Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.85
DOI: https://doi.org/10.24867/21BE21Stojadinov

ПРЕДВИЂАЊЕ ТОКА КАРИЈЕРЕ ТЕНИСЕРА ПОМОЋУ ТЕХНИКА АНАЛИЗЕ ПОДАТАКА

PREDICTING THE FLOW OF TENNIS PLAYER'S CARIER USING DATA MINING

Светлана Стојадинов, *Факултет техничких наука, Нови Сад*

Област – РАЧУНАРСТВО И АУТОМАТИКА

Кратак садржај – *У овом раду описан је процес кластеризације методом к-средњих вредности на основу изабраних физичких карактеристика и стилова игре тенисера. За сваки кластер креирана је регресиона крива, која указује на зависност успеха играча кроз сезоне. Цео процес валидиран је leave-one-out методом. Сви кораци система су детаљно описани и визуализовани.*

Кључне речи: *тенис, анализа података, кластеровање, регресија*

Abstract – *This paper describes the process of clustering tennis players based on selected physical characteristics and playing styles. A regression curve was created for each cluster, which indicates the dependence of the players' success through the seasons. The entire process was validated using the leave-one-out method. All steps of the system are described and visualized in detail.*

Keywords: *tennis, data mining, clustering, regression*

1. УВОД

У данашње време веома је упадљива човекова потреба за физичком активношћу и бављењем спортом, као начином за побољшање квалитета живота, елементом за одмор, забаву, рекреацију и дружење. Позитивно утиче на усвајање здравог начина живота и један је од кључних фактора за здрав живот. Од чисте љубави према неком спорту веома је лако препознати пут према професионалном бављењу одређеног спорта.

Спортови, попут тениса, захтевају посебан начин живота и максималну посвећеност и одрицање и то не само на тренинзима и у мечевима, него сваког дана. Представља стил живота, а не обавезу. Свакодневним улагањем себе у одређени спорт и одређена поља углавном доводи до нових амбиција и веома утиче на формирање такмичарског духа, као и тежње да се постане најуспешнији. Сразмерно популарности тениса као спорта и саме тежње тенисера да постану најбољи, расте и знатижеља љубитеља спорта који тенисер ће и успети да оствари успех у ближој будућности.

Проблем којим се бави овај рад јесте груписање и анализа играча тениса на основу физичких особина и карактеристика стила игре ради предикције каријере играча. Само груписање играча врши се кластеризацијом, која служи за проналажење сличних каријера на основу особина које нису унапред специфициране. Подаци над којима се врши груписање су претходно процесирани и прилагођени за даљу обраду.

Кластеризација се врши методом к-средњих вредности, којом се унапред дефинише у колико кластера ће бити распоређене јединке. Поред тога, детаљном анализом група сличних играча добијена је и крива старења за сваку од њих, која указује на зависност успеха играча кроз сезоне. Издвајањем одређеног играча из кластера могуће је проверити колико одступа од криве. Рачунањем средње вредности одступања проверава се колико су одступања велика.

2. ПОДАЦИ

За сврхе овог пројекта коришћена су два скупа података [1, 2] из којих су селектовани само потребни атрибути. Након селекције потребних особина, скупови података су спојени на основу заједничког идентификатора.

2.1. Подаци о играчима

Први скуп [1] обухвата податке који садрже вредности о карактеристикама сваког играча појединачно. Међу атрибутима овог скупа налазе се лични подаци као што су име, презиме, информације везане за место и датум рођења, место становања и прва сезона када је одређени играч постао професионалац. Овај скуп података садржи такође и веома корисне информације о физичким особинама и податке који описују стил игре играча, што се сматра да има највећи утицај на каријеру коју ће он имати као професионалац. Подаци о играчима искоришћени су за сврху прављења кластера тенисера на основу одабраних особина.

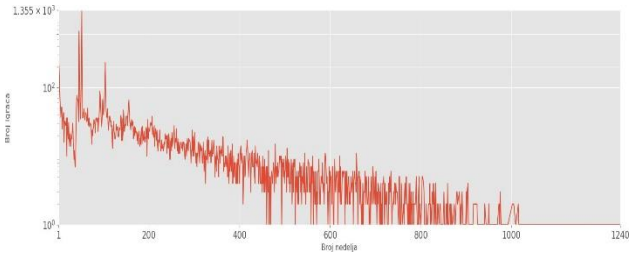
Висина и тежина изражене су у две, односно три различите јединице мере (у фунтама и килограмима, односно стопама, инчима и центиметрима) са прецизношћу једне децимале. Као две карактеристике које описују стил игре наведени су којом руком играч држи рекет и при томе су могуће вредности лево или десно. Међутим друга карактеристика стила игре је навођење доминантне руке играча у свакодневном животу (дакле, леваци или дешњаци), при чему постоје они који се изјашњавају да су им обе руке доминантне. Овај скуп бележи податке о преко 10.000 играча. Важно је напоменути да овај скуп садржи

НАПОМЕНА:
Овај рад проистекао је из мастер рада чији ментор је био др Александар Ковачевић, ред. проф.

недостајуће вредности као и очигледно нетачне вредности (на пример тежина од 0 килограма). Ови проблеми, али и уочавање и уклањање података који одступају ван граница прихватљивости решени су претпроцесирањем података и њиховом даљом припремом за примену у алгоритмима и визуализацији.

2.2. Подаци о недељном стању ранг листе

Други скуп [2] података садржи недељне прегледе стања *АТП* ранг листе. Скуп не садржи податке за сваку недељу - густина забележених недеља варира у сезонама. Најмања густина забележених недеља је у почетним годинама скупа података.



Слика 1. *Зависност броја недеља и броја играча*

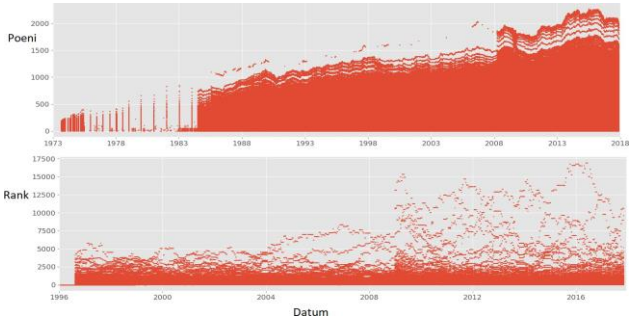
На слици 1. приказан је граф који представља зависност евидентираних играча и броја забележених недеља у току каријере. Атрибути који су коришћени из овог скупа су датум почетка недеље чије стање анализира и позиција одређеног играча у тој недељи. Подаци тенисера о недељном положају на *АТП* ранг листи се бележе од 1973. године, док се подаци о недељним *АТП* поенима тенисера бележе тек од 1996. године.
Од почетка бележења података о поенима, начин бодовања различитих категорија турнира се мењао. Ово представља велики фактор при анализи каријера играча у прелазним периодима (када престане примена једног система бодовања, а наступи употреба другог система бодовања), као и при упоређивању каријера играча из различитих периода. Разликују се системи бодовања из периода од 1996. до 1998. године, затим 1999. године, од 2000. до 2008. године и од 2009. године до данас.

Услед ових промена у начину бодовања различитих категорија турнира, графиконе података са положајем на *АТП* ранг листи и поенима играча карактеришу скокови на прелазима са једног система у други. Најочигледнији је скок између 2008 и 2009. године.

На слици 2. визуализовани су подаци о рангу (положају на *АТП* ранг листи) и *АТП* поени играча, како би се увидели скокови на прелазима система, као и густина забележених недеља кроз године.

3. ПРЕТХОДНА РЕШЕЊА

Рад се ослања на два претходна истраживања на тему анализе каријера тенисера и генерално спортиста. Један од циљева радова је описивање каријера спортиста путем регресионе криве, како би утврдиле законитости у каријерама спортиста.



Слика 2. *Зависност ранг/поена и времена*

Рад [3] се ослања на скуп података о положају тенисера на *АТП* ранг листи од 1973. до 2017. Рад тежи да утврди просечну криву перформанси професионалних тенисера из периода од 1974 до 2014, као и када ће тенисери достићи врхунац своје каријере. Такође описује рад алгоритма за предвиђање броја *АТП* поена у наредним годинама.

Утврђено је да већина тенисера достиже врхунац каријере у двадесет петој години живота. Такође, већина тенисера постиже највишу позицију на *АТП* ранг листи након 9 година каријере. Утврђена је крива којом се описује успех играча у току каријере.

Рад [4] се бави предвиђањем године живота спортиста када ће постићи врхунац каријере, за различите дисциплине. Анализирани су подаци са олимпијских игара, као и подаци о успеху из сваке дисциплине, за различите године. Посматрани су средња вредност година, стандардна девијација и модуо година за сваку спортску дисциплину, као и категорију у оквиру бејзбола.

Конкретно за тенис, утврђено је да тенисери у просеку постигну врхунац каријере са 24 године по модулу година, односно 25.43 по просеку година. Тенисерке достижу врхунац каријере са 23 године (модул година), односно 24.46 (просек година).

4. ОБРАДА ПОДАТАКА

Правилан одабир података, њихова обрада и доступност алата за њихову прераду и анализу је кључан фактор у радовима за анализу и обраду података. Рад са великом количином података доводи до великих организационих изазова. Подаци који се користе у обрадама, могу бити прикупљени на различите начине. Типови података могу бити представљени кроз различите типове. Радови, који се баве темама предлагања метода за прикупљање и обраду доступних података, су засновани на томе да објасне једноставне и ефикасне методе које служе за одабир информација од важности за даљу примену и њихово коришћење.

У овом раду велика пажња је посвећена детаљној и прецизној селекцији и прикупљању података, затим њиховој трансформацији, јер подаци представљају почетни корак за долазак до резултата. Битни кораци при припреми података су: одбацивање непотпуних података, одбацивање нерелевантних атрибута, трансформација типова података у други облик итд. Један од изазова у овом раду био је спајање података из два различита извора.

4. КЛАСТЕРИЗАЦИЈА

Већ припремљени подаци били су улазни параметар за почетак обраде и коришћење одговарајућих алгоритама.

Замисао овог сегмента рада је да се групишу и касније анализирају играчи са сличним карактеристикама са жељом да се покажу сличности и у њиховим каријерама без обзира на еру у којој су рођени и играју као професионалци. Узимајући ово у обзир, први корак је представљао процес кластеризације. За те потребе искоришћен је метод к-средњих вредности користећи библиотеку *scikit-learn* која је део програмског језика *python* у коме је имплементиран рад.

У зависности од вредности параметра *к* који се задаје, играчи су били класификовани у задат број кластера. У овом раду за број кластера изабрана је вредност 10. Кораке доделе и ажурирања је потребно понављати док се не постигне апсолутна конвергенција или достигне максималан број итерација.

5. РЕГРЕСИЈА

Подаци о недељном положају играча на *АТП* ранг листи, као и подаци настали из кластеризације играча према њиховим физичким особинама, користе се у регресионој анализи тока каријере тенисера. Циљ овог дела рада је да се покажу потребни кораци за добијање довољно добре регресионе криве која описује ток каријере тенисера. Регресиона функција тежи да најбоље могуће опише све каријере играча једног кластера, при чему је издвојена каријера првог играча ради валидације.

5.1. Претпроцесирање за регресију

Један од проблема који је уочен у процесу припреме података за регресију, је тај што постоје играчи који се појављују у скупу података играча, али нису ни једном наведени у ранг листи неке од седмица. Као решење на настали проблем играчи из датотека кластера, за које не постоје подаци о положају на *АТП* ранг листи или *АТП* поенима, су уклоњени из скупа. Такође се одбацују и подаци играча за које није сигурно да ли су њихове каријере у целости забележене у подацима.

То су играчи чији подаци се налазе у првој и последњој години података о недељном положају на *АТП* ранг листи, односно 1973. и 2017. године. Играчи за које је забележено мање од 10 недеља су изостављени из скупа јер се нису довољно дуго такмичили као професионалци и сматрају се нерелевантним.

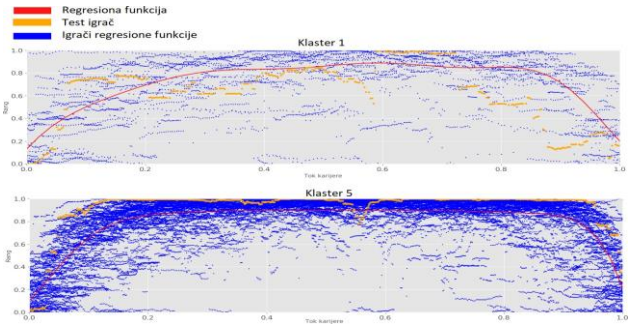
5.2. Кораци

Потребно је да се криве тока каријере играча из одређеног кластера преклопе, независно од датума када су забележени подаци, када је започета или завршена каријера играча, као и колико дуго је трајала каријера. Тачније, потребно је да се криве преклапају по апсциси, која описује временски ток каријере. Слично, потребно је да се подаци преклапају и по ординати, односно независно од положаја на *АТП* ранг листи које је постигао играч у току каријере. Нормализацијом се уклањају наведене зависности, односно своде да вредности од 0 до 1.

Нормализовани подаци о положају играча на *АТП* ранг листи играча у кластеру се издвајају у посебан скуп. Записи о положају на *АТП* ранг листи се затим сортирају по временској оси, у зависности од нормализованог тренутка у каријери када је податак забележен.

5.3. Валидација регресије

Тачност регресије се проверава тако што се подаци првог играча из кластера пореде са регресионом кривом насталом од података осталих играча. Због тога се за све играче (осим првог играча) у кластеру проналази регресиона функција. Одступање регресионе криве од података првог играча се посматра као средња вредност разлика положаја на *АТП* ранг листи валидационог играча и генерисаних положаја на *АТП* ранг листи. Ово одступање се користи као валидација.



Слика 3. *Регресиона крива*

6. РЕЗУЛТАТИ

Слика 3. приказује каријере и регресиону криву играча из два већа кластера. Наранџастим маркерима приказани су подаци првог играча у кластеру, плавим маркерима приказани су подаци свих осталих играча у кластеру, а црвеном линијом приказана је регресиона крива играча у кластеру (осим првог играча). Горњи графикон одговара првом кластеру, а средње одступање криве од података првог играча је мање од 18.13% висине каријере. Доњи графикон одговара петом кластеру, а за овај кластер је добијено средње одступање које је мање од 13.21% висине каријере.

Великом одступању података првог играча у кластеру, у односу на регресиону функцију кластера којем играч припада, доприноси неколико чинилаца. Због ових чиниоца, каријера играча се ретко кад може описати правилном кривом, јер је карактерише шум. Овај шум се испољава у скоковима и падовима положаја на *АТП* ранг листи, вишим положајима на почетку и крају каријере у односу на средину каријере, као и недостатком података.

Један од најчешћих чинилаца шума у криви тока каријере играча је што играчи често у току каријере направе паузу услед повреде или других здравствених разлога. Такве паузе трају од неколико месеци, па до неколико година. Услед паузе, положај на *АТП* ранг листи и поени доживљавају огроман пад. Каријера играча може и да се прекине у тренутку повреде или

другог личног догађаја. Такве каријере карактеришу високи положаји на *АТП* ранг листи на крају каријере, што битно утиче и на облик регресионе функције.

Други чинилац шума су промене у начину рангирања играча. Како је раније напоменуто, због ових промена система бодовања играча, каријере играча карактеришу степенице, односно континуални скокови или подови ранга.

Те степенице се јављају на прелазним годинама, где престаје употреба једног система бодовања и рангирања, а почиње употреба другог система. Регресиона крива не може да предвиди овакве поремећаје у току каријере играча.

Трећи чинилац великог одступања регресионе криве од каријере валидационог играча је чињеница да неке каријере уопште не прате криву. Каријере играча који су у току каријере постигли слабе резултате не карактерише раст ранга и поена у току почетка каријере, а ни пад ранга и поена на крају каријере. У току таквих каријера постоје чак и забележени падови на нижи ранг, од првог забележеног ранга.

Овакве каријере су валидне, услед тога што постоји довољни број убележених положаја на *АТП* ранг листи, као и дужина каријере није занемарљива. Због тога, овакви играчи се не смеју избацити из скупа играча којим се генерише регресиона функција. Последица узимања у обзир оваквих играча је што регресиона крива лошије описује играче са правилним каријерама.

Услед свих ових чинилаца, на слици 3. се примећује да постоје положаји на *АТП* ранг листи који су далеко од регресионе криве (плави маркери). Подаци који се налазе испод криве највероватније чине повреде и сличне паузе у току каријере.

Положаји изнад почетка регресионе криве највероватније чине играчи са неправилним каријерама. Положаји изнад краја криве чине играчи који су нагло завршили каријеру, или имају неправилне криве каријере. Овакав шум је битно померио врх каријере према нижим положајима, а почетак и крај каријере ка вишим положајима.

7. ЗАКЉУЧАК

Овим радом објашњени су поступци при реализацији пројекта који се бавио анализом тениских каријера почевши од 1973. године до 2017. године. Играчи су груписани према својим физичким особинама и стилу игре и према добијеним кластерима је одређена карактеристична крива за просечну каријеру сваког кластера. Ова крива представља такозвану крива старења и служи за предвиђање каријере новог играча који би припадао том кластеру. У добијеним резултатима примећен је шум који је оправдан ситуацијама као што су повреде или нагли завршетак каријере. Оваква анализа могла би да се искористи у сврхе одабира играча који ће бити заштитно лице одређених спонзора или при неком од облика предвиђања успешности одређеног играча у сезони.

8. ЛИТЕРАТУРА

[1] Association of Tennis Professionals (ATP), Github https://github.com/serve-and-volley/atp-world-tour-tennis- data.git.

[2] Association of Tennis Professionals (ATP), ATP World tour www.atpworldtour.com

[4] Miha Mlakar, Tea Tušar , “Analyzing and Predicting Peak Performance Age of Professional Tennis Players” , Jožef Stefan Institute Jamova cesta 39 SI-1000 Ljubljana, Slovenia, pp 1-4, 2017.

[3] “Peak Performance and Age Among Superathletes: Track and Field, Swimming, Baseball, Tennis, and Golf”, Richard Schulz, Christine Curnow, Jurnal of Gerontology: PSYCHOLOGICAL SCIENCES, Vol. 43, No. 5, P113 – 120, 1988.

Кратка биографија:

Светлана Стојадинов рођена је у Зрењанину 1994. године. Мастер рад на Факултету техничких наука из исте области одбранила је 2022. године.

контакт: stojadinov.svetlana@gmail.com