Uvod u složeno pretraživanje podataka

Modeliranje usmjerenih mreža niskim rangom

29. studenoga 2018.

Ana Parać
Sara Pužar
Petra Rožić
Paula Vujasić

Sadržaj

Sažetak

U današnje vrijeme sve su popularnije tzv. mreže povjerenja gdje se korisnici međusobno ocjenjuju na način da međusobno iskazuju povjerenje, odnosno nepovjerenje. Pokazuje se da prema podacima o međusobnim interakcijama korisnike možemo podijeliti u više grupa tako da između pojedine grupe vlada povjerenje te između grupa nepovjerenje. Zbog toga je mreže povjerenja moguće modelirati niskim rangom. Pojavljuje se problem: kako predvidjeti kako će jedan korisnik ocijeniti drugog, tj. kako bi ga ocijenio da ga ocjenjuje? Ovaj problem ćemo u radu nazivati problemom nepotpunih predznaka, a moguće ga je riješiti nadopunjavanjem, ili faktorizacijom matrice. Osim predviđanja o međusobnom ocjenjivanju, može se pokazati korisnim i grupirati korisnike, iako ne znamo sve veze među njima. Zato ćemo se baviti i problemom klasteriranja.

1 Opis problema

1.1 Uvod

Društvo sve više teži digitalizaciji i sve su više zastupljene društvene mreže. Možemo ih reprezentirati pomoću grafova gdje čvorovi predstavljaju korisnike, a bridovi veze među njima. Mi ćemo proučavati sve popularnije mreže povjerenja (*trust networks*) gdje korisnici izražavaju povjerenje, odnosno nepovjerenje jedni prema drugima. Takve mreže se mogu modelirati usmjerenim grafovima, gdje težina brida iznosi +1 ili -1. Primjer takve mreže jest Wikipedija, gdje korisnici mogu ocijeniti druge korisnike, kao zadovoljavajuće (takvu bi vezu modelirali s +1) ili nezadovoljavajuće (modelirali bi je s -1) .

Jedan od važnijih pojmova vezanih uz usmjerene mreže je strukturalna ravnoteža (*structural balance*) . Koncept strukturalne ravnoteže nalaže da su korisnici skloni pratiti pravila "neprijatelj mog prijatelja je moj neprijatelj" i "neprijatelj mog neprijatelja je moj prijatelj". Takav princip koristan je u problemu određivanja nepoznatih veza gdje o prirodi nepoznatih odnosa možemo zaključiti na temelju poznatih informacija o vezama. Ovakav koncept podrazumijeva da su korisnici podijeljeni u dvije grupe (gdje se u grupi korisnici međusobno podržavaju, a između grupa ne podržavaju), što se ne pokazuje dobrim za modeliranje stvarnih odnosa. Zato se uvodi pojam slabe ravnoteže (weak balance), koja dijeli korisnike u veći broj grupa. U ovom se seminaru bavimo modeliranjem niskim rangom (low rank model) kojeg su predložili Cho-Jui Hsieh, Kai-Yang Chiang i Inderjit S. Dhillon (1). Pomoću njega rješavamo problem nepotpunih predznaka (sign inference) i problem klasteriranja.

1.2 Matematičko modeliranje problema

Usmjerenu mrežu prikazujemo kao graf $G = (V, E, A)$, gdje je V skup svih vrhova veličine n, E je skup svih bridova veličine m te je $A \in \mathbb{R}^{n\times n}$ matrica susjedstva koja ima sljedeći oblik:

$$A_{ij} = \begin{cases} 1 & \text{ako } i\text{-ti korisnik vjeruje } j\text{-tom korisniku} \\ -1, & \text{ako } i\text{-ti korisnik ne vjeruje } j\text{-tom korisniku} \\ 0, & \text{ako nije poznato vjeruje li } i\text{-ti korisnik } j\text{-tom korisniku} \end{cases} \tag{1}$$

Uvodimo skup poznatih podataka s Ω, odnosno $(i,j) \in \Omega$ ako i samo ako $A_{ij} \neq 0$. Mreža je potpuna ako su svi elementi od A različiti od nule.

U nastavku formalno definiramo pojmove strukturalne ravnoteže i slabe ravnoteže.

Definicija 1 (Strukturalna ravnoteža za potpune grafove). Potpuna usmjerena mreža je strukturalno uravnotežena ako sve trijade u grafu imaju jedno od navedenog

- sve pozitivne bridove
- samo jedan pozitivan brid

Definicija 2 (Slaba ravnoteža za potpune grafove). Potpuna usmjerena mreža je slabo uravnotežena ako ne postoji trijada u mreži koja sadrži dva pozitivna i jedan negativan brid.

Teorem 3 (Globalna jaka strukturalna ravnoteža)**.** *Potpuna usmjerena mreža je strukturalno uravnotežena ako i samo ako su svi bridovi pozitivni ili vrhovi mogu biti podijeljeni u dvije grupe, takve da su bridovi unutar grupe pozitivni, a bridovi između grupa negativni.*

Teorem 4 (Globalna slaba strukturalna ravnoteža)**.** *Potpuna usmjerena mreža je slabo strukturalno uravnotežena ako i samo ako su svi bridovi pozitivni ili vrhovi mogu biti podijeljeni u više grupa, takvih da su bridovi unutar grupe pozitivni, a bridovi između grupa negativni.*

Sljedeći teorem pokazuje da matrica susjedstva A potpune k-slabo uravnotežene mreže ima nizak rang. S odgovarajućim preslagivanjem vrhova, A se može reprezentirati s blok-dijagonalnom matricom gdje su elementi u dijagonalnim blokovima $+1$, a na ostalim mjestima -1.

Teorem 5 (Struktura niskog ranga usmjerenih mreža)**.** *Matrica susjedstva A potpune k-slabo uravnotežene mreže ima rang 1 ako je $k \leq 2$, i rang k ako je $k > 2$.*

$$\begin{bmatrix} 1 & -1 & -1 & -1 & -1 & -1 \\ -1 & 1 & 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & -1 & -1 & -1 \\ -1 & -1 & -1 & 1 & -1 & -1 \\ -1 & -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix}$$

Figura 1. Prikaz strukture niskog ranga 4-slabo uravnotežene mreže

Budući da većina stvarnih mreža nije potpuna, definiramo slabu ravnotežu za generalne grafove, odnosno mreže koje mogu i ne moraju biti potpune.

Definicija 6. (Slaba ravnoteža za generalne grafove) Usmjerena mreža je slabo uravnotežena ako i samo ako je moguće dodati nepoznate bridove tako da rezultirajuća potpuna mreža bude slabo uravnotežena.

2 Metode rješavanja

Problem nadopunjavanja nepotpunih predznaka usmjerenih mreža može se riješiti na dva načina :

- Određivanjem predznaka pomoću nadopunjavanja matrice
- Određivanjem predznaka pomoću faktorizacije matrice

U radu detaljnije obrađujemo određivanje predznaka pomoću faktorizacije matrice budući da smo implementirale tu metodu. Navedenu metodu izabrale smo jer ona ima najbolje performanse u svim slučajevima, čak i za neuniformno distribuirane mreže, kao što je većina pravih mreža (npr. *Slashdot*).

U nastavku također obrađujemo klasteriranje.

2.1 Određivanje predznaka pomoću nadopunjavanja matrice

Za danu usmjerenu mrežu A s poznatim bridovima $A_{ij}, (i,j) \in \Omega$ želimo naći potpunu matricu X dodjeljujući $+1$ ili -1 svakom nepoznatom bridu tako da rezultirajuća matrica bude niskog ranga. Formalno problem zapisujemo :

$$\begin{aligned} &\text{minimizacija ranga}(X) \text{ tako da} \\ &X_{ij} = A_{ij}, \forall (i,j) \in \Omega \\ &X_{ij} \in \{+1, -1\}, \forall (i,j) \notin \Omega \end{aligned} \tag{2}$$

Poznato je da je problem (2) NP-težak, zato pronalazimo približno rješenje tako što problem minimizacije ranga mijenjamo problemom minimizacije norme (*trace norm*) od X. Zato relaksirani problem zapisujemo na sljedeći način :

$$\begin{aligned} &\text{minimizacija } \|X\|_* \text{ tako da} \\ &X_{ij} = A_{ij}, \forall (i,j) \in \Omega \end{aligned} \tag{3}$$

Ispostavlja se da u određenim uvjetima, rješavajući (3) možemo točno odrediti nepoznate predznake povezane mreže. Ti uvjeti podrazumijevaju da su podaci uniformno raspoređeni (međutim, to kod pravih mreža često nije slučaj).

Poznato je da problem (3) može biti riješen algoritmima SDP (*SemiDefinite Programming*) ili SVP (Singular Value Projection)

2.2 Određivanje predznaka pomoću faktorizacije matrice

Problem popunjavanja podataka koji nedostaju u povezanim mrežama svest ćemo na faktorizaciju matrice baziranoj na gradijentu. U pristupu faktorizacije matrica razmatrat ćemo sljedeći problem (4):

$$\min_{W,H \in \mathbb{R}^{k \times n}} \sum_{(i,j) \in \Omega} (A_{ij} - (W^T H)_{ij})^2 + \lambda \|W\|_F^2 + \lambda \|H\|_F^2 \tag{4}$$

Kvadrat razlike u prvom članu prethodne jednadžbe (4) teži postavljanju elemenata matrice $W^T H$ na $+1$ ili -1. Međutim, nas više zanima istovjetnost predznaka ij-tog elementa

u matricama $W^T H$ i A nego njihova međusobna razlika. U razmatranju/pisanju algoritama pozabavit ćemo se i time. Stoga, umjesto kvadrata razlike, koristit ćemo funkciju koja kažnjava neistovjetnost predznaka. Sada problem možemo zapisati u sljedećoj formi (5).

$$\min_{W,H \in \mathbb{R}^{k \times n}} \sum_{(i,j) \in \Omega} \ell(A_{ij} - (W^T H)_{ij}) + \lambda \|W\|_F^2 + \lambda \|H\|_F^2 \tag{5}$$

pri čemu je funkciju greške ℓ možemo razmatrati kao funkcije *sigmoid* i *square-hinge*.

$$\ell_{sigmoid}(x, y) = 1/1 + \exp(xy) \tag{6}$$

$$\ell_{square-hinge}(x, y) = (max(0, 1 - xy))^2 \tag{7}$$

Postoje dvije glavne tehnike za riješavanje problema (5): ALS (*Alternating Least Squares*) i SGD (*Stochastic Gradient Descent*). ALS se može koristiti samo kod funkcije greške reprezentirane kvadratom razlike. Nama je cilj analizirati sve funkcije greške. Stoga kako bi riješili problem (5) koristit ćemo metodu stohastičkog gradijentnog spusta (SGD).

2.2.1 Metoda SGD

Za svaku iteraciju slučajnim odabirom odaberemo ij-ti element te ažuriramo samo i-ti stupac matrice W i j-ti stupac matrice H, odnosno w_i i h_j. Pravila ažuriranja stupaca dana su sljedećim jednadžbama (8), (9)

$$w_i \leftarrow w_i - \eta \left(\frac{\partial \ell(A_{ij}, (W^T H)_{ij})}{\partial w_i} + \lambda w_i \right) \tag{8}$$

$$h_j \leftarrow h_j - \eta \left(\frac{\partial \ell(A_{ij}, (W^T H)_{ij})}{\partial h_j} + \lambda h_j \right) \tag{9}$$

pri čemu je η prethodno definirani parametar. Njegova vrijednost ne smije biti prevelika kako se optimum ne bi preskočio, ali ne smije biti ni premala jer bi bio potreban prevelik broj iteracija.

2.3 Klasteriranje

Kako smo već napomenule, stvarne usmjerene mreže često imaju slabu strukturalnu ravnotežu te je stoga cilj klasteriranja pronalazak k particija takvih da su unutar svake particije vrhovi čiji su međusobni bridovi većinom pozitivni, a bridovi između vrhova uz različitih particija većinom negativni.

Jedan od načina klasteriranja nepotpune usmjerene mreže A je sljedeći:

- nadopunimo nepoznate predznake matrice koja predstavlja mrežu, tako nadopunjenu matricu označimo s X
- kreiramo matricu U pomoću prvih k svojstvenih vektora od X
- provedemo standardni algoritam klasterizacije na U

3 Opis implementacije algoritma

Za praktični dio projekta odabrale smo baze:

- Bitcoin dataset
- Epinion dataset

Kako bi započele implementaciju SGD algoritma prvo smo naše baze podataka prilagodili zahtjevima budućih algoritama. Kreirali smo matricu susjedstva (A $\in \mathbb{R}^{n \times n}$) te smo nju faktorizirale pomoću MF algoritma.

Faktorizacija matrice (MF):
$input : A$ A$\in \mathbb{R}^{n \times n}$
begin
$k = expectedRankOfMatrix$
$W = rand([-1, 1], k, n)$
$H = rand([-1, 1], k, n)$
for $iter := 1$ to $iterNum$ step 1 do
$z = rand(1, size(W))$
$element = findNonNull(A)$
$i = element(z)$
$j = element(z)$
$eij = 1 - A(i, j) * dot(W(:, i), H(:, j))$
$W(:, i) = W(:, i) - eta * (-A(i, j) * e_{ij} * H(:, j) + lambda * W(:, i))$
$H(:, j) = H(:, j) - eta * (-A(i, j) * e_{ij} * W(:, i) + lambda * H(:, j))$
end
$return : W^T * H$
end

Nakon faktorizacije dobili smo matrice W $\in \mathbb{R}^{n \times k}$ i H $\in \mathbb{R}^{n \times k}$, pri čemu je $W^T H$ naša nova tražena matrica (newMatrix $\in \mathbb{R}^{n \times n}$). Između matrica A i newMatrix računali smo greške. Funkcije računanja graška prezentirane su u sljedećim pseudokodovima.

Greška udaljenosti:
$input : A, newMatrixA$
begin
$element = findNonNull(A)$
$errorNum = 0$
for $i := 1$ to $size(element)$ step 1 do
$errorNum =$
$errorNum + pow(A(element(i).x, element(i).y) - newMatrixA(element(i).x, element(i).y, 2)$
end
$return : errorNum$
end

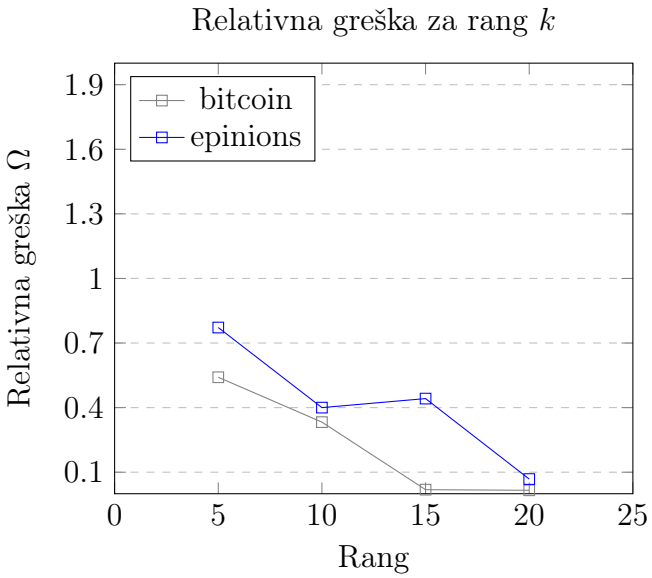
Greška predznaka:
$input : A, newMatrixA$
begin
  $element = findNonNull(A)$
  $errorNum = 0$
  for $i := 1$ to $size(element)$ step 1 do
    if $sgn(A(element(i).x, element(i).y))! = sgn(newMatrixA(element(i).x, element(i).y))$
      $errorNum++$
    end
  end
  $return : errorNum$
end

Norm greška:
$input : A, W, H$
begin
  $element = findNonNull(A)$
  for $i := 1$ to $size(element)$ step 1 do
    $vektor = W(:, element(i).x)^T * H(:, element(i).y) - A(element(i).x, element(i).y);$
  end
  $errorNum = normf(vector)/normf(A)$
  $return : errorNum$
end

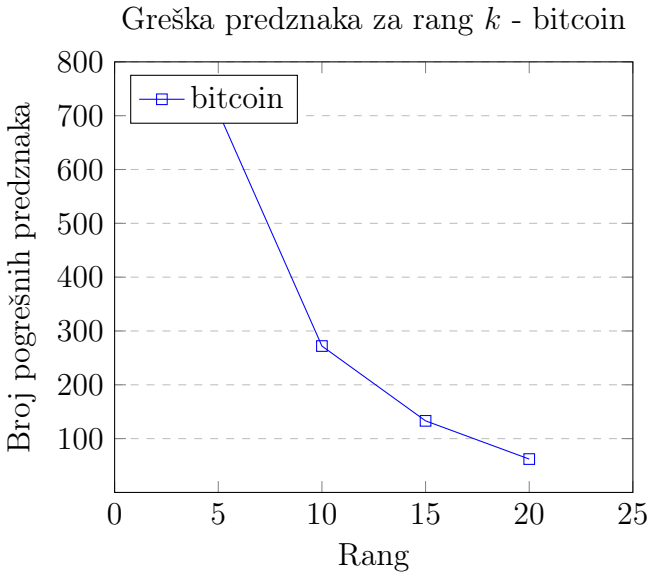
4 Rezultati testiranja

U rezultatima prikazujemo rezultate testiranja algoritma faktorizacije matrice na podacima preuzetih sa stranice *https://snap.stanford.edu/*. Algoritam je testiran na podatcima o korisnicima koji trguju bitcoin valutom te si međusobno iskazuju povjerenje ili nepovjerenje te na podacima sa stranice epinions.

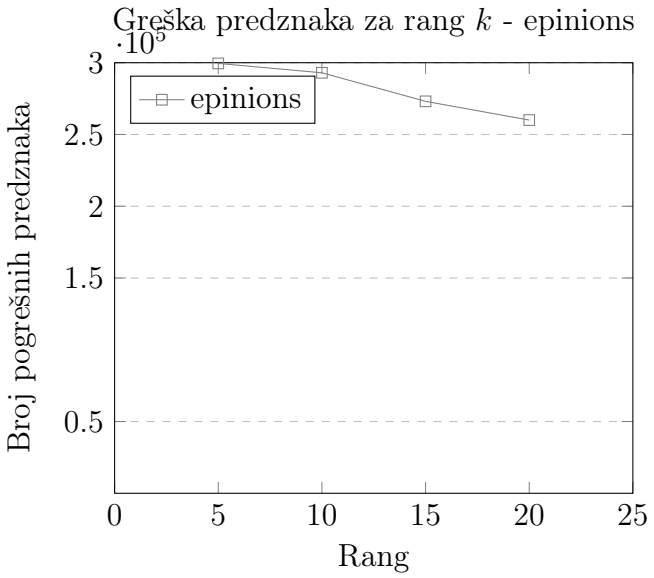
Sljedećim grafom prikazujemo relativnu grešku Ω za oba skupa podataka za rangove $5, 10, 15$ i 20 u verziji algoritma koja koristi *square hinge* kao funkciju gubitka podataka. (Ta je verzija dala najbolje rezultate.)



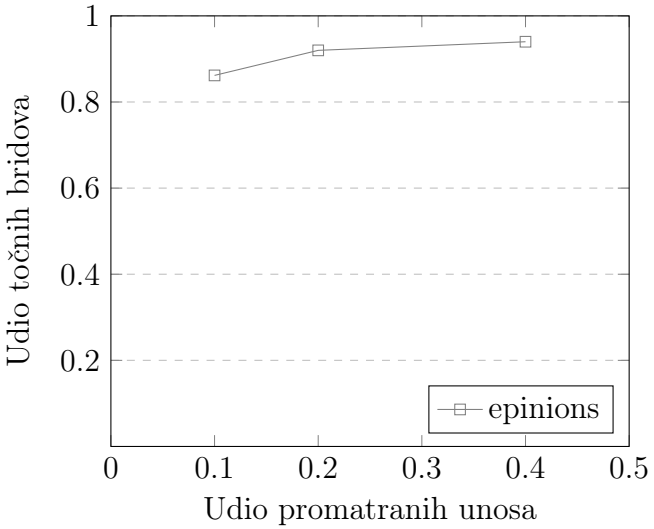
Sljedećim grafom prikazujemo grešku izraženu kao broj krivo dobivenih predznaka za *bitcoin* podataka za rangove $5, 10, 15$ i 20 u verziji algoritma koja koristi *square hinge* kao funkciju gubitka podataka.



Zbog razlike u količini podataka koja uzrokuje znatno drugačije rezultate funkcije greške za ovaj test, rezultate za *epinions* prikazujemo na posebnom grafu.



Rezultate testiranja algoritma klasteriranja s uniformnim odabirom uzoraka prikazujemo sljedećim grafom. Testiranje smo provele na sintetičkim podacima generiranim slučajnim odabirom iz 10-slabo uravnotežene mreže gdje je veličina svake grupe 100.



5 Literatura

Literatura

[1] Hsieh, Cho-Jui,Chiang, Kay-Yang, Dhillon, Inderjit S.: *Low Rank Modeling of Signed Networks*,

[2] Bauckhage, Christian: *k-Means Clustering Is Matrix Factorization*

[3] Hsieh, Cho-Jui,Chiang, Kay-Yang, Dhillon, Inderjit, Natarajan, Nagarajan: *Prediction and Clustering in Signed Networks: A Local to Global Perspective*

[4] *Gradient descent and stochastic gradient descent from scratch* https://gluon.mxnet.io/chapter06$_o$*ptimization*/$gd - sgd - scratch.html$