

**Uč enje
pravila**

Predavač:

dr. Dragan Gamberger

E-mail dragan.gamberger@irb.hr

tel. 4561 142

Ulaz: Podaci moraju biti prikazani tablično

varijable (atributi, deskriptori)

	IME	STAR.	SPOL	OBRAZ	ZANIM	TEZINA	PRIHOD	PUSAC
primjeri (instance)	jan	30	muski	nize	radnik	27.3	14000	da
	janko	55.5	muski	srednje	radnik	90	20000	ne
	zora	?	zenski	visoko	ucitelj	65.2	1000	ne
	tanja	18	zenski	srednje	student	55.1	0	ne
	tom	70	muski	visoko	?	60	9000	da
	tomi	35	muski	srednje	prof	33	16000	ne
	stev	42.2	muski	nize	vozac	27	7500	da
	marc	29	muski	?	konobar	31	8300	da

nominalni
(kategorički,
string)

ordinalni
(integer)

numerički (float)

Učenje pravila

A) Poznata nam je ciljna varijabla i ona je u obliku 2 ili više klasa (idealno dvije klase).
Zanimaju nas modeli (pravila) koji razdvajaju te klase

-> prediktivno učenje

AKO (ZAN=radnik) \wedge (TEZINA>75) -> PUSAC

B) Nije definirana ciljna varijabla i zanimaju nas sve značajne veze (pravila) među varijablama koje postoje

-> deskriptivno (asocijativno) učenje

PUSAC=da \wedge SPOL=muski (često se pojavljuje zajedno)

Asocijativno učenje pravila

Postupak je razvijen originalno za analizu
"potrošačke košarice"

osoba A je kupila: mlijeko, kruh, novine

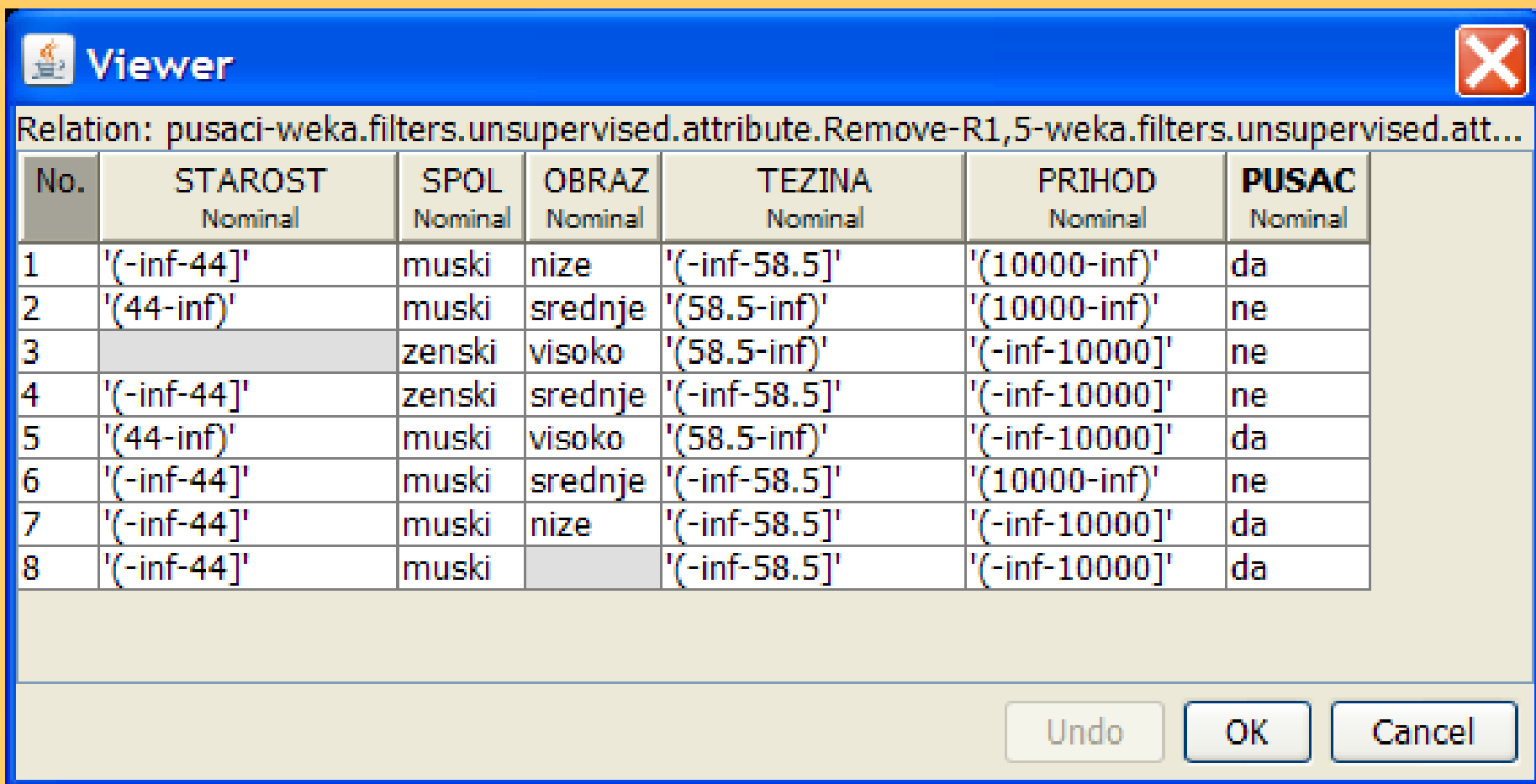
osoba B je kupila: jabuke, tijesto, maslac, ulje

.....

Dobro: otkriva sve značajne veze
radi brzo za velike skupove podataka
može poslužiti i kao osnova za
prediktivna pravila

Loše: obično smo zatrpani ogromnim brojem
rezultata
može raditi samo sa kategoričkim
atributima

Priprema podataka



Relation: pusaci-weka.filters.unsupervised.attribute.Remove-R1,5-weka.filters.unsupervised.att...

No.	STAROST Nominal	SPOL Nominal	OBRAZ Nominal	TEZINA Nominal	PRIHOD Nominal	PUSAC Nominal
1	'(-inf-44]'	muski	nize	'(-inf-58.5]'	'(10000-inf)'	da
2	'(44-inf)'	muski	srednje	'(58.5-inf)'	'(10000-inf)'	ne
3		zenski	visoko	'(58.5-inf)'	'(-inf-10000]'	ne
4	'(-inf-44]'	zenski	srednje	'(-inf-58.5]'	'(-inf-10000]'	ne
5	'(44-inf)'	muski	visoko	'(58.5-inf)'	'(-inf-10000]'	da
6	'(-inf-44]'	muski	srednje	'(-inf-58.5]'	'(10000-inf)'	ne
7	'(-inf-44]'	muski	nize	'(-inf-58.5]'	'(-inf-10000]'	da
8	'(-inf-44]'	muski		'(-inf-58.5]'	'(-inf-10000]'	da

Undo OK Cancel

Izbačeni su atributi IME i ZANIMANJE a
STAROST, TEZINA i PRIHOD su kategorizirani

Rezultat: 10 asocijacija

Weka Explorer

Preprocess Classify Cluster **Associate** Select attributes Visualize

Associator

Choose **Apriori** -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Start Stop

Result list (right-click for details)

16:01:28 - Apriori

Associator output

```
Size of set of large itemsets L(2): 12
Size of set of large itemsets L(3): 6
Size of set of large itemsets L(4): 1

Best rules found:

1. TEZINA='(-inf-58.5]' 5 ==> STAROST='(-inf-44]' 5    conf:(1)
2. STAROST='(-inf-44]' 5 ==> TEZINA='(-inf-58.5]' 5    conf:(1)
3. PUSAC=da 4 ==> SPOL=muski 4    conf:(1)
4. SPOL=muski TEZINA='(-inf-58.5]' 4 ==> STAROST='(-inf-44]' 4    conf:(1)
5. STAROST='(-inf-44]' SPOL=muski 4 ==> TEZINA='(-inf-58.5]' 4    conf:(1)
6. PRIHOD='(10000-inf)' 3 ==> SPOL=muski 3    conf:(1)
7. OBRAZ=srednje 3 ==> PUSAC=ne 3    conf:(1)
8. STAROST='(-inf-44]' PUSAC=da 3 ==> SPOL=muski 3    conf:(1)
9. TEZINA='(-inf-58.5]' PRIHOD='(-inf-10000]' 3 ==> STAROST='(-inf-44]' 3    conf:(1)
10. STAROST='(-inf-44]' PRIHOD='(-inf-10000]' 3 ==> TEZINA='(-inf-58.5]' 3    conf:(1)
```

Status

OK

Log x 0

Postupak generiranja asocijacijskih pravila

1) generirati skup svih čestih skupova

kruh - mlijeko
jabuke - tijesto
kruh - mlijeko - novine

1. sloj kruh, mlijeko, jabuke, tijesto, novine, ...
2. sloj kruh-mlijeko, kruh-novine, mlijeko-novine, jabuke-tijesto,
3. sloj kruh-mlijeko-novine.

postupak je jednostavan i brz

korisnik mora definirati minimalnu podršku
($n_c/N = 20\%$)

ako neka kategorija nije česta sama onda ona
ne može graditi niti složene česte kategorije !!

Postupak generiranja asocijacijskih pravila

2) Za svaki česti skup sa barem dvije kategorije provjerava se

$$\frac{N(\text{kruh} - \text{mlijeko} - \text{novine})}{N(\text{kruh} - \text{mlijeko})} = \text{pouzdanost (confidence)}$$

Sve relacije veće pouzdanosti (recimo od 75%) generiraju pravilo oblika

AKO kruh ^ mlijeko -> novine

Rezultat: drugi dio

Weka Explorer

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

Associator

Choose **Apriori** -N 30 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Start Stop

Result list (right-click for details)

- 16:01:28 - Apriori
- 16:04:20 - Apriori**

Associator output

```
9. TEZINA='(-inf-58.5]' PRIHOD='(-inf-10000]' 3 ==> STAROST='(-inf-44]' 3    conf:(1)
10. STAROST='(-inf-44]' PRIHOD='(-inf-10000]' 3 ==> TEZINA='(-inf-58.5]' 3    conf:(1)
11. TEZINA='(-inf-58.5]' PUSAC=da 3 ==> STAROST='(-inf-44]' 3    conf:(1)
12. STAROST='(-inf-44]' PUSAC=da 3 ==> TEZINA='(-inf-58.5]' 3    conf:(1)
13. TEZINA='(-inf-58.5]' PUSAC=da 3 ==> SPOL=muski 3    conf:(1)
14. PRIHOD='(-inf-10000]' PUSAC=da 3 ==> SPOL=muski 3    conf:(1)
15. SPOL=muski PRIHOD='(-inf-10000]' 3 ==> PUSAC=da 3    conf:(1)
16. SPOL=muski TEZINA='(-inf-58.5]' PUSAC=da 3 ==> STAROST='(-inf-44]' 3    conf:(1)
17. STAROST='(-inf-44]' TEZINA='(-inf-58.5]' PUSAC=da 3 ==> SPOL=muski 3    conf:(1)
18. STAROST='(-inf-44]' SPOL=muski PUSAC=da 3 ==> TEZINA='(-inf-58.5]' 3    conf:(1)
19. TEZINA='(-inf-58.5]' PUSAC=da 3 ==> STAROST='(-inf-44]' SPOL=muski 3    conf:(1)
20. STAROST='(-inf-44]' PUSAC=da 3 ==> SPOL=muski TEZINA='(-inf-58.5]' 3    conf:(1)
21. OBRAZ=nize 2 ==> STAROST='(-inf-44]' 2    conf:(1)
22. STAROST='(44-inf)' 2 ==> SPOL=muski 2    conf:(1)
23. STAROST='(44-inf)' 2 ==> TEZINA='(58.5-inf)' 2    conf:(1)
24. OBRAZ=nize 2 ==> SPOL=muski 2    conf:(1)
25. SPOL=zenski 2 ==> PRIHOD='(-inf-10000]' 2    conf:(1)
26. SPOL=zenski 2 ==> PUSAC=ne 2    conf:(1)
27. OBRAZ=nize 2 ==> TEZINA='(-inf-58.5]' 2    conf:(1)
28. OBRAZ=nize 2 ==> PUSAC=da 2    conf:(1)
29. OBRAZ=visoko 2 ==> TEZINA='(58.5-inf)' 2    conf:(1)
```

Status

OK

Log x 0

Imamo ciljno varijablu

IME	STAR.	SPOL	OBRAZ	ZANIM	TEZINA	PRIHOD	PUSAC
jan	30	muski	nize	radnik	27.3	14000	da
janko	55.5	muski	srednje	radnik	90	20000	ne
zora	?	zenski	visoko	ucitelj	65.2	1000	ne
tanja	18	zenski	srednje	student	55.1	0	ne
tom	70	muski	visoko	?	60	9000	da
tomi	35	muski	srednje	prof	33	16000	ne
stev	42.2	muski	nize	vozac	27	7500	da
marc	29	muski	?	konobar	31	8300	da

Pravilo je i prediktivni model i izbor bitnih atributa

IME	STAR.	SPOL	OBRAZ	ZANIM	TEZINA	PRIHOD	PUSAC
jan	30	muski	nize	radnik	27.3	14000	da
janko	55.5	muski	srednje	radnik	90	20000	ne
zora	?	zenski	visoko	ucitelj	65.2	1000	ne
tanja	18	zenski	srednje	student	55.1	0	ne
tom	70	muski	visoko	?	60	9000	da
tomi	35	muski	srednje	prof	33	16000	ne
stev	42.2	muski	nize	vozac	27	7500	da
marc	29	muski	?	konobar	31	8300	da

PUSAC ako SPOL jednak muski I PRIHOD manji od 15000

Učenje pravila je učenje konceptata

$$\text{Pusac} \text{ AKO } \text{svojstvo1} \wedge \text{svojstvo2} \wedge \dots \text{svojstvoN}$$

} pravilo

$$\text{Pusac}$$

$$\text{svojstvo1}$$

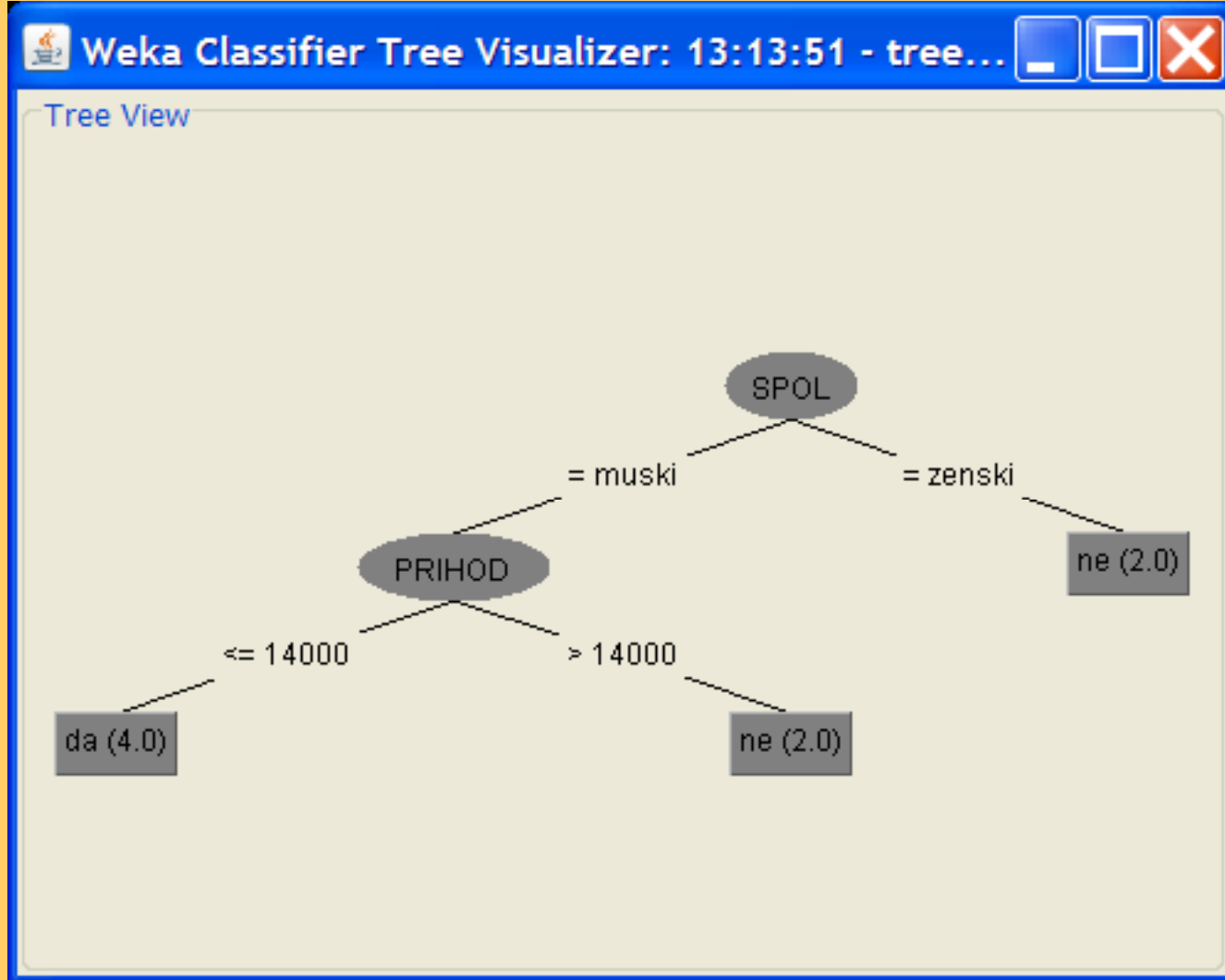
$$\text{svojstvo2}$$

$$\text{svojstvoN}$$

}

Boolove varijable sa vrijednostima 0 ili 1 (istinito/neistinito)

Stablo odlučivanja



PUSAC ako SPOL jednak muski I PRIHOD manji od 15000

Razlike

stabla odlučivanja

- Ciljni atribut ima 2 ili više klasa
- Odlučuje se na osnovi vrijednosti atributa
 - Lako je pretvoriti stablo u skup pravila

pravila

- Ciljni atribut može imati **samo 2 klase**
- Odlučuje se na osnovi istinito/neistino vrijednosti **svojstava**
 - Prikaz pravilima je **koncizniji**

Stablo sa 10
čvorova i 21
listom



Ako $A=3 \wedge B=3$ onda klasa x
Ako $C=3 \wedge D=3$ onda klasa x
Inače klasa y

Postupak učenja pravila

1. Konstruiraj sva potencijalno zanimljiva svojstva primjera (starost <30, obrazovanje=visoko, spol=muski, prihod>30000)
2. Kombiniraj svojstva tako da dobiješ pravilo optimalnih svojstava za odabranu klasu (Ako $sv1 \wedge sv2 \wedge sv3 \rightarrow$ klasa X)
3. Ponavljaaj učenje pravila i konstruiraj listu pravila

Ako $sv1 \wedge sv2 \wedge sv3 \rightarrow$ klasa X
Ako $sv4 \wedge sv5 \wedge sv6 \rightarrow$ klasa X
Ako $sv7 \wedge sv8 \rightarrow$ klasa Y
Inače klasa Z

1) Odaberi jednu klasu

Pretpostavimo da imamo primjere klasa X , Y , i Z .

Ako odaberemo klasu X , onda nam svi primjeri te klase postaju pozitivni primjeri a primjeri klasa Y, Z negativni primjeri. (Imamo problem učenja koncepta klase X).

Kada završimo sa klasom X , onda odaberemo klasu Y . Tada su njeni primjeri pozitivni a primjeri klasa X, Z negativni. Postupak ponavljamo ...

Zadnju klasu ne trebamo učiti već vrijedi "Inače Z ".

Uči koncept odabrane klase

Imamo P pozitivnih primjera i N negativnih primjera. ($P > 0$ i $N > 0$)

- 1) Generiraj svojstva pozitivne klase
- 2) Konstruiraj pravilo koje pokriva čim više pozitivnih a čim manje negativnih primjera (idealno sve pozitivne i niti jedan negativni primjer)
- 3) Dodaj pravilo u listu pravila
- 4) Obriši sve primjere koji zadovoljavaju odabrano pravilo (i pozitivne primjere i negativne primjere*)
- 5) Završi kada nema više pozitivnih primjera

Učenje pravila – dvije petlje

Vanjska - izaberi jednu po jednu klasu
(kada imamo dvije klase Pušač da/ne
onda izaberemo jednu od njih a druga
je pokrivena sa dijelom "Inače ...")

Unutarnja - za svaku klasu generiraj jedno ili
više pravila (ponavljanjem koraka 2-5)

Korak 1 – svojstva pozitivne klase

IME	STAR.	SPOL	OBRAZ	ZANIM	TEZINA	PRIHOD	PUSAC
jan	30	muski	nize	radnik	27.3	14000	da
janko	55.5	muski	srednje	radnik	90	20000	ne
zora	?	zenski	visoko	ucitelj	65.2	1000	ne
tanja	18	zenski	srednje	student	55.1	0	ne
tom	70	muski	visoko	?	60	9000	da
tomi	35	muski	srednje	prof	33	16000	ne
stev	42.2	muski	nize	vozac	27	7500	da
marc	29	muski	?	konobar	31	8300	da

Idemo varijablu po varijablu i za svaku napravimo po jedno ili više svojstava

Korak 1 – svojstva pozitivne klase

IME	STAR.	SPOL	OBRAZ	ZANIM	TEZINA	PRIHOD	PUSAC
jan	30	muski	nize	radnik	27.3	14000	da
janko	55.5	muski	srednje	radnik	90	20000	ne
zora	?	zenski	visoko	ucitelj	65.2	1000	ne
tanja	18	zenski	srednje	student	55.1	0	ne
tom	70	muski	visoko	?	60	9000	da
tomi	35	muski	srednje	prof	33	16000	ne
stev	42.2	muski	nize	vozac	27	7500	da
marc	29	muski	?	konobar	31	8300	da

Kategorička varijabla:

IME=jan, IME=tom, IME=stev, IME=marc

IME#janko, IME#zora, IME#tanja, IME#tomi

Korak 1 – svojstva pozitivne klase

IME	STAR.	SPOL	OBRAZ	ZANIM	TEZINA	PRIHOD	PUSAC
jan	30	muski	nize	radnik	27.3	14000	da
janko	55.5	muski	srednje	radnik	90	20000	ne
zora	?	zenski	visoko	ucitelj	65.2	1000	ne
tanja	18	zenski	srednje	student	55.1	0	ne
tom	70	muski	visoko	?	60	9000	da
tomi	35	muski	srednje	prof	33	16000	ne
stev	42.2	muski	nize	vozac	27	7500	da
marc	29	muski	?	konobar	31	8300	da

18

29

30

35

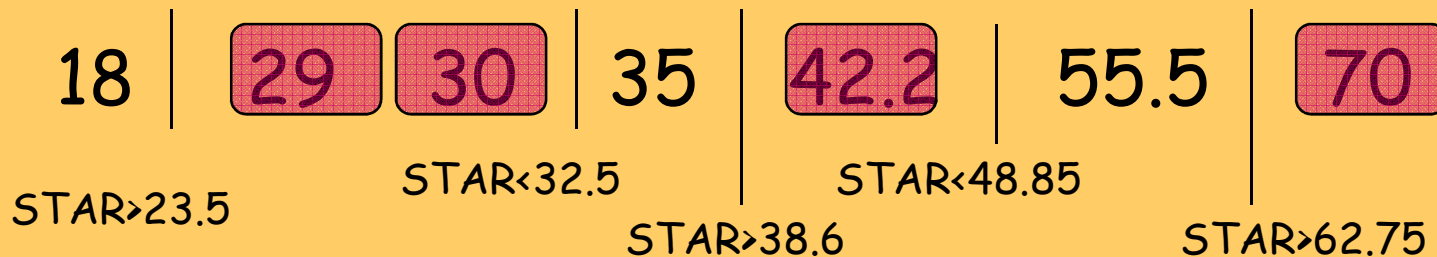
42.2

55.5

70

Korak 1 – svojstva pozitivne klase

IME	STAR.	SPOL	OBRAZ	ZANIM	TEZINA	PRIHOD	PUSAC
jan	30	muski	nize	radnik	27.3	14000	da
janko	55.5	muski	srednje	radnik	90	20000	ne
zora	?	zenski	visoko	ucitelj	65.2	1000	ne
tanja	18	zenski	srednje	student	55.1	0	ne
tom	70	muski	visoko	?	60	9000	da
tomi	35	muski	srednje	prof	33	16000	ne
stev	42.2	muski	nize	vozac	27	7500	da
marc	29	muski	?	konobar	31	8300	da



Korak 1 – svojstva pozitivne klase

IME	STAR.	SPOL	OBRAZ	ZANIM	TEZINA	PRIHOD	PUSAC
jan	30	muski	nize	radnik	27.3	14000	da
janko	55.5	muski	srednje	radnik	90	20000	ne
zora	?	zenski	visoko	ucitelj	65.2	1000	ne
tanja	18	zensk	srednje	student	55.1	0	ne
tom	70	muski	visoko	?	60	9000	da
tomi	35	muski	srednje	prof	33	16000	ne
stev	42.2	musk	nize	vozac	27	7500	da
marc	29	musk	?	konobar	31	8300	da

SPOL=muski, SPOL#zenski, SPOL#muski

Korak 1 – svojstva pozitivne klase

IME	STAR.	SPOL	OBRAZ	ZANIM	TEZINA	PRIHOD	PUSAC
jan	30	muski	nize	radnik	27.3	14000	da
janko	55.5	muski	srednje	radnik	90	20000	ne
zora	?	zenski	visoko	ucitelj	65.2	1000	ne
tanja	18	zenski	srednje	student	55.1	0	ne
tom	70	muski	visoko	?	60	9000	da
tomi	35	muski	srednje	prof	33	16000	ne
stev	42.2	muski	nize	vozac	27	7500	da
marc	29	muski	?	konobar	31	8300	da

nize->1 srednje->2 visoko->3

OBRAZ=1, OBRAZ=3, OBRAZ#2, OBRAZ#3

1 | 2 | 3 OBRAZ<2, OBRAZ>2

!!

Korak 2 - kvaliteta svojstava

IME	STAR.	SPOL	OBRAZ	ZANIM	TEZINA	PRIHOD	PUSAC	SPOL=muski
jan	30	muski	nize	radnik	27.3	14000	da	*
janko	55.5	muski	srednje	radnik	90	20000	ne	-
zora	?	zenski	visoko	ucitelj	65.2	1000	ne	*
tanja	18	zenski	srednje	student	55.1	0	ne	*
tom	70	muski	visoko	?	60	9000	da	*
tomi	35	muski	srednje	prof	33	16000	ne	-
stev	42.2	muski	nize	vozac	27	7500	da	*
marc	29	muski	?	konobar	31	8300	da	*

- * korisno za predikciju pozitivnih (TP -true positive)
- * korisno za odstranjivanje negativnih (TN -true neg.)

SPOL=muski ima kvalitetu 4 tp i 2 tn (nije idealno ali nije loše)

Matrica konfuzije

Matrica konfuzije

predvidjeno

da ne

<u>stva</u>	da	TP	FN
<u>rno</u>	ne	FP	TN

TP - true positive

TN - true negative

FP - false positive

FN - false negative

Idealno je kada su i FP i FN jednaki 0

Pravilo ili svojstvo su beskorisni ako je $TP=0$ ili $TN=0$

10 0
0 10

0 10
0 10

Da li je bolje imati više TP ili TN određeno je evaluacijskom mjerom !

$\begin{matrix} 7 & 3 \\ 0 & 10 \end{matrix}$ ili $\begin{matrix} 9 & 1 \\ 1 & 9 \end{matrix}$

Mjere kvalitete pravila (svojstava)

točnost = $(|TP| + |TN|) / |E|$ (E ukupan broj primjera)

senzitivnost = $|TP| / |P|$

specifičnost = $|TN| / |N|$

podrška = $|TP| / |E|$

pouzdanost = $|TP| / (|TP| + |FP|)$ = preciznost

učestalost greške = $(|FP| + |FN|) / |E|$

lift = $(|TP| / |P|) / ((|TP| + |FP|) / |E|)$

Korak 2 - postupak odabira svojstva

Nakon što smo odredili sva potencijalno zanimljiva svojstva (SPOL=muski, STAR<48.85) za svakog od njih odredimo broj TP i TN, izračunamo $(TP+TN)/E$, te odaberemo ono sa najvećim omjerom !

U našem primjeru je najbolje svojstvo SPOL=muski

I sa njim započinjemo pravilo:

Pusac AKO SPOL=muski ^

Primjer – svojstva pozitivne klase

IME	STAR.	SPOL	OBRAZ	ZANIM	TEZINA	PRIHOD	PUSAC	STAR<48.85
jan	30	muski	nize	radnik	27.3	14000	da	TP
janko	55.5	muski	srednje	radnik	90	20000	ne	TN
zora	?	zenski	visoko	ucitelj	65.2	1000	ne	-
tanja	18	zenski	srednje	student	55.1	0	ne	fp
tom	70	muski	visoko	?	60	9000	da	fn
tomi	35	muski	srednje	prof	33	16000	ne	fp
stev	42.2	muski	nize	vozac	27	7500	da	TP
marc	29	muski	?	konobar	31	8300	da	TP

STAR<48.85 ima kvalitetu 3 tp i 1 tn (nije idealno i lošije je od SPOL=muski)

Korak 2 – privremeno eliminiraj primjere koji ne zadovoljavaju odabrano svojstvo

Pusac AKO SPOL=muski ^

IME	STAR.	SPOL	OBRAZ	ZANIM	TEZINA	PRIHOD	PUSAC
jan	30	muski	nize	radnik	27.3	14000	da
janko	55.5	muski	srednje	radnik	90	20000	ne
zora	?	zenski	visoko	ucitelj	65.2	1000	ne
tanja	18	zenski	srednje	student	55.1	0	ne
tom	70	muski	visoko	?	60	9000	da
tomi	35	muski	srednje	prof	33	16000	ne
stev	42.2	muski	nize	vozac	27	7500	da
marc	29	muski	?	konobar	31	8300	da

briši

Ponovi postupak na reduciranom skupu primjera

IME	STAR.	SPOL	OBRAZ	ZANIM	TEZINA	PRIHOD	PUSAC	PRIHOD <15000
jan	30	muski	nize	radnik	27.3	14000	da	TP
janko	55.5	muski	srednje	radnik	90	20000	ne	TN
tom	70	muski	visoko	?	60	9000	da	TP
tomi	35	muski	srednje	prof	33	16000	ne	TN
stev	42.2	muski	nize	vozac	27	7500	da	TP
marc	29	muski	?	konobar	31	8300	da	TP

Svojstvo $PRIHOD < 15000$ je idealno sa 4 tp i 2 tn

Pusac AKO $SPOL = muski \wedge PRIHOD < 15000$

Kraj prvog pravila

S ovim je završilo prvo pravilo.

Brišemo sve* primjere koji zadovoljavaju ovo pravilo.

Ako je ostalo još i pozitivnih i negativnih primjera, počinjemo konstruirati novo pravilo na osnovi primjera koji su ostali.

Postupak ponavljamo dok god još ima primjera i možemo naći zadovoljavajuće pravilo.

Učenje pravila – dvije petlje

Vanjska - izaberi jednu po jednu klasu
(kada imamo dvije klase Pušač da/ne
onda izaberemo jednu od njih a druga
je pokrivena sa dijelom "Inače ...")

Unutarnja - za svaku klasu generiraj jedno ili više pravila
(ponavljanjem koraka 2-5)



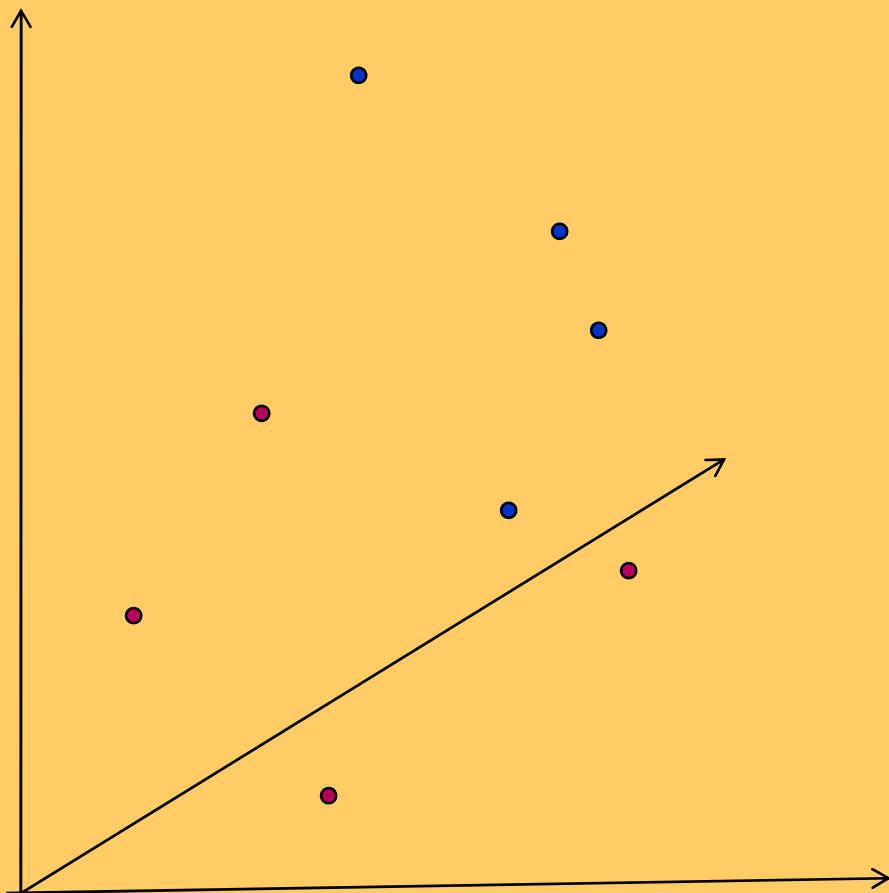
Ovo je učenje pravila ALI da bi to bilo strojno
učenje potrebno je još NEŠTO !!

Cilj strojnog učenja

IME	STAR.	SPOL	OBRAZ	ZANIM	TEZINA	PRIHOD	PUSAC
jan	30	muski	nize	radnik	27.3	14000	da
janko	55.5	muski	srednje	radnik	90	20000	ne
zora	?	zenski	visoko	ucitelj	65.2	1000	ne
tanja	18	zenski	srednje	student	55.1	0	ne
tom	70	muski	visoko	?	60	9000	da
tomi	35	muski	srednje	prof	33	16000	ne
stev	42.2	muski	nize	vozac	27	7500	da
marc	29	muski	?	konobar	31	8300	da

8 primjera sa 7 nezavisnih varijabli predstavljaju 8 točaka u sedmo-dimenzionalnom prostoru za koje znamo klasifikaciju. Cilj je napraviti model (pravilo) koje će omogućiti predvidjeti klasifikaciju ogromnog broja točaka u tom prostoru.

Cilj strojnog učenja



Mi učenjem pravila znamo napraviti hipotezu (teoriju) koja je točna za svih 8 točaka. Ali postoji vrlo velik broj raznih teorija koje su sve točne za tih 8 točaka a različite s obzirom na ostale točke.

Osnovno pitanje strojnog učenja:

Koja od svih mogućih teorija će imati najtočniju predikciju na ostalim točkama ?

Cilj strojnog učenja

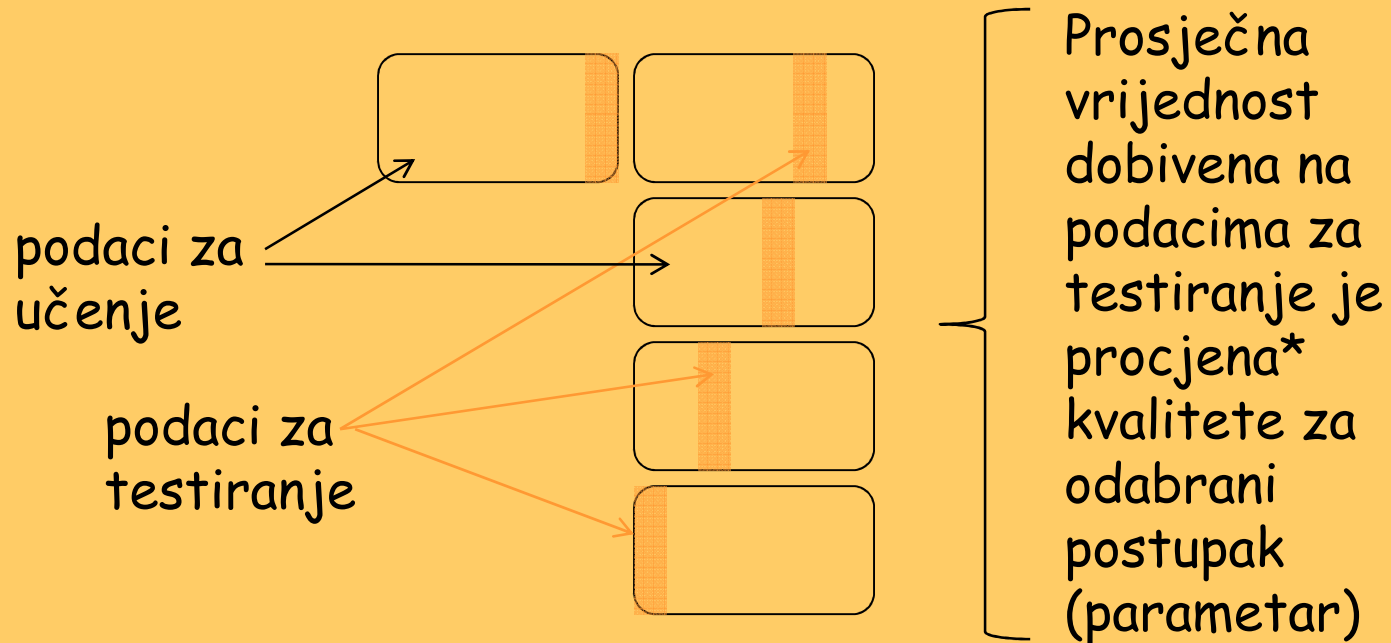
Znači: mi ne želimo bilo koje pravilo već pravilo koje će imati idealnu predikciju na neviđenim točkama

Problem nema egzaktno rješenje jer mi za neku određenu domenu ne znamo vrijednosti klasifikacije u nepoznatim točkama pa i ne možemo usporediti točnost raznih pravila.

Rješenje: 1) iskustvo sa sličnim domenama za koje smo testirali postupak. 2) procjena krosvalidacijom

Krosvalidacija

Uvijek kada postoji mogućnost izbora vrijednosti nekog parametra u postupku učenja pravila preporuča se provjeriti koja je optimalna vrijednost za dani problem



* Procjena je dobra ako je budući skup za testiranje statistički ekvivalentan trenutno raspoloživom skupu na kojem se izvodi krosvalidacija

Popravke postupka generiranja pravila - ograničenja

U postupak uvodimo razna ograničenja i nadopune s ciljem da generirano pravilo ima (ako ne najbolju onda barem) dobru prediktivnu točnost:

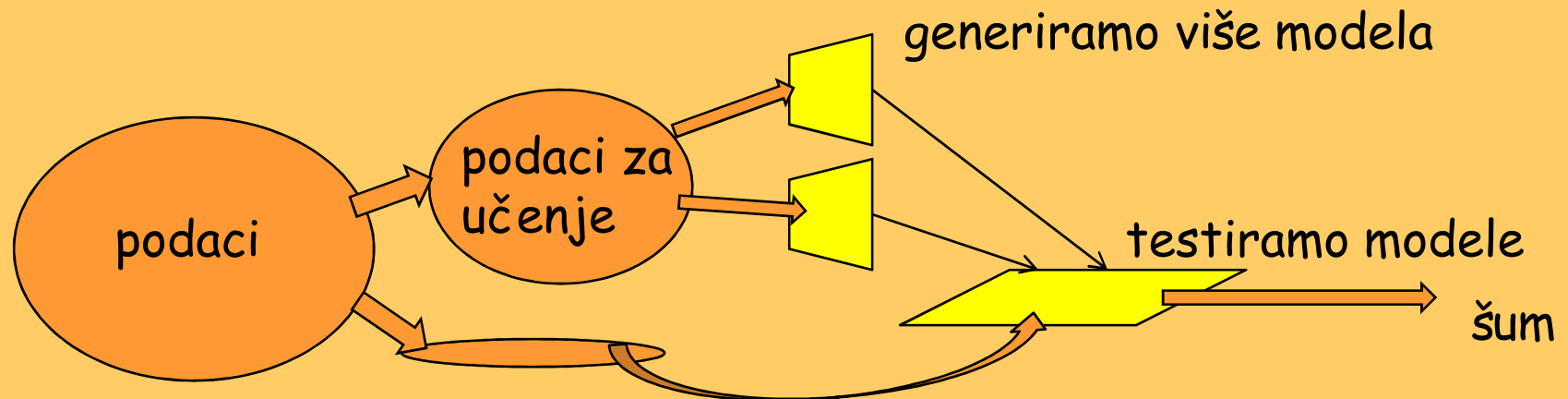
- Ograničimo izbor samo na svojstva koja imaju relativno dobre tp i tn vrijednosti (*feature relevancy*).
- Ograničimo da se ne može generirati pravilo koje pokriva jako mali broj pozitivnih primjera (*pre-pruning*).
- Ograničimo najveću dopuštenu dužinu pravila na 4-6 svojstava (*complexity restriction*).

Popravke postupka generiranja pravila - nadopune

- Uvodimo različite mjere kvalitete svojstava i njihovih kombinacija (*complexity measures*).
- Uvodimo postupak izbacivanja varijabli koje su nerelevantne (*attribut selection*).
- Uvodimo konstruiranje složenih svojstava (*feature construction*)
- Uvodimo postupak da u svakoj iteraciji odabiremo ne jedno već N ($N=10-500$) najboljih svojstava te onda za njih sve tražimo najbolju slijedeću iteraciju (*beam search*).
- Uvodimo postupak eliminacije primjera koji potencijalno predstavljaju šum (greške ili izuzetke) (*noise detection and elimination*).

Otkrivanje šuma - filtriranjem

Za raspoložive podskupove raspoloživih podataka napravimo raznim postupcima strojnog učenja više modela koje testiramo na primjerima koji nisu korišteni za učenje. Ako većina modela pogrešno klasificira određeni primjer, možemo pretpostaviti da je problem u tom primjeru i da on predstavlja šum



Otkrivanje šuma – saturacijom skupa podataka

- Za originalni skup primjera izmjerimo kompleksnost skupa pravila koji zadovoljavaju sve primjere.
- Zatim jedan po jedan eliminiramo primjere i za svaki tako dobiveni skup izmjerimo kompleksnost skupa pravila
- Ako se otkrije da se kompleksnost bitno smanjila u nekom slučaju, eliminirani primjer je potencijalni šum. On se eliminira i postupak se ponavlja.

Kompleksnost skupa pravila možemo odrediti i bez da generiramo sva pravila tako da odredimo minimalni broj svojstava potrebnih da bi se generirala hipoteza točna za sve primjere.

Učenje pravila cjelovit postupak

Vanjska petlja - izaberi jednu po jednu klasu

Unutarnja petlja - za svaku klasu generiraj jedno ili više pravila (ponavljanjem slijedećih koraka)

- 1) Generiraj sva svojstva a zatim odbaci nerelevantna svojstva
- 2) Odredi minimalni skup svojstava potrebnih za generiranje apsolutno točne hipoteze
- 3) Iterativno odbaci primjere koji omogućuju da se minimalni skup bitno smanji
- 4) Odaberi po kriteriju kvalitete mali skup (*beam*) najboljih svojstava
- 5) Traži optimalne konjunkcije svojstava iz skupa sa svim svojstvima
- 6) Najbolje kombinacije uključi u skup
- 7) Iterativno ponavljaj dok raste kvaliteta, dužina pravila nije prevelika i pravilo zadovoljava dovoljno velik broj pozitivnih primjera
- 8) Na kraju izdvoji konjunkciju svojstava (pravilo) najveće kvalitete
- 9) Obriši sve primjere koji zadovoljavaju odabrano pravilo te ponovi proceduru od točke 4) dok god ima dovoljno pozitivnih i negativnih primjera te je moguće generirati kvalitetna pravila.

Zaključak

- Za dani skup primjera nije problem naći skup pravila koji zadovoljavaju sve primjere već je problem naći skup pravila optimalne prediktivne kvalitete.
- Ponekad pravila koja i nisu točna za sve poznate primjere imaju bolju prediktivnu točnost na budućim primjerima (šum u podacima) !
- Ne postoji jedan idealni postupak učenja pravila koji garantira optimalnu prediktivnu točnost. To je predmet istraživanja i stalnog usavršavanja (izbor načina mjerenja kvalitete svojstava, otkrivanje šuma, preselekcija varijabli, složenost rješenja) a danas praktično postoji i primjenjuje se više raznih postupaka.

Poznati sustavi za učenje pravila

- CN2 uključuje dva razna postupka, jedan zasnovan na entropiji direktno može rješavati problem više klasa
- AQ3 klasični postupak pokrivanja primjera
- RIPPER najmoderniji, uključuje popravljivanje skupa pravila na osnovi procjene sveobuhvatne informatičke kompleksnosti modela s ciljem njenog smanjivanja



INSTITUT RUDJER BOŠKOVIĆ



in English

Dobro došli na DMS - poslužitelj za analizu podataka

DMS (skraćeno od engleskog Data Mining Server) je mrežna usluga namijenjena analizi podataka na osnovi indukcije znanja. Usluga je zamišljena tako da korisnici šalju podatke na naš poslužitelj, kompletna obrada se izvršava na poslužitelju a rezultat se prikazuje na računalu korisnika. Preglednik (browser) je sredstvo komunikacije korisnika i poslužitelja. Vaši podaci biti će analizirani sa ILLM (skraćeno od engleskog Inductive Learning by Logic Minimization) sustavom koji je osmišljen i realiziran u Laboratoriju za informacijske sustave Zavoda za elektroniku Instituta R. Bošković.

Poslije prikaza rezultata, podaci se automatski brišu s našeg poslužitelja. Potpuno poštujemo privatnost i sigurnost podataka primljenih od korisnika, ali vas molimo da pažljivo pročitate **informaciju o sigurnosti** kao i **pravne uvjete korištenja**.

Poslužitelj sadrži obrazovni materijal, pregledne tekstove i pokazivače na druge slične adrese. Struktura poslužitelja je prikazana u slijedećoj tablici koja uključuje i najvažnije ulazne točke za korisnike.

Novi korisnici	Ovaj dio sadrži uvodna objašnjenja i opis pripreme podataka.
Iskusni korisnici	Ako već imate pripremljene podatke u traženom obliku, možete odmah započeti analizu.
Pregled područja	Ako vas zanima inteligentna obrada podataka i otkrivanje znanja, ovaj dio nudi kratki uvod u područje.
Primjedbe	Ovdje možete dati primjedbe ili kontaktirati autore u vezi nekog problema.
DMS projekt	Informacije o razvoju ovog poslužitelja, zahvale, autori i ILLM reference mogu se naći u ovom dijelu.

zahvale i citiranje
autori

dms.irb.hr