Predikcija ishoda fudbalskih utakmica

Uvod

Predikcija utakmica se sprovodi u okviru engleske Premier lige koja je poznata kao vrlo nepredvidiva liga koju prate ljudi širom sveta. Predikcija se odnosi na odabir jednog od tri ishoda (1,X,2) za 2 odabrane ekipe. Rešenje problema je predstavljeno kroz 6 algoritama iz oblasti mašinskog učenja i neuronskih mreža

Prikupljanje i obrada podataka

Podaci su izvorno prikupljeni sa sajta
http://www.football-data.co.uk/englandm.php
Podaci su neobrađeni I sortirani u torke po fajlovima za svaku sezonu posebno. Fajlovi sadrže preko 70 obeležja za svaku torku.

I. Trening I test podaci

Podaci koji su obuhvaćeni su podaci od sezone 2006./2007.
Trening podaci predstavljaju svi podaci do sezone 2016./2017.
Test podatke predstavlja sezona 2016./2017.
Procenat uspešnosti je procentualno izražen brojem pogođenih ishoda za sve utakmice iz test podataka.

II. Obrada numeričkih podataka

Svi podaci se skupljeni u jedan fajl. Odatle uzimamo samo osobine koje su nam od interesa za algoritme. Pošto su obeležja rasprostranjena po različitim opsezima potrebna je izvršiti normalizaciju tih podataka, odnosno svesti ih na isti opseg.

III. Obrada nenumeričkih podataka

Jedini nenumerički podatak koji smo koristili jesu sudije. Obradu smo vršili tako što smo odredili koliko puta je ekipa pobedila kada je sudio određeni sudija.

Uticaj sudija na rezultat vršen je na osnovu sledeće funkcije:
zbirDomaciPobede*2 > (zbirGostiPobede + zbirNereseno)

```
R Beeby,0,0,1
C Pawson,27,19,13
B Knight,4,0,0
A Wiley,68,32,44
M Clattenburg,122,60,71
P Walton,63,43,41
L Probert,73,32,40
G Poll,29,14,17
M Dean,138,89,86
S Tanner,19,2,4
N Swarbrick,33,25,29
H Webb,122,69,74
D Gallaghe,0,0,0
G Scott,1,0,3
S Hooper,0,0,0
L Mason,100,46,57
S Attwell,20,20,8
P Tierney,0,3,1
J Moss,56,21,36
A Taylor,53,33,50
M Riley,50,27,24
K Friend,59,34,42
I Williamson,0,0,0
S Bennett,61,36,41
U Rennie,23,13,13
Mn Atkinson,1,0,0
M Jones,79,47,45
C Foy,114,47,65
R Madley,19,12,11
St Bennett,0,0,0
M Halsey,69,33,56
P Dowd,116,64,69
A Marriner,102,59,62
A D'Urso,0,0,0
D Gallagh,0,0,0
R Styles,51,26,33
D Gallagher,18,8,5
R East,28,13,10
K Stroud,7,5,4
M Atkinson,138,66,83
M Oliver,64,46,30
```

Metodologija

I. Normalizacija podataka

Normalizacija podataka je vršena po sledećoj forumli:

$$B = ((A - \min(A)) / (\max(A) - \min(A))) * (D - C) + C$$

Obeležje A se normalizuje na opseg od C-D

```
Div = League Division
Date = Match Date (dd/mm/yy)
HomeTeam = Home Team
AwayTeam = Away Team
FTHG = Full Time Home Team Goals
FTAG = Full Time Away Team Goals
FTR = Full Time Result (H=Home Win, D=Draw, A=Away Win)
HTHG = Half Time Home Team Goals
HTAG = Half Time Away Team Goals
HTR = Half Time Result (H=Home Win, D=Draw, A=Away Win)


B1 = (A - 0) / (9 - 0) * ( 10 - (-10) ) + (-10)         -gol domacin
B1 = (A - 0) / (7 - 0) * ( 10 - (-10) ) + (-10)         -gol gost
0.5*B1 = ((A - 0) / (5 - 0) * ( 10 - (-10) ) + (-10))*0.5       -poluvreme gol domacin
0.5*B1 = ((A - 0) / (5 - 0) * ( 10 - (-10) ) + (-10))*0.5       -poluvreme gol gost
B1 = (A - 1) / (43 - 1) * ( 10 - (-10) ) + (-10)                -sutevi domacin
B1 = (A - 0) / (30 - 0) * ( 10 - (-10) ) + (-10)                -sutevi gost
B1 = (A - 0) / (24 - 0) * ( 10 - (-10) ) + (-10)                -sutevi u okvir domacin
B1 = (A - 0) / (20 - 0) * ( 10 - (-10) ) + (-10)                -sutevi u okvir gost
0.125*B1 = ((A - 1) / (33 - 1) * ( 10 - (-10) ) + (-10))*0.125 i                -faul domacin
0.125*B1 = ((A - 1) / (26 - 1) * ( 10 - (-10) ) + (-10))*0.125  j       -faul gost
0.125*B1 = ((A - 0) / (20 - 0) * ( 10 - (-10) ) + (-10))*0.125  k       -korne domacin
0.125*B1 = ((A - 0) / (19 - 0) * ( 10 - (-10) ) + (-10))*0.125  l       -korne gost
0.125*B1 = ((A - 0) / (7 - 0) * ( 10 - (-10) ) + (-10))*0.125   m       -faul domacin
0.125*B1 = ((A - 0) / (9 - 0) * ( 10 - (-10) ) + (-10))*0.125   n       -faul gost
0.125*B1 = ((A - 0) / (2 - 0) * ( 10 - (-10) ) + (-10))*0.125   o       -korne domacin
0.125*B1 = ((A - 0) / (2 - 0) * ( 10 - (-10) ) + (-10))*0.125   p       -korne gost
B1 = (((10 - (-10))/(17 - 1.09)) * (A - 17) + 10)*(-1)   - kvota domacin
B1 = (((10 - (-10))/(29 - 1.16)) * (A - 29) + 10)*(-1)   - kvota gost


Match Statistics (where available)
Attendance = Crowd Attendance
Referee = Match Referee
HS = Home Team Shots
AS = Away Team Shots
HST = Home Team Shots on Target
AST = Away Team Shots on Target
HHW = Home Team Hit Woodwork
AHW = Away Team Hit Woodwork
HC = Home Team Corners
AC = Away Team Corners
HF = Home Team Fouls Committed
AF = Away Team Fouls Committed
HO = Home Team Offsides
AO = Away Team Offsides
HY = Home Team Yellow Cards
```

II. Selekcija osobina I dodatni korektivni faktori

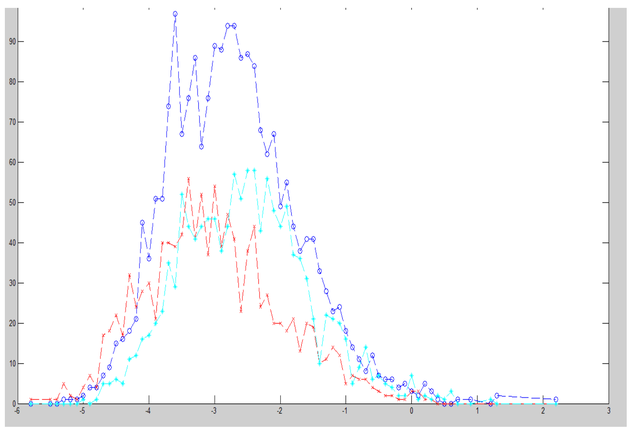
Za La Granževu interpolaciju, regresiju I kNN metodu koriti se meč rejting. Njega računamo na osnovu srednje vrednosti svih numeričkih osobina.

Izuzetak od pravila jesu meč rejtinzi za klubove koji imaju višegodišnju tradiciju u ligi. Njihove rejtinge množimo korektivnim faktorom, koji im daje prednost u odnosu na druge ekipe,

Za Tree Decision I Random Forrest koristi se I nenumerički podatak, dok se za neuronsku mrežu koristi rejting ekipa iz igrice Fifa 17.

III. La Granževa Interpolacija i Regresija Metodom najmanjih kvadrata

Plotovanjem podataka dobijaju se sledeće krive:



*Tamno plava pobede domaćina
*Svetlo plava pobede gosta
*Crvena nerešeno

*X osa predstavlja meč rejting
*Y osa predstavlja koliko puta se desio meč rejting za određeni ishod

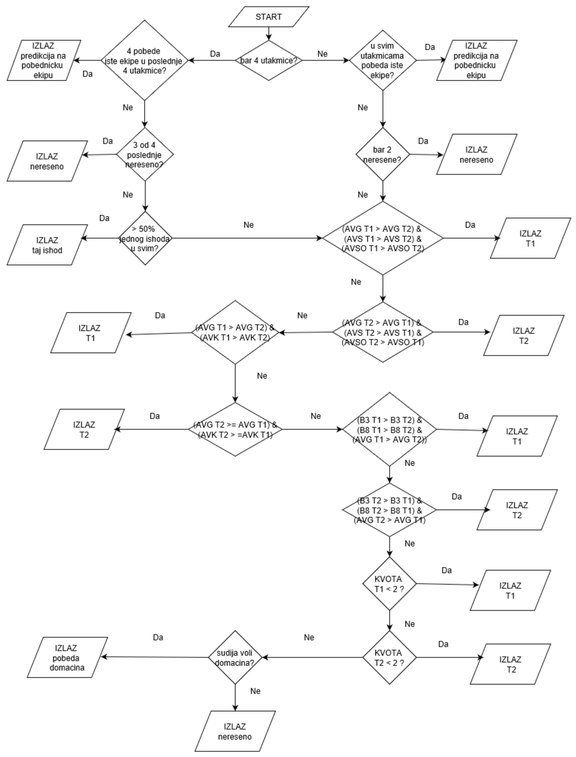
Metode pokušavaju da prilagode krivu datim ulazima. La Granževa interpolacija pokušava da prilagodi krivu prolaskom kroz sve tačke, dok regresija to ne zahteva.

IV. KNN

KNN metoda nalazi K najbližih tačaka koje imaju najblže meč rejtinge, nalazi njihovu aritmetičku sredinu I na osnovu toga određuje rezultat.
K je promenljivo I moguće je menjati ga u okviru algoritma

V. Tree Decision I Random forrest

Tree decision je algoritam iz oblasti mašinskog učenja koji ima više nivoa na kojima pokušava da donese odluku o pripadnosti ulaznog seta podataka određenoj grupi. U svakom čvoru podaci prolaze kroz selekciju po određenom obeležju/obeležjima. Čvorovi odnosno osobine mogu da imaju veću ili manju uspešnost. Nju je moguće izračunati matematički preko entropija (mere neuređenosti skupa).



Formula za računanje entropije :

$$Entropy(S) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

Random Forrest algoritam radi na osnovu slučajnog pogađanja ulaznog seta utakmica, kao I slučanog pogađanja metoda koje će se koristiti. Formira se ansambl klasifikatora koji daju pojedinačna rešenja. Računa se aritmetička sredina dobijenih predikcija.

Zbog mogućnosti da su ekipe igrale malo utakmica, uzima se slučajnim odabirom 2 utakmice iz skupa svih odigranih utakmica.i bira se n puta jedna od mogućih 6 funkcija.

Algoritam je fleksibilan I lako se može promeniti broj klasifikatora koji se formiraju.

Korišćene su sledeće funkcije:
1. Razlika u golovima
2. Razlika u šutevima na gol
3. Razlika u šutevima u okvir gola
4. Razlika u kvotama
5. Razlika u kornerima
6. Da li sudija preferira domaćine

VI. Neuronska mreža

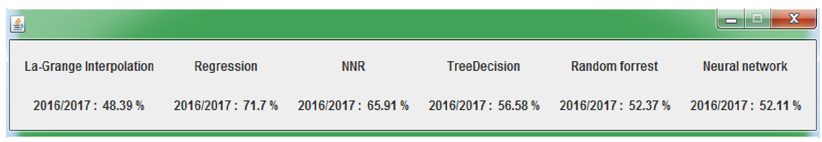
Korišćena je sigmoidna funkcija kao aktivaciona funkcija I jedan skriveni sloj. Rešenje je fleksibilno I lako se može izmeniti arhitektura neuronske mreže I prilagoditi I drugačijim ulaznim podacima. Takođe, moguće je eksperimentisati I sa slojevima I brojem neurona u njima.

Ima 8 ulaza I 3 izlaza. Trening podaci su zadati u manjem obimu nego ostalim metodama jer bi učenje takvog modela trajalo previše dugo. Svakako da je ovo jedna od tačaka koja predstavlja prostor za napredak.

Verifikacija

Svaka od metoda daje zadovoljavajuće rezultate ako imamo u vidu da je engleska liga vrlo nepredvidiva I da se kao odluke uzimaju sve predikcije. Čak I one koje su oko 35-40% (algoritmi LaGranž, Regresija I KNN).
Od mogućih ishoda, metode najteže pogađaju nerešen ishod I tu se dešava najviše promašaja. Najviše pogodaka se dešava na predikciju pobede domaćina.

Analiza uspeha vršena je na osnovu procentualnog broja pogodaka rezultata sezone 2016/2017, na osnovu podataka od 2006.-2016 godine. Kao najbolja metoda, pokazala se Regresija.
Njena prednost se vidi posmatranjem vizuelizacije podataka. Regresiona kriva pobeda domaćina je znatno “viša” od ostale dve I rezultati su dobro razdvojeni.
Budući da se oko 40-50% utakmica godišnje završi pobedom domaćina, preostalih 20% je odabir između dve krive koje su slične.



La-Grange Interpolation	Regression	NNR	TreeDecision	Random forrest	Neural network
2016/2017 : 48.39 %	2016/2017 : 71.7 %	2016/2017 : 65.91 %	2016/2017 : 56.58 %	2016/2017 : 52.37 %	2016/2017 : 52.11 %

Profesor: Đorđe Obradović
Asistent: Miroslav Kondić
Studenti: Marko Tagliavia(RA 136/2014), Tijana Lalošević(RA 3/2014)