PRIMENA ALGORITMA C4.5.1 U DRUŠTVENIM MREŽAMA
APPLICATION OF C4.5.1 ALGORITHM IN SOCIAL NETWORKS

Bojan Jelača

Elektrotehnički fakultet

Univerzitet u Beogradu

Sadržaj –*U ovom radu prikazano je kako se poznati data mining algoritam C4.5 može modifikovati radi poboljšanja. Zatim je pokazano i kako se predstavljena modifikacija može primeniti u društvenih mrežama, da bi se otkrilo da li je neka osoba u vezi ili ne.*

Abstract *– This document shows how to modify well-known data mining algorithm C4.5 in order to improve it. After that, it is shown how the given modification can be applied in social networks in order to predict if a person is in a relationship or not.*

1. Uvod

U današnjem svetu, društvene mreže su od izuzetnog značaja. One drastično utiču na društveni status i položaj osobe kao individue, kao i na međusobne odnose između osoba. Takođe, društvene mreže pokazuju izuzetan uticaj na popularnost i zastupljenost određenih brendova, kompanija, organizacija itd.

Prva ideja o društvenim mrežama pojavila se još 1890-ih, a razvili su je Emil Durkajm i Ferdinand Teni u svojim teorijama koje su se bavile istraživanjem društvenih grupa. Društvene mreže kakve danas poznajemo javile su se 90-ih godina XX veka, dok se većina danas popularnih društvenih mreža pojavila u prvoj deceniji trećeg milenijuma.

Nezavisno sa razvojem društvenih mreža, razvijale su se i tehnike data mining-a. Prva ideja o ovim tehnikama javila se 1960-ih godina, da bi pojam „data mining“ bio uveden 90-ih godina. Od tada se razvilo mnogo algoritama koji se mogu primeniti na velike skupove podataka da bi se došlo do naizgled nepoznatih, tzv. skrivenih podataka.

Pokazalo se da su društvene mreže pogodan teren za primenu algoritama data mining-a. Razvijeno je mnogo teorija i sproveden veliki broj eksperimenata koji ispituju kako bi se tehnike data mining-a mogle primeniti na podatke koji se mogu pronaći u društvenim mrežama. Takođe, same društvene mreže koriste brojne tehnike data mining-a da bi otkrile skrivene veze između svojih korisnika i na taj način unapredile same sebe.

2. Kratak opis algoritma C4.5 i uvođenje algoritma C4.5.1

Algoritam C4.5 je algoritam koji na osnovu trening skupa podataka generiše stablo odlučivanja. Generisano stablo se može primenjivati da bi se otkrio jedan nepoznati parametar novog podatka. Na Slika 1 Primer trening skupa podataka prikazan je primer trening skupa koji se može koristiti.

Customer	Savings	Assets	Income ($1000s)	Credit Risk
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	High	Medium	25	Bad
4	Medium	Medium	50	Good
5	Low	Medium	100	Good
6	High	High	25	Good
7	Low	Low	25	Bad
8	Medium	Medium	75	Good

Slika 1 Primer trening skupa podataka

U nastavku teksta je ukratko opisan algoritam C4.5.

Svaki parametar podatka (jedna kolona u tabeli) se posmatra kao kandidat za razdvajanje skupa. Nazvaćemo pomatrani parametar S. Neka parametar S deli početni skup podataka T u nekoliko podskupova T_1, T_2, ..., T_k. Redukcija entropije skupa parametra S se računa po sledećoj formuli:

$$ER(S) = H(T) - Hs(T) \quad (1)$$

H(T) predstavlja entropiju skupa T i računa se po formuli:

$$H(X) = -\sum p_j \log_2(p_j) \quad (2)$$

U datoj formuli p_j predstavlja procenat podataka u skupu T koji imaju jedan od mogućih ishoda, gde j uzima sve vrednosti od 1 do N, gde je N broj broj mogućih ishoda.

Hs(T) je ukupna entropija koja se dobija kada se skup T podeli na podskupove. Računa se po formuli:

$$Hs(T) = -\sum_{i=1}^{k} P_i \, Hs(T_i) \quad (3)$$

P_i je procenat podataka u skupu T_i. k predstavlja ukupan broj podskupova na koji parametar S deli skup T.

2.1. Problem sa algoritmom C4.5

Može se primetiti da se u jednačini (1) koristi logaritam za osnovu 2. Ovo je dobro ukoliko je broj mogućih ishoda parametra koji želimo da otkrijemo u novom podatku jednak 2. Ipak, ukoliko je broj mogućih ishoda veći, nastaje problem. Naime, ukoliko parametarkandidat S kreira takve podskupove da jedan podskup ima podjednak broj podataka sa svakim ishodom, očekivano je da entropija takvog podskupa bude 1. Međutim, ukoliko se koristi logaritam za osnovu 2, to neće biti slučaj.

Na primer, ukoliko podskup T_i sadrži 3 podatka, od čega $podatak_1$ ima $ishod_1$, $podatak_2$ ima $ishod_2$, a $podatak_3$ ima $ishod_3$, ovaj podskup ne daje nikakvu informaciju o mogućem ishodu njegovih elemenata, pa je očekivano da entropija bude 1. Međutim, ukoliko se koristi logaritam za osnovu 2, to neće biti slučaj.

2.2. Uvođenje algoritma C4.5.1

Rešenje problema navedenog u poglavlju Problem sa algoritmom C4.5 je da se umesto logaritma za osnovu 2, koristi logaritam za osnovu N, gde je N ukupan broj mogućih ishoda.

Na ovaj način bi se u gore pomenutim skupu T_i dobila entropija 1, ukoliko bi se koristio algoritam za osnovu 3 pri računanju entropije.

Ovo je upravo srž algoritma C4.5.1 i njegovo poboljšanje u odnosu na algoritam C4.5.

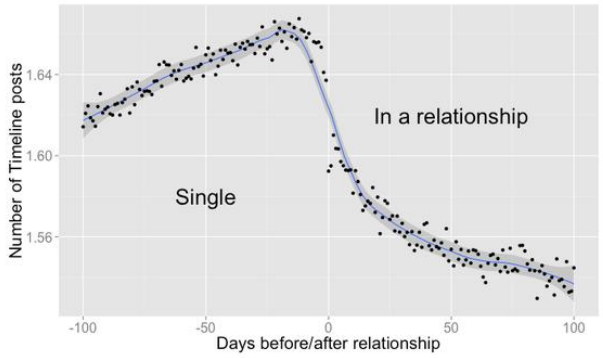
3. Korišćenje algoritma C4.5.1 da bi se predvidelo da li je neko u vezi

U ovom poglavlju opisan je algoritam, koji koristi C4.5.1 algoritam kao data mining tehniku, kako bi predvideo da li je neko u vezi ili ne, a kasnije i sa kojom osobom je u vezi.

3.1. Predviđanje da li je neko vezi (prvi prolaz algoritma C4.5.1)

Da bi se predvidelo da je neko u vezi, treba pratiti njegove aktivnosti na društvenoj mreži. Glavno pitanje koje se nameće jeste na osnovu čega se može zaključiti, prateći nečiji profil, da li je osoba u vezi. Konkretnije, koje parametre pratiti i ispitivati.

Statistika je pokazala da jedan parametar koji može indicirati da je osoba u vezi jeste broj objava te osobe na društvenoj mreži. Naime, kao što je pokazano u [1], broj objava osobe na društvenoj mreži blago ili je blizak konstantnoj vrednosti, a zatim naglo opadne nakon stupanja u vezu, što je i prikazano grafikom na narednoj Slika 2.



Slika 2 Broj objava osobe pre i posle stupanja u vezu

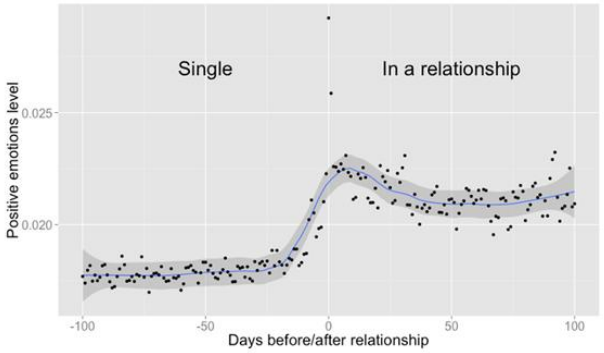
Ostaje još pitanje kako izraziti broj objava. Za to se koristi sledeća formula:

$$AP(t) = NP(t)/TON(t) \quad (4)$$

U navedenoj formuli, NP(t) je broj objava do trenutka t, dok je TON(t) ukupno vreme provedeno na mreži do trenutka t. Radi lakšeg prikazivanja rezultata, uzimamo da AP(t) može imati samo 3 vrednosti, i to LOW

(manjeod 0.25), MEDIUM (između 0.25 i 0.75) I HIGH (preko 0.75).

Drugi parametar koji se može koristiti kao indikator jeste nivo pozitivnih emocija u objavama. Naime, kao što se može i videti na grafiku na Slika 3, nivo pozitivnih emocija drastično poraste od trenutka stupanja u vezu.

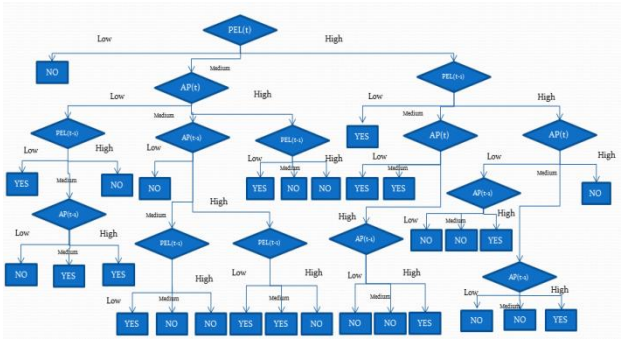


Slika 3 Nivo pozitvnih emocija pre i posle stupanja u vezu

Izmeriti nivo pozitivnih emocija je jako teško za mašinu, i u tu svrhu se moraju koristiti neke metode konceptualnog modelinga koje su objašnjene u [2]. Ponovo, radi lakšeg predstavljanja rezultata, uzeto je da nivo pozitivnih emocija (PEL) može imati vrednosti LOW, MEDIUM, HIGH.

Oba navedena parametra treba razmatrati barem u 2 vremenska trenutka, da bi se moglo zaključiti da li krive imaju željeni oblik.

Stablo koje je algoritam C4.5.1 generisao na osnovu trening skupa podataka prikazano je na narednoj Slika 4.



Slika 4 Stablo za odlučivanje da li je osoba u vezi

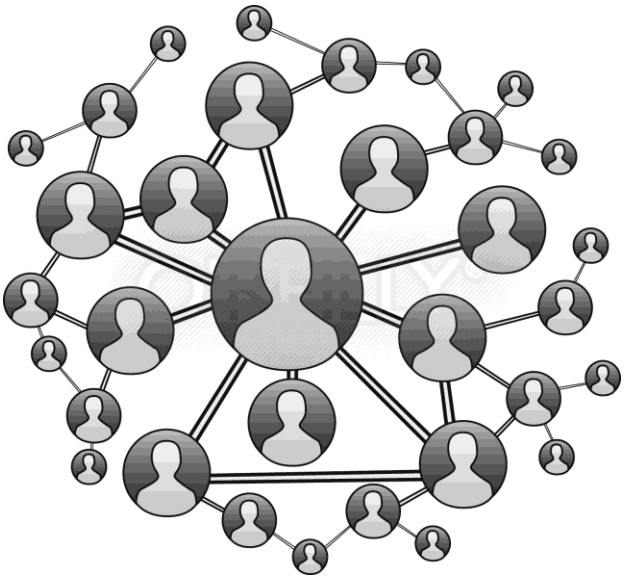
Može se videti da je dominantan parametar nivo pozitivnih emocija u sadašnjem trenutku i da je on najbolji pokazatelj da li je osoba u vezi. Takođe, kao što je i očekivano, može se videti da je približno isti broj listova koji pokazuju da je osoba u vezi i broj listova koji pokazuju suprotno.

Nakon treniranja C4.5.1 algoritma na predefinisanom trening skupu podataka, može se bilo koja osoba podvrgnuti algoritmu i sa velikom tačnošću proveriti da li je osoba u vezi ili ne.

3.2. Pronalaženje kandidata

Sledeće pitanje koje se nameće jeste kako pronaći osobe koje bi mogle da budu u vezi sa osobom na koju je primenjen prvi prolaz algoritma C4.5.1. Naravno, za očekivati je da to bude osobo koja je prijatelj sa posmatranom osobom.

Na Slika 5 prikazan je način kako su organizovani korisnici društvenih mreža.



Slika 5 Graf korisnika određene društvene mreže

Čvorovi grafa predstavljaju korisnike, dok su grane grafa relacije prijateljstva između korisnika. Ukoliko je relacija prijateljstva na društvenoj mreži obavezno obostrana, kao što je slučaj sa Fejsbukom, na primer, onda su grane dvostmerne, kao što je i slučaj na slici iznad. Sa druge strane, ako je relacija prijateljstva u formi praćenja, kao što je slučaj sa Tviterom, onda bi grane bile jednosmerne.

Kako izabrati kandidate? Od svih prijatelja posmatrane osobe, treba izdvojiti sve one osobe čiji pol odgovara interesovanjima posmatrane osobe, kao i čijim interesovanjima odgovara pol posmatrane osobe. Takođe, kandidati takođe treba da budu u vezi. Ovde je potrebno i

na kandidate primeniti prvi prolaz algoritma C4.5.1, što može oduzeti malo vremena, ali ipak značajno skraćuje listu kandidata. Na kraju, kandidati treba da ispunjavaju uslov da količina međusobnih aktivnosti sa posmatranom osobom bude veća od 0.5.

Količina međusobnih aktivnosti između osobe A i osobe B se može izračunati na sledeći način:

$$AMA(A,B) = \frac{Act\,(A,B)+Act\,(B,A)}{Act\,(A)+Act\,(B)} \quad (5)$$

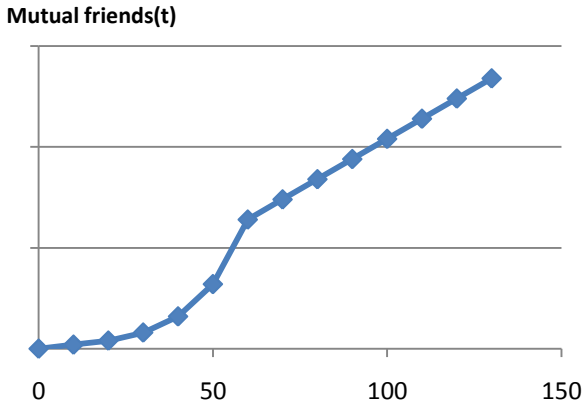
U datoj formuli, funkcija *Act(X,Y)*predstavlja količinu aktivnosti osobe X prema osobi Y, dok Act(X) presdstavlja ukupnu aktivnost osobe X. Pod aktivnošću se podrazumevaju objave, komentari, slike, poruke i sl.

3.3. Provera kandidata (drugi prolaz algoritma C4.5.1)

Nakon izdvajanja kandidata iz skupa prijatelja, ostalo je još da se, iz skupa izdvojenih kandidata, izdvoji osoba koja je u vezi sa posmatranom osobom.

U ovu svrhu se ponovo koristi algoritam C4.5.1. Pre toga, treba utvrditi koji parametri mogu pokazati da su dve osove u vezi.

Statistički podaci pokazuju da broj zajedničkih prijatelja osoba koje su u vezi ima oblik kao na grafiku sa naredne Slika 6.



Slika 6 Zavisnost broja zajedničkih prijatelja od vremena

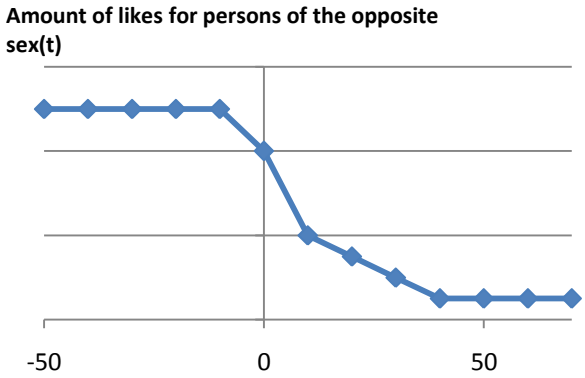
Kao što se može videti, broj zajedničkih prijatelja raste eksponencijalno na početku veze, a zatim linearno.

Broj zajedničkih prijatelja bi mogao da bude jedan parametar za utvrđivanje da li su dve osobe u vezi. On se može izraziti sledećom formulom:

$$MFG(A,B,t) = |Friends(A,t) \cap Friends(B,t)| \quad (6)$$

U navedenoj formuli, funkcija *Friends(X,t)*predstavlja skup svih prijatelja osobe X u trenutku t. Radi lakšeg predstavljanja rezultata, uzeto je da parametar MFG može imati vrednosti EXPONENTIAL, LINEAR i OTHER, gde OTHER predstavlja sve druge oblike.

Drugi parametar koji je od značaja jeste broj sviđanja (popularnih lajkova) prema osobama suprotnog pola. Naime, nakon stupanja u vezu, osobe pokazuju tendenciju smanjivanja broja sviđanja prema osobama suprotnog pola, što se može videti na Slika 7.



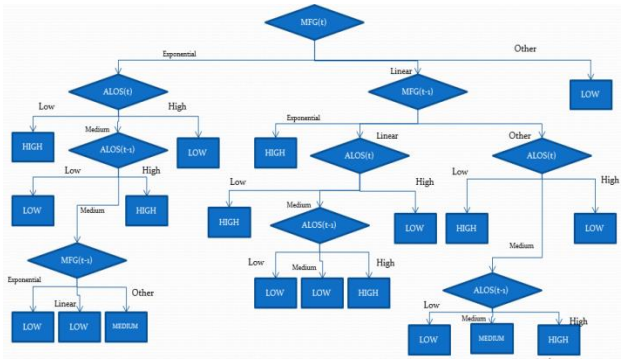
Slika 7 Zavisnost broja sviđanja prema osobama suprotnog pola od vremena

Količina sviđanja prema osobama suprotnog pola može se izraziti na sledeći način:

$$ALOS(A,B,t) = \frac{NLOS(A,t)}{NL(A,t)} + \frac{NLOS(B,t)}{NL(B,t)} \quad (7)$$

U navedenoj formuli, funkcija *NLOS(X,t)* daje broj sviđanja osobe X prema osobama suprotnog pola u trenutku t, dok funkcija *NL(X,t)* daje ukupan broj sviđanja osobe X u trenutku t. Treba napomenuti da se, pod terminom „osoba suprotnog pola“, misli na osobe čiji pol odgovara interesovanjima osobe na koju se formula primenjuje. I za ovaj parametar se uzimaju vrednosti LOW (ispod 0.25), MEDIUM (između 0.25 i 0.75) i HIGH (preko 0.75), radi lakšeg predstavljanja rezultata.

Nakon treniranja algoritma C4.5.1 na predefinisanom skupu podataka, generisano je stablo odlučivanja prikazano na Slika 8.



Slika 8 Stablo za odlučivanje verovatnoće da su dve osobe u vezi

Može se primetiti da je najdominantniji parametar rast zajedničkih prijatelja u sadašnjem trenutku. Takođe, treba primetiti da postoje tri moguća ishoda, LOW, MEDIUM i HIGH, što implicira da je pri generisanju stabla odlučivanja, za računanje entropije korišćen logaritam za osnovu 3. Kao što se i očekivalo, broj listova za svaki mogući ishod je približno isti.

Nakon treniranja algoritma C4.5.1, generisano stablo odlučivanje se može primeniti na posmatranu osobu i na svakog od kandidata, da bi se utvrdilo sa kim je posmatrana osoba u vezi.

4. Primena opisanog algoritma

Jedan vid primene opisanog algoritma bi mogao da bude u komercijalne svrhe. Naime, u zavisnosti od toga da li je osoba u vezi ili ne, razlikuju se načini pristupa toj osobi prilikom reklamiranja nekog proizvoda, brenda, organizacije i sl.

Drugi vid primene jeste od strane samih društvenih mreža. Svakoj društvenoj mreži je bitno da ima što više informacija o svojim korisnicima kako bi se što više prilagodila svakom pojedinačnom korisniku, a samim tim unapredila sebe, što bi povećalo broj njenih korisnika.

Još jedan od vidova primene ovog algoritma jeste prilikom procenjivanja osobe sa kojom se razmatra saradnja. Na primer, činjenica da li je osoba u vezi ili ne može biti značajna prilikom zapošljavanja te osobe, jer se iz te činjenice mogu doneti zaključci o lojalnosti, posvećenosti i sličnim osobinama.

5. Zaključak

U ovom radu pokazano je kako se algoritam C4.5 može modifikovati radi poboljšanja preciznosti, pri čemu se ne gubi na složenosti samog algoritma. Zbog toga bi predstavljeni algoritam C4.5.1 mogao da se koristi tamo gde se koristi C4.5. U nekim situacijama bi pokazao veću preciznost i generisao manje stablo nego C4.5.

Opisana primena algoritma C4.5.1 u društvenim mrežama daje veoma dobre i precizne rezultate. Jedina veća mana opisanog algoritma je njegova sporost pri prikupljanju podataka. Da bi se to izbeglo, moguće je izbaciti neke manje bitne delove, ali bi se onda i preciznost algoritma neznatno smanjila.

Problematika obrađena ovom temom je u ekspanziji, pošto broj društvenih mreža, a samim tim i njihovih korisnika, svakodnevno raste i svedoci smo velikih promena u društvenim odnosima u poslednjih 10-ak godina. Zbog toga ovaj algoritam, ili neka njegova unapređena verzija, može biti od velikog značaja u budućnosti.

Literatura

[1] Carlos Greg Diuk „The Formation of Love“, Facebook Data Science, 2014.

[2] Grega Jakus, Veljko Milutinović, Sanida Omerović, Sašo Tomazić „Concepts, Ontologies and Knowledge Representation“, 2014.