Nikola Kušlaković

Predviđanje veza u društvenim mrežama korišćenjem atributa čvorova i topoloških metrika sličnosti

U ovom radu istraženo je kojim podacima je najbolje trenirati klasifikator koji će da predviđa prijateljstva u društvenim mrežama. Upoređene su tačnosti klasifikatorâ koji su trenirani pomoću podataka o korisnicima (atributi čvorova) i pomoću metrika sličnosti (topologija mreže). Nad manjom mrežom boljim su se pokazale metrike sličnosti, jer klasifikatori trenirani njima daju prosečnu tačnost od 75.6%, dok klasifikatori trenirani atributima čvorova imaju prosečnu tačnost od 60.9%. Nad mrežom sa znatno više čvorova i veza situacija je obrnuta: klasifikatori trenirani metrikama sličnosti daju tačnost od 81.5%, a klasifikatori trenirani atributima čvorova 87.2%. Iz toga zaključujemo da je za treniranje klasifikatora nad velikim mrežama bolje koristiti atribute čvorova nego metrike sličnosti. Urađena je i analiza atributa čvorova i metrika sličnosti na osnovu koje se može videti koji su atributi čvorova i koje metrike sličnosti najbolji za predviđanje.

Uvod

Društvena mreža je matematički model socijalnih interakcija. Svaki korisnik društvene mreže pretstavlja jedan čvor, a svako prijateljstvo između korisnika predstavlja vezu ili granu grafa. U zavisnosti od vrste društvene mreže, graf može biti usmeren ili neusmeren, a veze između čvorova mogu biti okarakterisane određenim brojnim vrednostima (težinama). Za čvorove možemo da kažemo da su susedni (prijatelji) kada između njih postoji veza. Svaki čvor i veza mogu da poseduju određene atribute koji ih karakterišu (npr. hobi korisnika, mesto rođenja, obrazovanje). U samom grafu postoji mnoštvo ostvarenih veza, ali mnogo više neostvarenih, a koje bi se mogle realizovati tokom nekog vremenskog perioda. Skoro svaki servis socijalnog umrežavanja poseduje mogućnost da korisniku predlaže ljude koje možda zna, i koje bi mogao da doda kao prijatalje. Analizom ovih potencijalnih veza u društvenim mrežama može se predvideti tok razvijanja mreže i učvrstiti veza između korisnika. Na osnovu evolucije mreže mogu se dobiti informacije o tome da li će i kada određeni korisnici postati prijatelji.

U nekim ranijim radovima (Al Hasan *et al*. 2006; Julian i Lu 2016) poredi se efikasnost klasifikatora kada se on trenira različitim tipovima atributa čvorova, ali ti rezultati se ne porede sa rezultatima koji se dobijaju kada se klasifikator trenira metrikama sličnosti, odnosno kada se za treniranje koristi samo topologija mreže. Takođe, u referentnim radovima (Liben-Nowell i Kleinberg 2004; Gao *et al*. 2015), uopšte nije korišćeno mašinsko učenje, već se metrike sličnosti koriste zasebno bez međusobnog kombinovanja. U većini radova se koriste i drugi tipovi društvenih mreža: mreža koautorstva naučnih radova, društvene mreže gde prijateljstva nisu obostrana i društvene mreže gde su prijateljstva okarakterisana težinima.

Cilj ovog rada je uporedna analiza rezultata dobijenih korišćenjem atributa čvorova mreže i metrika sličnosti, u kojima su prijateljstva obostrana i nisu okarakterisana nikakvim težinama. Atributi čvorova, kao i metrike sličnosti, su dodatno analizirani, kako bi se videlo koje grupe at-

Nikola Kušlaković (2002), Novi Sad, učenik 2. razreda Gimnazije „Isidora Sekulić" u Novom Sadu

MENTORI:

Miloš Savić, Prirodno-matematički fakultet Univerziteta u Novom Sadu

Nikola Bebić, student Prirodno-matematičkog fakulteta Univerziteta u Novom Sadu

ributa i koje metrike imaju najveći uticaj u predviđanju prijateljstava. Korišćena su tri klasifikaciona modela: logistička regresija, algoritam *k* najbližih komšija i metod šume odlučivanja.

Materijal i metode

Opis korišćenog skupa podataka

Skup podataka koji je korišćen u ovom istraživanju predstavlja mali deo društvene mreže Facebook (Leskovec i Krevl 2014), i sastoji se od 10 grafova. Svaki graf je formiran tako što je uzet jedan ego čvor (čvor sa mnogo prijatelja) i potom zapisane sve veze njegovih prijatelja, a ego čvor obrisan. Svaki čvor unutar ovih 10 grafova poseduje određene atribute, koji su zapisani preko niza nula i jedinica (binarni atributi). Nula označava da čvor ne poseduje neki atribut, a jedinica da ga poseduje. Grafovi imaju različit broj i tip atributa koje čvorovi mogu da poseduju (npr. u jednom grafu čvor maksimalno može da poseduje samo 70 atributa, dok u drugom može da ih ima i do 500).

Skup podataka je u potpunosti anonimizovan, tako da nije moguće povezati podatke sa ljudima na koje se odnose. Svaki atribut u ovom skupu je okarakterisan jedinstvenom vrednošću, na osnovu koje je moguće međusobno razlikovati atribute. Jedino što može da se pročita iz skupa podataka jeste tip atributa (npr. znamo da se atribut odnosi na rodni grad osobe, ali ne znamo da li je taj rodni grad Novi Sad, Beograd ili neki drugi grad). Osim onsovnih atributa o osobi (ime, prezime, datum rođenja i pol), imamo i atribute koji se odnose na zanimanje i obrazovanje osobe. Objašnjenja za atribute čvorova koji se odnose na obrazovanje su sledeća: godina obrazovanja predstavlja razred koji osoba trenutno pohađa, obrazovni tip nam govori da li osoba pohađa osnovnu školu, srednju školu ili fakultet, stepen obrazovanja predstavlja vrstu diplome koju osoba poseduje, a obrazovno usmerenje smer koji pohađa u srednjoj školi ili na fakultetu. Što se tiče atributa vezanih za posao imamo: pozicija koju osoba ima na svom poslu, lokacija gde je osoba odlazi na posao i datumi kada je osoba počela i kada završava sa radom tog posla (odlazi u penziju). U skupu podataka se pojavljuju i atributi koji se odnose na rodni grad osobe, trenutnu lokaciju stanovanja osobe, deo države (region) iz koje je osoba i broj jezika koji osoba zna.

Atribute je takođe moguće grupisati, i posmatrati kao celinu (npr. gledamo sve atribute koji označavaju zanimanje kao jednu celinu). Dosta čvorova ima samo po jedan atribut iz svake grupe (npr. čvor A može da bude samo iz jednog grada od mogućih šest), dok neki imaju više atributa iz iste grupe (npr. čvor B može da zna više jezika). Takođe, bitno je navesti da je moguće spojiti svih 10 grafova u jedan veliki graf, čime se dobija mreža od 4039 čvorova i 88234 veza, gde svaki čvor (korisnik) poseduje preko 1200 atributa.

Metrike sličnosti

Da bismo uopšte predviđali veze, moramo nekako da ih rangiramo. Za to se koriste odgovarajuće metrike koje tim vezama daju određenu vrednost na osnovu kojih se mogu rangirati. Ima ih više i svaka od njih ima svoje prednosti i mane. Svaka od ovih metrika matematički pokušava da modeluje društvene odnose i načine na koje ljudi ostvaruju prijateljstva. Ove metrike se najčešće koriste samo kada analiziramo topologiju mreže, odnosno kada ne obraćamo pažnju na atribute čvorova i veza, već samo na povezanost čvorova i broj čvorova i veza. U ovom radu korišćeno je sledećih pet metrika:

1. Broj zajedničkih komšija (number of common neighbors) – vrednost koja predstavlja broj zajedničkih prijatelja (suseda) između dva čvora. U stvarnom svetu ljudi koji se ne poznaju mogu se upoznati preko zajedničkih prijatelja.

2. Žakarov koeficijent (Jaccard 1912) – količnik broja zajedničkih prijatelja dva čvora i broja elemenata skupa koji je unija njihovih prijatelja.

3. Preferencijalno povezivanje (preferential attachment) – proizvod ukupnog broja prijatelja za dva zadata čvora. Ljudi koji imaju puno prijatelja, bez problema će ostvariti nova prijateljstva.

4. Adamič-Adar indeks (Adamic i Adar 2003) – koristi se najčešće u cikličnim mrežama gde je izražena pojava prostih ciklusa (graf u kojem može da se krene od izabranog čvora i da se opet dođe do njega u tri koraka). Računa se po

formuli $f(x, y) = \sum_{v \in N(x,y)} \frac{1}{\log n(v)}$ – gde su x i y čvorovi čuju vezu ispitujemo, a $N(x, y)$ skup zajedničkih prijatelja čvorova x i y. Čvor v je jedan od čvorova iz skupa $N(x, y)$, a $n(v)$ je ukupan broj prijatelja čvora v. Situacija u realnom svetu koju ova metrika pokušava da modeluje bi bila sledeća: ako imamo osobe A i B koje su jako dobri prijatelji sa istom osobom C i osoba C ima jako malo prijatelja, velika je verovatnoća da će tad osoba C upoznati međusobno te dve osobe na nekom okupljanju, jer su joj to jedini prijatelji. Što osoba C ima više prijatelja, statistički je manji njen značaj kao zajedničkog prijatelja, odnosno manja je verovatnoća da će međusobno upoznati baš osobe A i B.

5. Alokacija resursa (resource allocation) – najčešće se koristi za optimizaciju računarskih sistema, ali može da se koristi i u cikličnim mrežama. Slična je Adamič-Adar metrici, samo što formula glasi $f(x, y) = \sum_{v \in N(x,y)} \frac{1}{n(v)}$.

Klasifikacioni modeli

Za treniranje klasifikatora korišćena su tri klasifikaciona modela: logistička regresija, algoritam k najbližih komšija i metod šume odlučivanja. Logistička regresija koristi sigmoid funkciju koju prilagođava u odnosu na podatke kada trenira klasifikator. Algoritam k najbližih komšija mapira podatke u n-dimenzinalnom prostoru, a potom za svaki novi podatak koji treba da klasifikuje računa rastojanje do k najbližih komšija (tačaka) na osnovu kojih tom podatku dodeljuje klasu. Metod šume odlučivanja se zasniva na generisanju određenog broja nasumičnih stabala odluke, gde svako stablo vrši klasifikaciju, i onda se klasa koja ima najviše glasova od strane stabala, dodeljuje podatku. Sva ova tri klasifikaciona modela treniraju binarni klasifikator. Binarni klasifikator treba da klasifikuje veze čvorova u dve grupe, u grupu veza za koje se predviđa da će se ostvariti (pozitivna klasa) i grupu veza za koju se predviđa da se neće ostvariti (negativna klasa). Vrednosti koje se uzimaju u obzir kada želimo da vidimo koliko je klasifikator dobar su tačnost, preciznost i odziv. Tačnost, kao najbitnija, predstavlja odnos svih tačno klasifikovanih elemenata i broja svih elemenata skupa. Preciznost je odnos tačno klasifikovanih elemenata klase i svih elemenata koji su bili klasifikovani kao ta klasa, a odziv predstavlja odnos tačno klasifikovanih elemenata klase i ukupnog broja elemenata te klase.

Za trening klasifikatora korišćena je unakrsna validacija. Unakrsna validacija na osnovu m parametra deli skup podataka na m delova. Prvi deo podeljenog skupa se koristi za testiranje, a ostatak za treniranje, potom drugi deo za testiranje, a ostatak za treniranje, itd. Dobija se m modela kojima je moguće odrediti tačnost, preciznost ili neku drugu meru. Korišćenjem unakrsne validacije svi podaci iz skupa koriste se i za testiranje i za treniranje. Konkretno, u ovom istraživanju korišćena je petostruka unakrsna validacija, odnosno skup podataka je svaki put bio podeljen na pet delova. Za svaki taj model izmereni su tačnost, preciznost i odziv. Preciznost i odziv odnose se na pozitivnu klasu, odnosno služe da se vidi koliko dobro klasifikator klasifikuje ostvarene veze.

Bitno je napomenuti da su za algoritam k najbližih komšija i šume odlučivanja rađena podešavanja hiperparametara, odnosno treniran je veliki broj klasifikatora različitim grupama parametara, i na osnovu toga zaključeno koje kombinacije parametara daju najbolje rezultate. Za algoritam k najbližih komšija ispitano je koja je vrednost parametra k (broj komšija) najbolja, a za šume odlučivanja ispitano je 6 parametara: broj stabala koji se generiše, maksimalna dužina stabla, maksimalan broj podataka potreban da bi se podelio unutrašnji čvor, minimalan broj podataka koji se nalaze u čvoru pre nego što je čvor podeljen, minimalan broj podataka dozvoljen u krajnjem čvoru i parametar koji određuje da li će se podaci više puta koristiti za generisanje stabala (bootstrap). Za svaki ovaj parametar kreirana je lista mogućih vrednosti i onda su ispitane različite kombinacije parametara.

Obrada grafova

Kako bi se dobili podaci koji će se koristiti za dalji rad, pre treniranja klasifikatora bilo je potrebno obraditi sve grafove. Prvo su kreirani podaci pomoću atributa čvorova: atributi su međusobno redom poređeni, i za svako poklapanje atributa u niz je upisivana vrednost 1, a u

suprotnom slučaju 0 (dva atributa se poklapaju ako oba čvora poseduju isti atribut). Svaki niz jedinica i nula odgovara vezi između čvorova čiji su atributi upoređeni. Ovaj postupak je urađen kako između čvorova koji imaju ostvarenu, tako i čvorova koji nemaju ostvarenu međusobnu vezu. Za neostvarene međusobne veze korišćeni su samo oni čvorovi koji imaju barem jednog zajedničkog prijatelja, odnosno čvorovi koji su relativno blizu, a nemaju sklopljeno prijateljstvo.

Sledeći korak bio je kreiranje podataka pomoću metrika sličnosti. Za rangiranje veza pomoću metrika sličnosti korišćena je samo topologija mreže, odnosno ignorisani su atributi čvorova. Kao i kod poređenja atributa, ovaj postupak je urađen i za čvorove koji imaju i za čvorove koji nemaju ostvarenu vezu (neostvarena veza je isto uzimana između čvorova koji imaju barem jednog zajedničkog prijatelja). Ove dve vrste veza zapravo predstavljaju klase u koje će binarni klasifikator svrstavati nove veze, i na taj način „zaključivati" da li dva korisnika treba da sklope prijateljstvo ili ne. Oba ova postupka kreiranja podataka, i pomoću atributa čvorova i pomoću metrika sličnosti, urađena su i na velikom grafu koji je kombinacija svih 10 manjih mreža. Nakon ovoga, za svaki skup podataka treniran je i testiran binarni klasifikator.

Za implementaciju ovog projekta korišćen je programski jezik Python 3.7 koji sadrži sve potrebne biblioteke za rad sa grafovima i upravljanje različitim tipovima podataka i fajlova.

Rezultati i diskusija

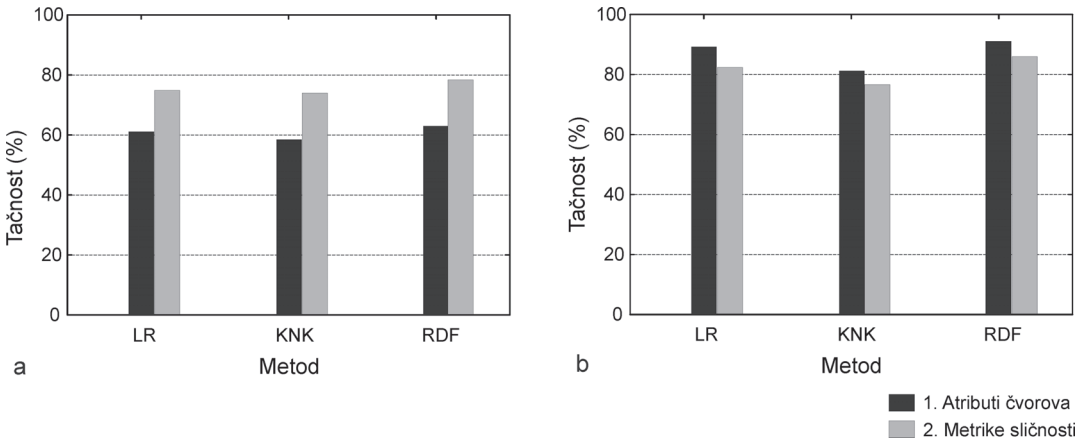
Prvi graf nad kojim su trenirani klasifikatori bio je jedan od 10 manjih grafova. Čine ga 347 čvorova i 5037 veza, pri čemu svaki čvor poseduje do 224 atributa. Na ovom grafu je pokazano kakvi se rezultati dobijaju kada se klasifikatori treniraju nad manjim mrežama, kada nema puno čvorova i veza. Vrednosti za tačnost klasifikatora, kao i vrednosti odziva i preciznosti prikazane su tabeli 1. Prosečna tačnost klasifikatora nad validacionim skupom kod šume odlučivanja

Tabela 1. Tačnost, preciznost i odziv za različite klasifikacione modele trenirane nad manjim grafom

	Atributi čvorova			Metrike sličnosti		
Klasifikacioni model	Tačnost (%)	Preciznost (%)	Odziv (%)	Tačnost (%)	Preciznost (%)	Odziv (%)
Logistička regresija	61.22	61.17	61.52	74.78	78.26	78.26
K najbližih komšija	58.46	59.71	52.20	73.84	76.49	69.04
Šuma odlučivanja	62.92	62.87	63.20	78.32	80.31	75.24

Tabela 2. Tačnost, preciznost i odziv za različite klasifikacione modele trenirane nad velikim grafom

	Atributi čvorova			Metrike sličnosti		
Klasifikacioni model	Tačnost (%)	Preciznost (%)	Odziv (%)	Tačnost (%)	Preciznost (%)	Odziv (%)
Logistička regresija	89.22	88.50	90.16	82.14	86.96	75.64
K najbližih komšija	81.20	76.36	90.40	76.58	81.46	68.84
Šuma odlučivanja	91.08	88.41	94.56	85.66	86.51	84.52



Slika 1. Uporedni prikaz tačnosti različitih klasifikacionih modela treniranih nad manjim grafom (a) i većim grtafom (b). Oznake: LR – logistička regresija, KNK – *k* najbližih komšija, RDF – slučajna šuma odlučivanja.

Figure 1. Comparison of accuracies for different classification models trained on the smaller graph (a) and the larger graph (b). Tags: LR – logistic regression, KNK – *k* nearest neighbors, RDF – random forest; 1 – node attributes, 2 – similarity metrics.

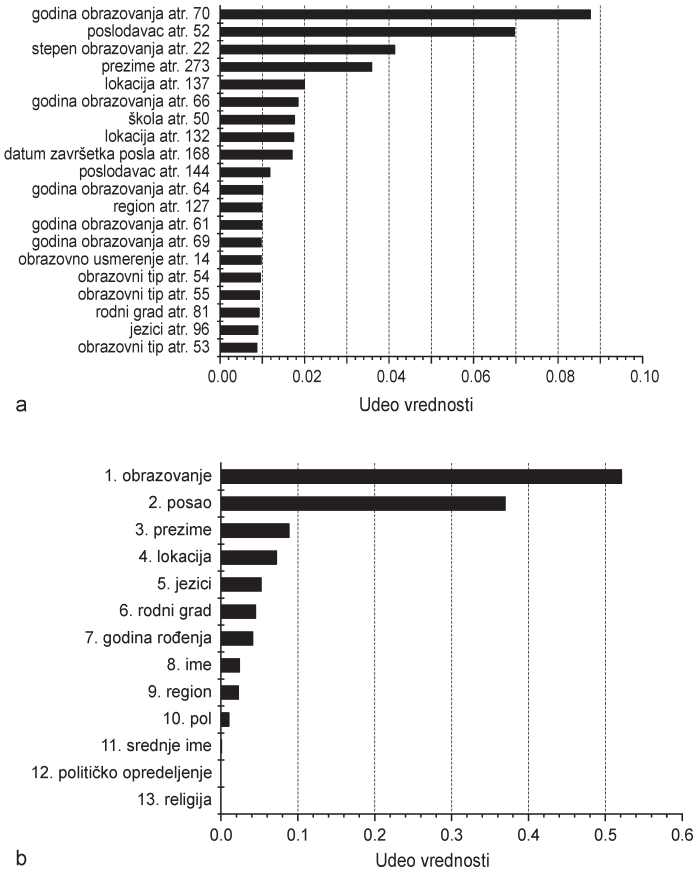
iznosila je 61.5% za atribute čvorova i 77.5% za metrike sličnosti. Svi ovi rezultati su dobijeni treniranjem nad malim uzorkom ovog grafa, konkretno korišćen je balansirani skup od 5000 veza – 2500 primeraka pozitivne i 2500 primeraka negativne klase.

Drugi graf nad kojim su trenirani klasifikatori bio je veliki graf, dobijen kombinacijom pomenutih 10 manjih grafova. Na ovom grafu pokazano je kakvi bi se rezultati dobili treniranjem klasifikatora nad velikim mrežama, odnosno mrežama koje imaju veliki broj čvorova i veliki broj veza. Korišćena mreža sadrži 4039 čvorova i 88234 veza, pri čemu svaki čvor poseduje do 1200 atributa. Dobijene vrednosti za tačnost klasifikatora, kao i vrednosti preciznosti i odziva, prikazane su u tabeli 2. Prosečna tačnost klasifikatora nad validacionim skupom kod šume odlučivanja iznosila je 86.6% za atribute čvorova i 85.3% za metrike sličnosti. Kao i kod prvog testiranja, nad manjim grafom korišćen je mali uzorak velikog grafa, odnosno balansirani skup podataka od 5000 veza.

Iz tabele 2 vidimo da za veći graf, korišćenjem atributa čvorova i metrika sličnosti, algoritam šume odlučivanja ima tačnost preko 90% i 85%. Zbog toga su za ovaj model urađena dodatna merenja kako bi se sagledalo koji su atributi čvorova, odnosno koje metrike sličnosti, najkorisniji za klasifikator. Na slici 2a rangirano je 20 najkvalitetnijih atributa čvorova (kada se gledaju nezavisno jedni od drugih), a na slici 3 svih pet metrika sličnosti. Na slici 2a je uz svaki atribut upisana i njegova jedinstvena vrednost iz skupa podataka. Na slici 2b rangirane su grupe atributa čvorova. Ovo rangiranje je urađeno na osnovu sumarnog udela svih atributa koji čine grupu.

Vidimo da u slučaju manje mreže treniranje metrikama sličnosti za sva tri klasifikaciona modela daje bolje rezultate u poređenju sa atributima čvorova (tabela 1 i slika 1). Ovo se moglo i pretpostaviti, zbog činjenice da su atributi koje čvorovi ove mreže poseduju po prirodi loši, a nema ih dovoljno da bi klasifikator mogao da nađe neku pravilnost među njima. Metrike sličnosti se ne oslanjaju na informacije o čvorovima, već samo na topologiju mreže. Tačnost nije velika, jer su klasifikatori trenirani nad malom mrežom, gde ima malo čvorova i veza, tako da ne može da se uoči jasna razlika između ostvarenih i neostvarenih veza.

S druge strane, rezultati dobijeni treniranjem nad većom mrežom (tabela 2 i slika 1) su znatno bolji od onih koji su dobijeni treniranjem nad

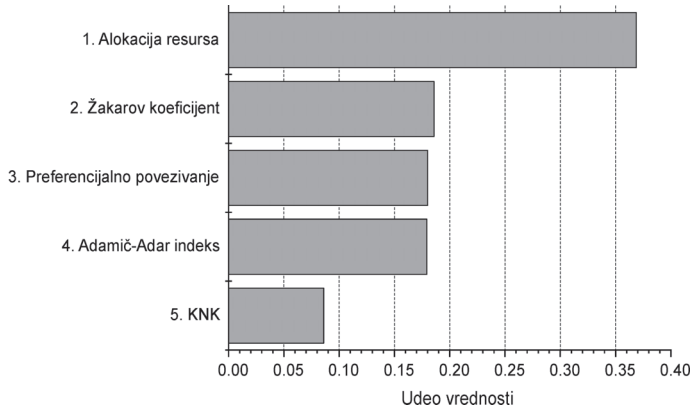


Slika 2. a) Uporedni prikaz najvažnijih atributa čvorova za model šume odlučivanja;
b) Najvažnije grupe atributa

Figure 2. a) Comparison of the importance of node attributes for the random forest model
b) Comparison of the importance of node attributes by groups for the random forest model:
1 – Education
2 – Occupation
3 – Last name
4 – Location
5 – Languages
6 – Hometown
7 – Birth year
8 – First name
9 – Locale
10 – Gender
11 – Middle name
12 – Political affiliation
13 – Religion

manjom mrežom. U ovom slučaju situacija je obrnuta – atributi čvorova daju bolje rezultate nego metrike sličnosti. Na osnovu ovih rezultata možemo da zaključimo da je mnogo bolje kada se koristi više atributa čvorova za treniranje klasifikatora i kad su ti atributi čvorova različite vrste (npr. atributi koji se odnose na obrazovanje, posao ili mesto stanovanja korisnika).

Što se tiče međusobnog rangiranja atributa čvorova (slika 2), vidimo da su najbolji oni atributi koji se odnose na obrazovanje i posao korisinka. Ovo ima smisla, jer će ljudi sa istim



Slika 3. Uporedni prikaz metrika sličnosti za model šume odlučivanja

Figure 3. Comparison of the importance of similarity metrics for the random forest model:
1. Resource allocation
2. Jaccard index
3. Preferential attachment
4. Adamic-Adar index
5. Common neighbors

stepenom obrazovanja (osnovna škola, srednja škola ili fakultet) najverovatnije lakše ostvariti prijateljstvo – srednjoškolci postaju prijatelji na mreži sa srednjoškolcima, a studenti sa studentima. Za posao važi slična priča, kolege iz istih kompanija i sa istim ili sličnim poslovnim opredeljenjima često se međusobno povezuju. Kada su u pitanju metrike sličnosti, vidimo da je metrika alokacije resursa najbolja, i to je verovatno posledica velike cikličnosti društvene mreže. Broj zajedničkih komšija ima najmanji udeo u poređenju sa ostalim metrikama previđanja veza, što znači da nije dovoljno da samo gledamo zajedničke prijatelje dva korisnika kada hoćemo da predvidimo da li će oni postati prijatelji.

Bitno je napomenuti da treniranje klasifikatora atributima čvorova traje znatno duže nego treniranje metrikama sličnosti. Ovo se dešava zato što je skup podataka sa atibutima čvorova znatno veće dimenzije (ima više parametara), dok skup podataka sa metrikama sličnosti uvek ima samo 5 parametara za svaku vezu.

Zaključak

Na osnovu dobijenih rezultata vidimo da su najbolji oni klasifikatori koji su trenirani nad velikom mrežom, uz korišćenje atributa čvorova, odnosno klasifikatori koji su trenirani podacima o samim korisnicima (npr. stepen obrazovanja, informacije o zanimanju ili mesto odakle korisnik potiče). Kod manje mreže bolje su se pokazali klasifikatori trenirani metrikama sličnosti nego atributima čvorova, ali njihova tačnost ne prelazi 80%.

Istraživanje je pokazalo da su najkorisniji oni atributi čvorova koji se odnose na posao i obrazovanje korisnika. To znači da su informacije o poslu kojim se korisnik bavi i njegovom obrazovanju veoma bitne, da bismo mu uspešno preporučili prijatelje na društvenoj mreži.

Kada su u pitanju metrike sličnosti, alokacija resursa se pokazala kao najbolja, dok metrika zajedničkih prijatelja ima najmanji značaj. Dakle, nije dovoljno koristiti samo broj zajedničkih prijatelja kada želimo da predvidimo prijateljstvo između dve osobe. Najbolje je kombinovati metrike sličnosti međusobno, jer ni jednu metriku nije dovoljno koristiti samu za sebe.

Od sva tri klasifikaciona modela, model šume odlučivanja se pokazao kao najbolji, što nam govori da je njega najbolje koristiti kada želimo da rešimo probleme ovakve prirode. Konkretno, model šume odlučivanja, treniran nad većom mrežom uz pomoć atributa čvorova, imao je tačnost od 91%.

Literatura

Adamic L. A., Adar E. 2003. Friends and neighbors on the Web. *Social Networks*, **25** (3): 211.

Al Hasan M., Chaoji V., Salem S., Zaki M. 2006. Link prediction using supervised learning. U *Proceedings of SDM 06: workshop on link analysis, counter-terrorism and security*. SIAM, str. 798-805.

Gao F., Musical K., Cooper C., Tsoka S. 2015. Link prediction methods and their accuracy for different social networks and network metrics. *Scientific Programming*, 2015: 1-13.

Jaccard P. 1912. The distribution of the flora in the alpine zone. *New phytologist*, **11** (2): 39.

Julian K., Lu W. 2016. Application of machine learning to link prediction. http://cs229.stanford.edu

Leskovec J., Krevl A. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data

Liben-Nowell D., Kleinberg J. 2004. The link prediction problem for social networks. https://www.cs.cornell.edu/home/kleinber/link-pred.pdf

Nikola Kušlaković

Prediction of Links in Social Networks by Using Node Attributes and Topological Similarity Metrics

This paper explores what data is best to train a classifier that predicts friendships in social networks. The accuracies of the classifiers that were trained by user data (node attributes) and similarity metrics (network topology) were compared. Similarity metrics were better for smaller networks, because classifiers trained by similarity metrics gave an average accuracy of 75.6% and

classifiers trained by node attributes an accuracy of 60.9%. For larger social networks, the situation is reversed, as classifiers trained by similarity metrics give an average accuracy of 81.5%, and classifiers trained by node attributes an accuracy of 87.2%. From this, we can conclude that it is better to use information about users rather than similarity metrics to train classifiers for large social networks.

Node attributes and similarity metrics were also analyzed to see which attributes and which metrics are the best in prediction. Attributes that are related to user education and occupation are the best for predicting friendships. As for the similarity metrics, the resource allocation metric is the most important metric in predictions that use similarity metrics.