D2.1

Konstrukcija značajki mrežnih podataka

Uvod

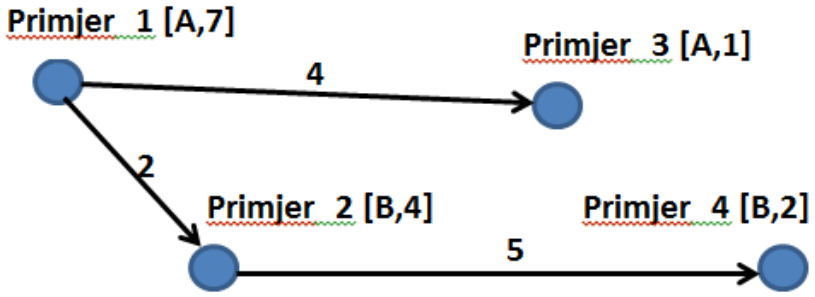
Tehnike strojnog učenja primjenjuju su na skupu primjera koji je opisan skupom atributa. Vrijednosti atributa mogu biti numeričke ili kategoričke, te po potrebi mogu imati i nepoznate vrijednosti. Bez obzira kako izgleda originalni problem, za primjenu tehnika strojnog učenja potrebno je originalni problem transformirati u skup primjera koji su opisani skupom atributa (značajki). Kvaliteta rezultata dobivenih strojnim učenjem u znatnoj mjeri ovisi o kvaliteti preslikavanja originalnog problema u značajke koji su ulazni podaci za postupke strojnog učenja.

Postoji više složenih formi ulaznih podataka kao što su vremenske i prostorne serije, relacijske baze i mrežni podaci. U tim oblicima zajedničko je da osim atributa koji su pridruženi svakom primjeru i koji se direktno mogu iskoristiti kao ulazni podaci za strojno učenje, postoje i informacije koje je potrebno transformirati u značajke. U vremenskim i prostornim serijama dodatne informacije se odnose na višestruko pojavljivanje istih objekata te je iz broja ponavljanja, njihove međusobne udaljenosti ili uzorka ponavljanja potrebno konstruirati značajke koje karakteriziraju seriju. U relacijskim bazama dodatne se informaciju odnose na činjenicu da su osnovni objekti koji su primjeri u postupku strojnog učenja povezani sa drugim objektima koji imaju svoje dodatne karakteristike. Problem konstruiranja značajki sastoji u tome da se karakteristike drugih objekata pravilno pridruže osnovnim objektima te na taj način obogate sadržaj koji se koristi u postupku otkrivanja znanja. U mrežnim podacima imamo slučaj da su objekti (primjeri) međusobno povezani raznim vezama. U domeni ekonomije objekti (primjeri) mogu biti države koje su definirane atributima kao što su njihova površina, broj stanovnika, postotak visokoobrazovanih, postotak obradive površine i slično a da su neke od tih država i dodatno međusobno politički ili prometno povezane, odnosno da postoje informacije o fizičkoj udaljenosti, viznom režimu, trgovini i broju turističkih putovanja.

Predmet ovog izvještaja je prikaz postupaka za konstrukciju značajki koji proizlaze iz međusobne veze primjera. Mreže i grafovi najznačajnije su strukture u kojima postoje veze između primjera a zadatak realiziranog sustava je općenitiji i odnosi se na konstrukciju značajki za bilo koji primjenu strojnog učenja gdje postoje veze između primjera. Realizirani sustav predstavlja konstrukciju značajki u relacijskim bazama ali samo u onima u kojima su relacije ograničene na isti skup objekata (primjera). Jedan od klasičnih relacijskih problema je otkrivanje koncepta „majka" u skupu primjera kojeg čine osobe i u

kojem postoji usmjerena veza roditelj-dijete između primjera. U tom slučaju koncept „majka" zadovoljavaju osobe (primjeri) koje su povezani relacijom roditelj-dijete sa barem jednim drugim primjerom kao početna točka veze a koje dodatno imaju svojstvo da su osobe ženskog spola. Zadatak sustava je otkrivanje takvih relacija ali i svih onih gdje interakcije između primjera predstavljaju bitnu karakteristiku primjera.

Primijenjeni pristup konstrukciji značajki u skladu je sa postojećim sustavima strojnog učenja i implementiran je kao nadogradnja sustava u kojem postoje samo atributi koji karakteriziraju pojedine primjere. Na Slici 1 prikazana ja mreža sa 4 čvora kojeg čine primjeri 1-4. Svaki čvor ima dva atributa od kojih je prvi kategorički a drugi numerički. Primjeri su povezani usmjerenom vezom koja ima numeričke vrijednosti.



Slika 1 Primjer mreže sa 4 čvora i usmjerenim vezama koje imaju vrijednost.

U Tablici 1 su prikazani podaci vezani na ovu mrežu. Svaki primjer prikazan je u svom retku. U lijevom dijelu tablice su podaci koji su vezani uz svaki čvor (atributi 1-2).

Tabela 1 Podaci za mrežu sa slike 1 kako ih unosi korisnik.

	Atribut 1	Atribut 2	Veza_1 Usmjerena sa vrijednošću, tip A
Primjer 1	A	7	2:2;2:4
Primjer 2	B	1	4:5
Primjer 3	A	4	
Primjer 4	B	2	

Novost je dodatak u obliku stupca koji je nazvan Veza_1. Sadržaj stupca su liste koje definiraju sa kojim primjerima je povezan primjer koji se nalazi u tom stupcu. Pošto neki primjer može biti povezan sa više drugih primjera (u prikazanoj mreži to je primjer 1 koji je povezan sa primjerima 2 i 4) sadržaj stupca mora biti lista koja sadrži popis svih primjera sa kojima on povezan. Lista može biti prazna ako veze ne postoji. Vrijednosti pridružene vezama specificiraju

se uz svaku vezu. Prvi broj u listi prije dvotočke specificira primjer prema kojem postoji veza a drugi broj iza dvotočke specificira vrijednost veze. U definiciji veze korisnik određuje da li je veza usmjerena ili neusmjerena.

Format zapisa prikazan u tablici je općenit jer se može koristiti i kada postoje samo podaci vezani na primjere, kada postoje samo podaci o vezama, te kada postoje i jedan i drugi tip podataka. Ako istovremeno postoji više tipova veza među primjerima takve veze se definiraju kao dodatni stupci čiji sadržaj su liste veza. Tada kažemo da je mreža višeslojna.

Očekivane primjene razvijene metodologije su u području ekonomije, socijalnih mreža, transporta, bioloških mreža i analize relacija.

U postupku razvoja ustanovljeno je postojanje četiri osnovnih tipova mreža (nepovezane, slabo povezan, jako povezane, potpuno povezane) te četiri vrste vrijednosti veza (A-D) koje imaju bitan utjecaja na način konstrukcije značajki. Iako su tipovi mreža i veza definirani općenito, tek korištenje sustava treba pokazati eventualno potrebu za uvođenjem i dodatnih relevantnih podskupina. U procesu razvoja ukazala se mogućnost za većom i manjom detaljnosti pri konstrukciji značajki. U ovoj fazi istraživanja nije moguće definirati univerzalno primjenjiv pristup koji će biti prihvatljiv za sve primjene. Zbog toga je postupak realiziran modularno sa mogućnošću utjecaja korisnika na konačni rezultat te sa mogućnošću nadogradnje sustava. Posebna pažnja je posvećena izgradnji modula koji konstruiraju značajke za osnovne slučajeve a koji se onda mogu iterativno i u kombinacijama koristiti i za složenije tipove veza.

Definicije pojmova

Cilj konstrukcije značajki je karakterizacija primjera sa atributima koji održavaju postojanje veza među primjerima. Konstruirane značajke mogu se koristiti za prediktivno razlikovanje primjera, za otkrivanje novog znanja o primjerima ili za grupiranje (klasteriranje) primjera. Kada se povezanost primjera interpretira kao mreža tada su primjeri čvorovi a veze između primjera su grane ili lukovi u takvoj mreži.

Povezanost primjera (čvorova) može biti usmjerena (napr. uvoz automobila iz države A u državu B) i neusmjerena (napr. postojanje kopnene granice između država). Veze mogu imati pridružene numeričke vrijednosti (napr. broj uvezenih/izvezenih automobila, dužina kopnene granice). U jednoj mreži može postojati višestruka povezanost čvorova. To se smatra višeslojnom mrežom u kojoj su u jednom sloju na primjer veze uvoza automobila a u drugom sloju veze s obzirom na postojanje kopnenih granica. Ako neki tip veze ima kategoričke vrijednosti (na primjer postojanje međudržavnih ugovora koja poprima vrijednost A ako je ugovor vezan za slobodnu trgovinu a vrijednost B ako je to vojni savez) tada se taj tip veze smatra višeslojnom mrežom u čijem jednom sloju su veze postojanje trgovinskih ugovora a u drugom sloju postojanje vojnih ugovora.

Uz svaki čvor mogu ali ne moraju biti pridružene vrijednosti koje karakteriziraju taj čvor. U primjeru kada su države čvorovi u mreži tada te vrijednosti mogu biti broj stanovnika i postotak obradive površine. Ovakve vrijednosti se nazivaju atributima čvorova. One mogu biti numeričke i kategoričke. Primjeri kategoričkih atributa država su naziv kontinenta na kojem se država nalazi i tip političke vlasti u državi koji može biti diktatura, kraljevina ili demokratska republika.

Numeričke vrijednosti atributa mogu imati utjecaj na značenje veza između čvorova. Na primjer, uvoz 1000 automobila godišnje u državu A može biti vrlo značajno ako se radi o državi sa malim brojem stanovnika a uvoz istog broja automobila za državu B sa vrlo velikim brojem stanovnika može biti beznačajno. Zbog toga se vrijednosti veza koriste i u svom normaliziranom obliku, na primjer kao broj uvezenih automobila normaliziran sa brojem stanovnika.

Povezana mreža je ona u kojoj postoji veza svakog čvora sa svim ostalim čvorovima pri čemu veze mogu biti direktne i indirektne preko drugih čvorova. Potpuno povezana mreža je ona u kojoj postoje direktne veze između svih parova čvorova. Jako povezana mreža je ona u kojoj postoji mnogo ali ne sve direktne veze. U ovoj realizaciji jako povezana mreža je ona koja ima više od polovice svih mogućih direktnih veza čvorova. Slabo povezana mreža je povezana mreža koja nije jako povezana. Nepovezana mreža je ona u kojoj postoji barem jedan par čvorova koji nisu povezani.

U slabo povezanim mrežama poseban značaj imaju centralni čvorovi koji osiguravaju povezanost onih čvorova koji nemaju direktnu vezu. U teoriji mreža

postoji više definicija centralnosti čvorova. U strojnom učenju koristimo sve te definicije a pojam centralnog čvora proširujemo tako da uključuje sve čvorove koji se ističu po bilo kojem kriteriju. To može biti centralan položaj s obzirom na bilo koji sloj višeslojne mreže a u primjeru sa državama kao čvorovima to može biti i država sa najvećim brojem stanovnika i država sa najvećom proizvodnjom sirove nafte. Praktično to znači da je u jednoj mreži možemo imati više pa čak i mnogo centralnih čvorova. Za konstrukciju značajki važan podatak je povezanost ili ne povezanost nekog čvora X sa centralnim čvorovima, odnosno udaljenost od čvora X do takvih čvorova.

Za svaki čvor X definiramo prvi i drugi krug čvorova. U prvom krugu su čvorovi sa kojima je X direktno povezan a u drugom krugu su čvorovi sa kojima je X povezan indirektno preko samo jednog čvora. Broj čvorova u prvom krugu se zove i stupanj čvora X.

Čvor X je redundantno povezan sa čvorom Y ako postoje barem dvije veze između X i Y a skupovi čvorova koji ih povezuju su disjunktni.

Tehnički pristup problemu

Forma ulaznih podataka

Za ulaz podataka koji uključuje i mrežnu povezanost između primjera koristimo standardnu formu uobičajenu u strojnom učenju koju proširujemo sa dodatnim tipom atributa. Taj dodatni tip zovemo mrežni atribut a njegov sadržaj je lista.
Za svaki sloj mreže dodajemo po jedan mrežni atribut.

Imena primjera odnosno čvorova su cjelobrojne vrijednosti od 1 na više. Lista koja je vrijednost nekog mrežnog atributa može biti prazna ako čvor nije u tom sloju povezan niti sa jednim drugim čvorom. Ako je veza bez pridružene vrijednosti onda se lista sastoji od točka zarezom odvojene liste cjelobrojnih vrijednosti koje označavaju čvorove sa kojima je čvor u tom retku povezan (na primjer u Tabeli 1 povezanost čvora 1 sa čvorovima 2 i 4 se označava sa listom 2;4 u retku broj 1 za dotični mrežni atribut. Ako je vezi pridružena numerička vrijednost onda lista ima formu parova brojeva odvojenih dvotočkom. U neusmjerenim vezama postojanje veze između čvora 1 prema čvoru 2 implicira i vezu čvora 2 prema 1 pa ju je dovoljno navesti samo jednom bilo u retku za čvor 1 ili u retku za čvor 2.

Formu prikazanu u Tabeli 1 moguće je koristiti kao općenitu formu jer u slučaju kada ne postoji mrežna povezanost primjera ostaje standardni prikaz atributa. Isto tako proširenu formu moguće je koristiti kada ne postoje standardni atributi već samo mrežna povezanost čvorova.

Za svaki mrežni atribut potrebno je specificirati da li se radi o usmjerenoj ili neusmjerenoj vezi, da li postoji pridružena numerička vrijednost, te koji je tip odnosno značenje numeričke vrijednosti. Trenutno razlikujemo 4 tipa vrijednosti veza. Definicije veze su pridružene uz ime mrežnog atributa.

Forma izlaznih podataka

Izlazni podaci su vrijednosti značajki. Za svaki primjer (čvor) određuje se onoliko vrijednosti koliko se konstruira značajki. Vrijednosti značajki su tipično numeričke vrijednosti ali one mogu biti i kategoričke ako vrijednosti čvorova uključuju kategoričke vrijednosti.

Vrijednost značajke ima nepoznatu vrijednost ako neki ulazni podaci imaju nepoznate vrijednosti ili ako iz raspoloživih ulaznih podataka nije moguće izračunati vrijednost značajke ili vrijednost značajke nije definirana za konkretni slučaj.

Neka značajka je beskorisna za karakterizaciju čvorova ako svi čvorovi imaju istu vrijednost za tu značajku. U slučaju postojanja nepoznatih vrijednosti, značajka

je beskorisna ako svi čvorovi za koje je vrijednost definirana imaju identičnu vrijednost. Ako neka značajka ima za sve primjere nepoznatu vrijednost ona je beskorisna. Beskorisne značajke se ne prikazuju u izlaznom skupu podataka.

Iako je uobičajena forma prikaza podataka tako da su primjeri redci a značajke odnosno atributi stupci, u konkretnoj implementaciji mi koristimo obrnutu formu: za svaku konstruiranu značajku u izlaznoj datoteci se generira po jedan redak. Taj redak ima stupaca koliko postoji različitih čvorova (primjera). Razlog za ovaj pristup je što konstrukcija značajki zamišljena modularno. Svaki modul konstruira svoje značajke koje je potrebno zatim integrirati u jednu datoteku i na njima izvesti postupak strojnog učenja. Kombiniranje značajki u ovom pristupu je jednostavno jer se one jednostavno dopisuju u istu datoteku. Jedini ozbiljan nedostatak je što je prije izvođenja strojnog učenja potrebno izvršiti konverziju u standardni oblik. U postupak generiranja velikog skupa pravila ugrađena je mogućnost prihvata podataka i u ovoj transponiranoj formi te konverzija nije potrebna.

Organizacija programa

Program za konstrukciju značajki je organiziran modularno. Svaki modul se poziva preko posebne opcije a po potrebi se dodatnim opcijama uključuju i vrijednosti koje usmjeravaju rad modula. Centralni program koji je zajednički za sve opcije sastoji se od dijela za čitanje ulaznih podataka i generiranje jednog retka u izlaznoj datoteci. Izlazni dio uključuje ispitivanje korisnosti generirane značajke i neispisivanje izlaza ako je značajka beskorisna.

Važno mjesto u programu zauzimaju potprogrami koji obavljaju neke poslove koji su zajedničke za više modula. U takve poslove spada računanje minimalnih veza u mreži te generiranje osnovnog skupa značajki. Računanje vrijednosti nekih složenih značajki svodi se na iterativno računanje osnovnih značajki za neke posebne slučajeve zamišljenih i reduciranih struktura mreža.

Tipovi numeričkih veza

Postoji više tipova veza. Veza može biti usmjerena i neusmjerena. Vezi može biti pridružena vrijednost.

Razumijevanje značenja veze ima bitan utjecaj na postupak generiranja značajki. Razlikujemo četiri tipa veze s tim da se tipovi A-C koriste samo za veze kojima je pridružena vrijednost a tip D za veze koje mogu ali ne moraju imati vrijednost.

Tip A
U ovom tipu udaljenost između čvorova (primjera) je veća ako je pridružena vrijednost vezama veća. Ako između nekih čvorova ne postoji veza ona se može interpretirati i kao veza beskonačne vrijednosti. Primjeri ovakve veze su geografska udaljenost i cijena prijevoza između država. Udaljenost dvaju indirektno povezanih čvorova jednaka je *sumi vrijednosti veza* na putu koji povezuje krajnje čvorove. Ako postoji više alternativnih veza onda se traži *minimalna* vrijednost.

Tip B
U ovom tipu, kao i u tipovima C i D, čvorovi su bolje povezani ako je vrijednost veze veće. Ako između nekih čvorova ne postoji veza ona se može interpretirati i kao veza vrijednosti jednake 0. Primjer ovakve veze je broj avionskih letova koji povezuju dva grada u nekom periodu vremena. U ovaj tip spadaju i socijalne mreže za koje je definiran intenzitet komunikacije između pojedinaca. Udaljenost dvaju indirektno povezanih čvorova jednaka je *najmanjoj vrijednosti veza* na putu koji povezuje krajnje čvorove. Ako postoji više potpuno odvojenih alternativnih veza onda je ukupna povezanost krajnjih čvorova jednaka *sumi* vrijednosti izračunatih za svaku alternativnu vezu.

Tip C
Ovaj tip je sličan prethodnom utoliko što su čvorovi bolje povezani ako su te vrijednosti veće. Primjer ovakve veze je izvoz visoko-tehnoloških proizvoda iz jedne u drugu zemlju. Udaljenost dvaju indirektno povezanih čvorova X i Y koji su povezani preko čvora Z računa se po formuli $z=x*y/(x+y)$ gdje je z vrijednost veze XY, x je vrijednost XZ a y je vrijednost ZY. Ako postoji više potpuno odvojenih alternativnih veza onda je ukupna povezanost krajnjih čvorova jednaka *sumi* vrijednosti izračunatih za svaku alternativnu vezu.

Tip D
Bitna značajka ovog tipa je nepostojanje indirektne povezanosti čvorova. Ako postoje vrijednosti veza onda velike vrijednosti označavaju jaku povezanost primjera kao i kod tipova B i C. Primjeri su veza roditelj-dijete, postojanje kopnene granice između država i izvoz automobila iz jedne u drugu državu.

Iako za tip D nemamo indirektnu povezanost čvorova njeno postojanje može biti vrlo značajno. Na primjer, kopnena granica između država XZ i ZY ne implicira kopnenu granicu između XY. Ali implicira kopnenu povezanost koja može biti važna za širenje životinja, ratnu opasnost, prometnu povezanost i slično. Znači da ima smisla i za tip D konstruirati značajke o indirektnoj povezanosti. Karakteristično je što osim broja među čvorova nemamo bolju mjeru udaljenosti tako povezanih čvorova. U analizi rezultata treba koristiti drugi naziv za takvu vezu: umjesto kopnene granice pojam kopnene povezanosti, umjesto veze roditelj-dijete pojam potomstva i predaka. Realizirani sustav je trenutno nepotpun u tom smislu jer ne postoji način definiranja naziva indirektnih veza za tip D.

Primjer izvoza/uvoza automobila za tip D pokazuje osjetljivost definiranja tipova veza i njihova korištenja u konstrukciji značajki. Očito je da se i izvoz automobila može tretirati kao izvoz visoko-tehnoloških proizvoda koji smo svrstali pod tip C. Jedina praktična razlika je što pod visoko-tehnološkim proizvodima smatramo reprodukcijski materijal pa će u slučaju povećanog izvoza iz zemlje Z u zemlju Y doći i do povećanog uvoza iz zemlje X u zemlju Z. Za automobile ne možemo očekivati takav efekt jer povećani izvoz iz zemlje Z u principu ne povlači povećani uvoz u zemlju Z. Ovaj primjer jasno pokazuje da se pri definiranju tipa veze treba pažljivo procijeniti stvarno značenje veze.

Zbog svoje složenosti vrijednosti veza za tipove A-C se računaju samo do čvorova u drugom krugu.

Pojam centralnosti čvorova u mreži

Važna karakteristika svakog čvora u mreži je njegova centralnost relativno prema drugim čvorovima. U užem smislu smislu svojstvo centralnosti označava dobru povezanost sa drugim čvorovima odnosno značaj nekog čvora za povezanost drugih čvorova. Osnovne definicije su da je centralan čvor onaj koji ima najveći broj veza (eng. degree centrality), da je centralan čvor onaj koji ima najmanju prosječnu udaljenost do drugih čvorova (eng. closeness), odnosno da je centralan čvor onaj kroz koji prolazi najveći broj najkraćih veza u mreži (eng. betweenness). Složenije mjere centralnosti su Katzova centralnost i PageRank centralnost koje u obzir uzimaju ne samo povezanost nekog čvora nego i značaj čvorova sa kojima je on povezan. Detaljne definicije svih ovih mjera centralnosti mogu se naći na https://en.wikipedia.org/wiki/Centrality.

Svaka od mjera centralnosti pridružuje numeričku vrijednost svakom čvoru. Izračunavanje vrijednosti za neku mjeru centralnosti za čvorove predstavlja generiranje nove značajke koja karakterizira čvorove. To znači da ako imamo pet različitih mjera centralnosti tada njihovim izračunavanjem dobivamo pet značajki čvorova.

Katzova centralnost i PageRank centralnost imaju veliku složenost računanja pa ih trenutno ne koristimo.

Mjere centralnosti omogućuju definiciju centralnih čvorova a nakon što su definirani centralni čvorovi tada se mogu definirati i značajke koje karakteriziraju udaljenosti svakog čvora X od definiranih centralnih čvorova. Zbog toga je važno odrediti primjeren skup centralnih čvorova.

U strojnom učenju centralnost definiramo općenito kao osobitost nekog čvora u bilo kojem pogledu. Neki čvor je centralan ako se ističe po bilo kojoj osobini koju je moguće kvantizirati. To praktično znači da se većina značajki koje se konstruiraju mogu iskoristiti i za definiranje centralnih čvorova. Dodatno, centralni čvorovi se mogu definirati i za vrijednosti atributa pridruženih čvorovima.

Isti čvor može biti centralan po više osnova. U tom slučaju generiranje značajki vezanih na takav centralni čvor radi se samo jednom. O konačnom izboru centralnih čvorova odlučuje korisnik.