Neural betwork - predviđanje sportskih rezultata korišćenjem neuronskih mreža

Mladen Canović, Aleksandar Muljaić

15. juni 2019

Sažetak

Klađenje na sportske događaje predstavlja multimilijardersko tržište koje se neprestano razvija. Sa druge strane, neuronske mreže, kao jedna od metoda mašinskog učenja, predstavljaju jednu od gorućih tema današnjice u domenu veštačke inteligencije. Ovaj rad spaja te dve oblasti i pokušava da uz pomoć jedne, iskoristi pogodnosti druge oblasti. Prikazana je implementacija neuronskih mreža za predviđanje rezultata fudbalskih utakmica. Projekat je rađen u okviru kursa Računarska inteligencija, koji se drži na Matematičkom fakultetu, Univerzitet u Beogradu.

Sadržaj

1 Uvod

Predviđanje ishoda sportskih događaja zavisi od mnogo faktora: trenutne forme tima, fizičke pripreme igrača, mesta odigravanja utakmice, vremenskih uslova, itd. Iz tih razloga, rešavanje tog problema predstavlja veliki izazov. U ovom projektu istražujemo ponašanje neuronskih mreža primenjenih na ovaj problem.

Osnovna ideja rada je da se na osnovu podataka o prethodno odigranim utakmicama, a pomoću neuronskih mreža predvide ishodi predstojećih utakmica. Korišćeni su podaci iz Engleske Premijer lige (en. *Premier League*) od sezone 2000/2001 do 2018/2019. Ovaj problem predstavlja problem nadgledanog učenja (eng. *supervised learning*) [4], gde je za svaki ulaz trening skupa poznat i njegov izlaz tj. za svaku utakmicu prilikom treniranja je poznat i krajnji ishod utakmice.

Osvrt na tržište klađenja današnjice, postojeće modele i pretpostavka o moći neuronskih mreža za rešavanje ovog problema dat je u poglavlju 2. U poglavlju 3 biće bliže opisani podaci koji se koriste, kao i metode korišćene za njihovu obradu i pripremu za neuronske mreže. O primeni i zapažanjima o efikasnosti neuronskih mreža za predviđanje rezultata biće reči u poglavlju 4.

2 Profitabilnost neuronskih mreža

Klađenje na sportske događaje sve više uzima maha u modernom svetu. Razvoj računara je doveo do širenja ovog tržišta, omogućavajući razvoj *online klađenja* [2]. Sa porastom broja korisnika usluga klađenja, dolazi i do usavršavanja statističkih modela, radi maksimizovanja profita kladionica, koje tim modelima raspolažu. Preciznost u predviđanju sportskih događaja, koju su dostigli aktuari, razvojem savremenih statističkih modela i korišćenjem velikog broja podataka, dostiže približno 72% i stoga ih je jako teško dostići [6].

Razvoj računara je uzrokovao napredak drugih podoblasti računarstva, poput veštačke inteligencije. Mašinsko učenje, grana veštačke inteligencije, predstavlja jednu od najaktuelnijih grana računarstva. Ima široku primenu. Mašinsko učenje koristi različite podskupove algoritama za rešavanje problema veštačke inteligencije. Jedan od podskupova algoritama mašinskog učenja predstavljaju algoritmi zasnovani na principu neuronskih mreža [4]. Jedno od prvih istraživanja na temu korišćenja veštačkih neuronskih mreža u svrhu predviđanja sportskih rezultata sproveo je M.C.Puruker, čiji je model predviđao rezultate Američke fudbalske lige (en. *National Football League (NFL)*). Postigao je preciznost od približno 61%, što je bio dobar rezultat, ali dosta slabiji od rezultata modela stručnjaka, koji su, između ostalog, imali pristup većoj količini podataka [5]. Trenutni rezultati pokazuju da će preciznost modela neuronskih mreža teško dostići postojeće modele, koje koriste kladionice za predviđanje ishoda i određivanje kvota za sportske događaje [6]. U nastavku rada, ovaj problem je obrađen detaljnije, na primeru fudbalskih utakmica.

3 Podaci

Problem koji ovaj projekat pokušava da reši je uobičajeni problem mašinskog učenja, problem klasifikacije [8]. Predviđanje sportskih rezultata je upravo jedan takav problem. U pitanju je problem klasifikacije sa

3 klase (pobeda domaćina, nerešen rezultat, pobeda gostujućeg tima) [3]. Da bi se pristupilo ovom problemu potrebna je velika količina podataka iz kojih bi se korišćenjem algoritama klasifikacije moglo doći do određenih zaključaka [8].

3.1 Izvor podataka

Podaci koji su korišćeni za treniranje neuronske mreže su preuzeti sa veb stranice www.football-data.co.uk, u *CSV* formatu. Sačinjeni su od raznih parametara i statistika svih odigranih utakmica Premijer Lige od sezone 2000/01 - 2018/19. Radi lakše obrade podataka korišćena je *Pandas* biblioteka [9].

Svaki red u *CSV* datoteci odgovara jednoj utakmici. Atributi iz početnog skupa vrednosti koji su korišćeni se mogu videti u listingu 1.

```
'HomeTeam': Ime domaceg tima,
'AwayTeam': Ime gostujuceg tima
'FTR': Konacan rezultat
'FTHG': Broj postignutih golova domaceg tima,
'FTAG': Broj postignutih golova gostujuceg tima,,
'HS': Broj suteva domaceg tima,
'AS': Broj suteva gostujuceg tima,
'HST': Broj suteva domaceg tima u okvir gola,
'AST': Broj suteva gostujuceg tima u okvir gola,
'WHH': Kvota na pobedu domaceg tima,
'WHD': Kvota na neresen rezultat,
'WHA': Kvota na pobedu gostujuceg tima
```

Listing 1: Izabrani parametri

3.2 Obrada podataka

Obrada podataka je vršena tako što je za svaku utakmicu izračunavana statistika svakog od rivala iz njegovih prethodnih n utakmica (izabrana vrednost parametra n je 10). Takođe, izračunata je statistika svih timova učesnika Premijer lige u tom periodu. Rezultat obrade podataka je smešten u odgovarajuće *JSON* datoteke, koje sadrže podatke koji su u formatu lakšem za kasniju upotrebu. Proces obrade podataka se vrši pre pokretanja procesa treniranja neuronske mreže. Jednom obrađene podatke nije potrebno ponovo obrađivati.

Koristeći odabrane atribute formira se vektor odgovarajuće dužine koji se sastoji od statistike domaćeg i gostujućeg tima. Prva polovina formiranog vektora odgovara statistici tima koji je na toj utakmici bio domaćin. Izračunata je na osnovu postignutih rezultata tog tima na osnovu poslednjih 10 utakmica koje je taj tim odigrao. Analogno tome, druga polovina odgovara statistici gostujućeg tima. Ukoliko za jedan od timova ne postoji istorija o prethodnih 10 utakmica ta utakmica se dalje ne razmatra.

U prvoj iteraciji neuronske mreže (4.2.1), ovaj vektor je korišćen kao ulaz za neuronsku mrežu. Za ostale iteracije vektor je dodatno obrađen, tako što se vrednosti koje se u njemu nalaze, koriste u formuli za računanje forme tima. Rezultat tog izračunavanja je vrednost u pokretnom zarezu. Za svaki tim, računa se snaga tog tima na domaćem, odnosno gostujućem terenu. Svi parametri se mogu podešavati korišćenjem koeficijenata koji su korišćeni za prilagođavanje parametara - što je koeficijent veći, značaj parametra raste u konačnoj formuli. Postavljanjem koeficijenta na 0, parametar se poništava. Izračunate vrednosti se koriste u procesu treniranja neuronske mreže.

3.3 Statistički parametri

Korišćene su sledeće osnovne ili izvedene statistike, koje su računate korišćenjem dotad obrađenih podataka. Za sve naredne podatke se podrazumeva da su prikupljeni iz prethodnih 10 utakmica u odnosu na posmatranu utakmicu.

- *home team wins* - Broj pobeda domaćeg tima.
- *home team draws* - Broj nerešenih mečeva domaćeg tima.
- *home team losses* - Broj poraza domaćeg tima.
- *home team goals scored* - Broj postignutih golova domaćeg tima.
- *home team goals conceded* - Broj primljenih golova domaćeg tima.
- *home team shots* - Broj šuteva domaćeg tima.
- *home team shots on target* - Broj šuteva domaćeg tima u okvir gola.
- *home team shots opposition* - Broj šuteva protivničkog tima.
- *home team shots opposition on target* - Broj šuteva protivničkog tima u okvir gola.
- **Analogni podaci prikupljeni su za gostujući tim posmatrane utakmice.**
- *result* - Ishod meča ('H' - Pobeda domaćina, 'D' - Nerešeno, 'A' - Pobeda gosta). Ovaj podatak je korišćen kao labela u procesu učenja.

Sve utakmice za koje postoje validni podaci i pre koje su oba tima odigrala 10 ili više utakmica, ulaze u skup podataka za treniranje neuronske mreže. Ostali podaci se odbacuju. Podaci se smeštaju u *JSON* datoteke radi lakše upotrebe u skriptu, koji je korišćen za treniranje i testiranje neuronske mreže. Podaci se parsiraju pokretanjem skripta *data_parser.py*. Rezultat izvršavanja ovog skripta su *JSON* datoteke *processed_teams_stats.json* i *processed_data.json*.

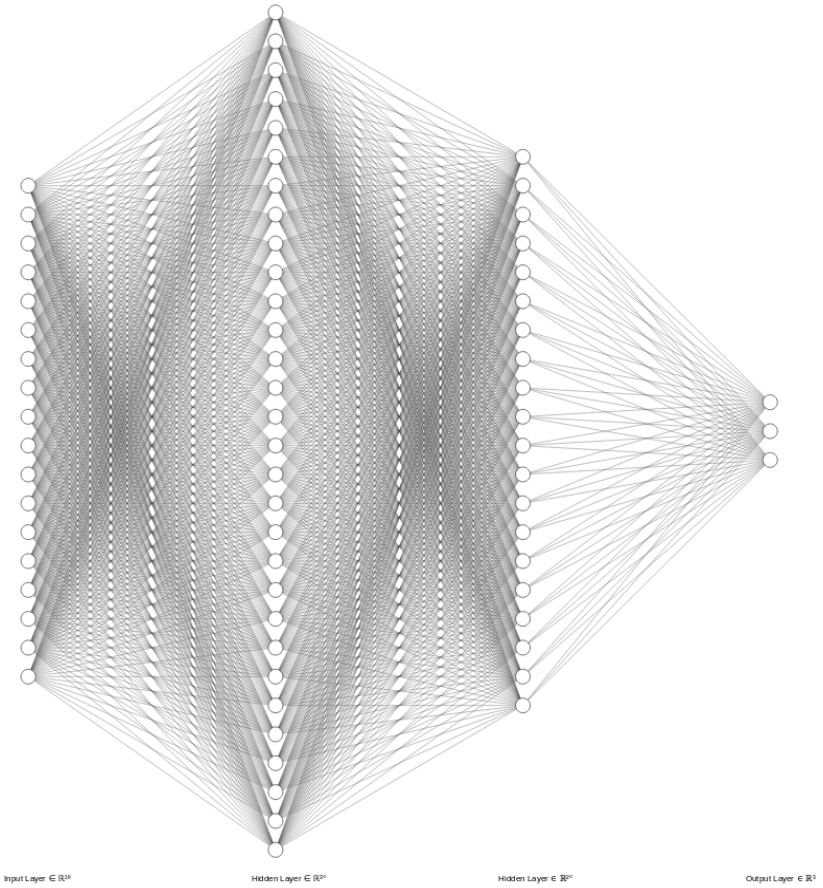
4 Treniranje i testiranje neuronske mreže

Projekat je rađen u Python programskom jeziku. Za kreiranje neuronske mreže korišćena je Keras biblioteka [1], kao i Hyperas za izbor optimalnih vrednosti parametara. Obrađeni podaci su podeljeni tako da je 70% korišćeno za treniranje, a 30% za testiranje.

4.1 Struktura neuronske mreže

Struktura mreže prikazana je na slici 1. Prolazila je kroz više iteracija, pa je broj čvorova ulaznog i skrivenih slojeva menjan kroz iteracije.

Sastoji se od 10-18 ulaznih čvorova gde svaki predstavlja jedan od atributa opisanih u odeljku 3. Sadrži dva skrivena sloja od kojih se svaki sastoji od po 10-30 čvorova i izlazni sloj od 3 čvora koji predstavljaju redom: pobedu domaćina, nerešeno ili pobedu gosta.



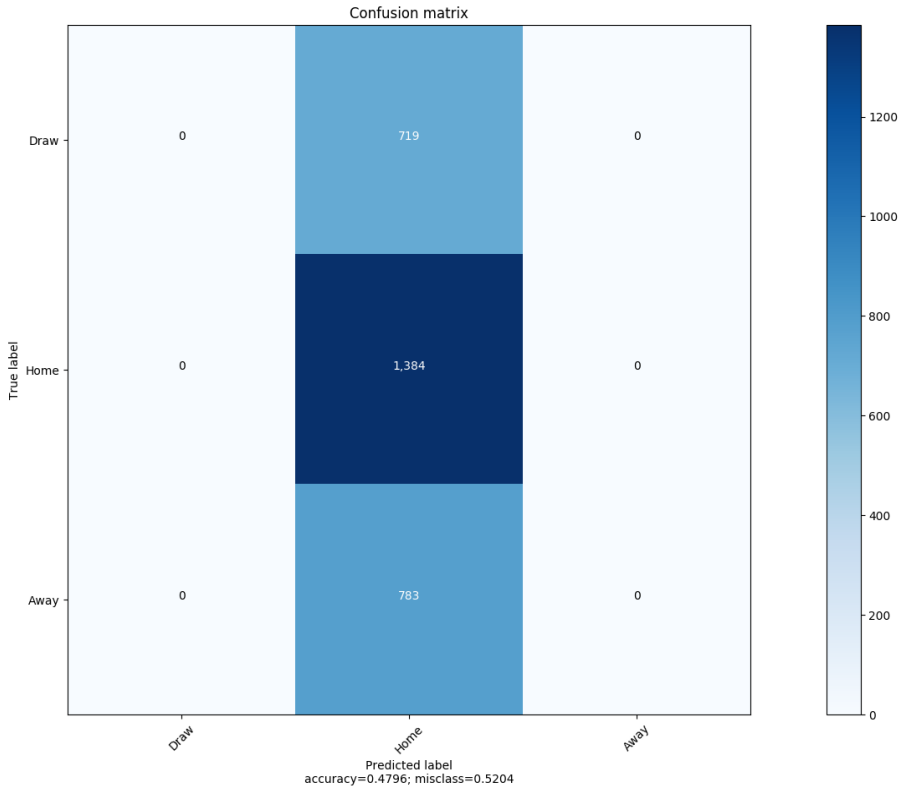
Slika 1: Struktura neuronske mreže

4.2 Iteracije neuronske mreže

U nastavku su prikazane iteracije naše neuronske mreže, uz dodavanje i promenu atributa sa ciljem dobijanja što boljih rezultata. Uprkos promenama i variranju parametara, preciznost nije umnogome poboljšana u odnosu na prvobitni model.

4.2.1 Prva

Prva iteracija je sačinjena od 10 ulaznih čvorova koja obuhvata broj pobeda, poraza, nerešenih utakmica kao i broj datih i primljenih golova, kako domaćina, tako i gosta. Dobijena preciznost na treningu je 0.456, a na testu je 0.4795. Matrica konfuzije ovog modela prikazana na slici 2 jasno pokazuje dominantnost predviđanja pobede domaćeg tima.



Slika 2: Matrica konfuzije

4.2.2 Druga

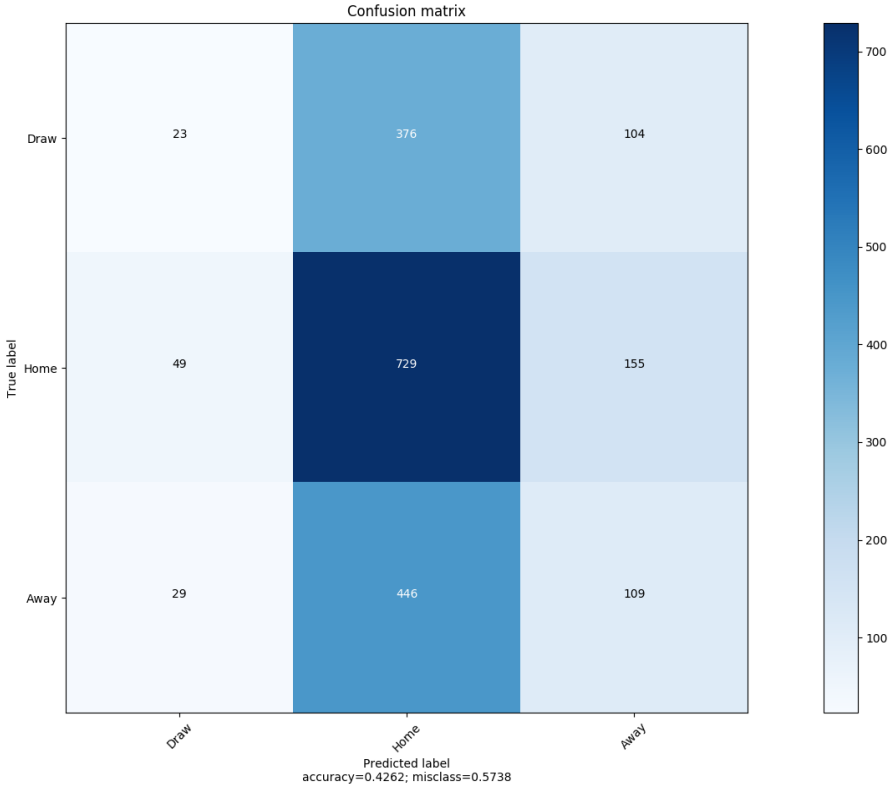
U drugoj iteraciji je dodato još 8 atributa: šutevi, šutevi u okvir gola, šutevi protivničkog tima, šutevi protivničkog tima u okvir gola. Atributi su dodati i za domaći i za gostujući tim. Cilj dodavanja je bio da se vidi kako će se neuronska mreža ponašati i da li će dodatni atributi koje smo smatrali relevantnim uticati na preciznost, kao i na izbor predviđene klase. Međutim, obe mere su ostale nepromenjene, kao i dominantna klasa.

4.2.3 Treća

Problem sa prethodne dve iteracije je što je model skoro isključivo predviđao *1*, tj. pobedu domaćeg tima. Zaključak je bio da su neke težine grana velike, što je prouzrokovalo nestabilnost mreže. Vrednost grana se minimalno menjala, bez uticaja na izlaz. Dodata je beč normalizacija (eng. *batch normalization*) na unutrašnje slojeve, koja normalizuje izlaz iz aktivacione funkcije. Ovim je početni problem dominantnosti jedne klase bio rešen, ali se preciznost na test skupu pogoršala, dok je na trening skupu ostala ista. Došlo je do preprilagođavanja. Povećan je broj epoha i uključena je *dropout* tehnika. *Dropout* je tehnika kojom se rešava problem preprilagođavanja. Ideja je da se neki, slučajno izabrani čvorovi, isključe prilikom treniranja [7]. Ovim je popravljena preciznost na test skupu, a na trening skupu je ostala nepromenjena, što pokazuje i matrica konfuzije na slici 3. Na slici se vidi da model i dalje najčešće predviđa pobedu domaćeg

tima (što je, zaključuje se, dominantna osobina ovog skupa podataka), ali da nakon izmena predviđa i ostale slučajeve, naspram prethodnih modela gde je na test skupu najčešće predviđao pobedu domaćeg tima 2.

Dobijena je niža preciznost u odnosu na početni model, u proseku, ali su predviđanja postala heterogena. Međutim, prilikom višestrukog testiranja primećena je nestabilnost u rezultatima. Preciznost na test skupu je opadala i do 0.32%, a rasla do 0.52%, dok se i dominantna klasa prilikom predviđanja menjala.



Slika 3: Matrica konfuzije

5 Zaključak

U ovom radu je prikazan način obrade, kao i rezultati dobijeni korišćenjem neuronskih mreža za predviđanje sportskih rezultata. Veliki je izazov odabrati odgovarajući skup atributa, kako bi se pristupilo rešavanju ovog problema. Iako daleko od rezultata koji postižu modeli razvijeni od strane vodećih timova na tržištu, rezultati su ohrabrujući za dalje istraživanje i unapređivanje postojećeg rešenja. Korišćenje statističkih, kao i drugih metoda mašinskog učenja, će u kombinaciji sa opisanim pristupom doprineti poboljšanju modela.

Literatura

[1] François Chollet. Keras, 2018. on-line at: https://keras.io/.

[2] Anthony Constantinou and Norman Fenton. Profiting from arbitrage and odds biases of the european football gambling market. *Journal of Gambling Business and Economics*, 2013.

[3] D. Harlili D. Prasitio. Predicting football match results with logistic regression, in: Proceedings of the 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), 2016.

[4] Andries P Engelbrecht. *Computational intelligence: an introduction*. John Wiley & Sons, 2007.

[5] M.C. Purucker. Neural network quarterbacking. *IEEE Potentials*, 15:9 – 15, 1996.

[6] Fadi Thabtah Rory P. Bunkera. A machine learning framework for sport result prediction. *Applied Computing and Informatics*, 2017.

[7] Nitish Srivastava and Geoffrey Hinton. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 2014.

[8] Pang-Ning Tan. *Introduction to data mining*. Pearson Education India, 2018.

[9] Wes McKinney & PyData Development Team. pandas: powerful Python data analysis toolkit, 2019. on-line at: https://pandas.pydata.org/pandas-docs/stable/.