Pristupi optimizaciji hiperparametara kod klasičnih algoritama mašinskog učenja bez nadzora na problemu detekcije anomalija

Milan Pavlović
Departman za matematiku i informatiku,
Prirodno-matematički fakultet,
Univerzitet u Novom Sadu
milan.pavlovic@dmi.uns.ac.rs

Mentor: Prof. dr Miloš Radovanović
Predmet: Seminar 3, II godina doktorskih studija
Projekat: Ideja rada je nastala tokom istraživanja na projektu COLLABS, Horizon 2020.
Kontekst: Koristi se skup podataka iz prethodno objavljenog rada [12], ali na problemu detekcije anomalija sa opštijim ciljem primene.
Publikacija: Planirana za proširenu verziju rada sa više skupova podataka i eventualno metrika.

Anomalije nastaju usled mehaničkih ili ljudskih grešaka, promena u ponašanju sistema ili prirodnim odstupanjima u populaciji. Njihov nastanak ne mora nužno ukazivati na grešku već to može biti nov, nepoznat ili dosad nevidjeni proces. Vremenom su razvijeni razni algoritmi za otkrivanje anomalija koji se mogu koristiti za sprečavanje potencijalnih negativnih posledica. U slučaju kada su oznake dostupne, pristup izbora algoritama za detekciju anomalija je jasno definisan poput tipa učenja, podele skupa podataka, izbora metrike, itd. Nejasnoće nastaju kada ne postoje dostupne oznake ili su parcijalno dostupne. Uglavnom, razvijene tehnike za izbor optimalnih hiperparametara su specifične za odredjene algoritme bez nadzora, što dovodi do nedostatka opštih tehnika i izazova prilikom izbora modela. Motivacija ovog rada jeste približiti se odgovoru na pitanje: *"Kako optimizovati hiperparametre nenadgledanih algoritama klasičnog mašinskog učenja na problemu detekcije anomalija?"*. U radu su uporedjena tri različita pristupa koja se fokusiraju na organizaciju validacionog skupa podataka u zavisnosti od dostupnih oznaka i na izbor metrike za traženje skupa optimalnih vrednosti hiperparametara.

---

1. UVOD

Detekcija anomalija je problem koji je duži period izučavan u različitim oblastima i domenima primene. Razne tehnike su uspešno razvijene za rešavanje ovog problema. Neke su opšte namene, dok su druge više specifične za odredjeni domen. Otkrivanje anomalija je kritičan zadatak u mnogim bezbedonosnim okruženjima i odnosi se na problem nalaženja obrazaca u podacima koji nisu u skladu sa očekivanim ponašanjem [4]. Anomalije su značajne i kritične informacije u raznim domenima primene i zato su interesantne za analizu. Anomalija je posmatranje koje značajno odstupa u odnosu na ostale članove uzorka u kom se pojavljuje, tako da izaziva sumnju da je generisano drugim procesom [5]. U odnosu na dostupnost oznaka, metode detekcije anomalija se dele u sledeće četiri kategorije [4].

Nadgledana detekcija anomalija (eng. *Supervised anomaly detection*) podrazumeva da skup za treniranje sadrži označene normalne i abnormalne instance. Problem se rešava binarnim klasifikatorom i uglavnom se dolazi do problema nebalansiranih klasa, zbog malog broja anomalija. Takodje, metoda podrazumeva da se anomalije ne menjaju tokom vremena jer klasifikatori pretpostavljaju odredjenu raspodelu podataka.

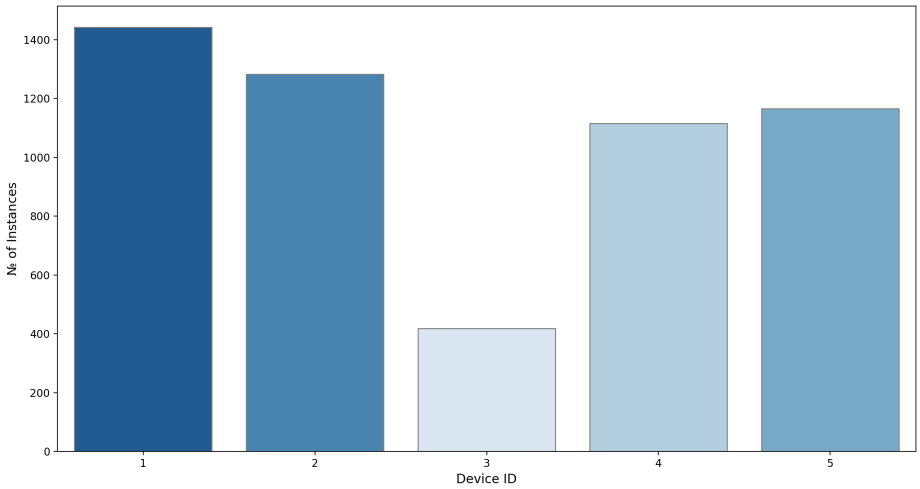
Slabo-nadgledana detekcija anomalija (eng. *Weakly-supervised anomaly detection*) pretpostavlja da postoje neke oznake klase anomalija, ali su te oznake parcijalne/nekompletne i/ili neprecizne [16]. Suština ovih metoda je da iskorišćavaju mali broj anomalija koje se mogu napraviti u realnom svetu uz relativno malu cenu koja zavisi od domena.

Polu-nadgledana detekcija anomalija (eng. *Semi-Supervised anomaly detection*) ili *detekcija noviteta* (eng. *Novelty detection*) [11] radi pod pretpostavkom da trening skup sadrži samo instance normalne klase. Ideja na kojoj se bazira ovaj pristup jeste pravljenje modela koji odgovara normalnom ponašanju. Ova metoda je pogodna za otkrivanje anomalija različitog tipa.

Nenadgledana detekcija anomalija (eng. *Unsupervised anomaly detection*) radi u nenadgledanom modu i uglavnom ne zahteva trening skup. To predstavlja glavni razlog zašto su ove metode najviše primenljive. Metode prave pretpostavku da su normalne instance dosta učestalije od anomalija, u suprotnom pate od velikog broja lažnih alarma.

U slučaju kada su oznake dostupne u skupu podataka, pristup izbora algoritama za detekciju anomalija je jasno definisan poput tipa učenja, podele skupa podataka, izbora metrike, itd. Nejasnoće nastaju kada ne postoje dostupne oznake ili su parcijalno dostupne [8,10,13]. Uglavnom, razvijene tehnike za izbor optimalnih hiperparametara su specifične za odredjene algoritme bez nadzora [14, 15], što dovodi do nedostatka opštih tehnika i izazova prilikom izbora modela. Istraživačkoj zajednici još uvek nedostaje uporedna univerzalna evaluacija, kao i zajednički javno dostupni skupovi podataka [8], dok se u praksi hiperparametri modela uglavnom ručno podešavaju.

Cilj ovog rada jeste uporedjivanje različitih pristupa za izbor nenadgledanih algoritama mašinskog učenja na problemu detekcije anomalija isključujući autoenkodere i algoritme klasterovanja. Hiperparametri predstavljaju parametre algoritma koji kontrolišu proces učenja i koji se ne menjaju tokom treniranja. Performanse modela tesno zavise od izabranog algoritma i od izbora njegovih hiperparametara. To je razlog zašto njihov izbor ima veliku ulogu pri razvijanju uspešnog modela. Opisani pristupi u ovom radu se fokusiraju na organizaciju validacionog skupa podataka u zavisnosti od dostupnih oznaka i na izbor metrike za traženje skupa optimalnih vrednosti hiperparametara. U narednom poglavlju su definisani i opisani različiti pristupi optimizacije hiperparametara koji su primenjeni u okviru eksperimenta. Zatim su opisani rezultati eksperimenta sa njihovom detaljnom analizom.



Grafikon 1: Raspodela skupa podataka

2. EKSPERIMENT

Eksperiment je postavljen tako da je svaki pristup procenjen na identičnom skupu za treniranje i testiranje. Dok skupovi podataka za validaciju teže da budu slični ili identični gde je to moguće. U narednoj podsekciji je opisan skup podataka na kom se zasniva eksperiment, zatim slede izabrani algoritmi i na kraju su opisani pristupi optimizacije hiperparametara.

2.1. SKUP PODATAKA

Skup podataka je generisan korišćenjem posebno dizajniranih 3GPP NB-IoT modula za generisanje bežičnih otisaka. Pet uredjaja je postavljeno u zatvorenu laboratoriju koja je oko 500m udaljena od makro-mobilne bazne stanice. Skup podataka je potom prikupljen korišćenjem softvera za prikupljanje prosečnih radio merenja. Za svaki od 5425 paketa snimljeno je 8 različitih merenja. [12]

Kako bi se napravio pogodan scenario za detekciju anomalija, pretpostavimo da su dva uredjaja sa najmanjim brojem signala nepoznata i/ili zlonamerna i da njihovi bežični otisci predstavljaju anomalije. Instance uredjaja 3 su obeležene oznakom ”1” i predstavljaju prvi tip anomalija, dok se instancama uredjaja 4 dodeljuje oznaka ”2” i predstavljaju drugi tip anomalija. Sa druge strane, preostalih tri uredjaja (1, 2 i 5) čine poznate uredjaje unutar sistema. Svi bežični otisci poznatih uredjaja predstavljaju normalne instance i njihove oznake su obeležene sa ”0”.

Nakon toga, 60% normalnih instanci je odredjeno za trening, dok preostalih 40% se deli tako što se 70% dodeljuje test skupu, a preostalih 30% validacionom skupu ukoliko postoji. Potom, test skupu su dodate sve anomalije tipa 2. Sa druge strane, validacioni skup je kreiran u zavisnosti od pristupa evaluacije, ali samo korišćenjem preostalih normalnih instanci i isključivo anomalija tipa 1. Time se dobija da je skup za testiranje identičan za sve pristupe. Pored toga, anomalija tipa 2 nikada nije vidjena tokom razvoj modela, što omogućava objektivnu i nezavisnu ocenu na test skupu. Polu-nadgledano učenje je implikacija ove podele, pošto se u trening skupu nalaze samo normalne instance.

2.2. ALGORITMI

Bitno je napomenuti da je proces traženja optimalnih hiperparametara pojednostavljen i da su izbačene dodatne tehnike poput selekcije ulaznih atributa, transformacije prostora karakteristika korišćenjem tehnika za smanjenje dimenzionalnosti i slično. Razlog tome je što nije moguće implementirati u svim pristupima dodatne tehnike na identičan način. Izuzetak je standardizacija ulaznih atributa kod algoritama koji se baziraju na računanju rastojanja izmedju instanci.

LOCAL OUTLIER FACTOR (LOF). Anomalični skor instance se naziva faktor lokalnog odstupanja i on meri lokalno odstupanje gustine date tačke u odnosu na njegove susede. Lokalno je po tome što anomalični skor zavisi od toga koliko je tačka izolovana u odnosu na okolno okruženje. Tačnije, lokalitet je dat pomoću k-najbližih suseda, čija se udaljenost koristi za procenu lokalne gustine. Uporedjivanjem lokalne gustine tačke sa lokalnim gustinama njegovih suseda mogu se identifikovati tačke koje imaju znatno nižu gustinu od svojih suseda i koje se smatraju anomalijama. [3]

ISOLATION FOREST (IFOREST). Šuma izolacije "izoluje" tačke nasumično birajući obeležje, a zatim nasumično birajući vrednost za podelu izmedju minimalne i maksimalne vrednosti izabranog obeležja. Pošto se rekurzivno particionisanje može predstaviti strukturom stabla, broj podela potrebnih za izolovanje posmatrane tačke je ekvivalentan dužini putanje od korenskog čvora do lista. Ova dužina puta, usrednjena u šumi takvih nasumičnih stabala, je mera normalnosti i funkcija odlučivanja. Nasumično particionisanje proizvodi primetno kraće putanje za anomalije. Stoga, kada šuma nasumičnih stabala ukupno proizvodi kraće dužine puta za odredjene tačke velika je verovatnoća da su to anomalije. [9]

ONE-CLASS SUPPORT VECTOR MACHINES (OCSVM). Uopšteno, ideja algoritam je da nauči granicu oko konture raspodele normalnih instanci skupa podataka. Zatim, ako tačke leže unutar graničnog podprostora, smatra se da potiču iz iste populacije. U suprotnom, ako leže van granice, možemo reći da su tačke anomalije. Jednoklasne mašine sa vektorima podrške predstavljaju prilagodjeni algoritam klasične mašine sa vektorima podrške za jednu klasu. [2]

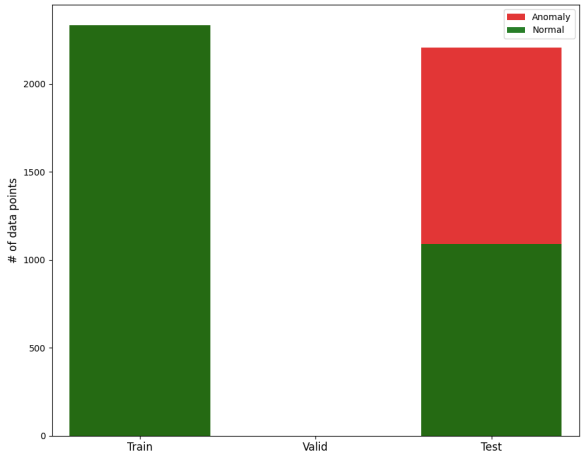
2.3. PRISTUPI

Pored ispitanih pristupa postoje i drugi, tako da lista nije ograničena. Jedan od pristupa koji nije naveden predlaže kreiranje veštačkih anomaličnih oznaka nad skupom normalnih tačaka. Potom se takve instance transformišu kako bi se povećao njihov procenat u skupu podataka. Zatim nad tako dobijenim skupom podataka se mogu koristiti metrike kao i kod nadgledanog učenja. Pored toga, postoji pristup koji eksploatiše samo-nadgledano (eng. *Self-supervised learning*) učenje koji je detaljnije opisan u radu [6]. Treća kategorija bi bila definisanje novih metrika, kao što je IREOS (*Internal, Relative Evaluation of Outlier Solutions*) metrika koja se izračunava bez konzumiranja oznaka i koja je opisana u radu [10]. U okviru ovog eksperimenta su ispitana sledeća tri pristupa.

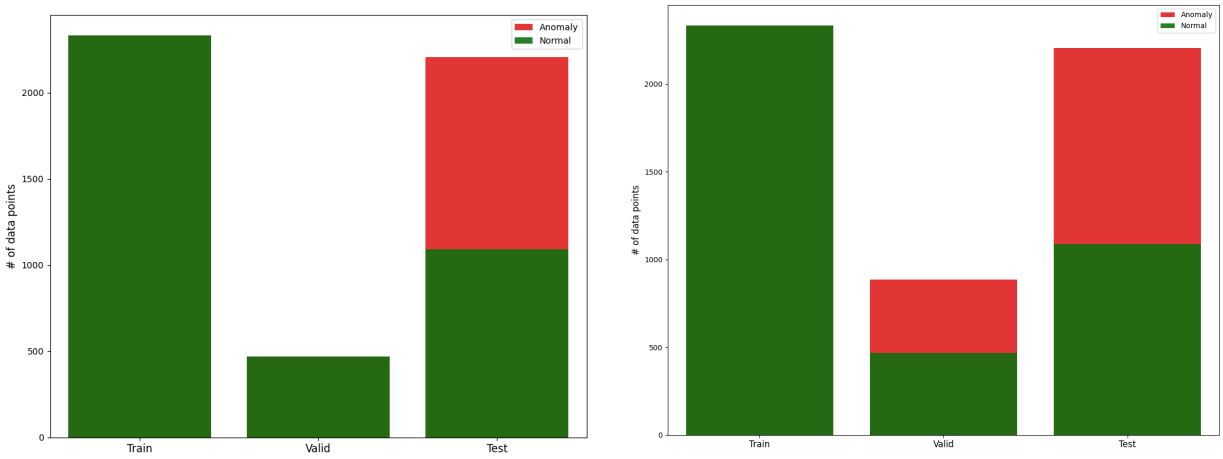
PRISTUP 1: PODRAZUMEVANI. Prvi pristup optimizovanja hiperparametara je konzervativan i zasniva se na korišćenju podrazumevanih vrednosti hiperparametara koje su se generalno dobro pokazale u praksi, naročito u odredjenom domenu. Podrazumevane vrednosti su definisane pomoću Pajton biblioteka *Scikit-learn* i *PyOD*. U ovom slučaju ne postoji validacioni skup podataka kao što se može videti na Grafikon 2.

PRISTUP 2: NENADGLEDANA-VALIDACIJA. Drugi pristup se zasniva na nenadgledanim metrikama, one koje se mogu izračunati bez oznaka i koje su uporedive sa *Receiver Operating Characteristic* (ROC) i *Precision-Recall* (PR) baziranim kriterijumima. Korišćene metrike se baziraju na *Excess-Mass* (EM) i *Mass-Volume* (MV) krivama koje su detaljnije opisane u radu [7]. Rad [7] tvrdi da se EM i MV metrike poklapaju u oko 80% slučaja sa ROC i PR metrikama. Ograničenja ovog pristupa jeste tip atributa, dakle svi atributi moraju biti neprekidni. Pored toga, postoje posebne verzije ovih metrika koje pokušavaju da prevazidju veliki broj dimenzija unutar skupa podataka. Posebne verzije metrika koje se zasnivaju na poduzorkovanju se preporučuju za skupove sa preko 8 atributa. U ovom slučaju, validacioni skup se sastoji samo od normalnih instanci nad kojima se računa EM metrika (Grafikon 3 a). Prilikom optimizacije je korišćena tehnikom nasumične pretrage [1].

PRISTUP 3: NADGLEDANA-VALIDACIJA. Poslednji pristup u ovom radu optimizovanja hiperparametara se bazira na ideji da je bolje kreirati mali broj anomalija i zatim koristiti standardne metrike evaluacije kao što su ROC AUC, PR AUC i druge. Ovaj pristup pretpostavlja da je moguće napraviti mali broj anomalija i da je njihova cena kreiranja vredna poboljšanja performansi modela nenadgledanog učenja. Odakle sledi da u ovom slučaju postoji validacionih skup koji se sastoji od normalnih instanci i anomalija prvog tipa, dok se trening skup sastoji samo od normalnih (Grafikon 3 b). Korišćena je ROC AUC metrika prilikom optimizacije zajedno sa tehnikom nasumične pretrage sa identičnom raspodelom hiperparametara kao u pristupu 2. Detaljniji opis ovog pristupa se može pronaći u radu [13].



Grafikon 2: Podela skup podataka za Pristup 1



a) Pristup 2

b) Pristup 3

Grafikon 3: Podela skupa podataka

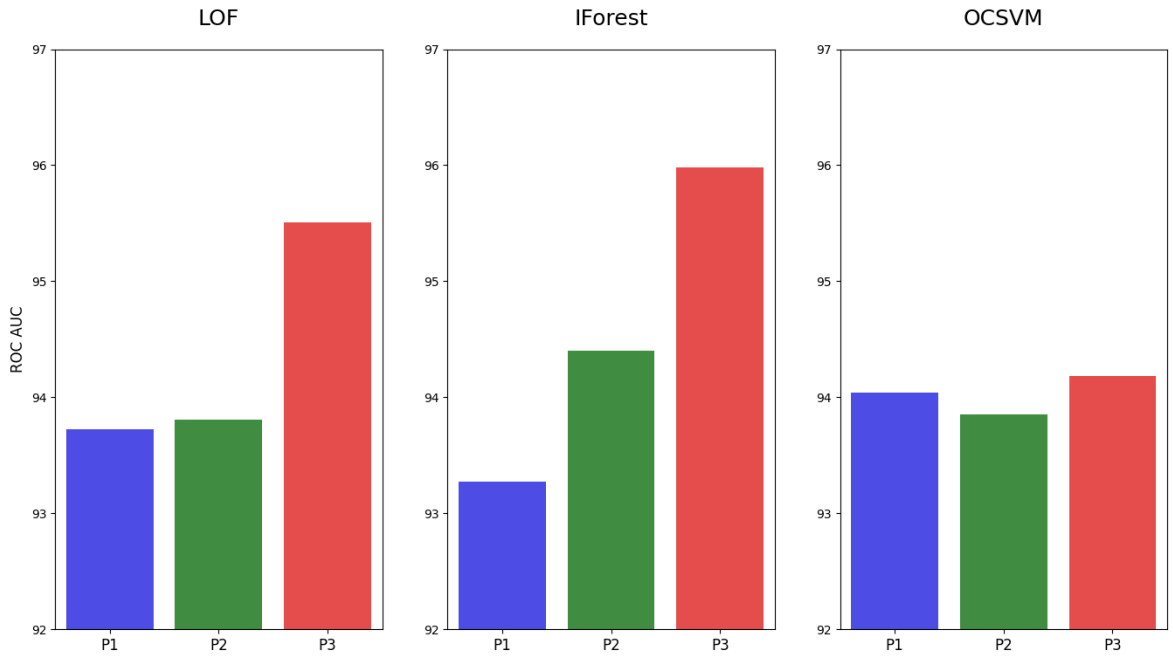
3. REZULTATI

Rezultati eksperimenta su sumirani na slici 4. U okviru prvog grafikona na slici 4 su opisani rezultati LOF algoritma za sva tri pristupa optimizovanja hiperparametara. Na drugom grafikonu se nalaze rezultati IForest algoritma, dok na poslednjem za OCSVM algoritam. Rezultati su mereni metrikama ROC AUC i PR AUC.

Na osnovu empirijskih dokaza, pristup 3 koji koristi anomalije u validacionom skupu je postigao najbolje rezultate kod sva tri algortima prema ROC i PR metrikama na test skupu. Potom, sledi pristup 2 koji optimizuje na validacionom skupu bez oznaka i koji je veoma sličan po performansama u odnosu na pristup 1 koji se bazira na podrazumevanim vrednostima hiperparametara. Kada se promeni perspektiva interpretacije rezultata i posmatraju se pristupi u kontekstu jednog algoritma, može se zaključiti da OCSVM radi izuzetno dobro sa već ispitanim vrednostima hiperparametara. Pored toga, dodatna optimizacija hiperparametara ne pravi veliki učinak. Što se tiče LOF algoritma, pristupi 1 i 2 su postigli slične rezultate, dok je pristup 3 značajno bolji. Pošto algoritam IForest veoma zavisi od tehnike optimizacije hiperparametara, može se zaključiti da algoritmi LOF i IForest značajno poboljšavaju svoje performanse ukoliko se kreiraju anomalije i potom koriste prilikom validacije algoritma.

4. ZAKLJUČAK

U okviru rada je izvršen eksperiment koji se sastoji od tri pristupa za optimizaciju hiperparametara u kontekstu detekcije anomalija. Prvi pristup se zasniva na korišćenju već ispitanih vrednosti hiperparametara i ne zahteva validacioni skup, niti optimizaciju hiperparametara. Intezivno se koristi u literaturi i smatra se pesimističkim. Drugi analiziran pristup se bazira na ideji nenadgledanih metrika. Metrike EM i MV se zasnivaju na ocenjivanju skor funkcije algoritama koristeći tehnike verovatnoće i gustinu raspodele skupa podataka. U okviru ovog pristupa je organizovan validacioni skup koji se sastoji samo od normalnih instanci. Ideja trećeg i poslednjeg pristupa se bazira na kreiranju dodatnih anomalija i korišćenju nadgledanih metrika (ROC i PR) prilikom traženja optimalnih hiperparametara.



Grafikon 4: Poredjenje algoritama za svaki pristup

Na osnovu empirijskih dokaza, pristup 3 koji koristi anomalije u validacionom skupu je postigao najbolje rezultate kod sva tri algortima prema ROC i PR metrikama na test skupu. Zatim se može zaključiti da OCSVM radi izuzetno dobro sa već ispitanim vrednostima hiperparametara. Pored toga, dodatna optimizacija hiperparametara ne pravi veliki učinak. Dok algoritmi LOF i IForest značajno poboljšavaju svoje performanse ukoliko se kreiraju anomalije i potom koriste u fazi validacije.

LITERATURA

(1) James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

(2) Abdenour Bounsiar and Michael G Madden. One-class support vector machines revisited. In *2014 International Conference on Information Science & Applications (ICISA)*, pages 1–4. IEEE, 2014.

(3) Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

(4) Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

(5) K Chen, SC Lu, and HS Teng. Adaptive real-time anomaly detection using inductively generated sequential patterns,". In *Fifth Intrusion Detection Workshop, SRI International, Menlo Park, CA*, 1990.

(6) Jan Diers and Christian Pigorsch. Self-supervised learning for outlier detection. *Stat*, 10(1):e322, 2021.

(7) Nicolas Goix. How to evaluate the quality of unsupervised anomaly detection algorithms? *arXiv preprint arXiv:1607.01152*, 2016.

(8) Markus Goldstein and Seiichi Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one*, 11(4):e0152173, 2016.

(9) Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

(10) Henrique O Marques, Ricardo JGB Campello, Arthur Zimek, and Jörg Sander. On the internal evaluation of unsupervised outlier detection. In *Proceedings of the 27th international conference on scientific and statistical database management*, pages 1–12, 2015.

(11) Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal processing*, 99:215–249, 2014.

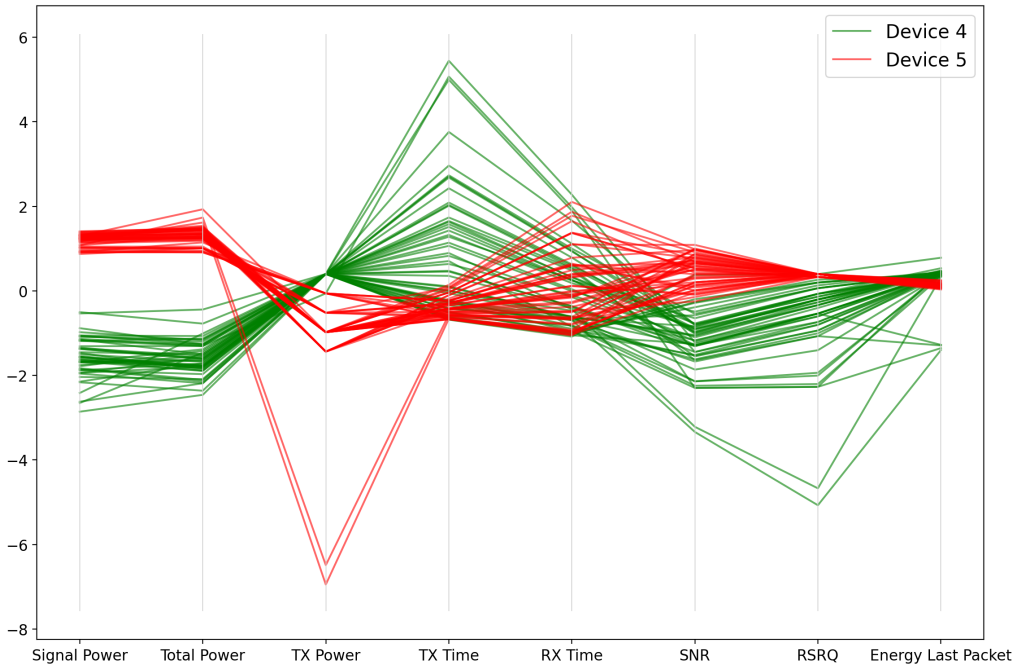
(12) Srdjan Sobot, Vukan Ninkovic, Dejan Vukobratovic, Milan Pavlovic, and Milos Radovanovic. Machine learning methods for device identification using wireless fingerprinting. In *2022 International Balkan Conference on Communications and Networking (BalkanCom)*, pages 183–188. IEEE, 2022.

(13) Jonas Soenen, Elia Van Wolputte, Lorenzo Perini, Vincent Vercruyssen, Wannes Meert, Jesse Davis, and Hendrik Blockeel. The effect of hyperparameter tuning on the comparative evaluation of unsupervised anomaly detection methods. In *Proceedings of the KDD*, volume 21, pages 1–9, 2021.

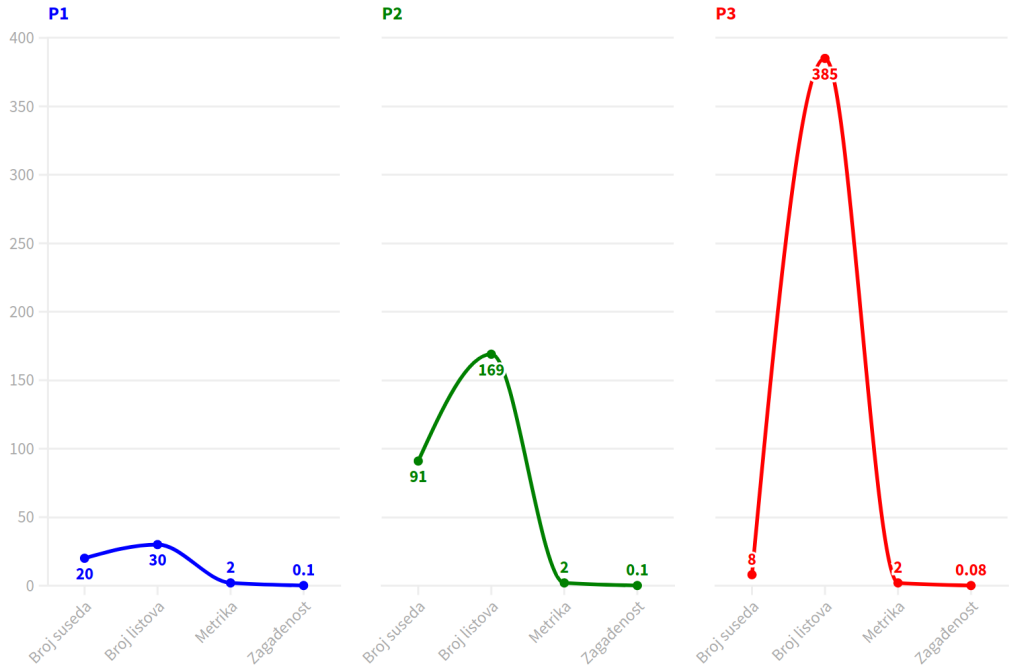
(14) Siqi Wang, Qiang Liu, En Zhu, Fatih Porikli, and Jianping Yin. Hyperparameter selection of one-class support vector machine by self-adaptive data shifting. *Pattern Recognition*, 74:198–211, 2018.

(15) Zekun Xu, Deovrat Kakde, and Arin Chaudhuri. Automatic hyperparameter tuning method for local outlier factor, with applications to anomaly detection. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4201–4207. IEEE, 2019.

(16) Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.



Grafikon 5: Grafikon paralelnih koordinata standardizovanih atributa. Svaka vertikalna linija predstavlja odgovarajući atribut, dok svaki bežični otisak je iscrtan horizontalno preko svih atributa. Uredjaji 4 i 5 su predstavljeni zelenom i crvenom bojom, sa po 50 otisaka.



Grafikon 6: Hiperparametri LOF algoritma