

UNIVERZITET CRNE GORE
PRIRODNO-MATEMATIČKI FAKULTET

Vijeću Prirodno-matematičkog fakulteta:

PREDMET: Ocjena o podobnosti teze master rada i kandidata Aleksandra Plamenca, specijaliste računarskih nauka

Na sjednici Vijeća Prirodno-matematičkog fakulteta Univerziteta Crne Gore, imenovani smo u komisiju za ocjenu podobnosti magistarske teze pod radnim nazivom: "Novelty detekcija primjenom hibridnih algoritama nad tekstualnim podacima", kandidata Aleksandra Plamenca, specijaliste računarskih nauka. Uvidom u priloženu dokumentaciju, komisija Vijeću podnosi sljedeći

IZVJEŠTAJ

Biografija kandidata:

Aleksandar Plamenac je rođen u Baru, 09.02.1993. godine. U Baru završava osnovnu školu, a zatim i gimnaziju "Niko Rolović". Školovanje nastavlja na Univerzitetu Crne Gore, na Prirodno-matematičkom fakultetu, studijski program Računarske nauke. Na ovom fakultetu stiče Bachelor diplomu, a odmah nakon nje i diplomu specijaliste računarskih nauka. Master studije upisuje 2016. godine. Polaže sve ispite predviđene planom i programom master studija na studijskom programu Računarske nauke. Time je stekao mogućnost izrade master rada.

Predmet istraživanja:

Problem *novelty* detekcije jeste problem klasifikacije i detektovanja entiteta (objekta, instance, događaja, signala ...), koji su novi i razlikuju se u odnosu na one podatke koji su dostupni pri treniranju nekog algoritma mašinskog učenja. U ovom istraživanju je predmet detekcija *novelty* instanci nad tekstualnim podacima.

Izazov u ovom problemu jeste to što *novelty* instance su rijetke, i ne predstavljaju normalno ponašanje nekog sistema, pa stoga ih nije moguće lako opisati pri treniranju algoritma mašinskog učenja, već je potrebno da u samom algoritmu postoji mehanizam koji će detektovati ovakvu instancu.

U ovom radu se izučava dati problem kako bi se uočilo da li hibridni algoritmi daju bolje rezultate u odnosu na tradicionalne algoritme koji se koriste za *novelty* detekciju. Pod hibridnim algoritmima se podrazumijeva algoritam koji inicijalno nije namijenjen za problem *novelty* detekcije, a dodatna osobina ovih algoritama jeste da predstavljaju kombinaciju više već postojećih algoritama i tehnika, uz određene modifikacije.

Motiv i cilj magistarskog rada:

Digitalizacijom podataka, dostupna je i ogromna količina podataka u obliku teksta. Ovo čini ekstrakciju informacija iz teksta vrlo aktuelnom temom. Motiv istraživanja upravo jeste i aktuelnost primjene raznih tehnika i algoritama mašinskog učenja nad ovim tipom podataka. Pored toga, primjena tehnika *novelty* detekcije je veoma široka, a neke od situacija u kojima se ove tehnike mogu primijeniti jesu sljedeće:

- Detekcija upada - detekcija bilo kakvih malicioznih napada na sistem, bilo da dolaze sa mreže, bilo da dolaze internim putem, preko virusa
- Detekcija prevare - detekcija neke vrste kriminalne aktivnosti, a posebno kod finansijskih organizacija ili osiguravajućih kuća. Vrlo su česti pokušaji prevare, stoga je i njihova detekcija od velike važnosti
- Primjena u medicini - Velika većina izvještaja, analiza, predstavlja tekstualni podatak. Njihovom analizom, moguće je doći do značajnih informacija o pacijentu
- Druge vrste anomalija na slikama, ili u tekstu - Postojanjem društvenih mreža, i njihovim sve većim korišćenjem, moguće je obraditi veliku količinu podataka koji predstavljaju sliku ili tekst. U ovom radu, biće i prikazana primjena *novelty* detekcije nad komentarima sa jednog lokalnog portala
- *Novelty* detekcija u edukativne svrhe

Samim tim, ovaj magistarski rad ima za cilj da dodatno doprinese razumijevanju problema *novelty* detekcije, sa akcentom na one tehnike koje se dominantno koriste nad tekstualnim podacima. U radu se prikazuju dva tipa algoritama: tradicionalni i hibridni, i rad ima za cilj da da odgovor na to da li hibridni algoritmi daju bolje performanse u odnosu na tradicionalne algoritme.

Pored toga, cilj je i definicija sistematičnog postupka, odnosno procedure, za konstrukciju hibridnog algoritma za *novelty* detekciju u domenu tekstualnih podataka polazeći od poznatih tehnika za klasterizaciju, klasifikaciju ili neki drugi problem mašinskog učenja.

Metode:

U radu će biti osmišljen eksperimentalni protokol, kao i sprovedena serija eksperimenata. Eksperimenti će podrazumijevati propuštanje skupova podataka kroz nekoliko algoritama. Ovi algoritmi se mogu podijeliti u dvije grupe, a to su:

- Statistički algoritmi
- Algoritmi klasterizacije i klasifikacije

Kod prve grupe algoritama se formira statistički model na osnovu podataka iz trening skupa, a zatim se posmatraju rezultati test primjera. Kod grupe algoritama za klasterizaciju i klasifikaciju će uz odgovarajuće modifikacije biti konstruisani hibridni algoritmi za koje će se takođe posmatrati performanse.

Principi po kojima će se posmatrati algoritmi jesu sljedeći:

- Princip robustnosti i razmjene (*principle of robustness and trade-off*) - algoritam treba da maksimizuje isključenje novih primjera, a da minimizuje isključenje poznatih primjera
- Princip uniformnog skaliranja podataka (*principle of uniform data scaling*) – Potrebno je da su svi test i trening podaci uniformno prikazani, tj. „leže“ u istom opsegu
- Princip minimizacije parametara (*principle of parameter minimization*) – Broj parametara koje unosi korisnik kako bi trenirao algoritam treba biti što manji
- Princip generalizacije (*principle of generalization*) – Sistem bi trebao biti sposoban za generalizaciju
- Princip nezavisnosti (*principle of independence*) – Podaci nad kojima se primjenjuje tehnika mašinskog učenja su uvijek opisani preko jedne, ili više karakteristika, kao što je recimo tačka u 2D modelu opisana sa dva realna broja, x i y koordinatom. Sistem ne bi trebao da zavisi od broja karakteristika kroz koje se posmatraju podaci, kao i od broja i prirode klasa koje postoje u podacima
- Princip prilagodljivosti (*principle of adaptability*) – Sistem koji prepoznaje nova posmatranja tokom testa bi trebao iskoristiti ove testove i za ponovno treniranje
- Princip složenosti (*principle of computational complexity*) – S obzirom da mnogi sistemi za novelty detekciju su online, odnosno da konstantno dobijaju nove informacije i

podatke sa Interneta, ili lokalne mreže, i rade u realnom vremenu, jako je bitan segment vremenska složenost samog algoritma

Važan segment posmatranja algoritama biće i vremenska složenost, odnosno vrijeme koje je potrebno algoritmu za trening, kao i za proces testiranja. Posmatraće se i vrijeme koje je algoritmu potrebno i u produkciji, tj. pri radu. Dodatni parametri koji će se koristiti pri procjeni performansi su *precision* i *recall* parametri, gdje se posmatraju odnosi lažno pozitivnih i lažno negativnih posmatranja u odnosu na skup podataka. Lažno pozitivno posmatranje predstavlja svako posmatranje koje algoritam mašinskog učenja klasifikuje kao "pozitivno" (u ovom slučaju, klasifikuje kao *novelty*), ali podatak to zaista i nije. Analogno, lažno negativno posmatranje je svako posmatranje koje algoritam klasifikuje kao "negativno", ali i u ovoj klasifikaciji, algoritam napravi grešku. Implementacija će biti urađena u programskom jeziku *python*, a generalna podjela skupa podataka će biti 80% za trening, a preostalih 20% za test algoritma.

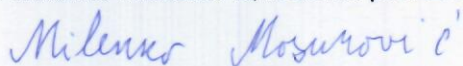
Zaključni stav i predlog:

Komisija pozitivno ocjenjuje podobnost teme i kandidata i predlaže Vijeću Prirodno-matematičkog fakulteta da odobri dalju izradu magistarskog rada "*Novelty* detekcija primjenom hibridnih algoritama nad tekstualnim podacima", kandidata Aleksandra Plamenca, specijaliste računarskih nauka. Komisija predlaže dr Sava Tomovića, redovnog profesora Prirodno-matematičkog fakulteta kao mentora pri izradi magistarske teze kandidata.

U Podgorici, 07.09.2022. godine

Članovi komisije:

dr Milenko Mosurović, redovni profesor



dr Aleksandar Popović, vanredni profesor



dr Savo Tomović, redovni profesor

