Klasteriranje zemalja

Matej Petrinović
Multivarijatna analiza

I. Uvod

Pretpostavimo da imamo sljedeću problematiku: Dani su nam podaci o državama svijeta, te je naš zadatak u tom skupu podataka pronaći zemlje koje su si na neki način najsličnije, tj. grupirati ih u određeni broj grupa odnosno klustera. Našu problematiku riješit ćemo s podacima poverty.txt koja sadrži podatke o 97 zemljama svijeta, a podaci su prikupljeni 1991. godine. Sljedeća tablica prikazuje nekoliko redova podataka.

Name	Natality	Mortality	Infant mortality rate	Male Exp. Life	Female Exp. Life	GDP
Hungary	11,6	13,4	14,8	65,4	73,8	2780
U.K.	13,6	11,5	8,4	72,2	77,9	16100
Austria	14,9	7,4	8	73,3	79,6	17000
Japan	9,9	6,7	4,5	75,9	81,8	25430
Canada	14,5	7,3	7,2	73	79,8	20470
U.S.A.	16,7	8,1	9,1	71,5	78,3	21790

Podaci koje tablica sadrži su sljedeći:

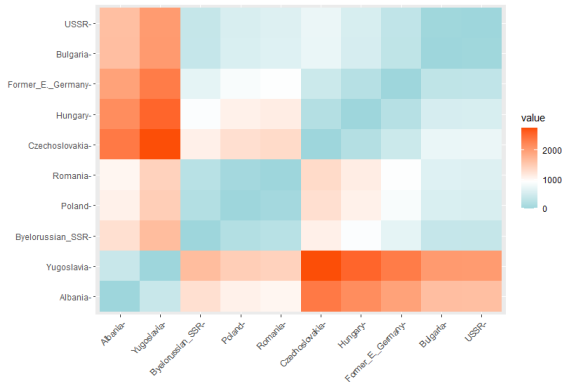
- Ime zemlje (Name)
- Stopa nataliteta (Natality)
- Stopa mortalite (Mortality)
- Stop smrtnosti dijece do god. dana (Infant mortality rate)
- Očekivani životni vijek za muškarce i žene ((Fe)Male Exp. Life)
- Bruto domaći proizvod (GDP)

II. Mjerenje udaljenosti

Razvrstavanje promatranja u skupine zahtijeva neke metode za izračunavanje udaljenosti ili sličnosti između svakog para promatranja. Rezultat ovog izračuna poznat je kao matrica udaljenosti. Za naše promatranje koristit ćemo Euklidsku udaljenost, koja je za par (x, y) definira kao:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Za naše podatke ($N = 10$) pogledajmo matricu udaljenosti:



III. Priprema podataka

Za obavljanje klaster analize, općenito, podatke treba pripremiti kako slijedi:

- Redovi su promatranja (pojedinci), a stupci varijable
- Sve nedostajuće vrijednosti u podacima moraju se ukloniti ili procijeniti
- Podaci moraju biti standardizirani (tj. skalirani) kako bi varijable bile usporedive

IV. K-Means Klasteriranje

K-Means najčešće je korišteni algoritam strojnog učenja bez nadzora za particioniranje datog skupa podataka u skup k grupa (tj. K klastera), gdje k predstavlja broj skupina koje je analitičar unaprijed odredio. Klasificira objekte u više grupa, tako da su objekti unutar istog klastera što sličniji, dok su objekti iz različitih klastera što različitiji (tj. klasna sličnost). U k-Means klasteriranju, svaka grupa predstavljena je svojim središtem (tj. centroidom) što odgovara srednjoj vrijednosti točaka dodijeljenih grupiranju.

Standardni algoritam definira ukupnu varijaciju unutar klastera kao zbroj kvadratnih udaljenosti euklidskih udaljenosti između određenog podatka i odgovarajućeg centroida:

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2 \qquad (1)$$

gdje su x_i podaci pridruženi klasteru C_k i μ_k centroid klaster C_k. Cilj problema je minimizirati ukupne varijaciju unutar klastera, tj. riješiti optimizacijski problem:

$$\sum_{j=1}^{k} W(C_j) = \sum_{j=1}^{k} \sum_{x_i \in C_j} (x_i - \mu_k)^2 \to \min \qquad (2)$$

Ukupni zbroj kvadrata unutar klastera mjeri kompaktnost klastera i želimo da on bude što manji.

V. Određivanje optimalnih klastera

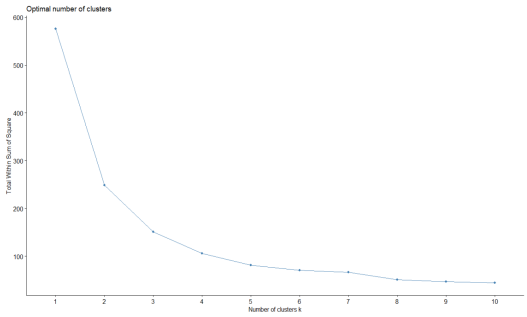
Obzirom da prije same uporabe algoritma, potrebno je specificirati koliko klastera želimo, tj. koliki je k? Postoje 3 popularne metode određivanja optimalnog broja k, a to su:

- Elbow Method
- Silhouette method
- Gap statistic

Bez ulaženja u teorijske opise samih metoda, prikazat ćemo to na primjeru naših podataka.

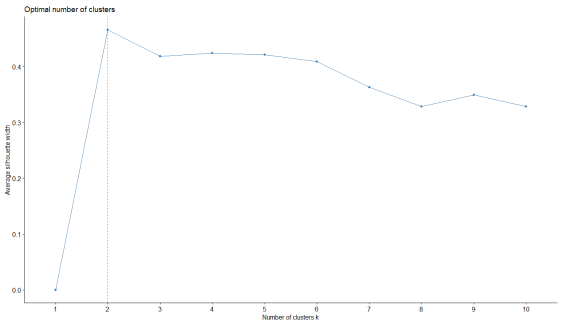
Elbow Metoda

Najjednostavnija metoda koja traži optimalan broj klastera kao riješene optimizacijskog problema (2). Iduća slika pokazuje izgled na korištenim podacima:



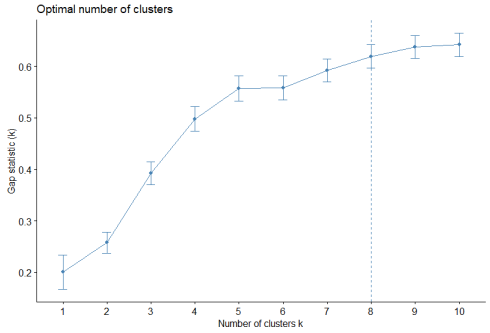
Silhouette Metoda

Metoda dređuje koliko dobro svaki objekt leži u svojoj grupi. Visoka prosječna Silhouette vrijednost ukazuje na dobro grupiranje. Metoda prosječne Silhouette vrijednosti izračunava prosječnu vrijednost promatranja za različite vrijednosti k. Optimalan broj k je onaj koji maksimizira prosječnu Silhouette vrijednost. Optimalan broj za naše podatke:



Gap Statistic Metoda

Metoda uspoređuje ukupnu varijaciju unutar klastera za različite vrijednosti k s njihovim očekivanim vrijednostima pod nultom referentnom raspodjelom podataka. Referentni skup podataka generira se pomoću Monte Carlo simulacija postupka uzorkovanja.



VI. Rezultati

Elbow i Gap metoda sugeriraju da je optimalan broj klaster $k = 8$, dok Silhouette metoda prikazuje da je optimalan $k = 2$. Rezultate ćemo grafički prikazati koristeći PCA za 2D sliku.