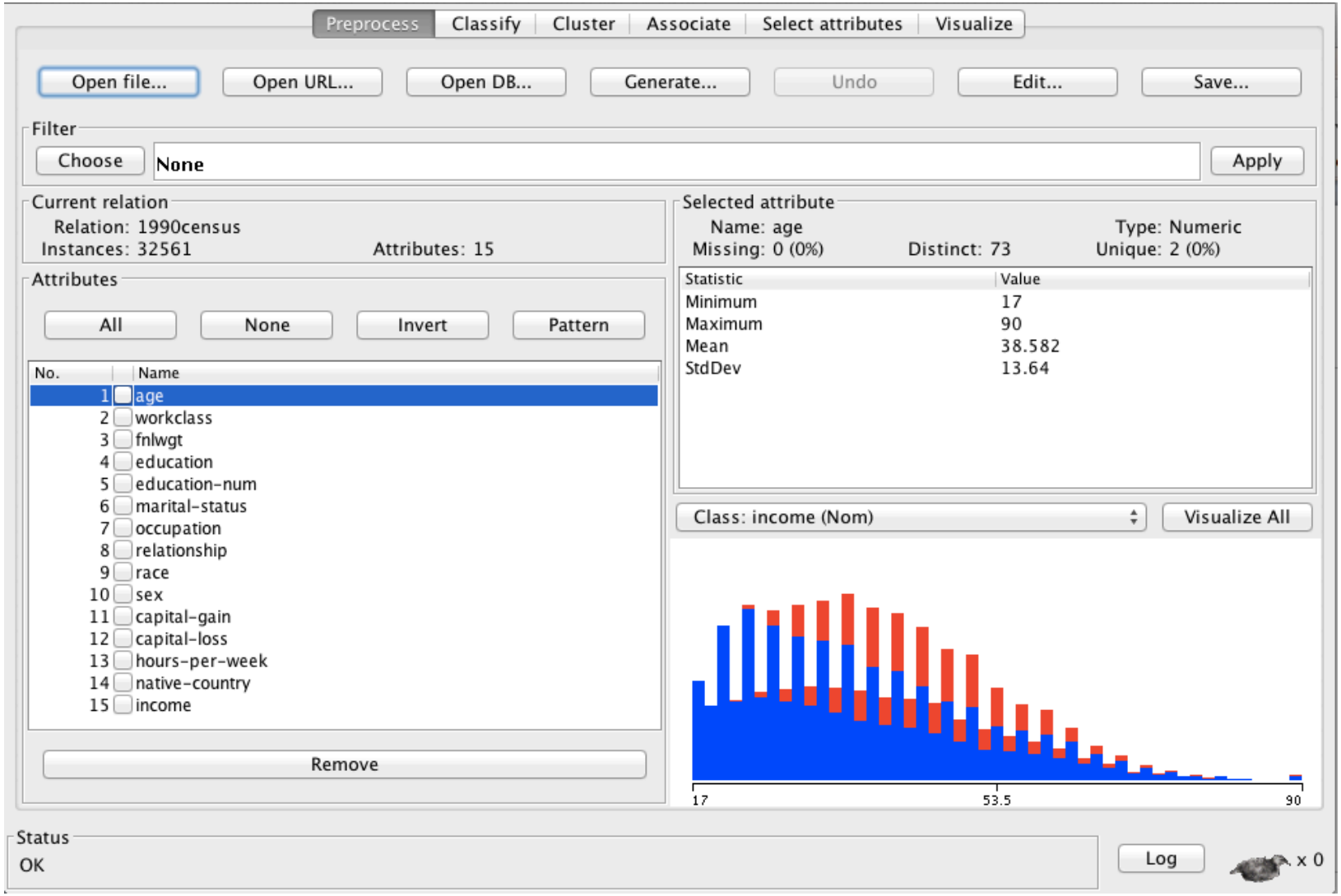
Pristupi selekcije atributa

Postoje dva pristupa:

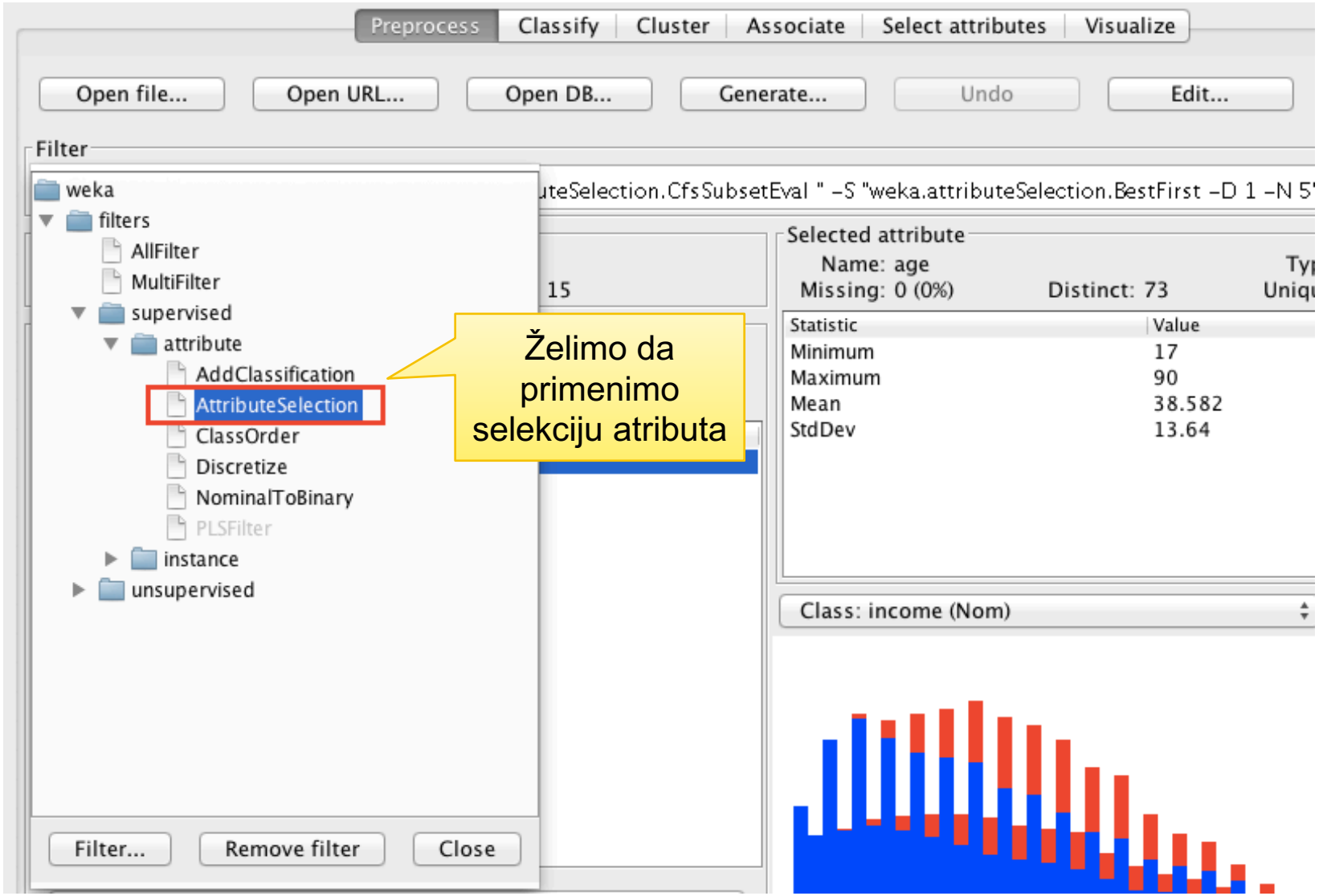
- *Filter* metoda – koriste se procene atributa na osnovu generalnih svojstava podataka. Svakom atributu se dodeli neki skor. Na osnovu skora se atributi rangiraju i zadržavaju ili odbacuju. Npr. mogu se koristiti Hi kvadrat test, informaciona dobit, stepen korelacije, itd..
- *Wrapper* metoda – podskupovi atributa se evaluiraju primenom algoritma mašinskog učenja koji će se koristiti nad skupom podataka. Naziv Wrapper se koristi iz razloga što je algoritam učenja "zapakovan" u samom procesu selekcije. Biće odabran onaj podskup atributa za koje dati algoritam učenja daje najbolje rezultate.

Primer selekcije atributa

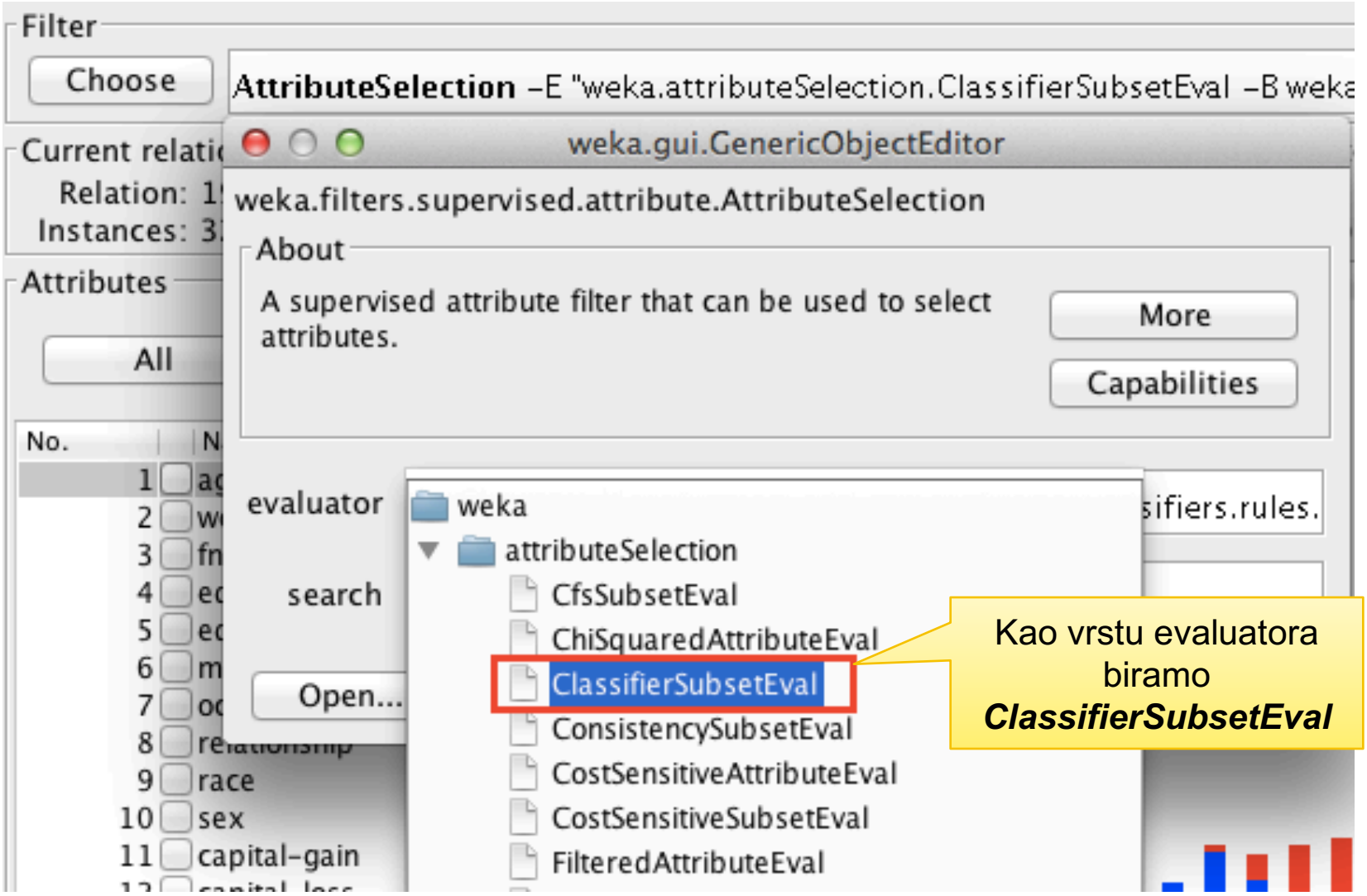
census90-income.arff



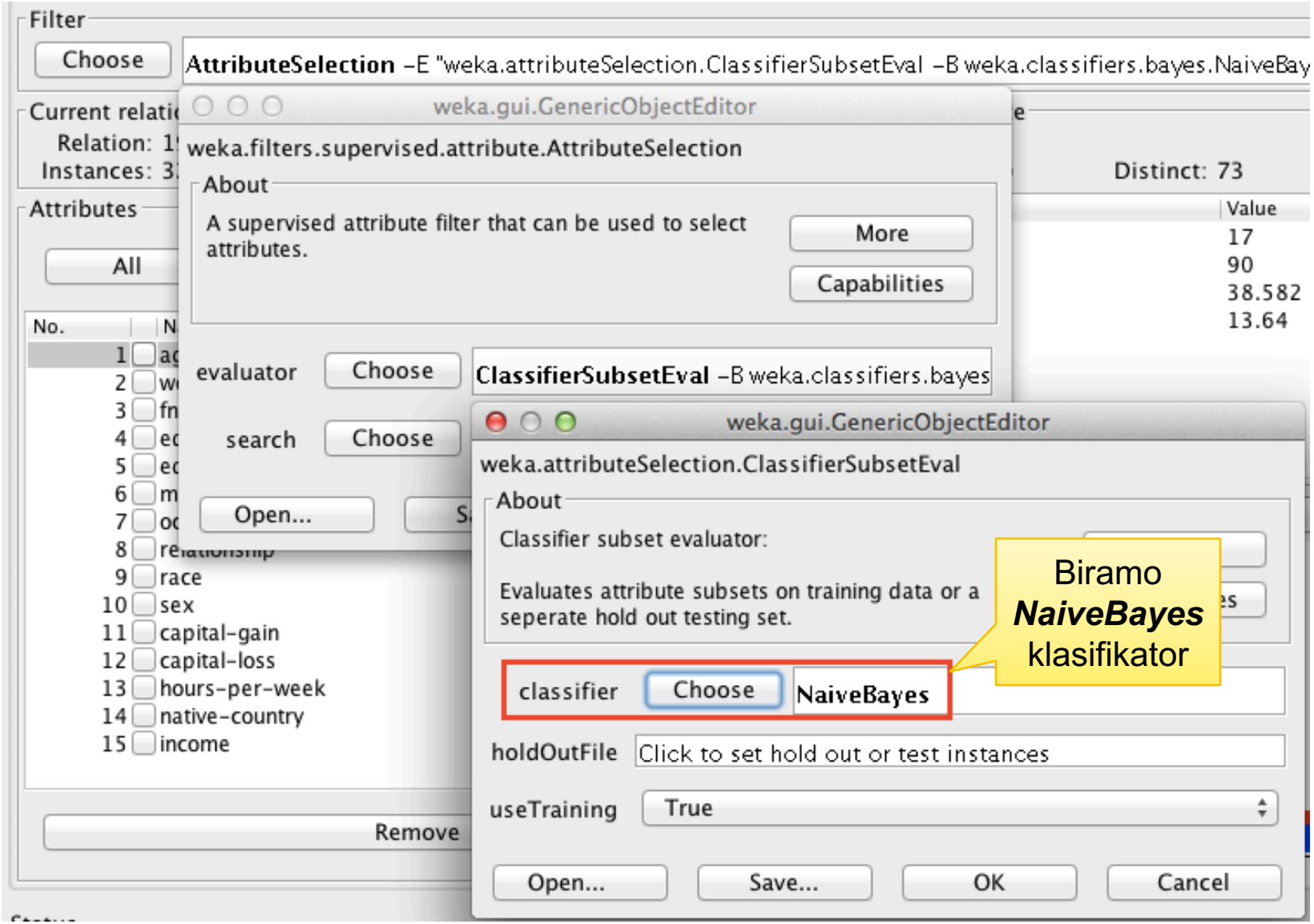
Primer selekcije atributa



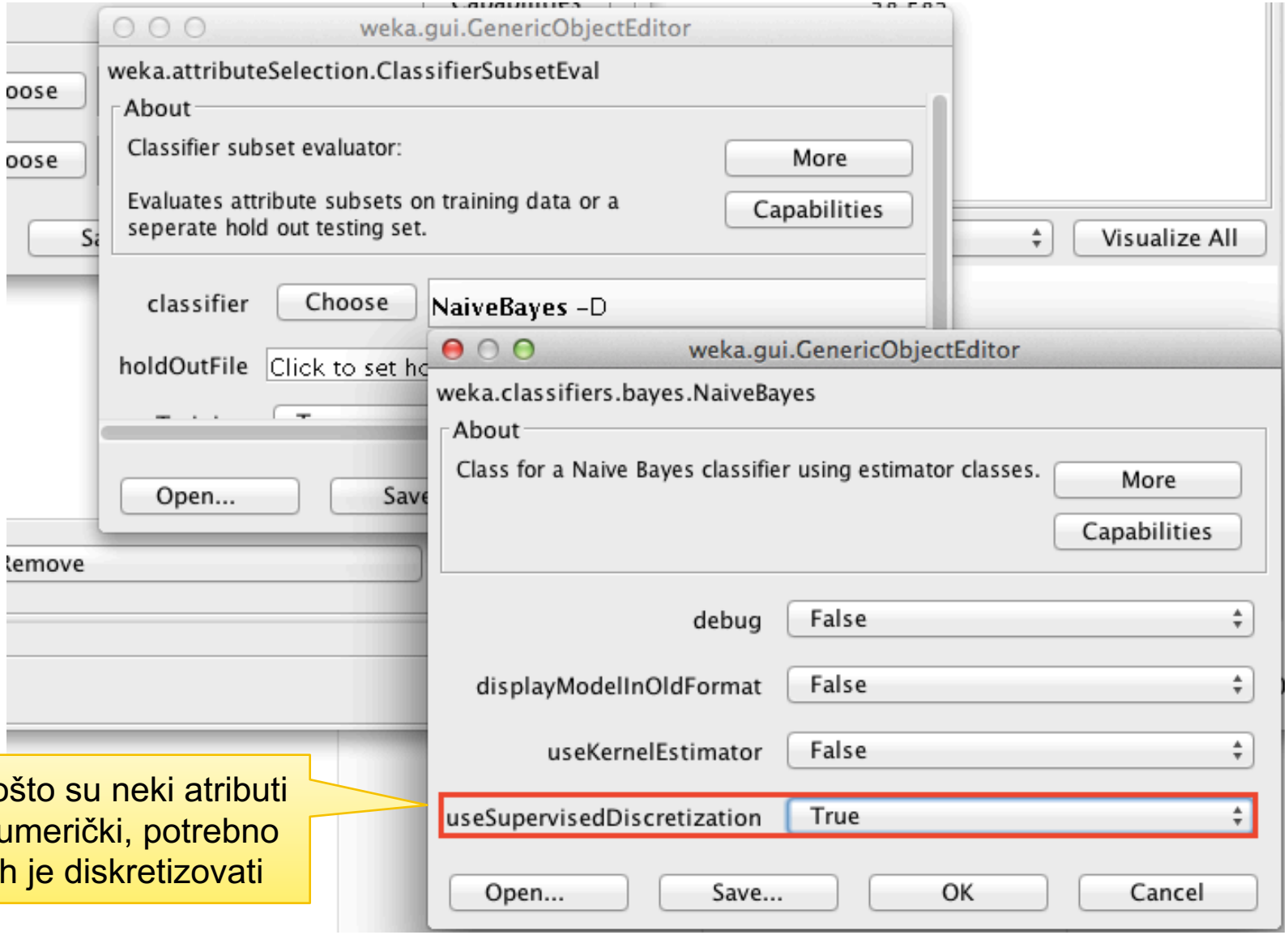
Primer selekcije atributa



Primer selekcije atributa

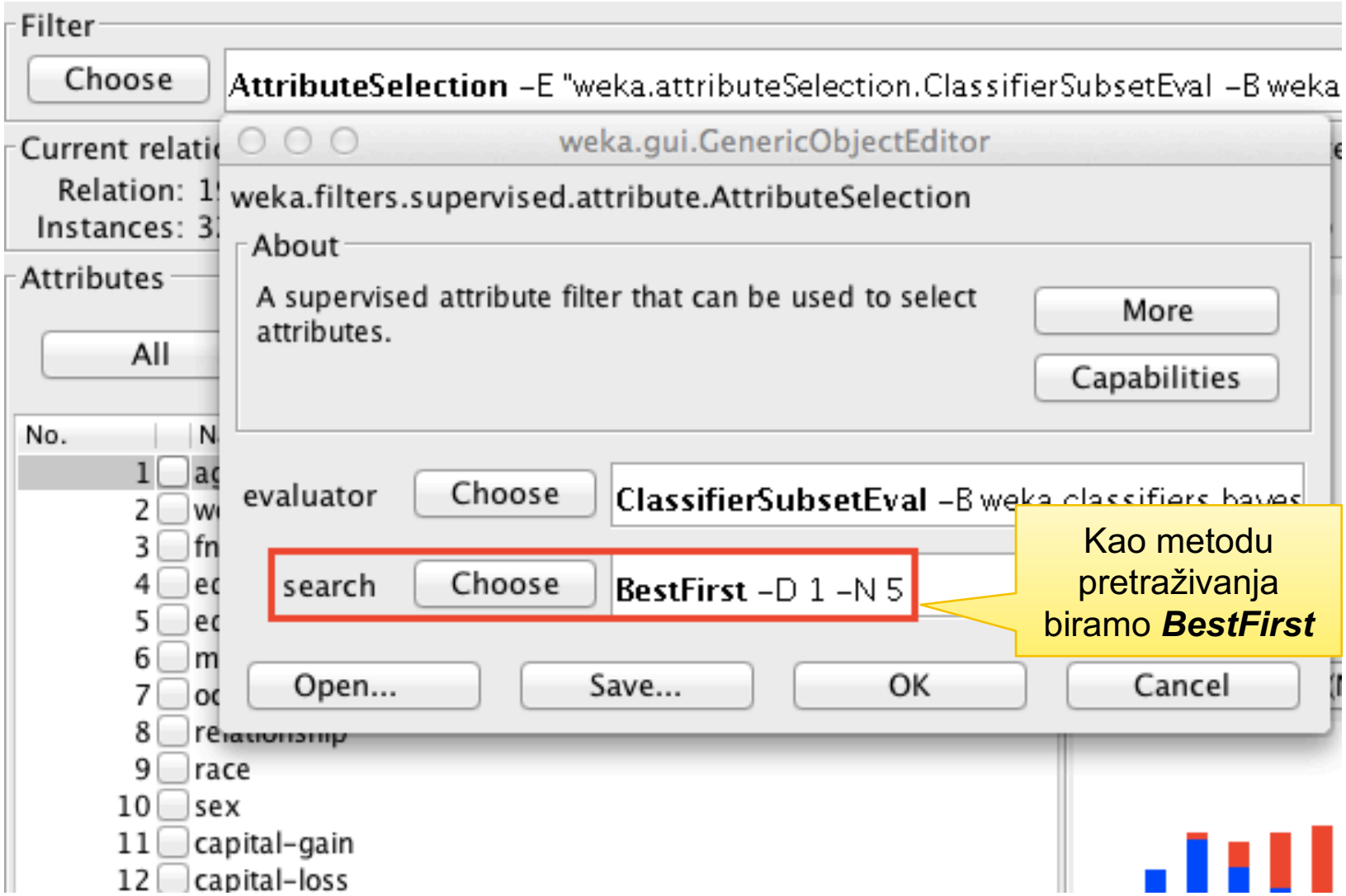


Primer selekcije atributa

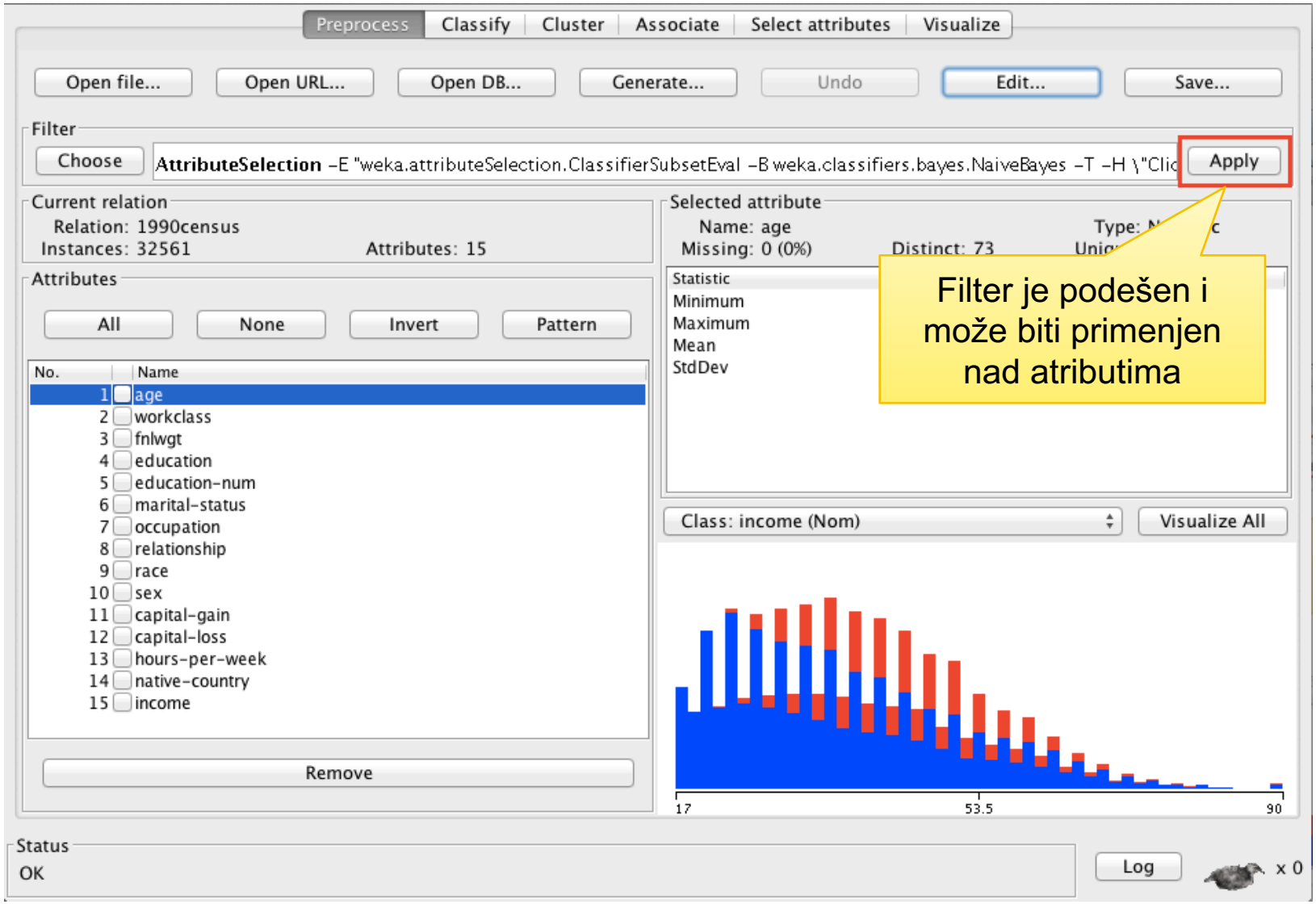


Pošto su neki atributi numerički, potrebno ih je diskretizovati

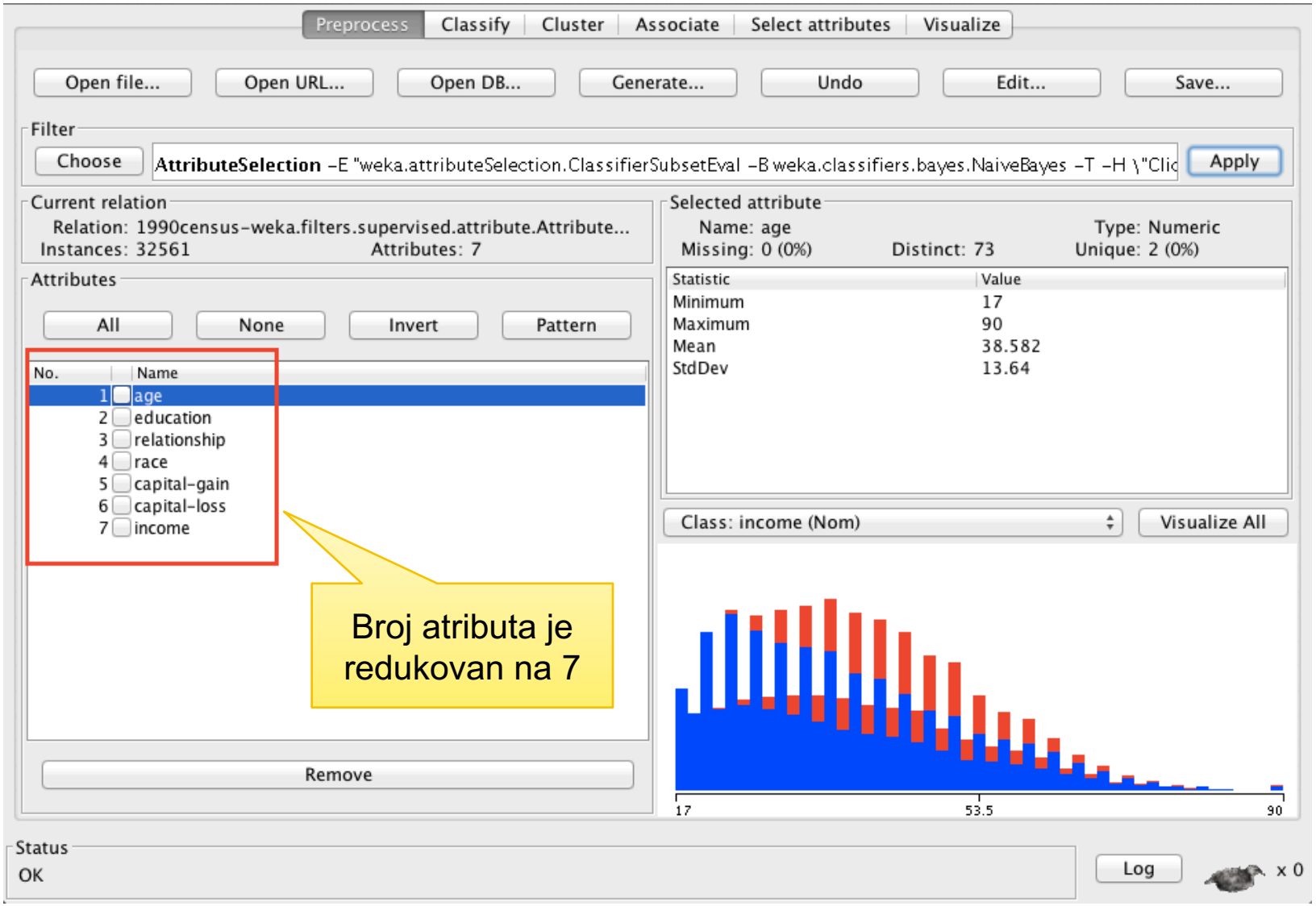
Primer selekcije atributa



Primer selekcije atributa

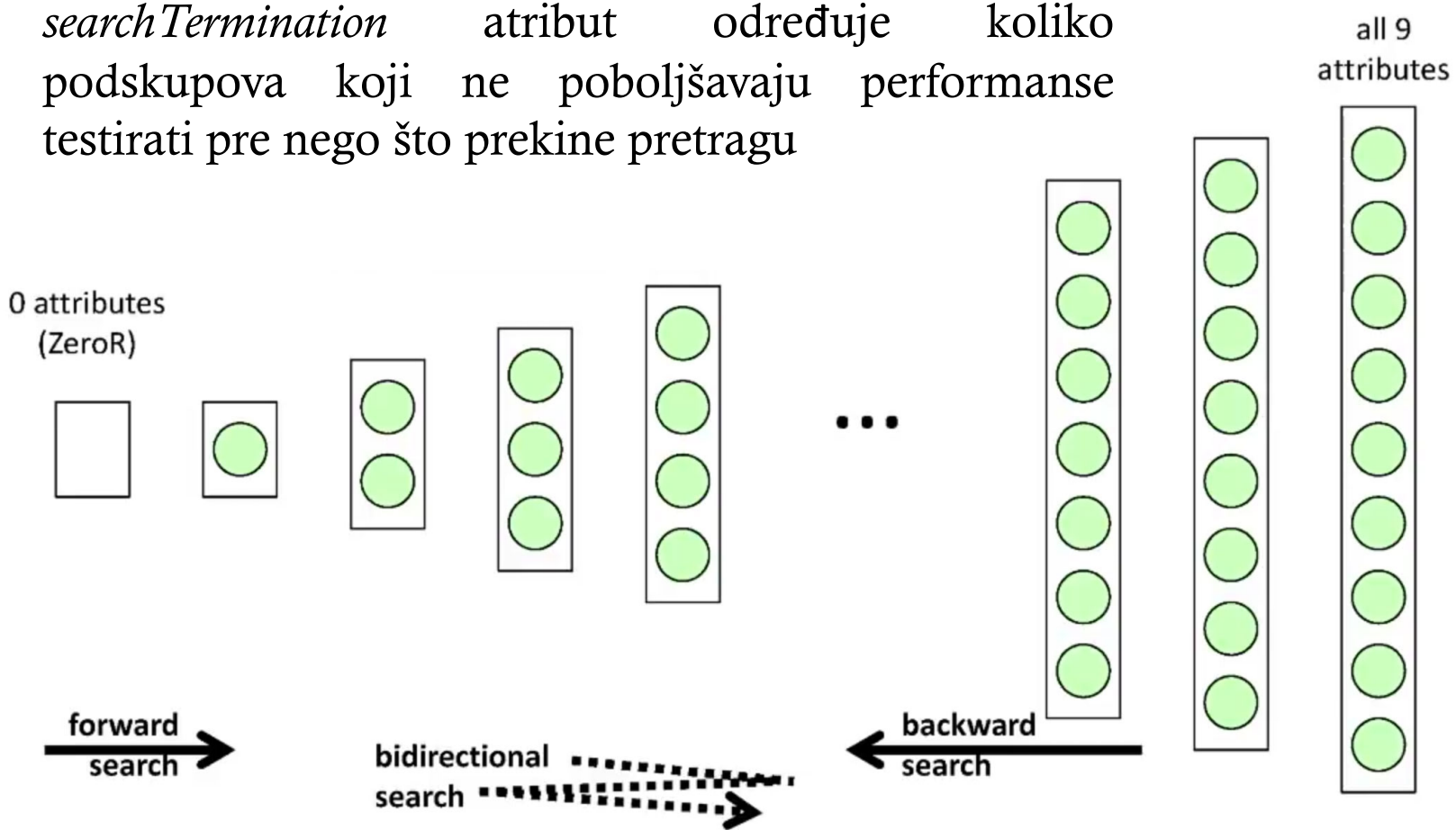


Primer selekcije atributa



Metod pretrage kod selekcije atributa

- Exhaustive search (512 podskupova atributa)
- Best First: *Forward*, *Backward*, *Bi-directional*
  - *searchTermination* atribut određuje koliko podskupova koji ne poboljšavaju performanse testirati pre nego što prekine pretragu



Preporuke i zahvalnice

"Data Mining with Weka" and "More Data Mining with Weka": MOOCs from the University of Waikato. A self-paced session of "Data Mining with Weka" runs until 23 October June 2015.

- Link: https://www.youtube.com/user/WekaMOOC/

(Anonimni) upitnik za vaše kritike, komentare, predloge:

http://goo.gl/cqdp3I

Pitanja?

NIKOLA MILIKIĆ

EMAIL: nikola.milikic@fon.bg.ac.rs

URL: http://nikola.milikic.info