Klasterizacija

NIKOLA MILIKIĆ

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

URL: http://nikola.milikic.info

Klasterizacija

Klasterizacija (eng. Clustering) spada u grupu tehnika nenadgledanog učenja i omogućava grupisanje instanci u grupe, gde unapred ne znamo koje su sve grupe moguće.

Grupe u koje se instance dele se nazivaju **klasteri**.

Kao rezultat klasterizacije svakoj instanci je dodeljen novi atribut koji predstavlja klaster kojoj pripada. Može se reći da je klasterovanje uspešno ukoliko su dobijeni klasteri smisleni i ukoliko se mogu imenovati.

K-Means

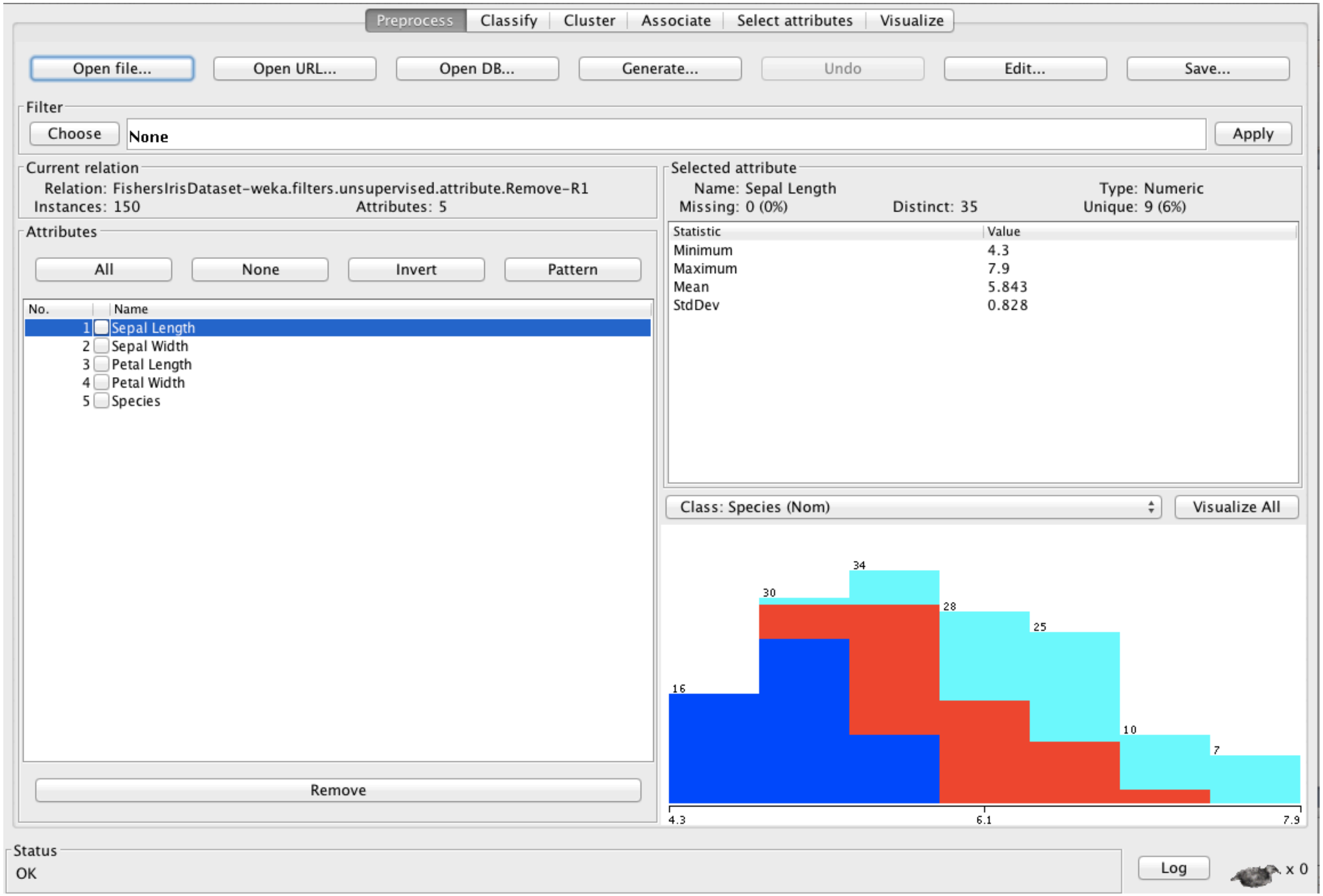
Koraci:

1) Inicijalni izbor težišta klastera, slučajnim izborom
   - težišta se biraju iz skupa instanci za trening, tj. K instanci za trening se nasumično izabere i proglasi za težišta
2) Ponoviti:
   1) Grupisanje po klasterima: za svaku instancu iz skupa za trening, i = 1,m, identifikovati najbliže težište i dodeliti instancu klasteru kome to težište pripada
   2) Pomeranje težišta: za svaki klaster izračunati novo težište uzimajući prosek instanci koje su dodeljene tom klasteru

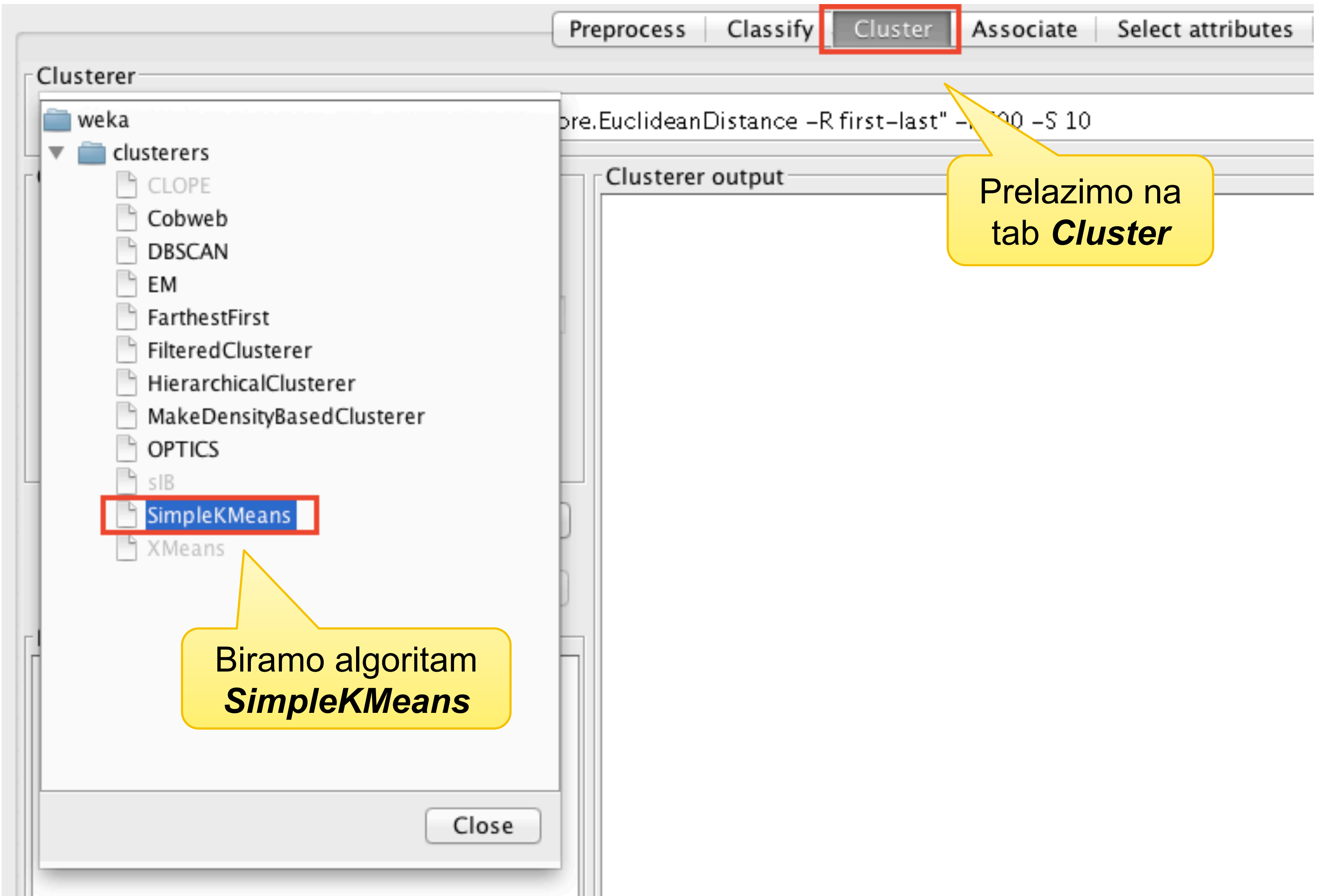
dok algoritam ne konvergira ili broj iteracija <= max

K-Means algoritam u Weka-i

FishersIrisDataset.arff



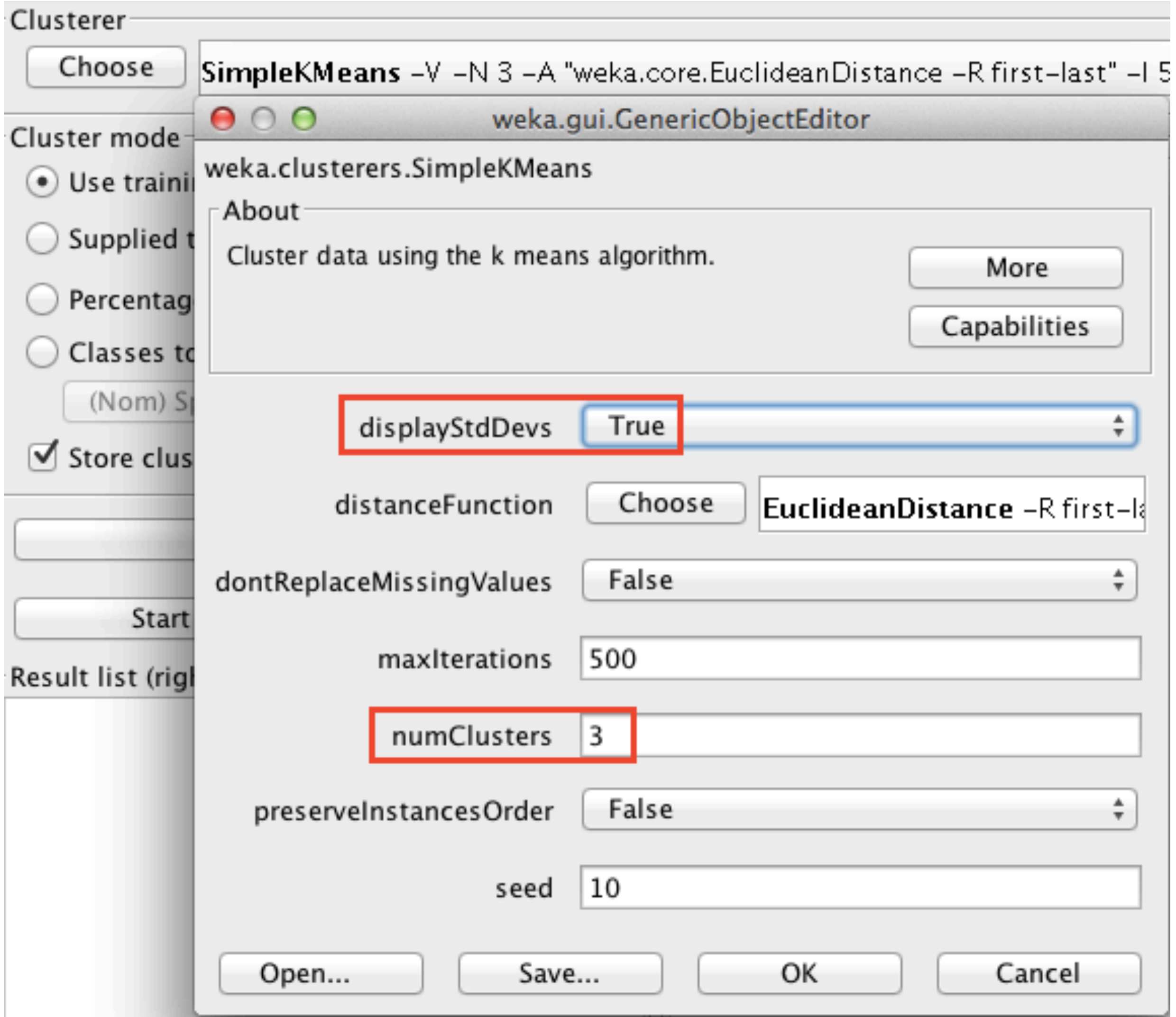
Odabir algoritma za klasterovanje



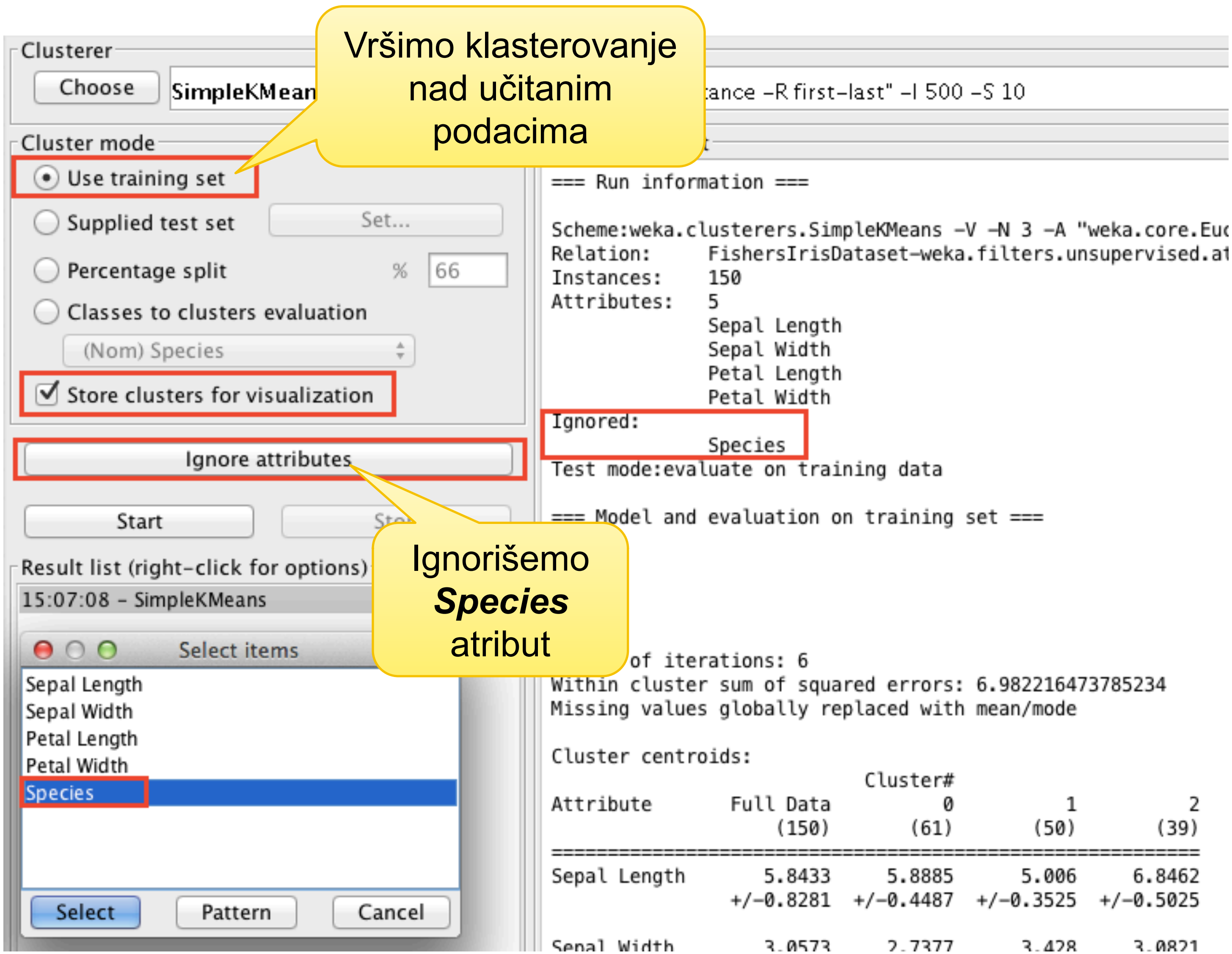
Podešavanje parametara

numClusters – broj željenih klastera; postavljamo na 3 jer imamo tri vrste

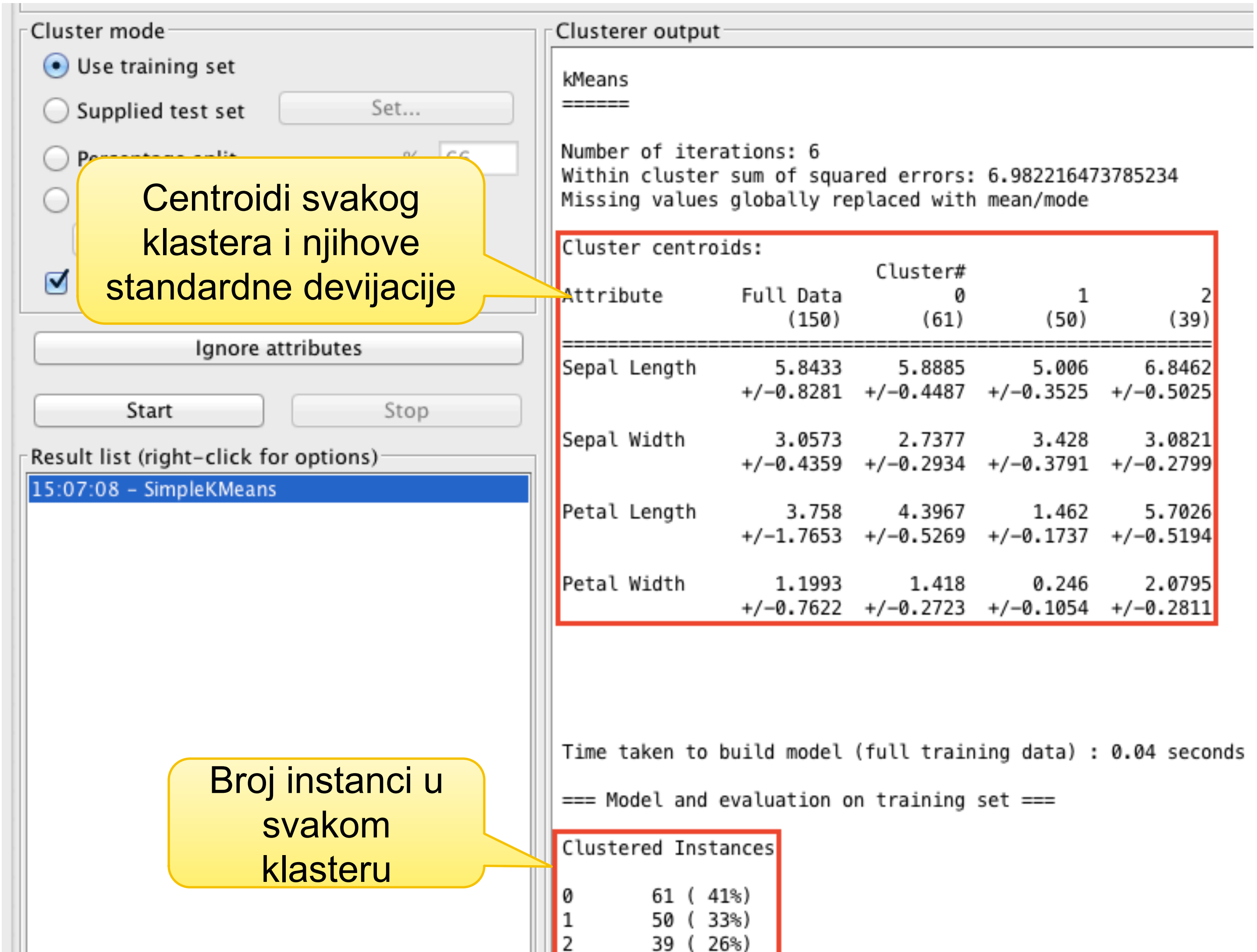
displayStdDevs– ako je *true*, onda će se ispisati vrednosti standardne devijacije



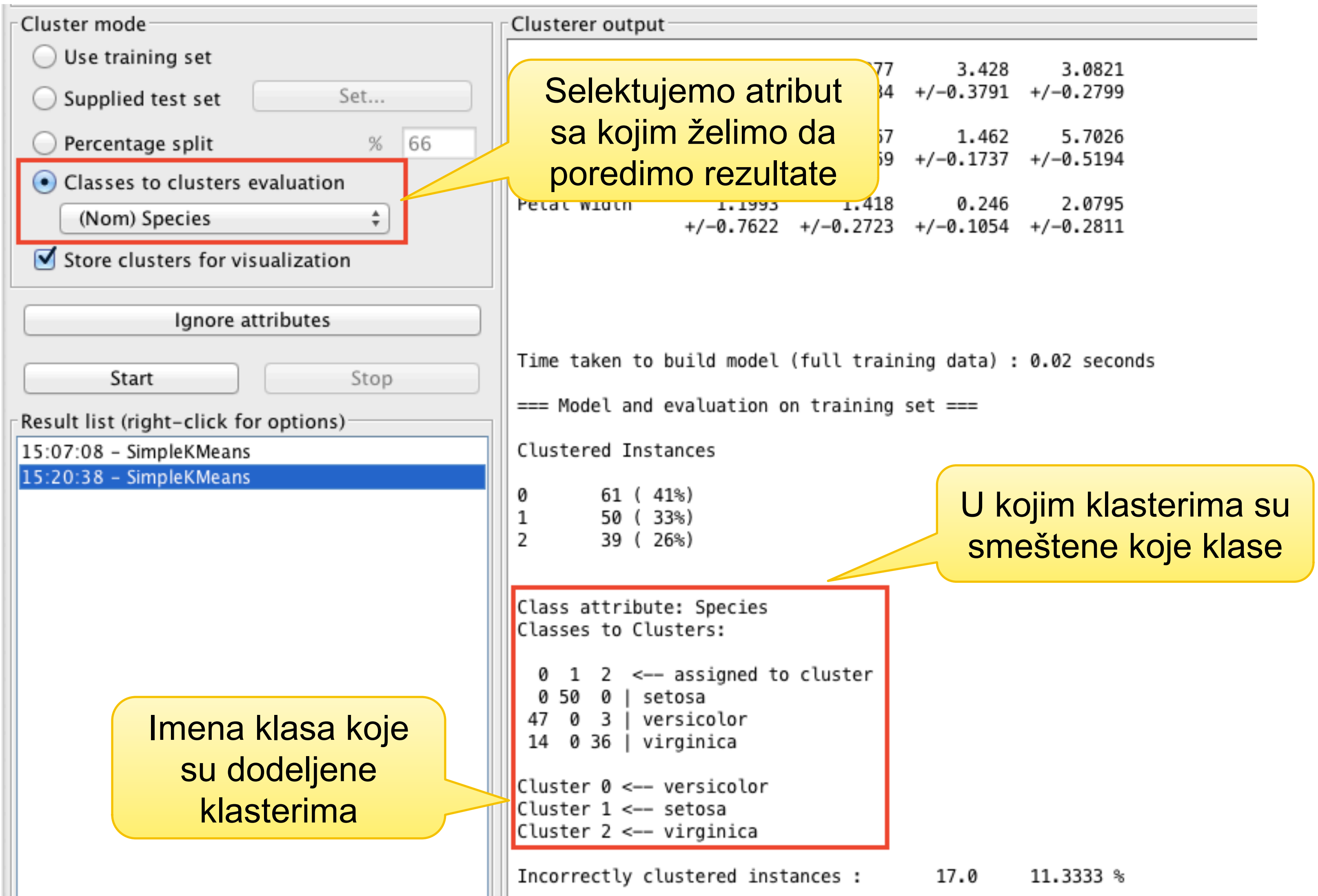
Pokretanje procesa klasterovanja



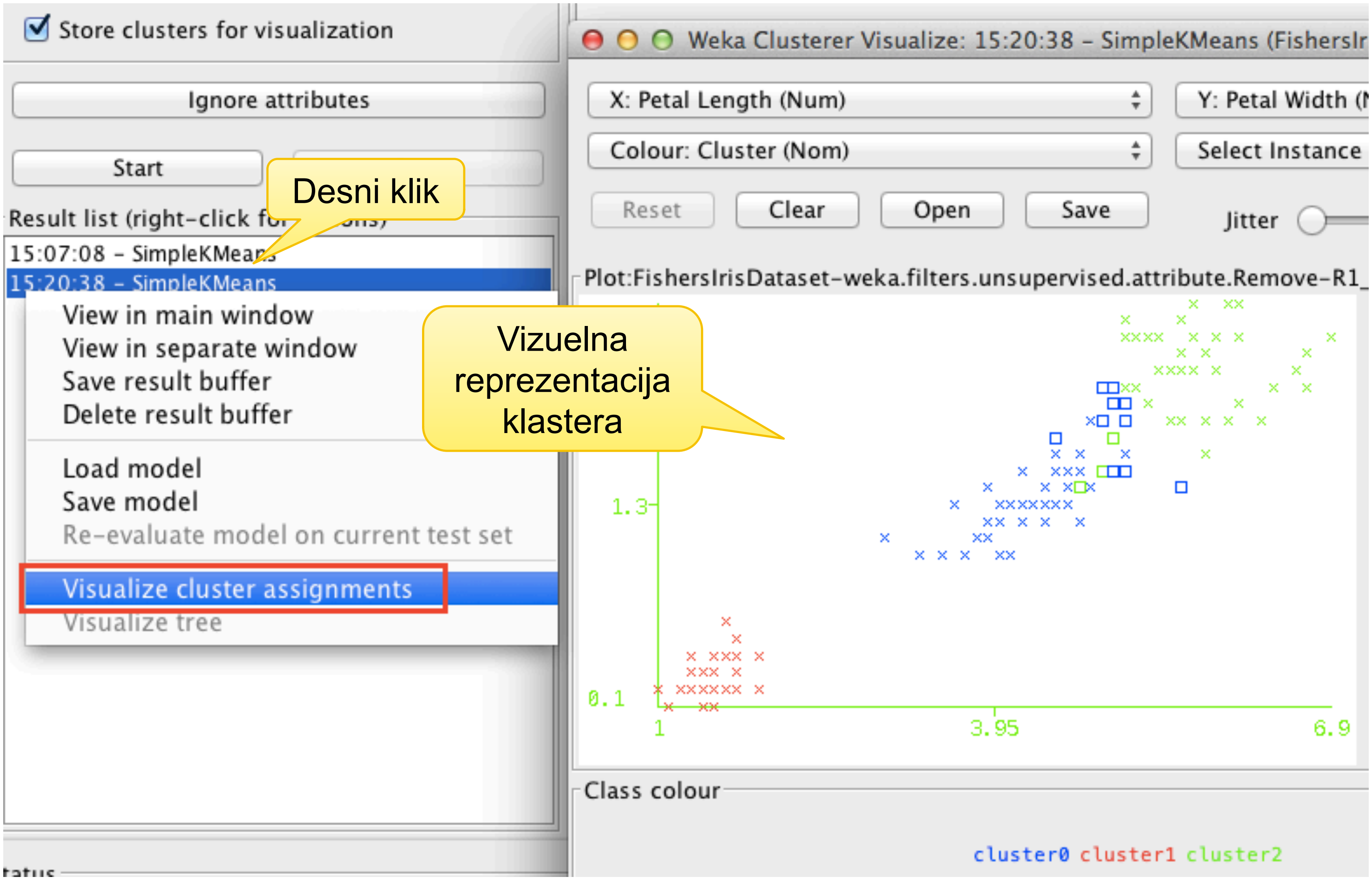
Rezultat klasterovanja



Evaluacija rezultata

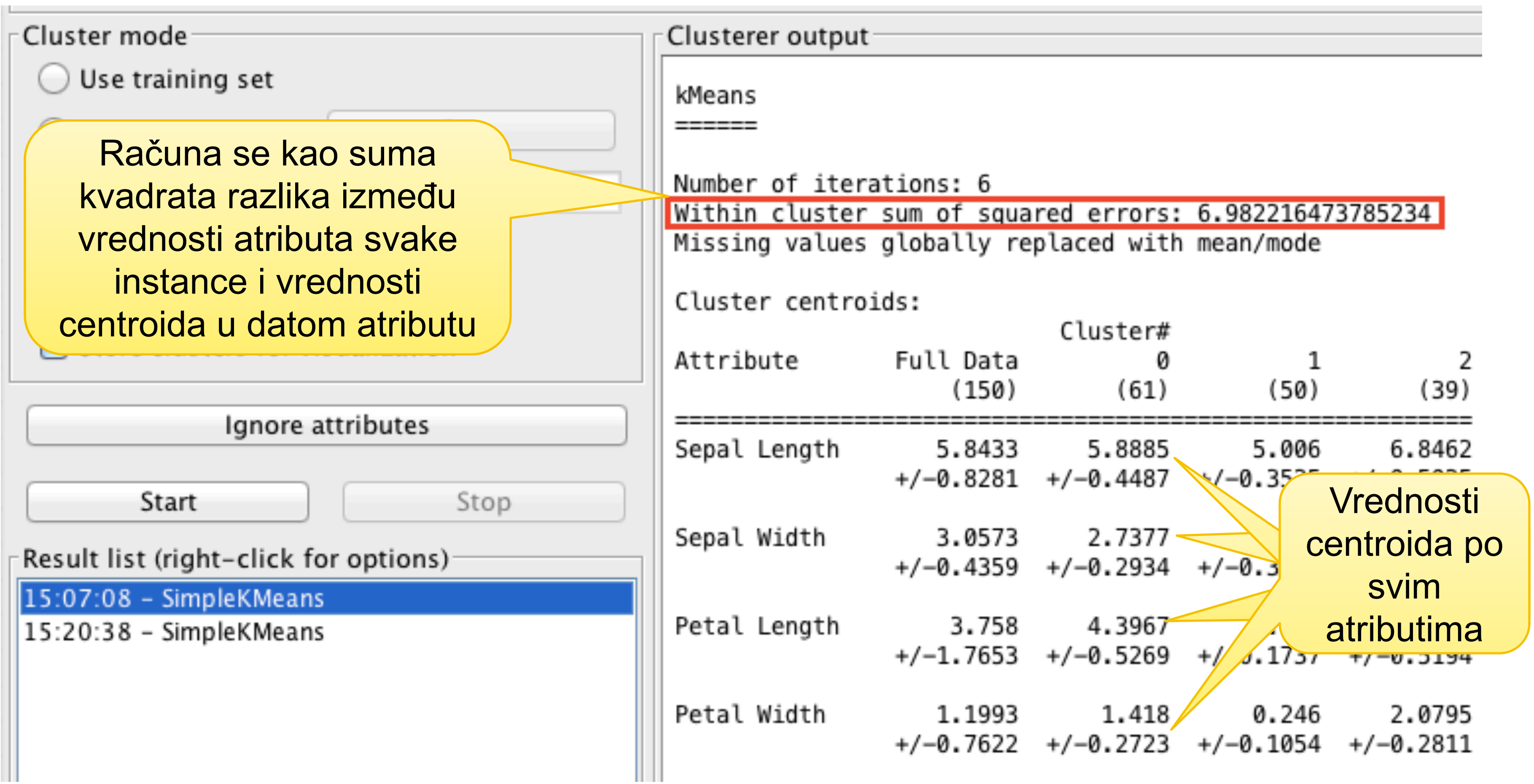


Vizuelizacija klastera



Procena uspešnosti klasterovanja

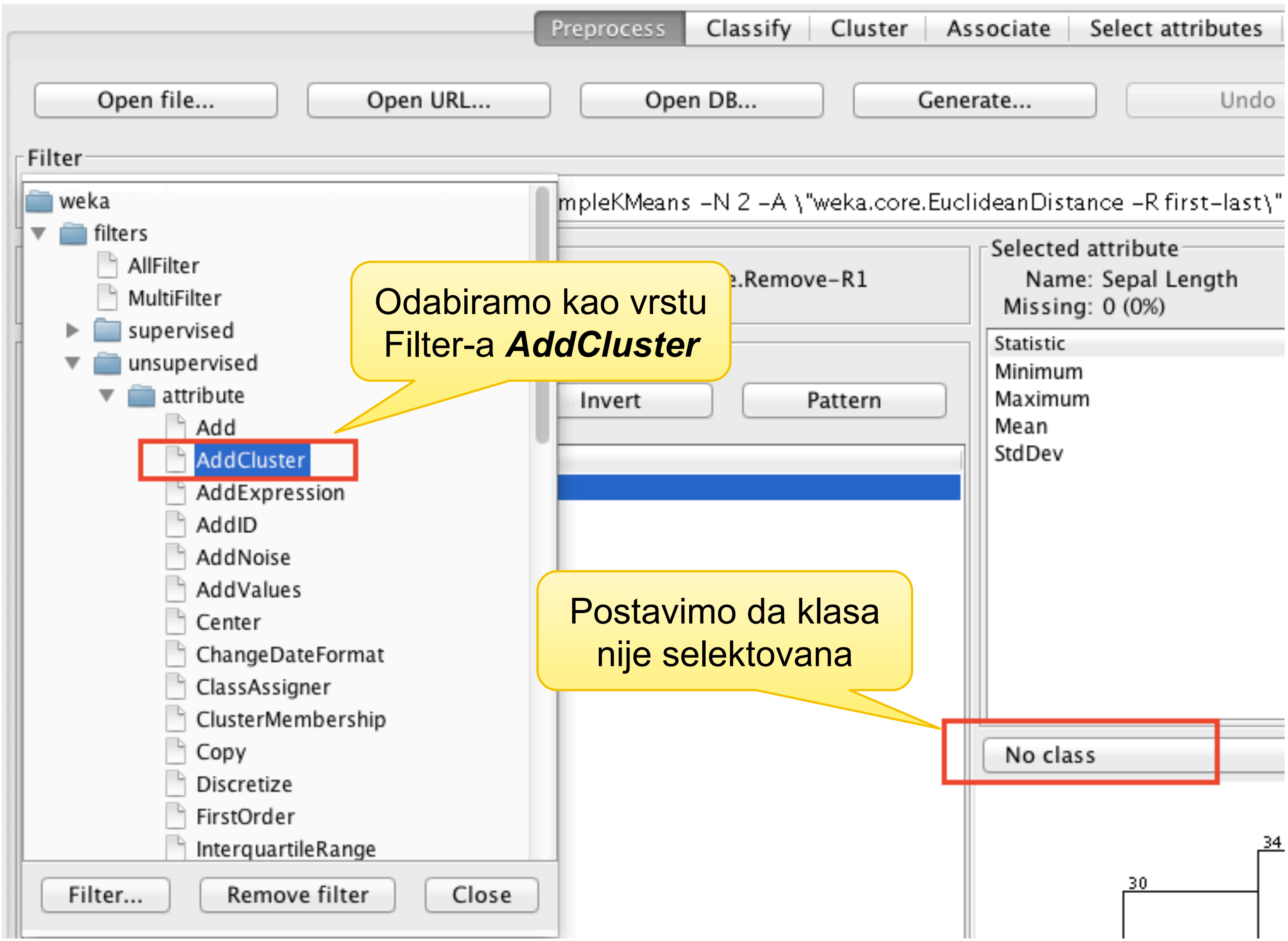
Within cluster sum of squared error (suma kvadrata greške unutar klastera) daje procenu kvaliteta dobijenih klastera



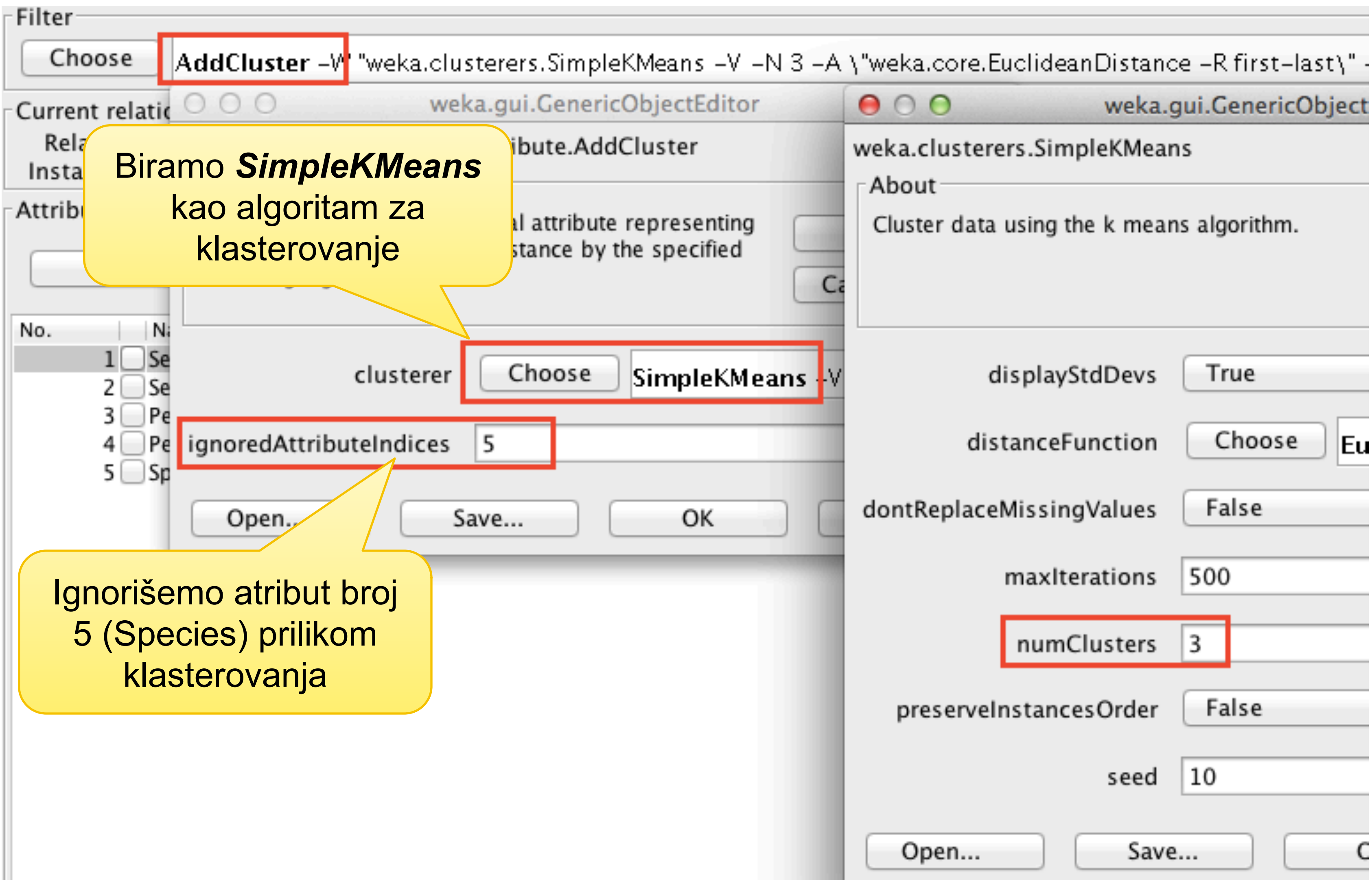
Kako proceniti dobar broj klastera?

Klasteri	Greška
1	55.6
2	12.1
3	7.0
4	5.5
5	5.0
6	4.8
7	4.7
8	4.2
9	4.1
10	3.6
20	1.7
50	0.6

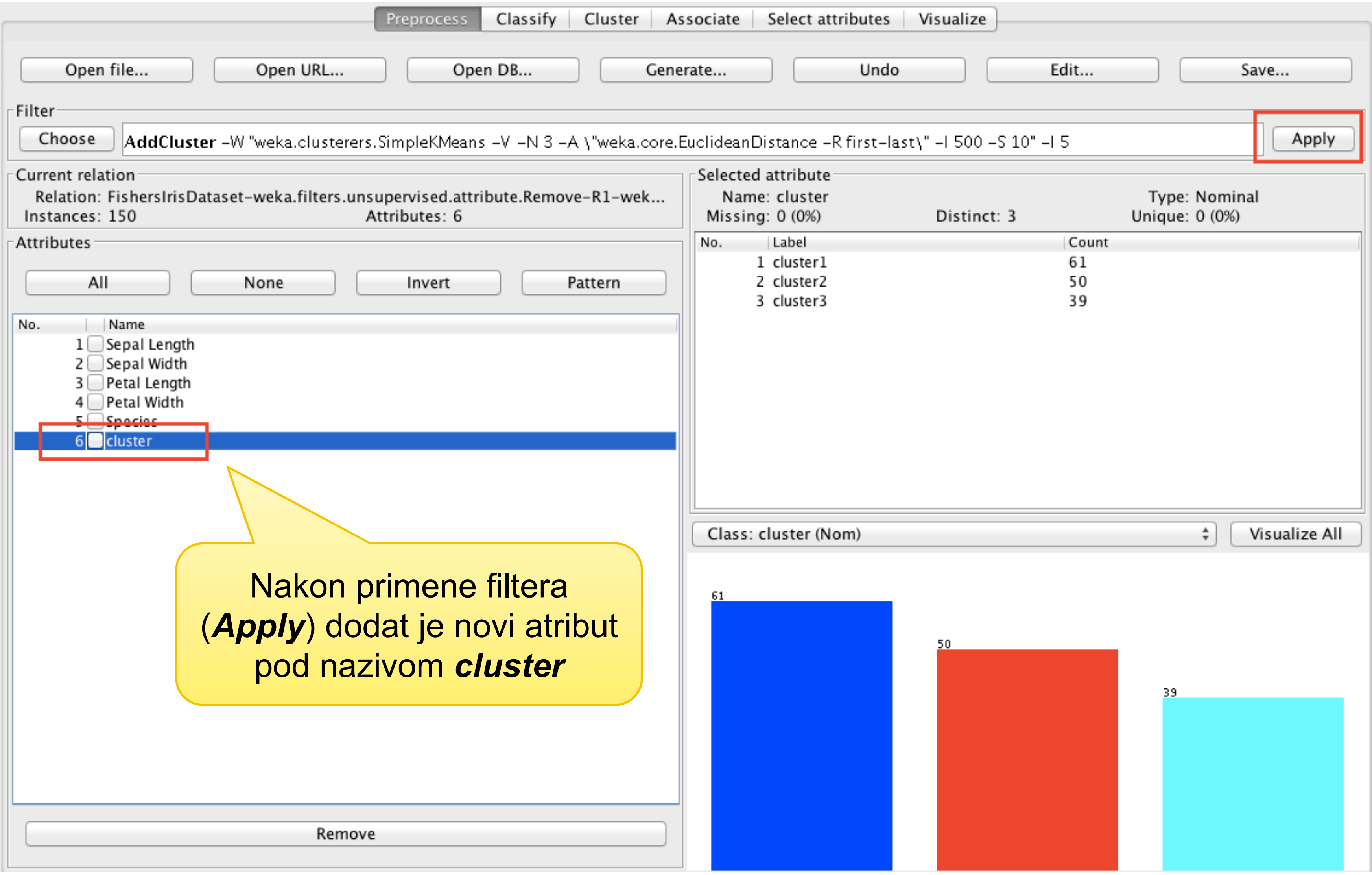
Korišćenje klastera za klasifikaciju



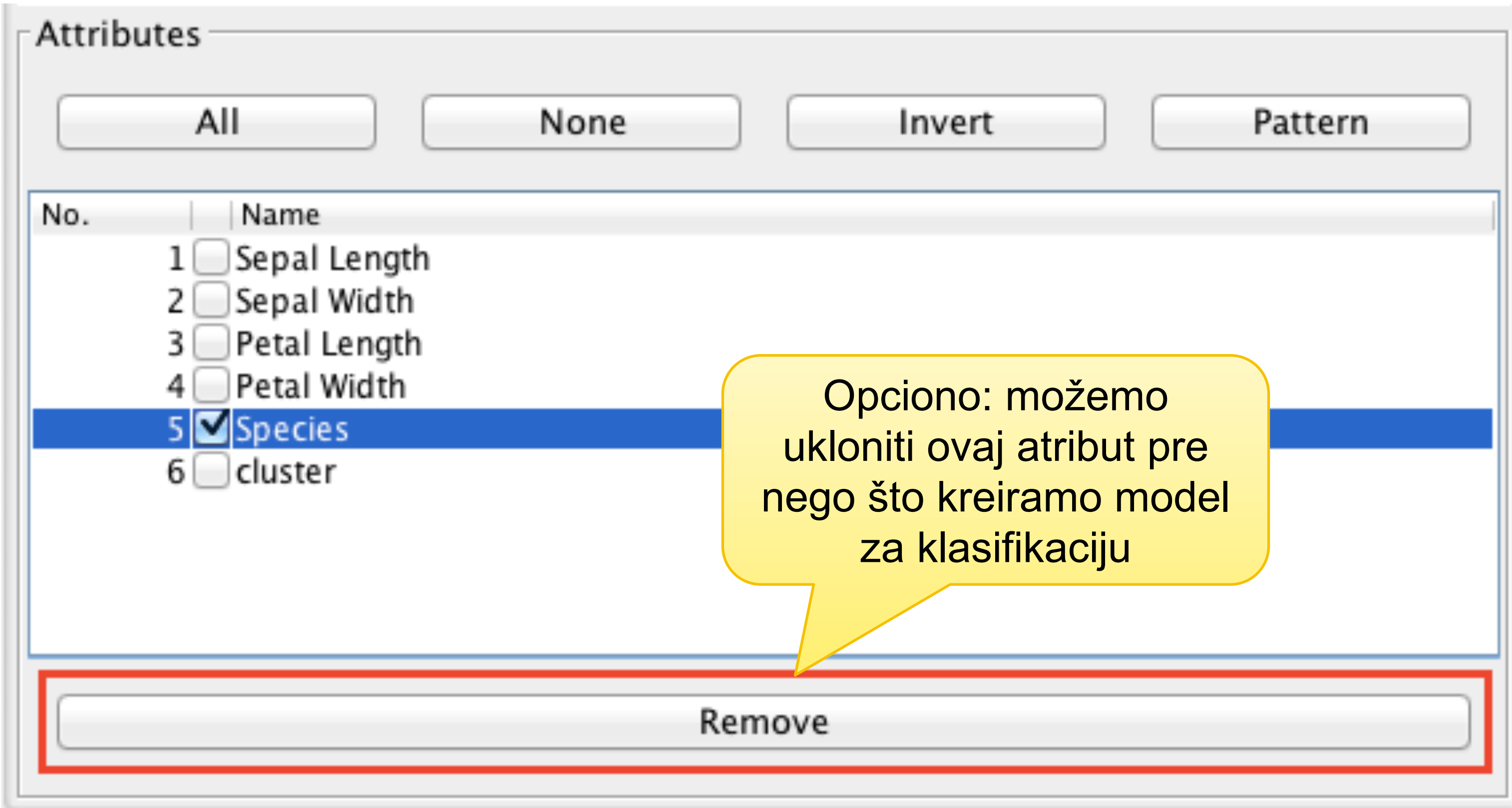
Korišćenje klastera za klasifikaciju



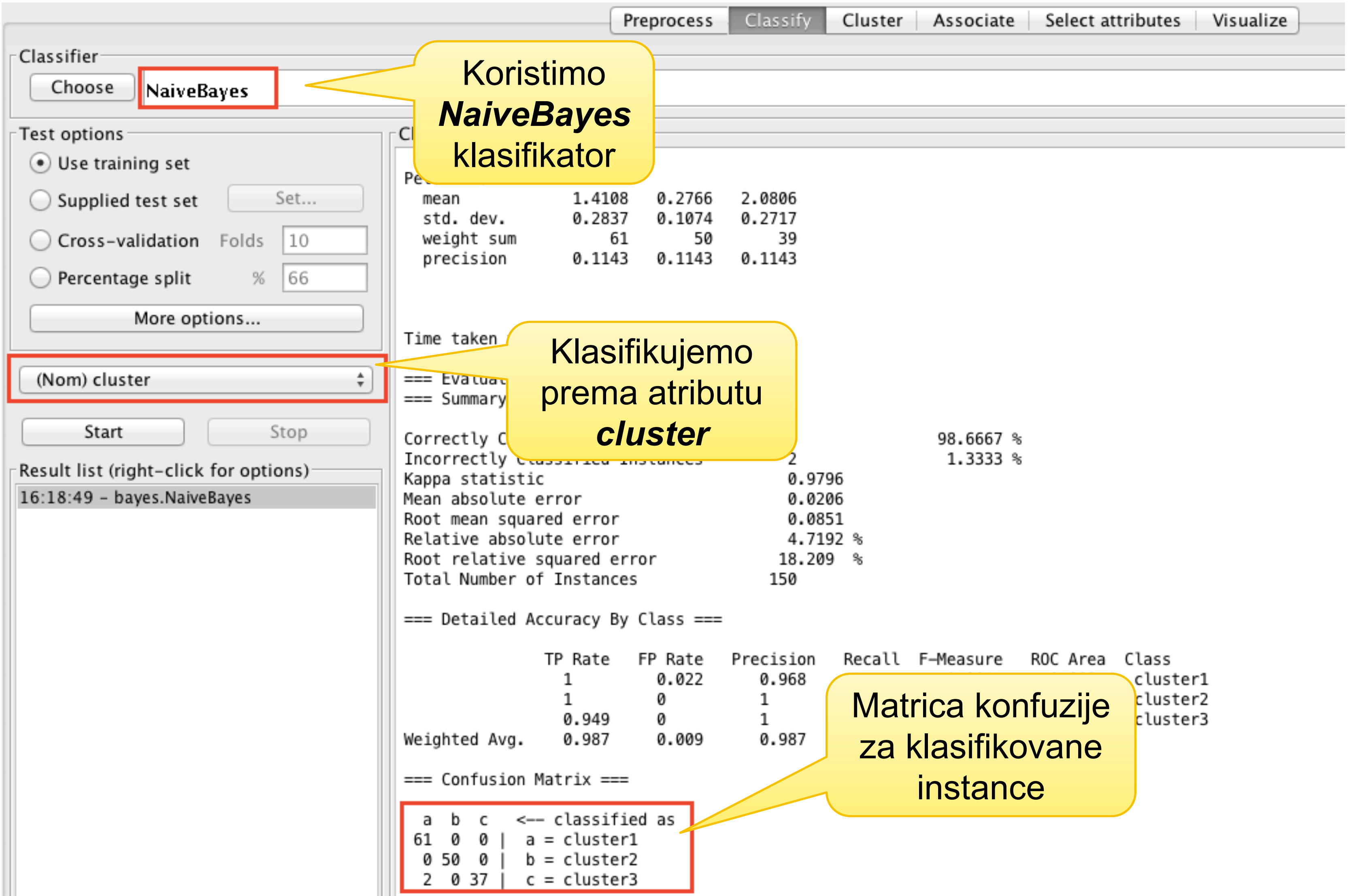
Korišćenje klastera za klasifikaciju



Korišćenje klastera za klasifikaciju



Korišćenje klastera za klasifikaciju



Expectation Maximization (EM)

Probabilistička klasterizacija - instance ne pripadaju jednom i samo jednom klasteru, već svaka instanca sa određenom verovatnoćom pripada svakom od klastera

Sastoji se iz dva koraka:

- E (expectation) korak – u ovom koraku podrazumevamo da znamo vrednosti parametara modela i na osnovu njih, za svaku instancu, računamo verovatnoću pripadanja svakom od klastera
- M (maximization) korak – na osnovu datih instanci, računamo (ponovo) vrednosti parametara modela; maksimizacija se odnosi na usklađivanje (parametara) modela sa datim podacima

Expectation Maximization (EM)

Postupak prilikom klasterovanja:

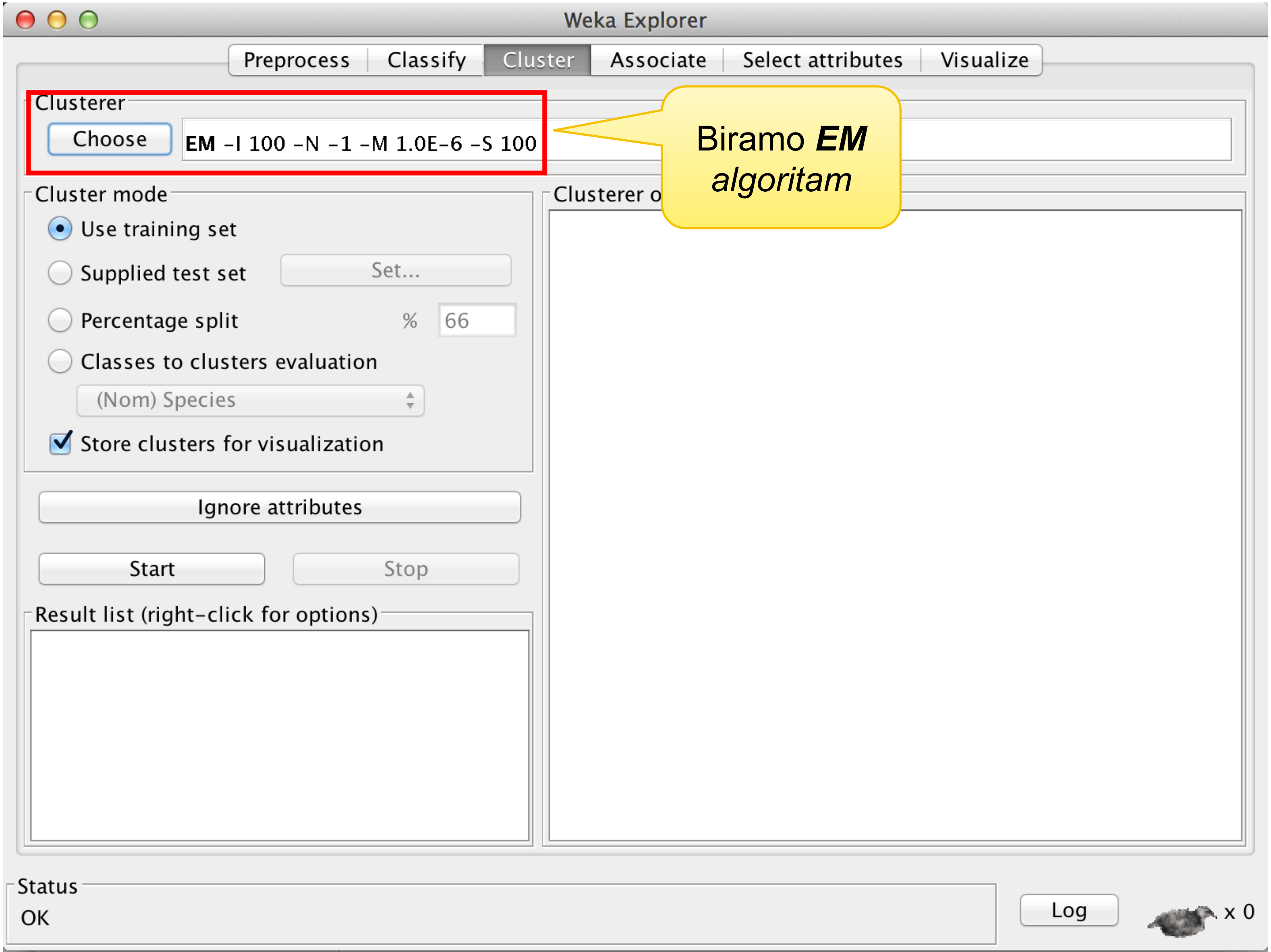
1) Inicijalno, definisati broj klastera (k) i nasumice izabrati vrednosti parametara modela (μ_i,σ_i, p_i, $i=1,k$)

2) Za date vrednosti parametara, za svaku instancu iz dataset-a, izračunati verovatnoću pripadanja svakom od klastera

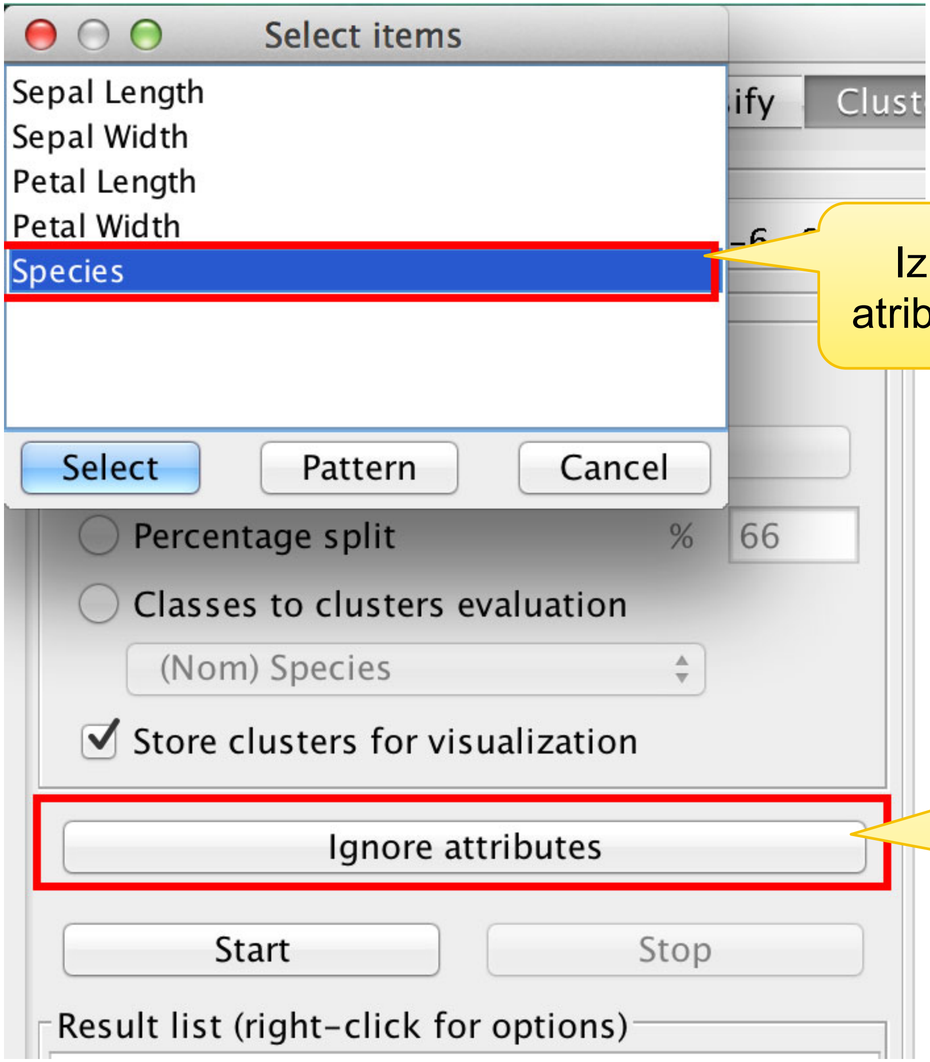
3) Na osnovu verovatnoća pripadnosti klasterima (instanci iz dataset-a), odrediti nove vrednosti parametara modela

Iterativno ponavljati korake 2) i 3) dok vrednosti parametara ne počnu da konvergiraju.

Korišćenje EM algoritma



Neuzimanje u obzir klase



Izuzimamo atribut **Species**

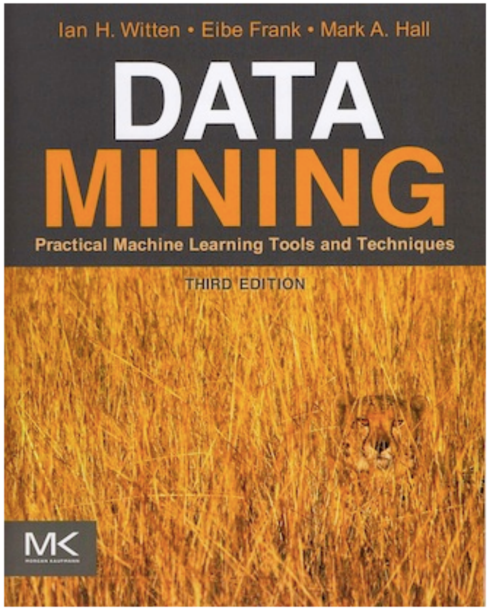
Selektovanje atributa koji neće biti korišćeni prilikom klasterovanja

Preporuke i zahvalnice

Weka Tutorials and Assignments @ The Technology Forge

- Link: http://www.technologyforge.net/WekaTutorials/

Witten, Ian H., Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. Elsevier, 2011.



(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