https://doi.org/10.37528/FTTE/9788673954318/POSTEL.2020.022

PREDIKCIJA BAZIRANA NA MAŠINSKOM UČENJU I SENZORSKIM PODACIMA

Snežana Mladenović[1], Ana Uzelac[2], Stefan Zdravković[3], Ivana Andrijanić[4]
[1]Univerzitet u Beogradu - Saobraćajni fakultet, snezanam@sf.bg.ac.rs
[2]Univerzitet u Beogradu - Saobraćajni fakultet, ana.uzelac@sf.bg.ac.rs
[3]Univerzitet u Beogradu - Saobraćajni fakultet, s.zdravkovic@sf.bg.ac.rs
[4] Javno preduzeće „Putevi Srbije“, Beograd, ivana.andrijanic@putevi-srbije.rs

Rezime: *U mnogim oblastima kontinuirano se prikupljaju podaci koristeći različite pametne tehnologije, među kojima su naročito zastupljeni senzori. Analiza senzorskih podataka zahteva pristupe koji će omogućiti da se otkriju zakonitosti u podacima koje nisu ni poznate, ni očigledne, a mogu biti korisne. U drumskom saobraćaju koriste se podaci o intenzitetu i strukturi saobraćajnih tokova, od faze planiranja putne infrastrukture, preko izgradnje, do njenog korišćenja i održavanja. Karakteristike saobraćajnih tokova rezultat su obrade podataka dobijenih brojanjem saobraćaja. U ovom radu istraživane su mogućnosti primene metode nadgledanog mašinskog učenja na saobraćajnim podacima prikupljenim od strane automatskih brojača saobraćaja, čiji je rad baziran na induktivnim petljama. U radu su predstavljeni rezultati predikcije karakteristika budućih saobraćajnih tokova na posmatranim deonicama puteva, dobijeni primenom različitih algoritama mašinskog učenja u softverskom alatu Weka.*

Ključne reči: *mašinsko učenje, Weka, Python, predikcija saobraćaja*

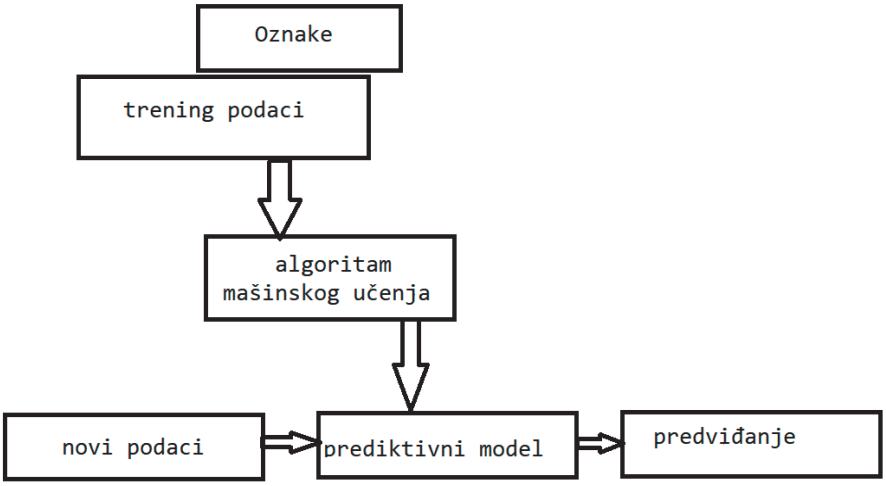
1. Uvod

Razvojem bežičnih tehnologija i sve većom primenom senzora generišu se ogromne količine podataka. Kao takvi, senzorski podaci u saobraćaju imaju *Big Data* obeležja i predstavljaju podesan izvor podataka za određivanje saobraćajnih pokazatelja koji se navode u zvaničnim statistikama i izveštajima. Predikcija na osnovu prikupljenih senzorskih podataka ima važnu ulogu u razvoju pametnih saobraćajnih sistema. Cilj praćenja drumskog saobraćaja je prikupljanje informacija o različitim učesnicima u saobraćaju, koje su u okviru pametnih saobraćajnih sistema neophodne za regulisanje i nadgledanje, bezbednijeg i ekološki prihvatljivijeg saobraćaja. Prikupljanje, upravljanje i kontrola velikog skupa senzorksih podataka donosi mnogo izazova u koje se pored ostalih može uvrsititi i kvalitet prikupljenih podataka. Automatsko brojanje vozila predstavlja jednu od mera koja se koristi u nadgledanju i kontroli saobraćajnog toka. U

zavisnosti od veličine i složenosti posmatranog saobraćajnog sistema koriste se različite metode i uređaji, kao što su: kamere za video nadzor, pneumo senzori, piezoelektrični senzori, induktivne petlje, magnetni senzori, akustični detektori, pasivni i aktivni infracrveni senzori, Doplerovi mikrotalasni senzori. Nadgledanje saobraćajnog toka korišćenjem magnetnih bežičnih senzorskih mreža predstavlja deo inovativne tehnologije; međutim i kod primene ove tehnologije javljaju se problemi. Ukoliko se primenjuju postojeći algoritmi skeniranja i odlučivanja, senzori zasnovani na zvuku mogu da imaju nedostatke u slučaju prisustva ostalih izvora buke koji mogu da uzrokuju greške u otkrivanju i brojanju vozila [1]. Druge tehnologije, poput merenja intenziteta saobraćaja i brzine vozila zasnovane na Doplerovim mikrotalasnim senzorima dovode u pitanje tehničku i ekonomsku opravdanost ovih aktivnosti [2]. Korišćenje akustičnih detektora u proceni drumskog saobraćaja se retko primenjuje u praksi. Međutim, postoje eksperimentalna istraživanja u ovoj oblasti [3, 4]. Integrisani senzori magnetnog polja imaju prednosti u odnosu na ostale senzorske tehnologije poput neosetljivosti na klimatske i vremenske prilike i niske cene [5]. Upotreba ove tehnologije uglavnom je zasnovana na anizotropno-magnetno otpornom (*eng. anisotropic magnetoresistance* - AMR) tipu senzora. S druge strane, snimanje saobraćaja pomoću video uređaja omogućava preciznije praćenje saobraćaja, monitoring više saobraćajnih traka i veći skup prikupljenih podataka. Međutim, nedostaci ove tehnologije mogu biti: neophodno periodično čišćenje sočiva, negativan uticaj loših vremenskih uslova na performanse uređaja, koji uz to zahtevaju i značajne računarske resurse [6]. Dodatno, kod sistema klasifikacije saobraćaja koji su zasnovani na video uređajima javljaju se problemi obrade velike količine prikupljenih podataka, čime se značajno produžava vreme otkrivanja i klasifikacije. Različite senzorske tehnologije imaju određene prednosti i nedostatke u pogledu troškova implementacije i održavanja, veličine, energetske potrošnje i lakoće instalacije. U narednim sekcijama rada istraživane su mogućnosti primene metode mašinskog učenja na saobraćajnim podacima prikupljenim pomoću automatskih brojača saobraćaja, čiji je rad zasnovan na tehnologiji induktivnih petlji. Ovu tehnologiju karakterišu: princip rada zasnovan na merenju induktivnosti, niski troškovi implementacije, velika energetska potrošnja i složenost same implementacije.

Sveprisutnost senzora dovodi do generisanja velike količine kako struktuiranih tako i nestruktuiranih podataka. Za prepoznavanje određenih pravilnosti i otkrivanje zakonitosti u senzorskim podacima pogodno je koristiti metodu mašinskog učenja. Mašinsko učenje obezbeđuje alate kojima se velike količine podataka mogu automatski analizirati. To je jedna od oblasti računarske analize podataka koja se poslednjih decenija najbrže razvija.

Postoje tri tipa mašinskog učenja: nadgledano učenje (eng. *supervised learning*), nenadgledano (eng. *unsupervised learning*) i učenje uz podsticaje (eng. *reinforcement learning*) [7]. Nadgledano učenje se primenjuje nad označenim podacima, ovakvo učenje daje nam direktno povratne informacije i ima mogućnost predviđanja. Kod nenadgledanog učenja nemamo označenih podataka, niti povratnih informacija, ali možemo da otkrijemo skrivene obrasce u podacima. Učenje uz podsticaje podrazumeva proces učenja gde model donosi određene odluke na osnovu pokušaja i grešaka, pri čemu svaka akcija donosi ili nagradu ili neku kaznu.



Slika 1. Tok procesa nadgledanog mašinskog učenja

U ovom radu su korišćene tehnike nadgledanog mašinskog učenja nad dostupnim skupom podataka sa ciljem da se uradi predikcija jednog od ciljnih atributa. Prvo je model istreniran koristeći označen skup podataka za učenje. Istrenirani model je nakon toga u mogućnosti da predvidi vrednosti ciljnog atributa na novom testnom skupu podataka. Tipičan tok procesa nadgledanog mašinskog učenja prikazan je na slici 1 [8]. Ovaj rad predstavlja nastavak ranijeg istraživanja dela autorskog tima ovog rada koje je opisano u [9].

2. Metodologija

Studija slučaja u ovom radu je zasnovana na skupu podataka koji su generisali izabrani automatski brojači saobraćaja, na državnim putevima i kategorije u Srbiji, u periodu od 01.01.2011. do 31.12.2018. godine. U radu je korišćen skup podataka koji je generisao jedan od brojača koji nosi oznaku 1270. Izabrani brojač se nalazi na putu broj 23, IB kategorije u mestu Preljina.

Početni skup podataka sadrži podatke o broju vozila, posebno za svaki smer, koje je svakog sata registrovao izabrani brojač saobraćaja. Dati brojač je za period od 8 godina generisao skup podataka koji se sastoji od ukupno 140.256 instanci, a svaka instanca sadrži sledeće atribute: *datum, dan u nedelji, sat, smer* i *broj vozila.* Cilj ovog istraživanja bio je da se na pomenuti skup podataka primeni tehnika nadgledanog mašinskog učenja i izvrši predikcija broja vozila na izabranom brojačkom mestu, po smerovima za svaki sat. Tačnije, cilj je bio da se izvrši predikcija intervala u kojem će se nalaziti ukupan broj vozila za svaki sat, u svakom danu, na izabranom brojačkom mestu. Ako je broj vozila imao vrednost između 0 i 100, taj interval je označen sa 1, ako je broj vozila iznosio između 101 i 200, to je bio interval sa brojem 2, itd. Ukupno je bilo 11 intervala.

Kao skup podataka za obučavanje modela izabrane su instance koje se odnose na period od 2011-2015. godine, dok su instance koje se odnose na period od 2016-2018.

godine korišćene za testiranje napravljenog modela mašinskog učenja. Tako je skup podataka za učenje sadržao 87.648 instanci, dok je skup za testiranje sadržao 52.608 instanci.

U prvom delu ovog istraživanja izvršena je priprema podataka s ciljem da se od početnih nesređenih skupova podataka dobiju podaci koji su pogodni za kasniju obradu. Od izvornih datoteka do datoteka koje su spremne za primenu različitih algoritama mašinskog učenja došlo se korišćenjem makroa za Excel napisanih u VBA (*Visual Basic for Applications*) i programa napisanih u programskom jeziku Python. Svaki konkretan broj vozila je zamenjen intervalom kom pripada i to je takođe urađeno u programskom jeziku Python.

Pripremljeni skup podataka za učenje sastoji se, očekivano, od 87.648 instanci. Cilj je bio da se projektuje atribut koji označava kom intervalu pripada broj vozila. Pošto je projektovani atribut numeričkog tipa, na njega su primenjeni svi regresioni algoritmi za mašinsko učenje, koji su dostupni u softverskom alatu Weka (*Waikato Environment for Knowledge Analysis*) [10], kao što su: linearna regresija, regresiono drvo, neuronske mreže, itd. Modeli su trenirani i ocenjeni 10-strukom unakrsnom validacijom na skupu podataka za učenje, a potom su najbolji od njih dodatno ocenjeni na skupu podataka za testiranje.

Za upoređivanje algoritama korišćene su sledeće metrike: koeficijent korelacije (1), srednja apsolutna greška (5) i srednja kvadratna greška (6). Ukupan broj instanci za testiranje je n; projektovane vrednosti na testnim instancama su $p_1, p_2, ..., p_n$; stvarne vrednosti su $a_1, a_2, ..., a_n$; $\overline{p}$ i $\overline{a}$ su srednje vrednosti projektovanih, odnosno stvarnih vrednosti.

$$koeficijent\ korelacije = \frac{S_{PA}}{\sqrt{S_p S_A}} \tag{1}$$

gde je

$$S_{PA} = \frac{\sum_{i=1}^{n} (p_i - \overline{p})(a_i - \overline{a})}{n-1} \tag{2}$$

$$S_p = \frac{\sum_{i=1}^{n} (p_i - \overline{p})^2}{n-1} \tag{3}$$

$$S_A = \frac{\sum_{i=1}^{n} (a_i - \overline{a})^2}{n-1} \tag{4}$$

$$srednja\ apsolutna\ greška = \frac{|p_1 - a_1| + ... + |p_n - a_n|}{n} \tag{5}$$

$$srednja\ kvadratna\ greška = \frac{(p_1 - a_1)^2 + ... + (p_n - a_n)^2}{n} \tag{6}$$

3. Rezultati i analiza rezultata

Obučavanje, validacija i testiranje modela mašinskog učenja obavljeni su u *data mining* softverskom alatu Weka. Weka je softver otvorenog koda, kreiran na Univerzitetu Waikato na Novom Zelandu. Ovaj softver predstavlja kolekciju algoritama mašinskog učenja koji se koriste u poslovima koji se bave otkrivanjem zakonitosti u podacima. Ovaj softver je izabran jer je lak za rukovanje, otvorenog je koda i poseduje sve alate koji su neophodni za izvršavanje poslova u procesu mašinskog učenja. Od svih testiranih algoritama, više algoritama je dalo dobre rezultate, a odabrano je pet najboljih.

Performanse pet najboljih algoritama mašinskog učenja, primenjenih na skupu podataka za učenje, a koje su dobijene korišćenjem 10-struke unakrsne validacije, prikazane su u Tabeli 1.

Tabela 1: Performanse najboljih modela mašinskog učenja primenjenih na skupu podataka za učenje

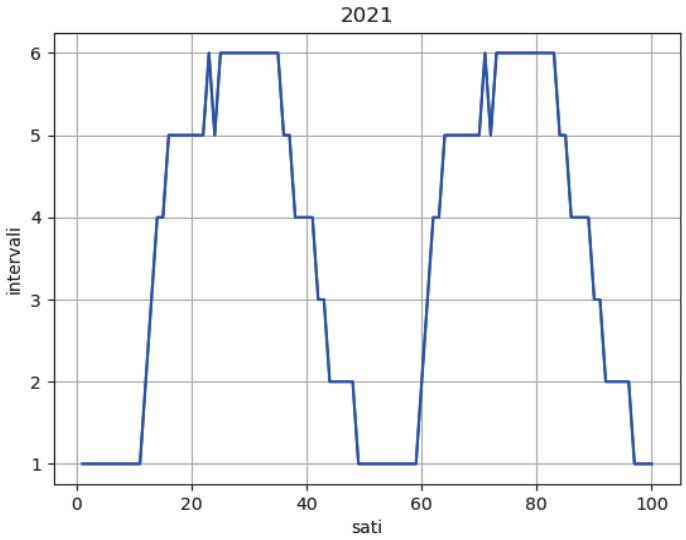
Algoritam za mašinsko učenje	Koeficijent korelacije	Srednja apsolutna greška	Srednja kvadratna greška
Random Forest	0.9620	0.3926	0.5438
Random Committee	0.9539	0.3972	0.6005
Bagging	0.9335	0.5137	0.7140
RandomTree	0.9348	0.9348	0.4464
Lazy IBk	0.9286	0.4720	0.7514

Radi određivanja modela koji najbolje rešava postavljeni problem, izvršena je evaluacija svih pet modela. Svi modeli rešavaju zadati problem, i performanse su približno iste. Najbolji koeficijent korelacije na skupu podataka za učenje ima *Random Forest* algoritam, a ima i najmanju srednju apsolutnu grešku. Kako bi se ostvarila što objektivnija procena performansi modela, evaluacija je sprovedena i nad novim, nepoznatim skupom podataka – skupom podataka za testiranje. Evaluacija je vršena korišćenjem istih metrika kao prilikom pravljenja modela. Performanse izabranih modela mašinskog učenja dobijene na skupu podataka za testiranje prikazane su u Tabeli 2. Iz tabele se može videti da je model baziran na Lazy IBk algoritmu pokazao malo bolje performanse (ima veći koeficijent korelacije, ali istovremeno i manju srednju apsolutnu i srednju kvadratnu grešku) od ostalih modela. Zato je model baziran na Lazy IBk algoritmu izabran kao najbolji.

Tabela 2: Performanse najboljih modela mašinskog učenja primenjenih na skupu podataka za testiranje

Algoritam za mašinsko učenje	Koeficijent korelacije	Srednja apsolutna greška	Srednja kvadratna greška
Random Forest	0.9128	0.8061	1.0850
Random Committee	0.9115	0.8062	1.0884
Bagging	0.9002	0.8591	1.1389
RandomTree	0.9115	0.8063	1.0884
Lazy IBk	0.9451	0.6795	0.9296

Korišćenjem modela koji je zasnovan na Lazy IBk algoritmu, urađena je i predikcija intervala kojem će pripadati broj vozila po smerovima na odabranoj lokaciji za svaki sat svakog dana u 2021. godini. Prvih 100 projektovanih vrednosti su vizuelizovane korišćenjem programskog jezika Python i prikazane su na slici 2.



Slika 2: Projektovani protok vozila po satu za oba smera u 2021. godini za brojač 1270, prikaz prvih 100 instanci

4. Zaključak

U sprovedenoj studiji slučaja obučavani su prediktivni modeli bazirani na regresionim algoritmima, i to: Lazy IBk (k-Nearest Neighbors), Random Forest, Random Tree i Random Committee. Na skupu podataka za treniranje, svi algoritmi su imali približne performanse, s tim da je Random Forest algoritam bio za nijansu bolji od ostalih. U procesu evaluacije modela na testnom skupu korišćenje istih metrika je ponovo pokazalo da su modeli prlibližno sličnih performansi, ali je ovog puta model zasnovan na algoritmu Lazy IBk pokazao rezultate koji su malo bolji od ostalih.

Rezultati ovog istraživanja podudaraju se sa rezultatima ranijeg istraživanja gde je rađena predikcija konkretnog broja vozila (a ne predikcija intervala u koju će taj broj da „upadne“) [9]. U [9] je Random Forest algoritam imao najbolje rezultate na trening skupu podataka dok je Lazy IBk algoritam pokazao najbolje rezultate na testnom skupu podataka.

Ono što je ostalo neistraženo a ima smisla da se uradi u budućnosti je primena klasterovanja kojim bi bili izdvojeni delovi dana koji predstavljaju “špiceve” sa najvećim brojem vozila, nasuprot onim delovima dana kada je protok vozila značajno manji.

Zahvalnost

Ovaj rad delimično je podržan od strane Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije, u okviru projekata pod brojevima 032025 i 036012.

Literatura

[1] M. A. Adnan, N. Sulaiman, N. I. Zainuddin, and T. B. H. T. Besar, "Vehicle speed measurement technique using various speed detection instrumentation". *BEIAC 2013 - 2013 IEEE Business Engineering and Industrial Applications Colloquium*, 2013, 668–672. https://doi.org/10.1109/BEIAC.2013.6560214

[2] A. Czyzewski, J. Kotus, and G. Szwoch, "Estimating traffic intensity employing passive acoustic radar and enhanced microwave doppler radar sensor". *Remote Sensing*, 2020, 12(1). https://doi.org/10.3390/rs12010110

[3] H. S. Fimbombaya, N. H. Mvungi, N.Y. Hamisi, & H.U. Iddi, "Enhanced Magnetic Wireless Sensor Network Algorithm for Traffic Flow Monitoring in Low-Speed Congested Traffic". *Journal of Electrical and Computer Engineering*, 2020, 1–8. https://doi.org/10.1155/2020/5875398

[4] Y. Na, Y. Guo, Q. Fu, and Y. Yan, "An acoustic traffic monitoring system: Design and implementation". Proceedings - *2015 IEEE 12th International Conference on Ubiquitous Intelligence and Computing, 2015 IEEE 12th International Conference on Advanced and Trusted Computing, 2015 IEEE 15th International Conference on Scalable Computing and Communications, 2015 IEEE International Conference on Cloud and Big Data Computing, 2015 IEEE International Conference on Internet of People and Associated Symposia/Workshops*, UIC-ATC-ScalCom-CBDCom-IoP 2015, 119–126. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.41

[5] X. Zu, S. Zhang, F. Guo, Q. Zhao, X. Zhang, X. You, H. Liu, B. Li, and X. Yuan, "Vehicle Counting and Moving Direction Identification Based on Small-Aperture Microphone Array". *Sensors* (Basel, Switzerland), 2017, 17(5), 1–11. https://doi.org/10.3390/s17051089

[6] S. R. E. Datondji, Y. Dupuis, P. Subirats, and P. Vasseur, "A survey of visionbased traffic monitoring of road intersections". *IEEE Transactions on Intelligent Transportation Systems*, 2016, 17(10): 2681–2698

[7] E. Alpaydin, *Introduction to Machine Learning*, Fourth edition. The MIT Press, 2020.

[8] S. Raschka, V. Mirjalili, *Python Machine Learning*, Third edition. Birmingham, UK, Packt Publishing, 2019.

[9] A. Uzelac, S. Zdravković, S. Janković, D. Mladenović, I. Andrijanić, "Predikcija časovnog protoka vozila korišćenjem metoda Big Data analitike". *Zbornik radova XLVII Simpozijuma o operacionim istraživanjima - SYM-OP-IS '20*, 2020, 293-296

[10]I. Witten, E. Frank, M. Hall, C. Pal, *Data Mining, Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers, 2016.

Abstract: *In many different areas, data is continuously collecting with the aim to extract useful patterns. In order to collect different data, various smart technologies together with different kind of sensors are used. Analysis of data obtained through sensors require approaches that are able to extract not previously known, useful and subtle patterns from data. Data related to intensity and structure of traffic flows are used in different phases in the field of road traffic: from planning of road infrastructure through construction to its use and maintenance. The characteristic of traffic flows is the result of processing the data obtained by traffic counting. In this paper, the possibilities of applying the method of supervised machine learning on traffic data collected by automatic traffic counters based on inductive loops, are investigated. The paper presents the results of prediction of the characteristics of future traffic flows on the observed road sections, obtained by applying different machine learning algorithms in the Weka software tool.*

Keywords: *machine learning, Weka, Python, traffic prediction*

PREDICTION BASED ON THE MACHINE LEARNING AND SENSOR DATA

Snežana Mladenović, Ana Uzelac, Stefan Zdravković, Ivana Andrijanić