

Strategije anotacije vozlišč pri polnadzorovanem učenju na grafih

Niko Uremović, Domen Kavran, Tjaša Repič, Niko Lukač

Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru

E-pošta: niko.uremovic@um.si, domen.kavran1@um.si, tjasa.repic1@um.si, niko.lukac@um.si

Node Annotation Strategies for Semi-Supervised Learning on Graphs

Graph neural networks (GNNs) are increasingly used in domains such as fraud and anomaly detection, where only a small fraction of nodes can realistically be labeled. This paper examines how the allocation of a fixed labeling budget affects the predictive performance of GNNs trained in a semi-supervised setting. We evaluate (i) the relationship between the fraction of labeled nodes and downstream performance, (ii) the trade-off between hard, abstaining, and soft labeling strategies under varying annotator reliability, modeled through a target area under the receiver operating characteristic curve (ROC-AUC), and (iii) the effect of label placement heuristics (random, hub-first, outlier-first and community-first) on predictive performance under a fixed labeling budget. Experiments are conducted across a benchmark suite of homogeneous and heterogeneous, homophilic and heterophilic graph datasets, using a graph convolutional network (GCN). Our results offer practical guidance on how to spend a limited labeling budget most effectively when annotating graph-structured data.

1 Uvod

Podatki, organizirani v obliki grafov, so vseprisotni: socialna omrežja, citatni grafi, omrežja transakcij, biološke interakcije in še mnogo drugih sistemov naravno opisujemo z vozlišči in povezavami med njimi [1]. Za razliko od tabelarnih podatkov, kjer so primeri med seboj neodvisni, graf eksplicitno kodira relacijsko strukturo med primeri, ki je klasični modeli strojnega učenja ne morejo izkoristiti [2]. Grafovske nevronske mreže (angl. *graph neural networks*, GNN) so se v zadnjih letih uveljavile kot standardno orodje za učenje na takih podatkih, saj z mehanizmom sporočanja med sosednjimi vozlišči zajamejo tako lastnosti posameznega vozlišča kot tudi strukturo njegove okolice [1].

Posebej pomembno je področje odkrivanja prevar, anomalij in kaznivih dejanj, kjer so entitete (uporabniki, transakcije, računi) praviloma med seboj povezane, sumljivo obnašanje pa se pogosto kaže prav v strukturi teh povezav, ne le v lastnostih posameznega vozlišča [3]. V takih okoljih je pridobivanje oznak (npr. potrjenih primerov prevare) drago, počasno in pogosto zahteva

strokovno znanje, zaradi česar je delež označenih vozlišč v praksi izjemno majhen [4].

Grafovsko strukturirani podatki imajo pri tem edinstveno lastnost: neoznačena vozlišča lahko kljub odsotnosti oznake bistveno prispevajo k informativnosti napovedi za ciljno (označeno) vozlišče, saj nosijo informacijo o strukturi okolice in se preko sporočanja med vozlišči prenašajo v reprezentacijo sosednjih, označenih vozlišč [5]. GNN je zato mogoče naučiti na celotnem grafu, tudi če je oznaka znana le za majhen del vozlišč; preostanek grafa še vedno prispeva k učenju strukturnih reprezentacij [5]. V praksi je tako polnadzorovano učenje (angl. *semi-supervised learning*) najnaravneje realizirano tako, da napovedni model izvede prehod naprej (angl. *forward pass*) čez celoten graf, napako pa nazaj razširimo (angl. *backpropagate*) le iz tistih vozlišč, ki imajo na voljo oznako [5].

V tem prispevku se osredotočimo na vprašanje, kako naj se omejen proračun za označevanje čim bolj izkoristi. Zanima nas, ali obstajajo prakse, smernice ali heuristike, ki lahko izboljšajo napovedno zmogljivost modela zgolj s strategijo razporeditve tega proračuna; brez spremembe arhitekture modela ali količine podatkov.

Vprašanje, katera vozlišča označiti pri omejenem proračunu, spada na področje aktivnega učenja (angl. *active learning*). Na grafih se izbira običajno opira na negotovost napovedi, reprezentativnost in strukturne lastnosti grafa, kot je centralnost [6, 7], novejša dela pa dajejo prednost raznolikosti izbranih vozlišč – prek partitioniranja grafa na skupnosti [8] ali maksimizacije vpliva izbranega vozlišča na preostanek grafa [9].

Vzporedno se obravnava tudi nezanesljiv, šumni anotator (angl. *noisy annotator*): možnost abstinence je bila v splošnem okolju prvič formalno uvedena kot lastna kategorija odziva [10], kasneje razširjena na kombinacijo poceni in zanesljivega anotatorja [11] ter prilagojena za nizek proračun za označevanje [12]. Na grafih se nezanesljivost upošteva pri izbiri vozlišč prek ocene zanesljivosti vpliva [13] ali pri obliki same oznake, kjer namesto trde razvrstitve zadostuje sproščen, mehkejši odziv [14].

Navedena dela praviloma obravnavajo eno samo idejo na ozkem naboru citatnih omrežij; v tem prispevku vprašanji proračuna in zanesljivosti anotatorja obravnavamo vzporedno, na širšem in raznolikejšem naboru podatkovnih množic. Prispevki tega dela so:

- Pregled in priprava nabora primerjalnih podatkovnih množic, ki zajemajo tako homogene kot heterogene grafe ter grafe z visoko in nizko homofilijo.
- Analiza razmerja med deležem označenih vozlišč (proračunom za označevanje) in napovedno zmožljivostjo modela.
- Primerjava strategij označevanja v razmerah, ko strokovnjak (anotator) ne more z gotovostjo določiti pravilnega razreda vozlišča (brez znatno večjega vložka časa). Primerjamo, ali je bolje (a) označiti vozlišče s trdo oznako kljub negotovosti in s tem tvegati napačno oznako, (b) vozlišče izpustiti iz označevanja (abstenenca) in s tem zmanjšati število oznak pri enakem proračunu, ali (c) podati mehko oznako, ki izraža ocenjeno verjetnost posameznega razreda. Pristope primerjamo pri različnih stopnjah "težavnosti" označevanja.
- Raziskava, ali lahko heuristike za razporeditev oznak po grafu izboljšajo kakovost naučenega modela pri enakem proračunu označevanja. Primerjamo naključno razporeditev, razporeditev po vozliščih z visoko stopnjo (*hub-first*), nizko stopnjo (*outlier-first*), ter razporeditev, skoncentrirano znotraj gosto povezanih skupnosti (*community-first*).

2 Metodologija

Kot osnovno arhitekturo GNN v eksperimentih uporabljamo grafovsko konvolucijsko mrežo (angl. *graph convolutional network*, GCN) [15], ki v vsakem koraku agregira značilke sosednjih vozlišč s simetričnim normaliziranim povzetkom in jih transformira z naučeno utežno matriko.

2.1 Vzorčenje sosedov po vzoru GraphSAGE

Učenje GNN na celotnem grafu (*full-batch*) postane pri velikih grafih računsko in pomnilniško neobvladljivo. Namesto tega uporabimo mini-paketno učenje po vzoru GraphSAGE [16]: za vsak paket ciljnih vozlišč vzorčimo le del njihove L -skokov oddaljene okolice, pri čemer za vsako plast ℓ vnaprej določimo največje število vzorčenih sosedov (*fanout* f_ℓ). Za validacijo in testiranje uporabimo $f_\ell = \infty$, kar zagotovi natančen, nestohastičen izračun napovedi.

2.2 Eksperimentalna postavitve

2.2.1 Delitev podatkov

Za vsako podatkovno množico ciljna (klasificirana) vozlišča razdelimo na učno, validacijsko in testno množico s stratificiranim vzorčenjem, ki ohranja deleže razredov znotraj vsake množice enake deležem v celotni ciljni populaciji vozlišč. V vseh eksperimentih uporabimo razmerje delitve 6 : 1 : 1 (učna : validacijska : testna množica), kar ustreza približno 75 % / 12,5 % / 12,5 % ciljnih vozlišč. Za zagotavljanje ponovljivosti rezultatov pri vseh eksperimentih, ki zahtevajo večkratne ponovitve, uporabimo enak nabor petih naključnih semen. Stratificirano vzorčenje je kontrolna predpostavka za namen

primerljivosti rezultatov; v praksi anotator porazdelitev razredov v celotni populaciji pogosto ne pozna vnaprej.

2.2.2 Optimizacija in učenje

Modele učimo z optimizatorjem Adam, z zgodnjo ustavitvijo (*early stopping*) glede na izbrano metriko na validacijski množici. Glavna metrika za zgodnjo ustavitvev in poročanje rezultatov je ROC-AUC (angl. Receiver Operating Characteristic-Area Under Curve), saj je robustna na neuravnoteženost razredov, ki je pri več naših podatkovnih množicah prisotna.

2.2.3 Iskanje hiperparametrov

Hiperparametre za vsako podatkovno množico poiščemo z orodjem Optuna [17], ki uporablja vzorčevalnik TPE (*Tree-structured Parzen Estimator*) za usmerjeno preiskovanje prostora hiperparametrov ter rezalnik (angl. *pruner*) HyperBand za zgodnjo prekinitev neobetavnih konfiguracij. Iskanje izvedemo pri 20-odstotnem deležu označenih vozlišč; najboljša konfiguracija (glede na ROC-AUC na validacijski množici) se nato uporabi v vseh nadaljnjih eksperimentih za to podatkovno množico.

3 Rezultati

3.1 Podatkovne množice

Za izvedbo eksperimentov uporabimo nabor javno dostopnih primerjalnih podatkovnih množic, ki zajemajo tako homogene kot heterogene grafe ter razpon od visoko homofilnih do heterofilnih grafov (Tabela 1).

Pri heterogeni množici RCDD (angl. *Risk Commodity Detection Dataset*) [23] vse tipe vozlišč sploščimo v homogen graf z dodanim *one-hot* kodiranjem tipa vozlišča. Oznake imajo le vozlišča tipa *item*; preostala vozlišča le prispevajo k širjenju informacije po grafu, niso pa del učnih ali testnih množic.

3.2 Eksperimenti

V raziskavi izvedemo preliminarno preiskovanje prostora optimalnih hiperparametrov in tri eksperimente na naslednjem redosledu:

Iskanje hiperparametrov: namen tega eksperimenta je poiskati optimalne hiperparametre modela za vsako podatkovno množico, ki jih nato uporabimo kot fiksne v vseh nadaljnjih treh eksperimentih, opisanih v nadaljevanju. Iskanje izvedemo z orodjem Optuna, kot je opisano v poglavju 2.2, pri fiksnem 20-odstotnem deležu označenih vozlišč in fiksnem naključnim semenom. Preiskovani prostor hiperparametrov vključuje število in širino skritih plasti, stopnjo osipa, hitrost učenja in utežni razpad.

Rezultati so povzeti v tabeli 2. Vsebuje najboljše dosežene mere ROC-AUC, PRC-AUC (ploščina pod krivuljo preciznost-priklic, angl. *area under the precision-recall curve*) in natančnost (angl. *Accuracy*) na testni množici pri 20-odstotnem deležu označenih vozlišč.

Tabela 1: Pregled uporabljenih podatkovnih množic. H = homofilen, HE = heterofilen, hom. = homogen graf, het. = heterogen graf, C = število razredov, F = število vhodnih značilk.

Množica	Domena	Graf	Tip	$ V $	$ E $	C	Naloga	F	Značilke
Cora_ML [18]	citatno omrežje	hom.	H	2 995	8 158	7	tema članka	2 879	BoW ključnih besed
CiteSeer [18]	citatno omrežje	hom.	H	4 230	5 337	6	tema članka	602	BoW ključnih besed
DBLP [18]	citatno omrežje	hom.	H	17 716	52 867	4	področje članka	1 639	BoW ključnih besed
PubMed [18]	citatno omrežje	hom.	H	19 717	44 324	3	tip sladkorne bolezni	500	TF-IDF izvlečkov
Coaut./CS [19]	soavtorstvo	hom.	H	18 333	81 894	15	področje razisk.	6 805	BoW objav
Coaut./Physics [19]	soavtorstvo	hom.	H	34 493	247 962	5	pod-disciplina	8 415	BoW objav
Amazon/Comp. [19]	e-trgovina	hom.	H	13 752	245 861	10	kat. izdelka	767	BoW ocen kupcev
Amazon/Photo [19]	e-trgovina	hom.	H	7 650	119 081	8	kat. izdelka	745	BoW ocen kupcev
WikiCS [20]	Wikipedia	hom.	H	11 701	215 863	10	področje CS članka	300	GloVe vložitve besedila
GitHub [21]	soc. omrežje	hom.	H	37 700	144 502	2	ML / splet. razv.	4 005	repozitoriji, delodajalec
Tolokers [22]	množičenje	hom.	HE	11 758	519 000	2	kakovost dela	10	lastnosti delavca
RCDD [23]	odkr. prevar	het.	HE	13 806 619	157 814 864	2	tvegano blago	256 + 7	naučene vložitve + tip vozlišča

Tabela 2: Najboljši rezultati iskanja hiperparametrov (eksperiment 0), 20 % označenih vozlišč, model GCN. Nev. = število nevronov na skrito plast, Pl. = število skritih plasti.

Množica	ROC	PRC	Acc	Nevr.	Pl.
Cora_ML	0,9570	0,8564	0,7916	128	4
CiteSeer	0,9597	0,8873	0,8353	256	4
DBLP	0,9285	0,7823	0,8013	256	4
PubMed	0,9479	0,9034	0,8350	256	2
Coaut./CS	0,9929	0,9304	0,9108	256	2
Coaut./Physics	0,9904	0,9633	0,9393	64	2
Amazon/Comp.	0,9794	0,8536	0,8411	128	2
Amazon/Photo	0,9878	0,9436	0,9169	128	2
WikiCS	0,9579	0,7941	0,7699	128	2
GitHub	0,8583	0,7108	0,8301	64	3
Tolokers	0,7608	0,4408	0,7929	256	4
RCDD	0,9625	0,8131	0,9610	128	3

Eksperiment 1: Razmerje med deležem oznak in zmogljivostjo: s tem eksperimentom raziščemo, kako se napovedna zmogljivost modela spreminja z deležem označenih vozlišč. Za vsako podatkovno množico uporabimo predhodno poiskane optimalne hiperparametre. Delež označenih vozlišč variiramo med vrednostmi {1 %, 5 %, 10 %, 20 %, 80 %}, pri čemer slednjo vrednost obravnavamo kot zgornjo mejo oz. osnovo. Za vsak delež izvedemo pet neodvisnih ponovitev z naključnimi semeni. Poročamo povprečne vrednosti mer ROC-AUC, PRC-AUC in točnosti čez pet ponovitev.

Tabela 3 prikazuje povprečno uspešnost napovednih modelov glede na velikost učne množice oz. količino označenih podatkov. Na Sliki 1 grafično prikažemo relativno uspešnost glede na v celoti označeno podatkovno množico; zgornji graf prikazuje ROC_r na posameznih množicah, spodnji pa povprečne vrednosti mer ROC_r , PRC_r in Acc_r čez vse primerjalne množice.

Tabela 3: Povprečni relativni faktorji zmogljivosti GCN glede na osnovo pri 80 % označenih vozlišč (eksperiment 1).

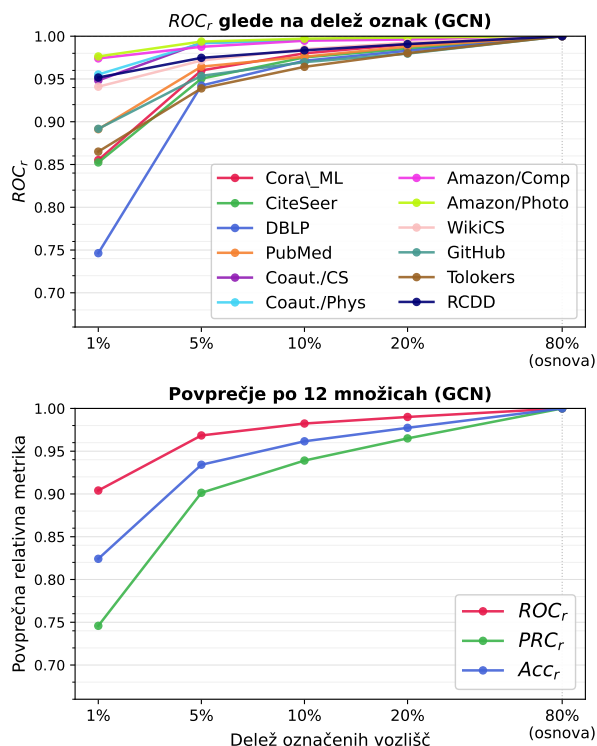
Delež oznak	ROC_r	PRC_r	Acc_r
1 %	$0,904 \pm 0,065$	$0,746 \pm 0,104$	$0,824 \pm 0,120$
5 %	$0,968 \pm 0,019$	$0,901 \pm 0,044$	$0,934 \pm 0,044$
10 %	$0,982 \pm 0,011$	$0,939 \pm 0,033$	$0,961 \pm 0,023$
20 %	$0,990 \pm 0,007$	$0,965 \pm 0,020$	$0,977 \pm 0,015$
80 %	1,000	1,000	1,000

Eksperiment 2: Trdo, abstinenčno in mehko označevanje: v praksi strokovnjak (anotator) pogosto ne more z gotovostjo določiti pravilnega razreda vozlišča, ne da bi za to porabil bistveno več časa. Ta eksperiment primerja tri strategije, kako naj anotator ravna v takih primerih, pri fiksnem 1-odstotnem ali 10-odstotnem deležu označenih vozlišč.

Anotatorja modeliramo kot šumni klasifikator, čigar prepričanje o razredu posameznega vozlišča je vzorec iz Dirichletove porazdelitve, ki je usmerjena proti pravemu razredu c z intenziteto $s \geq 0$:

$$\alpha_c = \alpha_0 + s, \quad \alpha_{c'} = \alpha_0 \quad (c' \neq c), \quad \mathbf{p} \sim \text{Dirichlet}(\boldsymbol{\alpha}) \quad (1)$$

kjer je α_0 bazna konstanta. V vseh eksperimentih uporabimo $\alpha_0 = 1$. Pri $s = 0$ je anotator brez diskriminativne moči (naključno ugibanje), pri $s \rightarrow \infty$ pa se prepričanje sesede v popoln (pravi) razred. Kakovost anotatorja izrazimo kot makro ROC-AUC vrednost ocenjenih verjetnosti $\mathbf{p}$ glede na prave razrede. Uporabljamo enako mero uspešnosti, kot jo uporabljamo za vrednotenje modela samega, kar omogoča neposredno primerjavo modela z anotatorjem. Vrednost s za ciljno vrednost ROC-AUC poiščemo z dvojiškim iskanjem na podlagi simulacije metode Monte Carlo.



Slika 1: Relativni ROC-AUC (ROC_r) glede na delež označenih vozlišč (GCN, 12 množic). Zgoraj: posamezne podatkovne množice. Spodaj: povprečje mer ROC_r , PRC_r in Acc_r . Vrednosti so normirane glede na osnovo pri 80 % označenih vozlišč.

Primerjane strategije: vse tri strategije izhajajo iz istega prepričanja anotatorja $\mathbf{p}$ pri dani vrednosti ROC-AUC; razlikujejo se le v tem, koliko tega prepričanja je razkritega modelu:

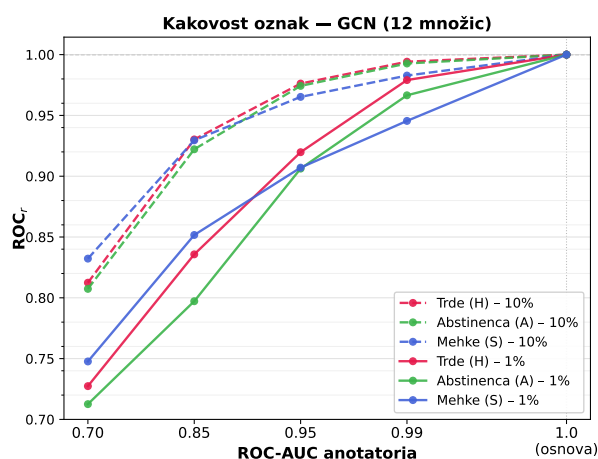
- **Trdo označevanje:** modelu razkrijemo le $\arg \max(\mathbf{p})$; anotator je prisiljen izbrati en razred, tudi če je negotov.
- **Abstinenca:** anotator odgovori le, kadar je njegovo lastno prepričanje dovolj visoko, tj. kadar velja $\max(\mathbf{p}) \geq \frac{1}{C} + \delta$, kjer je $\frac{1}{C}$ verjetnost naključnega ugibanja in δ fiksna marža (privzeto $\delta = 0,20$). Pri preostalih vozliščih se vzdrži, ta vozlišča so torej izključena iz učenja.
- **Mehko označevanje:** modelu razkrijemo celotno porazdelitev $\mathbf{p}$ kot ciljno vrednost.

Funkcija izgube: za mehke oznake uporabimo posplošeno križno entropijo:

$$\mathcal{L} = - \sum_{c=1}^C p_c \log \hat{y}_c \quad (2)$$

kjer je $\hat{y}_c = \text{softmax}(\mathbf{z})_c$ napovedana verjetnost razreda c , p_c pa pripadajoča komponenta ciljne porazdelitve. Za trde oznake se izraz (2) numerično ujema s standardno križno entropijo.

Tabeli 4 in 5 primerjata tri strategije označevanja pri različnih vrednostih ROC-AUC simuliranega anotatorja, pri fiksnem 1-odstotnem oz. 10-odstotnem deležu označenih vozlišč. Pri obeh deležih označenih vozlišč se najuspešnejši pristopi razlikujejo med podatkovnimi množicami. Na ravni povprečja več množic lahko povzamemo, da je v primeru suverena anotatorja ($ROC\text{-}AUC \geq 0,95$) bolje podati trde oznake (H), v nasprotnem primeru težavnejšega anotiranja pa mehke oznake (S). Izpuščanje oznak (A) pod pragom prepričanja, se tako pri 1 % kot tudi 10 %, izkaže za manj uspešno. Povprečni rezultati čez primerjalne množice so grafično prikazani na Sliki 2.



Slika 2: Povprečni relativni ROC-AUC (ROC_r) glede na ROC-AUC simuliranega anotatorja (GCN, 12 množic). Vrednosti so normirane glede na popolnega anotatorja (osnova = 1). Črtkano: 10 % označenih vozlišč. Polne linije: 1 % označenih vozlišč.

Tabela 4: Eksperiment 2 — kakovost oznak, GCN, 1 % označenih vozlišč. Povprečni $ROC_r \pm \sigma$, $PRC_r \pm \sigma$, $Acc_r \pm \sigma$ čez 12 množic glede na osnovo pri popolnem anotatorju. Krepko: najboljša strategija pri danem anotatorju in metriki.

Strategija	Anotator ROC	ROC_r	PRC_r	Acc_r
(H) Trde oznake	0,70	$0,73 \pm 0,09$	$0,45 \pm 0,20$	$0,52 \pm 0,24$
	0,85	$0,84 \pm 0,07$	$0,61 \pm 0,18$	$0,68 \pm 0,18$
	0,95	$0,92 \pm 0,04$	$0,78 \pm 0,12$	$0,84 \pm 0,10$
	0,99	$0,98 \pm 0,01$	$0,93 \pm 0,04$	$0,95 \pm 0,05$
(A) Abstinenca	0,70	$0,71 \pm 0,11$	$0,45 \pm 0,20$	$0,51 \pm 0,27$
	0,85	$0,80 \pm 0,12$	$0,58 \pm 0,23$	$0,63 \pm 0,28$
	0,95	$0,91 \pm 0,08$	$0,77 \pm 0,19$	$0,80 \pm 0,16$
	0,99	$0,97 \pm 0,06$	$0,90 \pm 0,13$	$0,95 \pm 0,13$
(S) Mehke oznake	0,70	$0,75 \pm 0,09$	$0,46 \pm 0,20$	$0,55 \pm 0,26$
	0,85	$0,85 \pm 0,06$	$0,62 \pm 0,18$	$0,68 \pm 0,18$
	0,95	$0,91 \pm 0,05$	$0,74 \pm 0,15$	$0,80 \pm 0,12$
	0,99	$0,95 \pm 0,04$	$0,84 \pm 0,13$	$0,90 \pm 0,11$

Eksperiment 3: Strategije razporeditve oznak: zadnji eksperiment raziskuje, ali lahko heuristike za izbiro katerih vozlišč ekspertnega označevanja (pri enakem proračunu in vseh oznakah trdih in pravilnih) izboljšajo

Tabela 5: Eksperiment 2 — kakovost oznak, GCN, 10 % označenih vozlišč. Povprečni $ROC_r \pm \sigma$, $PRC_r \pm \sigma$, $Acc_r \pm \sigma$ čez 12 množic glede na osnovo pri popolnem anotatorju. Krepko: najboljša strategija pri danem anotatorju in metriki.

Strategija	Anotator ROC	ROC_r	PRC_r	Acc_r
(H) Trde oznake	0,70	$0,81 \pm 0,07$	$0,54 \pm 0,20$	$0,62 \pm 0,19$
	0,85	$0,93 \pm 0,03$	$0,77 \pm 0,13$	$0,85 \pm 0,11$
	0,95	$0,98 \pm 0,01$	$0,92 \pm 0,04$	$0,95 \pm 0,04$
	0,99	$0,99 \pm 0,00$	$0,98 \pm 0,01$	$0,98 \pm 0,01$
(A) Abstinenca	0,70	$0,81 \pm 0,10$	$0,55 \pm 0,23$	$0,64 \pm 0,24$
	0,85	$0,92 \pm 0,06$	$0,77 \pm 0,16$	$0,83 \pm 0,15$
	0,95	$0,97 \pm 0,02$	$0,92 \pm 0,07$	$0,94 \pm 0,05$
	0,99	$0,99 \pm 0,00$	$0,97 \pm 0,03$	$0,98 \pm 0,02$
(S) Mehke oznake	0,70	$0,83 \pm 0,07$	$0,56 \pm 0,19$	$0,66 \pm 0,20$
	0,85	$0,93 \pm 0,04$	$0,78 \pm 0,15$	$0,84 \pm 0,12$
	0,95	$0,97 \pm 0,02$	$0,89 \pm 0,10$	$0,93 \pm 0,07$
	0,99	$0,98 \pm 0,01$	$0,95 \pm 0,06$	$0,96 \pm 0,04$

napovedno zmogljivost modela. Proračun oznak je za vsako strategijo dodeljen po pravih razredih ciljnih vozlišč; gre za nadzorovano eksperimentalno predpostavko, uvedeno z namenom zagotavljanja primerljivosti strategij razporeditve, saj v praktični anotacijski nalogi pravi razredi vozlišč pred označevanjem praviloma niso znani. Primerjamo štiri strategije pri dveh proračunih, 1 % in 10 % ciljnih vozlišč:

- **Naključna (random):** enakomerno naključno vzorčenje znotraj vsakega razreda.
- **Vozlišča z visoko stopnjo (hub-first):** prednost imajo vozlišča z visoko vrednostjo PageRank.
- **Vozlišča z nizko stopnjo (outlier-first):** prednost imajo vozlišča z nizko vrednostjo PageRank.
- **Skoncentrirano (community-first):** namesto razpršitve oznak po grafu jih anotator skoncentrira v tesno območje.

Ne glede na strategijo izbire vsaka strategija prejme enak proračun oznak na razred (izračunan enkrat, sorazmerno z deležem posameznega razreda med ciljnim vozlišči), tako da se strategije razlikujejo izključno po tem, katera vozlišča znotraj posameznega razreda izberejo, ne pa tudi po tem, koliko jih izberejo. To zagotavlja, da so razlike v končni napovedni zmogljivosti pripisljive izključno strategiji razporeditve, ne naključnim razlikam v ravnoesju razredov. Za vsako strategijo in proračun izvedemo pet ponovitev z naključnimi semeni $\{0, 1, 2, 3, 4\}$, kjer naključnost vpliva na izbiro znotraj strategije (npr. katero vozlišče je izhodišče pri širjenju, katere skupnosti so izbrane).

Tabela 6 primerja tri strategije razporeditve oznak pri proračunih 1 % in 10 %. Pri pregledu rezultatov s proračunom 10 % opazimo, da izbira vozlišč bistveno ne vpliva na uspešnost naučenih modelov. Pri 1-odstotnem proračunu doseže strategija *hub-first* izboljšavo, *community-first* pa poslabšanje napovedne

uspešnosti. Uspešnost prve pripisujemo izbiri vozlišč z gosto povezano okolico, ki privede do raznolikih predstavitev vozlišč v mini-paketnem učenju. Po drugi strani, neuspeh *community-first* pripisujemo ponavljanju istih predstavitev zaradi skoncentriranosti oznak v "raziskanem" delu grafa.

Tabela 6: Eksperiment 3 — strategije razporeditve oznak, GCN. Povprečni faktor glede na naključno razporeditev čez 12 množic ($\pm \sigma$). **Modro** $> 1,005$, **črno** $0,99 \leq x < 1,005$, **rdeče** $< 0,99$.

Delež oznak	Strategija	ROC_r	PRC_r	Acc_r
1 %	Hub	$1,017 \pm 0,03$	$1,030 \pm 0,07$	$1,065 \pm 0,12$
	Outlier	$1,001 \pm 0,03$	$0,992 \pm 0,07$	$1,026 \pm 0,08$
	Community	$0,971 \pm 0,05$	$0,931 \pm 0,08$	$0,943 \pm 0,12$
10 %	Hub	$1,000 \pm 0,00$	$1,000 \pm 0,00$	$1,014 \pm 0,04$
	Outlier	$1,000 \pm 0,00$	$1,001 \pm 0,01$	$1,010 \pm 0,02$
	Community	$0,999 \pm 0,00$	$1,000 \pm 0,01$	$1,012 \pm 0,04$

4 Zaključek

V prispevku smo raziskali, kako razporeditev omejenega proračuna za označevanje vpliva na napovedno zmogljivost grafovskih nevronske mreže v polnadzorovanem okolju.

Pri označevanju s trdimi, abstinenčnimi in mehki oznakami ugotovimo, da so mehke oznake najprimernejše za težavne anotacijske naloge ter trde za preprostejše naloge z višjo verodostojnostjo. Abstinenca oz. izpust oznak pod pragom se je izkazala za manj uspešen pristop.

Z uporabo hevristik pri izbiri vozlišč za anotacijo so največji pridobitki (kot tudi izgube) napovedne uspešnosti mogoči pri nizkem anotacijskem proračunu. Za uspešno hevristiko se izkaže prioritizacija vozlišč z visoko vrednostjo PageRank, ki opazno pripomore k raznolikosti predstavitev vozlišč v polnadzorovanem učenju.

Literatura

- [1] Petar Veličković. Everything is connected: Graph neural networks. *Current Opinion in Structural Biology*, 79:102538, 2023.
- [2] Fan Liang, Cheng Qian, Wei Yu, David Griffith, and Nada Golmie. Survey of graph neural networks and applications. *Wireless Communications and Mobile Computing*, 2022(1):9261537, 2022.
- [3] Dawei Cheng, Yao Zou, Sheng Xiang, and Changjun Jiang. Graph neural networks for financial fraud detection: a review. *Frontiers of Computer Science*, 19(9):199609, 2025.
- [4] Ludivia Hernandez Aros, Luisa Ximena Bustamante Molano, Fernando Gutierrez-Portela, John Johver Moreno Hernandez, and Mario Samuel Rodríguez Barrero. Financial fraud detection through the application of machine learning techniques: a literature review. *Humanities and Social Sciences Communications*, 11(1):1–22, 2024.

- [5] Zixing Song, Xiangli Yang, Zenglin Xu, and Irwin King. Graph-based semi-supervised learning: A comprehensive review. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8174–8194, 2022.
- [6] Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. Active learning for graph embedding. *arXiv preprint arXiv:1705.05085*, 2017.
- [7] Li Gao, Hong Yang, Chuan Zhou, Jia Wu, Shirui Pan, and Yue Hu. Active discriminative network representation learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2142–2148, 2018.
- [8] Jiaqi Ma, Ziqiao Ma, Joyce Chai, and Qiaozhu Mei. Partition-based active learning for graph neural networks. *arXiv preprint arXiv:2201.09391*, 2022.
- [9] Wentao Zhang, Zhi Yang, Yexin Wang, Yu Shen, Yang Li, Liang Wang, and Bin Cui. Grain: Improving data efficiency of graph neural networks via diversified influence maximization. *Proceedings of the VLDB Endowment*, 14(11):2473–2482, 2021.
- [10] Songbai Yan, Kamalika Chaudhuri, and Tara Javidi. Active learning from noisy and abstention feedback. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2015.
- [11] Chicheng Zhang and Kamalika Chaudhuri. Active learning from weak and strong labelers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2015.
- [12] Netta Shafir, Guy Hacohen, and Daphna Weinshall. Active learning with a noisy annotator. *arXiv preprint arXiv:2504.04506*, 2025.
- [13] Wentao Zhang, Yexin Wang, Zhenbang You, Meng Cao, Ping Huang, Jiulong Shan, Zhi Yang, and Bin Cui. Rim: Reliable influence-based active learning on graphs. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.
- [14] Wentao Zhang, Yexin Wang, Zhenbang You, Meng Cao, Ping Huang, Jiulong Shan, Zhi Yang, and Bin Cui. Information gain propagation: A new way to graph active learning with soft labels. *arXiv preprint arXiv:2203.01093*, 2022.
- [15] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017.
- [16] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [17] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.
- [18] Aleksandar Bojchevski and Stephan Günnemann. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking, 2018.
- [19] Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation, 2019.
- [20] Péter Mernyei and Cătălina Cangea. Wiki-cs: A wikipedia-based benchmark for graph neural networks, 2022.
- [21] Benedek Rozemberczki, Carl Allen, and Rik Sarkar. Multi-scale attributed node embedding, 2021.
- [22] Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of gnns under heterophily: Are we really making progress?, 2024.
- [23] Yijian Liu, Hongyi Zhang, Cheng Yang, Ao Li, Yugang Ji, Luhao Zhang, Tao Li, Jinyu Yang, Tianyu Zhao, Juan Yang, et al. Datasets and interfaces for benchmarking heterogeneous graph neural networks. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pages 5346–5350, 2023.