

Klasterizacija

NIKOLA MILIKIĆ

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

URL: <http://nikola.milikic.info>

Klasterizacija

Klasterizacija (eng. Clustering) spada u grupu tehnika nенадгledanог učenja i omogуćava 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 algoritam u Weka-i

FishersIrisDataset.arff

Preprocess Classify Cluster Associate Select attributes Visualize

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

Filter Choose None Apply

Current relation
Relation: FishersIrisDataset-weka.filters.unsupervised.attribute.Remove-R1
Instances: 150 Attributes: 5

Attributes
All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> Sepal Length
2	<input type="checkbox"/> Sepal Width
3	<input type="checkbox"/> Petal Length
4	<input type="checkbox"/> Petal Width
5	<input type="checkbox"/> Species

Remove

Selected attribute
Name: Sepal Length Type: Numeric
Missing: 0 (0%) Distinct: 35 Unique: 9 (6%)
Statistic Value
Minimum 4.3
Maximum 7.9
Mean 5.843
StdDev 0.828

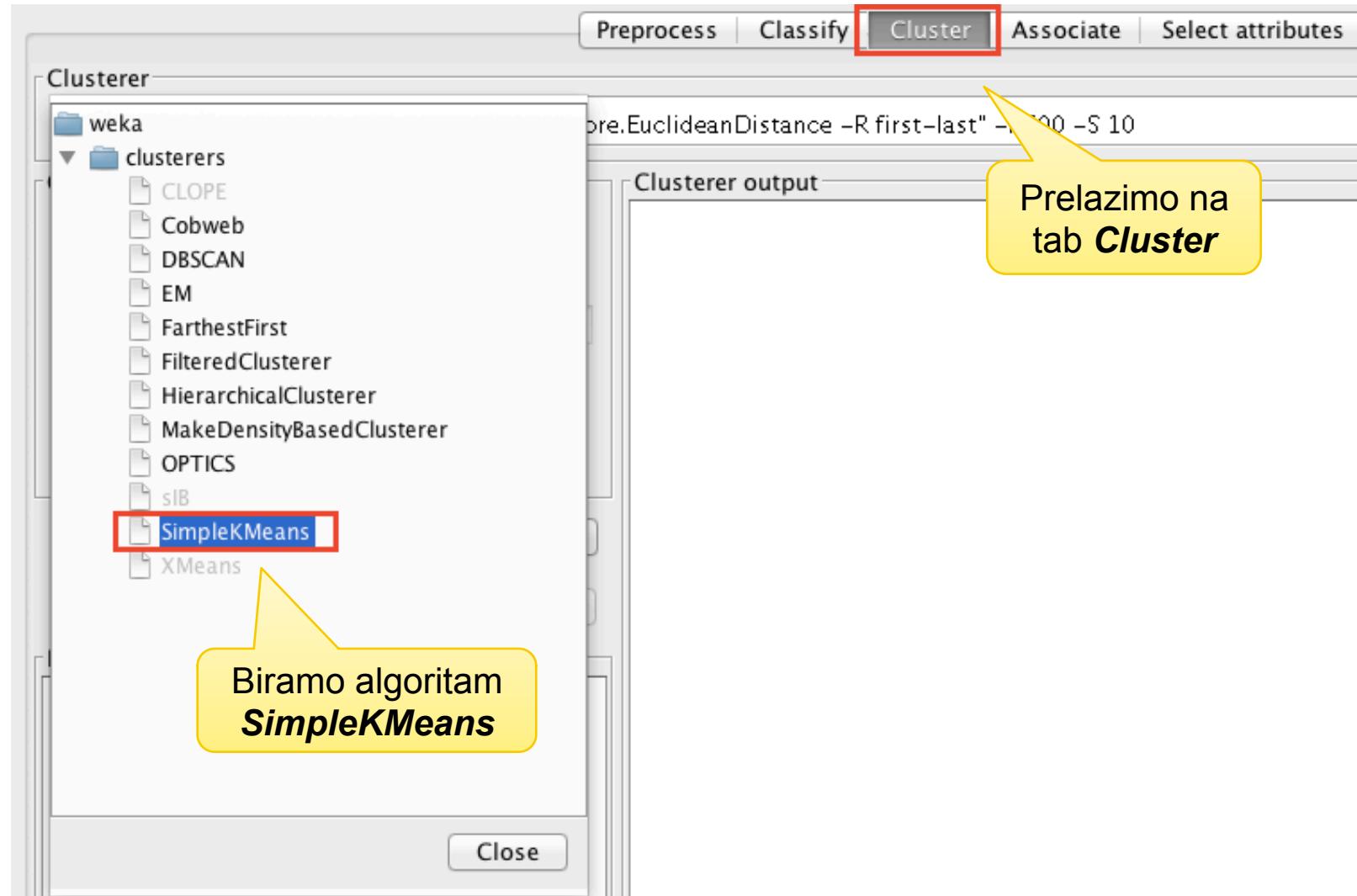
Class: Species (Nom) Visualize All

4.3 6.1 7.9

16 30 34 28 25 10 7

Status: OK Log x 0

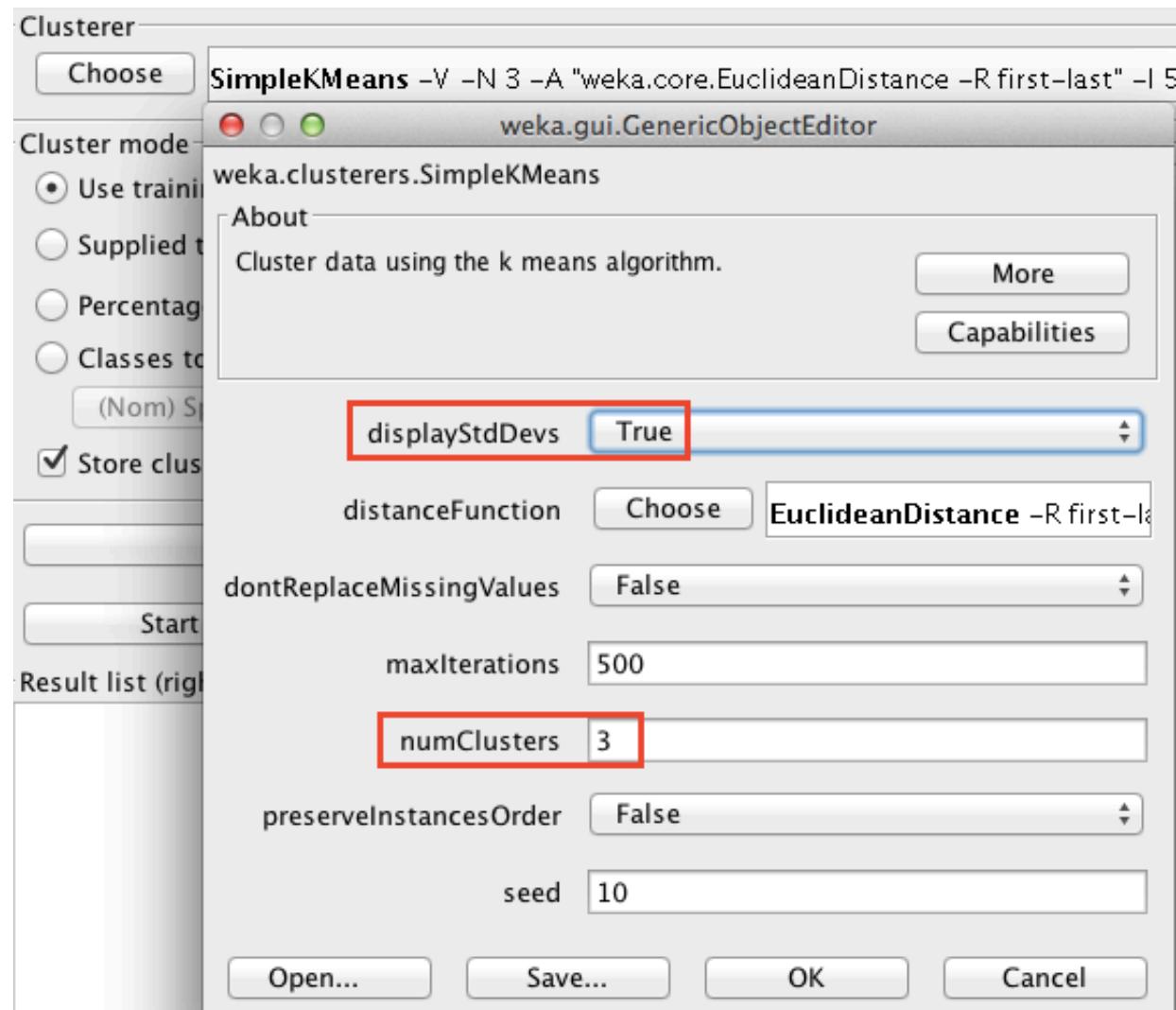
Odabir algoritma za klasterovanje



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

Vršimo klasterovanje nad učitanim podacima

Cluster mode

- Use training set
- Supplied test set
- Percentage split % 66
- Classes to clusters evaluation

Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

15:07:08 – SimpleKMeans

Select items

- Sepal Length
- Sepal Width
- Petal Length
- Petal Width
- Species**

== Run information ==

Scheme:weka.clusterers.SimpleKMeans -V -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Relation: FishersIrisDataset-weka.filters.unsupervised.attribute

Instances: 150

Attributes: 5

Sepal Length

Sepal Width

Petal Length

Petal Width

Ignored: Species

Test mode: evaluate on training data

== Model and evaluation on training set ==

of iterations: 6

Within cluster sum of squared errors: 6.982216473785234

Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (150)	Cluster# 0 (61)	Cluster# 1 (50)	Cluster# 2 (39)
Sepal Length	5.8433 +/-0.8281	5.8885 +/-0.4487	5.006 +/-0.3525	6.8462 +/-0.5025
Sepal Width	3.0573	2.7377	2.428	3.0821

Rezultat klasterovanja

Cluster mode

Use training set

Supplied test set Set...

Percentage split % 66

Random split

Calculate standard deviation

Centroidi svakog klastera i njihove standardne devijacije

Ignore attributes

Start Stop

Result list (right-click for options)

15:07:08 – SimpleKMeans

Clusterer output

kMeans

====

Number of iterations: 6

Within cluster sum of squared errors: 6.982216473785234

Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (150)	Cluster# 0 (61)	1 (50)	2 (39)
Sepal Length	5.8433 +/-0.8281	5.8885 +/-0.4487	5.006 +/-0.3525	6.8462 +/-0.5025
Sepal Width	3.0573 +/-0.4359	2.7377 +/-0.2934	3.428 +/-0.3791	3.0821 +/-0.2799
Petal Length	3.758 +/-1.7653	4.3967 +/-0.5269	1.462 +/-0.1737	5.7026 +/-0.5194
Petal Width	1.1993 +/-0.7622	1.418 +/-0.2723	0.246 +/-0.1054	2.0795 +/-0.2811

Time taken to build model (full training data) : 0.04 seconds

== Model and evaluation on training set ==

Clustered Instances

0	61 (41%)
1	50 (33%)
2	39 (26%)

Broj instanci u svakom klasteru

Evaluacija rezultata

Cluster mode

- Use training set
- Supplied test set
- Percentage split % 66
- Classes to clusters evaluation

(Nom) Species

Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

- 15:07:08 - SimpleKMeans
- 15:20:38 - SimpleKMeans

Selektujemo atribut sa kojim želimo da poredimo rezultate

Time taken to build model (full training data) : 0.02 seconds

== Model and evaluation on training set ==

Clustered Instances

0	61	(41%)
1	50	(33%)
2	39	(26%)

U kojim klasterima su smeštene koje klase

Imena klasa koje su dodeljene klasterima

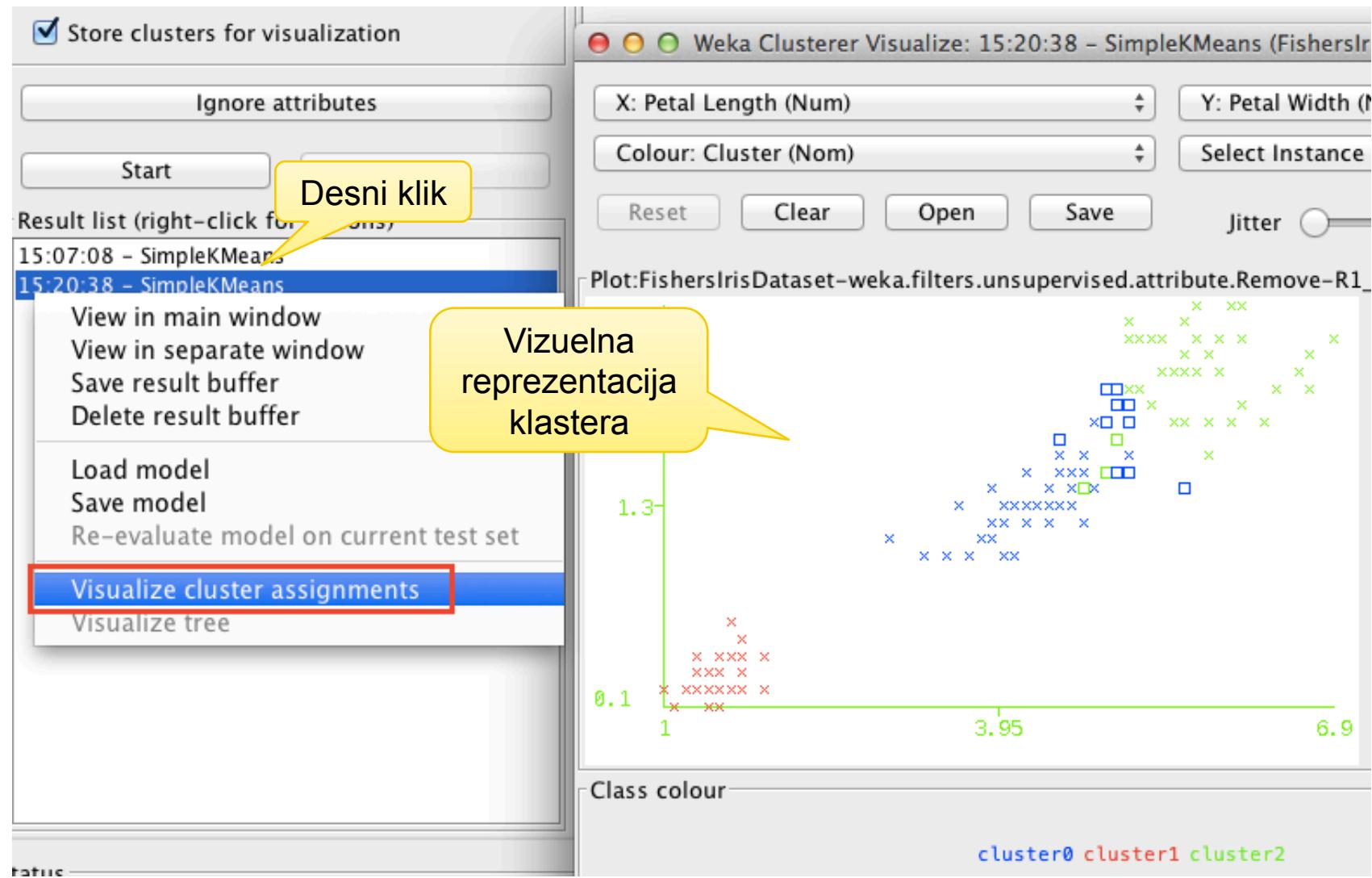
Class attribute: Species
Classes to Clusters:

0	1	2	<-- assigned to cluster
0	50	0	setosa
47	0	3	versicolor
14	0	36	virginica

Cluster 0 <-- versicolor
Cluster 1 <-- setosa
Cluster 2 <-- virginica

Incorrectly clustered instances : 17.0 11.3333 %

Vizuelizacija klastera



Procena uspešnosti klasterovanja

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

Cluster mode

Use training set

Računa se kao suma kvadrata razlika između vrednosti atributa svake instance i vrednosti centroida u datom atributu

Ignore attributes

Start Stop

Result list (right-click for options)

15:07:08 - SimpleKMeans
15:20:38 - SimpleKMeans

Clusterer output

kMeans

====

Number of iterations: 6

Within cluster sum of squared errors: 6.982216473785234

Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (150)	Cluster# 0 (61)	1 (50)	2 (39)
Sepal Length	5.8433 +/-0.8281	5.8885 +/-0.4487	5.006 +/-0.3525	6.8462 +/-0.5025
Sepal Width	3.0573 +/-0.4359	2.7377 +/-0.2934	+/-0.3	
Petal Length	3.758 +/-1.7653	4.3967 +/-0.5269	+/-0.1737 +/-0.5194	
Petal Width	1.1993 +/-0.7622	1.418 +/-0.2723	0.246 +/-0.1054	2.0795 +/-0.2811

Vrednosti centroida po svim atributima

```
Within cluster sum of squared errors: 6.982216473785234
```

Attribute	Full Data (150)	Cluster# 0 (61)	1 (50)	2 (39)
Sepal Length	5.8433 +/-0.8281	5.8885 +/-0.4487	5.006 +/-0.3525	6.8462 +/-0.5025
Sepal Width	3.0573 +/-0.4359	2.7377 +/-0.2934	+/-0.3	
Petal Length	3.758 +/-1.7653	4.3967 +/-0.5269	+/-0.1737 +/-0.5194	
Petal Width	1.1993 +/-0.7622	1.418 +/-0.2723	0.246 +/-0.1054	2.0795 +/-0.2811

Kako proceniti dobar broj klastera?



Korišćenje klastera za klasifikaciju

The screenshot shows the Weka Filter interface. The top menu bar includes Preprocess, Classify, Cluster, Associate, and Select attributes. Below the menu are buttons for Open file..., Open URL..., Open DB..., Generate..., and Undo. A 'Filter' section on the left lists categories like weka, filters, supervised, unsupervised, and attribute. Under attribute, the 'AddCluster' option is highlighted with a red box and a yellow callout bubble containing the text: "Odabiramo kao vrstu Filter-a **AddCluster**". To the right, a list of filters includes Remove-R1, Invert, and Pattern. A 'Selected attribute' panel shows Sepal Length with 0% missing data. A 'Statistic' panel lists Minimum, Maximum, Mean, and StdDev. At the bottom, a 'No class' button is highlighted with a red box and a yellow callout bubble containing the text: "Postavimo da klasa nije selektovana".

Preprocess Classify Cluster Associate Select attributes

Open file... Open URL... Open DB... Generate... Undo

Filter

weka

filters

- AllFilter
- MultiFilter
- supervised
- unsupervised
- attribute
 - Add
 - AddCluster**
 - AddExpression
 - AddID
 - AddNoise
 - AddValues
 - Center
 - ChangeDateFormat
 - ClassAssigner
 - ClusterMembership
 - Copy
 - Discretize
 - FirstOrder
 - InterquartileRange

Invert Pattern

Selected attribute

Name: Sepal Length
Missing: 0 (0%)

Statistic

Minimum
Maximum
Mean
StdDev

Odabiramo kao vrstu Filter-a **AddCluster**

Postavimo da klasa nije selektovana

No class

Filter... Remove filter Close

30 34

Korišćenje klastera za klasifikaciju

The screenshot shows the Weka interface for configuring a classification model. On the left, the 'GenericObjectEditor' window is open, showing the 'AddCluster' step of a pipeline. A yellow callout points to the 'clusterer' field, which is set to 'SimpleKMeans'. Another yellow callout points to the 'ignoredAttributeIndices' field, which contains the value '5'. On the right, a separate 'GenericObjectEditor' window is open for the 'SimpleKMeans' class, showing its configuration options. The 'numClusters' field is highlighted with a red box and set to '3'.

Biramo **SimpleKMeans** kao algoritam za klasterovanje

Ignorišemo atribut broj 5 (Species) prilikom klasterovanja

weka.clusterers.SimpleKMeans

About

Cluster data using the k means algorithm.

displayStdDevs True

distanceFunction Choose Eu

dontReplaceMissingValues False

maxIterations 500

numClusters 3

preserveInstancesOrder False

seed 10

Korišćenje klastera za klasifikaciju

Screenshot of the Weka interface showing the Preprocess tab selected. The 'Selected attribute' panel shows a new attribute 'cluster' with three distinct values: cluster1 (Count: 61), cluster2 (Count: 50), and cluster3 (Count: 39). A yellow callout box points to the 'cluster' entry in the list of selected attributes.

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

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

Filter

Choose AddCluster -W "weka.clusterers.SimpleKMeans -V -N 3 -A \\"weka.core.EuclideanDistance -R first-last\\\" -I 500 -S 10" -I 5 **Apply**

Current relation

Relation: FishersIrisDataset-weka.filters.unsupervised.attribute.Remove-R1-weka... Instances: 150 Attributes: 6

Attributes

All None Invert Pattern

No.	Name
1	<input type="checkbox"/> Sepal Length
2	<input type="checkbox"/> Sepal Width
3	<input type="checkbox"/> Petal Length
4	<input type="checkbox"/> Petal Width
5	<input type="checkbox"/> Species
6	<input checked="" type="checkbox"/> cluster

Selected attribute

Name: cluster
Missing: 0 (0%) Distinct: 3 Type: Nominal Unique: 0 (0%)

No.	Label	Count
1	cluster1	61
2	cluster2	50
3	cluster3	39

Class: cluster (Nom) **Visualize All**

Nakon primene filtera (**Apply**) dodat je novi atribut pod nazivom **cluster**

Remove

Korišćenje klastera za klasifikaciju

Attributes

All None Invert Pattern

No.	Name
1	<input type="checkbox"/> Sepal Length
2	<input type="checkbox"/> Sepal Width
3	<input type="checkbox"/> Petal Length
4	<input type="checkbox"/> Petal Width
5	<input checked="" type="checkbox"/> Species
6	<input type="checkbox"/> cluster

Opciono: možemo ukloniti ovaj atribut pre nego što kreiramo model za klasifikaciju

Remove

Korišćenje klastera za klasifikaciju

Classifier Preprocess Classify Cluster Associate Select attributes Visualize

Choose **NaiveBayes**

Test options

Use training set

Supplied test set Set...

Cross-validation Folds 10

Percentage split % 66

More options...

(Nom) cluster

Start Stop

Result list (right-click for options)

16:18:49 - bayes.NaiveBayes

Koristimo NaiveBayes klasifikator

Time taken

==== Evaluation by class ====
cluster1 98.6667 %
cluster2 1.3333 %

Correctly Classified Instances 150
Incorrectly Classified Instances 2
Kappa statistic 0.9796
Mean absolute error 0.0206
Root mean squared error 0.0851
Relative absolute error 4.7192 %
Root relative squared error 18.209 %
Total Number of Instances 150

==== Detailed Accuracy By Class ====

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
cluster1	1	0.022	0.968				cluster1
cluster2	1	0	1				cluster2
cluster3	0.949	0	1				cluster3
Weighted Avg.	0.987	0.009	0.987				

==== Confusion Matrix ====

	a	b	c	<-- classified as
a	61	0	0	a = cluster1
b	0	50	0	b = cluster2
c	2	0	37	c = cluster3

Matrica konfuzije za klasifikovane instance

Expectation Maximization (EM)

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

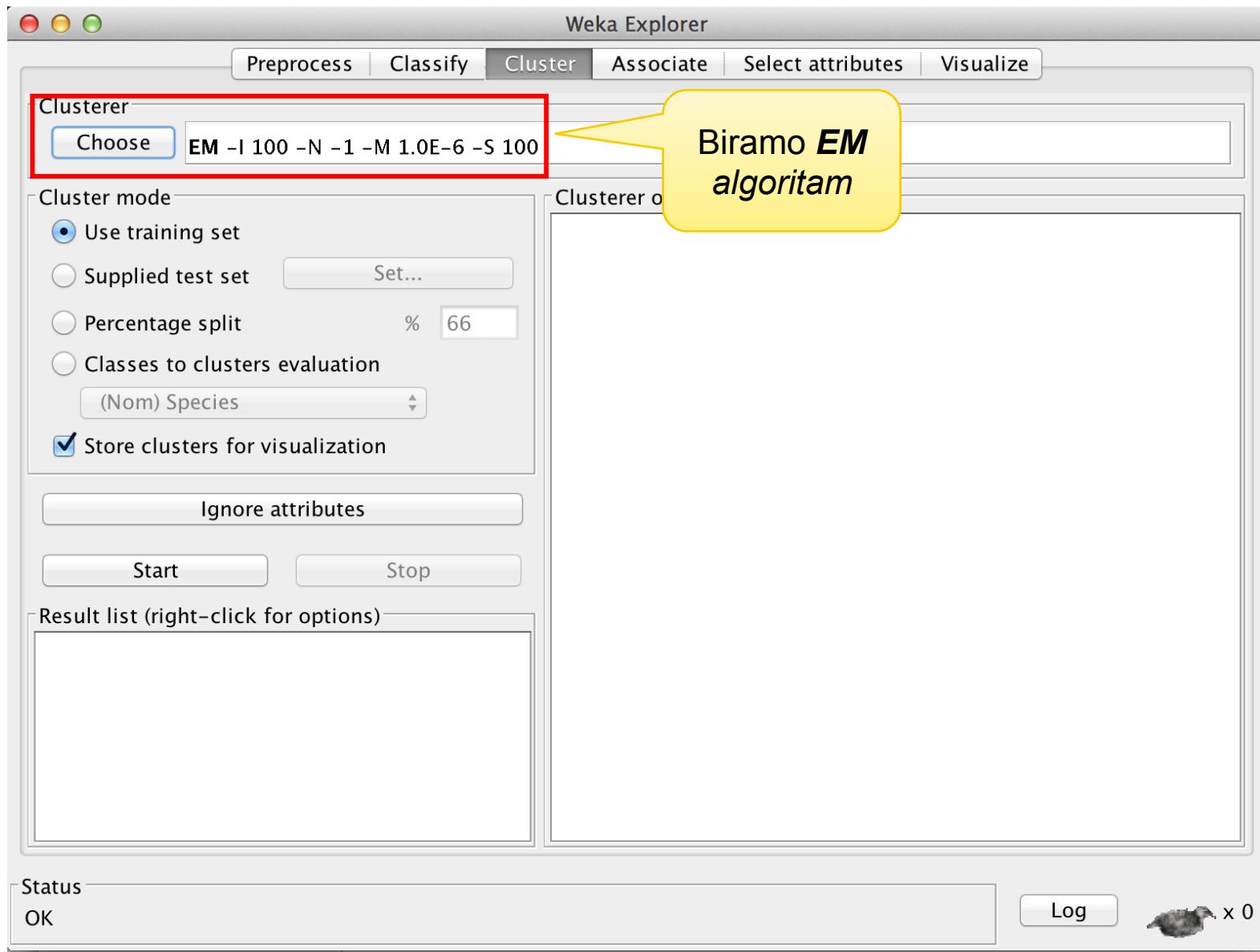
Expectation Maximization (EM)

Postupak prilikom klasterovanja:

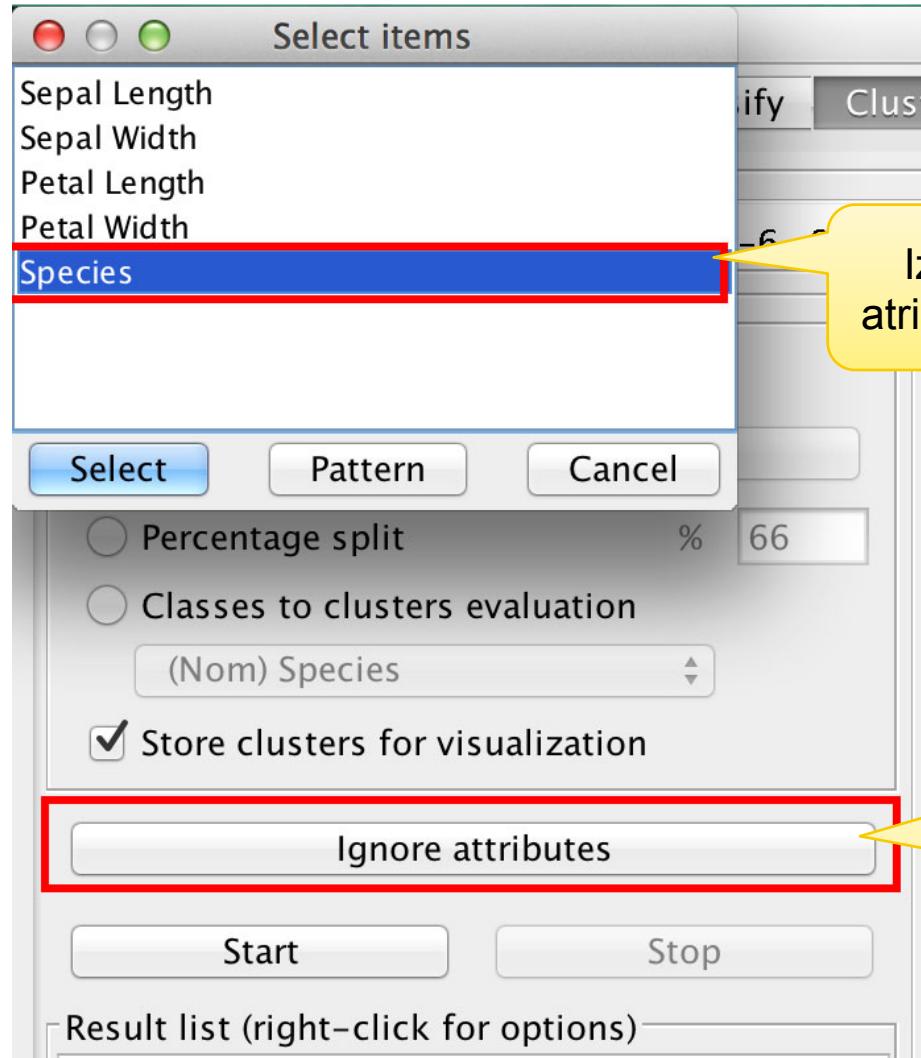
- 1) Inicijalno, definisati broj klastera (k) I nasumice izabrati vrednosti parametara modela ($\mu_i, \sigma_i, p_i, i=1,k$)
- 2) Za date vrednosti parametara, za svaku instancu iz dataset-a, izračunati verovatnoću pripadanja svakom od klastera
- 3) Na osnovu verovatnoća pripadnosti klasterima (instanci iz dataset-a), odrediti nove vrednosti parametara modela

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

Korišćenje EM algoritma



Ne uzimanje u obzir klase



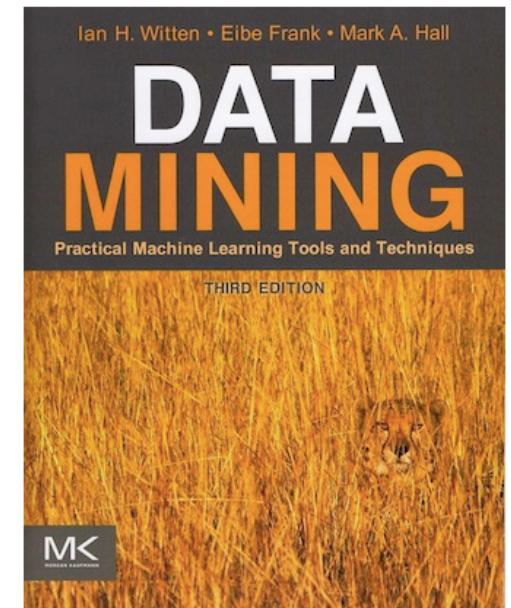
Selektovanje atributa koji neće biti korišćeni prilikom klasterovanja

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>