Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.65
DOI: https://doi.org/10.24867/11BE16Jankov

STOHASTIČKI BLOK MODEL I KLASIFIKACIJA

STOCHASTIC BLOCK MODEL AND CLASSIFICATION

Vladimir Jankov, Dragana Bajović, Željen Trpovski, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *Analiza stohastičkog blok modela u smislu grafova i kako možemo iskoristiti landing verovatnoće za klasifikaciju. U ovom radu grupisali smo centroide različitih grafova dobijenih stohastičkim blok modelom i utvrdili da njihovi centroidi landing verovatnoća teže jednoj tački [6].*

Ključne reči: *Grafovi, Stohastički blok model, klasifikacija*

Abstract – *An analysis of the stochastic block model for graphs and how we can use the landing probabilities for classification. In this work we grouped the centroids of different graphs generated with the stochasitc block model and concluded that their landing probabilities converge to a single point.*

Keywords: *Graphs, Stochastic block model, classification*

1. UVOD

U ovom radu analizirali smo stohastički blok model i kako se on može primeniti na grafove. Dali smo primer i definiciju stohastičkog blok modela i sliku grafa generisanog *SBM* modelom.
Pored toga, analizirali smo interakcije unutar čvorova kroz njihove *landing* verovatnoće.
Generisali smo više različitih grafova sa istim parametrima na osnovu stohastičkog blok modela i utvrdili kako se njihovi predstavnici ponašaju.
U radu je najviše akcenat stavljen na *landing* verovatnoće i njihovu interpretaciju.

2. STOHASTIČKI BLOK MODEL

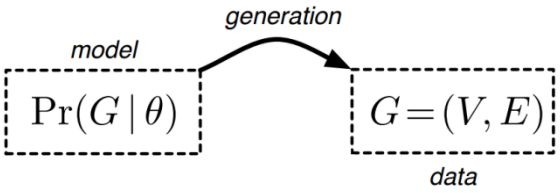
Stohastički blok model je probabilistički ili generativni model koji dodeljuje probabilistčku vrednost za svaki par čvorova *i,j*. Generativni modeli su veoma efikasni načini kodiranja određenih pretpostavki o načinu interakcija nepoznatih parametara pri formiranju veza između čvorova.

Prednost ovakvih modela jeste u tome sto nam pružaju mogućnost da koristimo „likelihood skorove“, koji su zasnovani na osnovnim principima statistike i verovatnoće kako bi poredili različite parametrizacije i druge slične modele.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentori su bili dr Dragana Bajović, docent i dr Željen Trpovski, vanr. prof.

Još jedna od prednosti jeste u tome što nam oni omogućavaju da predstavimo svet eksplicitno, za razliku od standardnih pristupa gde bismo morali da nekim algoritmom tako nešto nadogradimo. Njihovi parametri često imaju direktnu interpretaciju iz neke pretpostavke o strukturi mreže grafa.

Kao i ostali generativni modeli, stohastički blok model definiše distribuciju nad grafom $P(G|\Theta)$, gde je Θ skup parametara nad grafom (verovatnoća da čvor pripada određenoj zajednici ili da postoji veza između 2 čvora (Slika 1)). Na osnovu parametara Θ moguće je izgenerisati celu strukturu grafa.



Slika 1. *Generisanje grafa na osnovu verovatnoća*

2.1. Definicija

U najosnovnijoj varijanti stohastički blok model se može definisati pomoću:

- k : skalara koji predstavlja broj grupacija unutar grafa.
- $z : n \times 1$ vektor skrivenih labela čvorova z_l je grupni index l-tog čvora.
- M : je $k \times k$ stohastička blok matrica, gde je M_{ij} verovatnoća da je čvor i povezan sa čvorom j.

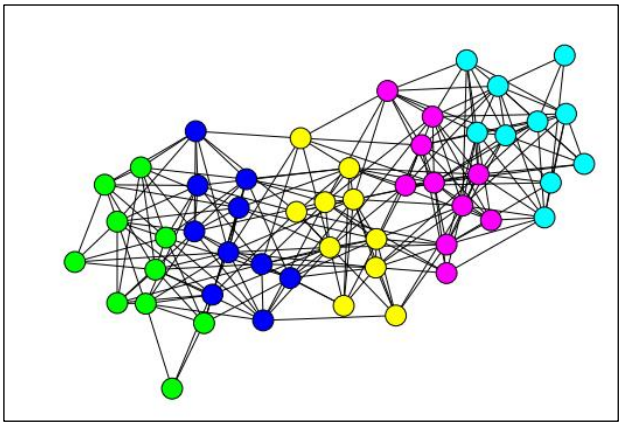
U samoj implementaciji stohastičkog blok modela, korisnik zadaje klasne verovatnoće, verovatnoće povezanosti dva čvora kada su iz iste klase p_{in} i kada nisu p_{out}. Na uniforman način se dodeljuju skrivene labele čvorovima i prisustvo veze između dva čvora [3]. M matrica sadrži vrednost p_{in} na glavnoj dijagonali, a na ostalim mestima p_{out}. Posmatrali smo model sa 2 klase. Za vrednosti p_{in} i p_{out} odabrali smo 0.8 i 0.2, a verovatnoće pripadanja bilo kojoj od dve klase jednake su.

2.2. Primer grafa i *M* matrice

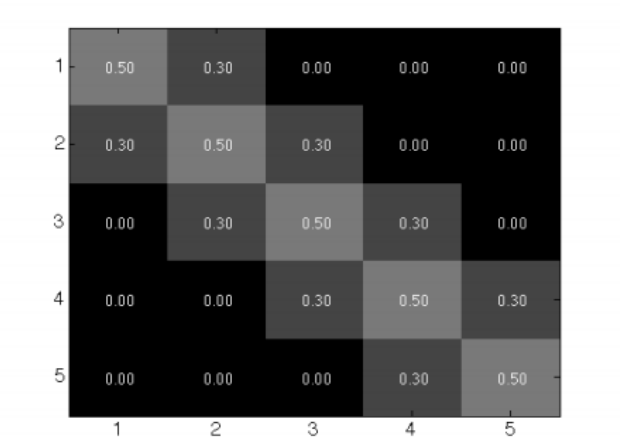
Primeri grafova dati su na modelima na kojima je prisutno 5 različitih grupa (Slika 2) i interakcije među njima su različite. U zavisnosti od matrice M moguće je naglasiti koje grupe su sličnije. Dobar primer takvih grafova bi bio graf gde je matrica M kod susednih grupa veća tj.

vrednosti pored glavne dijagonale su značajno veće od ostalih grupa, a ipak manje nego na glavnoj dijagonali.

Prikaz ovakvog grafa bi mogle biti grupe ljudi na društvenim mrežama koje često komuniciraju ili predstava podataka sa realnih senzorskih mreža (očitavanje temperature) gde će bliži čvorovi imati sličnije vrednosti pa stoga će tu između njih postojati veza. Kod udaljenijih senzora, vrednosti će biti različite pa samim time neće postojati veza između tih čvorova.



Slika 2. *Uređen graf sa 5 zajednica (grupacija)*



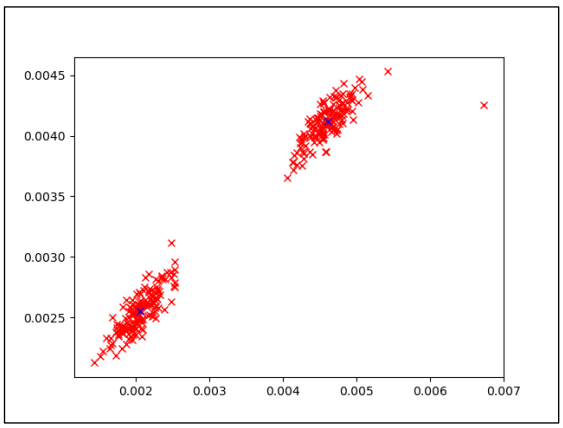
Slika 3. *M matrica grafa sa slike 2.*

3. METODOLOGIJA I ANALIZA

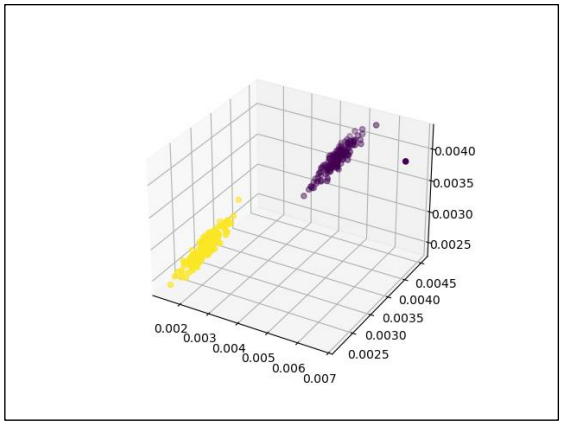
Fokusirali smo se na određivanje *landing* verovatnoća r_k^v. One predstavljaju verovatnoću da ćemo se naći u *v*-tom čvoru nakon *k* koraka počevši od nekog fiksnog čvora.

Grupisali smo više *landing* verovatnoća od prve pa do dobijene nakon *K* koraka u jedan vektor r^v tj. vektor r^v sadrži sve verovatnoće (r^v_1, r^v_2,..., r^v_K).

Kako bismo uočili pravilnosti na grafiku (Slika 3) smo prikazali vektor (r^v_2, r^v_3) i (r^v_2, r^v_3 , r^v_4). Za ovu analizu izgenerisali smo graf preko stohastičkog blok modela sa 300 čvorova i za verovatnoće p_{in} i p_{out} smo uzeli vrednosti 0.8 i 0.2. Verovatnoće za skrivene labele su jednake (klasne verovatnoće).



Slika 3. *Landing verovatnoće* (r^v_2, r^v_3) *grafa sa 300 čvorova*



Slika 4. *Landing verovatnoće* (r^v_2, r^v_3 , r^v_4) *grafa sa 300 čvorova*

Ponavljajući ovaj postupak uočili smo pravilnost da centroidi, različitih grafova generisanih stohastičkim blok modelom sa istim parametrima, teže ka jednoj tački (Slika 4). Centroidi se računaju prema sledećoj formuli:

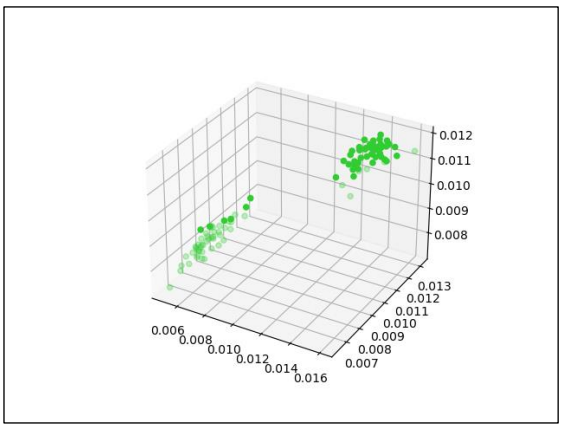
$$a\hat{} = \frac{\sum_{r^v \in c_1} r^v}{n_1} \quad (1)$$

$$b\hat{} = \frac{\sum_{r^v \in c_2} r^v}{n_2} \quad (2)$$

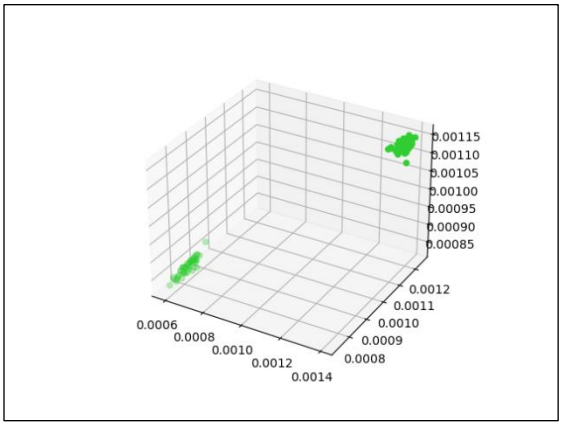
Napravili smo 50 grafova i na 3D grafiku prikazali njihove centroide kao i u prethodnom eksperimentu. Imali smo dve realizacije, kada imamo 100 čvorova (Slika 5) i kada imamo 1000 čvorova (Slika 6). Primetili smo da povećanjem broja čvorova u grafu njihovi centroidi teže ka elipsoidnom obliku i da njih možemo interpretirati kao predstavnike za čvorove koji pripadaju određenoj klasi. Ova tvrdnja je dokazana u radu [1] i tvrdi da centroidi grafova zavise od verovatnoća p_{in} i p_{out}. Razlika centroida $\omega\hat{} = a\hat{} - b\hat{}$ se može aproksimirati sledećim izrazom:

$$\psi\hat{} = \frac{1}{N}\left(\frac{p - P_{out}}{p + P_{out}}\right)^k \quad (3)$$

gde je *k* broj skokova u grafu.



Slika 5. *Centroidi landing verovatnoća grafova sa 100 čvorova*



Slika 6. *Centroidi landing verovatnoća grafova sa 1000 čvorova*

Iz plotova se može jasno uočiti tvrdnja da teže ka elipsoidu što je dokazano u radu [1]. Prednost ovakvog pristupa ogleda se u tome da nam nisu neophodne skrivene labele kako bismo vršili klasifikaciju čvorova, dovoljno nam je da poznajemo osobine grafa kako bi estimirali verovantoće p_{in} i p_{out}. Uz poznate verovatnoće, možemo izračunati centroide i samim tim vršiti klasifikaciju. Dok kod metoda kao što su k-means moramo poznavati labele čvorova, izračunavati distance i iterativno dolaziti do centroida, ovom metodom možemo da ih izračunamo i samim time koristimo za klasifikaciju.

4. KLASIFIKACIJA

Kako bi klasifikacija bila izvršena neophodno je formirati koeficijente ω diskriminantne funkcije. Polazi se od Fišerove formulacije diskriminantne funkcije gde su nam neophodni centroidi i varijanse labeliranih čvorova u grafu. Definišemo verovatnoće da za zadat vektor landing verovatnoća $r = (r_1, ..., r_K)$ čvor pripada klasi 0 ili 1:

$$P(z=1) \propto |\sum_a| \quad \frac{-1}{2} e^{\frac{-1}{2}(r-a)^T \sum_a^{-1}(r-a)} \tag{4}$$

$$P(z=1) \propto \sum_b \vee \quad \frac{-1}{2} e^{\frac{-1}{2}(r-b)^T \sum_b^{-1}(r-b)} \tag{5}$$

Pod uslovom da su poznati parametri stohastičkog blok modela, verovatnoća da neki čvor pripada klasi a je jednaka broju čvorova u klasi a podeljena sa ukupnim brojem čvorova u klasi a. Sa tim poznatim parametrima, logaritamski odnos verovatnoća je jednak:

$$g(r) = ln \frac{P(z=1)P(z=1)}{P(z=0)P(z=0)} = \omega^T r + r^T W r + \omega_0 \tag{7}$$

Konstanta ω_0 se može odbaciti jer je ona jednaka za sve čvorove i dalje neće uticati na klasifikaciju. Ako pretpostavimo da su kovarijansne matrice jednake i dijagonalne ($\Sigma_a = \Sigma_b = \sigma^2 I$), dobijemo oblik prvobitne geometrijske diskriminantne funkcije:

$$g_1(r) = \sigma^{-2}(a-b)^T r + C \tag{8}$$

Parametri ovih jednačina uglavnom nisu poznati, i moraju da se estimiraju. Za estimaciju parametra stohastičkog blok modela G((n_a.n_b), P) gde su verovatnoće p_{11} i p_{22} jednake p_{in}, a verovatnoće p_{12} i p_{21} jednake p_{out}. Ako nam je poznat matrica susedstva grafa, parametri p_{in} i p_{out} se estimiraju pomoću sledećih formula

Parametri ovih jednačina uglavnom nisu poznati, i moraju da se estimiraju. Za estimaciju parametra stohastičkog blok modela G((n_a.n_b), P) gde su verovatnoće p_{11} i p_{22} jednake p_{in}, a verovatnoće p_{12} i p_{21} jednake p_{out}. Ako nam je poznat matrica susedstva grafa, parametri p_{in} i p_{out} se estimiraju pomoću sledećih formula [2]:

$$\hat{p}_{out} = \frac{(s_3 - s_2 s_3)m_1^3 + (s_2^3 - s_3)m_2 m_1 + (s_3 s_2 - \cdot}{(m_1^2 - m_2)(2s_2^3 - 3s_3 s_2 + s_3)} \tag{9}$$

$$\hat{p} = \frac{m_1 + (s_2 - 1)\hat{p}_{out}}{s_2} \tag{10}$$

gde su s_2, s_3, m_1, m_2, m_3 parametri dobijeni pomoću matrice susedstva i broja pripadnika određenih klasa:

$$s_2 = n_a^2 + n_b^2 \tag{11}$$

$$s_3 = n_a^3 + n_b^3 \tag{12}$$

$$m_1 = \frac{1}{n(n-1)} \sum_{i,j=1, i \neq j} A_{ij} \tag{13}$$

$$m_2 = \frac{1}{n(n-1)(n-2)} \sum_{i,j,k=1, i \neq j \neq k} A_{ij} A_{ik} \tag{14}$$

$$m_3 = \frac{1}{n(n-1)(n-2)} \sum_{i,j,k=1, i \neq j \neq k} A_{ij} A_{ik} A_{jk} \tag{15}$$

Nepoznata kovarijansna matrica koja se koristi u Fišerovoj diskriminantoj funkciji, sa estimiranim parametrima $\hat{p_{in}}$, $\hat{p_{out}}$ i gde je $n_a = n_b$, može se estimirati

ponavljanjem simulacije grafa sa tim parametrima putem Monte Carlo metode i izračunavanjem centroida $a\hat{}(j)$ za svaku od realizacija grafa G, gde je *j* indeks Monte Carlo simulacije [4,7]. Kovarijansne matrice Σ_a računa se po sledećoj formuli:

$$\Sigma_a = \sum_{j=1}^{J} \frac{\left(a\hat{}_k^{(j)} - \underline{a}_k\right)\left(a\hat{}_k^{(j)} - \underline{a}_k\right)^T}{J} \qquad (16)$$

gde se $\underline{a}_k = \frac{\sum_{j=1}^{J} a\hat{}_k^{(j)}}{J}$ iJ je ukupan broj Monte Carlo simulacija. Isto tako se estimira kovarijansna matrica Σ_b centroida $b\hat{}$.

Pored ovog pristupa, uz poznavanje labela trening skupa podataka, moguće je izračunati vrednosti za p_{in} i p_{out}. Na osnovu matrice susedstva i poznatih labela tih čvorova verovatnoću p_{in} izračunavamo tako što saberemo jedinice iz matrice susedstva kod čvorova iz istih klasa i podelimo sa ukupnim brojem mogućih veza unutar klase [5], dok p_{out} izračunavamo kao broj veza van klase podeljeno sa mogućim brojem veza van klase.

5. ZAKLJUČAK

Na osnovu prethodnih analiza može se zaključiti da kod grafova koji se ponašaju kao stohastički blok model mogu se odrediti skrivene zajednice na osnovu njihovih *landing* verovatnoća. Centroidi reprezentativnih klasa su jasno uočljivi, lako se izračunavaju i mogu se iskoristiti za klasifikaciju.

Postoje mogućnosti predstave raznih skupova podataka preko grafova. Ako se podaci prestavljeni preko grafa ponašaju na sličan način kao stohastički blok model, postoji mogućnost pravljenja efektivnih klasifikatora podataka.

Ovakva vrsta klasifikacije može biti veoma pogodna za klasifikaciju zajednica na društvenim mrežama, s tim da se ove zajednice ponašaju na sličan način.

Prednost ovakvog klasifikatora jeste to da parametri p_{in} i p_{out} mogu da se estimiraju analitički na osnovu matrice susedstva. U našem slučaju to nije bilo pogodno koristiti zbog haotičnosti podataka i nepredvidljivosti generisanja matrice susedstva, dok u slučajevima gde su jasnije definisana susedstva, moguće je koristiti ovakve formule.

6. LITERATURA

[1] Isabel Kloumann, " Block Models and Personalized PageRank", *Proc. National Academy of Sciences, 114(1) 33-38, 3 January 2017*

[2] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall, " Exact recovery in the stohastic block model", *IEEE Transactions on, 62(1):471–487, 2016.*

[3] http://vtsns.edu.rs/wp-content/uploads/2019/01/Prim-metode-modelovanja-rizika-eksp_Skripta-I-deo.pdf (pristupljeno u septembru 2020.)

[4] James P Bagrow, " Evaluating local community methods in networks", *Journal of Statistical Mechanics: Theory and Experiment, 2008(05):P05001, 2008.*

[5] A.E. Bryson, Y.C. Ho, *"Applied Optimal Control"*, New York, Wiley, 1975.

[6] Emmanuel Abbe, " Community Detection and Stochastic Block Models: Recent Developments", *Journal of Machine Learning Research 18 (2018) 1-86.*

[7] Paul Erdos and Alfred Renyi, " On random graphs", *Publ. Math. Debrecen., 6:290–297, 1959*

Kratka biografija:

Vladimir Jankov rođen je u Zrenjaninu 1996. god. Master rad na Fakultetu tehničkih nauka iz oblasti Elektrotehnike i računarstva – Komunikacione tehnologije i obrade signala odbranio je 2020.god. kontakt: vlajkojjj@gmail.com

Dragana Bajović diplomirala je na Smeru za automatiku na Elektrotehničkom fakultetu, Univerzitet u Beogradu, 2007. Doktorirala je 2013. u okviru dualnog programa između Univerziteta Karnegi Melon, Pitsburg, SAD, i Visokog tehničkog instituta u Lisabonu, Portugal. Dr. Bajović je takođe diplomirala na Fakultetu muzičke umetnosti, Univerzitet u Beogradu, na Odseku za opštu muzičku pedagogiju, 2012.

Željen Trpovski rođen je u Rijeci 1957. godine Doktorirao je na Fakultetu tehničkih nauka 1998. godine. Oblast interesovanja su telekomunikacije i obrada signala.