UNIVERZITET U NOVOM SADU
PRIRODNO MATEMATIČKI FAKULTET
DEPARTMAN ZA MATEMATIKU I INFORMATIKU

Nenad Rakić

PRIMENA METODA MAŠINSKOG UČENJA ZA RANGIRANJE INDIVIDUALNIH SPOSOBNOSTI

-master rad-

Mentor: prof. dr Dušan Jakovetić

Novi Sad, 2020.

ZAHVALNICA

Želeo bih se zahvalim mentoru dr Dušanu Jakovetiću na strpljenu, savetima i sugestijama tokom izrade master rada.

Takođe, želeo bih da se zahvalim članovima komisije dr Nataši Krejić i dr Nataši Krklec Jerinkić na prenesenom znanju tokom studija.

Posebnu zahvalnost dugujem roditeljima na bezuslovnoj podršci koju su mi pružili tokom svih ovih godina studiranja, posebno majci na strpljenju u iščekivanju odbrane master rada.

APSTRAKT

U ovo savremeno tehnološko doba gde nam je dostupan ogroman broj podatka, adekvatna analiza dovodi do precizinijh zaključaka pri rešavanju određenog problema. Matematika i primena matematičkih metoda omogućava i olakšava donošenje preciznijih i adekvatnijih odluka. Još u ranom periodu sam se interesovao za matematiku i fudbal, naravno ne sluteći da oni mogu biti povezani na bilo koji način. Tokom studiranja sam uvideo da se matematika prožima kroz veliki broj oblasti, pa tako i u fudbalu. Shvativši široku povezanost matematike i fudbala, odlučio sam da otkrijem bar deo te povezanosti i prenesem svoja saznanja i primenim svoja znanja kroz ovaj master rad.

Ovoj master rad je posvećen analizi elemenata fudbalske igre, vrednovanju dodavanja u fudbalu, primeni metoda mašinskog učenja, procenjivanju individualnih sposobnosti na osnovu dodavanja i na osnovu grupnog poređenja. Za procenu vrednovanja i predviđanja individualnih sposobnosti se koriste i primenjuju određeni metodi mašinskog učenja. Na osnovu metoda mašinskog učenja se prave modeli koji su značajni u analizi i predikciji. Značajnost praćenja i predviđanja vrednosti u današnje vreme je velika, jer kako se usavršava i napreduje tehonologija, tako napreduje i fudbal kao igra. Na osnovu vrednovanja dodavanja, pas igre i poseda može se odrediti da li je ekipa ofanzivnija ili defanzivnija. Mogu se rangirati igrači u zavisnosti od njihovih dodavačkih sposobnosti, mogu se skautirati igrači od kojih preti najveća opasnost i u kom delu terena je protivnička ekipa najopasnija. Zbog toga se velika važnost pridaje i lokacija dodavanja. Vrednovanje dodavanja je značajno za procenjivanje tehničkih sposobnosti igrača, sposobnost igranja tačnog pasa, završnog dodavanja ili veštine za postizanje gola. Na osnovu ovih fudbalskih karakteristika određuje se tržišna vrednost igrača, moguće je upoređivanje igrača na istim pozicijama kao i upoređivanje dve fudbalske ekipe. Vrednovanje dodavanja može biti korisno i za predviđanje krajnjih ishoda utakmice kao i drugih statističkih elemenata koji se javljaju tokom jedne fudbalske utakmice.

SADRŽAJ

1. UVOD 4
2. TEORIJSKE OSNOVE: MAŠINSKO UČENJE 8
2.1. UVOD U MAŠINSKO UČENJE 8
2.1.1 K-SREDNJE PARCIJALNA KLASTERIZACIJA 16
2.1.2 BAJESOV KLASIFIKATOR 22
2.1.3 METODA PODRŽAVAJUĆIH VEKTORA (MPV) 23
2.2 MERA SLIČNOSTI 31
2.2.1 DINAMIČKO VREMENSKO SAVIJANJE 31
2.2.2 FREČETOVA MERA DISTANCE 34
2.2.3 NAJDUŽI ZAJEDNIČKI PODNIZ 35
2.3 K-NAJBLIŽI SUSEDI 35
3. OPIS PODATAKA 37
3.1 SKUPOVI PODATAKA 37
3.2 PODELA SKUPA PODATAKA 38
4. FUDBALSKA ANALIZA 39
4.1 PODELA TERENA 39
4.2 PERIOD POSEDOVANJA 40
5. PRISTUPI 43
5.1 ZONSKI-ORIJENTISANO VREDNOVANJE DODAVANJA 43
5.1.1 PODELA TERENA 44
5.1.2 VREDNOVANJE ZONE KORISTEĆI TRENING PODATKE 45
5.1.3 ODREĐIVANJE VREDNOSTI DODAVANJA 46
5.2 PAS-ORIJENTISANO VREDNOVANJE DODAVANJA 46
5.2.1 DEFINISANJE MERE DISTANCE 47
5.2.2 ODREĐIVANJE ISHODA PERIODA POSEDOVANJA 50
5.2.3 POD-KLASTERIZACIJA DODAVANJA 50
5.2.4 ODREĐIVANJE VREDNOSTI DODAVANJA 51
5.3 NIZOVNO-ORIJENTISANO VREDNOVANJE DODAVANJA 52

5.3.1 ODREĐIVANJE MERE DISTANCE ZA NOV 53
5.3.2 RAZDVAJANJE PERIODA U POD-PERIODE 55
5.3.3 POD-KLASTERI POD-PERIODA POSEDOVANJA 55
5.3.4 RAČUNANJE DISTANCE I ODREĐIVANJE K-NAJBLIŽIH SUSEDA 56
5.3.5 VREDNOVANJE DODAVANJA 56
6. INDIVIDUALNA RANGIRANJA ZASNOVANA NA GRUPNOM POREĐENJU 59
6.1 BINARNI ISHOD I MERNJE ISHODA 59
6.2 POREĐENJE SA BINARNIM ISHODIMA 61
6.2.1 METODA NAJMANJIH KVADRATA (MNK) 62
6.2.2 MAKSIMALNA VERODOSTOJNOST (MAX.VER) 65
6.3 POREĐENJE SA MERENIM ISHODIMA 67
6.3.1 MODEL NORMALNE RASPODELE (MNR) 67
6.3.2 MODEL RASPODELE EKSTREMNE VREDNOSTI (MREV) 69
6.4 PROCENJIVANJE PRISTUPA I GREŠAKA U RANGIRANJU 70
6.4.1 GREŠKE ZA SPOSOBNOST I RANGIRANJE 70
6.4.2 UPOREĐIVANJE PRISTUPA 72
7. ZAKLJUČAK 74

1. UVOD

Fudbal je kako mnogi smatraju “najvažnija sporedna stvar na svetu”. Osim što predstavlja sportsku igru, fudbal je odraz socijalnog i kulturnog stanja i razvoja određene države. O tome nam govori razvoj infrastrukture kroz vremenske periode i način ponašanja ljudi na stadionu. Milioni ljudi širom planete svakodnevno prate fudbalske utakmice, reportaže i iz minuta u minut prate vesti sa evropske i svetske fudbalske scene. Sa nestrpljenjem se isčekuje svaka naredna utakmica kako u najvećim klupskim prvenstvima tako i na reprezentativnom nivou. Koliko je fudbal zastupljen i popularan govori podatak da je finale Lige šampiona 2018. između Livepula i Totenhema pratilo oko 350 miliona gledalaca širom sveta. Najgledanije klupsko nacionalno takmičenje je Premijer liga. Engleska se smatra za kolevku fudbala, koji danas gledamo. Prema statističkim podacima sa zvaničnog sajta Premier lige (www.premierleague.com) fudbalsko prvenstvo ovog takmičenja je pratilo oko 1,35 miliona gledalaca širom planete. To donosi i ogroman priliv novca od televizijskih prava i sponzora. Veliki prinosi omogućavaju i veće ulaganje kapitala kluba za skautiranje i kupovinu najboljih igrača na svetskog fudbalskoj pijaci. Pošto se radi o višemilionskim ulaganjima u transfere igrača potrebno je dobro razmotriti sve aspekte koji utiču da li će se uložiti potreban novac u dovođenje igrača. U takvim procenama daje se veliki značaj podacima koji su korisni pri donošenju značajnih odluka.

Mašinsko učenje je veoma korisna i primenljiva oblast nauke. Koristi se u analizi i izučavanju raznih naučnih dostignuća u kojima se radi sa velikim brojem podataka. Smatra se da je mašinsko učenje nezaobilazna i neizbežna oblast u analitičkim i tehnoloskim dostignućima. Osnovni princip koji se primenjuje u mašinskom učenju je korišćenje algoritma i algoritamske logike, uočavanje zakonitosti i pravila između sličnih podataka.

Korišćenje podataka je značajno i zastupljeno u sportu, naročito u sadašnjem vremenu modernizacije i tehnologije. Kao najpopularnija i najzastupljenija sportska oblast, fudbal se sve više i više razmatra kao oblast istraživanja u matematici i informatici, na osnovu dostupnosti velikog broja podataka, koji se prikupljaju tokom jedne fudbalske utakmice. Dostupnost velikog broja podataka omogućava menadžerskim kompanijama da prate i procenjuju individualni kvalitet fudbalera, kao i kvalitet fudbalskih ekipa.

Tokom fudbalske utakmice, najčešći tip događaja koji igrač ostvari sa loptom je dodavanje. To je značajno za proučavanje, zato što se fudbalska filozofija i taktika zasnivanju na dodavanju i pas igri. Proučavanje dodavanja u modernom fudbalu je sve značajnije. Zahvaljujući velikom broju dostupnih podataka vrše se razna istraživanja vezana za dodavanja tokom jedne utakmice. Problem podataka o događaju sa loptom je u tome što je na osnovu njih teško tačno proceniti dodavačke sposobnosti igrača. Ti podaci ne ukazuju koliko je dodavanje dobro, nego označavaju samo lokaciju dodavanja, njegovo poreklo i destinaciju.

Za procenu dodavačkih sposobnosti fudbalera koriste se određeni modeli koji su fokusirani na kompletnoj oceni fudbalerovih dodavanja ili su fokusirani na broju asistencija (dodavanje posle kojeg je postignut gol) i završnih dodavanja (dodavanje posle kojeg se fudbaler nađe u gol šansi ali ne mora da postigne gol). Na primer, dodavanje lopte između dva odbrambena igrača ili između odbrambenog igrača i igrača sredine terena je lakše nego dodavanje igrača iz sredine terena ka napadaču, gde je potrebna veća veština i sposobnost, ali je i veća mogućnost greške. Kada se razmatra kompletna ocena, dodavanja su jednako vrednovana, mada logički posmatrano ova dve vrste dodavanja ne bi trebalo da uzimaju istu vrednost. Kako su asistencije i završno dodavanje ređi događaji u toku jedne fudbalske utakmice, njihovim korišćenjem gubi se mnogo podataka u određivanju dodavačkih sposobnosti fudbalera. Tehnike mašinskog učenja se koriste da bi se dobila metrika za vrednovanje dodavanja u fudbalu.

Kako bi se model mogao koristiti za rangiranje fudbalera, cilj je napraviti model koji dodeljuje vrednost za svako dodavanje. Rezultujući model vrednuje dodavanja prema njihovom očekivanom doprinosu na ishod utakmice (ECOM). Ovaj model omogućava rangiranje igrača na osnovu njihovog doprinosa na ishod utakmice. Fokusirali smo se na dodavanja koja imaju uticaj na stvaranje prilike za postizanje gola. Preciznost i težina dodavanja nisu uzeti u obzir, zbog nedostatka tačnosti praćena pozicije igrača na terenu. Dodavanja, koja razmenjuju saigrači, je teško deklarisati da su dobra ili loša. Dodavanje je dobro ako se saigrač nađe u gol šansi. Međutim, ako napadač ne postigne gol to dodavanje neće biti nagrađeno kao što bi bilo da je postignut gol, jer na kraju utakmice to dodavanje nije uticalo na krajnji rezultat. Postizanje gola je redak događaj na fudbalskoj utakmici. Koristi se očekivani gol-model koji određuje verovatnoću, da li će se gol pokušaj pretvoriti u pogodak.

Prikupljanjem, analiziranjem, modeliranjem i procenjivanjem svih ovih svojstava vidi se uska povezanost između matematike i fudbala. Pomoću raznih metoda matematičkih pristupa moguće je uvideti koliki uticaj ima matematika na fudbal, zato se u narednim tezama pokušava dati odgovor na sledeća pitanja:

1. "U kojoj meri su metode mašinskog učenja sposobne da odrede igrača koji ima najveći uticaj na ishod tima?"
2. "Da li se može napraviti model koji pronalazi igrače sa sličnim fudbalskim sposobnostima?"
3. " Na koji način se može upoređivati posedovanje lopte i kako se može računati u određenim delovima terena?"
4. " Da li postoji model za otkrivanje stila igre određenog tima i da li se mogu naći timovi iz različitih liga koji imaju sličan stil igre?"
5. "Da li se tržišna vrednost igrača i timova, može proceniti na osnovu dodavačkih sposobnosti igrača?"
6. "Na koji način se na osnovu rezultata timova mogu rangirati igrači?"

U drugom delu se govori o mašinskom učenju. Obrađene su neke od metoda mašinskog učenja. Na osnovu tih metoda definiše se mera sličnosti i kreiraju modeli za predikciju. Oni se primenjuju za rangiranje individualnih sposobnosti igrača. Metode mašinskog učenja koje se obrađuju u ovom radu su: dinamčko vremensko savijanje, Frečetova mera distance, najduži zajednički podniz, k-najbliži susedi.

Treći deo se bavi podelom fudbalskog terena na zone. Proučava podeljene zone, lokaciju dodavanja i definiše se period posedovanja.

U četvrtom delu se govori o načinu prikupljanja podataka i njihova selekcija tokom fudbalske utakmice. Takođe, govori se o uticaju statistike na rangiranje igrača i klubova.

U petom delu akcenat je stavljen na pristupima za određivanje vrednosti dodavanja. U zavisnosti na koji način će se vrednovati dodavanje koriste se tri pristupa vrednovanja: zonski-orijentisano vrednovanje, pas-orijentisano vrednovanje i nizovno-orijentisano vrednovanje.

U šestom odeljku se objašnjavaju pristupi koji razmatraju kako na osnovu ishoda grupnog poređenja rangirati individualne sposobnosti igrača. Definisana su dva tipa pristupa grupnog poređenja: binarni ishod i merenje ishoda.

2. TEORIJSKE OSNOVE: MAŠINSKO UČENJE

U ovom poglavlju se govori o mašinskom učenju, jednoj od značajnih naučnih podoblasti veštačke inteligencije. U ovom tehnološkom dobu modernizacije ogromna je primena mašinskog učenja. Kao nezaobilazna metoda, mašinsko učenje ima i veliku primenu u analiziranju podataka, određivanju i upoređivanju sličnosti događaja. Zato se u ovom radu i najveća pažnja posvećuje primeni mašinskog učenja.

2.1. UVOD U MAŠINSKO UČENJE

Mašinsko učenje predstavlja podoblast veštačke inteligencije. Nije egzaktna nauka. To podrazumeva da je nemoguće napraviti "savršen" model pristupom mašinskog učenja, koji dovodi do apsolutno tačnog rešenja za svaki pojedinačni slučaj. Težnja primene mašinskog učenja je da se od većeg broja modela izabere onaj koji nudi najbolja rešenja za veliki broj slučajeva. Mašinsko učenje omogućava računarima da uče bez eksplicitnog programiranja i obezbeđuje tehnike kojima se velike količine podataka mogu automatski analizirati. Osnovna ideja je da se koristi algoritam, umesto pisanja posebnog koda za neki problem. Algoritam pravi svoju logiku, otkriva pravila i zakonitosti između podataka i na kraju donosi zaključke koji mogu biti korisni u analiziranju podataka. Mašinsko učenje predstavlja skup tehnika koje generalizuju postojeće znanje nad novim podacima. Generalizacija je proces u kojem se znanje stečeno na osnovu prethodnog iskustva prenosi na novi, do tada ne analizirani skup podataka. Jednostavnije rečeno, mašinsko učenje omogućava da se na osnovu dostupnih podataka dobije dobro utrenirani model koji će se koristiti pri davanju odgovara za nove pojave i probleme. Korišćenje tehnika mašinskog učenja je veoma korisno u slučajevima gde: algoritamska rešenja nisu na raspolaganju, postoji nedostatak formalnih modela ili postoji problem razumevanja složenih funkcija, kao i za otkrivanje novih odnosa među pojmovima.

DEFINICIJA 1. [5]

Za kompijuterski program se kaže da uči iz iskustva E vezanog za zadatak T i meru performansi P, ukoliko se njegove performance na zadatku T, merene metrikom P unapređuju sa iskustvom E.

(Tom Mitchell 1998)

Na primeru za program koji obeležava kreditne zahteve kao dobre i loše, biće objašnjena definicija 1.

- Zadatak (T): klasifikacija kredita na dobre i loše
- Iskustvo (E): posmatranje procesa deklarisanja kreditnih rizika na dobre i loše
- Performansa (P): predstavlja procenat dobro klasifikovanih zahteva

Primena mašinskog učenja je sve više zastupljenija i obuhvata veliki broj oblasti, kao što su:

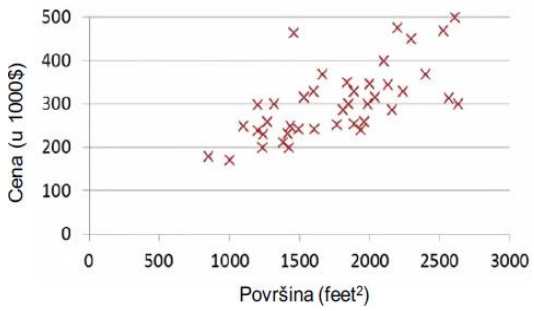
- Kategorizacija teksta prema temi, iskazanim stavovima…
- Autonomna vozila
- Segmentacija tržišta
- Razumevanje govornog jezika
- Mašinsko prevođenje teksta, itd…

Dva osnovna oblika mašinskog učenja su:

1. Nadgledano učenje
2. Nenadgledano učenje

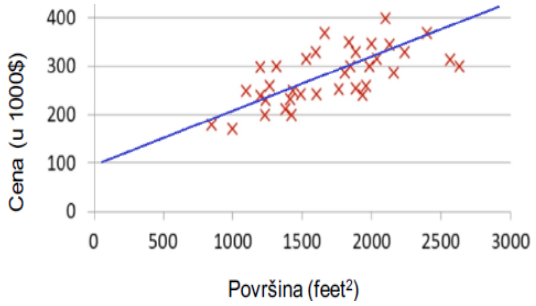
Nadgledano učenje se zasniva na tehnici da se algoritmu daje skup ulaznih podataka $(x_1, x_2, \dots, x_n)$ i skup željenih, izlaznih podataka $(y_1, \dots, y_n)$, tako da svaki ulazni podatak x_i ima odgovarajuci izlazni podatak y_i. Na osnovu ulaznih podataka algoritam treba da shvati kako da dobije odgovarajući izlazni podatak, tj. rešenje problema. Zadatak nadgledanog učenja je da se dobije model koji naučeno znanje primenjuje na novim ulaznim podacima i dobija odgovarajuću, željenu vrednost.

Kao primer, za nadgledano učenje može se posmatrati predviđanje cena nekretnina na osnovu površine nekretnina. Dati su podaci, slika 1 površina nekretnine (x) i cena nekretnine (y), kao podaci koji služe za proces učenja algoritma.



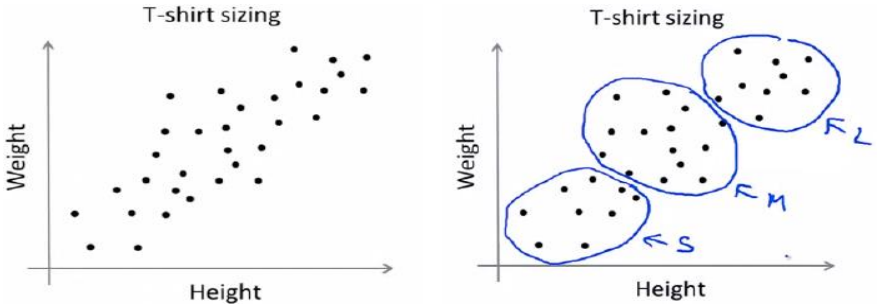
SLIKA 1: IZVOR [1]

Model koji treba da uči i zaključuje na osnovu dostupnih podataka je funkcija linearne regresije, slika 2. Ona ima oblik $h(x) = a + bx$, gde se koeficijenti a i b trebaju proceniti u tom procesu učenja.



SLIKA 2: IZVOR [1]

Kod nenadgledanog učenja nemamo izlazne podatke, postoje samo ulazni. Zadatak modela (algoritma) je da otkrije pravila, obrazce i zakonitosti koji važe između ulaznih podataka. Primer za nenadgledano učenje je određivanje veličine majice na osnovu visine i težine, slika 3.



SLIKA 3: IZOR [9]

Da bi se dobio odgovarajući model za rešavanje nekog problema, potrebno je pratiti sledeće osnovne korake u procesu mašinskog učenja:

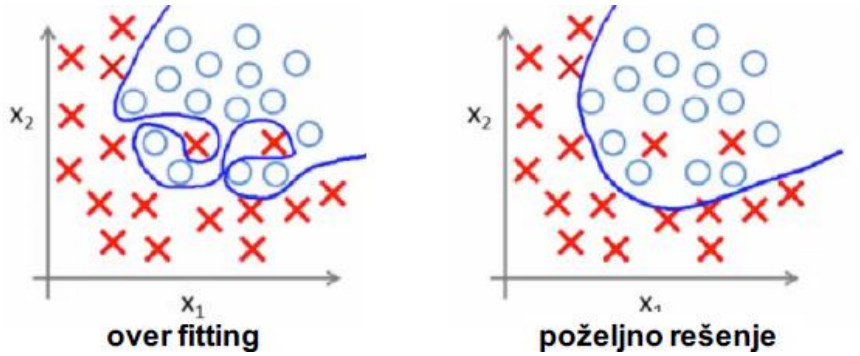
- Prikupljanje podataka potrebnih za formiranje skupa podataka, koji se koriste za treniranje, validaciju i testiranje modela mašinskog učenja
- Priprema podataka
- Analiza rezultirajućeg skupa podataka
- Izbor jednog ili više modela mašinskog učenja
- Obuka, konfiguracija i evaluacija kreiranih modela
- Izbor modela koji će se koristiti i njegovo testiranje

Selekcija podataka u procesu mašinskog učenja se vrši na slučajan način tako da se za trening koristi 60% podataka, za validaciju 20% i za testiranje 20% podataka. Trening podaci služe za obuku, konfiguraciju i razvoj jednog ili više modela. Podaci za validaciju se koriste za poređenje performansi kako bi se:

- Izabrao najbolji model od više kandidata
- Odredila optimalna konfiguracija parametara odabranog modela
- Izbegli problemi over/under-fitting-a.

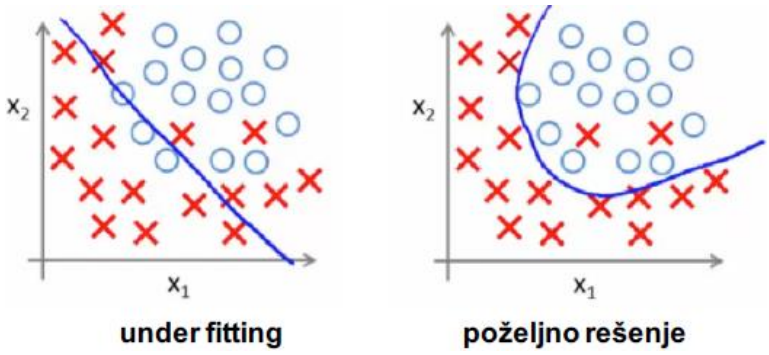
Over/under-fitting su dve bitne pojave koje je potrebno razmotriti pri kreiranju modela.

Over-fitting [9] iliti model prevelikog podudaranja odnosi se na situaciju u kojoj model savršeno nauči da vrši predikciju za podatke iz trening skupa, ali ima slabu sposobnost predikcije za nove podatke koji se bar malo razlikuju od naučenih. Drugim rečima, model utreniran na osnovu trening skupa neće dovoljno dobro generalizovati podatke izvan njega.



SLIKA 4: IZVOR [9]

Under-fitting [9] iliti problem nedovoljnog podudaranja predstavlja slabu prilagođenost modela. Kod under-fitting model je jednostavan i ne uspeva dobro da utrenira podatke iz trening skupa. Model ima slabe performanse na trening skupu pa ga je nemoguće koristiti i za testiranje na novim podacima.



SLIKA 5: IZVOR [9]

Na kraju, podaci za testiranje služe kao procena uspešnosti izabranog modela na novim podacima, koji nisu korišćeni ni za trening ni za validaciju modela.

Jedna od osnova mašinskog učenja je selekcija atributa. Atributi su komponente koje služe da verodostojno opišu pojavu. Suštinski atributi predstavljaju osobine koje karakterišu pojavu (instancu) i na osnovu njih se izdvajaju podaci koji su najoptimalniji u analizi i budućem predviđanju. Vrednosti atributa su uglavnom numeričke ali mogu biti i kategoričke, odnosno mogu predstavljati imena nekih kategorija kojima se ne može dodeliti neka specificirana numerička vrednost. Veliki je izazov odabir najkorisnijii atributa za opisivanje neke pojave. Na primer, atributi za određivanje cena nekretnina mogu biti:

- Površina i lokacija nekretnine
- Broj soba
- Tip gradnje
- Način grejanja….

Još jedna značajna karakteristika mašinskog učenja je analiza greške. Ona se koristi za uočavanje greške, posmatranjem primera gde je model napravio grešku, i pomaže da se te greške isprave i uklone iz modela. To se može uraditi na sledeći način:

- Identifikovanjem suvišnih atributa
- Identifikovanjem atributa koji nedostaju
- Podešavanjem parametara modela…

Mašinsko učenje rešava tri osnovna oblika problema koji imaju razne modifikcije i implementacije, a to su:

- Regresija
- Klasifikacija
- Klasterizacija

Regresija, oblik nagledanog učenja, predstavlja metodu kojim se opisuje odnos i povezanost zavisne primenljive Y i nezavisnih promenljivih (x_1, …, x_n). Regresija se koristi za predviđanje novih pojava na osnovu prethodnog iskustva, tj. predviđanje zavisne promenljive na osnovu skupa nezavisnih promenljivih. Predviđanje cena nekretnina je klasičan primer regresijsog modela. Regresija se koristi kada je potrebno dobiti brojne vrednosti.

Klasifikacija je oblik nadgledanog učenja, koja predstavlja jedan od zadataka mašinskog učenja sa ciljem da rasporedi nepoznate pojave u jednu od unapred ponuđenih kategorija (klasa). Kategorijama se ne može smisleno dodeliti numeričke vrednosti ili uređenje. Potrebno je odrediti vrednost atributa kategorije. Klasifikacijom pojava se određuje sličnost između posmatrane pojave sa već unapred kategorizovanim pojavama. Sličnost dve pojave se određuje analizom njihovih atributa. Cilj klasifikacije je da se na osnovu vrednosti atributa kategorizovane pojave napravi model koji će verodostojno klasifikovati nove pojave. Atribut klase može da ima konačan broj diskretnih vrednosti ali pojava uzima samo jednu vrednost atributa koja daje najbolju ocenu za određivanje kategorije. Broj kategorija (klasa) mora biti unapred poznat i ograničen. Klasifikacija ima brojnu primenu. Koristi se u dijagnostici bolesti, odabiru najbolje terapije, odobravanje kreditnih zahteva klijenata, analizi slike, analizi glasa…

Klasterizacija je jedan od oblika nenadgledanog učenja. Zadatak klasterizacije je grupisanje pojava tako da za svaku pojavu važi da je sličnija pojavama iz svoje grupe (klastera) nego pojavama iz nekog drugog klastera. Suštinski sličnost podataka u okviru klastera je maksimalna a minimalna sa podacima drugih klastera. Algoritam teži da razdvoji ceo skup podataka na više homogenih klastera. Sličnost između dve pojave se utvrđuje merenjem:

- Sličnosti (koeficijent korelacije…)
- Udaljenosti (Euklidska razdaljina…)

Kod klasterizacije se ne zna unapred tačan broj klastera. Primena klaster analize je širokog spektra. Koristi se za: segmentaciju tržišta, redukciju podataka, testiranje novog proizvoda na tržištu po gradovima, identifikaciji korisnika koje karakterišu slični oblici interakcije sa sadržajima nekog web sajta, klasterizacija korisnika na osnovu demografskih podataka…

Postoji više metoda za određivanje sličnosti izmedju određenih objekata. U ovom radu će biti predstavljene samo neke od njih su:

- Metode klasterizacije (nenadgledne metode)
- Bajesov klasifikator (nadgledne metode)
- Metoda podržavajućih vektora (nadgledne metode)

Metoda klasterizacije (grupisanje) ima za cilj da organizuje objekte u klastere tako da grupisani podaci imaju slična svojstva. Na osnovu tako grupisanih objekata može da se dođe do korisnih zaključaka posmatranih objekata. Metode klasterizacije predstavljaju skup metodologija za automatsko klasifikovanje uzoraka tako da se u istom klasteru nalaze objekti sa sličnim svojstvima a razlikuju od objekata u drugim klasterima. Primena metoda klasterizacije je zastupljena u mnogobrojnim oblastima:

- Inžinjerstvu, u analizi podataka u cilju primene u industriji, robotici,
- Inteligentna analiza,
- Istraživanje tržišta: klasterizacija kupaca sa sličnim ponašanjem na osnovu neke baze podataka.

Koristi se u različite svrhe a posebno za:

- Redukciju podataka
- Predikciju zasnovanoj na klasterima.

DEFINICIJA 2. [5]

Klasteri predstavljaju neprekidnu oblast prostora sa velikom gustinom tačaka, odvojena od drugih, istih takvih oblasti sa oblastima prostora sa relativno malom gustinom tačaka.

Klaster opisan na ovakav način se često zove i prirodni klaster.

Neka je X skup podataka definisan kao:

$$X = \{x_1, x_2, \dots, x_n\}$$

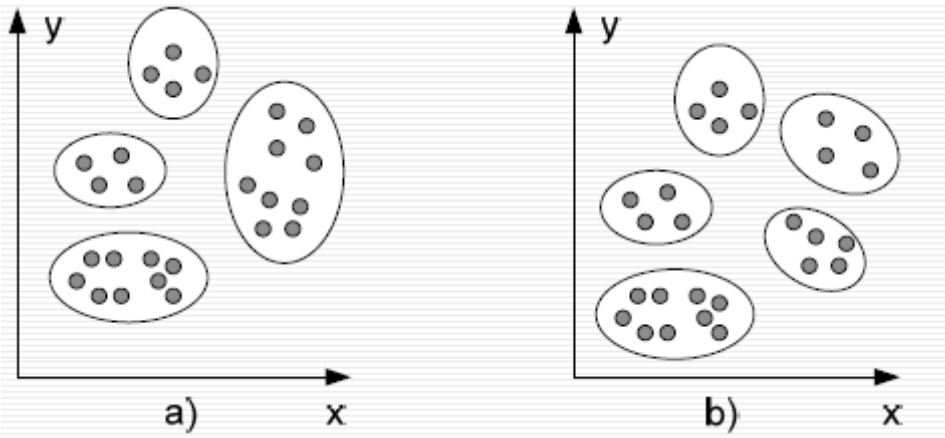
Klasterizacija skupa X predstavlja njegovu podelu u k klastera G_1, G_2,…, G_k tako da su zadovoljena sledeća tri uslova:

$$G_i \neq \{\}, \quad i = 1, 2, \dots, k$$

$$\bigcup_{i=1}^{k} G_i = X$$

$$G_i \cap G_j = \{\}, \quad i \neq j, \; j = 1,2, \dots, k$$

Na sledećim slikama se vidi način klasterizacije objekata gde se u jednom klasteru nalaze objekti koji imaju slične osobine.

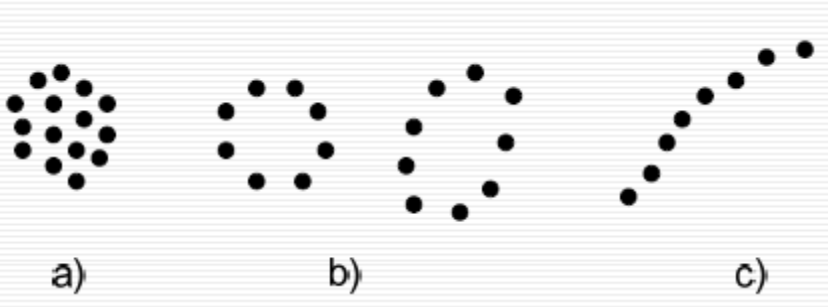


SLIKA 6: IZVOR [5]

Postoji više vrsta klasterovanja podataka, zavisno od strukture podataka klasterizacija može biti:

a) Kompaktna
b) Sferična i elipsoidna
c) Izdužena.

Na sledećoj slici je prikazan izgled gore pomenutih vrsta klasterizacije.



SLIKA 7: IZVOR [5]

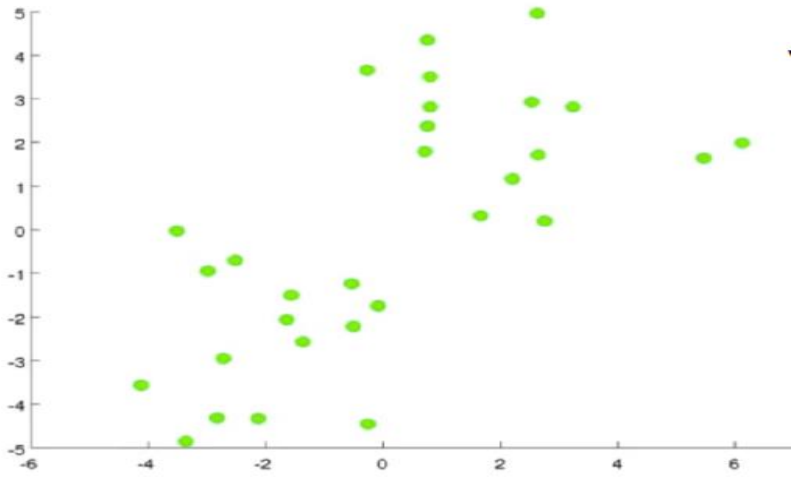
Da bi se objekti pravilno klasterizovali potrebno je uraditi pripremu podataka i izvršiti potrebnu proceduru. Osnovni koraci koji se primenjuju prilikom klasterizacije objekata su:

- Biranje atributa objekata
- Određivanje mere sličnosti/razlike
- Kriterijum klasterizacije podataka
- Algoritam klasterizacije podataka
- Validacija rezultata
- Interpretacija rezultata

2.1.1 K-SREDNJE PARCIJALNA KLASTERIZACIJA

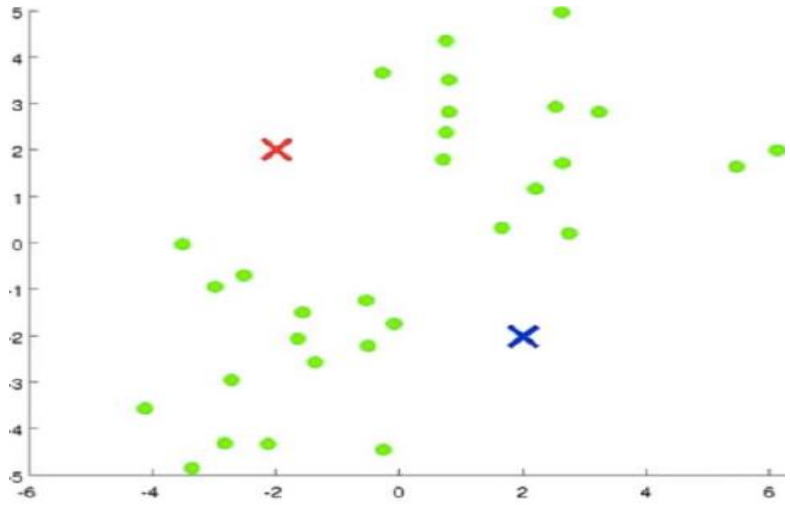
K-srednje parcijalna klasterizacija je jedna od najpoznatijih i najjednostavnijih algoritama klasterizacije, za grupisanje objekata sličnih karakteristika. Na sledećem primeru će se slikovito opisatii način funkcionisanja ove metode.

1) Neka su dati ulazni podaci objekata koje je potrebno klasterizovati.



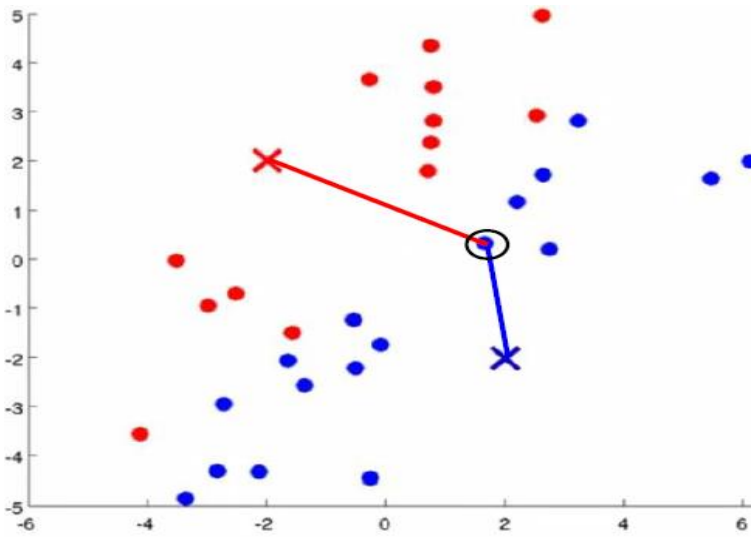
SLIKA 8: IZVOR [8]

2) Inicijalizacija: inicijalni izbor težišta klastera (K=2) metodom slučajnog izbora.



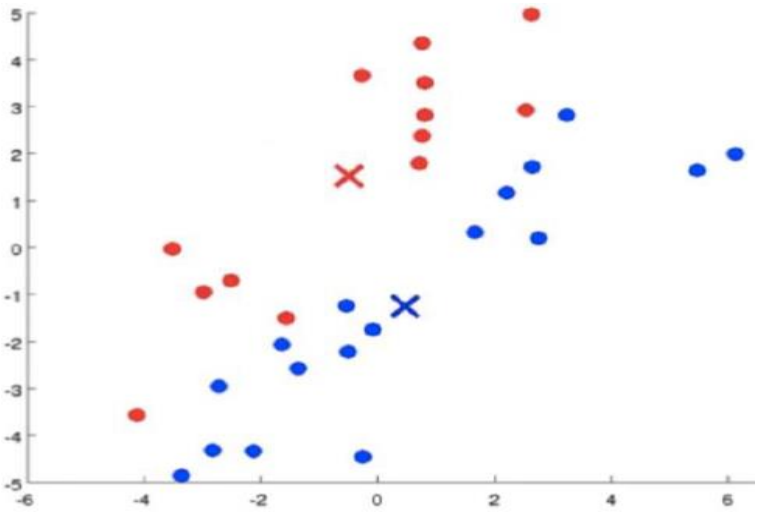
SLIKA 9: IZVOR [8]

3) Iteracija 1, korak 1: razvrstavanje instanci na osnovu udaljenosti od težišta klastera.



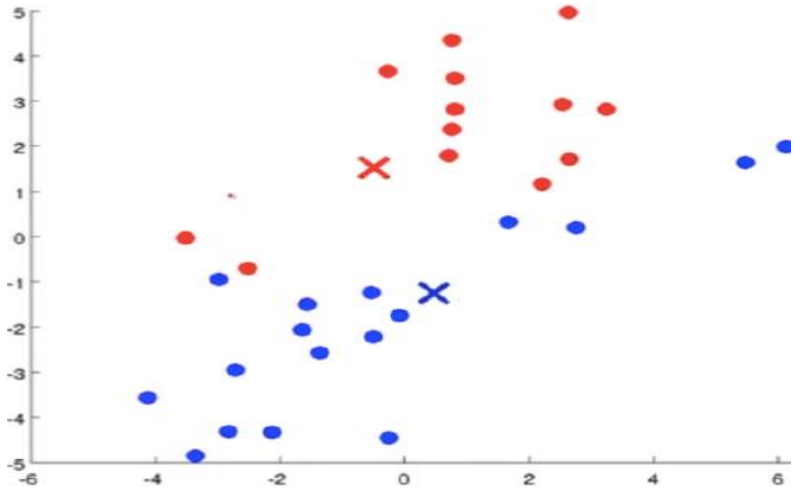
SLIKA 10: IZVOR [8]

4) Iteracija 1, korak 2: određivanje novog težišta za svaki klaster, na osnovu proseka vrednosti instanci u datom klasteru.



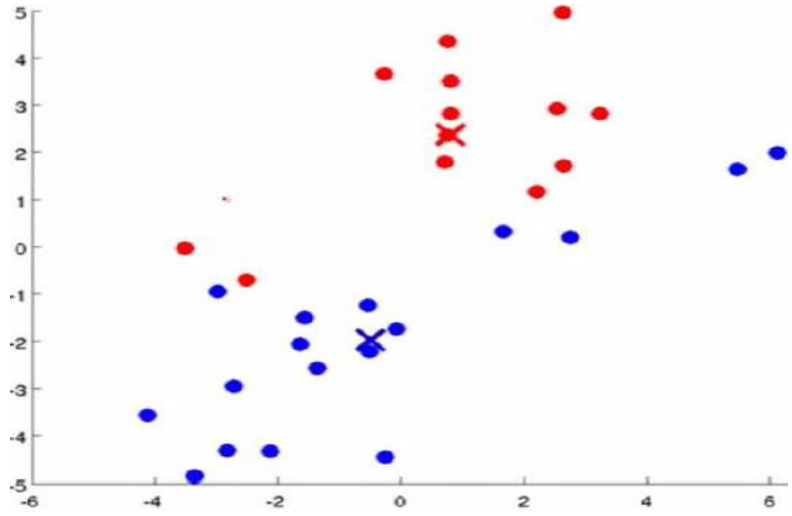
SLIKA 11: IZVOR [8]

5) Iteracija 2, korak 1: ponovno razvrstavanje instance po klasterima na osnovu udaljenosti od težišta klastera.



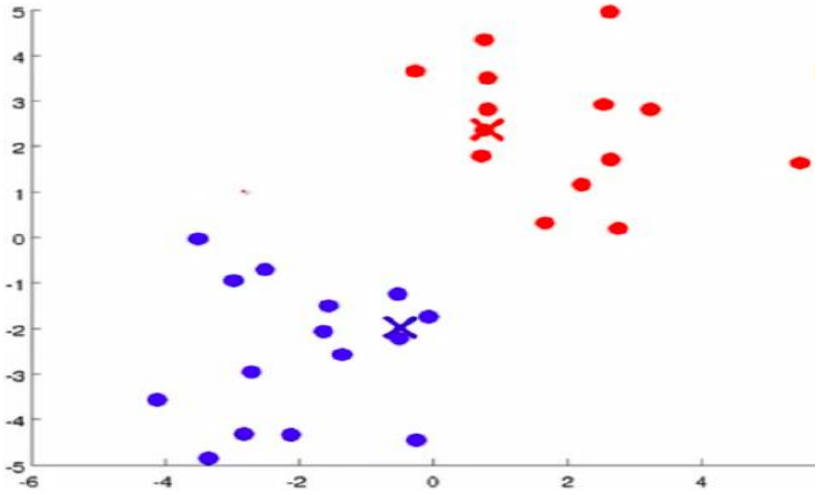
SLIKA 12: IZVOR [8]

6) Iteracija 2, korak 2: ponovno određivanje novog težišta za svaki klaster.



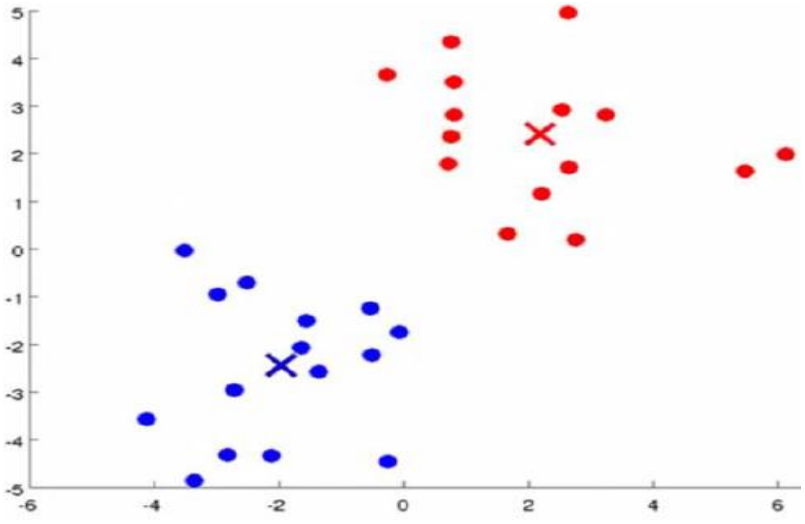
SLIKA 13: IZVOR [8]

7) Iteracija 3, korak 1: ponovno razvrstavanje instance po klasterima na osnovu udaljenosti od težišta klastera.



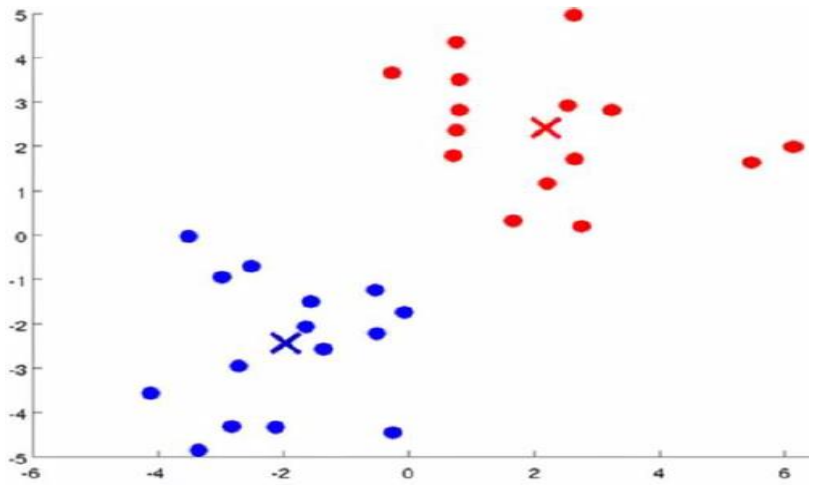
SLIKA 14:IZVOR[8]

8) Iteracija 3, korak 2: ponovno određivanje novog težišta za svaki klaster na osnovu proseku instance.



SLIKA 15: IZVOR [8]

9) Algoritam već konvergira: dalje iteracije neće dovesti do značajne promene i proces se zaustavlja. Dobijaju se dva klastera gde su objekti u jednom klasteru sličnih karakteristika a razlikuju se od karakteristika objekata u drugom klasteru.



SLIKA 16: IZVOR [8]

Glavna pretpostavka *k*-mean klasterizacije je da funkcija pripadnosti klastera μ_{ij} može imati vrednosti 0 ili 1

$$\mu_{ij} = \{0,1\},\ j = 1, \ldots, k$$

$$\sum_{j=1}^{k} \mu_{ij} = 1$$

Celokupna populacaju na kojoj se radi *k*-mean klasterizacija se deli u *K* klastera:

$$C_1, \ldots, C_K.$$

Svaki klaster sadrži n_k uzoraka

$$\sum n_k = N\ k = 1, \ldots, K$$

Srednja vrednost u algoritmu (M_k) predstavlja težišnu tačku odnosno prosečnu vrednost objekata u klasteru (C_k) i definiše se:

$$M_k = (\frac{1}{n_k}) \sum_{i=1}^{n_k} x_{ik}.$$

Kvadratna greška klastera C_k je suma kvadratnih razdaljina između svakog uzorka u klasteru (x_{ik}) i srednje vrednosti u algoritmu:

$$e_k^2 = ||x_k - M_k||_2^2 .$$

Ukupna kvadratna greška celog prostora koji sadrži svih *K* klastera:

$$E_k^2 = \sum_{i=1}^{n_k} e_k^2$$

Prednost *K*-mean klasterizacije je u tome što je jednostavan za implementaciju i ne zahteva mnogo vremena za izvršavanje algoritma.

Problem sa *K*-mean klasterizacijom je neizvesnost u određivanju broja klastera i podešavanjem srednje vrednosti algoritma usled lošeg izbora inicijalne particije.

2.1.2 BAJESOV KLASIFIKATOR

Neka je skup nezavisnih hipoteza [14]:

$$H(k) = \{H_i(k), i = 0,1, \dots, N\}$$

P (H_i) je početna verovatnoća hipoteze H_i pre eksperimenta i naziva se još apriorna verovatnoća. Ukoliko ne postoji apriorna verovatnoća onda se svim hipotezama dodeljuje jednaka početna verovatnoća. Nakon dodeljivanja početne verovatnoće, izvodi se eksperiment čiji je rezultat događaj X. $P(X)$ je verovatnoća pojavljivanja događaja X bez obzira na to koja je hipoteza ispravna.

$P(X/H_i)$ je uslovna verovatnoća pojavljivanja X uz uslov ispravnosti hipoteze H_i.

$P(H_i/X)$ je uslovna verovatnoća ispravnosti hipoteze H_i nakon pojavljivanja događaja X. Ovako definisana verovatnoća se zove a'posterior verovatnoća koja predstavlja tačnost hipoteze H_i nakon što se dogodio događaj X.

Teorema (Totalna verovatnoća): [14]

Pretpostavlja se da su hipoteze $H_1, H_2, \ldots, H_n$ međusobno isključivi. Verovatnoća događaja X se računa kao suma proizvoda verovatnoća svake hipoteze sa verovatnoćom događaja uz uslov ispravnosti hipoteza.

$$P(X) = \sum_{i=1}^{N} P(X/H_i)\mathrm{P}\,(H_i)$$

Formula za Bajesovu a'posteriornu verovatnoću je definisana na sledeći način:

$$P(H_i/X) = \frac{P(X/H_i)\,\mathrm{P}(H_i)}{\mathrm{P(X)}} = \frac{P(X/H_i)\,\mathrm{P}(H_i)}{\sum_{i=1}^{N} P(X/H_i)\,\mathrm{P}(H_i)}.$$

Bajesov klasifikator spada u grupu statističkih parametarskih klasifikatora. Vektor atributa se predstavlja kao stohastička promenljiva čija raspodela zavisi od klase uzoraka. Na osnovu toga se može koristiti Bajesova teorema u klasifikaciji.

Pretpostavimo da postoji skup od m uzoraka $S = \{S_1, \ldots, S_m\}$, gde je svaki uzorak S_i predstavljen kao n-dimenzionalni vektor $\{x_1, x_2, \ldots, x_n\}$ gde x_i predstavlja atribut uzorka.

Neka je definisano k klasa $k_1, k_2, \ldots, k_k$ i neka svaki uzorak pripada jednoj od ovih klasa. Zatim, neka je dat dodatni uzorak novi X za koji se ne zna kojoj klasi pripada. Verovatnoća da neki uzorak pripada određenoj klasi se izračunava na sledeći način:

$$P\,(k_j/X) = \frac{P(X/k_j)\mathrm{P}(k_j)}{\mathrm{P(X)}}$$

2.1.3 METODA PODRŽAVAJUĆIH VEKTORA (MPV)

Metoda podržavajućih vektora je tehnika automatskog generisanja klasifikatora. Predviđena je za klasifikaciju dve kategorije. Vremenom je razvijana tehnika i dobijeni su razni pristupi koji omogućavaju klasifikaciju više od dve kategorije. Metoda podržavajućih vektora je zasnovana na

ideji vektorskih prostora. Tako da je ova tehnika dobra za manipulisanje velikim brojem podataka, tj. veoma je korisna kada je broj dimenzija podataka veliki [17].

Primena MPV je zastupljena u raznim oblastima:

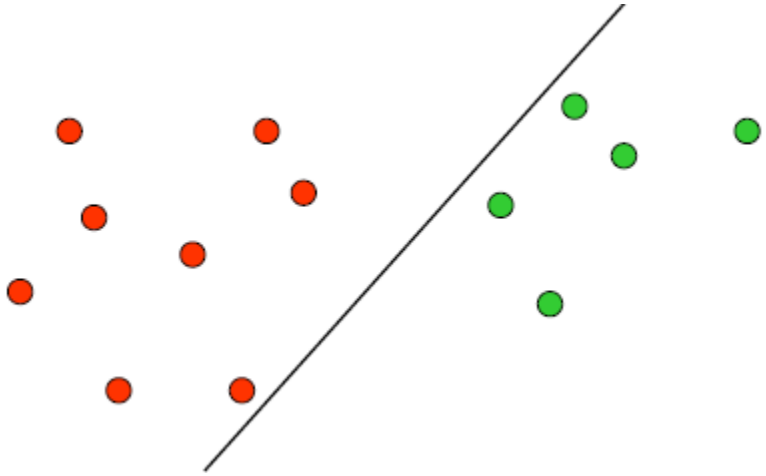
- U računarstvu za prepoznavanje zvuka
- U mobilnoj robotici za manipulaciju sa senzorskim informacijama
- Genetici, za simulaciju ponašanja novih generacija gena
- Za predviđanje tržišta u ekonomiji…

Metoda podržavajućih vektora je linearni klasifikator koji pronalazi hiperravan koja razdvaja dve klase. Može se pronaći beskonačno mnogo hiperravni. Potrebno je pronaci hiperravan koja najbolje razdvaja tj., klasifikuje dve klase.

Dve najznačajnije vrste MPV klasifikatora su:

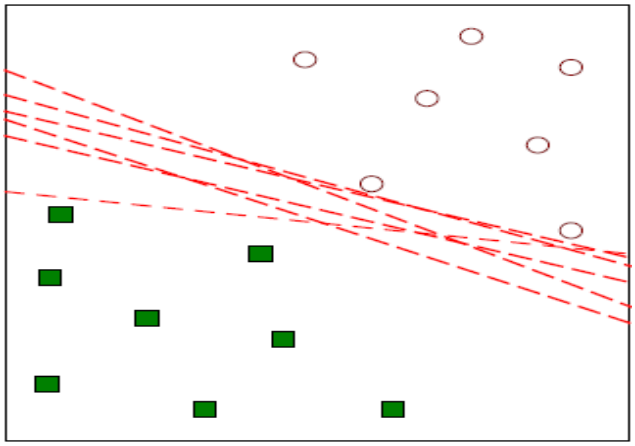
- Linearni klasifikator (separabilni i neseparabilni)
- Nelinearni klasifikator.

1) Linearni klasifikator predstavlja linearnu hiperravan koja razdvaja podatke u dve u klase. Ta hiperravan pomoću koje se vrši klasifikacija se zove još i granica odluke. Podaci iz iste klase se nalaze sa jedne strane hiperravni i razlikuju se od podataka iz druge klase koji se nalaze na suprotnoj strani hiperravni.



SLIKA 17: IZVOR [17]

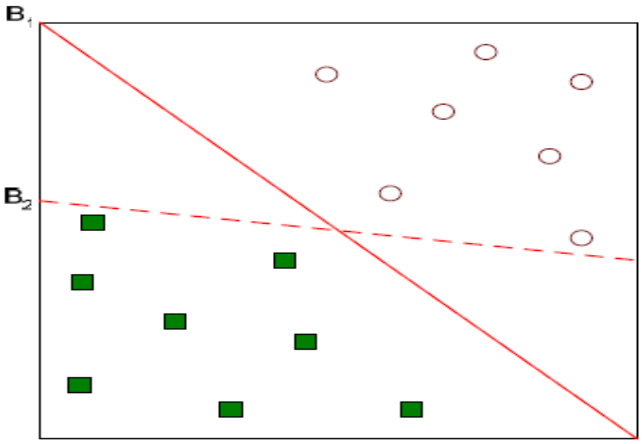
Prva faza je trening faza. U trening fazi vrši se pronalaženje optimalne ravni za razdvajanje podataka. Podaci se mogu razvojiti na više načina, i mnoge ravni mogu adekvatno razdvojiti podatke u klase ali je potrebno pronaći ravan koja najbolje klasifikuje podatke. Na sledećoj slici je prikazano da više različitih ravni isto klasifikuju podatke.



SLIKA 18: IZVOR [17]

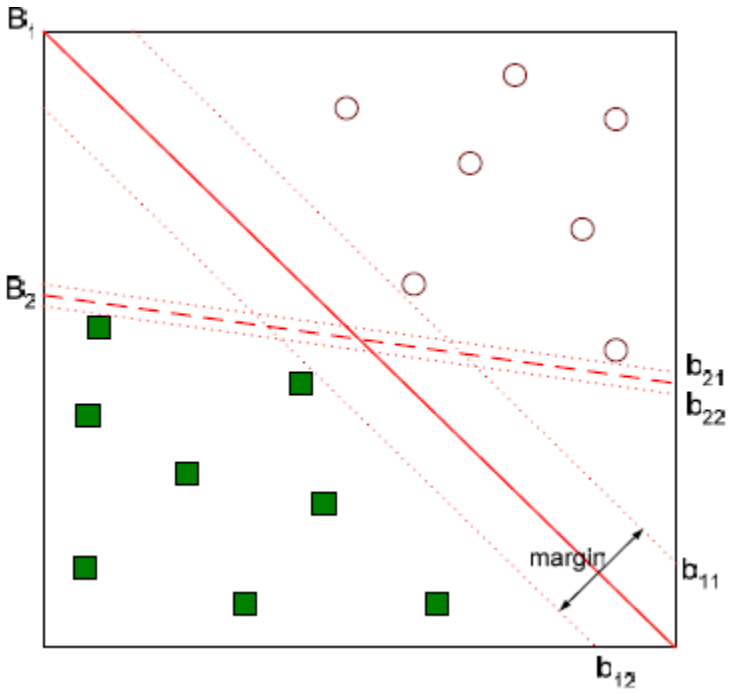
Cilj metode podržavajućih vektora je da od svih ravni nađe optimalnu ravan koja najbolje razdvaja podatke. To se postiže tako što se pronalazi hiperravan koja maksimizuje marginu određene ravni. U sledećim primerima se slikovito opisuje način na koji se dobija maksimizacija margine određene ravni.

Neka na proizvoljanom linearno separabilnom skupu podatak, nalaze dve ravni B_1 i B_2 koje dele podatke u dve klase na isti način.



SLIKA 19. IZVOR [17]

Sada je potrebno odrediti koja ravan je bolja za klasifikaciju, tj. maksimizovati margine ravni. Maksimizacijom margine ravni se dobija hiperravan. Hiperravan predstavlja najveće rastojanje od ravni do podataka u različitim klasama. Metoda podržavajućih vekotora predstavlja model koji maksimizuje hiperravan i tačke koje su blizu potencijalne linije razdvajanje. Na osnovu hiperravni se zaključuje koja se ravan koriti kao granica odluke B_1 ili B_2.



SLIKA 20: IZVOR [17]

Sa slike se vidi da je ravan B_1 bolja od ravni B_2 i da će se ona korisiti kao granica odluke na datom skupu podataka. Ravan B_1 je bolja od ravni B_2 zato što ima veću marginu, tj. veću hiperravan. Ravan koja razdvaja podatke u dve klase se može prikazati jednačinom:

$W * x + b = 0$, a W i b su parametri jednačine.

U daljem toku, se definiše postavka i rešavanje modela metode podržavajućih vektora.

Neka je dat prostor dimenzije d i neka je svaki podatak i predstavljen vektorom

$$x_i = (x^1, \dots, x^d).$$

Metoda podržavajućih vektora je binarni klasifikator i ona pripada skupu {-1, 1}

$$y_i \in \{-1,1\}.$$

Kao što je već rečeno hiperravan se dobija rešavanjem jednačine:

$$W * x + b = 0,$$

gde je *x* je skup od $\mathcal{N}$ trening podataka, a *W* i *b* su parametri koje je potrebno izračunati uz date uslove:

$$\min_{w} \frac{||W||^2}{2},$$

$$y_i(Wx_i + b) \geq 1$$

Parametri *W* i *b* se dobijaju rešavanjem jednačine Lagranžovih množitelja. Upotrebom jednačine Lagranžovih množitelja se dobija nova funkcija koja se minimizira:

$$L_p = \frac{1}{2}\left|\left|W\right|\right|2 - \sum_{i=1}^{N} \lambda_i(y_i(Wx_i + b) - 1)$$

λ_i su parametri Lagranžovih množitelja za koje važe sledeći uslovi:

$$\lambda_i \geq 0,$$

$$\lambda_i[y_i(Wx_i + b) - 1] = 0.$$

Vektor x_i se zovu vektori potpore i za njih važi $y_i^*(W^*x_i + b) = 1$. Vektori potpore x_i su vektori koji se nalaze na hiperravnima b_{i1} i b_{i2}.

Vrednost za λ_i se dobija rešavanjem parcijalnih izvoda.

$$\frac{\partial L_p}{W} = 0 \Rightarrow W = \sum_{i=1}^{N} \lambda_i y_i x_i$$

$$\frac{\partial L_p}{b} = 0 \Rightarrow \sum_{i=1}^{N} \lambda_i y_i = 0.$$

Rešavanjem parcijalnih izvoda i njihovim uvrštavanjem u Lagranžovu jednačinu dobija se:

$$L_p = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j \, y_i y_j x_i x_j .$$

Kada se odredi granica odluke na osnovu test podataka, novi podaci se klasifikuju na osnovu funkcije:

$$f(x) = sign\left(\left(\sum_{i=1}^{N} \lambda_i y_i \, x_i \, x\right) + b\right)$$

2) Linearno separabilni podaci su idealan slučaj za primenu metode podržavajućih vektora. Drugi, ne tako idealan, slučaj kod metode podržavajućih vektora je kada podaci nisu linearno separabilni. U stvarnosti često dolazi do greške u podacima, tako da dolazi do greške u klasifikaciji podataka. Problem se javlja što nekada ne može tačno da se odredi granica između dve klase, pa podaci iz jedne klase upadaju u drugu klasu. Zbog toga se kod linearno neseparabilnih podataka uvodi meka margina. Uvođenjem meke margine se tolerišu male greške pri klasifikaciji.

Neka je $\xi_i > 0$, $i = 0,\ldots, N$. Vrednosti ξ_i se zovu fiktivne promenljive, i predstavljaju grešku pri klasifikaciji. Sada se dobijaju novi uslovi za obučavanje linearnog neseparabilnog modela podržavajućih vektora:

$$\mathrm{W}x_i + b \geq 1 - \xi \; ako \; je \; y_i = 1$$

$$\mathrm{W}x_i + b \leq -1 + \xi \; ako \; je \; y_i = -1.$$

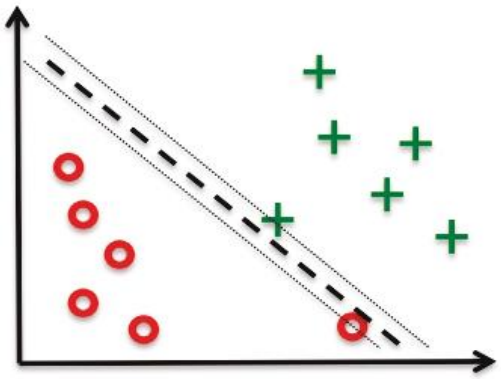
Problem se javlja kada je ξ_i veliko, i onda se dobijaju jako široke margine koje pogrešno klasifikuju veliki broj podataka. Da bi se izbegli ovakvi problemi uvodi se regularizacioni parametar *C*, koji smanjuje grešku klasifikacije, pa se dobija sledeća formula, na osnovu koje klasifikuju podaci:

$$f(W) = \frac{\| W \|^2}{2} + C\left(\sum_{i=1}^{N} \xi \mathrm{i}\right)^k$$

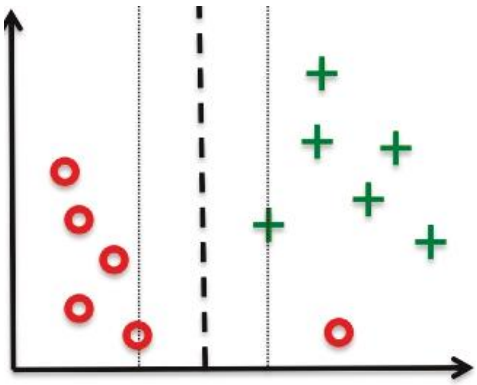
Za $k = 1$ dobija se funkcija Langražovih množitelja koja ima novi oblik:

$$L_p = \frac{1}{2}||W||2 - \sum_{i=1}^{N} \lambda_i (y_i (W x_i + b) - 1) + C\left(\sum_{i=1}^{N} \xi i\right)^k - \sum_{i=1}^{N} \xi i \mu_i$$

Na sledećim slikama se prikazuje kako regularizacioni parametar utiče na klasifikaciju podataka. Na prvoj slici parametar *C* uzima veliku vrednost i dobija se hiperravan koja vrši dobru klasifikaciju, dok se na drugoj slici parametru *C* dodeljuje mala vrednost koja ne utiče dovoljno dobro na optimizaciju granice odluke i dobija se lošija klasifikacija.



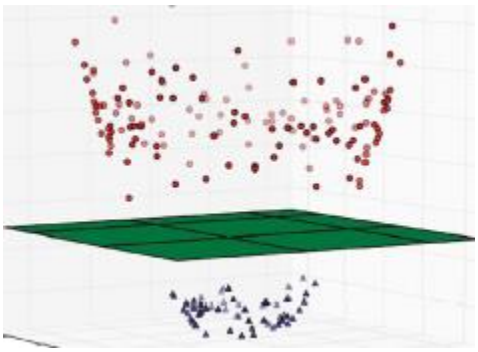
SLIKA 21: IZVOR [10]



SLIKA 22: IZVOR [10]

3) Pored linearnih MPV klasifikatora postoje i klasifikatori sa nelinearnom granicom odluke. Kada podatke nije moguće linearno razdvojiti u zadatom prostoru, podaci se preslikavaju u prostor sa većim brojem dimenzija. Tada se nelinearni separator transformiše linearnim, i onda se podaci mogu podeliti linearni separatorom Problem kod nelineaarnih klasifikatora je u tome što transformacija podataka računski veoma zahtevna. Na osnovu sledećih slika se može videti

osnovna ideja nelinearnih MPV klasifikatora i transformacija iz jednog prostora u drugi koja ima veći broj dimenzija.



SLIKA 23: IZVOR [10]

Dalje je potrebno definisati problem transformacije iz jednog prostora u drugi, i rešavanje problema linearnog modela sa novim uslovima.

Neka je $\Phi(x)$ funkcija koja služi za transformaciju.

Novi uslovi za rešavanje linearnog problema se zapisuju na sledeći način:

$$\min_{w} \frac{\| W \|^2}{2}$$

$$y_i(W\ \Phi(x_i) + b) \geq 1, \qquad i = 1, \dots, N$$

Dok jednačina Lagranžovih množitelja ima sledeći oblik:

$$L_D = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i\, \lambda_j y_i y_j\ \Phi(x_i)\ \Phi\left(x_j\right)$$

Prilikom implementacije nelinearnih MPV klasifikatora javljaju se sledeći problemi:

- Kako formirati funkciju Φ(x) da bi podaci u novom prostoru bili linearno separabilni,
- Ako je prostor velike dimenzinalnosti izračunavanje sklaranog proizvoda $\Phi(x_i) * \Phi\left(x_j\right)$ je veoma komplikovano i zahtevno.

Rešavanje ovih problema se vrši upotrebom Kernel funkcije, tj. funkcije jezgra.

Kernel funkcija je funkcija kojom se izražava skalarni proizvodi u nekom proširenom prostoru. Iz jednačine Lagranžovih množitelja sledi da nije potrebno tačno definisati funkciju $\Phi(x)$, nego je potrebno znati njihov skalarni proizvod $\Phi(x_i) * \Phi(x_j)$ za sve podatke iz prvobitno datog prostora. Zato se koristi Kernel funkcija, i pomoću nje se može izračunati taj skalarni proizvod. Veza između skalarnog proizvoda $x_i * x_j$ i $\Phi(x_i) * \Phi(x_j)$ se može zapisati na sledeći način:

$$K(x_i, x_j) = \Phi(x_i) * \Phi(x_j).$$

Na osnovu dosadašnjeg razmatranja se vidi da je metoda podržavajućih vektora veoma korisna i ima veoma dobre preformanse u klasifikaciji velikog broja podataka. MPV ima manju sklonost ka overfitting-u u odnosu na druge modele. Prednost MPV je i u mogućnosti upotrebe Kernela za rešavanje problema u izračunavanje skalarnog proizvoda.

Mana MPV klasifikacije je u tome što je znatno sporija od drugih metoda, posebno kod nelinearnih separatora gde korišćenje Kernel funkcije značajno usporava rad algoritma. Takođe, mana ovog modela je osetljivost na outlier-e kod separatora sa čvrstom marginom.

2.2 MERA SLIČNOSTI

Mera sličnosti predstavlja stvarno vrednovanu funkciju koja govori o sličnosti između dva objekta. U fudbalu mera sličnosti značajno doprinosi u merenju sličnosti između perioda posedovanja, a to su 3 mere distance za merenje sličnosti između perioda posedovanja [1]:

1) Dinamičko vremensko savijanje
2) Frečetovo rastojanje
3) Najduža zajednička distanca.

2.2.1 DINAMIČKO VREMENSKO SAVIJANJE

Dinamičko vremensko savijanje (DVS) je tehnika koja se koristi u analizi vremenskih serija. U suštini, DVS je metod koji se koristi za traženje najsličnih nizova iz skupa vremenskih serija. DVS

je jedno od najčešće korišćenih mera sličnosti koja omogućava merenje distance između dve vremenske serije koje variraju u brzini i dužini. Cilj DVS je pronalaženje optimalne povezanosti između dva niza. Ta optimalna povezanost se dobija tako što se traži najmanja putanja (optimalna kriva) u matrici između dva niza, vremenske serije [1].

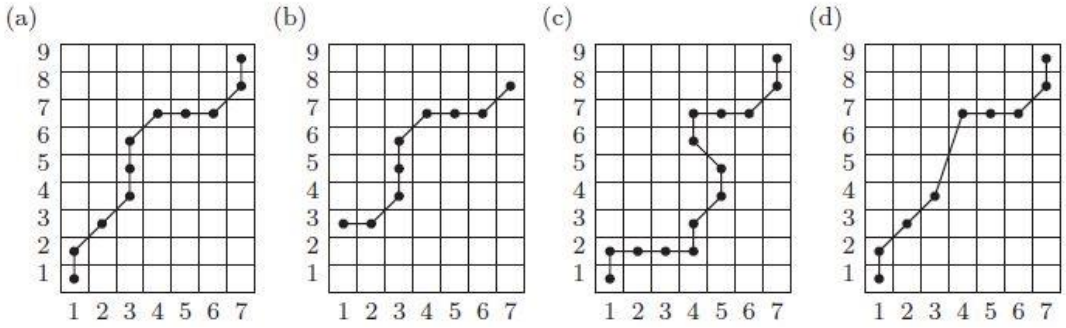
Radi lakšeg razumevanja, pretpostavimo da su nam date dve vremeske serije,

$$S_1 = [a_1, \dots, a_{l_1}] \text{ i } S_2 = [b_1, \dots, b_{l_2}].$$

Neka je $C \in \mathbb{R}^{l_1 x l_2}$ matrica, koja služi za poređenje ova dva niza, gde je $C(a_n, b_m) = \sqrt{(a_n - b_m)^2}$, $n = 1,\dots,l_1$ i $m = 1,\dots,l_2$. Matrica predstavlja Euklidsku distancu između svih tačaka oba niza. Izlomljena putanja, $W = [(w_1^1, w_1^2), \dots\dots, (w_L^1, w_L^2)]$ je putanja kroz matricu, za koju važe sledeći uslovi:

- granični uslov: $(w_1^1, w_1^2) = (1, 1)$ i $(w_L^1, w_L^2) = (l_1, l_2)$
- monotonost: $w_1^1 \leq w_2^1 \dots \leq w_L^1$ i $w_1^2 \leq w_2^2 \dots \leq w_L^2$
- uslov neprekidnosti: $w_{l+1} - w_l \in \{(1,0), (1,1), (0,1)\}$, $l = 1, \dots, L\text{-}1$.

Na slici 24. će se pokazati 4 primera izlomljene putanje u matrici.



SLIKA 24: IZVOR [1]

Na slici (a) je prikazana putanja u matrici koja zadovoljana sva tri uslova. Na slici (b) nije ispunjen granični uslov, na slici (c) nije ispunjen uslov monotonosti dok na slici (d) nije ispunjen uslov neprekidnosti.

Optimalna izlomljena kriva je putanja kroz matricu koja predstavlja minimalnu ukupnu vrednost nizova, koje se upoređuju u matrici. DVS razdaljina je određena optimalnom izlomljenom krivom:

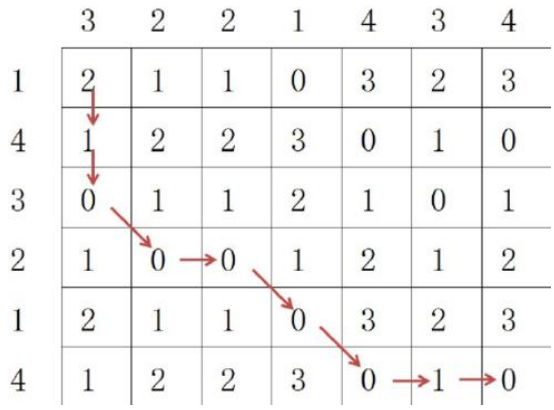
$$\text{DVS}(S_1, S_2) = \min\left\{\sum_{l=1}^{L}\sqrt{\left(a_{w_l^1} - b_{w_l^2}\right)^2} \mid W \textit{ je izlomljena kriva}\right\}.$$

Neka je $D(n, m) := \text{DSV}(S_1(1:n), S_2(1:m))$, gde je $S_1(1:n) = (a_1, \ldots, a_n)$ za $n = 1,\ldots,l_1$ i $S_2(1:m) = (b_1, \ldots, b_m)$ za $m = 1,\ldots,l_2$. Izlomljena putanja se onda može naći pomoću sledece jednačine:

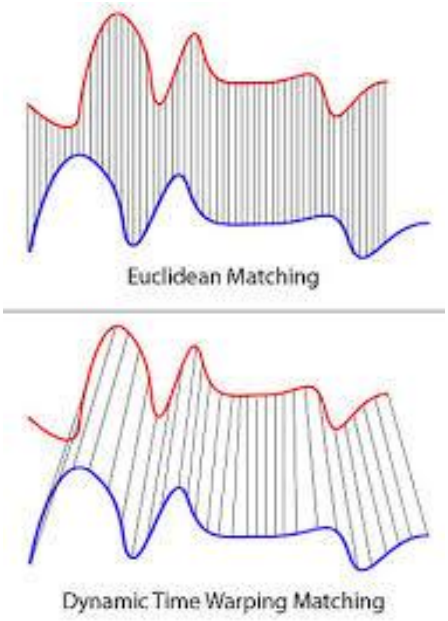
$$D(m, n) = C(a_n, b_m) + \min\{D(n\text{-}1, m\text{-}1), D(n\text{-}1, m), D(n, m\text{-}1)\}$$

$$\text{za } 1 < n < l_1 \text{ i } 1 < m < l_2.$$

Slika 25. predstavlja, matricu u kojoj se nalazi optimalna izlomljena putanja. Prvo se izračunavaju vrednosti u matrici pomoću prethodno navedene jednačine, posle se odredi ukupna minimalna vrednost, koja predstavlja optimalnu izlomljenu krivu.



SLIKA 25: IZVOR [15]



SLIKA 26: IZVOR [15]

Na slici 26. je predstavljena razlika između Euklidske razdaljine i DVS razdaljine između dva niza. DVS daje veću snagu izračunavanja sličnosti. Pomoću DVS se mogu porediti nizovi različite dužine. DVS razdaljina omogućava poređenje više tačaka iz jednog niza sa jednom tačkom drugog niza, dok Euklidska razdaljina može da poredi samo par-tačaka iz nizova u određenom vremenskom trenutku.

2.2.2 FREČETOVA MERA DISTANCE

Frečetova mera distance je mera razdaljine koja meri sličnosti između krivih koja uzima u obzir lokaciju i redosled tačaka na krivama. Da bi se lakše objasnila Frečetova distanca koristiće se primer kretanja psa i njegovog vlasinka. Tokom kretanja pas ide jednom putanjom i vlasnik drugom putanjom. Oni su povezani povocem. Oboje, i pas i vlasnik se kreću neprekidno praveći različite putanje, tj. krive, koje imaju svoju početnu i krajnju tačku. Kretanje mora biti isključivo unapred. Brzina kretanja može da varira. Frečetova distanca te dve putanje predstavlja najmanju dužinu povoca koji je dovoljan da se spoje te dve putanje.

Frečetovom distancom se želi videti u kojoj meri ona određuje sličnosti između perioda posedovanja [12].

2.2.3 NAJDUŽI ZAJEDNIČKI PODNIZ

Najduži zajednički podniz (NZP) je algoritamski problem u kojem je potrebno odrediti najduži podniz koji je zajednički svim nizovima u skupu nizova. Podniz nekog niza se dobija tako što se uklanja određeni broj simbola, a oni koji ostaju zadržavaju svoj originalni međusobni redosled [12].

Na primer, imamo dva niza, X = (N, E, N, A, D) i Y = (N, N, A, B), najduži zajednički podniz ova dva niza je Z = (N, N, A).

NZP je još jedan način na koji se može meriti sličnost dva niza. Ovom metodom se traži broj zajedničkih tačaka nizova. Vrednost koja se dobija da bi se procenila sličnost dva niza je između 0 i 1. Što je dobijana vrednost veća, tj. bliža 1 to su dva niza sličnija. Mera sličnosti (mera distance) između dve sekvence (P i Q) se dobija na sledeći način:

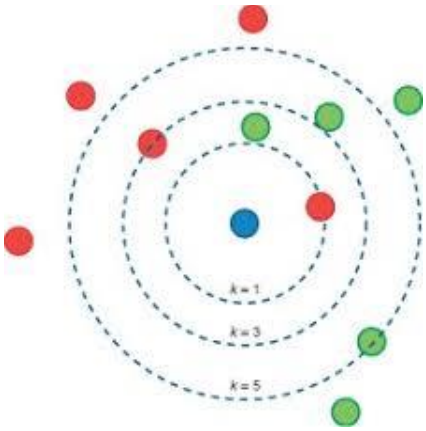
$$D(P,Q) = 1 - \frac{\#"zajedničke"\ tačke\ u\ P\ i\ Q}{\min\{dužina(P), dužina(Q)\}}.$$

2.3 K-NAJBLIŽI SUSEDI

Algoritam *k*-najbližih suseda je neparametarski algoritam koji se koristi za klasifikaciju, gde se za svaki objekat traži *k*-najbližih objekata. To je algoritam koji se koristi za otkrivanje zakonitosti u podacima, kada je potrebno definisati ishod nekog problema. Kada je potrebno objekat dodeliti nekoj klasi, potrebno je posmatrati *k*-najbližih suseda tog objekta i na osnovu njihovih sličnosti doneti zaključak o klasi novog objekta. Ako je $k = 1$ onda se objekat dodeljuje klasi njemu najbližeg suseda. Na sledećem primeru će se objasniti kako funkcioniše algoritam *k*-najbližih suseda [16].

Potrebno je da se proceni da li će se novi igrač uklopiti u ekipu (klasifikacija). Na osnovu nekih osobina igrača koje predstavljaju ulazne atribute, dobijaju se izlazni atributi, koji govore da li se igrač uklopio u ekipu ili nije. Na slici 27, crveni krug govori da se igrač nije uklopio u ekipu a zeleni da jeste. Pri donošenju odluke da li će se novi igrač uklopiti ili ne, posmatramo njegove najbliže susede. Za $k = 1$ novi igrač se neće uklopiti u novu ekipu, jer mu je crveni krug najbliži.

Za k = 3 se takođe neće uklopiti, jer se u skupu najbližih suseda nalaze 2 crvena i 1 zeleni krug, a ako se uzme k = 5, novi igrač će se uklopiti u ekipu jer se skup od 5-najbližih suseda sastoji od 3 zelena i 2 crvena kruga.



SLIKA 27: IZVOR [16]

3. OPIS PODATAKA

U ovom poglavlju se govori o skupu podataka koji se koristi u istraživanju i način kako kombinujemo te podatke. Posebna pažnja se posvećuje lokaciji gde se određeno dodavanje dešava. Jako bitno je da li se to desilo na svojoj ili protivničkoj polovini. Važni skupovi podataka koji se objašnjavaju a koji su važni u istraživanjima: trening, validacioni i test skupovi podataka.

3.1 SKUPOVI PODATAKA

Jedan od najvažnijih i najobimnijih skupova podataka je onaj koji sadrži podatke o događajima koji se dešavaju tokom utakmice. Skup podataka sadrži sledeće događaje: dodavanje, šut, oduzimanje lopte, faul, aut, korner, odbrana golmana, penal, vreme dešavanja događaja, deo terena gde se događaj desio, da li je postignut gol... Tokom praćenja i korišćenja ovih podataka značajne su nam i informacije koje opisuju neki određeni tip podataka. Na primer, za dodavanje je bitna lokacija, da li je dodavanje asistencija, završni pas ili neko drugo obično dodavanje. Kod događaja koji se deklariše kao šansa za gol bitno je: kojim delom tela je upućen udarac ka golu da li je šut izblokiran, da li je gol, da li je golman odbranio, da li je lopta pogodila okvir gola ili je otišla van okvira. Ovakva vrsta podataka, koja obeleži jednu fudbalsku utakmicu se prikuplja manuelno od strane ljudi, pa postoji mogućnost greške ali je ona zanemarljivo mala.

Takođe, važni podaci za jednu fudbalsku utakmicu su: sastav ekipa (koji igrači počinju utakmicu i koji igrači sede na klupi za izmene), formacija ekipa (4-3-3, 4-2-3-1, 3-5-2…), izmene koje su izvršene u toku utakmice i na kraju rezultat utakmice. Kombinacijom sastava, formacije i izmena dobijamo minutažu i poziciju svakog igrača. Pozicija svakog pojedinačnog igrača se izračunava na osnovu prosečne pozicije koji taj igrač provede na određenom delu terena. U današnje vreme, velika pažnja se posvećuje statistici i brojevima ostvarenim na nekoj utakmici. Takođe, prati se i koliko su igrači pretrčali na jednoj utakmici, koliko je uspešnih a koliko neuspešnih dodavanja ostvarila neka ekipa. Ide se i do takvih detalja da se prati u kojem minutu određena ekipa najčešce postiže gol, koliko je prošečno šansi potrebno ekipi da bi se postigao gol.

Podaci o učinku igrača na utakmici su veoma značajni. Na osnovu tih statističkih podataka prate se individualne karakteristike igrača, njihova dodavačka sposobnost i druge veštine koje poseduje taj igrač. Glavni sajt za praćenje transfer vrednosti igrača je Transfermarket (https://www.transfermarkt.com/).

Iz dosadašnjeg izlaganja smo videli da je rangiranje igrača veoma korisno u raznim procenama i predviđanjima u fudbalu. Pored rangiranja igrača jako je bitno i rangiranje fudbalskih klubova. Igrajući u svojim ligama i na međunarodnoj sceni, fudbalski klubovi dobijaju određne bodove odnosno koeficijente. Ti koeficijenti zavise od rezultata koje je određeni fudbalski klub ostvario i u zavisnosti da li igra sa slabijim ili jačim protivnikom. Iz utakmice u utakmicu se menjaju ti koeficijenti, što nam govori i u kakvoj rezultatskoj formi se nalazi neka ekipa. ELO rangiranje klubova predstavlja rangiranje svih fudbalskih klubova na svetu. ELO rangiranje je preciznije u dodeljivanju koeficijenta fudbalskim ekipama, i postepeno zamenjuje FIFA rangiranje. Računanje koeficijenata se vrši tako da zbir bodova ekipa koje međusobno igraju ostaje isti, samo što se jednima dodaje u slučaju pobede a drugima oduzima isti taj broj bodova. Rangiranje klubova je usko povezano i sa igračima. Što je klub bolje rangiran, to znači da ima kvalitetniji tim, odnosno da ima bolje igrače. Ako tim ima bolje igrače, njihova tržišna vrednost je veća.

3.2 PODELA SKUPA PODATAKA

Prvo je potrebno sakupiti podatke, koji se kasnije analiziraju. Analizirani i usklađeni podaci služe za građenje predikcionog modela. Radi efikasnije procene podaci se dele u tri skupa, a to su: trening skupovi podataka, validacioni skupovi i test skupovi. Raspodela podataka se uglavnom vrši na sledeći način: 60% podataka čini trening skup, 20% validacioni skup i 20% test skup. Ovi skupovi podataka, kao i njihovo značenje su detaljnije opisani u delu gde se govori o metodama mašinskog učenja.

4. FUDBALSKA ANALIZA

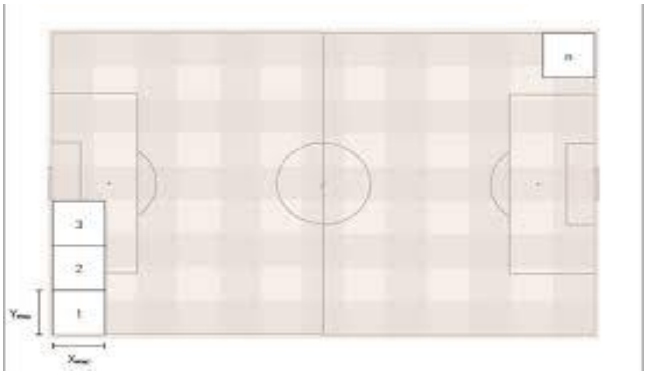
Radi lakše analize podataka potrebno je analizirati i podeliti fudbalski teren. Fudbalski tereni mogu biti različitih dimenzija, zato je potrebno naći univerzalnu jedinicu za separaciju terena na zone. Raščlanjivanjem terena na zone omogućeno je lakše praćenje parametara i elemenata koji olakšavaju izračunavanje statističkih i matematičkih kalkulacija za analizu klasifikacije. Posedovanje lopte je značajan segment u fudbalskoj analizi. Zato se obraća velika pažnja na period posedovanja lopte u fudbalu.

4.1 PODELA TERENA

Fudbalski tereni nisu istih dimenzija. Dužina terena varira u rasponu od 90 metara do 120 metara, a širina od 64 do 75 metara. Standardna dimenzija fudbalskog terena u internacionalnim utakmicama je: dužina 105 metara a širina 68 metara. Radi lakšeg prikupljanja i računanja podataka fudbalski teren je podeljen u zone jednake veličine. Sa X se obležava dužina zone a sa Y se obeležava širina zone. Fudbalski teren se može podeliti na n jednakih zona. Broj jednakih zona se računa na sledeći način:

$$n = \frac{105}{X} * \frac{68}{Y}.$$

Slika 28 pokazuje kako je teren podeljen na n jednakih zona.



SLIKA 28: IZVOR [1]

Na primer, ako je teren podeljen na 18 jednakih zona, tada je $n = 18$, a način podele terena na 18 zona je prikazan na slici 29.

SLIKA 29: IZVOR [3]

Važno je naglasiti da se teren deli u zone radi lakšeg sakupljanja podataka i klasifikacije. Na osnovu podele terena može se prikazati lokacija dodavanja, po kojoj strani terena ekipa najviše napada i u kom delu terena se najbolje snalazi. Tako se može se odrediti kretanje igrača i taktička zamisao ekipe. Na osnovu slike 29 se vidi, da su zone 1-6 defanzivne zone, od 7-12 zone sredine terena, a 13-18 su napadačke zone. Najopasnije zone, odnosno zone iz kojih je najveća mogućnost postizanja gola su 14 i 17.

4.2 PERIOD POSEDOVANJA

Period posedovanja je period uzastopnog dodavanja lopte pod kontrolom jednog tima, koji počinje ubacivanjem lopte iz auta, korner, posle faula…, a završava se šutem u gol, faulom, izlaskom lopte iz terena… Period posedovanja predstavlja vremenski period gde se lopta nalazi u posedu jedne ekipe, sve dok ona ne pređe u posed protivnčke ekipe. Period posedovanja je bitna karakteristika u razmatranju jedne fudbalske ekipe. Gol može biti postignut nakon više uzastopnih dodavanja, nakon izvođenja slobodnog udarca, greške protivničkog igrača. Suštinski, način postizanja gola nije krucijalna činjenica, ali ako se ti događaji pre postizanja gola učestalo ponavaljaju to nam je veoma značajno jer govori o karakteristikama tima. To se može prikazati na primeru Barselonine igre. Oni često postižu gol posle kratkih uzastopnih dodavanja, što znači da se taktika bazira na pas igri u ofanzivi.

Dakle, na osnovu prethodnog, jedna fudbalska utakmica se može posmatrati po periodima posedovanja a pojedinačni period posedovanja se može posmatrati kao niz određenih dešavanja (dodavanje, faul, korner ,…). Zbog preciznosti u radu, treba tačno definisati period posedovanja.

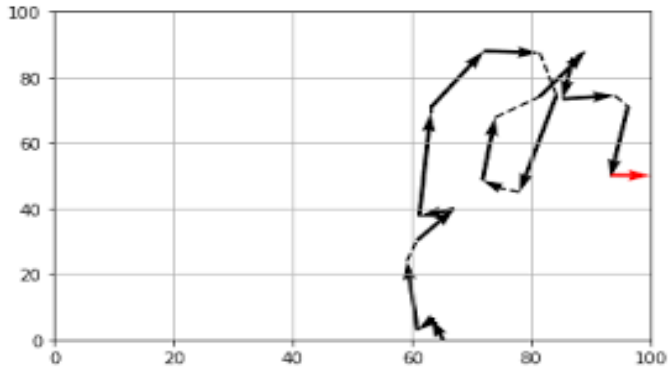
Period posedovanja počinje i završava se sledećim događajima:

- početak i kraj perioda utakmice (prvo poluvreme, drugo poluvreme, produžetak)
- lopta izlazi van terena
- protivnički tim dotakne loptu
- postigne se gol.

Ishod, odnosno način završetka, perioda posedovanja je takođe jedna od karakteristika koja nam govori o kvalitetu jednog fudbalskog tima ali i o toku fudbalske utakmice. Zato je potrebno definistati i ishod tog perioda. Ishod perioda posedovanja može biti sledeći:

- gubitak poseda lopte
- određeni period utakmice je završen
- desio se pokušaj za gol (gol je postignut ili nije)
- načinjen je faul i dosuđen je slobodan udarac
- loptu je protivnik izbacio sa terena i ekipi koja je imala posed dosuđen je aut ili korner.

Na slici 30 je prikazan period posedovanja koji počinje sa ubacivanjem lopte u teren iz auta a završava se golom.



SLIKA 30: IZVOR [1]

Potrebno je definisati i vrednosti za ishode perioda posedovanja. Kada se izgubi lopta, odnosno kada dođe do prekida poseda jedne ekipe, dodeljuje se nula periodu posedovanja. Vrednost nula, se uzima i kada lopta izađe sa terena, kada se desi faul. Vrednost nula se dodeljuje za završetak perioda posedovanja i početak novog perioda.

Ishod perioda posedovanja koji ima najveću vrednost, je kada dođe do pokušaja da se postigne gol. Vrednovanje, pokušaja da se postigne gol, se vrši pomoću očekivane gol metrike, koja je često korišćena metrika u fudbalskoj analizia uveo ju je Caley (2013). Ova metrika uzima u obzir različite parametere kako bi se izračunala verovatnoća da određeni broj pokušaja bude pretvoren u postignut gol. Najčešće korišćeni parametri za dobijanje modela očekivanog postizanja gola su distanca sa koje se upućuje šut i ugao u odnosu na gol odakle je šut upućen. Postoji još mnogo parametara koji se koriste za dobijanje modela. Parametri se koriste na osnovu određenih situacija iz igre kao što su korner, slobodan udarac, penal, odbitak. Takođe se mogu koristiti i parametri koji govore na koji način je šut upućen (nogom, glavom, nekim drugim delom tela).

Da bi se napravio model za dodeljivanje vrednosti za dodavanje, prvo se napravi model očekivanog gola, koji vrednuje svaku priliku za postizanje gola. Dodavanjima, koja dovode do prilike za postizanje gola, se mogu odrediti vrednosti. Problemi mogu da se jave kada se uzme u obzir samo postignut gol, jer se gube važne informacije. Igrač koji odigra perfektan završni pas a njegov saigrač ne postigne gol, ne dobija adekvatne vrednost. Nasuprot tome, ako igrač postigne gol iz teške situacije onda igrač koji mu je odigrao loptu dobija veću vrednost nego što zaslužuje.

Parametri za određivanje očekivanog gol modela se zasnivaju na lokaciji gde se stvorila prilika za postizanje gola.

5. PRISTUPI

U ovom delu se razmatraju pristupi koji omogućavaju lakše vrednovanje dodavanja. Uvode se tri pristupa za vrednovanje: zonski-orijentisano vrednovanje (ZOV), pas-orijentisano vrednovanje (POV) i nizovno-orijentisano vrednovanje (NOV).

- ZOV koristi vrednosti posedovanja lopte u određenim zonama na terenu kako bi se vrednovalo dodavanje.
- POV se odnosi na pojedinačna dodavanja. Meri sličnost između pojedinačnih dodavanja koristeći mere distance.
- Dodavanje kod NOV se vrednuje na osnovu njegovog uticaja na ishod perioda posedovanja.

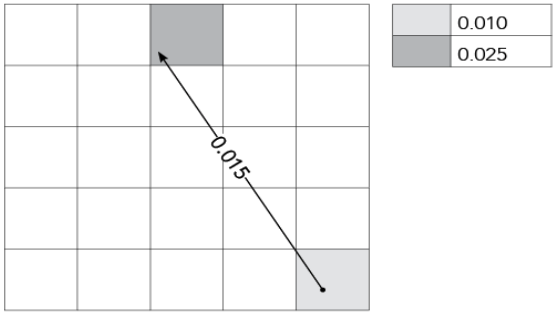
5.1 ZONSKI-ORIJENTISANO VREDNOVANJE DODAVANJA

Ovaj pristup deli teren na zone. Te zone moraju biti jednake veličine. Svakoj zoni se dodeljuje određena vrednost koja se izračunava pomoću dva načina:

- pravilo 15 sekundi
- model očekivanog gola.

Pravilo 15 sekundi govori da je svaka zona vredovana tako da se pretpostavlja da je gol postignut iz određene zone u roku od 15 sekundi. Na primer, neka je fudbalski teren podeljen na 18 zona. Lopta se nalazi u zoni 14 i roku od 15 sekundi gol mora biti postignut da bi važilo ovo pravilo. Model očekivanog gola predstavlja vrednovanje svake zone pod pretpostavkom da svaki šut koji bude upućen iz te zone bude realizovan kao pogodak.

Za oba ova pravila, dodavanje se vrednuje na isti način, tako što od vrednosti krajnje zone oduzmemo vrednost zone iz koje dolazi dodavanje. Slika 31 prikazuje ZOV pristup.



SLIKA 31: IZVOR [1]

Dodavanje dato iz svetlo sive zone (poreklo dodavanja) koje se završava u tamno sivoj zoni (destinacija dodavanja) ima vrednost 0.015, tako što od vrednosti destinacije oduzmemo vrednost porekla dodavanja: 0.025 - 0.010 = 0.015.

Definisana su tri koraka koja se služe za izračunavanje vrednosti dodavanja ZOV pristupom.

1. Deljenje terena u zone
2. Određivanje vrednosti zone koristeći trening podatke, pravila 15 sekundi ili modela očekivanog gola.
3. Za svako dodavanje iz test skupa podataka određujemo njegovu ZOV, tako što oduzmemo krajnju lokaciju dodavanja i početnu lokaciju.

5.1.1 PODELA TERENA

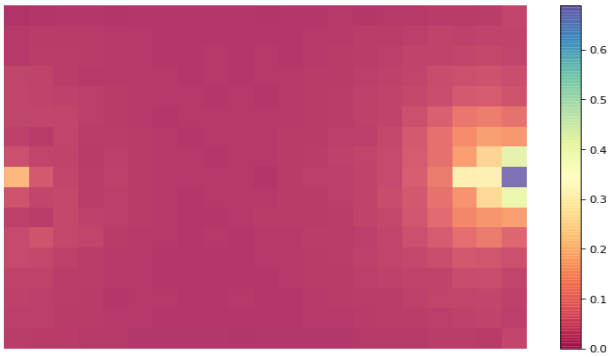
Teren se uglavnom deli na neparan broj zona po širini i dužini. Sa takvom podelom dobijamo zone koje obuhvataju liniju sredine terena, centar terena i zonu koje se nalaze tačno ispred gola. Uzima se u obzir dve dimenzije zona. Jedan dimenzija je veličine 5 x 4 metara a za drugu dimenziju se uzima veličina 1 x 1 metar. U zavisnosti koju od dimenzija uzmemo imaćemo i različit broj zona. Dimenzije zona se biraju na osnovu boljeg rezultata koji se dobija na osnovu dva već pomenuta pravila, pravilo 15 sekundi i model očekivanog gola.

5.1.2 VREDNOVANJE ZONE KORISTEĆI TRENING PODATKE

Kao što je već rečeno, pravilo 15 sekundi je definisano kao verovatnoća da gol bude postignut iz neke zone u vremenskom rasponu od 15 sekundi, tj. da gol bude postignut u roku od 15 sekundi od kada je lopta bila u određenoj zoni. Neka je G_{z_i} ukupan broj postignutih golova za koji važi ovo pravilo, gde $z_i, i = 1, \dots, n$ predstavlja zonu iz koje je gol postignut. Neka T_{z_i} označava ukupan broj pokušaja da se postigne gol iz zone $z_i, i = 1, \dots, n$. Na osnovu broja pokušaja i broja postignutih golova može se izračunati vrednost te zone. Vrednost određene zone z_i se računa na sledeći način:

$$V(15) = \frac{G_{z_i}}{T_{z_i}}$$

Kada bi upoređivali zone, na koje je teren podeljen, videlo bi se da zone koje su oko protivničkog gola imaju veće vrednosti u odnosu na zone koje su udaljenije, jer je u tim zonama veća verovatnoća za postizanje gola u roku od 15 sekundi. Još jedna zona koja ima neočekivano veću vrednost je zona ispred sopstvenog gola. To je zato što kada golman ili neko od igrača ispucaju oduzetu loptu na protivničku polovinu, a protivnička ekipa nije uspela da se dobro poziciono postavi na terenu, započinje se kontra-napad koji može dovesti do postizanja gola u narednih 15 sekundi. Vrednosti zona na koje je teren podeljen su prikazane na slici 32.



SLIKA 32: IZVOR [1]

Drugi metod izračunavanja vrednosti zona je model očekivanog gola. Model očekivanog gola predstavlja verovatnoću da gol bude postignut kada se iz neke zone $z_{i,}\ za\ i = 1, \dots, n,$ uputi udarac ka golu. Sa OGM (z_i) će se označiti vrednost zone z_i koja se dobija pomoću ovog modela.

5.1.3 ODREĐIVANJE VREDNOSTI DODAVANJA

Dodavanje iz test skupa podataka obeležavamo sa p_j. Zona iz koje je dato dodavanje se zove zona porekla i obeležavamo je sa z_{o_j}, a zona u kojoj se završava dodavanje se zove zona destinacije i obeležava se sa z_{d_j}. Na sledeći način se izračunava dodavanje p_j:

$$\text{ZOV15}_{p_j} = \text{V15}(z_{d_j}) - \text{V15}(z_{o_j})$$

$$\text{ZOVOG}M_{p_j} = \text{OGM}(z_{d_j}) - \text{OGM}(z_{o_j}),$$

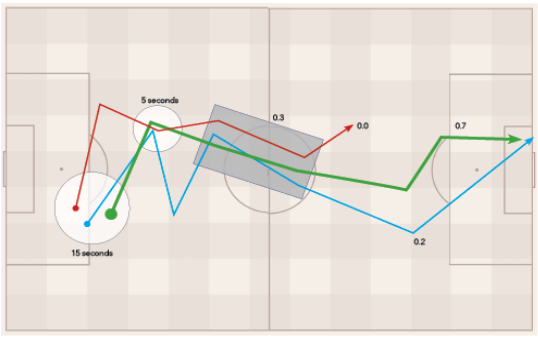
gde ZOV15_{p_j} predstavalja vrednost dodavanja p_j koristeći pravilo 15 sekundi, a $\text{ZOVOG}M_{p_j}$ predstavalja vrednost dodavanja p_j koristeći model očekivanog gola.

5.2 PAS-ORIJENTISANO VREDNOVANJE DODAVANJA

POV pristup se zasniva na merenju sličnosti između dodavanja. Tehnike koje se koriste za merenje sličnosti su mere distance. Neke mere distance su već spomenute u prethodnim oblastima. Dalje se vrednuje dodavanje na osnovu prosečne vrednosti ishoda perioda posedovanja najsličnijih dodavanja. Definisana su četiri koraka iz kojih se sastoji POV pristup, koja dovode do izračunavanja vrednosti dodavanja koristeći:

1. Definisati meru distance za merenje sličnosti dodavanja
2. Za svako dodavanje iz trening skupa određujemo ishode perioda posedovanja kojem to dodavanje pripada
3. Podelimo dodavanja iz trening skupa preko različitih pod-klastera uzimajući u obzir njihovo poreklo i destinaciju
4. Za svako dodavanje iz test skupa određujemo njegov pod-klaster i nalazimo *k*-najbliže susede u pod-klasteru koji se određuju na osnovu mere distance. Dodavanju je dodeljena prosečna vrednost ishoda susednih perioda posedovanja.

Sledeća slika će ilustovari uopšteni princip rada POV pristupa.



SLIKA 33: IZVOR [1]

Na slici 33 su predstavljena 3 perioda posedovanja, koja se sastoje od više različitih dodavanja. Crvenom bojom je predstavljen period posedovanja koji se završava gubitkom poseda i zato je ishod ovog perioda jednak 0. Zeleni i plavi period posedovanja se završavaju udarcem ka golu i dobijaju se očekivane gol vrednosti, 0,7 za zeleni period i 0,2 za plavi period posedovanja. Vrednost dodavanja u sivom pravougaoniku je 0,3. Ta vrednost se dobija kada se izračuna prosečna vrednost ishoda (PVI) perioda posedovanja,

$$PVI = (0+0,7+0,2)/3 = 0,3.$$

Dodavanja u sivom praougaoniku imaju slične geometrijske karakteristike a lopta se nalazi u skoro istoj lokaciji 5 i 15 sekundi pre nego što su dodavanja u sivom pavougaoniku bila izvršena. Zbog toga su ova 3 dodavanja slična i dodeljuje im se izračunata vrednost od 0,3.

5.2.1 DEFINISANJE MERE DISTANCE

Mera distance, kao što je već rečeno, služi za merenje sličnosti između individualnih dodavanja, da se za svako dodavanje pronađe njemu najsličnija dodavanja. Problem kod merenja sličnosti je u tome što se samo na osnovu podataka jako teško može izmeriti sličnost između dva dodavanja, jer se ne uključuje pozicija igrača na terenu i da li su dobro pozicionirani na terenu. Na primer, kada je pitanju kontra-napad igrači protivničke ekipe nisu formacijski dobro pozicionirani, u odnosu kada se lopta ubacuje posle gol-auta gde protivnički igrači uspevaju da se formacijski pozicioniraju. Zbog toga se, upoređuju samo dodavanja koja su usledila iz kontra-napada. Da bi

se sličnost između dodavanja što bolje izmerila uvodi se nova mera distance. Ona se sastoji od sledećih 5 svojstava:

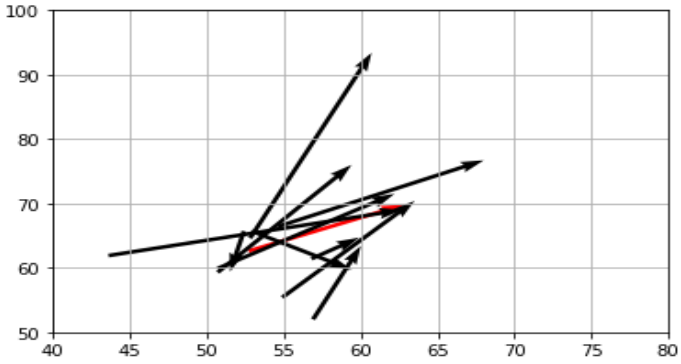
1. Razlika u dužini dodavanja (w_l)
2. Euklidska razdaljina između porekla dodavanja (w_o)
3. Euklidska razdaljina između destinacije dodavanja (w_d)
4. Lokacija lopte 5 sekundi pre nego što je dodavanje izvršeno (w_5)
5. Lokacija lopte 15 sekundi pre nego što je dodavanje izvršeno (w_{15})

Pored ovih svojstava potrebno je još uvesti i težinske koeficijente da bi se moglo izvršiti merenje distance. Težinski koeficijenti nam određuju koliki je udeo iliti koliku značajnost ima neki od ovih 5 svojstava u određivanju mere distance za dodavanja. Prvo je potrebno analizirati i uporediti težinske koeficijente na trening skupu i onda na osnovu analiza, koje vrše stručna lica, izberemo koeficijente koji se bolje prilagođavanju, bolje određuju i pronalaze slična dodavanja. U sledećoj tabeli su data dva skupa težinskih koeficijenata koja se analiziraju i bira se onaj skup koji daje bolje rezultate.

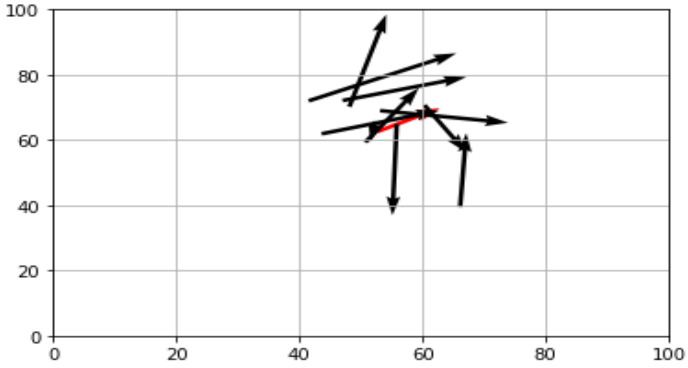
	w_0	w_d	w_l	w_5	w_{15}
W_1	¼	1/4	1/4	1/6	1/12
W_2	1/5	1/5	1/5	1/5	1/5

Tabela 1: Izvor [1]

U prvom skupu W_1, veća važnost se pridaje geometriskim karakteristikama, poreklu, destinaciji i dužini dodavanja. Zbog toga oni imaju imaju veće koeficijente u odnosu na lokaciju lopte 5 i 15 sekundi pre nego što je dodavanje izvršeno. Iz skupa W_1, na osnovu težinskih koeficijenta se vidi da lokacija lopte 5 sekundi pre dodavanja više govori od dodavanju nego lokacija lopte 15 sekundi pre dodavanja. U drugom skupu W_2, imamo jednake koeficijente, što govori da svako svojstvo podjednako utiče na određivanje mere distance. Analizom ova dva skupa težinskih koeficijenata dobija se da je W_1 bolje prilagodljiv. To znači da W_1 daje bolji rezultat u određivanju mere distance. Ta mera distance bolje upoređuje dodavanja i omogućava preciznije određivanje sličnih dodavanja. Slike 34 i 35, pokazuju 10 najbližih suseda istog dodavanja, gde su korišćeni W_1 i W_2 težinski skupovi za određivanje mere distance.



SLIKA 34: IZVOR [1]



SLIKA 35: IZVOR [1]

Na slici 34, crvenom bojom je označeno dodavanje u odnosu na koje se meri distanca i određuje 11 najbližih suseda. Ova mera distance koristi skup W_1 za težinske koeficijenate. Na slici 35, je za to isto dodavanje prikazano 10 najbližih suseda ali se za meru distance koristi W_2 težinski skup koeficijenta. Posmatrajući ove dve slike i rezultate dobijene korišćenjem ova dva težinska skupa vidi se da W_1 daje bolje rezultate, kao što se i ranije zaključilo. Na slici 34, se dobijaju manje distance između uporedivih dodavanja i sličnija su, dok na slici 35 dodavanja su malo razbacanija i različitija u odnosu na dodavanje označeno crvenom bojom.

5.2.2 ODREĐIVANJE ISHODA PERIODA POSEDOVANJA

U ovom koraku se govori da je za svako dodavanje iz trening skupa potrebno odrediti ishod perioda posedovanja kojem to dodavanje pripada. Ishod perioda posedovanja zavisi od toga da li se on završio udarcem na gol i ako jeste koja je očekivana vrednost postizanja gola. Na ovaj način svako dodavanje ima vrednost a to je vrednost njegovog perioda posedovanja.

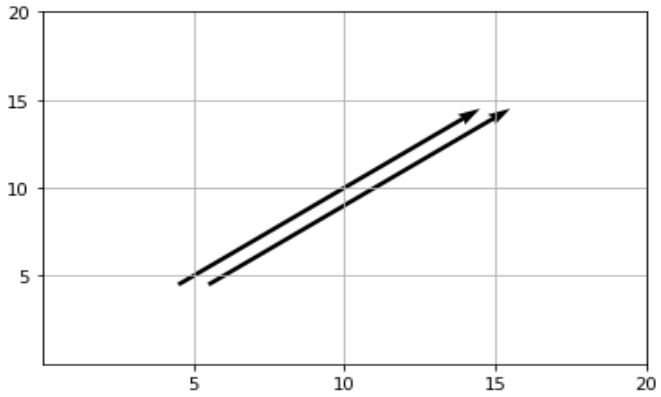
5.2.3 POD-KLASTERIZACIJA DODAVANJA

Računanje distance između velikog broja dodavanje nije moguće izvršiti u nekom razumnom vremenskom roku. Određivanje sličnosti među milionima dodavanja je nepraktično i oduzima puno vremena. Zbog toga je potrebno uvesti pod-klastere. Dodavanja se dodeljuju različitim pod-klasterima. Na taj način lakše izračunavamo distance između dodavanja i određujemo sličnost između njih.

Pod-klasteri se prave samo na osnovu porekla i destinacije dodavanja. Uzimaju se u obzir zone gde se nalaze poreklo i destinacija dodavanja i onda se ta dodavanja dodeljuju određenim pod-klasterima. Na primer, ako dva dodavanja imaju sličnu zonu porekla i sličnu zonu destinacije, ta dva dodavanja će biti dodeljena istom pod-klasteru. Samim tim ako su zone porekla i destinacije dodavanja slični onda je i dužina dodavanja slična, što je takođe relevantna karakteristika za određivanje vrednosti dodavanja. Dakle, dobijamo sličnu polaznu i krajnju tačku putanje lopte, što nam govori da su u istom pod-klasteru smeštena dodavanja čija je težina odigravanja skoro ista. Pod-klastere pravimo tako da budu jednake veličine, tj. da sadrže jednak broj dodavanja. Pažnja se obraća da pod-klasteri ne sadrže preveliki broj dodavanja ali i da nemamo preveliki broj pod-klastera. Potrebno je naći određeni balanas između veličine pod-klastera i broja pod-klastera, radi lakšeg izračunavanja. Za veličinu zona na koje je teren podeljen, uzima se dimenzija 5 x 4 metara. Ova dimezija se koristi zbog veće obuhvatnosti zone. Ako bi se koristila dimenzija 1 x 1 metar dešavalo bi se, u mnogo većoj meri, da slična dodavanja završe u različitim pod-klasterima, što nam usporava i otežava izračunavanje distance.

Generalno, problem kod pod-klastera je u situacijama kada se dodavanje nalazi u blizini granice zone. Problem je što dva slična dodavanja koja imaju različitu zonu porekla i destinacije mogu da završe u različitim pod-klasterima, kao što je prikazano na slici 36, a to nije željeni ishod

klasterizacije. Da bi se izbegao ovaj problem moramo naći rešenje, kako da odredimo kojoj će zoni pripadati dodavanje, da bi mogli klasterizovati slična dodavanja u isti pod-klaster. Ako se poreko i destinacija dodavanja nalaze blizu granica zona, onda se oni dodeljuju u više zona. Kada je razdaljina od susedne zone manja od već unapred određenog praga, onda se poreklo i destinacija dodeljuju toj susednoj zoni, a kada se poreklo i destinacija nalaze u samom uglu zone onda se oni mogu dodeliti u 4 najbliže zone. Na osnovu ovog se odredi kojoj će zoni pripadati poreko i destinacija dodavanja i u koji pod-klaster će dodavanje biti stavljeno.



SLIKA 36: IZVOR [3]

5.2.4 ODREĐIVANJE VREDNOSTI DODAVANJA

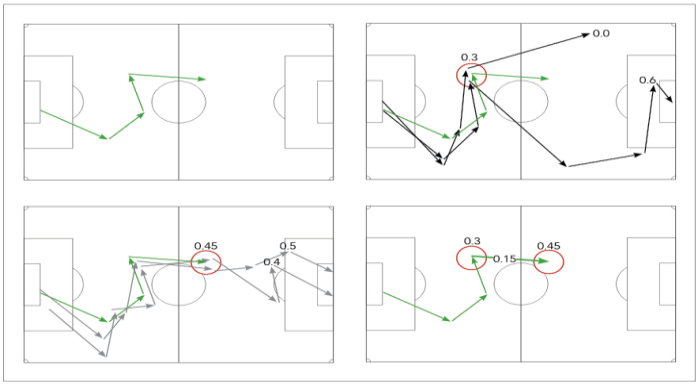
Kada su kreirane pod-klasteri ostaje još samo da se odrede vrednosti za dodavanje. Vrednosti za dodavanja se dobijaju kao prosečna vrednost očekivanih ishoda perioda posedovanja, njihovih *k*-najbližih suseda. Za računanje *k*-najbližih suseda se koristi gore pomenuta, novo uvedena mera distance. U ovom radu je $k = 100$, što znači da sličnost dodavanja merimo u odnosu na 100 najbližih suseda. POV pristup se koristi za dobijanje vrednosti dodavanja na osnovu prosečne vrednosti perioda posedovanja za *k*-najbližih suseda.

5.3 NIZOVNO-ORIJENTISANO VREDNOVANJE DODAVANJA

Treći pristup vrednovanja dodavanja je nizovno-orijentisano vrednovanje (NOV). Ovaj pristup uzima u obzir periode posedovanja i pod-periode perioda posedovanja. Potrebno je vrednovati pod-period posedovanja na osnovu njegovog očekivanog ishoda. Kada je vrednovan pod-period posedovanja može se vrednovati i dodavanje na osnovu njegovog uticaja na ishod perioda posedovanja kome to dodavanje pripada. U ovom pristupu koristimo 3 mere distance:

- DVS mera distance
- Frečetova mera distance
- NZP mera distance.

Ove mere nam služe za poređenje pod-perioda i merenje sličnosti između njih. Na osnovu rezultata iz trening skupa biramo meru koja je najbolja, tj. najbolje određuje sličnost između dodavanja pod-perioda posedovanja. Nakon ovog upoređivanja i određivanja očekivanog ishoda pod-perioda posedovanja, može se odrediti vrednost za dodavanje. Na slici 37, je prikazano vrednovanje dodavanja NOV pristupom.



SLIKA 37: IZVOR [1]

Vrednuje se poslednje dodavanje iz perioda posedovanja sa prve slike. Na drugoj slici imamo dva slična perioda posedovanja sa očekivanim ishodima, 0 i 0,6. Računanjem prosečne vrednosti dobijamo, da je vrednost dva slična dodavanja (dva najbliža suseda) iz dva perioda posedovanja jednak 0,3. Ta dva slična dodavanja se završavaju na skoro istom mestu gde počinje dodavnje koje treba izračunati (posmatra se crveni krug). Na trećoj slici opet imamo neka dva perioda

posedovanja koja u sebi sadrže dva slična dodavanja. Periodi posedovanja imaju očekivane vrednosti 0,4 i 0,5. Kao i do sada vrednosti sličnih dodavanja se dobijaju kao prosečna vrednost i u ovom slučaju iznosi 0,45. U crvenom krugu na trećoj slici vidimo da dva uporediva dodavanja i dodavanje koje treba vrednovati imaju istu destinaciju. Na kraju, kada smo dobili početnu i krajnju vrednost dodavanja, može se izračunati vrednost za to dodavanje. Vrednost se dobija kada od krajnje vrednosti oduzme početna, tj. 0,45 - 0,30 = 0,15. Može se zaključiti, da se određeno dodavanje vrednuje kao razlika između pod-perioda koji se završava sa tim dodavanje i pod-perioda koji se završava neposredno pre tog dodavanja.

Kao i prethodni pristupi i NOV pristup se sastoji od nekoliko svojstava, koji su objašnjeni u sledećim koracima:

1. Odrediti koja se mera distance koristi za upoređivanje pod-perioda posedovanja
2. Razdvojiti periode u pod-periode posedovanja i vrednovati ih
3. Grupisati pod-periode trening i test skupa po poreklu i destinaciji
4. Izračunati distancu između pod-perioda i pronaći 100 najbližih suseda
5. Vrednovati dodavanje

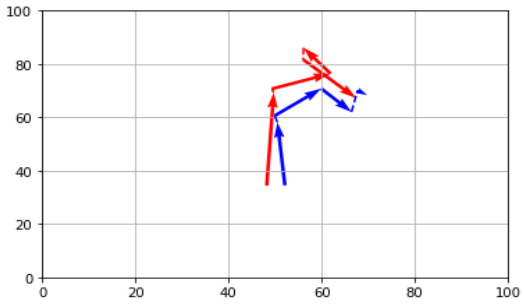
5.3.1 ODREĐIVANJE MERE DISTANCE ZA NOV

Niz dodavanja koji čine jedan period posedovanja može se posmatrati kao vremeska serija, gde svako dodavanje ima poreklo, destinaciju i vremenski okvir. To nam govori da se dodavanja razlikuju u dužini i vremenu odigravanja, a pa se i periodi posedovanja razlikuju po vremenu i dužini. Da bi se izmerila distanca između dodavanja, odnosno perioda posedovanja koristimo tri distance:

- Dinamičko vremensko savijanje (DVS)
- Frečetova mera distance
- Najduži zajednički podniz (NZP).

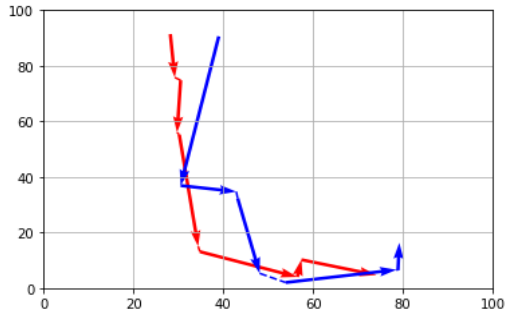
Od te tri distance biramo jednu koja daje najbolje rezultate, koja određuje najsličnija dodavanja. Pri analizi mera distance koristi se podskup od 1000 perioda posedovanja iz trening skupa, a svaki taj period posedovanja se mora sastojati od najmanje 5 perioda dodavanja. Na slici 37, 38 i 39 se

vidi kakve rezultate daju sve tri mere distance i koja od ove tri mere je najbolja za pronalaženje najbližih suseda.



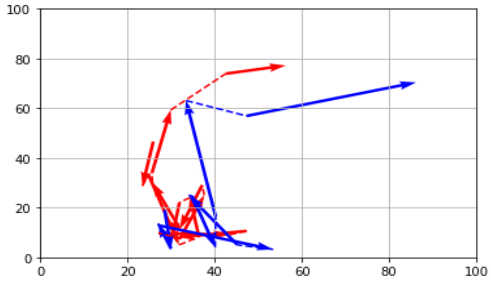
SLIKA 38: IZVOR [1]

Na slici 38, su prikazana dva perioda posedovanja, čija razdaljina je merana DVS distancom. Sa slike se može primetiti da su oba ova perioda posedovanja veoma slična, što govori da DVS mera distance daje dobre rezultate za upoređivanje i određivanje sličnosti.



SLIKA 39: IZVOR [1]

Na slici 39, distanca dva perioda poseda se meri pomoću Frečetove distance. Posmatrajući ovu sliku vidimo da i Frečetova distanca daje dobre rezultate i da su ova dva perioda posedovanja veoma slična, što znači da je i Frečetova distanca dobra za određivanje *k*-najbližih suseda.



SLIKA 40: IZVOR [1]

Na slici 40, na osnovu poređenja dva perioda posedovanja, može se primetiti da nema prevelike sličnosti između njih, kao što je bilo u prethodna dva slučaja. Za merenje razdaljine između dva perioda posedovanja koristi se NZP distanca, koja nije dovoljno dobra mera sličnosti izmedju dodavanja, tj. perioda posedovanja koja se sastoje od tih dodavanja.

Može se zaključiti da su DVS i Frečetova distanca dobre mere za određivanje sličnosti i za određivanje *k*-najbližih suseda dodavanja, dok NZP mera distance nije dovoljno dobra i ona se ne koristi za merenje distance.

5.3.2 RAZDVAJANJE PERIODA U POD-PERIODE

U ovom koraku se periodi posedovanja razdvajaju u pod-periode, pa onda vrednujemo dobijene pod-periode. Pod-periodi imaju isti početak kao i periodi posedovanja ali njihova dužina se sastoji od najmanje dva dodavanja. Pod-period posedovanja se vrednuje na isti način kao što je vrednovan period posedovanja. Pod-period se vrednuje pomoću očekivanog ishoda pod-perioda posedovanja. Njegova dužina ne može da bude veća od dužine perioda posedovanja. Na primer, neka je S_i period posedovanja, koji se sastoji od niza dodavanja $[p_1, \dots, p_{n_i}]$, gde je n_i ukupan broj dodavanja koja čine jedan period posedovanja. Onda se pod-period ovakvog perioda posedovanja S_i, označava na sledeći način $S_i(r) = [p_1, \dots, p_r]$, gde je r = 2, …, n_i.

5.3.3 POD-KLASTERI POD-PERIODA POSEDOVANJA

Potrebno je pod-periode posedovanja podeliti u pod-klastere, tako da se slični pod-periodi nalaze u istim pod-klasterima. Na ovaj način lakše i brže računamo vrednosti za pod-periode. Da bi pod-

periode svrastali u iste pod-klastere uzimaju se u obzir početna i krajnja tačka niza pod-perioda. Ovde je bitno šta se dešava između tih početnih i krajnjih tačaka, jer može da se desi da neki pod-periodi imaju sličnu početnu i krajnju tačku ali da se sastoje od različitog broja dodavanja koji međusobno nisu slični. Ovde koristimo veće zone nego u POV pristupu da se ne bi dešavalo da slični pod-periodi posedovanja završe u različitim pod-klasterima. Za veličine zona su uzete dimenzije 10 x 8 metara. Koristi se prag od 2 metra i ukoliko se pod-period nalazi u okviru 2 metra od granice zone, pod-period posedovanja se dodeljuje većem broju zona, odnosno većem broju pod-klastera.

5.3.4 RAČUNANJE DISTANCE I ODREĐIVANJE K-NAJBLIŽIH SUSEDA

Kao što je već utvrđeno koriste se dve distance za merenje sličnosti perioda posedovanja: DVS mera distance i Frečetova mera distance. U koraku 5.3.3 su pod-periodi podeljeni u pod-klastere, za trening skup i za test skup. Nekada zbog blizine početne i krajnje tačke granici zona, dešava se da pod-period pripada većem broju zona. Tada se računa distanca između pod-perioda za više pod-klastera i onda se dobija kojem pod-klasteru tačno pripada taj pod-period. Na ovaj način se pronalaze *k*-najbliži susedi pod-perioda posedovanja, gde je k = 100.

5.3.5 VREDNOVANJE DODAVANJA

Kada su određeni najbliži susedi za svaki pod-period posedovanja test skupa, mogu se odrediti vrednosti za svaki pod-period. Vrednost za neki pod-period se računa kao prosečna vrednost ishoda pod-perioda posedovanja njegovih najbližih suseda. Kada su vrednovani svi pod-periodi onda je moguće izračunati vrednost pojedinačnih dodavanja. Neka sa V (S_i) obeležimo vrednost pod-perioda S_i(r) = $[p_1, \dots, p_j, \dots, p_r]$, onda se vrednost pojedinačnog dodavanja p_j računa na sledeći način:

$$\text{NOV}\ (p_j) = \text{V}\ (S_i(j)) - \text{V}\ (S_i(j-1))$$

Vrednost pojedinačnog dodavanja je razlika između pod-perioda koji se završava sa tim dodavanjem i pod-perioda koji se završava u tački gde počinje to dodavanje.

Primenom ovih 5 korak u NOV pristupu, njihovom analizom i izračunavanjem dobijaju se vrednosti za pojedinačna dodavanja. Pri izračunavanju ovih vrednosti najveća pažnja se posvećuje upoređivanju pod-perioda posedovanja i određivanju njihove sličnosti. Na osnovu tog upoređivanja mogu se identifikovati timovi koji imaju sličan stil igre. Na taj način se lakše analiziraju protivnički timovi, njihova taktika i napadačke sposobnosti. U sledećoj tabeli će biti prikazano 6 grupa u koje su podeljenje ekipe sa sličnim stilom igre. Ova podela se dobila korišćenjem DVS mere distance.

Grupa1	**Grupa2**
Barcelona, Napoli, PSG, Manchester City, Juventus, Bayern Munchen	Sunderland, WBA, FSV Mainz 05, FC Augsburg, SV Darmstadt 98
Grupa3	**Grupa4**
Real Madrid, Arsenal, Chelsea, Manchester United, Liverpool, Nice, Fiorentina, AS Roma, Sevilla, Lyon, Marseille, Tottenham Hotspur	AC Milan, Inter, Atletiko Madrid, Monaco, Torino, Atalanta, Bologna, Lazio, Sampdoria, Real Sociedad, Lille, Everton, Empoli, Bournemouth
Grupa5	**Grupa6**
Udinese, Eintracht Frankfurt, Granada CF, Watford, Leicester City, Eibar, FC Koln, Crystal Palace, Toulouse	Valencia, Chievo, Swansea City, Palermo, Bayer 04 Leverkusen, Schalke 04, Deportivo de La Coruna, Hull City, Real Betis, Malaga, Nantes, West Ham United

Tabela 2: Izvor [1]

Timovi iz grupe 1, su poznati po svom specifičnom stilu igre, puno kratkih dodavanja, veliki procenat poseda lopte, izrazito ofanzivne ekipe. Grupu 2, čine ekipe koje su pretežno defanzivne, uglavnom se bore za opstanak u svojm ligama i prepuštaju posed i igru drugim ekipama. Timovi iz grupe 3, su slični timovima iz prve grupe. Takođe su ofanzivne ekipe, ostvaruju veliki broj dodavanja tokom utakmice ali se igra više zasniva po bočnoj strani terena i centaršutevima. Grupa 4, se sastoji od ekipa koje su pri vrhu u gornjem delu sredine tabele. Tu su ekipe koje kada igraju protiv slabijih ekipa igraju ofanzivno, ali kad igraju protiv jačih ekipa su u podređenijem položaju, posed prepuštaju boljim ekipama i igraju defanzivnije. Na kraju ekipe iz grupe 5 i 6 su ekipe iz donjeg dela sredine tabele. One nemaju neki lep stil igre uglavnom su defanzivne i iz kontra-napada pokušavaju da postignu gol.

Korišćenjem ova tri pristupa: ZOV, POV i NOV dobijamo tri načina za vrednovanje dodavanja u fudbalu. Svaki od ova tri pristupa ima svoje prednosti i mane. ZOV pristup je najjednostavniji i

najbrži za izračunavanje. POV pristup koristi vrednosti ishoda perioda posedovanja da bi se izračunale vrednosti najbližih suseda dodavanja. NOV pristup za vrednovanje dodavanja je najdetaljniji, analizira periode i pod-periode posedovanja. NOV pristup je dobar za pronalaženje dobrih dodavača. Razlika u POV i NOV pristupu je u tome što se u NOV meri sličnost pod-perioda posedovanja i dodavanja dobijaju vrednost na osnovu njihovog uticaja na taj period posedovanja, dok se POV pristupom meri samo sličnost individualnih dodavanja. Razlika između ZOV i POV pristupa je u tome što, ZOV pristup ne razmatra da li slična dodavanja dovode do prilike za postizanje gola, dok POV pristup uzima to u razmatranje. Svaki od ovih pristupa sadrži karakteristike koje su potrebne da bi se izračunalo vrednovanje dodavanje. Neke od tih karakteristika su sadržane u sva tri pristupa a neke samo u određenim pristupima. U sledećoj tabeli su date karakteristike koje su potrebne za vrednovanje dodavanja određenim pristupom.

Karakteristika	ZOV	POV	NOV
podela terana na zone	X	X	X
očekivani gol model	X	X	X
pravilo 15 sekundi	X		
sličnost dodavanja		X	
sličnost perioda posedovanja			X
vrednost pre i posle dodavanja	X		X
vrednovanje perioda posedovanja		X	X
pod-klasterizacija		X	X
k-najbliži susedi		X	X
k-klasterizacija			X

Tabela 3: Izvor [1]

6. INDIVIDUALNA RANGIRANJA ZASNOVANA NA GRUPNOM POREĐENJU

U ovom odeljku se predstavlja traženje novih pristupa, kojima se daje odgovor na pitanje: “kako rangirati pojedinca na osnovu rezultata grupnog poređenja”?!! Cilj je rangirati individualne sposobnosti igrača na osnovu rezultata koje postiže tim za koji nastupa taj igrač. Postoje dva tipa rezultata koja služe za grupno (timsko) poređenje:

- binarni ishod
- merenje ishoda.

Pored ovog pristupa koriste se još i pristupi:

- metode najmanjih kvadrata,
- metode maksimalne verodostojnosti,
- primenjivanje normalne raspodele.

6.1 BINARNI ISHOD I MERNJE ISHODA

Binarni ishod se definiše samo kao pobeda ili poraz tima, dok merenje ishoda predstavalja tačan rezultat koje je ostvario tim u jednoj utakmici. Na primer:

U odigranoj utakmici Mančester vs Čelzi je pobedio Mančester i to predstavlja binarni ishod. Kada bi se napisalo: U odigranoj utakmici Mančester – Čelzi je pobedio Mančester rezultatom 4:2, to bi označavalo merenje ishoda.

Da bi se došlo do grupnog poređenja potrebno je prvo krenuti od upoređivanja individualnih parova. Upoređivnje individualnih parova predstavlja upoređivanje pojedinca sa preostalim pojedincima i na osnovu upoređivanja se dobija koji od njih je najbolji. Na primer:

Posmatraju se 4 fudbalera (A, B, C, D) i treba da uporedimo koji je među njima najbolji, koristi se upoređivanje individualnih parova. Potrebno je uporediti svakog fudbalera sa svakim i na osnovu rezultata doneti zaključak koji je najbolji. U tabeli 4 biće prikazano kako se to upoređivanje vrši.

	A	B	C	D
A		0	0	0
B	1		0	0
C	1	1		
D	1	1	1	0
Rezultat	3	2	1	0

Tabela 4

Upoređivanje se vrši na sledeći način: fudbalera A iz kolone uporedim sa ostalim fudbalerima iz vrste (B, C, D). Broj 1 predstavlja da je jedan fudbaler bolji od drugog a broj 0 da nije bolji. Iz prve kolone se vidi da je fudbaler A bolji od preostala tri fudbalera, dok fudbaler D ima sve nule, tako da je on najlošiji.

Postoji više modela za izračunavanje individualnih sposobnsti, upoređivanjem parova a jedan od najpopularnijih i najkorišćenijih je Bradley – Terry model. Pretpostavlja se da postoji *k* pojedinaca, čije sposobnosti su date ne-negativnim vektorom $p = [p_1,\ p_2, \dots, p_k\ ]^T$. Verovatnoća da pojedinac *i* pobedi pojedinca *j* se dobija na sledeći način:

$$P\ (i \text{ pobeđuje } j) = \frac{p_i}{p_i + p_j} \qquad (1)$$

gde se ocene za vektor *p* dobijaju rešavanjem maksimalne verodostojnosti

$$\min_p - \sum_{i \neq j} n_{ij} \log \frac{p_i}{p_i + p_j} \qquad (2)$$

$$\sum_{j=1}^{k} p_j = 1,\ \ p_j \geq 0, j = 1, \dots, k, \qquad (3)$$

gde n_{ij} predstavlja broj koliko puta je pojedinac *i* pobedio *j*.

Nakon objašnjenja pojedinačnog upoređivanja, može se preći na razmatranje grupnog upoređivanja. Neka je, *k* pojedinaca $\{1, \dots, k\}$ i *m* poređenja. *i*-to poređenje uključuje podskup I_i, koji se razdvaja u dva protivnička tima, I_i^+ i I_i^-. Ta dva tima se upoređuju $n_i = n_i^+ + n_i^-$ puta, gde je n_i^+ broj pobeda tima I_i^+, dok je n_i^- broj pobeda tima I_i^-. Grupno upoređivanje predstavlja proširenje jednačine (1), odnosno generalizaciju Bradley – Terry modela (B-T), koje je dato jednačinom:

$$\mathrm{P}(I_i^+ \text{ pobeđuje } I_i^-) = \frac{\sum_{j:j\in I_i^+} p_j}{\sum_{j:j\in I_i} p_j}. \qquad (4)$$

Ovaj model govori da je, sposobnost tima jednaka sumi njegovih članova, što je usko povezano i sa jednom fudbalskom ekipom čiji kvalitet zavisi od kvaliteta njenih igrača. Ocenjivanje parametara individualih sposobnosti u grupnom poređenju se vrši minimizacijom negativne log-verodostojnosti.

$$\min_p - \sum_{i=1}^{m}(n_i^+ \log\frac{\sum_{j:j\in I_i^+} p_j}{\sum_{j:j\in I_i} p_j} + n_i^- \log\frac{\sum_{j:j\in I_i^-} p_j}{\sum_{j:j\in I_i} p_j}) \qquad (5)$$

$$\sum_{j=1}^{k} p_j = 1,\ \ p_j \geq 0, j = 1, \dots, k\ . \qquad (6)$$

Oba ova modela, Bradley-Terry model i generalizovani Bradley-Terry model za rezultat dobijaju binarni ishod, uključeni su samo pobeda ili poraz određenog tima.

6.2 POREĐENJE SA BINARNIM ISHODIMA

Neka su individualne sposobnosti označene vektorom $v \in \boldsymbol{R}^k$, $-\infty < v_s < \infty$ gde je $s = 1,\dots, k$. Sposobnost timova, za $I_i^+ i\ I_i^-$, se definišu kao suma indivdualnih sposobnosti članova tog tima,

$$T_i^+ \equiv \sum_{s:s\in I_i^+} v_s \text{ i } T_i^- \equiv \sum_{s:s\in I_i^-} v_s. \qquad (7)$$

Uvode se promenljive Y_i^+ i Y_i^-, $1 \leq i \leq m$. One predstavljaju promenljive koje predstavljaju sposobnosti timova pa se može zapisati

$$\mathrm{P}\ (I_i^+ \text{ pobeđuje } I_i^-) \equiv P(Y_i^+ - Y_i^- > 0). \qquad (8)$$

Raspodela za Y_i^+ i Y_i^- je nepoznata, ali zbog jednostavnijeg računanja izvoda prepostavljeno je da promenljive imaju duplu-eksponencijalnu raspodelu

$$P(Y_i^+ \leq \mathrm{y}) = \exp\ (-e^{-(y-T_i^+)}). \qquad (9)$$

Pod pretpostavkom da su Y_i^+ i Y_i^- nezavisne promenljive i na osnovu jednačina (8) i (9) se dobija:

$$\mathrm{P}(Y_i^+ - Y_i^- > 0) \equiv \int_{-\infty}^{\infty} \int_{y^-}^{\infty} de^{-e^{-\left(y^+ - T_i^+\right)}} \, de^{-e^{-(y^- - T_i^-)}} \tag{10}$$

Neka je

$$x^+ \equiv e^{-(y^+ - T_i^+)} \text{ i } x^- \equiv e^{-(y^- - T_i^-)} \tag{11}$$

Zbog toga se dobija

$$de^{-e^{-(y^+ - T_i^+)}} = -e^{-x^+} \, dx^+ \text{ i } de^{-e^{-(y^- - T_i^-)}} = -e^{-x^-} \, dx^-. \tag{12}$$

Onda,

$$\int_0^{\infty} -e^{-x^-} \int_0^{x^- e^{T_i^+ - T_i^-}} -e^{-x^+} \, dx^+ \, dx^- = \frac{e^{T_i^+}}{e^{T_i^+ + T_i^-}}. \tag{14}$$

Rešavajući ove jednačine dobija se da je predloženi model za binarni ishod jednak:

$$\mathrm{P}\ (I_i^+ \text{ pobeđuje } I_i^-) = \frac{e^{T_i^+}}{e^{T_i^+ + T_i^-}}. \tag{15}$$

Analizirajući ovaj model procenjuju se vrednosti za vektor *v*.

6.2.1 METODA NAJMANJIH KVADRATA (MNK)

Kao što je već rečeno n_i^+ i n_i^- je broj pobeda upoređenih timova I_i^+i I_i^-, respektivno [2]. Iz (15) dobija se

$$\frac{e^{T_i^+}}{e^{T_i^+ + T_i^-}} \approx \frac{n_i^+}{n_i^+ + n_i^-} \tag{16}$$

pa je onda

$$e^{T_i^+ - T_i^-} = \frac{e^{T_i^+}}{e^{T_i^-}} \approx \frac{n_i^+}{n_i^-}. \tag{17}$$

Ako je $n_i^+ \neq 0$ i $n_i^- \neq 0$, ocenjivanje parametara vektora *v*, individualnih sposobnosti, dobija se rešavanjem jednačine

$$\min_v \sum_{i=1}^{m} \left((T_i^+ - T_i^-) - \log \frac{n_i^+}{n_i^-} \right)^2 . \tag{18}$$

U slučaju da je $n_i^+ = 0$ ili $n_i^- = 0$, rešenje problema se dobija dodavanjem malog broj na n_i^+ i n_i^-. Ova ideja je opšte prihvaćena i često se koristi kada se pojavljuje deljenje sa nulom. Da bi se jednačina (18) uprostila prvo se uvodi vektor $d \in \boldsymbol{R}^m$, koji definiše $\log \frac{n_i^+}{n_i^-}$, $d \equiv \log \frac{n_i^+}{n_i^-}$. Posle toga se pravi matrica poređenja, gde $G \in \boldsymbol{R}^{m \times n}$ i dobija se:

$$G_{ij} \equiv \begin{cases} 1, & ako\ j \in I_i^+ \\ -1, & ako\ j \in I_i^- \\ 0, & ako\ j \notin I_i \end{cases} \tag{19}$$

Na primeru ekipe za bridž će se objasniti matrica *G*. Za bridž je potrebno četiri igrača, gde dva igrača iz istog tima igraju protiv dva igrača iz drugog tima. Matrica poređenja *G* će pokazati koji igrači igraju u istoj ekipi i protiv koja dva igrača igraju.

$$\mathrm{G} = \begin{bmatrix} 1 & 1 & -1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 & -1 & -1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \tag{20}$$

Posmatra se prva vrsta matrice, ona govori da prvi i drugi igrač igraju protiv trećeg i četvrtog igrača. Iz druge vrste se vidi da prvi i drugi igrač igraju protiv petog i šestog…

Sada kada je sve definisano može se jednačina (18) zapisati na sledeći način:

$$\min_v (Gv - d)^T (Gv - d). \tag{21}$$

Rešavanjem ovog sistema dobija se sledeći linearni sistem iz kojeg se dobijaju vrednosti za vektor *v*:

$$G^T G v = G^T d. \tag{22}$$

Ako $G^T G$ nije invertibilna, linearni sistem (22) može imati višestruka rešenja, što dovodi do više različitih rangiranja. Potrebno je da $G^T G$ bude invertibilna a to je moguće ako važi sledeća teorema:

Teorema 1. **[2]**

$G^T G$ je invertibilna ako i samo ako rang (G) = k.

Dokaz.

Ako je rang (G) $< k, G^T G$ nije punog ranga, tj. nije invertibilna. Ako je rang(G) = k, na osnovu dekompozicije singularne vrednosti dekomponuje se matrica G na sledeći način:

$$G = UQV^T \tag{23}$$

Gde su $U \in \boldsymbol{R}^{m\times k}$, $V \in \boldsymbol{R}^{k\times k}$ ortonormalne i $Q \in \boldsymbol{R}^{k\times k}$ dijagonalna sa

$$Q_{ii} \neq 0, i = 1, \dots, k. \tag{24}$$

Dobija se da je

$$G^T G = VQU^T UQV^T = VQ^2V^T \tag{25}$$

invertibilna.

Rezultat pokazuje da članove timova treba često menjati tokom poređenja, tako da su individualne sposobnosti jedinstveno određene. Analiziranjem ekstremnih slučajeva kada nekoliko igrača uvek pripada istom timu, dolazi do pojave višestrukog rangiranja. Da bi se izbegle ovakve situacije, koristi se regularizacija. Regularizacija služi:

- za regulisanje modela, da se izbegnu višestruka rešenja
- da smanji grešku generalizacije
- da model radi dobro, kako na trening podacima tako i na test podacima.

Sad je u jednačinu (21) potrebno dodati parameter regularizacije μvv^t, gde je μ mali pozitivan broj koji naziva hiperparametar parametrizacije i dobija se sledeća jednačina:

$$\min_v (Gv - d)^T (Gv - d) + \mu vv^t \tag{26}$$

Uvođenjem parametrizacija dobija se jedinstveno rešenje:

$$(G^T G + \mu I)^{-1} G^T d. \tag{27}$$

6.2.2 MAKSIMALNA VERODOSTOJNOST (MAX.VER)

Maksimalna verodostojnost je jedna od metoda za određivanje nepoznatih parametara. Ideja metoda je da se vrednost parametra izabere tako da verodostojnost modela bude maksimalna, tj. verovatnoća realizacije dobijenog uzorka bude najveća. Neka su poređenja nezavisna, a zna se da je minimum negativne log-verodostojne funkcije ekvivalentan funkciji maksimalne verodostojnosti, tako da se može koristiti negativna log-verodostojna funkcija [2]:

$$\mathrm{L(v)} \equiv -\sum_{i=1}^{m} \left(n_i^+ \log \frac{e^{T_i^+}}{e^{T_i^+} + e^{T_i^-}} + n_i^- \log \frac{e^{T_i^-}}{e^{T_i^+} + e^{T_i^-}} \right) \tag{28}$$

i nepoznati parameter *v* se dobija rešavanjem

$$\arg\min_{v} L(v). \tag{29}$$

Kako je log-verodostojna funkcija konkavna onda je *L(v)* konveksna. Međutim, ako *L(v)* nije striktno konveksna višestruki globalni minimum može rezultirati višestrukim rangiranjem. Zato je data sledeća teorema koja daje potreban i dovoljan uslov za striktnu konveksnost.

***Teorema 2*. [2]**

L(v) je striktno konveksna ako i samo ako rang(*G*) = *k*.

Dokaz.

Prvo se *L(v)* zapiše kao

$$L(v) \equiv -\sum_{i=1}^{m} (n_i^+ T_i^+ + n_i^- T_i^-) + \sum_{i=1}^{m} n_i \log(e^{T_i^+} + e^{T_i^-}) \tag{30}$$

Prva suma je konveksna, a za drugu sumu potrebno je pokazati konveksnost. Koristeći Holderovu nejednakost dobija se:

$$\sum_{i=1}^{m} n_i \log(e^{\alpha T_i^+ + (1-\alpha)\tilde{T}_i^+} + e^{\alpha T_i^- + (1-\alpha)\tilde{T}_i^-}) \quad (31)$$

$$= \sum_{i=1}^{m} n_i \log((e^{T_i^+})^{\alpha}(e^{\tilde{T}_i^+})^{1-\alpha} + (e^{T_i^-})^{\alpha}(e^{\tilde{T}_i^-})^{1-\alpha}) \quad (32)$$

$$\leq \sum_{i=1}^{m} n_i \log(e^{T_i^+} + e^{T_i^-})^{\alpha} (e^{\tilde{T}_i^+} + e^{\tilde{T}_i^-})^{1-\alpha} \quad (33)$$

$$= \sum_{i=1}^{m} n_i \alpha(e^{T_i^+} + e^{T_i^-}) + \sum_{i=1}^{m} n_i (1-\alpha)(e^{\tilde{T}_i^+} + e^{\tilde{T}_i^-}) \quad (34)$$

za svako v, $\tilde{v}_i$ $\alpha \in (0,1)$ i jednakost važi ako i samo ako

$$T_i^+ - T_i^- = \tilde{T}_i^+ + \tilde{T}_i^- \; \forall i \quad (35)$$

pa se može zapisati

$$G(v - \tilde{v}) = 0. \quad (36)$$

Ako je rang(G) = k, onda $G(v - \tilde{v}) = 0$ važi ako i samo ako je $v = \tilde{v}$, pa je $L(v)$ je strogo konveksna.

Ako je $L(v)$ strogo konveksna, onda jednakost (34) važi ako i samo ako je $v = \tilde{v}$, pa se dobija

$$G(v - \tilde{v}) = 0 \Leftrightarrow v = \tilde{v}, \quad (37)$$

što implicira da je rang(G) = k.

Da bi se obezbedilo jedinstveno rešenje koristi se parametar regularizacije. Parametar regularizacije ima sledeći oblik:

$$\mu \sum_{i=1}^{k} (e^{v_s} + e^{-v_s}), \quad (38)$$

gde je μ mali pozitivan broj.

Ovaj parameter regularizacije mora biti koveksan i mora imati jedinsteni minimum kada je $v_s = 0$, za $s = 1, \ldots, k$. Sada kad je definisan i parametar regularizacije, funkcija negativne log-verodostojnosti ima sledeći oblik:

$$L(v) \equiv -\sum_{i=1}^{m} \left(n_i^+ \log \frac{e^{T_i^+}}{e^{T_i^+} + e^{T_i^-}} + n_i^- \log \frac{e^{T_i^-}}{e^{T_i^+} + e^{T_i^-}} \right) + \mu \sum_{i=1}^{k} (e^{v_s} + e^{-v_s}). \quad (39)$$

Individualne sposobnosti v se dobijaju traženjem jedinstvenog globalnog minimum, rešavajući

$$\operatorname{argmin}_{v} L(v). \quad (40)$$

6.3 POREĐENJE SA MERENIM ISHODIMA

Individualne sposobnosti se procenjuju i na osnovu analize merenih ishoda. Za razliku od binarnog ishoda sada je bitno kojima rezultatom je neka ekipa pobedila. Kao promenljive koje govore o učinku timova koriste se Y_i^+ i Y_i^-. n_i^+ i n_i^- sada označavaju tačan rezultat timova I_i^+ i I_i^-. Cilj ovog pristupa je modeliranje razlike u rezultatu n_i^+ - n_i^-, umesto modeliranja samog rezultata. n_i^+ - n_i^- razlika u rezultatu predstavlja razliku u realizaciji promenljivih $Y_i^+ - Y_i^-$. U ovom pristupu, za procenu individualne sposobnosti koriste se dve raspodele:

- normalana raspodela
- raspodela ekstrmnih vrednosti.

6.3.1 MODEL NORMALNE RASPODELE (MNR)

Neka slučajna promenljiva Y_i ima normalnu raspodelu [2]

$$Y_i \sim N(v_i, \sigma^2),\ i = 1, \dots, k. \quad (41)$$

Razlika između pojedinca i i j može se predstaviti kao razlika u realizaciji $Y_i - Y_j$. Pretpostavlja se, da su promenljive Y_i i Y_j nezavisne, za sve i = 1, …, k, onda se dobija raspodela za $Y_i - Y_j$

$$Y_i - Y_j \sim N(v_i - v_j, 2\sigma^2) \quad (42)$$

a individulane sposobnosti se procenjuju metodom maksimalne verodostojnosti.

Neka su sada Y_i^+ *i* Y_i^- slučajne promenljive koje predstavlaju učinak dva tima. One su nezavisne i imaju normalnu raspodelu, pa se dobija

$$Y_i^+ \sim N(T_i^+, \sigma^2), Y_i^- \sim N(T_i^-, \sigma^2) \quad (43)$$

i

$$Y_i^+ - Y_i^- \sim N(T_i^+ - T_i^-, \sigma^2). \quad (44)$$

Radi lakšeg zapisa definiše se vektor b:

$$b_i \equiv n_i^+ - n_i^-. \quad (45)$$

Kada je definisana normalna raspodela i kada je definisan vektor *b*, onda negativna log-verodostojna funkcija ima sledeći oblik

$$\mathrm{L}(v, \sigma) = \log\sigma + \frac{1}{4\sigma^2}\sum_{i=1}^{m}(T_i^+ - T_i^- - (n_i^+ - n_i^-))^2 \quad (46)$$

$$= \log\sigma + \frac{(Gv-b)^T(Gv-b)}{4\sigma^2} \quad (47)$$

gde je *G* matrica poređenja. Procenjene vrednosti vekotra *v* za individualne sposobnosti se dobijaju metodom maksimalne verodostojnosti. Rešavajući parcijalne izvode $\frac{\partial L(v,\sigma)}{\partial v_s} = 0$, za $\forall s$, kao rešenje dobija se sistem linearnih jednačina

$$G^T G v = G^T b. \quad (48)$$

gde je vektor individualnih sposobnosti

$$\bar{v} \equiv (G^T G)^{-1} G^T b. \quad (49)$$

Da bi se izbeglo višestruko rangiranje i dobilo jedinstveno rešenje i u ovaj pristup uvodimo parametar regularizacije, pa jednačina ima novi oblik

$$\min_v L(v, \sigma) + \frac{\mu}{4\sigma^2} v^t v \quad (50)$$

gde je hiperparametar μ mali pozitivan broj. Za novo regularizovano rešenje dobija se sisitem linearnih jednačina

$$(G^T G + + \mu I)v = G^T b \quad (51)$$

pa se dobija jedinstveno rešenje za vektor *v*

$$\bar{v} \equiv (G^T G + \mu I)^{-1} G^T b \quad (52)$$

gde je $G^T G$ invertibilno.

6.3.2 MODEL RASPODELE EKSTREMNE VREDNOSTI (MREV)

Posmatraju se slučajne promenljive Y_i^+ *i* Y_i^-, tako da razlika u realizaciji Y_i^+- Y_i^- ima raspodelu za ekstremne vrednosti kao i kod binarnih ishoda. Na osvnovu izvoda koji je izračunat u odeljku 6.2 dobija se [2]

$$P(Y_i^+ - Y_i^- \leq y) = \frac{e^{T_i^-}}{e^{T_i^+ - y} + e^{T_i^-}} \tag{53}$$

i onda funkcija gustine ima sledeći oblik

$$f_{Y_i^+ - Y_i^-}(y) = \frac{e^{T_i^- + T_i^+ - y}}{(e^{T_i^+ - y} + e^{T_i^-})^2}. \tag{54}$$

Negativna log-verodostojna funkcija sada izgleda

$$L(v) = -\sum_{i=1}^{m} \log \frac{e^{T_i^- + T_i^+ - (n_i^+ - n_i^-)}}{(e^{T_i^+ - (n_i^+ - n_i^-)} + e^{T_i^-})^2}. \tag{55}$$

Ova funkcija je striktno konveksna. Ako ne važe uslovi striktne konveksnosti kao u delu 6.2.2 onda se javlja problem višestrukog rešenja a samim tim i pojava višestrukog rangiranja. Kao i ranije ovaj problem se rešava uvođenjem parametra regularizacije koji je takođe konveksan i ima sledeći oblik

$$\mu \sum_{s=1}^{k} (e^{v_s} + e^{-v_s}). \tag{56}$$

Sada se dobija regularizovana negativnu log-verodostojnu funkciju koja ima sledeći oblik

$$\min_v L(v) \equiv -\sum_{i=1}^{m} \log \frac{e^{T_i^- + T_i^+ - (n_i^+ - n_i^-)}}{\left(e^{T_i^+ - (n_i^+ - n_i^-)} + e^{T_i^-}\right)^2} + \mu \sum_{s=1}^{k} (e^{v_s} + e^{-v_s}). \tag{57}$$

Rešenje funkcije se dobija određivanjem jedinstvenog globalnog minimuma, koji se dobija rešavanjem parcijalnih izvoda $\frac{\partial L(v)}{\partial v_s} = 0$, za sve $s = 1,\ldots, k$.

6.4 PROCENJIVANJE PRISTUPA I GREŠAKA U RANGIRANJU

Rangiranje u sportu ima veliki značaj. Najvažnije dve karakteristike rangiranja su:

1. Sumiranje relativnih preformansi igrača ili timova na osnovu ishoda, tako da se može lakše uočiti razlika između kvalitetnijih i lošijih, timova ili igrača
2. Predviđanje budućih ishoda na osnovu dostupnih podataka iz prošlosti.

Ove dve karakteristike se mogu povezati sa minimizacijom empirijske greške i sa minimizacijom greške generalizacije. Za karakteristiku (1.), rangiranje mora biti prilagodljivo, tj. u čvrstoj vezi sa dostupnim podacima tako da se minimizira greška na trening podacima. Kod karakteristike (2.), rangiranje treba dobro da predvidi buduće ishode, što govori da je potrebno minimizirati grešku na novim podacima, koji još nisu korišćeni. Ovako definisane karakteristike se mogu predstaviti još i kao: emprijski kriterijum i kriterijum generalizacije. Oni daju odgovor na sledeća pitanja?

1. Empirijske kriterijum: koliko dobro su procenjena sposobnost i rangiranje prilagođeni dostupnim podacima iz trening skupa?
2. Kriterijum generalizacije: koliko dobro procenjena sposobnost i rangiranje predviđaju ishode na osnovu novih podataka?

Treba jasno napraviti razliku između individualnih sposobnosti i individualnog rangiranja. Individualne sposobnosti daju rangiranje, dok se iz rangiranja ne dobija individualna sposobnost. Ako postoje individualne sposobnosti može se izračunati grupna sposobnost, koja se računa kao suma individualnih sposobnosti. Nasuprot individualnoj sposobnosti gde je jasan njihov odnos u grupi, za individualna rangiranja odnos između njih u grupi nije prilično jasan, pa sposobnost grupe nije direktno uočljiva.

6.4.1 GREŠKE ZA SPOSOBNOST I RANGIRANJE

U ovom delu se objašnjava kako da se izračunavaju greške pri procenjivanju sposobnosti i rangiranja. Takođe će se objasniti veza između tih grešaka i gore navedenih kriterijuma.

Pri procenjivanju sposobnost i rangiranja javljaju se greške tj. odstupanja od stvarnih vrednosti. Razmatraju se dve greške:

- greška grupnog poređenja (GGP)
- greška grupnog ranga (GGR).

Neka su$\{(I_1^+, I_1^-, n_1^+, n_1^-), \dots, (I_1^+, I_1^-, n_1^+, n_1^-)\}$ grupna poređenja i njihovi ishodi, a vektor $v \in \boldsymbol{R}^k$je vektor indivudualnih sposobnosti. Greška grupnog poređenja predstavlja grešku u predviđanju grupnog upoređivanja, gde se sposobnost grupe definiše pojedinačnom sposobnošću (v) i može se zapisati

$$GGP(v) \equiv \frac{\sum_{i=1}^{m} I\{(n_i^+ - n_i^-)(T_i^+ - T_i^-) \le 0\}}{m} \tag{58}$$

gde je I $\{\cdot\}$ indikator funkcije; T_i^+ *i* T_i^- su grupne sposobnosti koje se predviđaju za grupe I_i^+ I_i^-. GGP predstavlja odnos između pogrešno predviđenih poređenja sa vektorom v i m poređenja. GGP uzima u obzir greške svih poređenja.

Druga greška je greška grupnog ranga. GGR predstavlja grešku predviđanja grupnog poređenja, gde se jačina grupe dobija na osnovu ranga pojedinca koji se nalazi u toj grupi. GGR računa greške samo onih poređenja u kojima se jačina neke grupe jasno može odrediti na osnovu ranga njenih članova. Formula za izračunavanje GGR se može zapisati kao

$$GGR(r) \equiv \frac{\sum_{i=1}^{m}\left(I\{n_i^+ > n_i^- \; i \; U_i^+ > L_i^-\} + I\{n_i^+ < n_i^- \; i \; L_i^+ < U_i^-\}\right)}{\sum_{i=1}^{m}\left(I\{U_i^+ > L_i^-\} + I\{L_i^+ < U_i^-\}\right)} \tag{59}$$

gde se r koristi za rangiranje a r_s predstavlja rang pojedinca s i gde su

$$U_i^+ \equiv \min_{j \in I_i^+} r_j, \quad U_i^- \equiv \min_{j \in I_i^-} r_j \tag{60}$$

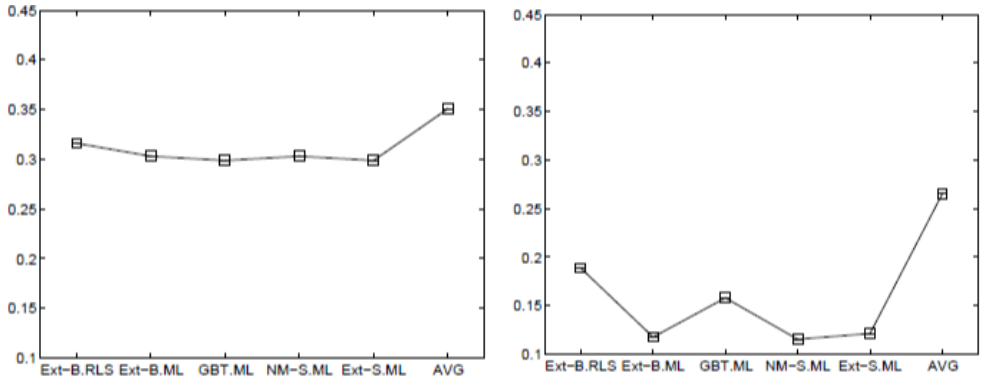
$$L_i^+ \equiv \max_{j \in I_i^+} r_j, \quad L_i^- \equiv \max_{j \in I_i^-} r_j. \tag{61}$$

U_i^+ *i* L_i^+označavaju najmanji i najveći rang u grupi I_i^+. U imeniocu je predstavljen broj poređenja gde svi članovi jedne grupe imaju veći (niži) rang nego članovi druge konkurentne grupe, dok je u brojiocu predstavljen broj pogrešnih predikcija, gde članovi pobedničke grupe imaju manji rang od članova poražene grupe.

6.4.2 UPOREĐIVANJE PRISTUPA

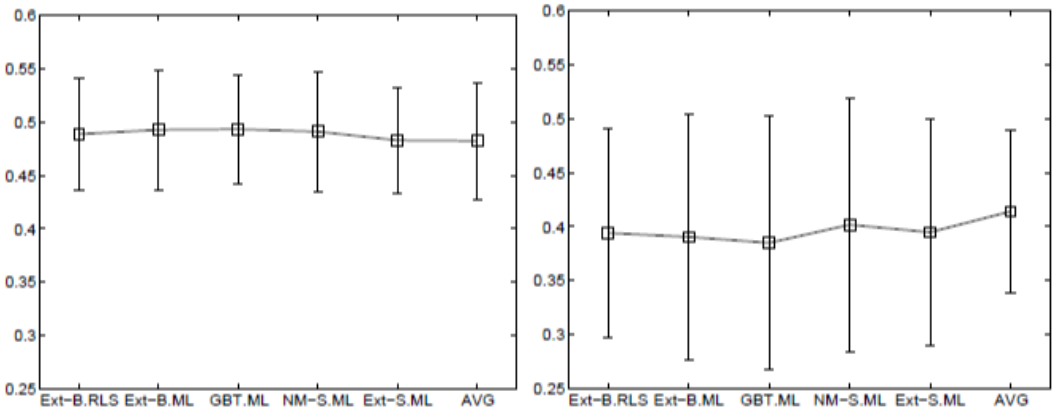
Kombinavanjem ove dve greške sa gore dva navedena kriterijuma dobija se kombinacija između njih: Empirijska GGP, Empirijska GGR, Generalizovana GGP i Generalizovana GGR. U okviru ovih kombinacija, vrši se procenjivanje i ocenjivanje najboljeg i najlošijeg pristupa u odnosu na GGP i GGR. Predloženi pristupi koji se međusobno porede čini šest pristupa, a to su: MNK, MAX.VER, MNR, MREV, B-T i AVG. Prvih pet pristupa je već objašnjeno, a AVG je jednostavan pristup za dobijanje individualnih sposobnosti. Sumiranjem tih individualnih sposobnosti mogu se dobiti i grupne sposobnosti.

Analize i rezultati koji su dobijeni upoređivanjem pristupa su prikazani na slikama 41 i 42. One prikazuju greške pri predviđanju svakog pristupa za empirijski i generalizovan kriterijum. Iz empirijskog kriterijuma, slika 41 (a) se vidi da AVG ima najveću GGP. Svi ostali pristupi su bolji, daju manju GGP nego AVG. Posle AVG najveću GGP ima MNK ali ona se ne razlikuje puno od GGP za ostale pristupe. Rezultati za GGR su mnogo bolji i greške pristupa su mnogo manje, slika 41 (b). To je zato što su podaci prilagodljivi, jer se radi samo sa podacima koji jasno određuju jačinu grupe. Uočljivo male GGR su za MAX.VER, MNR i MREV dok AVG i za ovu grešku ima najveće odstupanje. Nasuprot emprijiskom kriterijumu, u kriterijumu generalizacije dobijamo slabe rezultate. Svi pristupi imaju velike GGP i GGR, slika 42 (a) i (b). GGR su malo manje u odnosu na GGP, ali je to zanemarljivo. U svim ovim kriterijumima AVG pristup može da odredi ishode za veći broj utakmica ali se javlja i veći broj grešaka. Cilj ovih predloženih pristupa je da teže ka što preciznijem rangiranju. Pod tim se podrazumeva da pristupi ne moraju proceniti jačinu grupne sposobnosti u svim poređenjima, ali kada je uspeju odrediti da ta procena bude što tačnija.



a) GGP b) GGR

Slika 41: Empirijski kriterijum za šest pristupa



a) GGP b) GGR

Slika 42: Kriterijum generalizacije za šest prisputa

7. ZAKLJUČAK

Svrha ovog istraživanja je da teorijski predstavi povezanost matematike i fudbala. Konkretno, u ovom radu se pažnja posvetila kako se na osnovu mašinskog učenja i nekih drugih pristupa mogu vrednovati dodavanja, kako se može vrednovati i šut ka golu. Najveća pažnja se posvetila kako na osnovu mašinskog učenja teorijiski možemo da izračunamo vrednosti za dodavanje. Videli smo da jedna od najvažnijih stavki je kako grupisati slična dodavanja u skupove. Uz pomoć metoda za merenje distanci i klasterizacije mi klasifikujemo koja dodavanja su slična, i onda se za ta dodavanja može izračunati njihova vrednost. Bitnu ulogu kada se posmatra celokupna slika vrednovanja dodavanja ima i podela terena na zone. Kada se teren podeli na zone lakše je klasifikovati slična dodavanja i odrediti u koji skup mogu da se svrstaju. Mali problem se stvara kad se početak ili kraj dodavanja nalazi na granici dve zone, tada je potrebno odrediti prag tolerancije i videti u kom skupu će to dodavanje imati najveću sličnost sa ostalim dodavanjima.

Tokom istraživanja se videlo da bitnu ulogu u vrednovanju dodavanja imaju i tri pristupa: zonski-orijentisano vrednovanje, pas-orijentisano vrednovanje i nizovno-orijentisano vrednovanje. U ovim pristupima krucijalna stvar koji god pristup da se koristi za vrednovanje dodavanja je taj da najprecizniji način za vrednovanje dodavanja je kada se dodavanje podeli na niz pod-perioda dodavanja. Tada se određuje vrednost za te pod-periode dodavanja i na osnovu vrednosti pod-perioda dobija se vrednost za celo dodavanje.

LITERATURA

[1] Lotte Bransen, Valuing passes in football using ball event data, Erasmus University Rotterdam, 2017;

[2] Tzu-Kuo Huang, Chih-Jen Lin, Ranking Individuals by Group Comparisons, Departman of Computer Science, Nattional Taiwan University, 2008;

[3] Joel Brooks, Matthew Ker, John Guttag, Developing a Data-Driven Ranking in Soccer Using Predictive Model Weights, Massachusetts Institite of Technology, Cambridge, 2016;

[4] Laszlo Gyarmati, Rade Stanojević, QPasss: aMerit-based Evaluation of Soccer Passes, Qatar Computing Research Institute, 2016;

[5] Jelena Jovanović, Mašinsko učenje, Departmant of Software Engineering, Univresity of Belgrade,"http://www.youtube.com/view_play_list?p=A89DCFA6ADACE599",

https://www.coursera.org/course/ml, 2016;

[6] Danijela Petrović , Mašinskoučenje, http://poincare.matf.bg.ac.rs/~danijela/VI/11_cas/p_12.pdf,

http://poincare.matf.bg.ac.rs/~danijela/VI/12_cas/p_13.pdf, 2016;

[7] Jasmina Đ. Novaković, Rešavanje klasifikacionih modela mašinskog učenja, fakultet tehničkih nauka, Univerzitet u Kragujevcu, 2013;

[8] Jelena Jovanović, Klasterizacija, Departmant of Software Engineering, Univresity of Belgrade, http://ai.fon.bg.ac.rs/wp-content/uploads/2015/04/Klasterizacija-2015.pdf, 2015;

[9] Bojan Furlan, Mašinsko učenje k-najbližih suseda, Elektrotehnički fakultet, Univerzitet u Beogradu, Discovering Knowledge in Data: An Introduction to Data Mining (Wiley, 2005) Larose D. poglavlja 2 i 5, 2012;

[10] Vuk Batanović, Linearna regresija, Obučavanje i evaluacija modela u nadgledanom mašinskom učenju, Elektrotehnični fakultet, univerzitet u Beogradu,

https://rti.etf.bg.ac.rs/rti/ms1psz/pdf/Linearna%20regresija.pdf, 2017;

[11] Kumar Vasimalla, Narasimham Challa, Manohar Naik, Efficient Dynamic Warping For Time Series Classification, Indian Journal Of Science and Technology, 2016;

[12] Sajt Vikipedija, https://sh.wikipedia.org/wiki/Ma%C5%A1insko_u%C4%8Denje,

https://en.wikipedia.org/wiki/Fr%C3%A9chet_distance,

https://sr.wikipedia.org/sr-ec/Pretra%C5%BEivanje_najbli%C5%BEeg_suseda;

[13] Josip Krapac, Regularizacija dubokih modela, Univerzitet u Zgrebu;

[14] Z. Lozanov-Crvenković, Statistika, Univerzitet u Novom Sadu 2012;

[15] Sajt, https://www.picswe.com/;

[16] Alok Raj Gupta, Simple and in depth introduction od k-NN, 2018;

[17] Milan M. Milosavljević, Support vector machine, Lekcija 5a SVM.pdf

Kratka biografija

Nenad Rakić je rođen 30. Juna 1990 godine u Sremskoj Mitrovici. Osnovnu školu "Veljko Dugošević" završio je u Rumi 2005 godine. Srednju školu, gimnaziju "Stevan Puzić" je završio u Rumi, 2009 godine. Zatim je upisao Prirodno-matematički fakultet, smer primenjena matematika, modul matematika finansija, gde je 2015. godine završio osnovne studije. Iste godine upisuje i master studije na istom fakultetu i završava sa prosekom 6,93.

Novi Sad, 2020

UNIVERZITET U NOVOM SADU PRIRODNO-MATEMATIČKI FAKULTET
KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj:

RBR

Identifikacioni broj:

IBR

Tip dokumentacije: monografska dokumentacija

TD

Tip zapisa: tekstualni štampani materijal

TZ

Vrsta rada: master rad

VR

Autor: Nenad Rakić

AU

Mentor: dr Dušan Jakovetić

MN

Naslov rada: Promena metoda mašinskog učenja za rangiranje individualnih sposobnosti

NR

Jezik publikacije: srpski (latinica)

JP

Jezik izvoda: s/e

JI

Zemlja publikovanja: Republika Srbija

ZP

Uže geografsko područje: Vojvodina

UGP

Godina: 2020.

GO

Izdavač: autorski reprint

IZ

Mesto i adresa: Novi Sad, Trg Dositeja Obradovića 4

MA

Fizički opis rada: 6 poglavlja, 77 strana, 17 lit. citata, 3 tabele, 42 slike

FO

Naučna oblast: matematika

NO

Naučna disciplina: primenjena matematika

ND

Ključne reči: Mašinsko učenje, fudbal, vrednovanje dodavanja, potklase, pristupi

PO

UDK

Čuva se: u biblioteci Departmana za matematiku i informatiku, Prirodno-matematičkog fakulteta, u Novom Sadu

ČU

Važna napomena:

VN

Izvod: Cilj rada je da pokaže da se pokaze povezanost i utićaj matematike na vrednovanje dodavanja u fudbalu. Rezultat je pronalazanje metoda i pristupa za vrednovanje dodavanja.

IZ

Datum prihvatanja teme od strane NN veća:

DP

Datum odbrane:

DO

Članovi komisije: dr Dušan Jakovetić, dr Nataša Krejić, dr Nataša Krklec Jerinkić

KO

Predsednik:

Član:

Član:

Član:

UNIVERSITY OF NOVI SAD
FACULTY OF SCIENCE
KEY WORDS DOCUMENTATION

Accession number:

ANO

Identification number:

INO

Document type: monograph type

DT

Type of record: printed text

TR

Contents code: master thesis

CC

Author: Nenad Rakić

AU

Mentor: dr Dušan Jakovetic

MN

Title: Changing machine learning methods to rank individual abilities

XI

Language of text: Serbian (latin)

LT

Language of abstract: s/e

LA

Country of publication: Republic of Serbia

CP

Locality of publication: Vojvodina

LP

Publication year: 2020.

PY

Publisher: author's reprint

PU

Publ. place: Novi Sad, Trg Dositeja Obradovica 4

PP

Physical description: 6 sections, 77 pages, 17 references, 3 tables, 42 figures

PD

Scientific field: mathematics

SF

Scientific discipline: applied mathematics

SD

Key words: Machine learning, football, evaluation of additions, subclasses, approaches

UC

Holding data: Department of Mathematics and Informatics' Library, Faculty of Sciences, Novi Sad

HD

Note:

N

Abstract: The aim of the paper is to show that mathematics is related and influenced by the value of adding football. The result is finding methods and approaches to evaluate addition.

AB

Accepted by the Scientific Board on:

ASB

Defended:

DE

Thesis defend board: dr Dušan Jakovetić, dr Nataša Krejić, dr Nataša Krklec Jerinkić

DB

President:

Member:

Member:

Member: