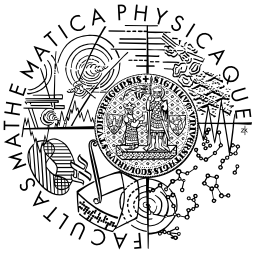

MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Daniel Šipoš

Predikce sportovních utkání pomocí neuronových sítí

Katedra softvéru a výuky informatiky

Vedoucí bakalářské práce:	Mgr. David Kuboň
Studijní program:	Informatika
Studijní obor:	Obecná informatika

Praha 2019

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Poděkování.

Název práce: Predikce sportovních utkání pomocí neuronových sítí

Autor: Daniel Šipoš

Katedra: Katedra softvéru a výuky informatiky

Vedoucí bakalářské práce: Mgr. David Kuboň, Katedra softvéru a výuky informatiky

Abstrakt: Futbal a tenis patria k najpopulárnejším športom na tejto planéte. Platí to hlavne vďaka jednoduchosti pravidiel a nenáročnosti na vybavenie. Obe športy môže hrať prakticky ktokoľvek. Sú ale ľudia, ktorým tieto športy idú lepšie ako ostatným, takzvaní profesionáli. Títo profesionáli potom chodia po rôznych turnajoch, resp. ligách, kde hrajú zápasy, aby sa ukázalo, kto je najlepší. Ľudská chamtivosť na jednej strane a závislosť na hazarde na druhej podporujú vznik rôznych spoločností, stávkových kancelárií, ktoré umožňujú tipovať výsledky týchto zápasov za peniaze. Táto práca sa zameriava na predpovedanie takýchto výsledkov futbalových líg a tenisových turnajov pomocou dvoch mierne odlišných druhov neurónových sietí, porovnanie jednotlivých predpovedných modelov a taktiež porovnanie predvídateľnosti futbalu a tenisu.

Klíčová slova: neurónová sieť rekurentná neurónová sieť športové stávkovanie športové kurzy futbal tenis

Title: Prediction of sports results using neural networks

Author: Daniel Šipoš

Department: Department of Software and Computer Science Education

Supervisor: Mgr. David Kuboň, Department of Software and Computer Science Education

Abstract: Abstract.

Keywords: neural network recurrent neural network football tenis sport betting sport odds

Obsah

Úvod

Šport je súčasťou zábavného priemyslu hlavne pre relatívnu nepredvítateľnosť jeho výsledkov. Stať sa môže v podstate čokoľvek. Vyhrať môže favorit udalosti alebo osoba/tím, od ktorej sa o vôbec neočakávalo. Môže začať pršať alebo na ihrisko vbehnúť exhibicionista s kontroverznou myšlienkou.

Táto nepredvídateľnosť podnietila vznik stávkových kancelárií, ktoré na tieto a na rôzne ďalšie udalosti vypisuje kurzy, ktoré v prípade, že tieto udalosti nastanú, zaručia stávkujúcemu výhru. Tieto stávkové kancelárie ročne zarábajú milióny tým, ako vypisujú kurzy, aby boli lákavé pre bežných ľudí. V podstate sa snažia uhádnuť, s akou pravdepodobnosťou nastane daná udalosť, napríklad predikovať výsledok. Stávkové kancelárie určite používajú na tieto odhady nejaké data, ale pravdepodobnosti daných udalostí zvykne predpovedať odborník, bookmaker. Je možné nájsť nejakú množinu dát, na základe ktorej vieme naučiť počítač predikovať výsledky jednotlivých športových udalostí s určitou presnosťou?

Súvisiace práce

V minulosti boli použité rôzne metódy na predikciu športových výsledkov. V roku 2005 sa o predpoveď 6 rôznych udalostí týkajúcich sa austrálskej kriketovej ligy a AFL, ligy v austrálskom futbale, pokúsil Bailey. Na austrálsky futbal použil data zo zápasov zo 100 sezón odohraných pred rokom 1997 a testoval to na zápasoch od sezóny 1997 do 2003 použitím rôznych modelov lineárnej regresie. Dokázal získať presnosť 66.7% (Bailey a kol., 2005).

V roku 2006 Joseph, Fenton a Neil vyskúšali viaceré druhy strojového učenia na predikciu výsledkov zápasov tímu Tottenham Hotspur F.C. v najvyššej anglickej futbalovej lige, Premier League, v sezónach 1995/1996 a 1996/1997. To znamená, že pracovali s datasetom o veľkosti 76 zápasov, z ktorej časť delili na trénovacie a časť na testovacie data. Použité metódy zahŕňali expertmi konštruované bayesovské siete, naivný bayesovký klasifikátor, rozhodovacie stromy a k-NN. Použili pri tom 30 príznakov, ale až 28 sa viazalo iba na to, či daný hráč nastúpil od začiatku na daný zápas alebo nie, zvyšné dva predstavovali silu súpera a miesto zápasu (či hral predikovaný tím na domácom štadióne alebo nie). V tomto prípade dosiahli bayesovské siete úspešnosť niečo vyše 59%, zvyšné metódy sa pohybovali medzi 30 – 38% pri disjunktných testovacích a trénovacích datach (Joseph a kol., 2006).

V roku 2011 sa dvojica Hucaljuk a Rakipović zameriavala na výber príznakov pri predikcii výsledkov futbalovej Ligy majstrov. Pracovali s datami z 96 zápasov, ktoré manuálne ohodnotili podľa 30 príznakov. Vybrané príznaky predstavovali formu oboch tímov v posledných 6 zápasoch, výsledok posledného vzájomného zápasu týchto dvoch tímov, postavenie v rebríčku, počet zranených hráčov a priemerný počet strelených a inkasovaných gólov. Neskôr zúžili počet príznakov na 20 a na novovzniknutý dataset bolo aplikovaných 6 rôznych metód strojového učenia, naivný bayesovský klasifikátor, bayesovské siete, LogitBoost, k-NN, Random forest a neurónové siete. Najvyššia dosiahnutá úspešnosť bola 68%, dosiahli ju neurónové siete (Hucaljuk a Rakipović, 2011).

V roku 2016 vyskúšali logistickú regresiu na predikciu výsledkov futbalovej

Premier League výskumníci okolo Prasetia. Stavali na výsledkoch svojich predchodcov a vybrali 4 príznaky, ktoré hrali v predchádzajúcich prácach najväčšiu rolu, konkrétne ohodnotenia pre obranu a útok, pre domácich aj hostí. Dosiahli úspešnosti v najlepšom prípade 69,5%.(Prasetio a kol., 2016)

V roku 2014 použili Igiri a Nwachukwu nástroj, ktorý zvaný Rapid Miner. Jeho úlohou bolo predikovať výsledky anglickej Premier League. Použité techniky boli popredná neurónová sieť a lineárna regresia. Neurónová sieť dosiahla úspešnosti 85%, lineárna regresia 93%. Je potrebné dodať, že neurónová sieť predpovedala všetky typy výsledkov (výhra domácich, prehra, remíza), zatiaľ čo regresia predpovedala len zápasy, ktoré sa v konečnom dôsledku skončili výhrou alebo prehrou domáceho celku, takže celková úspešnosť bola o niečo nižšia. Autori dodali, že ak sa predpokladá, že zápas môže skončiť aj remízou, tak neurónové siete mali lepšie výsledky. K predikcii použili rôzne príznaky vrátane kurzov, priemerný počet striel, striel na bránu, rohových kopov, ale aj abstraktnejšie príznaky ako ofenzívna/defenzívna sila mužstva a ohodnotenie sily jednotlivých hráčov a kvality manažéra (Igiri a Nwachukwu, 2014).

V tom istom roku sa Shin a Gasparyan pokúsili nájsť nové metódy predikcie. Navrhli použiť data z videohry FIFA 2015 na predikciu španielskej La Ligy. Použitie tohto návrhu odôvodnili tým, že vydavatelia videohier v dnešnej dobe pracujú na tom, aby boli ich hry čo možno najreálnejšie. To sa hlavne týka športových hier, kde je dôležité, aby bol každý hráč ohodnotený čo možno najpresnejšie, aby sa to podobalo realite. FIFA 2015 používa rôzne atribúty na ohodnotenie hráča ako napríklad zrýchlenie, strely z diaľky alebo reflexy pre post brankára. Tieto data sa získavajú oveľa jednoduchšie ako z iných zdrojov. Autori vytvorili dva typy modelov: učenie s učiteľom (supervised learning) a bez učiteľa (unsupervised learning). Pri učení s učiteľom vytvorili 2 prístupy, reálny prediktor, ktorý využíval reálne data a virtuálny prediktor, ktorý využíval práve data z popísanej videohry. Obe využívali logistickú regresiu a metódu podporných vektorov (support-vector machine). Reálny prediktor dosiahol úspešnosť 75%, virtuálny 80%, čo podľa autorov dokazuje, že data získané z videohier sa dajú používať aj v reálnom svete. Učenie bez učiteľa analyzovalo stratégie tímov podľa typov hráčov, ktorí sú v danom tíme pomocou k-means clusteringu. Zistili, že lepšie tímy zvyknú mať útočnejšie stratégie a slabšie tímy dokážu uhrať lepšie výsledky proti silnejším tímom, ak majú defenzívnejšiu stratégiu. (Shin a Gasparyan, 2014).

Čím sa táto práca líši

V tejto práci budeme predikovať futbal a tenis pomocou popredných a rekurentných neurónových sietí. Tenis nie je predikovaný v žiadnej z prác spomínaných v predchádzajúcej sekcii. Futbal je síce predikovaný, ale ani raz štýlom, aký bude prezentovaný v tejto práci.

Väčšina prác má oveľa menšiu trénovaciu vzorku pre siete. Pre túto prácu boli použité informácie z viac ako 5000 futbalových zápasov, z toho trénovacia množina tvorila viac ako 3000 vstupov pre každú ligu. Pre tenis obsahuje dataset viac ako 6200 riadkov a je vytvorený z informácií z viac ako 55000 zápasov.

Ďalšou vecou, ktorou sa táto práca odlišuje od ostatných je to, aké data sú použité. V tejto práci budú použité výhradne výsledky a prostredie zápasov, z ktorých sú následne kalkulované ostatné informácie. Nebudú použité abstraktné

data ako sila hráčov alebo tímu ani ohodnotenia žiadnych hráčov ako ani data o počte rohových kopov, žltých kariet, es alebo nevynútených chýb.

1. Základné pojmy

1.1 Futbal

Futbal je šport, pri ktorom na hracej ploche, futbalovom ihrisku, proti sebe nastúpia dva jedenásťčlenné tímy s cieľom skórovať čo najviac gólov a inkasovať čo najmenej. Na ihrisku je vždy najviac jedna lopta, hráči ju ovládajú prevažne nohami. Gól nastáva, keď jeden s tímov pošle loptu celým objemom za bránkovú čiaru do priestoru medzi bránkové tyče, teda do súperovej bránky, vrámci pravidiel. Víťazom sa stáva tím, ktorý strelí viac gólov ako súper. Ak je počet vstrelených gólov pre obe zúčastnené strany rovnaký, nastáva remíza (Táborský, 2004).

1.1.1 Futbalové ligy

(football league rules)

1.2 Tenis

Tenis je šport tímov súperiacich proti sebe, skladajúcich sa z jedného alebo dvoch ľudí, hrajúcich proti sebe na tenisovom kurte. Zápasy sa delia na dvojhry, teda zápasy dvoch jednočlenných tímov, a štvorhry, zápasy dvoch dvojčlenných tímov. Hlavným cieľom tenisu je použiť tenisovú raketu na zahratie loptičky na súperovu stranu kurtu jedným úderom tak, aby mala súperiaca strana, čo najväčší problém ho vrátiť naspäť (Koromházová, 2008). Ak sa to jednému z tímov nepodarí v súlade s pravidlami, súper získa bod. Tím, ktorý získa 4 body, získa hru. Ak obe tímy získajú po 3 body skôr, ako jeden z nich získa 4, hru získa tím, ktorý získa o 2 body viac ako súper. Tím, ktorý skôr získa 6 hier, získa sadu. Ak nastane stav 5:5, hru získa tím, ktorý získa 7 hier. Zápas sa hrá na dve alebo tri víťazné sady, toto číslo je vždy určené vopred. V tejto práci nás budú hlavne zaujímať dvojhry, teda zápasy jeden proti jednému. (Táborský, 2005).

1.2.1 Okruh ATP

(ATP rules)

1.3 Porovnanie futbalu a tenisu

Z predchádzajúcich kapitol je zrejmé, že futbal a tenis majú veľa spoločných a veľa rozdielnych vlastností. Futbal je kontaktný šport, teda protihráči sú často vo fyzickom kontakte medzi sebou, zatiaľ čo pri tenise sú protihráči vždy na opačných stranách tenisového kurtu. Rozdielny je aj počet hráčov v jednom tíme, vo futbale je maximálny počet hráčov hrajúcich v jednom momente za jeden tím 11,

v tenise to je buď jeden alebo dvaja. Spoločný je napríklad fakt, že sa jedná o loptový šport. Na druhej strane, vo futbale je povolené loptu zasiahnuť ktoroukoľvek časťou tela okrem rúk (s výnimkou brankára), v tenise je zakázané dotknúť sa tenisovej loptičky akoukoľvek časťou tela, loptičku je povolené zahrať len tenisovou raketou. Ďalším rozdielom je hrací čas. Vo futbale má každý zápas fixnú dĺžku (2 polčasy po 45 minút s maximálne 15 minútovou prestávkou medzi nimi), rozhodca na konci každého polčasu nadstaví čas, po ktorý sa nehralo kvôli rôznym prerušeniam v hre (Táborský, 2004). V tenise môže zápas vďaka pravidlám trvať od desiatok minút do niekoľko hodín (Koromházová, 2008).

1.4 Kurzy stávkových na kancelárií

Kurzové stávky sú stávky na akýkoľvek jav, na ktorý vypíše daná stávková kancelária kurz. Kurzy stanovuje bookmaker podľa toho, aká je pravdepodobnosť, že daný jav nastane, kde platí, že čím nižší kurz, tým je vyššia pravdepodobnosť nastania daného javu. Väčšinou sa tieto javy týkajú nejakej športovej udalosti, napríklad nejaké futbalové zápasy alebo automobilové preteky, ale stávkové kancelárie vypisujú kurzy aj na nešportové udalosti, kde medzi tie známejšie patria prezidentské voľby (Kyseľ, 2018) alebo ohlásenie mena novorodeného princa v kráľovskej rodine, kde zvyknú byť vypísané kurzy napríklad na pohlavie, meno novorodenca alebo presný dátum narodenia (Mansaray, 2019).

Na javy, na ktoré sú vypísané kurzy môže potom zákazník staviť istú sumu peňazí , vklad, obvykle tak, že vloží tento vklad do stávkovej kancelárie. Ak daný jav nastane, zákazník dostane od tejto stávkovej kancelárie výhru, ktorá predstavuje výsledok vynásobenia daného kurzu vkladom. Ak daný jav nenastane, vklad prepadá v prospech stávkovej kancelárie. Pre príklad si vezmime tipovanie výsledku futbalového zápasu Slovensko - Česká republika, ktorý sa odohral dňa 13.10.2018. Podľa internetového portálu OddsPortal.com bol priemerný vypísaný kurz na tip domáci (v tomto prípade Slovensko) 2,06, na tip hostia (Česká republika) 3,86 a na tip remíza 3,35. Zápas skončil výhrou hostí, čo znamená, že ak by sme boli stavili 100 korún na tento výsledok, tak by sme si boli odniesli zo stávkovej kancelárie 386 korún (3,86 * 100 = 386), čo predstavuje zárobok 286 korún, pretože 100 korún predstavuje vklad. Ak by sme boli stavili 100 korún na výhru domácich alebo na remízu, tak by sme boli prehrali celý vklad.

Stávkovanie je hazardná hra, obľúbená práve preto, že každý hráč môže vyhrať a vie aj ovplyvniť svoju pravdepodobnosť úspechu tým, že danú udalosť pozná (Netík, 2005).

2. Neurónové siete

Neural networks are composed of nodes or units (see Figure 18.19) connected by directed links. A link from unit i to unit j serves to propagate the activation a_i from i to j. Each link also has a numeric weight $w_{i,j}$ associated with it, which determines the strength and sign of the connection. Just as in linear regression models, each unit has a dummy input $a_0 = 1$ with an associated weight $w_{0,j}$. Each unit j first computes a weighted sum of its inputs:

$$in_j = \sum_{i=0}^{n} w_{i,j} a_i$$

Then it applies anactivation function g to this sum to derive the output:

$$a_j = g(in_j) = g\left(\sum_{i=0}^{n} w_{i,j} a_i\right)$$

The activation function g is typically either a hard threshold (Figure 18.17(a)), in which case the unit is called a perceptron, or a logistic function (Figure 18.17(b)), in which case the term sigmoid perceptron is sometimes used. Both of these nonlinear activation function ensure the important property that the entire network of units can represent a nonlinear function (seeExercise 18.22). As mentioned in the discussion of logistic regression (page 725), the logistic activation function has the added advantage of being differentiable. Having decided on the mathematical model for individual "neurons," the next task is to connect them together to form a network. There are two fundamentally distinct ways to do this.

A feed-forward network has connections only in one direction—that is, it forms a directed acyclic graph. Every node receives input from "upstream" nodes and delivers output to "downstream" nodes; there are no loops. A feed-forward network represents a function of its current input; thus, it has no internal state other than the weights themselves.

A recurrent network, on the other hand, feeds its outputs back into its own inputs. This means that the activation levels of the network form a dynamical system that may reach a stable state or exhibit oscillations or even chaotic behavior. Moreover, the response of the network to a given input depends on its initial state, which may depend on previous inputs. Hence, recurrent networks (unlike feed-forward networks) can support short-term memory. This makes them more interesting as models of the brain, but also more difficult to understand.

Feed-forward networks are usually arranged in layers, such that each unit receives input only from units in the immediately preceding layer. In the next two subsections, we will look at single-layer networks, in which every unit connects directly from the network's inputs to its outputs, and multilayer networks, which have one or more layers of hidden units that are not connected to the outputs of the network.

2.1 Popredné neurónové siete

2.2 Rekurentné neurónové siete

Vo všeobecnosti možno za rekurentnú sieť považovať akúkoľvek neurónovú sieť, v ktorej istá podmnožina neurónov (rekurentné neuróny) jes chopná uchovať informáciu o svojich aktiváciách v predošlých časoch pre výpočet aktivácií neurónov v čase t+1. "Odpamätané" hodnoty sa objavia v čase t+1 ako aktivácie tzv. kontextových neurónov (Kvasnička a kol., 2002)

3. Datasety

Data, s ktorými budeme pracovať, sú výhradne len výsledky a konečné stavy jednotlivých zápasov.

3.1 Futbal

Pre futbal data predstavujú pre každú ligu dataset všetkých zápasov odohraných len vrámci ligy za pár posledných sezón. Nebudeme používať žiadne data informujúce o hráčoch, ktorí sú v oficiálnej súpiske na zápas ani data priamo len o základnej zostave na daný zápas. Taktiež vzhľadom na to, že tímy v jednotlivých ligách hrajú zápasy aj mimo ligy, prinajmenšom zápasy v ligovom pohári, tak nebudú použité ani informácie o oddychu pred daným zápasom, teda koľko dní pred zápasom mali zúčastnené tímy voľno.

Dataset pre každú ligu je tabuľka, kde riadky predstavujú jednotlivé zápasy zoradené podľa dátumu, v ktorom bol zápas odohraný, zostupne. Stĺpce sú v poradí:

1. Jednoznačný názov domáceho tímu (nemusí byť celý názov, stačí skrátený, ale jednoznačný a, pokiaľ možno, v celom datasete konzistentný),
2. Jednoznačný názov hosťujúceho tímu,
3. Id zápasu,
4. Ligové kolo, v ktorom sa zápas odohral (0, ak sa nevie),
5. Id domáceho tímu,
6. Id hosťujúceho tímu,
7. Počet gólov strelených domácim tímom v zápase,
8. Počet gólov strelených hosťujúcim tímom v zápase,
9. Dátum zápasu,
10. Sezóna,
11. Kurz na výhru domácich,
12. Kurz na remízu,
13. Kurz na výhru hostí.

Tento dataset potom predáme programu *DataMaker.exe* (TODO!) písanom v jazyku C#, ktorý pretransformuje tieto data na vstupné neuróny pre neurónovú sieť. Všetkých vstupných neurónov je 44, v poradí:

1. htW,
2. htD,
3. htL,
4. htGFpG,

5. htGApG,
6. atW,
7. atD,
8. atL,
9. atGFpG,
10. atGApG,
11. htHW,
12. htHD,
13. htHL,
14. htHGFpG,
15. htHGApG,
16. atAW,
17. atAD,
18. atAL,
19. atAGFpG,
20. atAGApG,
21. hFW,
22. hFD,
23. hFL,
24. hFGF,
25. hFGA,
26. aFW,
27. aFD,
28. aFL,
29. aFGF,
30. aFGA,
31. MW,
32. MD,
33. ML,
34. MGF,
35. MGA,
36. MhW,
37. MhD,
38. MhL,
39. MhGF,
40. MhGA,
41. htLTS,
42. atLTS,
43. dFS,
44. dFCS.

Vysvetlivky: Prefixy: h[t] – domáci tím, a[t] – hosťujúci tím, F – forma (posledných 5 zápasov), M – posledných 5 vzájomných zápasov oboch daných tímov, Mh – posledných 5 vzájomných zápasov oboch tímov hrané na ihrisku domáceho tímu. Sufixy: W – počet výher, D – počet remíz, L – počet prehier, GF[pG] – počet strelených gólov prepočítaných na zápas, GA[pG] – počet inkasovaných gólov prepočítaných na zápas, LTS – dlhodobá sila tímu (priemerný počet bodov tímu v posledných sezónach v lige). Ďalšie: dFS – rozdiel v skóre formy oboch tímov, dFCS – rozdiel v momentálnom skóre formy oboch tímov. Skóre formy oboch tímov je vypočítané ako počet bodov súpera v posledných 5 zápasov pre tím vynásobený počtom bodov získaných z daného zápasu. Momentálne skóre formy funguje podobne s výnimkou toho, že to je prepočítavané pred momentálnym zápasom, zatiaľ čo skóre sa počíta v momente ukončenia zápasu.

Skóre je pokus čo najlepšie ohodnotiť formu tímu jedným údajom. V sekcii (Vylepšovanie siete) (TODO!) sa budeme snažiť znížiť počet vstupných neurónov

a ponechať len tie, ktoré sú dôležité. Čím väčší počet vstupných neurónov, tým väčšia je šanca, že sieť sa pokúsi medzi datami nájsť nejakú súvislosť, ktorá tam nie je, čo môže pri testovacích datach vyústiť v nesprávne výsledky (pretrénovanie dat) (TODO! cit).

Posledné 3 stĺpce tohto súboru predstavujú kurzy na dané výsledky. Tieto ale nie sú pri trénovaní siete využívané.

Program tiež vytvorí ďalší súbor, ktorý obsahuje testovacie data, teda data, ktoré sa nevyužívajú pri trénovaní siete, ale len pri vyhodnocovaní výsledkov. Tieto data sú v rovnakom poradí a musia obsahovať kurzy na dané výsledky a aj výsledok zápasu vo forme troch stĺpcov v poradí domáci, remíza, hostia, kde výsledok, ktorý nastal je ohodnotený 1, zvyšné sú 0. Je to potrebné pre vyhodnocovanie, pretože neurónová sieť bude mať 3 výstupné neuróny v rovnakom poradí a predikciu ohodnotí na 1.

Data v jednom súbore predstavujú pár posledných sezón a prvú polovicu sezóny 2018/2019, ktorá predstavuje všetky odohrané zápasy od začiatku sezóny až po odohratie posledného zápasu pred začiatkom kola, ktoré je numericky už v druhej polovici sezóny. Napríklad najvyššia anglická futbalová liga Premier League má 38 kôl každú sezónu, do úvahy sa bude brať posledných pár sezón pred sezónou 2018/2019 a všetky zápasy odohrané pred prvým zápasom 20. kola sezóny 2018/2019 (s výnimkou predohrávok, teda zápasov, ktoré boli preložené na dátum pred dátumom, v ktorom daný zápas figuroval v predsezónnom rozpise zápasov). Túto hranicu pre každú predikovanú ligu uvediem ručne do zdrojového kódu programu DataMaker.exe, pretože neviem o nejakom reálnom funkčnom algoritme, ktorý by to vedel s absolútnou istotou určiť a predstavuje to len jednu sezónu pre 5 líg.

Tento súbor je potom predaný programu v0.py (TODO!), ktorý data pripraví, vytvorí neurónovú sieť s danými parametrami (bližšie o presných parametroch v kapitole Príprava siete) a naučí ju dané data, ktoré nakoniec vyhodnotí podľa rôznych kritérií ako dôvera v daný tip alebo kurzovo vyrovnané zápasy, teda zápasy, kde na výhru domácich a vyhru hostí je dostatočne podobný kurz.

3.2 Tenis

Pre tenis budeme používať data pre najlepších 100 hráčov na začiačtku každého roka v rebríčku ATP. Data v tomto prípade predstavujú zápasy z turnajov typu ATP 250, ATP 500, ATP Masters 1000, Grandslam, Finals, Nextgen Finals a ... .

Dataset je tabuľka, každý zápas predstavuje jeden riadok tabuľky, zápasy sú zoradené do turnajov od najskôr odohraných turnajov po tie najbližšie súčasnosti (ak sa obe turnaje začali a končili hrať v rovnaký deň, tak sú v ľubovoľnom poradí, nie je možné, aby poradie zmenilo nejaké data, pretože nie je možné hrať na dvoch turnajoch takéhoto typu zároveň). Zápasy v turnajoch sú zoradené od finále po prvé kolo, teda intuitívne opačne. V tomto prípade na poradí nezáleží, dôležité je, že je v tom systém. Program na spracovanie dat (ATPDataMaker.exe) si tie poradie dat upraví tak, aby mu vyhovovali. Stĺpce tabuľky sú v poradí:

1. Názov turnaja,

2. Počet bodov, ktoré víťaz obrdží za výhru v turnaji (ak to je neznáme, tak je tam nápis N/A)
3. Rok, v ktorom sa turnaj odohral,
4. Povrch kurtov na turnaji (tvrdý, antukový alebo trávnatý povrch),
5. Meno víťaza zápasu,
6. Meno hráča, ktorý zápas prehral,
7. Kolo turnaja, v ktorom sa zápas odohral od najdôležitejšieho (1 značí finále, 2 semifinále, apod.),
8. ID zápasu,
9. ID víťaza,
10. ID porazeného hráča,
11. Počet setov, ktoré v zápase získal víťaz,
12. Počet setov, ktoré v zápase získal porazený hráč,
13. Počet hier, ktoré v zápase získal víťaz v jednotlivých setoch oddelené znakom |,
14. Počet hier, ktoré v zápase získal porazený hráč v jednotlivých setoch oddelené znakom |.

ID hráčov sa nachádzajú v ďalšom súbore (*atpranking.csv*), ktorý sa predáva aplikácii na tvorbu vstupných neurónov do neurónovej siete. Tento súbor obsahuje ID jednotlivých hráčov, ich mená a ich poradie v koncoročných rebríčkoch hodnotenia ATP za roky 1999–2018. Poradie berieme len ak sa hráč umiestnil na miestach 1–100.

Zápasy obsiahnuté v súbore *atpresults.csv* sú len zápasy, v ktorých aspoň jeden hráč bol na konci aspoň raz v daných rokoch na miestach 1–100 v hodnotení ATP. Predikovať sa budú len zápasy medzi takýmito hráčmi, ale kvôli rôznym výpočtom je potrebné mať všetky data o takýchto hráčoch z turnajov, ktoré sú obsiahnuté v súbore.

Tieto datasety sa potom predajú súboru *ATPDataMaker.exe*, ktorý ich pretransformuje na data pre vstupné neuróny neurónových sietí. Všetkých vstupných neurónov je 37, súbor ku každému vstupnému neurónu vydá aj očakávaný výstup (1?, 2?) a pre predikovanú časť dodá aj kurzy stávkových kancelárií na daný výsledok (1B, 2B). Výstupné súbory majú teda stĺpce v poradí:

1. 1W
2. 1L
3. 1GDpS
4. 2W
5. 2L
6. 2GDpS
7. 1FW
8. 1FL

9. 1FGDpS
10. 2FW
11. 2FL
12. 2FGDpS
13. 1SW
14. 1SL
15. 1SGDpS
16. 2SW
17. 2SL
18. 2SGDpS
19. 1SFW
20. 1SFL
21. 1SFGDpS
22. 2SFW
23. 2SFL
24. 2SFGDpS
25. 1MW
26. 1ML
27. 1MGDpS
28. 1MSW
29. 1MSL
30. 1MSGDpS
31. 1R
32. 2R
33. H
34. C
35. G
36. dSc
37. dSSc
38. 1?
39. 2?

Vysvetlivky: Prefixy: 1,2 – hráči, M – vzájomné zápasy (z pohľadu hráča 1), S – povrch (zápasy hráča na tomto povrchu), F – forma (posledných 10 zápasov). Suffixy: W – počet výhier*, L – počet prehier*, GDpS – priemerný rozdiel v počte získaných hier v sete v prospech daného hráča, R – poradie v rebríčku, ? – výsledok (ak hráč vyhral - 1, inak 0) Zvyšné: H – tvrdý povrch, C – antuka, G – trávnatý povrch, dSc - rozdiel v skóre** medzi hráčmi (z pohľadu hráča 1), dSSc - rozdiel v povrchovom skóre** (z pohľadu hráča 1)
* - v danej sezóne, s výnimkou, ak predchádza prefix SF - povrch si uchováva formu hráča aj z minulej sezóny, ak bola braná do úvahy ** - skóre je pokus ohodnotiť silu víťazstva, berie do úvahy formu, teda posledných 10 zápasov a počíta sa ako $(150 - rank) \cdot point$, kde rank je poradie súpera v poslednom koncoročnom rebríčku ATP a point je 1, ak hráč vyhral, 0, ak vyhral súper. Ak súper nebol v Top 100 rebríčka ATP na konci predchádzajúceho roka, tak za jeho rank je dosadené číslo 130. To je len preto, lebo teoreticky má dané víťazstvo hodnotu, musí byť teda nejak ohodnotené lepšie ako ľubovoľná prehra, ktorá je ohodnotená hodnotou 0.

Skóre je pokus výraznejšie ohodnotiť formu hráča ako len počtom výhier a prehier. Pri pokusoch a vylaďovaní siete budeme v sekcii (Vylepšovanie siete) (TODO!) selektovať dané vstupné neuróny podľa rôznych kritérií a vyskúšame tiež aj ako sa bude sieť správať, ak nahradíme všetky stĺpce obsahujúce data o forme rozdielom v skóre. Teoreticky tým ušetríme 6 vstupných neurónov, príliš veľa vstupných neurónov môže viesť k rýchlejšiemu pretrénovaniu siete (TODO!

cit), čomu sa budeme snažiť zabrániť selektovaním len tých dôležitých neurónov.

4. Stavba siete

5. Dokumentácia

Z programátorského hľadiska je práca rozdelená na tri časti. Prvú časť predstavuje získavanie výsledkov a kurzov jednotlivých zápasov. Druhú časť programu predstavuje transformácia dat na údaje priamo vložiteľné do vstupných neurónov daných neurónových sietí. Poslednú časť tvorí stavba daného typu neurónovej siete pre daný šport.

Transformačná časť rozdelí data na 3 časti, trénovacie data, trénovacie data (pre optimalizovanie siete používané ako testovacie) a testovacie data. Je teda zaručené, že žiadna sieť neuvidí testovacie data vopred pred finálnym vyhodnocovaním.

Údaje, ktoré sa objavia vo výstupe sú celková úspešnosť a celkový zisk, úspešnosť a zisk siete pri vyrovnaných zápasoch (a vyrovnaný zápas považujem zápas, kde kurzy na výhru jedného alebo druhého tímu sa líšia najviac o 1) a úspešnosť a zisk siete pri výhradnom tipovaní zápasov, na ktoré máme istú dôveru (od hodnoty, ktorá je počítaná ako rozdiel dvoch najvyšších čísel, ktoré sieť vydá na výstup, jednoducho povedané, rozdiel najpravdepodobnejšej a druhej najpravdepodobnejšej možnosti výsledku zápasu z hľadiska siete). Táto hodnota dôvery bola tiež vyoptimalizovaná pre každú sieť/program osobitne.

5.1 Futbal

Podmienkou pre futbal je získať všetky zápasy sezóny pre danú ligu. Musia byť všetky, pretože v ďalšej časti sa počíta na základe už odohraných zápasov a jeden zápas by mohol skresliť výsledky. Jednotlivé ligy boli teda vyberané nielen na základe kvality, ale aj na základe toho, že v pár posledných sezónach sa ani raz nestalo, že zápas musel byť z nejakého hľadiska udelený kontumačne (awarded) jednému z tímov (ako sa napríklad stalo vo francúzskej lige - nejaký zdroj) alebo celá sezóna bola poznačená korupčným škandálom ako v prípade talianskej ligy v sezóne xxxx (zdroj). Takéto výsledky by nemuseli skresliť stavbu neurónovej siete, ale všeobecne je lepšie, ak sa takýmto situáciám vyhneme.

Údaje o týchto zápasoch sa dajú stiahnuť jednoducho, spustením programu oddscaper.py a zadaním skratky danej ligy pre futbal. (TODO!) Skratky sú:

1. ENG - najvyššia anglická liga (Premier League)
2. GER - najvyššia nemecká liga (Bundesliga)
3. SPA - najvyššia španielska liga (La Liga)
4. BUL -
5. QAT -

Program stiahne všetky výsledky a kurzy pre všetky zápasy všetkých kompletných sezón tej-ktorej ligy zo stránky www.oddsportal.com.

Výstup tohto programu predáme programu *DataMaker.exe* (teda ako prvý parameter programu DataMaker.exe je potrebné predať cestu k súboru, ktorý je výstupom súboru oddscraper.py, tento súbor sa volá rovnako ako skratka danej

ligy s príponou *.csv*), ktorý je písaný v jazyku C# a pretransformuje tieto data na vstupné neuróny pre neurónovú sieť. Všetkých vstupných neurónov je 44. Presné poradie aj popis sa dá nájsť v sekcii Prílohy (Príloha A1). Tieto údaje boli vybrané špecificky aj s pomocou súvisiacich prác ako údaje, ktoré popisujú stav oboch tímov, ktoré hrajú proti sebe zápas. Bonus predstavujú vstupy označené ako skóre, tieto boli vytvorené mnou ako pokus o jednoduchý a presnejší popis formy pomocou jedného údaju namiesto 10. Ak bude mať teda jeden z týchto neurónov (alebo obe spoločne) úspech, tak bude možné skrátiť počet vstupných neurónov o 10. Pre upresnenie, výstupom súboru je opäť tabuľka formátu csv, názov je zložený zo skratky pre názov danej ligy a slova *input*.

Cestu na dané súbory potom ako prvé tri parametre (v poradí, v akom sú uvedené v úvode sekcie) predáme programu *ffnnfootball.py* alebo *rnnfootball.py* podľa toho, či chceme, aby dané údaje vyhodnocovala popredná rekurentná neurónová sieť. Výsledky vypíše na štandardný výstup a uloží ich aj do logu, ktorý pozostáva z typu siete, názvu ligy a časovej známky vo formáte *txt*.

Hodnoty dôvery pre ffnnfootball.py a rnnfootball.py sú

TODO!!!

resp.

TODO!!!

5.2 Tenis

Údaje o tenisových zápasoch sú predpripravené v súbore atpresults.csv.

Podmienkou pre tenis je získať všetky zápasy každého turnaja ATP typu 500, 1000 a Grand Slam, kde hraje aspoň jeden hráč z Top 100 rebríčka ATP pre danú sezónu. Dôvodom je fakt, že predikujeme zápasy týchto turnajov medzi hráčmi z Top 100 rebríčka ATP, ale pre týchto hráčov počítame ich momentálnu formu, takže sú pre nás dôležité aj zápasy, ktoré odohrajú proti hráčom mimo Top 100. Vzhľadom na relatívnu kvalitu turnajov ATP 250 a fakt, že množstvo hráčov z Top 100 sa pravidelne zúčastňuje aj týchto turnajov. Čo teda logicky znamená, že niekedy nastúpia dvaja takíto hráči aj proti sebe. Takže zoberieme do úvahy aj tieto turnaje (vzájomné zápasy medzi jednotlivými hráčmi na takto ohodnotených turnajoch tiež patria medzi údaje, z ktorých sa stávajú vstupné neuróny (odkaz na prílohu s tenisom (A2?))).

V tenise sa nemôžeme vyhnúť zápasom, ktoré boli nejakým spôsobom udelené jednému z hráčov, či už bez boja alebo po skreči súpera v priebehu zápasu, pretože zranenia sú súčasťou profesionálneho športu. Vo futbale sa to obvykle rieši prestriedaním zraneného hráča, v tenise to, prirodzene, nie je možné. Pre potreby tejto práce máme dve možnosti, buď môžeme tieto zápasy úplne ignorovať alebo ich môžeme započítavať do niektorých oblastí vstupu (ako napríklad forma alebo vzájomné zápasy) a ignorovať inde (predpovedať takéto výsledky je možno nápad pre inú prácu). Pre potreby tejto práce budeme tieto zápasy úplne ignorovať, čo znamená, že sa nevyskytnú v trénovacích ani testovacích datach. Samozrejme, má

to svoje výhody aj nevýhody. Výhodou je, že výsledky budú reálne odzrkadľovať presnosť siete na zápasoch, ktoré sa odohrali a skončili. Predpovedať zranenie nie je cieľom tejto práce. Ďalšou výhodou je spravodlivosť oblastí vstupu ako forma a vzájomné zápasy, pretože sa tam berú len zápasy, ktoré sa dohrali dokonca, takže tieto čísla sa v žiadnom okamihu nenafukujú. Napríklad ak hráč natrafí počas turnaja na dvoch/troch súperov, ktorí sa vzdajú, tak by sa mu vo forme ukázali tieto víťazstvá, aj keď to neboli plnohodnotné výhry. Nevýhodou je, že výsledky nemusia ukazovať reálne výsledky v praxi (pred zápasom nevieme určiť, či sa hráč zraní, ale sieť aj tak vydá svoju predpoveď, aj keď nebola na tieto údaje trénovaná).

Pre tabuľku *atpresults.csv* je postup podobný. Túto tabuľku je potrebné predať programu *ATPDataMaker.exe* (opäť ako prvý parameter je potrebené predať cestu k tejto tabuľke). Program je opäť písaný v jazyku C# a opäť pretransformuje data na vstupné neuróny pre neurónovú sieť. Všetkých vstupných údajov (počet stĺpcov tabuľky) je 37. Ich presné poradie a popis sa dá nájsť v sekcii Prílohy (Príloha A2).

Cestu na dané súbory potom ako prvé tri parametre (v poradí, v akom sú uvedené v úvode sekcie) predáme programu *ffnnatp.py* alebo *rnnatp.py* podľa toho, či chceme, aby dané údaje vyhodnocovala popredná rekurentná neurónová sieť. Výsledky vypíše na štandardný výstup a uloží ich aj do logu, ktorý je vo formáte *txt* a ktorého názov pozostáva z typu siete, slova *atp* a časovej známky.

Hodnoty dôvery pre ffnnatp.py a rnnatp.py sú

TODO!!

resp.

TODO!!

Záver

Zoznam použitej literatúry

Bailey, M. J. a kol. (2005). *Predicting sporting outcomes: A statistical approach*. PhD thesis, Faculty of Life and Social Sciences, Swinburne University of Technology.

Gers, F. A., Schraudolph, N. N. a Schmidhuber, J. (2002). Learning precise timing with lstm recurrent networks. *Journal of machine learning research*, **3**(Aug), 115–143.

Hucaljuk, J. a Rakipović, A. (2011). Predicting football scores using machine learning techniques. In *2011 Proceedings of the 34th International Convention MIPRO*, pages 1623–1627. IEEE.

Igiri, C. P. a Nwachukwu, E. O. (2014). An improved prediction system for football a match result. *IOSR Journal of Engineering (IOSRJEN)*, **4**(12), 12–20.

Joseph, A., Fenton, N. E. a Neil, M. (2006). Predicting football results using bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, **19**(7), 544–553.

Koromházová, V. (2008). *Jak dokonale zvládnout tenis*. Grada Publishing as. ISBN 978-80-247-2316-7.

Kvasnička, V., Beňušková, Ľ., Pospíchal, J., Farkaš, I., Tiňo, P. a Kráľ, A. (2002). Úvod do teórie neurónových sietí. URL `http://ics.upjs.sk/~novotnyr/home/skola/neuronove_siete/nn_kvasnicka/Uvod%20do%20NS.pdf`. [cit. 2019-05-20].

Kyseľ, T. (2018). Prezidentské voľby 2019: Stávkové kancelárie veria viac harabinovi ako bugárovi. URL `https://www.aktuality.sk/clanok/599690/prezidentske-volby-2019-stavkove-kancelarie-veria-viac-harabinovi-ako-bugarc` [cit. 2019-05-09].

Mansaray, J. (2019). Any day now - odds on for ímminentťroyal baby birth. URL `https://www.reuters.com/article/us-britain-royals-baby-betting-idUSKCN1S7418`. [cit. 2019-05-09].

Netík, M. (2005). Jak sázet s pomocí internetu (1.). URL `https://www.lupa.cz/clanky/jak-sazet-s-pomoci-internetu-1/`. [cit. 2019-04-28].

Olah, C. (2015). Understanding lstm networks. URL `https://colah.github.io/posts/2015-08-Understanding-LSTMs/`. [cit. 2019-05-20].

Prasetio, D. a kol. (2016). Predicting football match results with logistic regression. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, pages 1–5. IEEE.

Shin, J. a Gasparyan, R. (2014). A novel way to soccer match prediction. *Stanford University: Department of Computer Science*.

TÁBORSKÝ, F. (2004). *Sportovní hry*. Grada Publishing as. ISBN 80-247-0875-2.

TÁBORSKÝ, F. (2005). *Sportovní hry 2: základní pravidla, organizace, historie*. Grada Publishing as. ISBN 80-247-1330-6.

Zoznam obrázkov

Zoznam tabuliek

Seznam použitých zkratek

A. Přílohy

A.1 První příloha