

TRENDOVI U POSLOVANJU

Kombinovana primena hijerarhijskih i nehijerarhijskih metoda klasterizacije u cilju segmentacije kupaca u jednom trgovinskom lancu

Sažetak

U ovom radu primenjene su hijerarhijske i nehijerarhijske metode klaster analize, u cilju utvrđivanja relativno homogenih grupa kupaca u jednom trgovinskom lancu tokom prethodnih pet godina. Nakon izdvajanja objekata i određenih atributa iz SQL Server baze podataka i određivanja skupa relevantnih varijabli koje će reprezentovati obeležja grupe, primenjen je K-means algoritam u MATLAB programskom okruženju, a kao mera sličnosti korišćen je kvadrat Euklidskog rastojanja. Na osnovu analize ukupnog broja kupljenih artikala, frekvencije kupovina, ukupnog prometa po kupcu i broja kontaktiranja kompanije prema kupcu, predložen je broj klastera, koji je tokom izvršavanja algoritma fiksiran, a u cilju detekcije razdvojenosti i kompaktnosti klastera, korišćen je Silhouette indeks. Korišćeno je fiksno pet klastera, a takođe je primenjen i hijerarhijski metod, da bi se postavljanjem granice na dendrogramu validirao pretpostavljeni broj klastera, a čime se takođe može eliminisati i problem uticaja slučajnog izbora početnog položaja centroida pri izvršavanju K-means algoritma. Dobijeni rezultati mogu se koristiti u smislu planiranja prodajnih kampanja, optimizacije troškova marketinga, predlaganja novih programa lojalnosti, u cilju boljeg razumevanja potrošačkog ponašanja, kao i pravljenja

posebnih planova poslovnih aktivnosti za svaki klaster pojedinačno.

Abstract

U ovom radu primenjene su hijerarhijske i nehijerarhijske metode klaster analize, u cilju utvrđivanja relativno homogenih grupa kupaca u jednom trgovinskom lancu tokom prethodnih pet godina. Nakon izdvajanja objekata i određenih atributa iz SQL Server baze podataka i određivanja skupa relevantnih varijabli koje će reprezentovati obeležja grupe, primenjen je K-means algoritam u MATLAB programskom okruženju, a kao mera sličnosti korišćen je kvadrat Euklidskog rastojanja. Na osnovu analize ukupnog broja kupljenih artikala, frekvencije kupovina, ukupnog prometa po kupcu i broja kontaktiranja kompanije prema kupcu, predložen je broj klastera, koji je tokom izvršavanja algoritma fiksiran, a u cilju detekcije razdvojenosti i kompaktnosti klastera, korišćen je Silhouette indeks. Korišćeno je fiksno pet klastera, a takođe je primenjen i hijerarhijski metod, da bi se postavljanjem granice na dendrogramu validirao pretpostavljeni broj klastera, a čime se takođe može eliminisati i problem uticaja slučajnog izbora početnog položaja centroida pri izvršavanju K-means algoritma. Dobijeni rezultati mogu se koristiti u smislu planiranja prodajnih kampanja, optimizacije troškova marketinga, predlaganja novih programa lojalnosti, u cilju boljeg razumevanja

potrošačkog ponašanja, kao i pravljenja posebnih planova poslovnih aktivnosti za svaki klaster pojedinačno.

Ključne reči: dendrogram; klaster analiza; K-means algoritam; segmentacija tržišta; Silhouette indeks

Keywords: dendrogram; cluster analysis; K-means algorithm; market segmentation; Silhouette index

1. Uvod

Osnovni cilj klaster analize je pronalaženje grupa objekata takvih da su objekti u grupi međusobno slični ili povezani, dok su objekti u različitim grupama međusobno različiti ili nepovezani. Metode klaster analize dele se na dve osnovne grupe: aglomerativne metode, koje grupišu taksonomske jedinice u grupe (klaster) po srodnim osobinama, i divizivne metode, koje razbijaju skup taksonomskih jedinica na više grupa (klastera) [1]–[3]. Metode klaster analize se takođe dele i na: hijerarhijske metode, koje daju niz sukcesivnih particija skupa odnosno podela na klaster, koji se obično prikazuje specijalnim dijagramom – dendrogramom, i nehijerarhijske metode, koje daju samo jednu, optimalnu podelu na klaster, čiji broj može, ali ne mora biti unapred poznat [4], [5].

Prilikom kategorizacije kupaca u jednom trgovinskom lancu, u ovom radu primenjeno je nehijerarhijsko particiono klasterovanje, sa ciljem podele skupa podataka (kupaca) u nepreklapajuće podskupove takve da je svaki podatak tačno u jednom podskupu. Predloženi su klasteri zasnovani na centru, što znači da klaster predstavlja skup objekata takvih da je bilo koji objekat u klasteru bliže (ili više sličan) centru klastera u odnosu na centre ostalih klastera. Centar klastera je uobičajeno centroid (prosek svih tačaka u klasteru) ili medoid (najreprezentativnija tačka u klasteru), a često se početni centroid bira na slučajan način [6]–[8], dok se broj klastera definiše pre procesa izvršavanja algoritma [9]–[11]. Na osnovu analize određenih atributa iz SQL Server baze podataka jednog trgovinskog lanca, dat je

predlog grupisanja kupaca koji imaju slično potrošačko ponašanje i funkcionalnosti [12], [13], prema određenim kategorijama, a na osnovu čega je predloženo pet klastera. Kao atributi od interesa korišćeni su: količina potrošenog novca, broj kupovina, kao i broj kontaktiranja sa klijentom, tokom proteklih pet godina.

U cilju validacije prertpostavljenog broja klastera, primenjen je i hijerarhijski metod klasterizacije, koji kao krajnji rezultat ima dendrogram, odnosno grafički prikaz klastra (grupa) u obliku stabla povezivanja. Prilikom primene hijerarhijskih metoda, prvo se vrše izračunavanja udaljenosti svih objekata međusobno, a zatim se grupe formiraju korišćenjem tehnika spajanja ili razdvajanja [5], [6]. Suprotno tome, kod nehijerarhijskih metoda vrši se raščlanjavanje, tako da je moguće premeštanje objekata iz jedne u drugu grupu tokom različitih faza analize.

Prilikom primene particionog klasterovanja korišćen je model sa prototipom, pri čemu su početni centroidi birani u početku slučajan način, a kasnije na osnovu podatka dobijenog na osnovu dendograma. Svaki podatak iz baze dodeljen je klasteru sa najbližim centroidom, a u svakoj iteraciji vršen je proračun položaja centroida za sve klaster. Broj klastera je postavljen na fiksnu vrednost, koja se tokom izvršavanja algoritma ne menja, dok se podaci iterativno razvrstavju u cilju postizanja što veće homogenosti klastera. Kao mera sličnosti korišćeno je Euklidsko rastojanje, a kao mera razdvojenosti, odnosno kompaktnosti klastera, korišćen je Silhouette indeks [14].

Opis algoritma, izbor skupa relevantnih varijabli i vizuelizacija dobijenih klastera prikazani su u drugom poglavlju. U trećem poglavlju prikazani su segmentirani kupci na dijagramima koji ilustruju povezanost broja kupovina, ukupne vrednosti kupovine, vrste proizvoda, kao i broja i načina ostvarenih komunikacija sa kupcem. U četvrtom poglavlju prikazani su rezultati hijerarhijskog klasterovanja i analizirane su mogućnosti postavljanja granice na dendrogramu koristeći standardno Euklidsko rastojanje i kosinusnu meru sličnosti. U zaključnom poglavlju

iznete su mogućnosti primene i dalje optimizacije predloženog algoritma.

2. Grupisanje kupaca primenom nehijerarhijskih metoda klaster analize

Primarni cilj grupisanja klaster analizom je podela objekata u dve ili više grupa na osnovu sličnosti određenih obeležja, predstavljenih klaster varijablama [5]. Homogenost unutar klastera podrazumeva da su podaci koji pripadaju istom klasteru u što većoj meri slični, dok heterogenost između različitih klastera podrazumeva da su podaci koji pripadaju različitim klasterima u što većoj meri različiti, pri čemu se koncept sličnosti određuje u zavisnosti od samih podataka. U ovom radu je korišćena Euklidska udaljenost između podataka kao mera sličnosti, a podaci o kupcima su na odgovarajući način gradirani, da bi raspon podataka dao razumne vrednosti udaljenosti između podataka. Centroidi, koji predstavljaju centre klastera, ne moraju biti podaci iz dataset-a, jer predstavljaju tačke koje su dobijene usrednjavanjem parametara tačaka koji pripadaju datim klasterima [10]. Kao izlazni podatak, dobija se takođe i pripadnost svakog podataka jednom od K klastera, čiji je broj definisan pre primene K -means algoritma klasterizacije.

Na osnovu analize određenih atributa iz SQL Server baze podataka jednog trgovinskog lanca, dat je predlog grupisanja kupaca koji imaju slično potrošačko ponašanje i funkcionalnosti, prema određenim kategorijama, a na osnovu čega je predložen broj klastera. Kao atributi od interesa korišćeni su: količina potrošenog novca, broj kupovina, vrsta kupljenog proizvoda, kao i broj i način kontaktiranja sa kupcem, tokom preteklih pet godina, a predloženo je pet klastera [12], [13]. Klaster 1 čine kupci koji često kupuju, ali ne troše mnogo novca. Ljubitelji su kupovine, a kompanija ih kontaktira povremeno. Zauzimaju skoro četvrtinu baze. Klaster 2 čine skromni i oprezni kupci, sa relativno malim brojem kupovina i relativno malom količinom novca koji ostavljaju. Najbrojniji su. Kompanija ih retko kontaktira, a predstavljaju najveći deo baze, nešto više od

polovine. Klaster 3 predstavljaju kupci koji ostavljaju dosta novca prilikom kupovine, a posebno u odnosu na manji (selektovani) broj kupovina. Kompanija ih često kontaktira, a čine malo više od 5% baze. Klaster 4 čine kupci koji vole da kupuju, često kupuju, ali ne tako skupe stvari. Kupuju uglavnom jeftinije proizvode i ne ostavljaju puno novca. Lojalni su i često dolaze, a kompanija ih često kontaktira. Čine nešto manje od 10% baze. Klaster 5 predstavljaju veoma bogati kupci koji ne dolaze često i ne kupuju često, ali kupuju vrlo skupe proizvode. Kompanija ih ne kontaktira često, a čine nešto manje od 10% baze. Tokom vršenja algoritma, broj klastera nije menjan.

U cilju detekcije razdvojenosti i kompaktnosti klastera, korišćen je Silhouette indeks, S . Neka je $X_T = \{X_1, \dots, X_N\}$ skup podataka, pri čemu $C = (C_1, \dots, C_K)$ predstavlja klasterovanje tog skupa na K klastera. Neka je $d(X_k, X_l)$ rastojanje između X_k i X_l . Neka je $C_j = \{X_1^j, \dots, X_{m_j}^j\}$ j -ti klaster, $j=1, \dots, K$, gde je $m_j = |C_j|$. Srednje rastojanje a_i^j između i -tog vektora u klasteru C_j i drugih vektora u istom klasteru određeno je sledećim izrazom:

$$a_i^j = \frac{1}{m_j - 1} \sum_{\substack{k=1 \\ k \neq i}}^{m_j} d(X_i^j, X_k^j), \quad i=1, \dots, m_j \quad (1)$$

Minimalno srednje rastojanje između i -tog vektora u klasteru C_j i svih vektora iz klastera C_k , $k=1, \dots, K$, $k \neq j$, određeno je sa:

$$b_i^j = \min_{\substack{n=1, \dots, K \\ n \neq j}} \left\{ \frac{1}{m_n} \sum_{\substack{k=1 \\ k \neq i}}^{m_n} d(X_i^j, X_k^n) \right\}, \quad i=1, \dots, m_j \quad (2)$$

pri čemu je d predstavlja rastojanje koje može biti računato kao Euklidsko, rastojanje Minkovskog, Manhattan rastojanje i sl. [3]. Širina siluete i -tog vektora klastera C_j određena je sa:

$$s_i^j = \frac{b_i^j - a_i^j}{\max(a_i^j, b_i^j)} \quad (3)$$

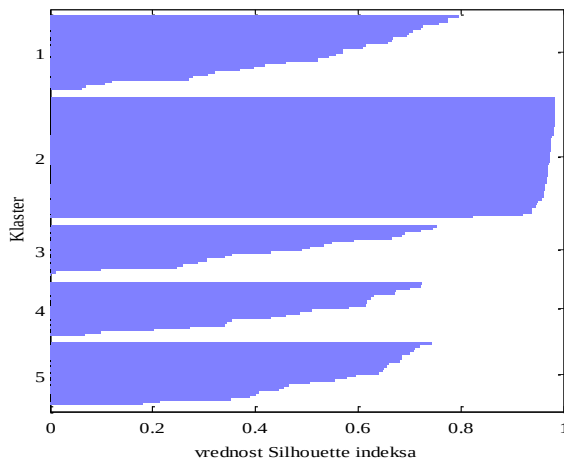
na osnovu čega sledi $-1 \leq s_i^j \leq 1$. Širina siluete klastera C_j određena je sa:

$$S_j = \frac{1}{m_j} \sum_{i=1}^{m_j} s_i^j \quad (4)$$

dok je globalni Silhouette indeks klasterovanja definisan sa:

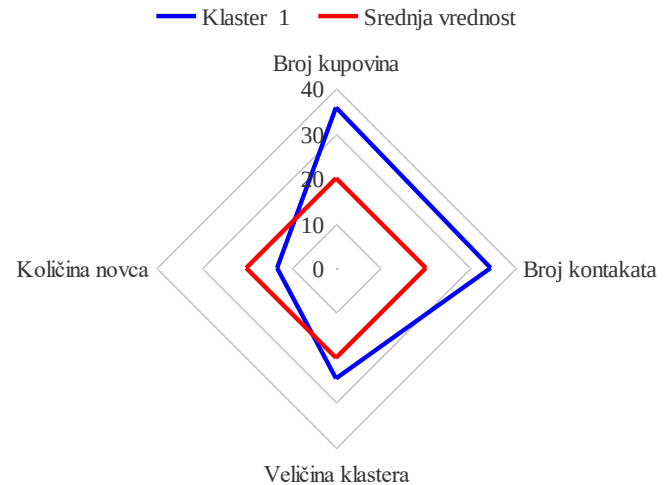
$$S = \frac{1}{K} \sum_{j=1}^K S_j \quad (5)$$

što ima za posledicu da širina klastera, kao i globalni Silhouette indeks klasterovanja, uzimaju vrednosti između -1 i 1. Vrednost Silhouette indeksa dobijena upotrebom silhouette funkcije u MATLAB programskom okruženju [14] za predložene klastere, prikazana je na Sl.1.



Slika 1. Vrednost Silhouette indekse za predložene klastere kupaca u trgovinskom lancu

Vizuelni prikaz klastera, u poređenju sa srednjom vrednošću relevantnih parametara, dat je na Sl.2, Sl.3, Sl.4, Sl.5. i Sl.6.



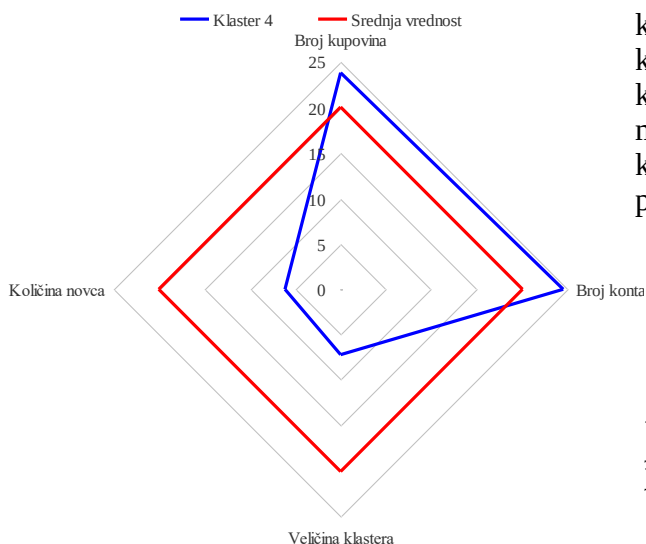
Slika2. Vizuelni prikaz klastera 1



Slika3. Vizuelni prikaz klastera 2



Slika 4. Vizuelni prikaz klastera 3



Slika 5. Vizuelni prikaz klстера 4



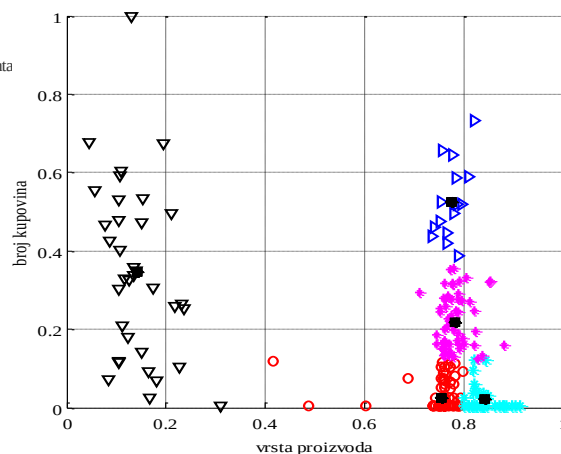
Slika 6. Vizuelni prikaz klстера 5

3. Prikaz segmentiranih kupaca u zavisnosti od relevantnih atributa

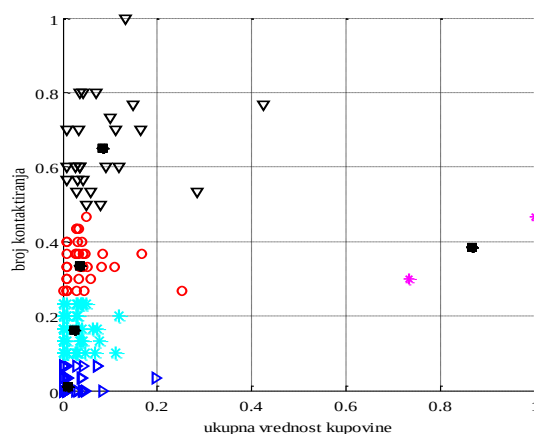
Vizuelna reprezentacija klastera, na dijagramima zavisnosti različitih atributa, data je na Sl.7, Sl.8, Sl.9. i Sl.10. Ovi dijagrami vizuelizuju klasifikaciju kupaca po klasterima, u zavisnosti od odnosa broja i iznosa kupovina, vrste proizvoda, kao i broja kontaktiranja kompanije prema kupcu.

Predložena analiza ima za cilj da se uoče i definišu neka rešenja bazirana na identifikaciji grupa unutar date baze kupaca, odnosno da se segmentacijom izdvoje grupe sa sličnim potrošačkim ponašanjem. Rezultati imaju primenu u smislu

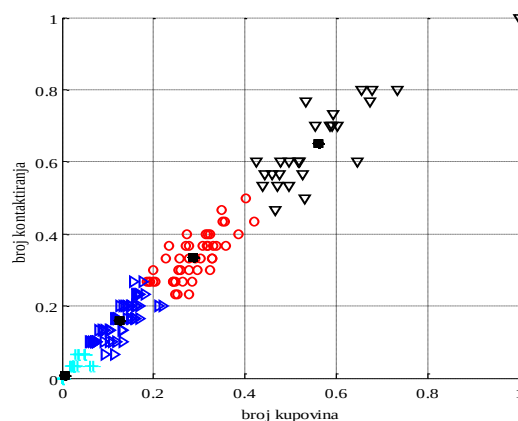
kvalitetnijeg planiranja angažovanja kupaca kroz odgovarajuće kanale i komunikaciju, kao i u smislu definisanja i formiranja mogućih lojaliti programa, s obzirom da su kod određenih klastera primećene sličnosti po nekim atributima.



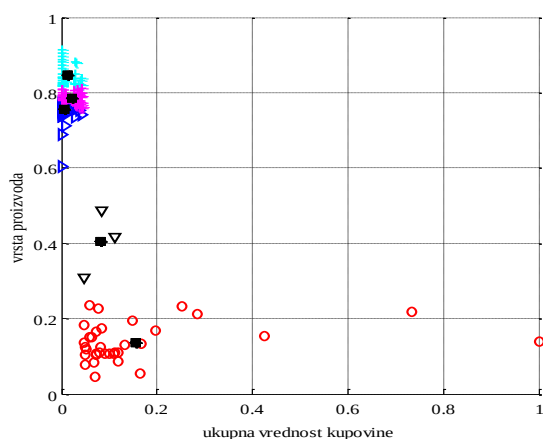
Slika 7. Vizuelni prikaz po klsterima broj kupovina – vrsta proizvoda



Slika 8. Vizuelni prikaz po klsterima broj kontaktiranja – ukupna vrednost kupovine



Slika 9. Vizuelni prikaz po klsterima broj kontaktiranja – broj kupovina



Slika 10. Vizuelni prikaz po klasterima vrsta proizvoda – vrednost kupovine

Prikazana raspodela po klasterima dobijena je kroz više uzastopnih izvršavanja programa, koja se u nekoj meri mogu razlikovati, s obzirom da je početni centroid biran na slučajan način, uz moguću specifikaciju minimalne udaljenosti. Najveći deo konvergencije u svim izvršavanjima ostvaren je u prvim nekoliko iteracija, a u samom programskom kodu maksimalan broj iteracija postavljen je na 40. Nisu primećene značajne razlike ni ako se ovaj broj smanji na 20. S obzirom se u svakoj iteraciji izračunavaju nove tačke grupisanja (centroidi), vizuelni prikaz raspodele po klasterima prikazuje se u vremenskim razmacima od 0.5 s, da bi se i vizuelno isparatilo moguće pomeranje podataka iz jednog klastera u drugi. Proces se odvija iterativno, sve do postizanja stabilnosti, što je realizovano kroz prvih nekoliko iteracija.

Pregled osnovnih karakteristika klastera dat je i u Tabeli 1., sa ciljem da se i kroz numeričke pokazatelje jasnije izdvoje sličnosti i razlike između određenih klastera, po ciljnim atributima. Uočava se sličnost klastera 3 i klastera 5 po ukupnoj vrednosti kupovina, dok su po broju kupovina najbližiji klaster 3 i klaster 4. Klaster 2, iako ga čini najveći deo baze (više od 50% svih kupaca iz baze), donosi samo 11% od ukupnog prihoda za razmatrani period.

Tabela 1. Numerički podaci o klasterima

	Broj kupovina	Broj kontakata	Veličina klastera	Uk.vredn. kupovina
Klaster1	453 [35.78%]	3331 [34.47%]	58 [24.47%]	112940348.5 [13.04%]
Klaster2	86 [6.81%]	601 [6.22%]	126 [53.16%]	95652496.1 [11.01%]
Klaster3	294 [23.21%]	2269 [23.48%]	15 [6.32%]	320227993.8 [36.88%]
Klaster4	301 [23.77%]	2364 [24.47%]	17 [7.17%]	53384435.9 [6.14%]
Klaster5	128 [10.40%]	1096 [11.34%]	21 [8.86%]	286062335.2 [32.94%]
Sr.vrdnost	252.4 [20%]	1932.2 [20%]	47.4 [20%]	173653521.9 [20%]

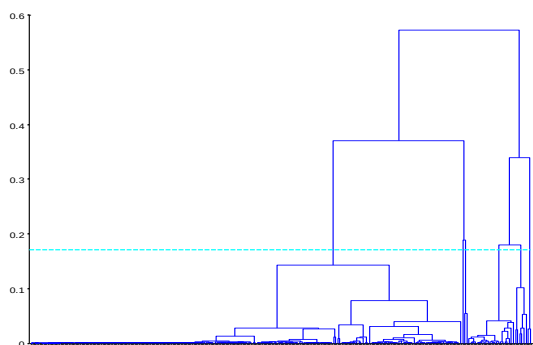
4. Validacija broja predloženih klastera postavljenjem granice na dendrogramu dobijenim hijerarhijskom metodom klasterizacije

Hijerarhijske metode klasterizacije kao krajnji rezultat daju dendrogram, koji predstavlja grafički prikaz klastera u obliku stabla povezivanja, pri čemu se prvo vrše izračunavanja udaljenosti svih objekata međusobno, a zatim se formiraju grupe koristeći tehnike spajanja ili razdvajanja [1]–[3]. Primenom hijerarhijskih metoda klasteri se formiraju korak po korak, sve dok se sve jedinice posmatranja ne nađu na dendrogramu, a nakon toga se može pristupiti određivanju broja klastera koji imaju značaj za analizu.

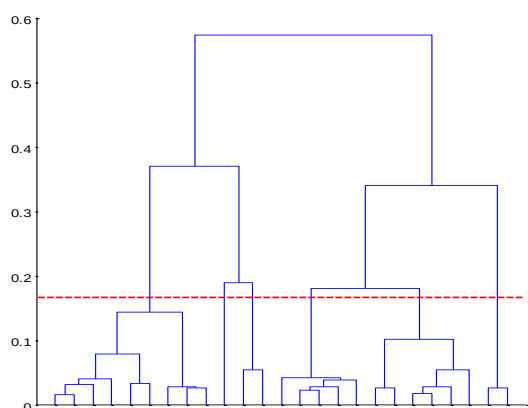
Hijerarhijsko stablo pogodno za ilustraciju rasporeda klastera, prikazano je na Sl.11., koristeći standardno Euklidsko rastojanje, dok je su rezultati hijerarhijskog grupisanja prikazani pomoću dendrograma dobijenog koristeći kosinusnu meru sličnosti [15], ilustrovani na Sl.12. Kosinusna sličnost između dva objekta d_1 i d_2 određena je sa:

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \|d_2\|} \quad (6)$$

pri čemu • označava skalarni proizvod.



Slika 11. Postavljanje subjektivne granice na dendrogramu dobijenim hijerarhijskom klasterizacijom koristeći standardno Euklidsko rastojanje



Slika 12. Postavljanje subjektivne granice na dendrogramu dobijenim hijerarhijskom klasterizacijom koristeći kosinusnu meru sličnosti

Na horizontalnoj osi prikazani su objekti, a na vertikalnoj rastojanje između objekata. Prilikom deljenja dendrograma u cilju dobijanja određenog broja grupa (klastera), uobičajeno se traže veći skokovi na vertikalnoj osi, kao što je prikazano na Sl.11. i Sl.12., gde je horizontalnim presecanjem dendrograma izdvojeno pet klastera.

Analizom horizontalnih preseka stabla može se uočiti napredak podele grafa na određene grupacije, a ova tehnika se može primeniti u različitim fazama istraživanja. U ovom radu je prilikom kreiranja stabla korišćeno jednostruko povezivanje (metod najbližeg suseda), a kao mera sličnosti

Euklidsko rastojanje i kosinusna mera sličnosti. Primena metode najudaljenijeg suseda, kao i metode prosečnog povezivanja koja koristi prosečnu udaljenost između svih objekata različitih grupacija, nisu dali značajno drugačije rezultate.

U okviru budućih istraživanja planirano je da u istraživanje budu uključene još neke varijable, kao što su datumi kupovina, sa ciljem izdvajanja sezonskih kupaca, zatim starost i pol kupaca, što bi mogao biti indikator za određivanje kvalitetnijeg načina angažovanja kupaca kroz odgovarajuće komunikacione kanale i bolje usmerenu komunikaciju sa kupcima, kao i targetiranje tržišta. Dostupnost demografskih podataka o klijentu, kao i poznavanje situacionih faktora, zatim pristupa kupovini i nivoa lojalnosti brendu, doprineli bi kvalitetnijoj segmentaciji, što u ovom radu nije realizovano, s obzirom da je iz baze bio doputan samo jedinstveni identifikacioni broj kupca.

5. Zaključak

Klaster analizom se na algoritamski jednostavan način određuju strukturalne karakteristike izmerenih svojstava objekata, grupisanih na matematičkoj, ali ne i statističkoj utemeljenosti, što čini klaster analizu vrlo korisnom i atraktivnom tehnikom istraživanja, iako nema osnove strogog statističkog zaključivanja. Osnovni nedostatak je to što rešenja nisu jedinstvena, a istraživač mora da proceni uticaj svake odluke uključene u izvođenje klaster analize.

Formiranje grupa, odnosno klastera, u slučaju analiziranog trgovinskog lanca, korisno je u cilju lakšeg upoređivanja klijenata, optimizacije troškova marketinga, osmišljavanja grupnih kampanja prodaje i predlaganja novih ili unapređenja postojećih lojaliti programa, a takođe i u cilju izdvajanja manje aktivnih ili odlazećih kupaca.

Od velikog značaja je i samo interpretiranje dobijenih rezultata klaster analize, što podrazumeva vrlo dobro poznavanje teorijske pozadine istraživanja, sa ciljem se dobijeni rezultati što bolje iskoriste za postavljanje novih ili potvrdu postojećih hipoteza u datom domenu istraživanja.

U ovom radu su, sa ciljem da se iskoriste neke prednosti i hijerarhijskih i nehijerarhijskih metoda klaster analize, korišćene obe tehnike sukcesivno, što je rezultovalo kvalitetnijom interpretacijom dobijenih grupa kupaca (klastera) i validacijom, da bi dobijeni klasteri bili što relevantniji za korisnike istraživanja.

Bibliografija

1. T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, 2009.
2. A. Agresti, Categorical Data Analysis, 2nd ed., Wiley, New York, 2002.
3. T. Anderson, An Introduction to Multivariate Statistical Analysis, 3rd ed., Wiley, New York, 2003.
4. C. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.
5. A. K. Jain, "Data clustering: 50 years beyond K-means", Pattern Recogn. Lett. 31, 2010, pp. 651-666.
6. M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm", Expert Syst. Appl., 2013, pp. 200-210.
7. R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a dataset via the gap statistic", Journal of the Royal Statistical Society, 2001, Series B. 32(2) pp. 411-423.
8. H. Stefanović, R. Veselinović, G. Bjelobaba, A. Savić, "An adaptive car number plate image segmentation using K-means clustering", Proceedings of Int. Scientific Conference on Information Technology and Data Related Research-SINTEZA 2018, 2018, pp. 74-78.
9. Z. Wang, Z. Xu, S. Liu, and Z. Yao, "Direct clustering analysis based on intuitionistic fuzzy implication", Applied Soft Computing 23, 2014, pp. 1-8.
10. T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient kmeans clustering algorithm: Analysis and implementation", IEEE Trans. Pattern Anal. Mach. Intell. 24(7), 2002, pp. 881-892.
11. J. Nayak, D.P. Kanungo, B. Naik, and H.S. Behera, "Evolutionary improved swarm-based hybrid K-means algorithm for cluster analysis", Proceedings of Int. Conf. on Computer and Communication Technologies, Springer, New Delhi, 2016, pp. 343-352.
12. C. McDaniel and R. Gates, Marketing Research, 10th ed, John Wiley & Sons, 2014.
13. C.-H. Cheng and Y.-S. Chen, "Classifying the segmentation of customer value via RFM model and RS theory", Expert Systems with Applications 36, 2009, pp. 4176-4184.
14. <https://www.mathworks.com/help/stats/silhouette.html>
15. <https://www.mathworks.com/help/stats/dendrogram.html>