UDC 519.237.8

OSNOVNE PREMISE ANALIZE GRUPISANJA

THE BASIC PREMISES OF GROUPING ANALYSIS

Nataša Papić-Blagojević[*]
Denis Bugar[**]

Sažetak: *U procesu statističke analize, radi rešavanja poslovnih i istraživačkih problema i obezbeđenja valjanih informacija za ceo analitički proces, veoma je važno jasno definisati sve njegove delove. Jedan deo tog analitičkog procesa čini i analiza grupisanja, koja primenom hijerarhijskih i nehijerarhijskih metoda vrši razvrstavanje objekata u grupe. Cilj analize grupisanja se ogleda u podeli datog seta podataka ili objekata u grupe. Ova podela podrazumeva da su objekti koji pripadaju jednoj grupi među sobom slični, ali i da su objekti iz različitih grupa znatno različiti. To je, ujedno, i smisao klaster analize. U radu su objašnjene procedure analize grupisanja i njihova šira primenjivost.*

Ključne reči: *klaster analiza, procedure*

Abstract: *The process of statistical analysis, as a very complex one, needs to be completly defined, with all its parts, in order to provide information for the entire analytical process. One part of that process is cluster analysis, that by using hierarchical and nonhierarchical methods classifies objects in groups. The aim of a cluster analysis is to divide a given set of data or objects into clusters. This partition means that the objects that belong to the same cluster should be as similar as possible and, also, objects or data that belong to different clusters should be as different as possible. That is, at the same time, the point of a cluster analysis. This paper describes some procedures of those analysis.*

Key words: *cluster analysis, procedures*

Uvod

U literaturi se pod različitim nazivima mogu sresti metode koje, po svojoj prirodi, možemo svrstati u analizu klasifikacije odnosno analizu grupisanja. Najčešće se koristi naziv *klaster analiza* (eng. *cluster analysis*) koji se kao takav odomaćio i u našoj literaturi. U suštini, analiza grupisanja je metod multivarijacione analize koji se koristi za razvrstavanje objekata u grupe.

Da bi se uopšte sprovela analiza grupisanja, neophodno je definisati *mere bliskosti* dva objekta na osnovu njihovih karakteristika. Na bazi mera bliskosti razvijeni su brojni postupci grupisanja objekata, koje možemo razvrstati u dve velike grupe: *hijerarhijski* i *nehijerarhijski metodi*. U osnovi hijerarhijskih metoda leži iterativan proces spajanja objekata u grupe tako da u sledećoj etapi spajamo objekte i prethodno formirane grupe. To zapravo znači da se jednom formirane grupe samo proširuju novim objektima prema specifičnostima izabranog kriterijuma, a da, istovremeno, ne postoji mogućnost prelaska objekta iz jedne u drugu grupu tokom postupka iteracije. Ovu mogućnost daju nehijerarhijski metodi grupisanja.

[*] Nataša Papić-Blagojević (saradnik u nastavi), Visoka poslovna škola strukovnih studija, Novi Sad, Srbija
[**] Denis Bugar (asistent), Visoka poslovna škola strukovnih studija, Novi Sad, Srbija

Ciljevi analize grupisanja

Primarni cilj analize grupisanja je podela objekata u dve ili više grupa na osnovu sličnosti određenih obeležja (klaster varijabla). Ova podela bi trebala da ima sledeće karakteristike:

- homogenost unutar klastera, to jest podaci koji pripadaju istom klasteru bi trebalo da su što sličniji,
- heterogenost između klastera, to jest podaci koji pripadaju različitim klasterima bi trebalo da su što je moguće više različiti.

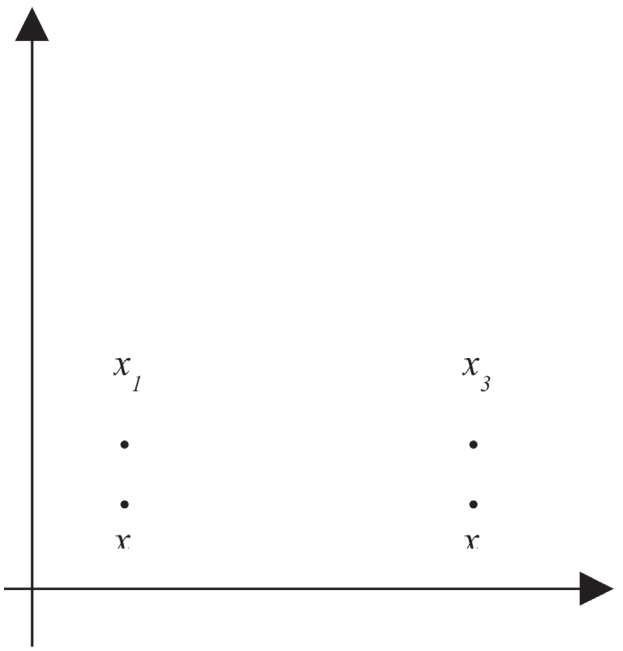
Pri formiranju homogenih grupa, istraživač može postići bilo koji od sledeća tri cilja:

(1) Taksonomija opisa. Najpoznatiji tradicionalni način korišćenja klaster analize je u istraživačke svrhe i za formiranje jednog taksonoma (**taksonomija je oblast sistematike koja se bavi proučavanjem principa, metoda i pravila klasifikacije). Klaster analiza, takođe, može generisati hipoteze koje se odnose na strukturu objekata. Ipak, iako se posmatra prvenstveno kao istraživačka tehnika, analiza grupisanja se može koristiti za potvrdu nečega već ustanovljenog.**

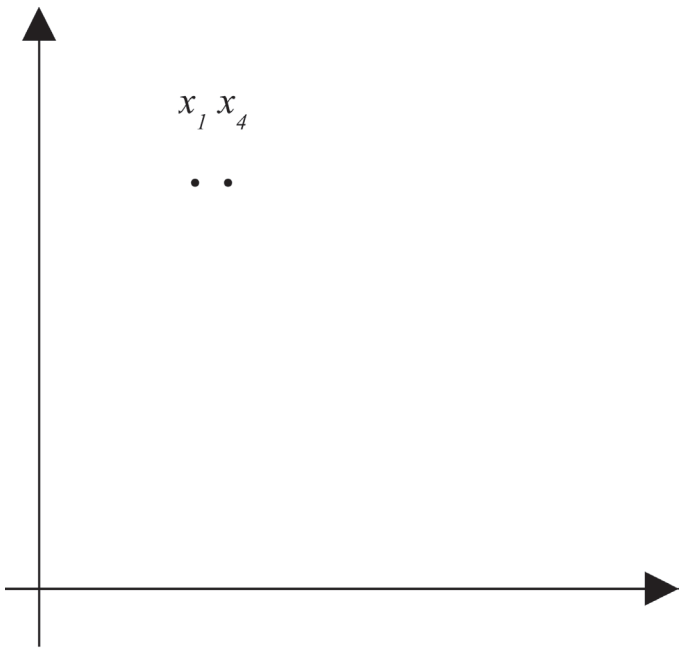
(2) Pojednostavljenje podataka. U toku izvođenja procesa taksonomije, klaster analiza takođe dolazi do pojednostavljenog načina posmatranja. Sa definisanom strukturom zapažanja, podaci mogu biti grupisani u cilju daljih analiza. Dok faktorska analiza pokušava da pruži „dimenzije" ili strukturu promenljivih, klaster analiza obavlja isto to sa posmatranjem. Pa umesto da se sva zapažanja posmatraju kao jedinstvena, ona će biti posmatrana kao članovi klastera i profilisana po svojim opštim karakteristikama.

(3) Identifikacija odnosa. Sa definisanim klasterima i osnovnom strukturom podataka u njima, istraživač objašnjava odnos između posmatranja koje nije bilo moguće sa individualnim posmatranjem. Znači, klaster analiza prikazuje odnos ili sličnosti i razlike koje prethodne analize nisu objavile.

Koncept „sličnosti" se određuje u zavisnosti od samih podataka. S obzirom da su podaci u većini slučajeva vektori stvarnih vrednosti, Euklidska udaljenost između podataka može poslužiti kao mera te različitosti. Treba uzeti u obzir da pojedinačne varijable (elementi vektora) mogu imati različit značaj. Posebno, raspon vrednosti bi trebalo da bude na odgovarajući način gradiran kako bi se dobile razumne vrednosti udaljenosti između podataka. Slike 2 i 3 ilustruju ovaj problem na veoma jednostavnom primeru.



Slika 1. Tačke podataka



Slika 2. Promena skale

Slika 1 pokazuje četiri tačke podataka koje očigledno mogu biti podeljene u dva klastera (x_1, x_2) i (x_3, x_4). Na *Slici* 3, iste tačke podataka su prikazane upotrebom drugačije skale, gde su jedinice na x-osi međusobno bliže, dok su udaljenije na y-osi. Efekat će čak biti jači ukoliko bi se uzelo u obzir hiljadu jedinica za x-osu i milion jedinica za y-osu. Dva klastera se, takođe, mogu prepoznati na *Slici* 2. Međutim, u ovom slučaju se kombinuju tačke x_1 i x_4, kao i x_2 sa x_3.

Nadalje, poteškoće sve više dolaze do izražaja, naročito ukoliko se pored stvarnih varijabli uzmu u obzir i celobrojne vrednosti ili čak apstraktne klase (na primer, tipovi automobila). Naravno, Euklidska udaljenost se može izračunati za celobrojne vrednosti. Ipak, celobrojne vrednosti kod varijabli mogu dovesti do podele klastera, gde se klaster jednostavno pridružuje svakom postojećem celobrojnom broju. To može biti značajno ili u potpunosti nepoželjno u zavisnosti od podataka i pitanja koja se istražuju. Brojevi se mogu pridružiti apstraktnim grupama i tada se Euklidska udaljenost može ponovo primeniti.

Euklidsko odstojanje, kao mera udaljenosti između varijabli, predstavlja specijalni slučaj tzv. *Minkowskog* odstojanja, koje je dato izrazom:

$$Mrs = \left[\sum_{j=1}^{p} \left| x_{rj} - x_{sj} \right|^{\lambda} \right]^{1/\lambda}$$

Za $\lambda = 2$ odstojanje Minkowskog se svodi na Euklidsko odstojanje. Pored ovog, postoji još i *Manhattan odstojanje*, tipa „gradskog bloka", kada je $\lambda = 1$, ali koje je ipak manje u upotrebi.

Metodi grupisanja

Nehijerarhijski metod grupisanja

Nehijerarhijski metod grupisanja objekata dozvoljava mogućnost premeštanja objekata iz ranije formiranih grupa. Do premeštanja objekata će doći ukoliko to sugeriše izabrani kriterijum optimalnosti. U primeni ovih metoda pretpostavlja se da je broj grupa unapred poznat ili ga, kao kod nekih metoda, variramo tokom postupka grupisanja.

Postupak nehijerarhijskog grupisanja započinje inicijalnom podelom skupa objekata u izabran broj grupa. Alternativa inicijalnoj podeli objekata u grupe je apriori određivanje inicijalne klice, odnosno inicijalnog centroida za svaku grupu. Zatim se odredi odstojanje između svakog objekta i svake grupe (inicijalnog centroida). Objekti se lociraju u grupe kojima su najbliže. Nakon pridruživanja objekata nekoj grupi, izračunava se centriod grupe iz koje je objekat „otišao" i grupe kojoj se objekat „pridružio". Ponovo za svaki objekat izračunavamo njegovo odstojanje od centroida grupa i vršimo preraspodelu objekata između grupa sve dotle dok izabrana funkcija kriterijuma to sugeriše.

Najpopularniji među nehijerarhijskim metodama je *metod k-sredina* (eng. *K-means method*). Prema njemu objekat pridružujemo grupi koja ima najbliži centroid (sredinu). Jedna od najčešće korišćenih metoda nehijerarhijskog grupisanja je K-mean algoritam koji podrazumeva sledeće korake:

(1) odredi se broj klasa u koje će se svrstati polazni skup entiteta;
(2) bira se početni centroid (prvih n, slučajnih n, datih n elemenata, datih n centroida);
(3) primenjuje se opšti postupak za bilo koju nehijerarhijsku metodu klasifikacije, a kriterijum zaustavljanja je: ponavljanje sukcesivnih rešenja, dostizanje maksimalnog broja iteracija ili postizanje zadate homogenosti.

U našem primeru, u kom su podaci uzeti iz Odeljenja za logistiku preduzeća FITCO d.o.o. za 2007. godinu uzorak se sastoji od 99 proizvoda, koje smo klaster analizom sveli na tri grupe ili klastera. U tom smislu korišćen je statistički paket SPSS for Windows, s tim što postoje i drugi softveri sličnih karakteristika koji takođe olakšavaju postupak grupisanja.

Broj klastera je izabran proizvoljno (3 klastera). Najpre su izabrani početni centroidi sa koordinatama klasa. Algoritam ove analize je sam izabrao slučajeve koji su dovoljno različiti, kao početne centre. Na osnovu unapred zadatih 10 iteracija, dobijeni su podaci o premeštanju entiteta iz grupe u grupu. Klasifikovanje je završeno u sedmoj iteraciji. Konačni centri klastera dati su u *Tabeli* 1.

Za svaki klaster, centar je aritmetička sredina svih varijabli izračunata na osnovu objekata koji čine klaster. Takvi centri klastera nazivaju se *inicijalni/početni klasterski centri* (eng. *Initial cluster centeres*).

Za svaki novi objekt računa se euklidska udaljenost od inicijalnih klasterskih centara i objekt se svrstava u najbliži klaster. Nakon pridruživanja svih novih objekata moguće je ponovo izračunati centre klastera, koji se sada nazivaju *konačni centri klastera* (eng. *Final Cluster Centers*). Naredna tabela ih prikazuje.

Tabela 1. Konačni centri klastera

	Klaster		
	1	2	3
prodata količina, april	53	276	761
prodata količina, avgust	54	235	517
prodata količina, decembar	62	303	977
vrsta proizvođača	6	5	10
proizvođač	3	3	3

Na kraju, *Tabela* 2. prikazuje koliko entiteta pripada svakom klasteru:

Tabela 2. Broj klastera u svakom centru

Klaster	1	81,000
	2	15,000
	3	3,000
Valid		99,000
Missing		.000

Na osnovu ovako dobijenih rezultata, grupe su dobile sledeće nazive:

1. **Proizvodi malog prodajnog obima** – proizvodi koji se slabo prodaju tokom cele godine;
2. **Proizvodi srednjeg prodajnog obima** – proizvodi čija je prodaja ujednačeno dobra tokom cele godine;
3. **Proizvodi velikog prodajnog obima –** proizvodi čija prodaja po obimu odskače, u pozitivnom smislu, tokom cele godine.

Hijerarhijski metodi grupisanja

Hijerarhijski metodi podrazumevaju izgradnju jedne hijerarhijske strukture nalik drvetu. U osnovi postoje dve vrste hijerarhijskog grupisanja podataka. Kod prve vrste, svaki objekat ili jedinica posmatranja, prema određenom kriterijumu, udružuje se u grupe. U narednim koracima, formiraju se nove grupe udruživanjem ranije formiranih grupa ili individualnih objekata. U narednim iteracijama, ne postoji mogućnost prelaska objekta iz jedne u drugu grupu, odnosno jednom udružene grupe ostaju zajedno. Ovakav vid klasifikovanja nazivamo *hijerarhijskim metodom udruživanja* jer suština leži u iterativnom postupku tokom koga, polazeći od *n* grupa, formiramo jednu grupu. Znači da se veličina grupa povećava, a smanjuje se njihov broj.

Druga grupa metoda, *hijerarhijski metodi deobe*, prelaze isti put, ali u suprotnom smeru. Polazeći od jedne grupe koja sadrži sve objekte, prema određenom kriterijumu, iz iste izdvajamo po jedan objekat ili grupu sve dok se ne formira onoliko grupa koliko ima individualnih objekata.

Najpopularniji metodi grupisanja pripadaju hijerarhijskim metodima udruživanja, a među njima se izdvajaju metodi povezivanja i to, *metod jednostrukog povezivanja*, *metod potpunog povezivanja*, *metod prosečnog povezivanja*, *metod centroida* i *Wardov metod*. Metodi hijerarhijskog udruživanja razlikuju se prema tome kako u drugoj fazi iterativnog postupka određuju međusobnu bliskost grupa.

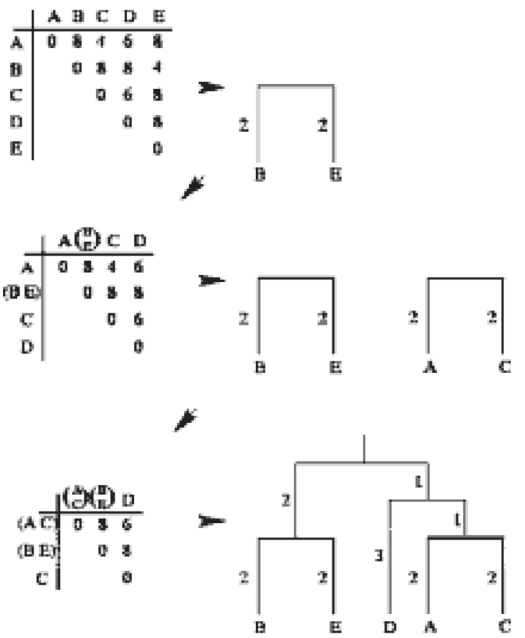
Neke karakteristike ove metode su:

- polazi se od matrice sličnosti među objektima;
- sukcesivno formiranje grupa može se prikazati grafički dijagramom u obliku stabla, koji se naziva dendrogram (grč. *dendros* – stablo);
- metoda zahteva n-1 koraka u formiranju grupa, budući da se na prvom koraku svi pojedinačni objekti tretiraju kao zasebne grupe; na kraju se dobija jedan klaster koji sadrži sve objekte;
- relativno je lako razumljiva širem krugu potencijalnih korisnika.

Pošto hijerarhijska metoda na kraju sve grupe spoji u jednu (ili u obrnutom slučaju početnu jedinstvenu grupu rasturi na entitete) kako znati kada prekinuti grupisanje tj. koliko grupa zadržati? U suštini, grupisanje treba prekinuti onda kada počne spajanje veoma udaljenih grupa ili, u obrnutom slučaju, kada dođe do rasturanja grupa na grupe koje nisu mnogo udaljene.

Hijerarhijski klaster algoritmi imaju za rezultat dendrogram koji reprezentuje grupisanje podataka i nivoe sličnosti na kojima dolazi do promene grupisanja. Dendrogram može biti prelomljen na različitim nivoima kako bi se na taj način izvršilo različito grupisanje podataka. Primer se može videti na *Slici* 3. Sa leve strane, udaljenost između uzoraka se može videti u obliku matrice različitosti. U ovom početnom stadijumu svaka tačka formira pojedinačni klaster.

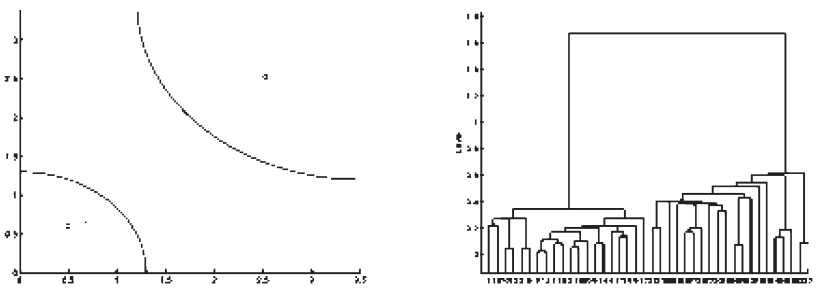
Prvi korak podrazumeva pronalaženje dva najsličnija klastera (dve najbliže tačke podataka). U ovom primeru, postoje dva para sa istom udaljenošću, pa se jedan bira svojevoljno (B i E). Oznake tačaka se upisuju i povezuju u skladu sa oblikom, gde je dužina vertikalne linije jednaka polovini udaljenosti. U drugom koraku, matrica različitosti se menja jer povezane tačke sada obrazuju novi klaster, pa je potrebno izračunati udaljenost između ovog novog klastera i onog prethodnog. Ovi koraci se ponavljaju sve dotle dok ne ostane jedan klaster ili dok se ne dostigne unapred određeni broj klastera.



Slika 3. Formiranje dendrograma

Većina hijerarhijskih klaster algoritama predstavalja različite oblike jednostrukog povezivanja, potpunog povezivanja i algoritama sa minimalnom varijansom. Algoritmi jednostrukog i potpunog povezivanja su među popularnijim.

Jednostavan primer se može videti na sledećoj slici. Sa leve strane, male tačke prikazuju originalne podatke. Jasno je da postoje dva, dobro razdvojena klastera. Rezultate jednostrukog povezivanja možemo pronaći na desnoj strani. Može se utvrditi da je udaljenost između podataka u desnom klasteru veća nego u levom, ali se, isto tako, klasteri mogu jasno razdvojiti.



Slika 4. Rezultati hijerarhijskog i deobnog grupisanja

Ova dva algoritma se razlikuju u načinu izražavanja sličnosti između para klastera. Kod metoda jednostrukog povezivanja, udaljenost između dva klastera je minimalna udaljenost između svih parova tačaka dva klastera (jedan uzorak iz prvog klastera, drugi iz drugog). Kod metoda potpunog povezivanja, udaljenost između dva klastera je maksimalna udaljenost između svih parova tačaka ova dva klastera. U drugom slučaju, dva klastera su spojena kako bi se obrazovao veći klaster zasnovan na kriterijumu minimalne udaljenosti.

Zaključak

Problem koji zbunjuje istraživače klaster analize je određivanje konačnog broja obrazovanih klastera (poznato kao stoping pravila). Nažalost, ne postoji objektivna standardna procedura i ne postoji interni statistički kriterijum za rešavanje ovog problema. Najveći nedostatak je to što istraživači moraju da uključe ad hock metode koje su inače relativno kompleksne. Jedna vrsta stoping pravila je relativno prosto istraživanje mera sličnosti ili rastojanja između klastera u svakom uzastopnom koraku, sa definisanim klaster rešenjima kada je mera sličnosti jedna određena vrednost. Jednostavniji primer za to se osvrće na veliki rast prosečnog rastojanja unutar klastera. Kada usledi jedan jači skok tada istraživači klaster rešenja pribegavaju logici koja je kombinacija znatnog pada u sličnostima. Time se u empirijskim istraživanjima dolazi do prilično tačnih odluka. Druga vrsta stoping pravila odnose se na jednu formu statističkih pravila gde se primenjuje adaptirani statistički test.

Postoji određeni broj specifičnih procedura koje su predložene ali se ni jedna nije pokazala kao najbolja u svim situacijama. Takođe, istraživači moraju dati čvrste procene, sa konceptom teorijskih odnosa koji može predložiti prirodan broj klastera. Može se pokrenuti proces u kojem određeni kriterijumi, na osnovu praktičnih ispitivanja, pokazuju da rezultati moraju biti pregledni i razumljivi za upotrebu kada se poseduje prirodan broj klastera, tj. od 3–6, i tada najbolje rešenje za ovaj broj klastera je izbor najbolje alternative posle njihove procene. U konačnoj analizi je verovatno najbolje da se uzme jedan broj klaster rešenja (npr. 2, 3, 4) i da se donese odluka sa alternativnim rešenjima, koristeći apriori kriterijume i praktičnu ocenu, zdrav razum ili teorijske ocene. Klaster rešenja će biti poboljšana kada se nađu rešenja za konceptualne aspekte problema.

Osnovni problem jeste gde povući crtu, tako da ostane optimalni broj klastera. Treba reći da ovaj problem nema zadovoljavajuće rešenje. Iterativne metode zahtevaju od korisnika da unapred odredi broj klastera. U statističkom smislu nulta-hipoteza o nepostojanju strukture unutar nekog skupa objekata nije sasvim jasna, pa ni smislena.

U društvenim naukama dominiraju dva pristupa određivanju broja klastera: heuristički pristup i formalni testovi. Prvi pristup je najčešći, a odnosi se na subjektivno postavljanje granice na dendrogramu dobijenom hijerahijskom klasterizacijom. Osnovni kriterijum jeste smislenost ili interpretabilnost dobijenog rešenja. Drugi način, podjednako subjektivan jeste analiza koeficijenata (koeficijenti fuzije) koji pokazuju sličnosti među klasterima pri sukcesivnom spajanju klastera. Naglo opadanje (ili povećanje vrednosti kod mera udaljenosti) ukazuje na manju povezanost među klasterima koji se spajaju. Nagli skok ukazuje na spajanje dva relativno različita klastera.

Na kraju, važno je zaključiti da klaster analiza daje istraživačima empirijsku i objektivnu metodu za izvođenje jednog od najbitnijih zadataka kao što je klasifikacija. Da li za svrhu uprošćavanja, istraživanja ili potvrde, analiza grupisanja je jedan vrlo moćan analitički aparat koji ima vrlo široku primenu. Ipak, ova tehnika povlači odgovornost istraživača, pa je nužna određena doza opreza prilikom njenog korišćenja. Ukoliko se pravilno koristi, ova analiza ima potencijal da otkrije podatke koji do tada nisu otkriveni pomoću drugih metoda. Takođe, pravilno rukovanje zahteva veliko znanje, kako se zbog loše upotrebe ne bi javili pogrešni rezultati i zaključci.

Literatura

[1] Abonyi, J., Feil, B., (2007) *Cluster Analysis for Data Mining and System Identification*, Berlin, Germany, Birkhauser Verlag, AG

[2] Hopner, F., Klawonn, F., (1999) Fuzzy *Cluster Analysis: Methods for Classification, Data Analysis and Image Recognation*, West Sussex, England, John Wiley & Sons Ltd

[3] Ivanović, B., (1988) *Grupisanje obeležja preko metoda automatske klasifikacije*, „Zbornik radova: II majski skup '88 sekcije za klasifikacije SSDJ-a, Mostar", Beograd, Savezni zavod za statistiku, Institut za statistiku

[4] Ivanović, B., (1977) *Teorija klasifikacije*, Beograd, Institut za ekonomiku industrije

[5] Kovačić, Z., (1998) *Multivarijaciona analiza*, Beograd, Ekonomski fakultet

[6] www.komunikacija.org.rs