PREDNOSTI I OGRANIČENJA METODA ZA KLASTERIRANJE PODATAKA U DIMENZIJI IDENTIFIKACIJE KLIJENATA

ANTE PANJKOTA
Odjel za ekonomiju i Pomorski odjel
Sveučilište u Zadru
Splitska 1, 23000 Zadar, Croatia
apanjkot@unizd.hr

IVA ŠALOV
Odjel za ekonomiju
Sveučilište u Zadru
Splitska 1, 23000 Zadar, Croatia
isalov@student.unizd.hr

ALEKSANDRA KRAJNOVIĆ
Odjel za ekonomiju
Sveučilište u Zadru
Splitska 1, 23000 Zadar, Croatia
akrajnov@unizd.hr

SAŽETAK

Brojni autori predložili su 4-dimenzionalni CRM model kojeg čine identifikacija klijenata, privlačenje klijenata, razvoj odnosa s klijentima i zadržavanje klijenata. Prednosti i ograničenja različitih klasterskih metoda pri identifikaciji klijenata koje je moguće naći u raspoloživoj akademskoj literaturi predmet su ovog rada. Najčešće korištene metode u toj dimenziji su k-means ili njegove izvedenice i različite metode tzv. mekog klasteriranja (npr. Fuzzy C-means), a poslovna područja u kojima se uobičajeno koriste za identifikaciju klijenata su maloprodaja, telekomunikacije, bankarstvo, turizam i energetika. Jednostavnost upotrebe, dobra i razumljiva segmentacija korisnika su osnovne prednosti navedenih pristupa, dok su problemi procjene broja klastera, nemogućnosti direktne ugradnje dinamičkog ponašanja klijenata u klasterske strukture i tzv. problem višedimenzionalnosti uočeni kao glavna ograničenja primjene metoda klasteriranja u dimenziji identifikacije klijenata. Ovaj rad daje nekoliko smjernica koje bi mogle biti korisne strategije u nadvladavanju uočenih ograničenja.

KLJUČNE RIJEČI: klasteriranje, identifikacija klijenata, k-means, meko klasteriranje, CRM

UVOD

Brojni autori predložili su 4-dimenzionalni CRM model s dimenzijom identifikacije korisnika, dimenzijom privlačenja korisnika, dimenzijom zadržavanja korisnika i dimenzijom razvoja odnosa s korisnicima [Swift, 2001], [Parvatiyar and Sheth, 2001] i [Kracklauer *et al.*, 2003]. Za druge CRM dimenzionalne modele dostupne u literaturi upućujemo na 5-dimenzionalni model [Josiassen, Assaf and Cvelbar, 2014], te na drugi 4-dimenzionalni model [Yim, Anderson and Swaminathan, 2004]. Budući da ovaj rad obrađuje primjenu klasterskih metoda rudarenja podataka u okviru CRM-a, odlučili smo se na prvi navedeni 4-dimenzionalni model koji je primijenjen u često citiranom preglednom radu [Ngai, Xiu and Chau, 2009].

U ovom radu od posebnog interesa je dimenzija identifikacije korisnika sa stajališta primjene različitih metoda klasteriranja. Osnovni preduvjet uspješne primjene bilo kojeg tehnološkog rješenja u bilo kojem području, pa tako i pri identifikaciji korisnika u CRM sustavu, jest poznavanje njegovih prednosti i ograničenja. Analizom raspoložive literature identificirani su najčešće korišteni algoritmi klasteriranja u toj dimenziji, poslovna područja učestale primjene istih te su utvrđene njihove prednosti i ograničenja u istaknutoj zadaći u okviru CRM sustava. Također, sintetizirane su osnovne preporuke za što efikasniju primjenu klasterskih algoritama u dimenziji identifikacije korisnika.

Važno je istaknuti kako promatranje čisto tehnološke komponente u ovom radu ne zanemaruje važnost holističkog pogleda na CRM koji strateški integrira CRM filozofiju u organizacijsku kulturu tvrtke i pripadajuće operativne procese podržane najnovijim tehnološkim rješenjima izbjegavajući time zamku neuspjeha CRM koncepta uslijed ograničavanja na isključivu primjenu tehnologije [Nasir, 2017]. Takav holistički pristup prelazi okvire ovog rada, te ujedno predstavlja jedno od njegovih važnijih ograničenja.

Rad je dalje koncipiran kako slijedi: cjelina 2 donosi osnovne teorijske osnove klasterskih metoda, cjelina 3 određuje metodološki okvir istraživanja, u cjelini 4 prikazani su najvažniji rezultati istraživanja, dok je značaj istih prodiskutiran u cjelini 5. Posljednju cjelinu čini zaključak rada.

OSNOVNA TEORIJSKA RAZMATRANJA UZ METODE KLASTERIRANJA PODATAKA

Metode klasteriranja podataka spadaju u grupu algoritama nenadziranog učenja (engl. unsupervised learning) koji se zasnivaju na pronalaženju pravilnosti skrivenih isključivo u samim ulaznim podacima bez ikakve informacije o vrijednostima izlaznih varijabli, u ovom slučaju oznaka klasa. Osnovna pretpostavka koja se nalazi iza svakog algoritma klasteriranja jest postojanje neke nepoznate zakonitosti po kojoj su objekti grupirani. Nepoznata zakonitost nadomješta se izabranom ili izabranim mjerama sličnosti prema kojoj se objekti nastoje razvrstati u područja prema vrijednostima indeksa sličnosti. Pojednostavljeno, konačno grupiranje postiže se optimizacijskim algoritmom funkcije indeksa sličnosti na način da se

dobije što viši stupanj homogenosti grupe objekata. Konačno grupiranje objekata, odnosno kvalitetu dobivenih klastera, potrebno je evaluirati nekom prikladnom mjerom od kojih su najpoznatije Davies-Bouldin indeks, Dunn indeks, Calinski-Harabasz indeks, Silhoutte koeficijenti i tzv. Gap statistika [Mathworks, 2017]. Detaljan pregled klasterskih metoda, pripadajućih indeksa sličnosti i optimizacijskih funkcija cilja, te evaluacijskih mjera moguće je naći u radovima [Xu, 2005; Xu and Tian, 2015].

Osnovna pitanja koja treba uzeti u razmatranje kada je riječ o primjeni klasterskih algoritama u bilo kojoj domeni definirana su u knjizi [Jain and Dubes, 1988], a sasvim su relevantna i danas: Što je klaster?; Koje ulazne atribute koristiti za klasteriranje?; Zahtjevaju li podaci normalizaciju?; Sadrže li podaci netipične vrijednosti (engl. outliers)?; Kako je definirana prikladna mjera sličnosti?; Koliki je broj klastera?; Koju metodu klasteriranja upotrijebiti za dani problem?; Posjeduju li podaci tendenciju prema stvaranju klastera?; Kakva je valjanost dobivenih klasterskih struktura?

Sličnim pitanjima bavi se i rad [Hiziroglu, 2013]: Kada je moguće segmentirati tržište?; Koje varijable određuju segmentaciju?; Koju metodu segmentacije izabrati za problem od interesa?; Na koji način odrediti cilj segmentacije?; Što predstavlja jedinice analize?; Kako izvršiti izbor uzorka?; Koji su izvori podataka?; Kako opisati podatke?; Postoji li potreba za standardizacijom/normalizacijom podataka?; Kako odrediti broj segmenata (klastera)?; Koliko je pouzdano dobiveno rješenje?

Navedene dvije reference s istaknutim pitanjima poslužit će nam za definiranje metodološkog okvira u razmatranju prednosti i ograničenja primjene metoda klasteriranja u dimenziji identifikacije klijenata.

METODOLOŠKI OKVIR ZA ODREĐIVANJE PREDNOSTI I OGRANIČENJA METODA KLASTERIRANJA U DIMENZIJI IDENTIFIKACIJE KLIJENATA

Prednosti i ograničenja primjene metoda klasteriranja u dimenziji identifikacije klijenata promatrana su izučavanjem raspoložive literature, preciznije analizom 27 znanstvenih članaka koji obrađuju problematiku od interesa (prilog 1, tablica 1). Elementi metodološkog okvira za analizu dobiveni su sintezom i indukcijom dobivenih rezultata, zaključaka i nalaza nekoliko radova [Jain and Dubes, 1988], [Hiziroglu, 2013], [Xu and Tian, 2015] i [Ngai, Xiu and Chau, 2009].

CILJ IDENTIFIKACIJE

U dimenziji identifikacije klijenata osnovno pitanje jest „Tko su naši klijenti"? U tom pogledu interesantno je ciljati na potencijalne klijente ili odrediti najprofitabilnije klijente, te analizirati vlastite klijente i klijente konkurencije prema lojalnosti [Thanuja, Venkateswarlu and Anjaneyulu, 2011]. Pored tih ciljeva identifikacija klijenata se vrši i zbog optimizacije resursa

[Wang *et al.*, 2014; Biscarri *et al.*, 2017] ili pak zbog predviđanja budućih prihoda po grupama korisnika [Khajvand and Tarokh, 2011].

POSLOVNO PODRUČJE PRIMJENE

Klasterske metode u identifikaciji klijenata koriste se u različitim područjima poslovanja, primjerice u telekomunikacijama [Hamka *et al.*, 2014]. Nedvojbeno je da uz neki manji broj osnovnih karakteristika koje su zajedničke svakom području poslovanja postoji relativno veliki broj specifičnosti po kojima se područja poslovanja međusobno znatno razlikuju. U slučajevima gdje područje poslovanja nije navedeno koristit ćemo pojam generička primjena pripadnog klasterskog algoritma u identifikaciji klijenata.

TIP PODATAKA

Algoritmi klasteriranja s obzirom na izabranu mjeru sličnosti i pridruženu optimizacijsku funkciju cilja direktno su ovisni o tipu podataka s kojim mogu operirati. S obzirom na tip podataka algoritme klasteriranja ćemo promatrati kroz prizmu kvantitativnih podataka (kontinuirani i diskretni), te kvalitativnih (nominalni i ordinalni) ili mješovitih podataka (koji sadrže kvalitativne i kvantitativne podatke). Sva tri tipa podataka mogu biti statički i dinamički.

OSJETLJIVOST NA NETIPIČNE VRIJEDNOSTI

Netipične vrijednosti mogu znatno utjecati na performanse algoritama klasteriranja na način da konačna klasterska struktura znatno iskrivljuje prirodni raspored objekata distorzirajući ga prema netipičnim vrijednostima [Hautamaki *et al.*, 2005].

ODREĐIVANJE BROJA KLASTERA

Broj klastera može biti određen od strane eksperta iz domene problema, dobiven postupkom vizualizacije podataka kada je to moguće, proračunat nekom od brojnih automatiziranih metoda [Jain, 2010] ili pak dobiven poluautomatiziranim metodama koje se temelje na grafičkom prikazu kriterija klasteriranja u odnosu na broj klastera [Kononenko and Kukar, 2007].

PROBLEM VIŠEDIMENZIONALNOSTI PODATAKA

Za veliki broj klasterskih algoritama kojima je sličnost definirana nekom mjerom udaljenosti (npr. Euklidska ili udaljenost Minkowskog) ili algoritama koji na bilo koji način implementiraju mjere udaljenosti u osnovni modus rada svojstveno je da su podložni problemu višedimenzionalnosti koji se u literaturi nalazi pod nazivom „Curse of dimensionality" [Har-Peled, Indyk and Motwani, 2012]. Riječ je o problemu pri kojem u slučaju broju dimenzija većem od 6 sve udaljenosti postaju relativno jednake, a ta jednakost je tim više izražena što je broj dimenzija viši.

IMPLEMENTACIJA

Postojanje gotovih paketa ili biblioteka koje implementiraju pojedini algoritam znatno olakšava njegovu primjenu, posebice što se često radi o softverski optimiziranim rješenjima. Primjerice,

u programskom paketu WEKA implementiran je čitav niz metoda za klasteriranje podataka [Sharma, Bajpai and Litoriya, 2012].

MOGUĆNOST OBJAŠNJENJA DOBIVENIH RJEŠENJA

Kakvu vrijednost konačna klasterska struktura ima za ekperte iz domene problema, odnosno koliko kvalitetno mogu objasniti pojave koje su generirale podatke iz samih klastera u velikoj mjeri ovisi o jednostavnosti razumijevanja dobivenih rješenja. Ovaj zahtjev u direktnoj je vezi s izborom ulaznih atributa veće deskriptivne sposobnosti [Greene and Cunningham, 2005].

PRETPROCESIRANJE PODATAKA

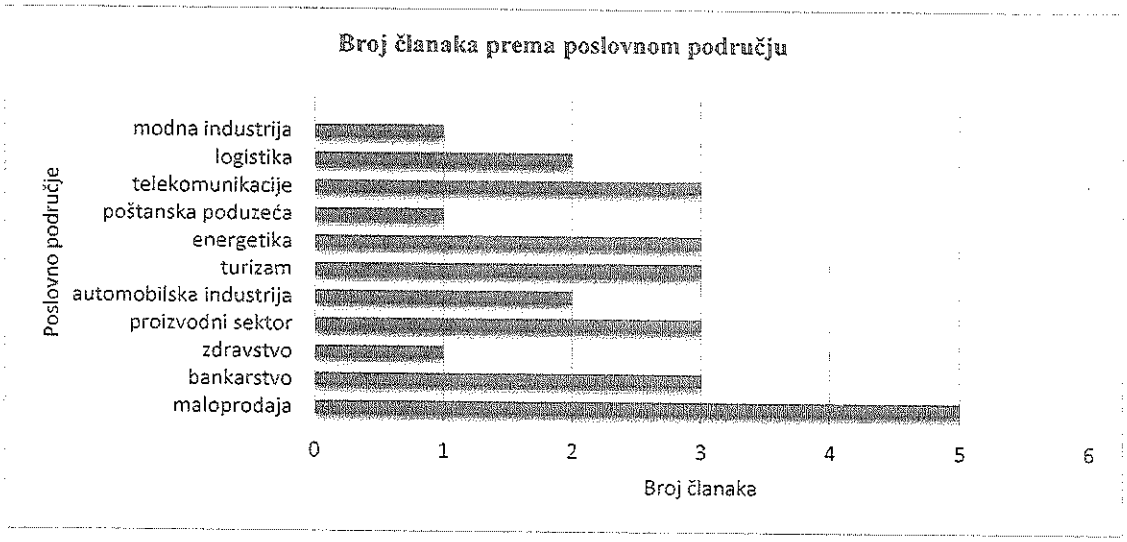
Skupovi podataka iz realnih domena pohranjeni u baze podataka ili u skladišta podataka često su karakterizirani vrijednostima koje nedostaju, sadrže smetnje, irelevantne ili redundantne atribute, netipične vrijednosti i nekonzistentnost u označavanju ili izražavanju vrijednosti što predstavlja osnovu za pretprocesiranje podataka. Različite mjere između vrijednosti atributa mogu predstavljati probleme za većinu klasterskih algoritama pa se u tom pogledu redovito vrši normalizacija ili standardizacija. U nekim primjenama pretprocesiranje podrazumijeva i smanjivanje broja dimenzija selekcijom ili ekstrakcijom atributa, te diskretizaciju ulaznih varijabli [Witten *et al.*, 2016].

REZULTATI

Potpuni podaci analize prema izdvojenim elementima metodološkog okvira iz 27 akademskih članaka prikazani su u Prilogu 1 Tablica 1. U ovoj cjelini prikazat ćemo samo najvažnije rezultate za ovaj rad.

Broj članaka prema području poslovanja dan je prikazom 1. Poslovno područje je direktno po sebi jasno i predstavljeno je nazivom područja kako su navedena od strane autora, dok u slučajevima kada nisu navedena podrazumijevamo da se radi o generičkoj primjeni klasteriranja u funkciji identifikacije klijenata.. Iz grafičkog prikaza lako je uočiti da se metode klasteriranja u dimenziji identifikacije korisnika prema analiziranoj literaturi najviše primjenjuju u maloprodaji, a potom podjednako u bankarstvu, proizvodnom sektoru, turizmu, energetici i telekomunikacijama. Najmanje je zabilježena primjena u zdravstvu, modnoj industriji i poštanskim poduzećima.

Prikaz 1 Grafički prikaz broja članaka po poslovnom području

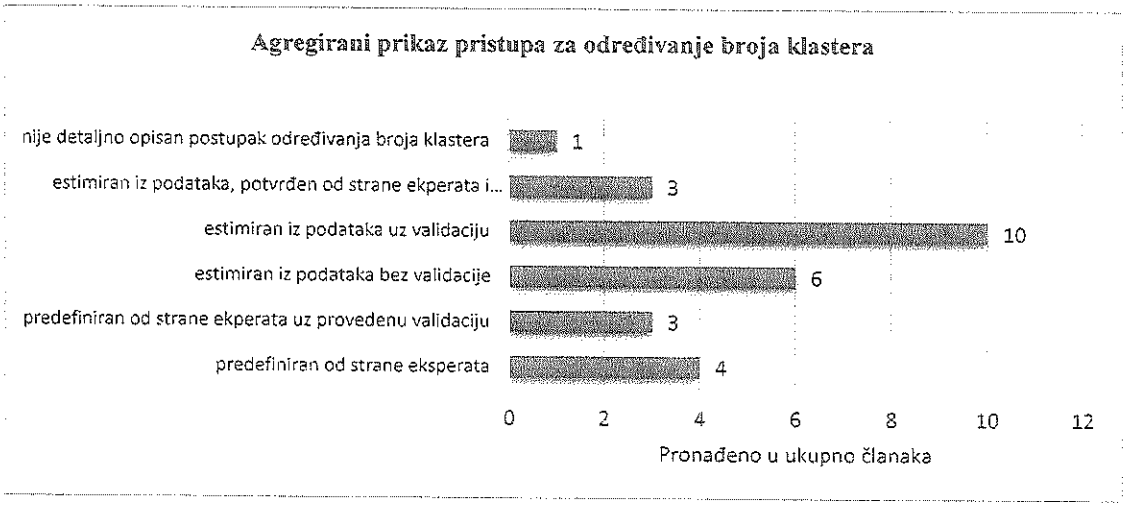


Prikaz 2 donosi agregirane podatke korištenih pristupa za određivanje broja klastera u analiziranoj literaturi. Lako uočavamo da je najviše radova s pristupom koji broj klastera određuje estimacijom iz podataka, a da pri tome ne koriste nikakav validacijski postupak. Samo u jednoj publikaciji nije opisan postupak određivanja broja klastera.

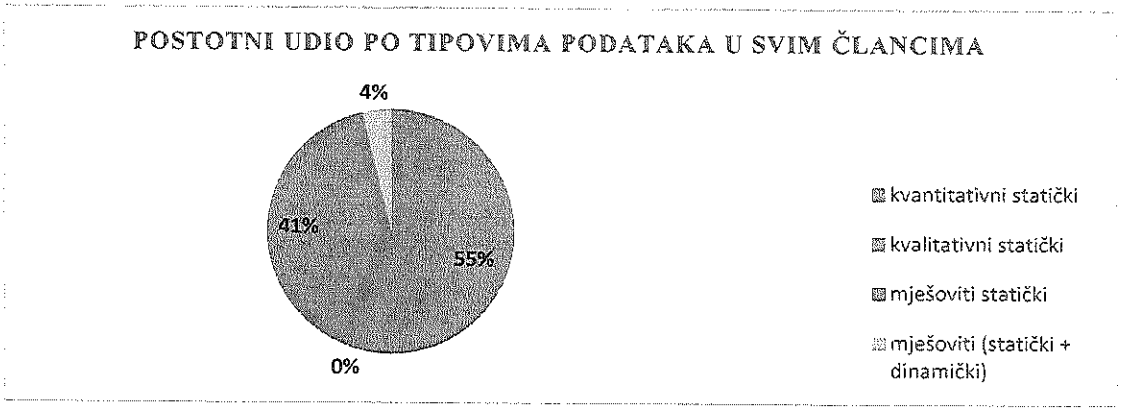
Postotni udio po tipovima podataka u svim člancima vidimo na prikazu 3. U analiziranoj literaturi najviše su zastupljeni problemi klasteriranja s kvantitativnim statičkim i mješovitim statičkim podacima. Ovdje statičnost određuje vrijednosti podataka koji su u biti povijesni podaci iz baza ili skladišta podataka, te u sebi ne uključuju bilo kakvu kontinuiranu vremensku ovisnost. Nasuprot njima, dinamički podaci uključuju vremensku ovisnost. Kvalitativni statički podaci nisu pronađeni u nijednom radu, a mješoviti sa statičkim i dinamičkim karakteristikama pojavljuju se samo u jednoj referenci od 27 analiziranih.

Osjetljivost promatranih klasterkih algoritama na problem višedimenzionalnosti koji značajno narušava pouzdanost dobivenih rješenja nalazi se na prikazu 4. Većina algoritama je veoma osjetljiva na problem višedimenzionalnosti, dok su prema dostupnim podacima u tri članka upotrijebljeni klasterski algoritmi koji su neosjetljivi na ovaj problem.

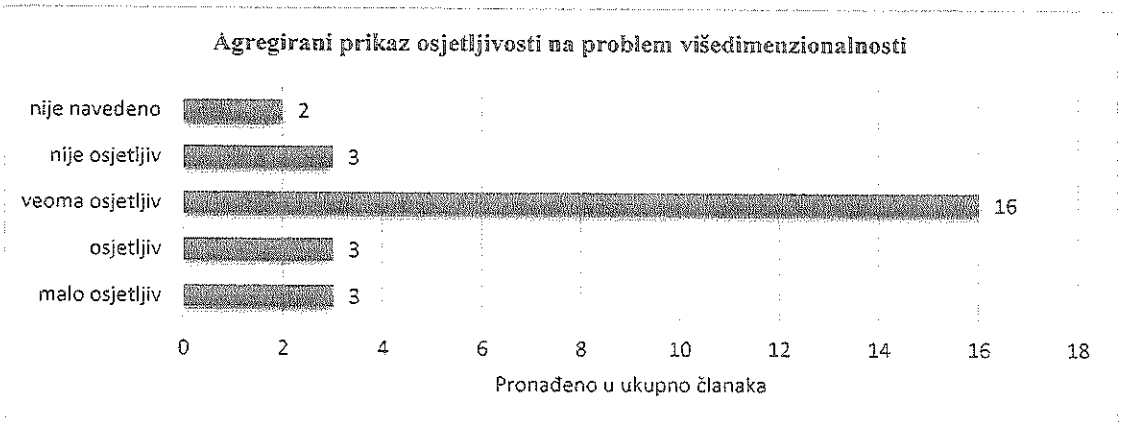
Prikaz 2 Grafički prikaz korištenih pristupa za određivanje broja klastera pronađenih u analiziranim člancima



Prikaz 3 Grafički prikaz postotnog udjela tipova podataka za klasteriranje u svih 27 članaka



Prikaz 4 Grafički prikaz agregiranih podataka iz 27 članaka o osjetljivosti klasterskih metoda u identifikaciji klijenata na problem višedimenzionalnosti



U svih 27 analiziranih članaka autori navode kao prednost jednostavnost objašnjenja dobivenih rješenja klasteriranja i mogućnost brze interpretacije istih od strane eksperata. U 21 promatranom slučaju provedeno je pretprocesiranje podatka, a za 6 nisu navedeni podaci o

pretprocesiranju podataka. U 12 radova se koristi k-means algoritam ili njegove izvedenice, a u 6 neka od metoda fuzzy klasteriranja čime su te dvije grupe algoritama najzastupljenije u dimenziji identifikacije klijenata na prikupljenom uzorku od 27 članka. U čak 18 radova autori su primijenili gotovo softversko rješenje ili paket koji implementira izabrani klasterski algoritam (vidi prilog 1).

RASPRAVA

Poznato je da e-trgovinu karakterizira visoka razina interakcije s krajnjim korisnicima što dovodi do mogućnosti bilježenja velikog broja podataka o potrošačkim navikama, sklonostima, interesima, potrebama i sl. [Akter and Wamba, 2016]. Upravo su to neki od najvažnijih razloga zašto su članci iz područja maloprodaje o primjeni klasterskih metoda u identifikaciji klijenata najzastupljeniji u prikupljenoj literaturi. Isti razlozi sektor bankarstva, turizma i telekomunikacija dovode na drugo mjesto zastupljenosti promatrane primjene, dok su se energetika i proizvodni sektor, protivno očekivanjima, našli u istoj grupi. Razloge je moguće tražiti u činjenici da se u tim područjima javlja sve veći interes za optimizacijom proizvodnih resursa u ovisnosti o klijentskim potrebama ili navikama što potvrđuju i podaci u tablici 1 iz priloga 1 (cilj identifikacije).

Izbor atributa kojim se vrši modeliranje klijenata u svrhu njihove identifikacije klasteriranjem ovisit će o specifičnostima promatranog poslovnog područja i ciljevima identifikacije, a sama deskriptivna snaga klasteriranja, pored specifičnosti metode klasteriranja, bit će veća što su ekspresivne mogućnosti skupa atributa veće. Samim tim, poželjno je uz kvantitativne tipove atributa (npr. RFM) koristiti i kvalitativne (npr. demografske) statičke prirode u slučajevima kada njihova vremenska ovisnost nije od interesa, odnosno dinamičke prirode kada je potrebno razmotriti dinamičku prirodu ponašanja klijenata (npr. vremenski slijed transakcija kupaca) u kojima se navedena dinamika treba uzeti u obzir [Liu and Chen, 2017].

Određivanje broja klastera krucijalan je faktor pri primjeni bilo koje klasterske metode [Witten *et al.*, 2016], što je potvrđeno i iz analizirane literature u ovom radu. Standardna procedura podrazumijeva određivanje broja klastera i njihovu validaciju. Da bi se metode klasteriranja mogle kvalitetno primijeniti u realnim slučajevima identifikacije klijenata nužno je još tim postupcima pridružiti mišljenje eksperata iz promatranog poslovnog područja.

Za klasterske algoritme koji kao indeks sličnosti koriste neku mjeru udaljenosti između objekata svojstveno je da su osjetljivi na problem višedimenzionalnosti [Xu and Tian, 2015]. Iz ovog rada proizlazi da je najveći broj primjena metoda klasteriranja u segmentaciji klijenata veoma osjetljiv na ovaj problem. Za takve slučajeve nužno je izvršiti smanjenje dimenzionalnosti – npr. metodom ekstrakcije ili selekcije atributa uz obveznu suradnju s ekspertima [Boutsidis *et al.*, 2015].

U 21 članku autori navode i opisuju postupak pretprocesiranja podatka prije samog postupka klasteriranja čime se stavlja naglasak na važnost tog postupka pri segmetaciji tržišta. Pretprocesiranje podrazumijeva uključanjanje netipičnih vrijednosti, filtriranje smetnji, odbacivanje irelevantnih i redundantih atributa, pretvorbu atributa iz jednog tipa u drugi,

standardizaciju/normalizaciju, nadomiještanje podataka koji nedostaju i slično [Witten *et al.*, 2016].

Uzimajući u obzir da su k-means i njegovi derivati, te fuzzy klasteriranje s izvedenicama najzastupljeniji u promatranoj literaturi to je primjena jednostavna jer postoji veliki broj gotovih, optimiziranih i provjerenih softverskih rješenja i paketa s njihovom implementacijom. Bilo bi korisno istražiti sposobnosti identifikacije klijenata primjenom i drugih metoda klasteriranja prema osobinama ulaznih atributa i s obzirom na prirodu promatranih problema [Xu and Tian, 2015].

Jednostavnost objašnjenja i visoka razumljivost dobivenih klasterskih struktura od strane eksperata se pokazala kao jedna od temeljnih prednosti primjene algoritama klasteriranja i u ovom radu (svih 27 članka) što se podudara sa svojstvima tih algoritama navedenim u osnovnoj literaturi [Kononenko and Kukar, 2007; Witten *et al.*, 2016].

ZAKLJUČAK

Dimenzija identifikacije klijenata iz 4-dimenzionalnog CRM modela predstavlja područje intenzivnih napora istraživačke zajednice. Ovo istraživanje je pokazalo da u promatranoj dimenziji svoju široku primjenu nalaze metode klasteriranja. Na uzorku od 27 relevantnih članka promatrane problematike u primjeni najviše dominiraju k-means i njegove izvedenice i različite verzije fuzzy klasteriranja. Maloprodaja, bankarstvo, telekomunikacije, turizam, energetika i proizvodnja predstavljaju najzastupljenija poslovna područja primjene klasterskih algoritama za segmentiranje klijenata.

Kao osnovne prednosti klasterskih algoritama u toj zadaći pokazale su se jednostavnost primjene i razumljivost dobivenih deskriptivnih rješenja. Prvo detektirano ograničenje je visoka osjetljivost na problem višedimenzionalnosti što je moguće razriješiti primjenom neke od metoda redukcije dimenzija ili odabirom klasterskog algoritam neosjetljivog na taj problem. Drugo zabilježeno ograničenje je nemogućnost ugradnje dinamičke prirode ponašanja klijenata u korištene algoritme klasteriranja. Odgovor na to ograničenje moguće je potražiti u prikladnom vremenskom modeliranju ulaznih atributa koji opisuju dinamički karakter problema uz apliciranje algoritama klasteriranja koji mogu raditi s takvim tipovima podataka. Treće ograničenje je problem složenosti i nestandardiziranosti postupka određivanja broja klastera koji bi se trebao sastojati od tri faze: estimacije broja klastera iz podataka, validacije dobivenog rješenja i revizije validiranog rješenja od strane eksperata iz promatranog poslovnog područja u suradnji s ekspertima za rudarenje podataka. Provjera valjanosti predloženih rješenja za nadvladavanje uočenih ograničenja primjene metoda klasteriranja u dimenziji identifikacije klijenata predstavlja pravce za buduća istraživanja.

REFERENCE

[1] Akter, S. and Wamba, S. F. (2016) 'Big data analytics in E-commerce: a systematic review and agenda for future research', *Electronic Markets*. Electronic Markets, 26(2),

pp. 173–194. doi: 10.1007/s12525-016-0219-0.

[2] Ansari, A. and Riasi, A. (2016) 'Customer Clustering Using a Combination of Fuzzy C-Means and Genetic Algorithms', *International Journal of Business and Management*, 11(7), p. 59. doi: 10.5539/ijbm.v11n7p59.

[3] Biscarri, F., Monedero, I., García, A., Guerrero, J. I. and León, C. (2017) 'Electricity clustering framework for automatic classification of customer loads', *Expert Systems with Applications*, 86(May), pp. 54–63. doi: 10.1016/j.eswa.2017.05.049.

[4] Boutsidis, C., Zouzias, A., Mahoney, M. W. and Drineas, P. (2015) 'Randomized Dimensionality Reduction for k -Means Clustering', 61(2), pp. 1045–1062.

[5] Brito, P. Q., Soares, C., Almeida, S., Monte, A. and Byvoet, M. (2015) 'Customer segmentation in a large database of an online customized fashion business', *Robotics and Computer-Integrated Manufacturing*. Elsevier, 36, pp. 93–100. doi: 10.1016/j.rcim.2014.12.014.

[6] Chen, Y. S., Cheng, C. H., Lai, C. J., Hsu, C. Y. and Syu, H. J. (2012) 'Identifying patients in target customer segments using a two-stage clustering-classification approach: A hospital-based assessment', *Computers in Biology and Medicine*. Elsevier, 42(2), pp. 213–221. doi: 10.1016/j.compbiomed.2011.11.010.

[7] D'Urso, P., Giovanni, L. De, Disegna, M. and ... (2013) 'Bagged Clustering and its application to tourism market segmentation', *Expert Systems with ...*, 40, pp. 4944–4956. Available at: http://www.sciencedirect.com/science/article/pii/S0957417413001553.

[8] Dursun, A. and Caber, M. (2016) 'Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis', *Tourism Management Perspectives*, 18, pp. 153–160. doi: 10.1016/j.tmp.2016.03.001.

[9] Greene, D. and Cunningham, P. (2005) 'Producing accurate interpretable clusters from high-dimensional data', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3721 LNAI, pp. 486–494. doi: 10.1007/11564126_49.

[10] Gupta, G., Aggarwal, H. and Rani, R. (2016) 'Segmentation of retail customers based on cluster analysis in building successful CRM', *International Journal of Business Information Systems*, 23(2), pp. 212–228. doi: 10.1504/IJBIS.2016.078907.

[11] Hamka, F., Bouwman, H., De Reuver, M. and Kroesen, M. (2014) 'Mobile customer segmentation based on smartphone measurement', *Telematics and Informatics*. Elsevier Ltd, 31(2), pp. 220–227. doi: 10.1016/j.tele.2013.08.006.

[12] Har-Peled, S., Indyk, P. and Motwani, R. (2012) 'Approximate Nearest Neighbor: Towards Removing the Curse of Dimensionality', *Theory of Computing*, 8(1), pp. 321–350. doi: 10.4086/toc.2012.v008a014.

[13] Hautamaki, V., Cherednichenk, S., Karkkainen, I., Kinnunen, T. and Franti, P. (2005) 'Improving K-Means by Outlier Removal', *Scia 2005, Lncs 3540*, pp. 978–987. doi: 10.1007/11499145_99.

[14] Hiziroglu, A. (2013) 'Soft computing applications in customer segmentation: State-of-art review and critique', *Expert Systems with Applications*. Elsevier Ltd, 40(16), pp. 6491–6507. doi: 10.1016/j.eswa.2013.05.052.

[15] Ho, G. T. S., Ip, W. H., Lee, C. K. M. and Mou, W. L. (2012) 'Customer grouping for better resources allocation using GA based clustering technique', *Expert Systems with Applications*, 39(2), pp. 1979–1987. doi: 10.1016/j.eswa.2011.08.045.

[16] Jain, A. K. (2010) 'Data clustering: 50 years beyond K-means', *Pattern Recognition Letters*. Elsevier B.V., 31(8), pp. 651–666. doi: 10.1016/j.patrec.2009.09.011.

[17] Jain, A. K. and Dubes, R. C. (1988) *Algorithms for clustering data*. Prentice-Hall, Inc.

[18] Josiassen, A., Assaf, A. G. and Cvelbar, L. K. (2014) 'CRM and the bottom line: Do all CRM dimensions affect firm performance?', *International Journal of Hospitality Management*. Elsevier Ltd, 36, pp. 130–136. doi: 10.1016/j.ijhm.2013.08.005.

[19] Kashwan, K. R. and Velu, C. M. (2013) 'Customer Segmentation Using Clustering and Data Mining Techniques', *International Journal of Computer Theory and Engineering*, (April), pp. 856–861. doi: 10.7763/IJCTE.2013.V5.811.

[20] Khajvand, M. and Tarokh, M. J. (2011) 'Estimating customer future value of different customer segments based on adapted RFM model in retail banking context', *Procedia Computer Science*. Elsevier, 3, pp. 1327–1332. doi: 10.1016/j.procs.2011.01.011.

[21] Kolarovszki, P., Tengler, J. and Majerčáková, M. (2016) 'The New Model of Customer Segmentation in Postal Enterprises', *Procedia - Social and Behavioral Sciences*, 230(May), pp. 121–127. doi: 10.1016/j.sbspro.2016.09.015.

[22] Kononenko, I. and Kukar, M. (2007) *Machine learning and data mining: introduction to principles and algorithms*. Horwood Publishing.

[23] Kracklauer, A. H., Mills, D. Q., Seifert, D. and Mills, D. Q. (2003) 'Collaborative customer relationship management'.

[24] Li, D. C., Dai, W. L. and Tseng, W. T. (2011) 'A two-stage clustering method to analyze customer characteristics to build discriminative customer management: A case of textile manufacturing business', *Expert Systems with Applications*. Elsevier Ltd, 38(6), pp. 7186–7191. doi: 10.1016/j.eswa.2010.12.041.

[25] Liu, Y. and Chen, Y. (2017) 'Customer Clustering Based on Customer Purchasing Sequence Data', *Int. Journal of Engineering Research and Application*, 7(1), pp. 49–58.

[26] López, J. J., Aguado, J. A., Martín, F., Muñoz, F., Rodríguez, A. and Ruiz, J. E. (2011) 'Hopfield-K-Means clustering algorithm: A proposal for the segmentation of electricity customers', *Electric Power Systems Research*, 81(2), pp. 716–724. doi: 10.1016/j.epsr.2010.10.036.

[27] Luo, Y., Cai, Q., Xi, H., Liu, Y. and Zhu, G. (2013) 'Customer segmentation for telecom with the k-means clustering method', *Information Technology Journal*, 12(3). doi: 10.3923/itj.2013.409.413.

[28] Mathworks, I. (2017) *Documentation: Statistics and Machine Learning Toolbox*. Available at: https://www.mathworks.com/help/ (Accessed: 18 August 2017).

[29] Miguéis, V. L., Camanho, A. S. and Falcão E Cunha, J. (2012) 'Customer data mining for lifestyle segmentation', *Expert Systems with Applications*, 39(10), pp. 9359–9366. doi: 10.1016/j.eswa.2012.02.133.

[30] Murray, P. W., Agard, B. and Barajas, M. A. (2015) 'Forecasting supply chain demand

by clustering customers', *IFAC-PapersOnLine*. Elsevier Ltd., 28(3), pp. 1834–1839. doi: 10.1016/j.ifacol.2015.06.353.

[31] Nasir, S. (2017) 'Customer Relationship Management as a Customer-Centric Business Strategy', in *Advertising and Branding: Concepts, Methodologies, Tools, and Applications*. IGI Global, pp. 649–685.

[32] Ngai, E. W. T., Xiu, L. and Chau, D. C. K. (2009) 'Application of data mining techniques in customer relationship management: A literature review and classification', *Expert Systems with Applications*. Elsevier Ltd, 36(2), pp. 2592–2602. doi: 10.1016/j.eswa.2008.02.021.

[33] Paker, N. and Vural, C. A. (2016) 'Customer segmentation for marinas: Evaluating marinas as destinations', *Tourism Management*, 56, pp. 156–171. doi: 10.1016/j.tourman.2016.03.024.

[34] Parvatiyar, A. and Sheth, J. N. (2001) 'Customer Relationship Management: Emerging Practice, Process, and Discipline', *Journal of Economic and Social Research*, 3(2), pp. 1–34. doi: 10.1007/s002280050537.

[35] Rajagopal, S. (2011) 'Customer Data Clustering Using Data Mining', *Database Management Systems*, 3(4), pp. 1–11. doi: 10.5121/ijdms.2011.3401.

[36] REN, H., ZHENG, Y. and WU, Y. rong (2009) 'Clustering analysis of telecommunication customers', *Journal of China Universities of Posts and Telecommunications*. The Journal of China Universities of Posts and Telecommunications, 16(2), p. 114–116,128. doi: 10.1016/S1005-8885(08)60214-9.

[37] Sharma, N., Bajpai, A. and Litoriya, R. (2012) 'Comparison the various clustering algorithms of weka tools', *International Journal of Emerging Technology and Advanced Engineering*, 2(5), pp. 73–80.

[38] Sheshasaayee, A. (2017) 'An Efficiency Analysis on the TPA Clustering Methods for Intelligent Customer Segmentation Dr.', in *IICIMIA 2017*, pp. 784–788.

[39] Swift, R. S. (2001) *Accelerating customer relationships: Using CRM and relationship technologies*. Prentice Hall Professional.

[40] Thanuja, V., Venkateswarlu, B. and Anjaneyulu, G. S. G. N. (2011) 'Applications of Data Mining in Customer Relationship Management', 2(3), pp. 423–433.

[41] Tsai, C. F., Hu, Y. H. and Lu, Y. H. (2015) 'Customer segmentation issues and strategies for an automobile dealership with two clustering techniques', *Expert Systems*, 32(1), pp. 65–76. doi: 10.1111/exsy.12056.

[42] Wang, C. H. (2009) 'Outlier identification and market segmentation using kernel-based clustering techniques', *Expert Systems with Applications*. Elsevier Ltd, 36(2 PART 2), pp. 3744–3750. doi: 10.1016/j.eswa.2008.02.037.

[43] Wang, Y.-J. (2010) 'A clustering method based on fuzzy equivalence relation for customer relationship management', *Expert Systems with Applications*. Elsevier Ltd, 37(9), pp. 6421–6428. doi: 10.1016/j.eswa.2010.02.076.

[44] Wang, Y., Ma, X., Lao, Y. and Wang, Y. (2014) 'A fuzzy-based customer clustering approach with hierarchical structure for logistics network optimization', *Expert Systems*

with Applications. Elsevier Ltd, 41(2), pp. 521–534. doi: 10.1016/j.eswa.2013.07.078.

[45] Witten, I. H., Frank, E., Hall, M. A. and Pal, C. J. (2016) *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[46] Xu, D. and Tian, Y. (2015) 'A Comprehensive Survey of Clustering Algorithms', *Annals of Data Science*. Springer Berlin Heidelberg, 2(2), pp. 165–193. doi: 10.1007/s40745-015-0040-1.

[47] Xu, R. (2005) 'Survey of clustering algorithms', *IEEE Transactions on Neural Networks*, 16(3), pp. 645–678. doi: 10.1109/TNN.2005.845141.

[48] Yim, F. H., Anderson, R. E. and Swaminathan, S. (2004) 'Customer Relationship Management: Its Dimensions and Effect on Customer Outcomes', *Journal of Personal Selling & Sales Management*, 24(4), pp. 263–278. doi: 10.1080/08853134.2004.10749037.

[49] Zhou, K., Yang, S. and Shao, Z. (2017) 'Household monthly electricity consumption pattern mining: A fuzzy clustering-based model and a case study', *Journal of Cleaner Production*. Elsevier Ltd, 141, pp. 900–908. doi: 10.1016/j.jclepro.2016.09.165.

PRILOG 1

Tablica 1 Rezultati analize 27 članaka koji obrađuju problematiku klasteriranja u dimenziji identifikacija klijenata

Cilj identifikacije	Poslovno područje	Tip podataka	Netipične vrijednosti	Broj klastera	Višedim.	Implementacija	Objašnjenje	pretprocesiranje	Članak
bolja alokacija resursa	proizvodnja	kvantitativni (statički)	velika osjetljivost	predefiniran od strane eksperata	veoma osjetljiv	jednostavna	jednostavno	nije provedeno pretprocesiranje	(Ho *et al.*, 2012)
segmentacija u cilju definiranja poslovnih strategija identifikacija zbog povećanja prodaje + zadržavanja postojećih, te privlačenja novih klijenata	maloprodaja	kvantitativni (statički)	nije poznato	predefiniran od strane eksperata uz provedenu validaciju	nije osjetljiv	jednostavna	jednostavno	pretprocesirani podaci	(Rajagopal, 2011)
	bankarstvo	mješoviti (statički)	velika osjetljivost	estimiran iz podataka bez validacije	veoma osjetljiv	jednostavna	jednostavno	pretprocesirani podaci	(Sheshasayee, 2017)
bolja alokacija resursa zdravstvenih usluga	zdravstvo	kvantitativni (statički)	velika osjetljivost	estimiran iz podataka bez validacije	veoma osjetljiv	nije navedeno	jednostavno	pretprocesirani podaci	(Chen *et al.*, 2012)
određivanje ponašanja klijenata za odabir prikladne marketinške strategije	proizvodnja	kvantitativni (statički)	velika osjetljivost	estimiran iz podataka uz validaciju	veoma osjetljiv	jednostavna	jednostavno	pretprocesirani podaci	(Li, Dai and Tseng, 2011)
generička segmentacija tržišta	tomobilska industr	kvantitativni (statički)	malo osjetljivost	estimiran iz podataka bez validacije	veoma osjetljiv	složena	jednostavno	pretprocesirani podaci	(Wang, 2009)
generička segmentacija tržišta	telekomnikacije	kvantitativni (statički)	nije osjetljiv	predefiniran od strane eksperata	malo osjetljiv	složena	jednostavno	pretprocesirani podaci	(REN, ZHENG and WU, 2009)
segmentacija tržišta u cilju određivanja marketinških strategija	tomobilska industr	mješoviti (statički)	velika osjetljivost	predefiniran od strane eksperata uz provedenu validaciju	veoma osjetljiv	umjerena	jednostavno	pretprocesirani podaci + PCA	(Tzai, Hu and Lu, 2015)
generička segmentacija tržišta	turizam	mješoviti (statički)	velika osjetljivost	estimiran iz podataka bez validacije	veoma osjetljiv	jednostavna	jednostavno	nije navedeno	(Dursun and Caber, 2016)
generička segmentacija tržišta	maloprodaja	mješoviti (statički + dinamički)	mala osjetljivost	estimiran iz podataka uz validaciju	nije osjetljiv	jednostavna	jednostavno	nije navedeno	(Liu and Chen, 2017)
optimizacija resursa	energetika	kvantitativni (statički)	mala osjetljivost	predefiniran od strane eksperata uz provedenu validaciju	osjetljiv	jednostavna	jednostavno	pretprocesirani podaci + selekcija atributa	(Biscani *et al.*, 2017)
segmentacija tržišta u cilju podrške odlučivanju u predviđanju i planiranju	maloprodaja	kvantitativni (statički)	velika osjetljivost	estimiran iz podataka uz validaciju	veoma osjetljiv	jednostavna	jednostavno	pretprocesirani podaci	(Kashwan and Velu, 2013)
generička segmentacija tržišta	poštanska poduzeća	mješoviti statički	nije poznato	predefiniran od strane eksperata	nije osjetljiv	umjerena	jednostavno	nije poznato	(Kolarevski, Tengler and Majerčáková, 2016)
segmentacija tržišta u cilju bolje marketinške strategije	telekomunikacije	kvantitivni (statički)	velika osjetljivost	predefiniran od strane eksperata	veoma osjetljiv	jednostavna	jednostavno	pretprocesirani podaci	(Luo *et al.*, 2013)
generička segmentacija tržišta	proizvodnja	kvantitativni statički	umjerena osjetljivost	estimiran iz podataka bez validacije	veoma osjetljiv	jednostavna	jednostavno	nije navedeno	(Ansari and Riasi, 2016)
segmentacija u cilju predviđanja potražnje	logistika	kvantitativni statički	velika osjetljivost	estimiran iz podataka uz validaciju	veoma osjetljiv	jednostavna	jednostavno	pretprocesirani podaci	(Murray, Agard and Barajas, 2015)
segmentacija tržišta u cilju određivanja životnog cila	maloprodaja	kvantitativni statički	mala osjetljivost	estimiran iz podataka bez validacije	malo osjetljiv	jednostavna	jednostavno	pretprocesirani podaci	(Migueis, Camanho and Falcão E Cunha, 2012)
segmentacija tržišta (marketinške aktivnosti i optimizacija proizvodnje)	modna industrija	mješoviti (statički)	mala osjetljivost	nije detaljno opisan postupak određivanja broja klastera	veoma osjetljiv	jednostavna	jednostavno	nije navedeno	(Brito *et al.*, 2015)
segmentacija zbog optimizacije	logistika	mješoviti (statički)	nije poznato	estimiran iz podataka, potvrđen od strane eksperata i provedena validacija	malo osjetljiv	umjerena	jednostavno	pretprocesirani podaci	(Wang *et al.*, 2014)
generička segmentacija tržišta	bankarstvo	mješoviti (statički)	nije poznato	estimiran iz podataka uz validaciju	nije navedeno	umjerena	jednostavno	pretprocesirani podaci	(Wang, 2010)
segmentacija u cilju razvoja bolje marketinške strategije	maloprodaja	mješoviti (statički)	nije poznato	estimiran iz podataka uz validaciju	nije navedeno	nije navedeno	jednostavno	pretprocesirani podaci	(Gupta, Aggarwal and Rana, 2016)
segmentacija u cilju estimacije buduće vrijednosti klijenata	bankarstvo	kvantitativni statički	velika osjetljivost	estimiran iz podataka uz validaciju	veoma osjetljiv	k-means	jednostavno	pretprocesirani podaci	(Khajvand and Tarokh, 2011)
generička segmentacija tržišta	telekomunikacije	mješoviti statički	mala osjetljivost	estimiran iz podataka, potvrđen od strane eksperata i provedena validacija	osjetljiv	LCA (Latent Class Analysis) Latent Gold 4.5 software	jednostavno	pretprocesirani podaci	(Hanka *et al.*, 2014)
generička segmentacija tržišta	turizam	mješoviti	velika osjetljivost	estimiran iz podataka, potvrđen od strane eksperata i provedena validacija	veoma osjetljiv	Hijerarhijski + K-means	jednostavno	pretprocesirani podaci	(Paker and Vural, 2016)
segmentacija u cilju kreiranja nove usluge	energetika	kvantitativni statički	velika osjetljivost	estimiran iz podataka uz validaciju	veoma osjetljiv	Hopfield-K-means	jednostavno	pretprocesirani podaci	(López *et al.*, 2011)
segmentacija (optimizacija potrošnje i razvoj fleksibilnih marketinških strategija)	energetika	kvantitativni statički	velika osjetljivost	estimiran iz podataka uz validaciju	veoma osjetljiv	Improved Fuzzy C-means (Matlab package)	jednostavno	pretprocesirani podaci	(Zhou, Yang and Shao, 2017)
generička segmentacija tržišta	turizam	mješoviti (statički)	mala osjetljivost	estimiran iz podataka uz validaciju	osjetljiv	umjerena	jednostavno	pretprocesirani podaci	(D'Urso *et al.*, 2013)