Matematički fakultet
Univerzitet u Beogradu

Podaci o muzejima

— seminarski rad —

Studenti	Bukurov Anja	1082/2016
	Stanković Vojislav	1080/2016
Predmet	Istraživanje podataka	
Školska godina	2016/2017	
Nastavnik	dr Mladen Nikolić	
Datum	25. juni 2017	

1 Zadatak

U ovom radu obrađeni su podaci o muzejima celog sveta (tripadvisor_museum_world.csv) koji se mogu preuzeti sa Muzeji. Podaci su obrađeni pomoću alata *KNIME*.

2 Opis podataka

Skup podataka o muzejima iz datoteke *tripadvisor_museum_world.csv* sadrži sledeće atribute:

- **Address** - adresa muzeja
- **Description** - opis muzeja
- **FeatureCount** - broj turističkih vodiča u kojima je muzej preporučen
- **Fee** - da li je ulaz slobodan ili se naplaćuje
- **Langtitude** - geografska dužina
- **Latitude** - geografska širina
- **LengthOfVisit** - preporučena dužina posete
- **MuseumName** - ime muzeja
- **PhoneNum** - broj telefona
- **Rank** - mesto na listi svih znamenitosti koje se mogu posetiti u gradu u kom se nalazi muzej
- **Rating** - prosečna ocena muzeja (0-5)
- **ReviewCount** - koliko ljudi je ocenilo muzej
- **TotalThingsToDo** - ukupan broj znamenitosti koje se mogu posetiti u gradu u kom se nalazi muzej

3 Prečišćavanje podataka

Nakon učitavanja podataka pomoću čvora CSV Reader, upotrebljeni sledeći čvorovi:

1. **Column Filter** - atributi *Address*, *PhoneNum*, *Langtitude* i *Latitude* su odbačeni jer ne nose korisne informacije. Atribut *Description* je odbačen jer je za njegovu analizu potreban program za obradu prirodnih jezika.

2. Pošto atributi *TotalThingsToDo* i *ReviewCount* sadrže celobrojne vrednosti ali su predstavljeni u obliku stringova izvršena je njihova konverzija u cele brojeve:

   - **String Replacer** je korišćen da se iz vrednosti atributa *TotalThingsToDo* i *ReviewCount* izbace zarezi.
   - **String To Number** je zatim korišćen da se vrednosti atributa *TotalThingsToDo* i *ReviewCount* prebace u brojeve.

3. **Sorter** - podaci su sortirani po nazivu muzeja

4. **GroupBy** - podaci su grupisani po nazivu muzeja kako bi se izbacila ponavljanja

4 Da li će muzej biti uključen u turistički vodič?

Pomoću linearne regresije pokušali smo da predvidimo u koliko će se turističkih vodiča naći muzej tj. predviđamo vrednosti atributa *FeatureCount*. Kombinacija atributa koja je dala najbolje rezultate je: *Fee*, *LenghtOfVisit*, *Rating* i *ReviewCount*. Iskoristili smo čvor **Numeric Scorer** da odredimo koliko je model dobar. Rezultati su sledeći:

- $R^2 = 0{,}454$
- *mean absolute error* = 1,558
- *mean squared error* = 4,664
- *root mean squared deviation* = 2,16
- *mean signed difference* = -0,218

Na osnovu rezultata zaključili smo da između ovih atributa ne postoji prava zavisnost.

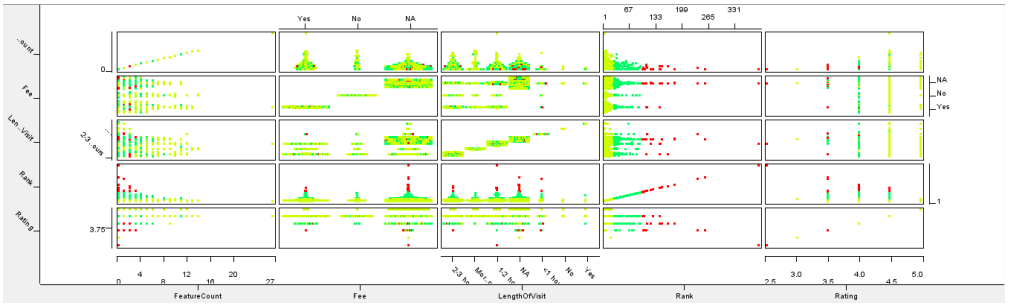
5 Da li postoji veza između Rating i ostalih atributa?

Koristeći pravila pridruživanja pokušali smo da pronađemo kombinacije atributa koje daju visoke vrednosti atributa Rating. Međutim, nismo dobili nijedno pravilo sa min_sup većim od 0,6 i min_conf većim od 0,8.

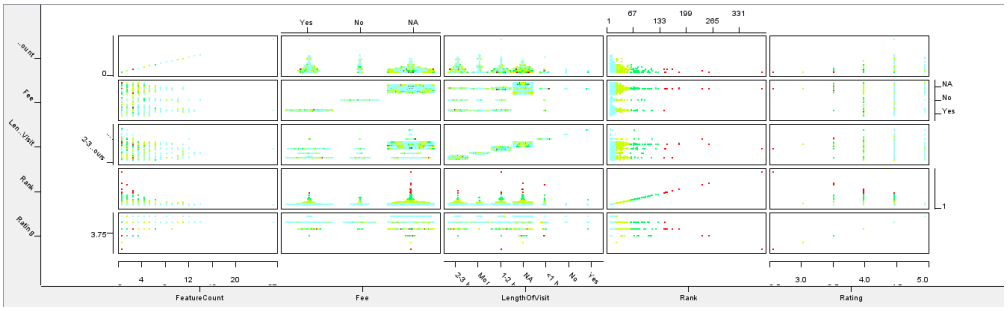
6 Da li podaci mogu da se grupišu na osnovu nekih atributa?

Pokušali smo da grupišemo podatke u klastere i da na osnovu tih grupa zaključimo šta utiče na visoke vrednosti atributa *Rating*. Na slikama 1, 2 i 3 prikazane su matrice sličnosti za 3, 4 i 5 klastera redom.

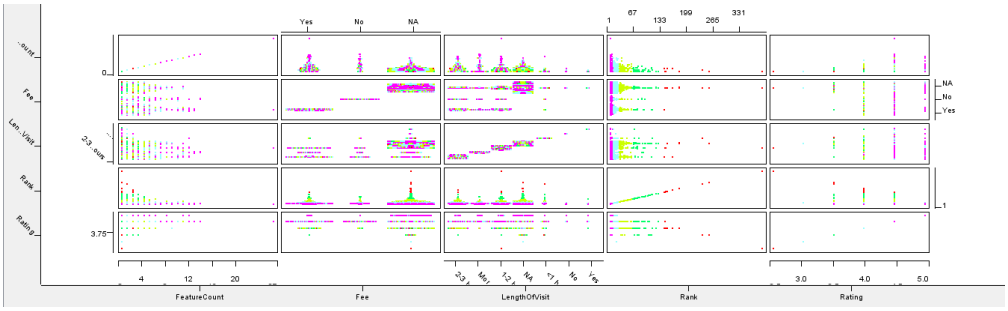
Nakon klasterovanja rezultate smo sortirali po klasterima a onda izdvojili redove sa najvišim vrednostima za Rating (5,0). Preostali redovi uglavnom pripadaju istom klasteru - tabela je prikazana na slici 4. Vidimo da su vrednosti ostalih atributa raznolike pa zaključujemo da se podaci ne mogu lepo grupisati tj. da ne postoji prava zavisnost između atributa.



Slika 1: Podaci grupisani u 3 klastera.



Slika 2: Podaci grupisani u 4 klastera.



Slika 3: Podaci grupisani u 5 klastera.

Row ID	Featur...	Fee	LengthOfVisit	Rank	Rating	Review...	Cluster
Row217	0	NA	NA	1	5	100	cluster_5
Row218	0	NA	NA	1	5	147	cluster_5
Row219	0	NA	NA	1	5	367	cluster_5
Row220	0	NA	NA	1	5	543	cluster_5
Row221	0	NA	NA	1	5	604	cluster_5
Row222	0	NA	NA	1	5	681	cluster_5
Row223	0	NA	NA	1	5	735	cluster_5
Row224	0	NA	NA	1	5	744	cluster_5
Row225	0	NA	NA	1	5	801	cluster_5
Row226	0	NA	NA	1	5	983	cluster_5
Row227	0	NA	NA	1	5	987	cluster_5
Row228	0	NA	NA	1	5	993	cluster_5
Row229	0	NA	NA	1	5	1441	cluster_5
Row230	0	NA	NA	1	5	1479	cluster_5
Row231	0	NA	NA	1	5	1769	cluster_5
Row232	0	NA	NA	1	5	1773	cluster_5
Row233	0	NA	NA	1	5	2076	cluster_5
Row234	0	NA	NA	1	5	2848	cluster_5
Row235	0	NA	NA	1	5	2884	cluster_5
Row236	0	NA	NA	1	5	2905	cluster_5
Row237	0	NA	NA	1	5	2942	cluster_5
Row238	0	NA	NA	1	5	3057	cluster_5
Row239	0	NA	NA	1	5	4589	cluster_5
Row240	0	NA	NA	1	5	6464	cluster_5
Row283	0	NA	NA	2	5	718	cluster_5
Row284	0	NA	NA	2	5	1984	cluster_5
Row301	0	NA	NA	3	5	880	cluster_5
Row302	0	NA	NA	3	5	1046	cluster_5
Row335	0	NA	NA	5	5	442	cluster_5
Row451	0	NA	No	2	5	433	cluster_5
Row453	0	NA	Yes	1	5	893	cluster_5
Row463	0	Yes	1-2 hours	1	5	626	cluster_5
Row470	0	Yes	2-3 hours	1	5	1406	cluster_5
Row488	1	NA	1-2 hours	2	5	1181	cluster_5
Row611	2	NA	More than 3 hours	1	5	3553	cluster_5
Row616	2	NA	NA	1	5	14080	cluster_5
Row625	2	NA	NA	5	5	1835	cluster_5
Row661	2	No	2-3 hours	6	5	542	cluster_5
Row752	3	Yes	2-3 hours	4	5	2769	cluster_5
Row771	4	NA	NA	2	5	5183	cluster_5

Slika 4: Deo tabele sa klasterovanim podacima.

7 Eksperimentalna klasifikacija na osnovu atributa Rating

Konvertovali smo vrednosti atributa Rating u string i pokušali klasifikaciju. Metod potpornih vektora nije uspevao ni za jedan kernel - konkretno nije pronađen potporni vektor za klasu 4,0.

Pomoću K najbližih suseda postigli smo bolje rezultate. Koristili smo **Interval Loop** kako bismo odredili najbolju vrednost za k. Rezultati su prikazani na slici 5. Vidimo da je najveća preciznost postignuta za k=8 i na osnovu toga smo klaisifkovali podatke sa preciznošću 73%. Matrica konfuzije prikazana je na slici 6. Vidimo da dobro predviđa klase 4,5 i 5,0 dok za klasu 4,0 mnogo greši.

Row ID	Accuracy	k	Iteration
Overall#0	0.69	3	0
Overall#1	0.697	4	1
Overall#2	0.708	5	2
Overall#3	0.73	6	3
Overall#4	0.723	7	4
Overall#5	0.734	8	5
Overall#6	0.719	9	6
Overall#7	0.73	10	7

Slika 5: Vrednost parametra k.

Row ID	4.0	4.5	5.0	3.5	2.5	3.0
4.0	3	45	0	0	0	0
4.5	5	197	0	0	0	0
5.0	1	17	1	0	0	0
3.5	0	5	0	0	0	0
2.5	0	0	0	0	0	0
3.0	0	0	0	0	0	0

Slika 6: Matrica konfuzije za K najbližih suseda.

8 Zaključak

Na osnovu dobijenih rezultata zaključili smo da među podacima nema zavisnosti i da nisu pogodni za klasifikaciju (jer nemamo prave klase po kojima bismo je vršili) niti bilo koju drugu metodu predviđanja.