

PRIMENA FAZI KLASTEROVANJA ZA KATEGORISANJE TENISKIH IGRAČA

CATEGORIZING TENNIS PLAYERS USING FUZZY CLUSTERING TECHNIQUE

LUKA TASIĆ¹, ANA POLEDICA², MILICA ZUKANOVIĆ³

¹ Connect The Dots, tasiicluka123@gmail.com

² Univerzitet u Beogradu – Fakultet organizacionih nauka, ana.poledica@fon.bg.ac.rs

³ Univerzitet u Beogradu – Fakultet organizacionih nauka, zukanovic.milica@gmail.com

Rezime: *Primena tehnika računarske inteligencije je sve više zastupljena u sportskoj analitici. Prilikom analize teniskih mečeva i igrača, postoji mnogo faktora koji se mogu razmatrati kako bi se utvrdilo koji će teniseri biti uspešniji. Osnovne statistike kao što su procenat prvog servisa, procenat osvojenih brejk lopti, procenat drugog servisa i druge daju uvid u stil igre i prednosti svakog igrača. Osnovni cilj ovog rada je da se na osnovu podataka o teniskim mečevima od 2000. do 2022. godine grupišu teniseri, utvrde ključne karakteristike tenisera i identifikuju klasteri prema stilu igre. Primenjen je algoritam fazi C-srednjih vrednosti koji je uspešno izdvojio 3 karakteristične grupe tenisera. Takođe, pokazalo se da postoje igrači koji prirodno imaju značajnu pripadnost više klastera koji su posebno razmatrani.*

Ključne reči: *fazi klasterovanje, fazi C-srednjih vrednosti, kategorisanje, tenis*

Abstract: *The application of computer intelligence techniques is increasingly present in sports analytics. When analyzing tennis matches and players, there are many factors that can be considered in order to determine which tennis players will be more successful. Basic statistics such as the percentage of the first serve, the percentage of break points won, the percentage of the second serve and others give an insight into the style of play and the advantages of each player. The main goal of this paper is to group tennis players based on data of tennis matches from 2000 to 2022, determine the key characteristics of tennis players and identify clusters according to the style of play. The fuzzy C-mean algorithm was applied, which successfully singled out 3 characteristic groups of tennis players. Also, it has been shown that there are players who naturally have a significant affiliation to multiple clusters that have been specifically considered.*

Keywords: *fuzzy clustering, fuzzy C-means, categorizing, tennis*

1. UVOD

Poslednjih godina metode računarske inteligencije i analize podataka postale su značajni alati u sportskoj analitici i zauzimaju značajnu ulogu kod sve većeg broja različitih sportskih disciplina npr. [3,6].

I u tenisu napredna analitika dobija sve veći značaj. Postoji veliki broj faktora koji mogu uticati na sam ishod meča na šta ukazuju i mnoge studije koje su se time bavile:

- Najbolji profesionalni rangovi igrača rođenih 1982. ili ranije bili su pozitivno povezani sa uzrastom u kojem su igrači zaradili svoj prvi ATP poen, a zatim ušli u prvih 100, što sugeriše da godine povezane sa ovim prekretnicama u rangiranju mogu imati određeni potencijal za predviđanje [14].
- Sila tokom teniskog udarca forhendom korelira sa veličinom hvata [12]. Autori su takođe otkrili korelaciju pola i iskustva sa udarnom silom.
- Tenis na travi zahteva specifične veštine za postizanje visokog ranga u svetu, kao što su veštine servisa i vraćanja servisa. Servis je najvažnija veština u tenisu i važniji je kod muškaraca i uspešniji je u dublu [7].
- Postoji veza između količine i nivoa konkurencije i kognitivna stručnosti nivo takmičenja mogu biti od ključnog značaja za razvoj stručnosti tenisera [8].

- U radu [17] su ispitani atributi i performanse 100 najboljih tenisera i izvršena je klaster analiza. Jedan klaster su igrači koji preuzimaju rizik, a drugi su igrači koji izbegavaju rizik. *K*-srednjih vrednosti (eng. *k*-Means) tehnika grupisanja je korišćena da bi se razumeli fizički atributi igrača tokom igre. Razmatrajući klastere otkriveno je da su duple greške i asovi važne ulazne varijable u predviđanju rizične prirode sportista na teniskoj travi.

Ovaj rad je nastao kao deo šireg istraživanja koje se bavi poređenjem različitih algoritama klasterovanja pri kategorizaciji teniskih igrača. Jedan od najčešćih algoritama koji se primenjuje u kategorizaciji sportista je algoritam *k*-srednjih vrednosti i uspešno je primenjen u klasterovanju igrača u košarci [15], američkom fudbalu [10], tenisu [16] itd. Nakon polaznih rezultata istraživanja za algoritam *k*-srednjih vrednosti, uvideli smo da postoji više karakterističnih igrača koji ne pripadaju strogo jednom klasteru, već poseduju karakteristike tenisera drugog klastera. Iako su tradicionalne tehnike tvrdog klasterovanja (eng. hard clustering) uspešne kada se klasteri mogu jasno razdvojiti i definisati, manje su pogodne u situacijama kada ne se ne mogu uspostaviti jasne granice i kada nam je važno da posebno pažnju posvetimo graničnim slučajevima. Tada je pogodnije koristiti metode mekog klasterovanja (eng. soft clustering) gde objekat može pripadati više klastera istovremeno.

U okviru ovog rada primenjen je algoritam fazi C-srednjih vrednosti, kako bismo obezbedili da za svaku instancu postoji nekoliko deskriptora pripadnosti klasterima. Podaci koji su korišćeni za analizu preuzeti su sa sajta Tennis Abstract¹ koji održava teniski fan Džef Sakman gde su javno dostupne statistike teniskih mečeva koje datiraju od 1968. godine. Za analizu posmatran je period od 2000. do 2022. godine. U okviru istraživanja analizirane su osnovne statistike koje daju uvid u stil igre i relativne prednosti svakog igrača kao što su procenat prvog servisa, procenat osvojenih poena na prvi servis, procenat osvojenih poena na drugi servis i druge. Skup podataka ne sadrži metrike nadigravanja kao što je dužina nadigravanja, da li je poen direktan (viner), osvojen iznuđenom, neiznuđenom greškom ili na mreži. Osnovna pretpostavka je da se primenom algoritma fazi C srednjih vrednosti mogu grupisati teniseri i pronaći obrasci različitih stilova igre koji karakterišu svaki klaster.

U nastavku rada, u okviru poglavlja 2 data je teorijska osnova algoritma fazi C-srednjih vrednosti. Zatim su opisani podaci i odabrani atributi koji su korišćeni kao ulaz u klasterovanje. U poglavlju 4 prikazani su rezultati fazi klasterovanja i analizirani dobijeni klasteri. Na kraju dat je zaključak istraživanja i predlozi za dalji rad.

2. KLASTEROVANJE METODOM FAZI C-SREDNJIH VREDNOSTI

Klasterovanje je metod za grupisanje opservacija na osnovu njihove sličnosti. Klaster analiza je značajna za mnoštvo oblasti poput prepoznavanja paterni, analize podataka, obrade slika i istraživanja tržišta [9]. Uopšteno govoreći, klasterovanje se može podeliti u dve podgrupe, tvrdo klasterovanje i meko klasterovanje [13]. Algoritmi tvrdog klasterovanja, svaku opservaciju dodeljuju samo jednom klasteru i jedan od široko korišćenih jeste algoritam K-srednjih vrednosti [2].

Od nedavno, meko klasterovanje pokazalo se kao superiornije u odnosu na tradicionalne tvrde algoritme za klasterovanje [18]. Kod mekog klasterovanja, samim tim i kod algoritma fazi C-srednjih vrednosti (eng. Fuzzy C-Means, FMC), svaka opservacija pripada svakom klasteru, iako su funkcije pripadnosti drugačije. Nivo pripadnosti kod pomenutih algoritma zavisi od blizine opservacije od centroida, odnosno od centra klastera. Što je opservacija bliža centroidu, veća je vrednost pripadanja i obrnuto. Svrha ovakvog načina klasterovanja je minimizacija funkcije cilja kako bi se dokazala sličnost između opservacija u klasteru i smanjila sličnost između opservacija u poređenju sa drugim klasterima [1]. Funkcija cilja može da se definiše na sledeći način:

$$J = \sum_{i=1}^n \sum_{j=1}^c \mu_{ij}^m \|x_i - r_j\|^2 \quad (1)$$

gde je n broj instanci, c broj klastera, x_i i -ta opservacija, r_j j -ti centroid, μ_{ij} funkcija pripadnosti i -te opservacije j -tom centroidu, gde funkcija ima vrednost od 0 do 1, uključujući te vrednosti za svako $i=1,2,\dots,n$ i svako $j=1,2,\dots,c$, m fazi stepen i $\|x_i - r_j\|$ Euklidsko rastojanje između i -te opservacije od j -tog centroida [5].

Prednosti fazi grupisanja u odnosu na klasične metode grupisanja se ogledaju u tome što pruža detaljnije informacije o podacima, kao i vrednosti funkcija pripadnosti koje su pogodne za komentarisanje, fleksibilan je u korišćenju odstojanja i prilagodljiv je numeričkoj optimizaciji [11]. S druge strane, ima i nedostatke. U

¹ www.tennisabstract.com

slučaju velikog broja opservacija i klastera, teško je sumirati i klasifikovati podatke prilikom mnogobrojnih izlaza. Štaviše, algoritmi fazi klasterovanja, koji se koriste kada postoji nesigurnost, generalno su komplikovani [4].

3. OPIS PROBLEMA I PODATAKA

Osnovni motiv ovog rada je primena tehnike fazi klasterovanja kako bi se izvršilo identifikovanje grupa teniskih igrača i analizirale njihove performanse i osobine. Izučavani problem ovog rada je analiza statistika teniskih mečeva kako bi se dalje pronašli obrasci različitih stilova igre za svaku grupu. Klasterovanje metodom fazi *C*-srednjih vrednosti je rađeno sa ciljem da se izdvoje najbolji teniseri, kako bi se utvrdilo koje od karakteristika su najbitnije za uspeh u tenisu. Takođe, posebna pažnja će biti posvećena graničnim slučajevima, odnosno igračima koji sa velikim stepenom pripadaju više klastera što je jedna od prednosti FCM-a.

Na osnovu preuzetih podataka² o teniskim mečevima odigranih od januara 2000. do marta 2022. godine, gde svaki red predstavlja jedan meč i uključuje statistike o pobedniku i gubitniku, napravljen je novi skup podataka za potrebe fazi klasterovanja. Izvršeno je filtriranje, prvo teniskih turnira, a potom i samih igrača. U obzir su uzeti samo turniri koji se igraju na tri dobijena seta, igrači koji imaju najmanje 50 mečeva na pomenutim turnirima i koji su ikad bili u prvih 50 pozicija na ATP listi. Usled promene pozicija na ATP listi, u dobijenom skupu podataka analizirana su 133 teniska igrača koji zadovoljavaju prethodne kriterijume.

Odabrani su atributi koji se smatraju osnovnim statistikama u teniskoj analitici:

- procenat osvojenih poena na prvi servis (A1)
- procenat osvojenih poena na drugi servis (A2)
- procenat as servisa (A3)
- procenat duplih grešaka (A4)
- prosečan broj gemova po meču (A5)
- procenat dobijenih tajbrejkova (A6)
- procenat spašenih brejk lopti (A7)
- procenat osvojenih brejkova (A8)

4. EKSPERIMENT I REZULTATI

Nakon kreiranja skupa podataka, sledeća faza obuhvatala je detaljniju analizu podataka i pripremu za fazi klasterovanje. U prvom koraku izvršena je standardna min-max normalizacija, kao i analiza nestandardnih opservacija koje su zadržane u skupu zbog poznatog porekla i značajnih informacija. Zasad detaljnijeg razumevanja podataka, izračunata je korelaciona matrica između atributa. Očekivano, procenat osvojenih poena na prvi servis (A1) je veoma visoko korelisan sa atributima: procenat asova (A3) (0.89) i procenat spašenih brejk lopti (A7) (0.7). Imajući u vidu da je procenat prvog servisa veoma bitan faktor u igri tenisera, analizirana su dva modela, sa i bez atributa koji opisuje procenat osvojenih poena na prvi servis (A1). Utvrđeno je da se izbacivanjem promenljive (A1) gube informacije, odnosno da je dobijena bolja struktura klastera sa uključenim svim atributima. Za određivanje optimalnog broja klastera korišćene su lakat metoda (eng. elbow method) i metoda siluete (eng. silhouette method). Primenom lakat metode utvrđeno je da je optimalan broj klastera tri, dok je metoda silueta pokazala dva klastera kao optimalan broj. Nakon analize, tri klastera su se pokazala kao odgovarajuća.

Klasterovanje je izvršeno metodom fazi *C*-srednjih vrednosti, dok je za određivanje distance primenjeno standardno Euklidsko odstojanje. Za potrebe ovog eksperimenta korišćen je programski jezik R, odnosno R studio i biblioteka *fcm*. S obzirom na relativno mali broj ulaznih atributa i opservacija, vremenska i prostorna ograničenja algoritma nisu bila kritična. Korišćene su podrazumevane vrednosti: parameter *m* je 2, maksimalan broj iteracija je 1000, kriterijum zaustavljanja je 10^{-9} , itd.

² https://github.com/JeffSackmann/tennis_atp

Sumirane vrednosti FCM-a, uključujući i promenljivu A1 date su u Tabeli 1. Pripadnost tenisera klasteru određena je na osnovu najveće vrednosti pripadnosti. Jaka pripadnost klasteru je definisana u slučaju kada je pripadnost veća ili jednaka 0.5. Centroidi klastera dati su u Tabeli 2. Slika klastera prikazana je na Slici 1.

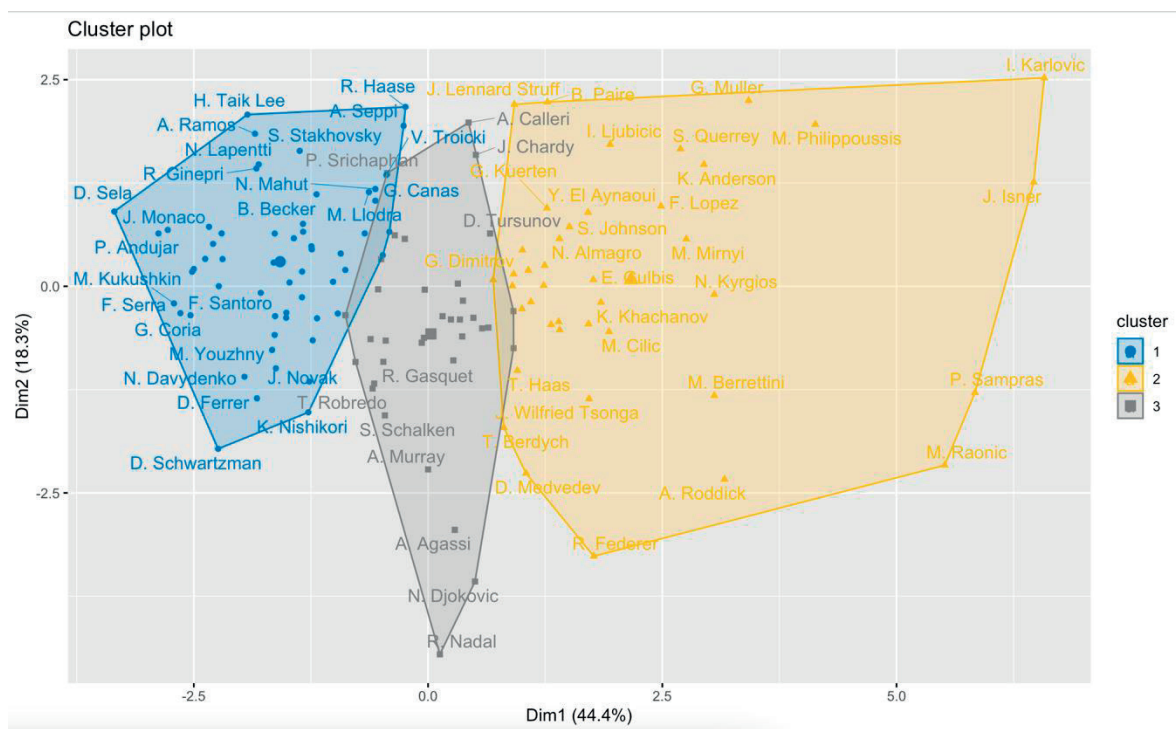
Tabela 1: Rezultati klasterovanja metodom FCM-a – metrike klastera

Metrike	Klaster 1	Klaster 2	Klaster 3
Broj tenisera	57	41	35
Najveća pripadnost	0.645	0.776	0.584
Najmanja pripadnost	0.389	0.395	0.359
Prosečna pripadnost	0.516	0.537	0.455

Tabela 2: Rezultati klasterovanja metodom FCM-a – centroidi klastera

Promenljive	A1	A2	A3	A4	A5	A6	A7	A8
Klaster 1	0.328	0.319	0.194	0.254	0.330	0.455	0.335	0.673
Klaster 2	0.629	0.476	0.433	0.312	0.474	0.558	0.555	0.548
Klaster 3	0.457	0.399	0.273	0.287	0.368	0.509	0.409	0.666

Na osnovu analize rezultata klasterovanja može se izneti nekoliko zaključaka. Pripadnike prvog klastera odlikuju loše statistike vezane za servis (A1, A2, A3) i nizak procenat spašenih brejk lopti (A7), dok je procenat osvojenih brejk lopti visok (A8), kao što su Švarcman, Ferer, Davidenko, Nišikori. Tenisere iz trećeg klastera, suprotno teniserima iz prvog klastera, odlikuju visoki procenti kada su u pitanju atributi vezani za servis (A1, A2, A3), samim tim i visok procenat osvojenih gemova po meču (A5), visok procenat spašenih brejk lopti (A7) i viši procenat duplih grešaka (A4), dok im je procenat osvojenih brejk lopti (A8) znatno lošiji poput Karlovića, Iznera, Raonića, Samprasa, Federera, itd. Pripadnike drugog klastera odlikuju visoke vrednosti kada su u pitanju procenat osvojenih poena na drugi servis (A2), osvojenih tajbrejkova (A6), sačuvanih brejk lopti (A7) i osvojenih brejk lopti (A8).



Slika 1: Prikaz klastera

Nakon detaljnije analize predstavnika klastera, Švarcman (klaster 1 – „Never surrenders“), Đokovića i Nadala (klaster 2 – „Mental rocks“) i Karlovića (klaster 3 – „Top servers“), izvedeno je nekoliko zaključaka. Teniseri iz prvog klastera su uglavnom ljudi nižeg rasta, u odnosu na tenisere iz drugog i trećeg klastera, koji čine mahom visoki teniseri. U skladu sa tim, teniseri iz prvog klastera gube dosta gemova na svoj servis, ali isto tako i prave dosta brejkova, suprotno teniserima iz trećeg klastera, koji svoju igru baziraju na servisu. Pripadnici drugog klastera, u poređenju sa ostala dva klastera, uzimaju najviše poena na svoj drugi servis,

procenat poena na prvi servis je manji od tenisera iz trećeg, ali veći od tenisera iz prvog klastera. Neki od ključnih momenata meča jesu trenuci kada se osvajaju ili brane brejk lopte, kao i igranje samih tajbrejkova. U skladu sa tim, može se zaključiti da teniseri iz drugog klastera imaju dobre ocene kod glavnih atributa za dobijanje meča. Kada pogledamo činjenicu da drugom klasteru pripadaju Đoković, teniser sa najviše nedelja na prvoj poziciji svetske ATP liste i Nadal, teniser sa najviše grend slem titula, potvrđujemo ovu grupu klastera.

Na slici klastera možemo uočiti i neke granične slučajeve. Neki od njih su na primer Grigor Dimitrov, teniser koji se nalazi na granici između drugog i trećeg klastera i Mikael Ljodra koji se nalazi na granici između prvog i drugog klastera. Dimitrov ima visok procenat osvojenih poena na prvi servis (A1) i visok procenat spašenih brejk lopti (A7), koje su odlike trećeg klastera. Dimitrov takođe ima i visok procenat osvojenih poena na drugi servis (A2) što je najviše odlika drugog klastera, dok su mu vrednosti ostalih atributa vezanih za drugi klaster niske. Mikael Ljodra takođe ima visok procenat osvojenih poena na drugi servis (A2) kao odliku drugog klastera, a kao odliku prvog klastera ima visok procenat osvojenih brejk lopti (A8).

U analiziranje je uključen i Rodžer Federer, kao preostali član velike trojke u svetu tenisa. Federer pripada trećem klasteru, ali je jako blizu drugog klastera (Tabela 3). Njegove statistike za servis su dobre, ali isto tako i ocene atributa koje karakterišu pripadnike drugog klastera. Imajući to u vidu, jasno je zašto je smatran za jednog od najvećih tenisera svih vremena. Pripadnost klasterima velike trojke prikazane su u Tabeli 3.

Tabela 3: Pripadnost klasterima Đokovića, Nadala i Federera

Pripadnosti klasteru	Klaster 1	Klaster 2	Klaster 3
R. Federer	0.230	0.333	0.436
R. Nadal	0.278	0.382	0.340
N. Đoković	0.309	0.369	0.322

5. ZAKLJUČAK

Osnovni cilj ovog rada bio je da odgovori na pitanja: kako grupisati tenisere i koji atributi su najznačajniji za uspešnost tenisera. Na osnovu statistika iz mečeva, primenom algoritma fazi C-srednjih vrednosti izvršeno je klasterovanje 133 teniska igrača. Kao rezultat dobijena su tri klastera koji su u nastavku identifikovani prema stilu igre: „Never surrenders“, „Mental rocks“ i „Top servers“. Time je primarni cilj istraživanja postignut.

Na osnovu pojedinih svetski poznatih kriterijuma, kao što su broj nedelja na prvom mestu i broj grend slem titula, izdvojeni su najbolji i na osnovu njih upoređene bitnosti i karakteristike klastera. U skladu sa informacijama koje smo dobili, može se primetiti da atributi koji karakterišu pripadnike drugog, pa onda i atributi trećeg klastera u kombinaciji sa atributima iz drugog klastera, mogu uzeti za najbitnije u teniskom svetu, jer se u njima nalaze najbolji svih vremena. Zahvaljujući algoritmu fazi klasterovanja, dobijeni su rezultati tenisera koji imaju značajnu pripadnost u više klastera i vrednosti atributa koji ih opisuju su posebno zanimljive za posmatranje.

Mogućnost za unapređenje istraživanja se ogleda u dodavanju atributa koje bi se postiglo uz širenje skupa podataka sa dodatnim statistikama kao što su broj direktnih poena iz forhenda i bekhenda, prosečno trajanje poena, broj grešaka i slično. Takođe, moglo bi se izvršiti poređenje u zavisnosti od podloge na kojoj se igra neki turnir, a pored toga dalji nastavak istraživanja može ići i u smeru primene drugih algoritama klasterovanja.

LITERATURA

- [1] Askari, S. (2021). Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development. *Expert Systems with Applications*, 165, 113856.
- [2] Bezdek, J. C. (2017). A primer on cluster analysis: 4 basic methods that (usually) work. First Edition Design Publishing.
- [3] Bonidia, R. P., Rodrigues, L. A., Avila-Santos, A. P., Sanches, D. S., & Brancher, J. D. (2018). Computational intelligence in sports: A systematic literature review. *Advances in Human-Computer Interaction*, 2018.
- [4] De Oliveira, J. V., & Pedrycz, W. (Eds.). (2007). *Advances in fuzzy clustering and its applications*. John Wiley & Sons.
- [5] Erilli, N. A., Öner, Y., Alakuş, K., & Tunç, T. (2010). Classifying Chess Players with Fuzzy Clustering Analysis in Fuzzy Data Using Eco Codes. In *1st International Symposium on Computing in Science & Engineering Proceedings Book* (pp. 1105-1110).

- [6] Fister Jr, I., Ljubič, K., Suganthan, P. N., Perc, M., & Fister, I. (2015). Computational intelligence in sports: challenges and opportunities within a new research domain. *Applied Mathematics and Computation*, 262, 178-186.
- [7] Furlong, J.D.G. (1995). The service in lawn tennis: how important is it?
- [8] Garcia-Gonzalez, L., Diminguez, A.M., Gil, A. & Moreno, M.P. (2014). Effect of decisions making and performance young tennis players: an applied research
- [9] Govindasamy, K., & Velmurugan, T. (2018). Analysis of student academic performance using clustering techniques. *International Journal of Pure and Applied Mathematics*, 119(15), 309-323.
- [10] Murray, N. P., & Hunfalvay, M. (2017). A comparison of visual search strategies of elite and non-elite tennis players through cluster analysis. *Journal of sports sciences*, 35(3), 241-246.
- [11] Næs, T., & Mevik, B. H. (1999). The flexibility of fuzzy clustering illustrated by examples. *Journal of Chemometrics*, 13(3-4), 435-444.
- [12] Ohguni, M., Aoki, M., Sato, H. & Imada, K. (2009). The effect of grip size on the hitting force during a soft tennis forehand stroke.
- [13] Ramamoorthy, V. (2019). Fuzzy C-Mean Clustering Using Data Mining. BookRix: Munich, Germany.
- [14] Reid, M. & Morris, C. (2013). ranking benchmarks of top 100 players in men's professional tennis.
- [15] Sampaio, J., McGarry, T., Calleja-González, J., Jiménez Sáiz, S., Schelling i del Alcázar, X., & Balciunas, M. (2015). Exploring game performance in the National Basketball Association using player tracking data. *PloS one*, 10(7), e0132894.
- [16] Shelly, Z., Burch, R. F., Tian, W., Strawderman, L., Piroli, A., & Bichey, C. (2020). Using K-means clustering to create training groups for elite American football student-athletes based on game demands. *International Journal of Kinesiology and Sports Science*, 8(2), 47-63.
- [17] Syed, A. M. (2019). Performance Attributes and Risk Taking of Players-Evidence from Lawn Tennis. *Journal of Sport and Games 2019*, 1(3), 13-20.
- [18] Vučičević, A., Milošević, P., Poledica, A. (2021). Selecting Case Study members using fuzzy C-means algorithm. *Proc. XLVIII International Symposium on Operational Research, SYMOPIS 2021*. Banja Koviljača, Sept. 20-23, 2021.