

OZNAL - Dátová analýza dát hráčov z hry FIFA 19

Timotej Zaťko and Tomáš Hoffer

Fakulta informatiky a informačných technológií STU v Bratislave Ilkovičova 2, 842 16
Bratislava 4

xzatkot1@stuba.sk xhoffer@stuba.sk

<https://www.fiit.stuba.sk/>

Abstract. Táto práca obsahuje analýzu dát hráčov z hry FIFA 19, opis dát a ich charakteristiky. V tejto práci skúmame vzťahy atribútov na predikovaný atribút trhovej ceny hráča a jeho hernú pozíciu na ihrisku.

Keywords: Analýza dát, FIFA 19, Strojové Učenie, Regresia, Klasifikácia

1 Opis problému a motivácia

Rozhodli sme sa pre analýzu dát hráčov z hry FIFA 19. V hre sa nachádza veľké množstvo hráčov z rôznych krajín, hrajúcich v rôznych súťažiach. Ich schopnosti v hre by mali odzrkadľovať ich schopnosti z reálneho sveta. Tvorcovia hry sa o to snažia vytvorením herných atribútov, akými sú napríklad rýchlosť šprintu, sila strely, zakončovanie alebo hlavičkovanie ktoré dokázali vyjadriť číselne. V projekte sa budeme snažiť na základe týchto atribútov klasifikovať hráčov do hernej pozície a predikovať ich trhovú hodnotu v hre. Keďže hra hráčom neukazuje vždy trhovú hodnotu hráča, ale len jeho atribúty, náš model môže byť užitočný pri určovaní výšky ponúkanej sumy za prestup hráča pri vyjednávaní v hre. Keďže v reálnom svete schopnosti hráčov (napr. zakončovanie) nie sú nijako numericky vyjadrené skúsime využiť to, že v hre vyjadrené sú, na zistenie, ktoré atribúty sú dôležité pre určité herné pozície.

2 Opis dát s charakteristika dát

Naša dátová sada obsahuje 18207 záznamov - hráčov, ktorý každý z nich má 87 atribútov. Z toho je 42 numerických a 45 kategorických.

2.1 Očistenie dát

Kvôli analýze bolo treba dátovú sadu, očistiť a urobiť predspracovanie niektorých atribútov. Bolo nutné spraviť nasledovné úpravy – konverzia a očistenie finančných hodnôt (napr. '\$77.5M', '\$1K'), konverzia mier (napr. '159lbs', '5'11'),

konverzia časových údajov (napr. ‘Jan 25, 2019’, ‘2018’) a rozdelenie niektorých atribútov na viac atribútov - niektoré atribúty obsahovali dva numerické atribúty. Taktiež atribút určujúci pozíciu hráča obsahoval až 38 rôznych herných pozícií, preto sme sa rozhodli tento atribút rozšíriť do ďalších dvoch atribútov v ktorých sme podobné pozície spojili čím vznikli dva atribúty s 13 resp. 4 hodnotami. Po týchto úpravách sa nám počet atribútov zmenil na 129 – 89 numerických a 40 kategorických atribútov.

Toto veľké množstvo atribútov sme zaradili do nasledovných kategórií, uvádzame k nim aj niektoré atribúty. Atribúty sme rozdelili aj na základe toho či jeho hodnotu musela hra nejakým spôsobom odvodiť z reálneho sveta.

- **človek** - meno, národnosť, výška, hmotnosť, vek...
- **futbalový hráč (reálny svet)** - názov klubu, číslo dresu, dátum príchodu do klubu, dĺžka kontraktu, plat hráča, hosťovský klub, herná pozícia, preferovaná noha
- **futbalový hráč (hra)** - trhovú hodnotu hráča, triky, pracovitosť v útoku/obrane, vhodnosť na určitú špecifickú pozíciu a potenciálny rast (78 atribútov), medzinárodná reputácia
- **futbalové schopnosti hráča určené hrou** (celkovo 34 atribútov) - krátke prihrávky, hlavičkovanie, zakončovanie...
- **iné atribúty hry** - logo klubu, vlajka (na základe národnosti), typ postavy (tj. typ herného modelu), typ tváre (tj. kvôli herného modelu)

2.2 Analýza chýbajúcich hodnôt

Ako prvé sme sa pozreli na chýbajúce hodnoty v našej dátovej sade. Početnosti sme vizualizovali stĺpcovým diagramom a vzťahy tepelnou mapou a dendrogramom. Zistili sme, že niektorým hráčom chýbajú hodnoty v atribútoch ako sú klub, plat, dĺžka kontraktu, číslo dresu či výkupná klauzula. Tieto hodnoty chýbajú väčšinou spoločne, o čom sme sa presvedčili v dendrograme a tepelnej mape. Je to aj logické, keďže hráč, ktorý nemá klub nemôže poberať plat alebo mať výkupnú klauzulu v kontrakte. Z našich zistení môžeme konštatovať, že ak hráčovi chýbajú hodnoty v atribúte, väčšinou sa nejedná o chybu v úplnosti dátovej sady.

2.3 Analýza z pohľadu pozície hráča

Analýzu sme ďalej realizovali z pohľadu dvoch atribútov, ktoré sa budeme snažiť predikovať - pozíciu hráča (kategorický atribút) a jeho trhovú hodnotu (numerický atribút). Dátová sada celkovo obsahuje 36 herných pozícií.

Na základe našich futbalových znalostí sme sa rozhodli zoskupiť podobné pozície do skupín, čím sme znížili počet rôznych hodnôt v atribúte ‘Position’ (Fig. 1). Môžeme pozorovať, že triedy nie sú vyvážené.

Pozíciu sa môžeme pokúsiť predikovať na základe 34 atribútov definujúcich futbalové schopnosti hráča určené hrou. Vizualizovali sme všetky dvojice týchto atribútov a pomocou ”scatter plot” -ov sme hľadali zhluky jednotlivých pozícií.



Fig. 1: Početnosti nových atribútov 'Position (4)' a 'Position (13)' po zoskupení podobných hodnôt z atribútu 'Position'.

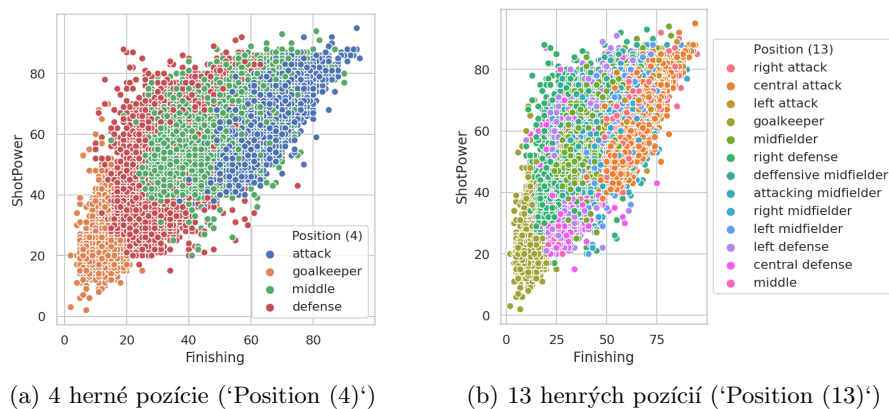
Ako príklad uvádzame vzťah medzi atribútmi 'Shot Power' (sila strely) a 'Finishing' (zakončovanie), kde môžeme pozorovať jednotlivé zhľuky podľa pozície hráča (Fig. 2).

Z prieskumnej analýzy tiež vyplýva, že pre konkrétne pozície hráčov sú typické určité čísla dresov. Pre brankárov (angl. goalkeeper) je typickým číslom dresu číslo 1 pričom toto číslo nemá priradený žiaden hráč na inej pozícii. Pre útočníkov (angl. attack) to je 9, pre obrancov (angl. defense) sú to čísla 2 - 6 a pre stredopoloárov 7, 8 a 10. Tento atribút môže byť veľmi dobrý na klasifikáciu pozície hráča, avšak my sa v prvom rade zamierame na predikciu z herných atribútov. Atribúty, ako napríklad číslo dresu nám môžu úlohu príliš zjednodušiť.

Ďalšou zaujímavosťou je, že v našej dátovej sade sa nachádzajú prevažne hráči ktorí preferujú pravú nohu, avšak na pozícii ľavého obrancu výrazne prevládajú hráči s preferovanou ľavou nohou. (Fig. 3).

2.4 Analýza z pohľadu trhovej hodnoty hráča

Pomocou Pearsonovho korelačného koeficientu sme hľadali korelácie medzi atribútom 'Value' a ostatnými numerickými atribútmi. Takmer lineárnu koreláciu (0.99) vykazuje atribút 'Release Cause' (Fig. 4). Hráči s vysokou trhovou hodnotou majú v dátovej sade podpísanú zmluvu s vyššou výkupnou klauzulou. Vysokú koreláciu taktiež vykazujú atribúty 'Overall' (0.631), 'Wage' (0.850) a 'International Reputation' (0.656).



(a) 4 herné pozície ('Position (4)')

(b) 13 herných pozícií ('Position (13)')

Fig. 2: Vzťah medzi atribútmi 'Shot Power' a 'Finishing'. Pre 4 herné pozície sú zhluky zreteľnejšie ako pre 13 herných pozícií.

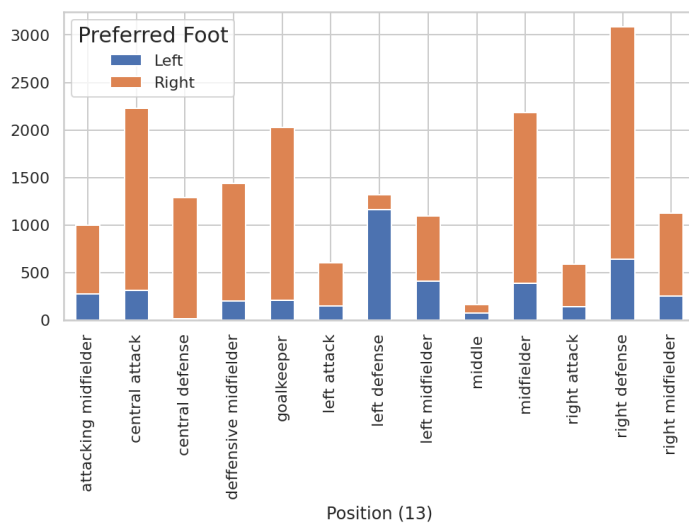


Fig. 3: Preferovaná noha hráčov podľa hernej pozície.

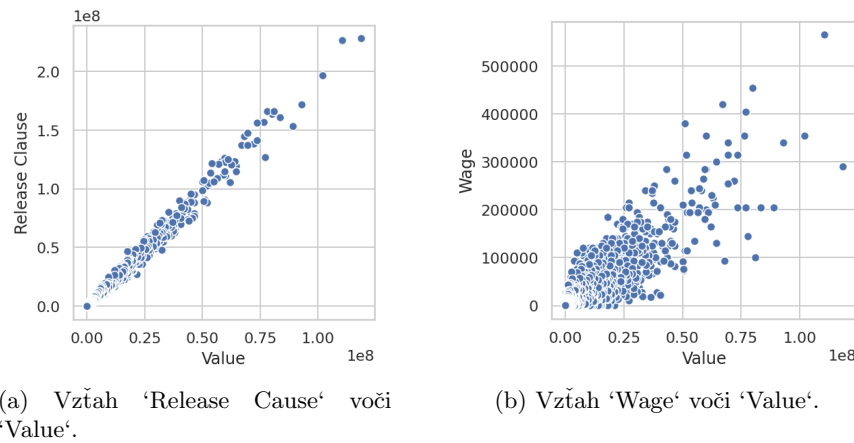


Fig. 4: Niektoré atribúty vykazujú vysokú mieru korelácie s atribútom 'Value'.

3 Definovanie úlohy objavovania znalostí

Rozhodli sme sa, že budeme vykonávať nasledujúce úlohy:

- predikcia hernej pozície hráča – všeobecnej (4 triedy), rozšírenej (13 tried)
 - *klasifikácia*
- predikcia hodnoty hráča (atribút s názvom 'Value') – *regresia*

Obe tieto úlohy budeme realizovať z atribútov určujúcich herné schopnosti hráča a následne zo všetkých atribútov. Výsledky porovnáme a očakávame, že model natrénovaný zo všetkých atribútov bude výrazne lepší, keďže niektoré atribúty výrazne ovplyvňujú predikovanú premennú a to – Release Clause - Value; Jersey Number - Position.

4 Predpokladaný scenár riešenia (problémy)

Predpokladáme, že bude potrebné vykonať nasledovné úlohy:

- predspracovanie kategorických hodnôt (tj. one-hot encoding)
- normalizácia dát
- over-sampling
- tréning modelu

Tréning modelu zahŕňa výber atribútov (angl. feature selection) a výber a tréning modelu. Na úlohu predikcie hodnoty hráča budeme pravdepodobne používať lineárnu regresiu/neurónovú sieť a na určenie hernej pozície hráča rozhodovací strom / náhodný les / SVM / neurónovú sieť.

5 Predspracovanie dát

Prvotnému predspracovaniu dát na účely analýzy sme sa venovali v kapitole 2.1 kde sme vykonali nasledovné transformácie dát:

- Transformácia hodnôt peňažných atribútov (napr. hodnoty ‘1.2 M’, ‘200 K’ sme transformovali na jednotný numerický tvar)
- Dátumové atribúty sme transformovali do jednotného formátu, reprezentovaného UNIX-ovou časovou pečiatkou
- Atribúty miery (napr. výška a hmotnosť hráča) sme previedli do rovnakých jednotiek.
- Špeciálne pozície hráča boli udávané v tvare reťazca *Value+Grow*. Atribúty špeciálnych pozícií sme rozdelili na dvojice aby nám vznikli numerické atribúty.
- Atribút *Work Rate* bol udávaný reťazcom v tvare *Attack/Defense*. Tento atribút sme tiež rozdelili na 2 numerické.
- Boolean atribúty udávané reťazcami sme transformovali na numerické 0/1
- Vytvorili sme 2 nové atribúty *Position(4)* a *Position(13)*, ktoré reprezentujú pozíciu hráča po zoskúpení pozícií, opísanom v kapitole 2.3
- Vytvorili sme atribút *Contract length*, ktorý reprezentuje dĺžku aktuálne podpísanej zmluvy hráča

Kategorické atribúty sme transformovali metódou ”One-Hot Encoding”. Numerické atribúty sa nachádzali v rôznych rozsahoch a jednotkách (napríklad hmotnosť v lb, výška v cm, rôzne schopnosti hráča v rozsahu $< 0, 100 >$). Rozhodli sme sa ich normalizovať do jednotného intervalu $< 0, 1 >$.

5.1 Chýbajúce hodnoty

V kapitole 2.2 sme opísali dôvod výskytu chýbajúcich hodnôt. Chýbajúce hodnoty opísané v kapitole 2.2 považujeme za opodstatnené a nebudeme ich nahrádzať. Tj. dané hodnoty chýbajú pretože majú chýbať a nie preto, že by boli dáta neúplné.

5.2 Vychýlené hodnoty

Rozdelenie číselných hodnôt v dátovej sade sme podrobne analyzovali a vizualizovali. Vychýlené hodnoty sú iba v niektorých atribútoch.

5.3 Problém nevyváženosti tried

Pri zoskúpení všetkých pozícií do 13 tried sme odhalili nevyváženosť jednotlivých tried (Fig.1), čo by mohlo negatívne ovplyvniť presnosť predikcie. Tento problém sme sa rozhodli riešiť metódou nazývanou *oversampling* a konkrétne algoritmom SMOTE [1], ktorý vytvára syntetické inštancie minoritných tried/y pomocou lineárnej kombinácie reálnych inštancií.

6 Výber atribútov

Predikciu pri oboch typoch úloh budeme realizovať pomocou 34 atribútov definujúcich futbalové schopnosti hráča určené hrou. Využijeme tiež demografické údaje, napr. výšku, hmotnosť a vek. K tejto základnej množine atribútov sme sa dopracovali na základe doménovej znalosti a rozsiahlej prieskumnej analýzy. V častiach 2.3 a 2.4 sa venujeme výberu týchto atribútov a opisujeme dôvody, prečo sme niektoré atribúty do tejto množiny nezahrnuli. Pri riešení klasifikačnej úlohy rozšírime základnú sadu atribútov o atribúty *Preferred Foot*, *Work Rate Attack* a *Work Rate Defense*. Pri regresnej úlohe využijeme dopočítaný atribút *Contract Length* a využijeme tiež medzinárodnú reputáciu hráča (angl. *International reputation*). Množinu vybraných atribútov sa pravdepodobne ešte pokúsime zredukovať (angl. feature selection). V závislosti od modelu zvolíme konkrétny prístup. Niektoré modely podporujú tzv. *feature importances*. Pri modeloch, ktoré to nepodporujú môžeme použiť rekurzívnu elimináciu atribútov alebo prírodou inšpirované algoritmy, ktoré v tejto oblasti dosahujú veľmi dobré výsledky [6].

7 Príbuzná práca

V našej doméne (avšak na dátach zo staršej verzie hry) Soto-Valero, C. [4] zisťoval, ktoré atribúty sú dôležité pre jednotlivé herné pozície. Využil PCA na zredukovanie dimenzionality zo všetkých herných atribútov a následne zhľukovaním klasifikoval hráčov do jednotlivých herných pozícií. Následne pomocou algoritmu *Gradient tree boosting* ohodnotil dôležitosť jednotlivých atribútov. Na klasifikovanie využil takzvané učenie bez učiteľa, čím sa jeho práca líši od tej našej.

Nazim R. a spol. sa snažia predikovať pozíciu hráča na základe podobných atribútov, avšak využívajú dátovú sadu zo športovej školy *Bukit Jalil Sports School*. Na klasifikáciu využívajú bayesovské siete, rozhodovacie stromy a algoritmus najbližšieho suseda. Hráči boli klasifikovaní do 10 tried. Vzhľadom na to, že dátová sada obsahovala len 100 prvkov sa autori rozhodli použiť "leave-one-out" validáciu [3].

Yaldo L. a Shamir L. sa snažia riešiť problém predikcie týždenného platu futbalového hráča pomocou 8 rôznych algoritmov strojového učenia na podobnej dátovej sade FIFA z roku 2016 [5]. Autori na predikciu využili napríklad aditívnu regresiu či rozhodovacie stromy. Platy hráčov sú predikované na základe ich pozície, ale aj atribútov hry, ktoré charakterizujú ich schopnosti. Zaujímavým zistením je fakt, že platy hráčov nezávisia len od ich schopností a pozície hráča, ale aj celkovej reputácie a jeho popularity. Známy hráč L. Messi by napríklad pri predikcii len na základe jeho schopností nedosiahol ani polovičný plat. Rozhodli sme sa teda medzinárodnú reputáciu hráča zahrnúť medzi črty použité na predikciu. Autori používajú Piersonov korelačný koeficient (CC) a priemernú absolútnu odchýlku (MAE) na meranie úspešnosti.

8 Výber metrík pre evaluáciu úspešnosti

Klasifikácia – keďže klasifikujeme do 4 resp. 13 tried, ktoré sú nevyvážené, budeme sledovať metriky **f1 micro** a **f1 macro**. Táto metrika je harmonickým priemerom *precision* a *recall* [2]. Nás zaujíma predovšetkým varianta **micro**, keďže máme nevyvážené triedy a chceme aby metrika brala do úvahy početnosti tried (tj. náš model nemusí byť taký dobrý v klasifikácii minoritných tried). Budeme tiež sledovať aj variantu **macro**, ktorá berie do úvahy všetky triedy rovnako.

Regresia – pri regresii budeme sledovať metriku s názvom **RMSE** (Root Mean Square Error).

9 Experimenty

9.1 Predikcia pozície hráča (klasifikácia)

Úlohu sme sa v prvej etape rozhodli riešiť pomocou viacerých jednoduchých modelov. Klasifikácia hráča do 4 pozícií už pri základných nastaveniach modelov vykazovala vysokú úspešnosť 1. Model sa najviac mýlil pri klasifikácii do triedy *middle*. Je to pochopiteľné, keďže stredopoliari sú ”medzi” útočníkmi a obrancami 5.

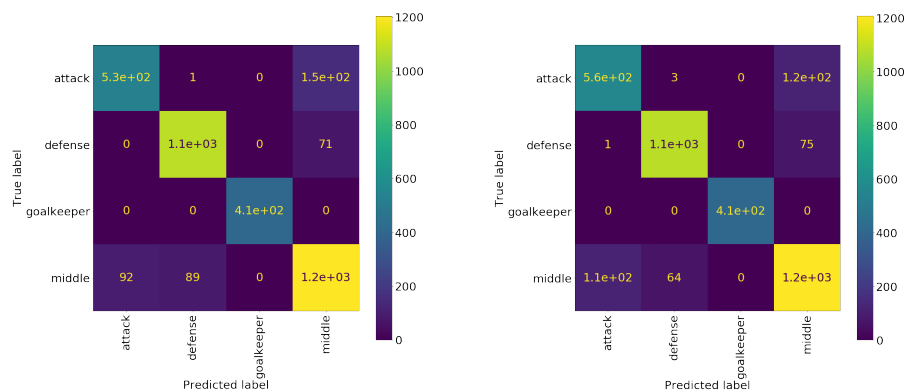
Pri klasifikácii do 13 tried pozície je úspešnosť jednoduchých modelov významne nižšia a niektoré triedy jednotlivé modely dokonca vôbec nepredikovali.

Model	Position (4) – f1 skóre						Position (13) – f1 skóre	
	Attack	Defense	GK	Middle	Micro	Macro	Micro	Macro
Rozhodovací strom	0.74	0.87	1	0.76	0.81	0.84	0.55	0.42
Logistická regresia	0.82	0.94	1	0.87	0.87	0.90	0.71	0.53
SVM	0.81	0.93	1	0.84	0.88	0.89	0.66	0.44
Náhodný les	0.81	0.93	1	0.86	0.89	0.89	0.68	0.50

Table 1: Úspešnosti jednoduchých modelov pri predikcii ‘Position 4’ a ‘Position 13’. Pre predikciu štyroch herných pozícií zobrazujeme aj metriku **f1** pre jednotlivé pozície. V oboch prípadoch bola najlepšia logistická regresia a náhodný les.

9.2 Predikcia trhovej hodnoty hráča (regresia)

Na riešenie tejto úlohy sme využili lineárnu regresiu. V tejto úlohe sa nám zatiaľ nepodarilo dosiahnuť výraznejšie výsledky (Obr. č.6).



(a) Matrica zmätenia pre model náhodný les.

(b) Matrica zmätenia pre model logistickej regresie.

Fig. 5: Matrice zmätenia pre dva z našich natrénovaných modelov, môžeme si všimnúť že model robí chyby pri klasifikácii do triedy *middle*.

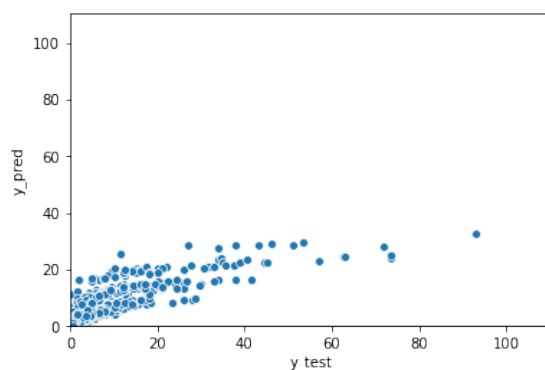


Fig. 6: Predikované hodnoty hráčov a reálne hodnoty hráčov v miliónoch. Zatiaľ sa nám nepodarilo dosiahnuť lepšie výsledky ($RMSE = 3,560331$).

9.3 Oversampling

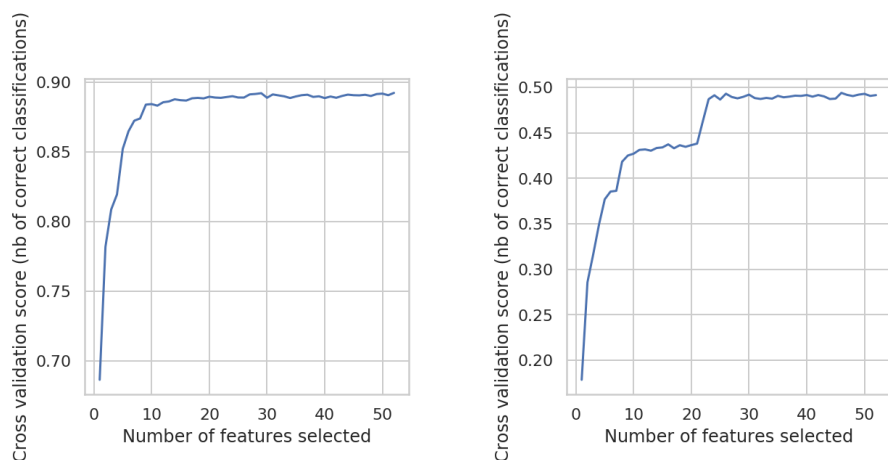
Keďže pri našej klasifikačnej úlohe sú jednotlivé triedy nevyvážené (Sekcia 5.3) pokúsili sme sa zvýšiť početnosti minoritných tried. Vyskúšali sme minoritné triedy dovozorkovávať náhodne a zároveň metódou SMOTE a jej variáciami (SVM-SMOTE, BorderlineSMOTE). Tiež sme skúšali rôzne kombinácie hyperparametrov týchto metód. Ako model sme použili náhodný les, keďže zo všetkých modelov sa trénoval najrýchlejšie a patril medzi najlepšie modely (Tab. 1). Pri klasifikácii do ‘Position 4’ a ‘Position 13’ oversampling modely výrazne nezlepšil. Pri všetkých modeloch sa zlepšila metrika *f1 macro* a zas shoršila *f1 micro*, tj. model sa zlepšil na klasifikácii minoritných tried na úkor mierneho zhoršenie klasifikácie tých majoritných. Najlepší model s použitím SMOTE bol lepší o 2 percentuálne body v metrike *f1 macro* a 0.4 percentuálneho bodu horší v metrike *f1 micro*. Použitím oversampling-u sme dosiahli iba mierne zlepšenie v klasifikácii.

9.4 Výber atribútov (angl. feature selection)

Jednotlivé atribúty sme zatiaľ vyberali iba na základe doménových znalostí a podľa nami definovanej úlohy objavovania znalostí (Sekcia 3). Keďže je týchto atribútov stále veľa pokúsili sme sa ich zredukovať, aby sme mali jednoduchší a ľahšie interpretovateľný model. Vyskúšali sme niekoľko spôsobov, a to konkrétne: odstránenie atribútov z nízkou varianciou, výber atribútov na základe jednorozmerných štatistických testov (angl. univariate feature selection), rekurzívnu elimináciu atribútov, výber na základe ich dôležitosť pre model (feature importances). Týmto jednotlivým metódam sme skúšali rôzne parametre a výber atribútov sme vykonávali pre klasifikačnú úlohu (klasifikátor - náhodný les) a aj pre regresnú úlohu (lineárna regresia).

Výber atribútov výrazne nezlepšil úspešnosť modelu, ale ani nezhoršil, použitím polovice zo všetkých atribútov sme dosiahli tak dobré výsledky ako keby sme ich použili všetky (Obr. 7). Pri regresii bol tento trend podobný.

Výberom atribútov sa nám podarilo natrénovať jednoduchšie modely s podobnou úspešnosťou ako bez výberu atribútov.



(a) Úspešnosť pri klasifikácii 'Position 4' podľa počtu použitých atribútov

(b) Úspešnosť pri klasifikácii 'Position 13' podľa počtu použitých atribútov

Fig. 7: Úspešnosti pri klasifikácii podľa počtu použitých atribútov. Graf vznikol použitím rekurzívnej eliminácie s krížovou validáciou ($k=2$) na klasifikátore náhodný les.

References

1. Alberto Fernandez, Salvador Garcia, Francisco Herrera, Nitesh V. Chawla: SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research* **61**, 863–905 (2018), <https://www.jair.org/index.php/jair/article/view/11192>
2. Opitz, J., Burst, S.: Macro f1 and macro f1. *arXiv preprint arXiv:1911.03347* (2019)
3. Razali, N., Mustapha, A., Yatim, F., Aziz, R.: Predicting player position for talent identification in association football. *IOP Conference Series: Materials Science and Engineering* **226**, 012087 (08 2017). <https://doi.org/10.1088/1757-899X/226/1/012087>
4. Soto-Valero, C.: A gaussian mixture clustering model for characterizing football players using the ea sports' fifa video game system. [modelo basado en agrupamiento de mixturas gaussianas para caracterizar futbolistas utilizando el sistema de videojuegos fifa de ea sports]. *RICYDE. Revista Internacional de Ciencias del Deporte*. doi:10.5232/ricyde **13**(49) (2017), <https://www.cafyd.com/REVISTA/ojs/index.php/ricyde/article/view/1165>
5. Yaldo, L., Shamir, L.: Computational estimation of football player wages. *International Journal of Computer Science in Sport* **16** (07 2017). <https://doi.org/10.1515/ijcss-2017-0002>
6. Zawbaa, H.M., Emary, E., Grosan, C., Snasel, V.: Large-dimensionality small-instance set feature selection: A hybrid bio-inspired heuristic approach. *Swarm and Evolutionary Computation* **42**, 29–42 (2018)