Klasterovanje FIFA 19 igrača

Seminarski rad u okviru kursa
Istraživanje podataka
Matematički fakultet

Jovan Dmitrović
jdmitrovic@gmail.com

18. jun 2019.

Sažetak

U ovom radu biće prikazana klasterovanje nad podacima iz video igre FIFA 19. Nad datim skupom podataka će biti izvršeno predprocesiranje, a zatim će biti prikazano korišćenje algoritma K-sredina, OPTICS, Kohonenovog algoritma, kao i korišćenje hijerarhijskog klasterovanja.
Ključne reči: *klasterovanje, fudbal, istraživanje podataka*

Sadržaj

1 Uvod

Serijal FIFA video igara je jedan od najdugovečnijih serijala sportskih igara osnovan od strane američke kompanije EA Sports. Od 1993. godine pa do danas, serijal FIFA je dobijao barem jednog novog člana godišnje; sa svakom novom FIFA igrom, rastao je i broj liga i timova implementiranih u igri, da bi se u poslednjoj iteraciji, FIFA-i 19, našlo preko 700 klubova iz preko 30 liga širom sveta.

U samoj igri je prisutan i veliki broj fudbalera, pri čemu svaki od njih ima svoje karakteristike, što je velika količina podataka. Cilj ovog rada je da pokuša da klasteruje igrače i da izvede zaključke o prepoznatim klasterima.

2 Podaci

Korišćeni podaci se mogu naći na sledećoj veb-strani. U igri postoji ukupno 18207 fudbalera sa po 89 atributa. Pored meta-atributa igrača kao što su nacionalnost ili klub za koji fudbaler nastupa, među atributima se nalaze i atributi vezani za sposobnosti igrača u igri (npr. brzina, jačina šuta i slično); takvi atributi imaju opseg $[1, 99]$.

2.1 Predprocesiranje podataka

Iako SPSS modeler obavlja veliki deo posla automatski, neki od atributa nisu u pogodnom obliku za rad nad njima:

- **Visina** - visina je navedena vodeći se imperijalnim sistemom, navodeći stope i inče. Ovaj način zapisa je nezgodan, pa je poželjno prebaciti podatke u metrički sistem, tj. u centimetre.
- **Masa** - masa je izražena u funtama. Zarad konzistentnosti, masa je prevedena u kilograme.
- **Vrednost igrača, plata i otkupna klauzula** - finansijski podaci su izraženi pomoću skraćenica, te su oni prevedeni u konkretne cele brojeve.
- **Ocene igrača po pozicijama** - podaci su uneti u obliku zbira $x+y$, pa su ti podaci sabrani.

Postoje i neki nominalni atributi koji neće biti uzeti u obzir kao što su: fotografija fudbalera, identifikacioni broj, zastava...

2.1.1 Predprocesiranje podataka u programskom jeziku Python

Kada je reč o primenjivanju klasterovanja u programskom jeziku *Python*, pored transformacija podataka pomenutih u 2.1, neophodno je izvršiti neke tehnike predprocesiranja ručno, za razliku od programa *(*SPSS Modeler). Ovo se u najvećoj meri odnosi na standardizaciju podataka, čiji je zadatak da sve numeričke atribute preslika u opseg $[0, 1]$. Standardizacija se u Python-u vrši uz pomoć `MinMaxScaler` funkcija iz paketa `sklearn.preprocessing`.

Kako se korišćenjem paketa `pandas` kategorički atributi ne prepoznaju automatski, neophodno ih je ručno transformisati u kategoričke podatke sa korektno označenim kategorijama. Nedostajuće vrednosti su zamenjene srednjom vrednošću atributa.

3 Klasterovanje

Za klasterovanje ovog skupa podataka biće upotrebljeno četiri algoritma:

- Algoritam K-sredina
- Kohonenov algoritam
- DBSCAN/OPTICS
- Algoritam hijerarhijskog klasterovanja

Prva dva će biti sprovedeno uz pomoć programa *IBM SPSS Modeler*, dok će druga dva biti implementirana u programskom jeziku *Python*. Za ocenjivanje kvaliteta klasterovanja biće korišćen **senka koeficijent** (engl. *Silhouette coefficient*), koji se računa, za svaku instancu i na sledeći način:

$$s_i = \frac{b - a}{max(a, b)}$$

U gorenavedenoj formuli a je prosečna udaljenost instance i od ostalih instanci u klaseru, dok je b prosečno rastojanje između instance i i instanci iz najbližeg susednog klastera. Senka koeficijent se nalazi u opsegu $[-1, 1]$; što je senka koeficijent bliži 1, to je klasterovanje gušće.

3.1 K-sredina

K-sredina (engl. *K-means*) je algoritam koji, na osnovu korisnički definisanog parametra k, nalazi k klastera i centre tih klastera. To znači da se taj parametar mora odabrati na neki način; iako postoje bolje metode za determinisanje najboljeg parametra, jednostavnosti radi će biti uzete sve vrednosti parametra od 2 do 5.

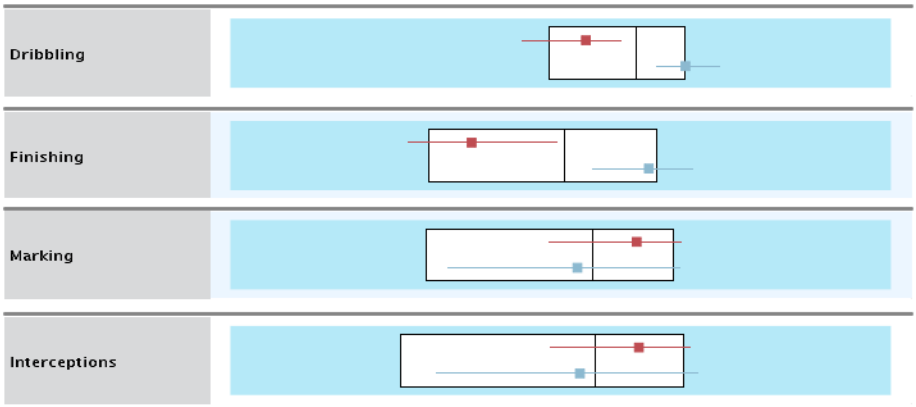
Tabela 1: Algoritam K-sredina za različite vrednosti k

k	Senka koeficijent
2	0.6
3	0.4
4	0.4
5	0.3

Iz priloženog eksperimenta, može se uočiti da algoritam K-sredina najbolje funkcioniše za $k = 2$. Ako se pogledaju dobijeni klasteri, postaje jasno da je ovaj metod odvojio golmane i ostale igrače. Naime, svi igrači imaju atribute koje karakterišu branjenje (kao što su npr. refleksi), ali igrači čija je prirodna pozicija na golu imaju mnogo bolje odgovarajuće atribute. Slično važi i u suprotnom slučaju, golmani nemaju dobre atribute van onih vezanih za branjenje.

Interesantniji slučajevi su kada je $k = 3$ ili $k = 4$. Kada postoje 3 klastera, igrači se dele na golmane, ofanzivne i defanzivne igrače. Na slici 1 su prikazani atributi koji prikazuju razlike između klastera ofanzivnih i defanzivnih igrača:

Vidimo da plavi klaster ima igrače koji su (u proseku) bolji u driblanju i realizaciji, dok crveni klaster se odlikuje boljim ocenama atributa vezanih za defanzivne zadatke, kao što su markiranje i „presecanje“ dodavanja. Međutim, podela nije savršena; recimo, bekovi se ne smatraju defanzivcima jer oni često imaju i ofanzivne zadatke.



Slika 1: Neki od korišćenih atributa za algoritam 3-sredina

Ako se uzme da je $k = 4$, dobijamo do podjednako zadovoljavajućeg klasterovanja, sudeći po senka koeficijentu. Ako postoje četiri klastera, očekivani rezultat bi bio da se dobiju klasične fudbalske linije: golmani, defanzivci, napadači i vezni red. Međutim, razlika između napadača i igrača veznog reda nije velika, tako da se dobiju neočekivani rezultati. Eksperimentalnim testiranjem klasterovanja gde je $k > 4$ dobijaju se sve lošiji rezultati.

3.2 Kohonenov algoritam

Model Kohonenovog algoritma je posebna vrsta modela neuronskih mreža koja mapira atribute u dvodimenzionalni prostor i grupiše slične instance. Ovaj model nema skrivene slojeve, već samo ulazni i izlazni sloj, pri čemu su čvorovi ulaznog sloja atributi, a čvorovi izlaznog sloja su klasteri.

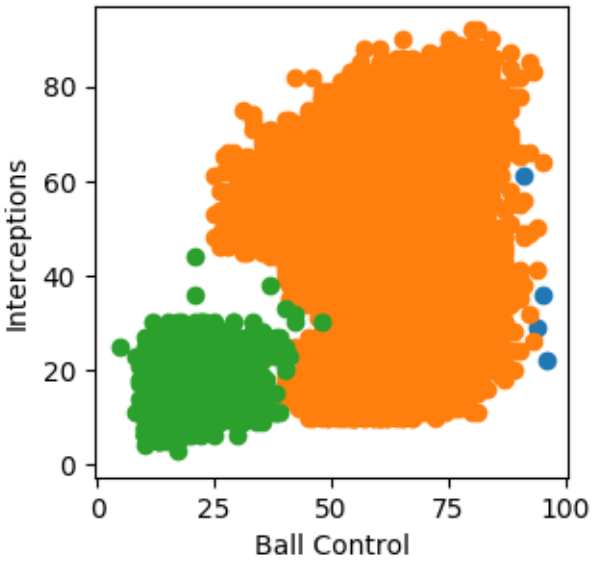
Korišćenjem ovog algoritma u programu *SPSS Modeler* nije dao dobre rezultate, bez obzira na prosleđene *width* i *length* atribute. Shodno tome, ovom algoritmu neće biti posvećeno mnogo pažnje.

3.3 DBSCAN/OPTICS

DBSCAN (**D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise) je algoritam klasterovanja zasnovan na gustini. Ovaj algoritam se ponaša dobro za klastere nepravilnih oblika i za elemente van granica.

U 2011. godini pojavio se i OPTICS algoritam[2] (**O**rdering **P**oints **T**o **I**dentify the **C**lustering **S**tructure). OPTICS se zasniva na istim osnovama kao i DBSCAN, ali donosi poboljšanje prilikom prepoznavanja klastera u prostorima manje gustine. U programskom jeziku *Python*, korišćenje OPTICS algoritma ima još jednu prednost u odnosu na DBSCAN: nije neophodno unositi parametar ε, jer će se on automatski izračunati. Korišćenje ovog algoritma se vrši uz pomoć funkcije `OPTICS` koja se nalazi u paketu `sklearn.cluster`.

Na slici 2 je prikazano korišćenje algoritma OPTICS sa minimumom od 50 uzoraka. Tačke obojene žutom i zelenom bojom predstavljaju klastere, dok su plave tačke prepoznate kao šum. Korišćenjem ovog algoritma dobija se senka koeficijent od 0.370271 što je umereno dobar rezultat.



Slika 2: Rezultati klasterovanja za algoritam OPTICS

3.4 Hijerarhijsko klasterovanje

Hijerarhijsko klasterovanje izdvaja skupove ugnježdenih klastera, tako da oni formiraju hijerarhijsko stablo. Postoje dva tipa hijerarhijskog klasterovanja: **klasterovanje spajanjem** i **klasterovanje razdvajanjem**. Razlika između ova dva tipa je u početnom pristupu klasterovanju; kod spajanja, svaka instanca se posmatra kao zasebni klaster, dok se kod razdvajanja sve instance nalaze u jednom klasteru. Prilikom klasterovanja ovog skupa podataka, biće korišćeno klasterovanje spajanjem.

Parametri neophodni za ovaj tip klasterovanja su broj klastera, mera bliskosti i kriterijum određivanja blizine klastera. Nakon korišćenja više različitih kombinacija parametara, Vardov metod[1] (engl. *Ward's method*) i Euklidsko rastojanje su se pokazali najboljim. U listingu 1 je prikazano korišćenje ovog algoritma u programskom jeziku *Python*.

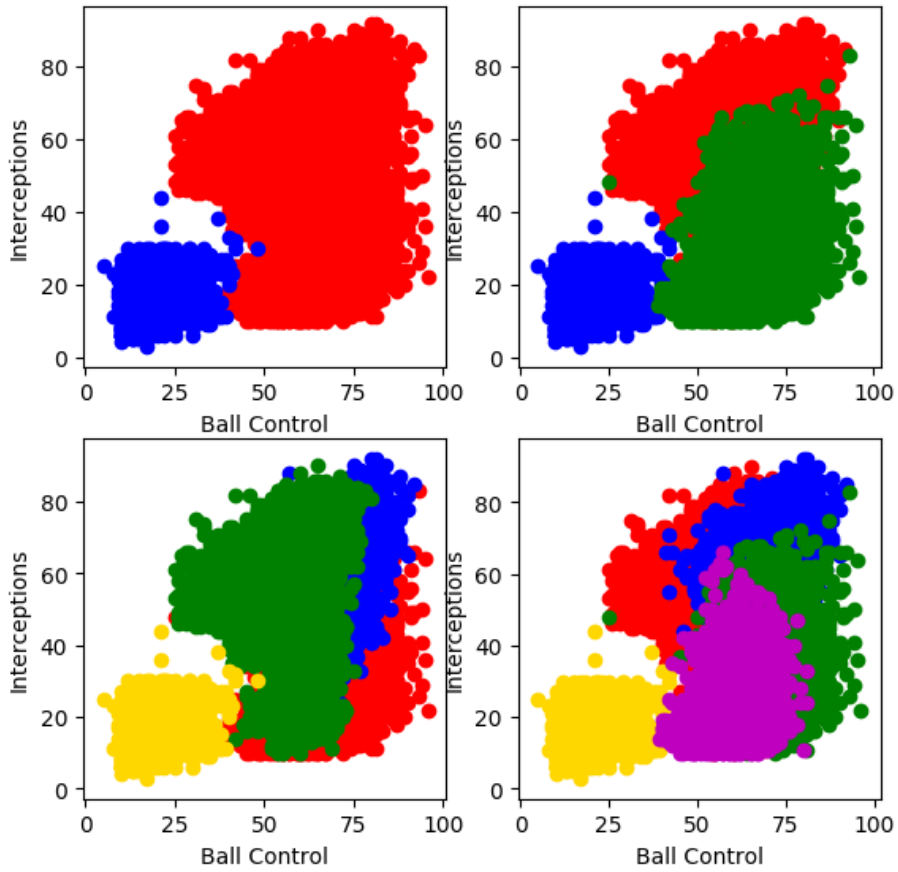
```
1000 est = AgglomerativeClustering(n_clusters=num_clusters, linkage='
         ward', affinity='euclidean')
     est.fit(scaled_df)
1002 df['labels'] = est.labels_

1004 print("Silhouette score: %f " % silhouette_score(scaled_df, est.
         labels_))
```

Listing 1: Hijerarhijsko klasterovanje u programskom jeziku Python

Međutim, rezultati velikim delom nisu zadovoljavajući: dobijaju se dobri rezultati isključivo za dva klastera, dok već kod tri klastera rezultati značajno opadaju. Na slici 3 su prikazani rezultati za do 5 klastera, gde su ordinati pridružene vrednosti kontrole lopte igrača, dok su apscisi pridružene vredosti presecanja lopte:

Slično kao i kod algoritma K-sredina, razdvajanje igrača na golmane i „spoljne igrače“ radi na zadovoljavajućem nivou. Klasterovanje sa više od dva klastera stvara probleme, tako da se senka koeficijent prepolovljava



Slika 3: Rezultati hijerarhijskog klasterovanja za 2, 3, 4 i 5 klastera

već kod tri klastera.

4 Zaključak

Kod svih prikazanih metoda klasterovanja, primetno je da klasteri igrača nisu jasno izdvojivi; najčešće postoji pouzdan način da se skup podeli na 2 klastera, ali prilikom povećavanja broja klastera, dobijaju se značajno lošiji rezultati. Kao što je napomenuto u tekstu, dva klastera u pitanju su golmani i ostali igrači; postoje velike razlike u vrednostima atributa golmana i ostalih fudbalera.

Razlog loših rezultata klasterovanja je dvojak: kao prvo, fudbaleri nisu strikno vezani za pozicije i svaki igrač ima jedinstven skup sposobnosti. Takođe, neke pozicije ne pripadaju tačno jodnoj „liniji“ tima (npr. bekovi, krila). Kao drugo, atributi igrača nisu eksperimentalno izmereni, niti se do njih došlo nekim naučnim putem; oni su uneti od strane razvojnog tima kompanije EA Sports, a preslikavanje sposobnosti fudbalera u numeričke atribute je možda i nemoguće.

U radu sa ovim skupom podataka, došlo se do zaključka da on nije

najpogodniji za primenu klasterovanja; štaviše, stiče se utisak da bi primena klasifikacije nad ovim podacima, gde bi klase bile pozicije fudbalera, bila bolji pristup za njihovo bolje razumevanje.

Literatura

[1] Joe H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 2011.

[2] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, and Arthur Zimek. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):231–240, 2011.