Prva domača naloga

Anže Pečar (63060257)

26. februar 2012

1 Uvod

Cilj prve domače naloge je bil seznaniti se s podatki, ki jih uporablja tekmovanje *JRS 2012 Data Mining Competition: Topical Classification of Biomedical Research Papers*. Podatke smo preučili tako, da smo prešteli primere in atribute, preverili kako redka je matrika in porazdelitve atributov prikazali na različne načine.

2 Rezultati

Koliko primerov in atributov vsebujejo podatki?

Podatki vsebujejo 2000 primerov in 10000 atributov.

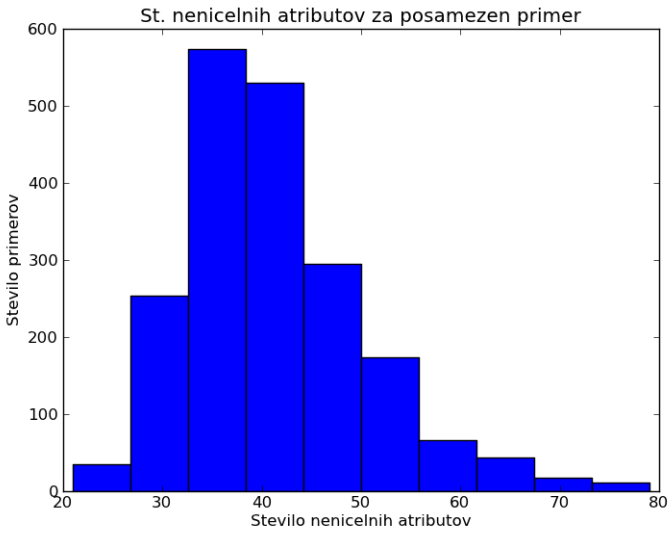
Kakšnega tipa so atributi?

Atributi so zvezni, saj zasedajo vrednosti na intervalu pozitivnih celih števil.

Kako redka je matrika oz. kakšen delež njenih elementov ima vrednost različen od 0?

Matrika je precej redka, saj ima le 0.41% elementov vrednost različno od nič.

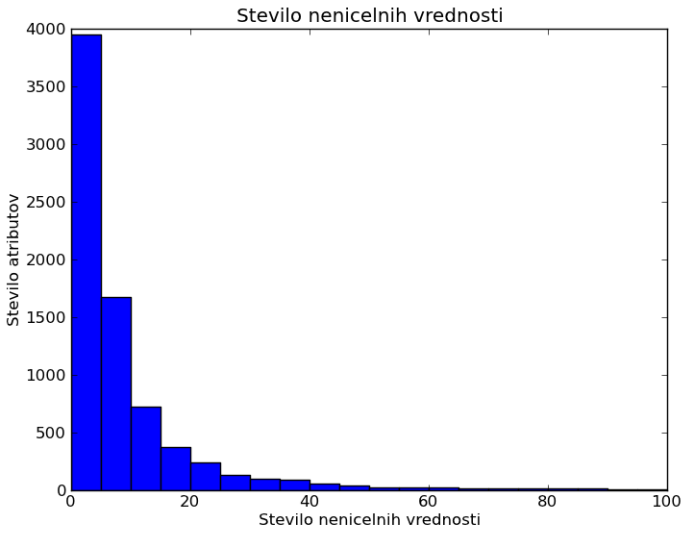
Koliko atributov ima vrednost različno od 0 za posamezen primer?



Slika 1: Prikaz atributov, ki imajo vrednost različno od 0 za posamezen primer

Iz Slike 1 je lepo razvidno, da ima največ primerov okoli 40 neničelnih atributov. Redki so primeri, ki imajo več kot 70 oz. manj kot 30 atributov.

V koliko primerih atribut zavzame neničelne vrednosti?



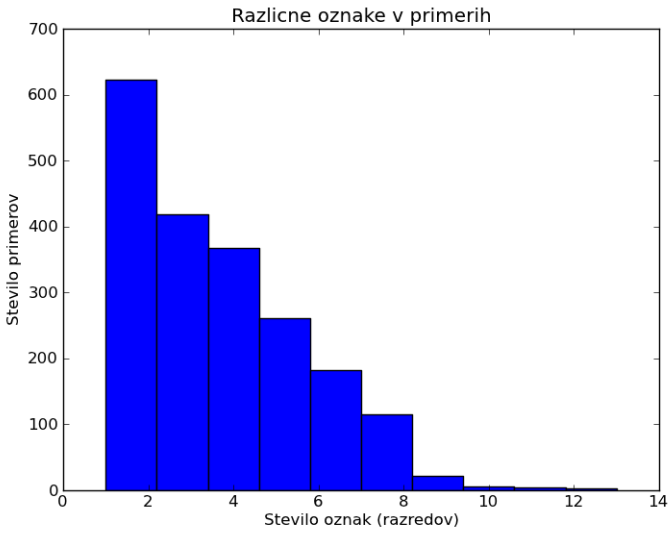
Slika 2: Primeri, kjer atribut zavzame neničelne vrednosti

Iz Slike 2 je razvidno, da so zelo redki atributi, ki se pojavijo v več kot 20 primerih. Graf sem omejil na intervalu od 0 do 100, saj ga ostale ničelne vrednosti naredijo manj preglednega.

Koliko je vseh različnih oznak (razredov) v podatkih?

Vseh različnih oznak v podatkih je 82.

S koliko različnimi oznakami so označeni primeri?



Slika 3: Različne oznake za primere

Na Sliki 3 lahko vidimo, da ima največ primerov po en razred oz. oznako. Primeri z več kot desetimi oznakami so redki.

Lastno vprašanje 1: Koliko atributov se ne pojavi v nobenem primeru?

Vprašanje je zanimivo, saj atribute, ki se ne pojavijo v nobenem primeru, lahko odstranimo in poenostavimo nabor podatkov. V našem primeru imamo imamo kar 2349 atributov, ki imajo pri vseh primerih vrednost 0.

Lastno vprašanje 2: Kateri atribut se največkrat pojavi v primerih? V kolikih primerih se pojavi?

V primerih se največkrat pojavi atribut 7877, ki se pojavi 537 krat.

Lastno vprašanje 3: Naštej 3 najbolj pogoste oznake? Kolikokrat se pojavijo?

Najpogostejše oznake so:

- ’40’ s 502,
- ’44’ s 494 in
- ’18’ s 428 ponovitvami.

3 Izjava o izdelavi domače naloge

Domačo nalogo in pripadajoče programe sem izdelal sam.