FairGAT

Fairness-Aware Graph Attention Networks

15/10/2024

About

- O. Deniz Kose, Yanning Shen
- University of California, Irvine
- ACM Transactions on Knowledge Discovery from Data, Vol. 18, Issue 7
- August 2024
- https://doi.org/10.1145/3645096

Introduction

From GNN to GAT

- Nodes used to contribute equally to embeddings
- GCN create node's weight based on its degree
- Better ways to represent neighborhood importance?
- Graph Attention Networks (Veličković et al. 2018)

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(W_2[WH_i^{k-1} \mid\mid WH_j^{k-1}]))}{\sum_{l \in N(i)} \exp(\text{LeakyReLU}(W_2[WH_i^{k-1} \mid WH_l^{k-1}]))} \qquad H_i^k = \sigma\left(\sum_{j \in N(i)} \alpha_{ij} WH_j^{k-1}\right)$$

Problem

Algorithmic Bias

- Disparity in results when a *sensitive class* is changed
- GNN can propagate and amplify Bias
- No relevant studies on Bias for GATs
- Context of node classification
- Need for expressing bias boundaries

Preliminaries

Some notation

Graph $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} := \{v_1, \ldots, v_n\}$ and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$

Adjacency matrix: $A \in \{0,1\}^{N \times N}$

Nodes feature matrix: $X \in \mathbb{R}^{N \times F}$ where F is the dim of the features

Sensitive attributes: $\mathbf{s} \in \{0,1\}^N$ a single binary attribute for each node

$\mathcal{S}_0, \mathcal{S}_1$ are the set of nodes with sensitive attribute 0 and 1 respectively

Preliminaries

Some notation

Inter-edge set: $\mathscr{E}^\chi := \{e_{ij} \mid v_i \in \mathscr{S}_a, v_j \in \mathscr{S}_b, a \neq b\}$

Intra-edge set: $\mathscr{E}^\omega := \{e_{ij} \mid v_i \in \mathscr{S}_a, v_j \in \mathscr{S}_b, a = b\}$

Set of nodes containing at least one inter-edge: $\mathscr{S}^\chi := \{v_i \mid \exists e_{ij} \in \mathscr{E}^\chi\}$

Set of nodes containing only intra-edges: $\mathscr{S}^\omega := \{v_i \mid \forall e_{ij} \in \mathscr{E}^\omega\}$

$\mathscr{S}_0^\chi := \mathscr{S}_0 \cap \mathscr{S}^\chi$ $\qquad$ $\mathscr{S}_1^\chi := \mathscr{S}_1 \cap \mathscr{S}^\chi$

Preliminaries

Attention weight for sensitive classes

Given $\mathcal{S}_i, \mathcal{S}_j$ such that $i \neq j$, and $v_k \in \mathcal{S}_j$

Attention weight assigned to neighbors of another sensitive group $\alpha_k^\chi := \sum_{a \in \mathcal{N}(k) \cap S_i} \alpha_{ka}$

not feasible to calculate for each v_k

The attention weight of a node from a certain $\mathcal{S}_i$ is shared. i.e., $\alpha^\chi := \alpha_k^\chi$

(not to be confused with the actual attention given from a node v_i to a node v_j; α_{ij})

Preliminaries

Classification disparity

$$\delta_{\hat{y}} := \left\| \text{mean}(\hat{y}_j \mid s_j = 0) - \text{mean}(\hat{y}_j \mid s_j = 1) \right\|_2$$

where $\hat{y}_j$ is the soft label prediction for node v_j (predicted *pdf*), while the mean($\cdot$, $\cdot$) function gets the sample mean value for the distribution

($\hat{c}_j$, the hard label prediction for node v_j will also be used)

Bias analysis

Axiom 1

Let the sample mean vectors for $\mathcal{S}_s$ with $s \in \{0,1\}$ be:

- $\bar{\mathbf{z}}_s^{l+1}$:= mean($\mathbf{z}_j^{l+1} \mid v_j \in \mathcal{S}_s$), where $\mathbf{z}_i^{l+1} := \sum_{j \in \mathcal{N}_i} \alpha_{ij}^l \mathbf{c}_j^{l+1}$ is the aggregation for node v_i at layer $l+1$
- $\bar{\mathbf{c}}_s^{l+1}$:= mean($\mathbf{c}_j^{l+1} \mid v_j \in \mathcal{S}_s$), where $\mathbf{c}_i^{l+1} := \mathbf{W}^l \mathbf{h}_i^l$

and let max($\cdot$, $\cdot$) outputs the element-wise maximum of the input vectors

1. $\| c_j^{l+1} - \bar{c}_s^{l+1} \|_\infty \leq (\Delta_c^{(s)})^{l+1}, \forall v_j \in \mathcal{S}_s$ with $s \in \{0,1\}$ where $\Delta_c^{l+1} = \max((\Delta_c^{(0)})^{l+1}, (\Delta_c^{(1)})^{l+1})$
2. $\| z_j^{l+1} - \bar{z}_s^{l+1} \|_\infty \leq (\Delta_z^{(s)})^{l+1}, \forall v_j \in \mathcal{S}_s$ with $s \in \{0,1\}$ where $\Delta_z^{l+1} = \max((\Delta_z^{(0)})^{l+1}, (\Delta_z^{(1)})^{l+1})$

where the delta terms correspond to the maximum deviation value at the l + 1 layer **between each sensitive class**

Bias analysis

Theorem

THEOREM 4.1. *The disparity between the representations of different sensitive groups that are output by the lth GAT layer, δ_h^{l+1}, can be upper bounded by*

$$\delta_h^{l+1} \leq L\Big(\underline{\sigma_{max}(\mathbf{W}^l)}\big|(R_1^{\chi}\underline{\alpha^{\chi}} + R_0^{\chi}\underline{\alpha^{\chi}} - 1)\big|\delta_h^l + 2\sqrt{N}\underline{\Delta_c^{l+1}} + 2\sqrt{N}\underline{\Delta_z^{l+1}}\Big), \tag{4}$$

where L is the Lipschitz constant of the utilized nonlinear activation, $\sigma_{max}(\cdot)$ denotes the largest singular value of the input matrix, and $R_1^{\chi} := \frac{|\mathcal{S}_1^{\chi}|}{|\mathcal{S}_1|}, R_0^{\chi} := \frac{|\mathcal{S}_0^{\chi}|}{|\mathcal{S}_0|}$.

Based on the above, a GAT architecture proposal is made

FairGAT

Objective: minimize the implied terms at the disparity upper bound

---

ALGORITHM 1: FairGAT

Data: $\mathcal{G} := (\mathcal{V}, \mathcal{E})$, $\mathbf{X}$, $\mathbf{s}$, α^{χ}_{max}, η

Result: $\hat{\mathbf{y}}$

S1. Employ fair attention learning described in Equation (10) at every attention layer.

S2. Apply spectral normalization to $\mathbf{W}^l$ at every layer l to ensure that $\sigma_{max}(\mathrm{SN}(\mathbf{W}^l)) = 1$.

S3. Scale representations $\mathbf{Z}^{l+1}$ at every l and $\mathbf{C}^{l+1}$ at every attention layer l by a factor η.

FairGAT

Fair Attention Learning

$$(\alpha^{\chi})^* = \min_{\alpha^{\chi}} \quad |R_1^{\chi}\alpha^{\chi} + R_0^{\chi}\alpha^{\chi} - 1| \quad \text{s.t.} \quad 0 \le \alpha^{\chi} \le \alpha^{\chi}_{max}. \longrightarrow (\alpha^{\chi})^* = \begin{cases} \alpha^{\chi}_{max}, & \text{if } R_1^{\chi} + R_0^{\chi} < \frac{1}{\alpha^{\chi}_{max}}, \\ \frac{1}{R_1^{\chi}+R_0^{\chi}}, & \text{else.} \end{cases}$$

where α_{max} is an hyperparameter limit less or equal to 1

(1) $e\left(\boldsymbol{h}_i^l, \boldsymbol{h}_j^l\right) = \mathrm{LReLU}\left((\boldsymbol{a}^l)^\top \cdot \left[\boldsymbol{W}^l \boldsymbol{h}_i^l \| \boldsymbol{W}^l \boldsymbol{h}_j^l\right]\right),$

(2)

$$\alpha_{ij}^l = \begin{cases} (\alpha^{\chi})^* \frac{\exp\left(e\left(\boldsymbol{h}_i^l, \boldsymbol{h}_j^l\right)\right)}{\sum_{j' \in \mathcal{N}_i \cap \mathcal{S}_q} \exp\left(e\left(\boldsymbol{h}_i^l, \boldsymbol{h}_{j'}^l\right)\right)}, & \text{if } v_i \in \mathcal{S}_p, v_j \in \mathcal{S}_q \text{ and } p \neq q \\ (\alpha^{\omega})^* \frac{\exp\left(e\left(\boldsymbol{h}_i^l, \boldsymbol{h}_j^l\right)\right)}{\sum_{j' \in \mathcal{N}_i \cap \mathcal{S}_q} \exp\left(e\left(\boldsymbol{h}_i^l, \boldsymbol{h}_{j'}^l\right)\right)}, & \text{if } v_i \in \mathcal{S}_p, v_j \in \mathcal{S}_q \text{ and } p = q \end{cases} \quad (10)$$

(3) $\boldsymbol{h}_i^{l+1} = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^l \cdot \boldsymbol{W}^l \boldsymbol{h}_j^l\right).$

FairGAT

Spectral normalization

The term W^l is normalized at every layer, $\sigma_{\text{max}}(\text{SN}(\mathbf{W}^l)) = 1$

The largest singular value, also known as the *spectral norm*, of a matrix $\mathbf{W} \in \mathbb{R}^{F_1 \times F_2}$ equals $\sigma_{max}(\mathbf{W}) = \max_{\boldsymbol{\xi} \in \mathbb{R}^{F_1}, \boldsymbol{\xi} \neq \mathbf{0}} \frac{\|\mathbf{W}\boldsymbol{\xi}\|_2}{\|\boldsymbol{\xi}\|_2}$. Consider the input–output relation $\hat{\mathbf{y}} = \sigma(\boldsymbol{W}\mathbf{h})$. Here, if a perturbation $\boldsymbol{\xi}$ is applied to the input, i.e., $\tilde{\mathbf{y}} = \sigma(\boldsymbol{W}(\mathbf{h} + \boldsymbol{\xi}))$, we have that

$$\frac{\|\tilde{y} - \hat{y}\|_2}{\|\xi\|_2} = \frac{\|\sigma(Wh) - \sigma(W(h+\xi))\|_2}{\|\xi\|_2} \leq \frac{L\|(W\mathbf{h}) - (W(\mathbf{h} + \boldsymbol{\xi}))\|_2}{\|\boldsymbol{\xi}\|_2} \leq L\frac{\|W\xi\|_2}{\|\xi\|_2} \leq L\sigma_{\text{max}}(W)$$

this also helps improve the robustness and generalizability of the model

FairGAT

Scaling representation

Maximal deviation influence disparity, so it get scaled by an hyperparameter η

$$\|\eta \mathbf{z}_j^{l+1} - \eta \mathbf{z}_s^{l+1}\|_\infty = \eta \|\mathbf{z}_j^{l+1} - \mathbf{z}_s^{l+1}\|_\infty \leq (\eta \Delta_z^{(s)})^{l+1}, \forall v_j \in S_s$$

η provides a trade-off between fairness and utility

Experimental Results

Table 1. Comparative Results

	Pokec-z			Pokec-n			Recidivism		
	Acc (%)	Δ_{SP} (%)	Δ_{EO} (%)	Acc (%)	Δ_{SP} (%)	Δ_{EO} (%)	Acc (%)	Δ_{SP} (%)	Δ_{EO} (%)
GAT	66.26 ± 0.9	3.63 ± 2.6	4.30 ± 2.4	67.50 ± 0.4	2.26 ± 2.8	3.56 ± 1.7	**95.63** ± 0.2	8.08 ± 1.7	2.09 ± 1.0
FairGNN	**67.71** ± 0.7	**2.27** ± 0.9	2.31 ± 1.0	65.81 ± 0.8	2.21 ± 1.4	2.97 ± 1.3	95.18 ± 0.2	7.31 ± 1.9	1.27 ± 1.0
EDITS	63.89 ± 0.7	3.27 ± 2.0	2.93 ± 2.2	63.47 ± 0.9	2.01 ± 1.5	2.48 ± 2.3	88.52 ± 0.6	**6.59** ± 2.1	1.73 ± 1.1
NIFTY	66.59 ± 0.8	4.21 ± 1.4	4.19 ± 2.7	**68.41** ± 1.5	1.41 ± 0.7	2.30 ± 1.8	88.52 ± 2.3	6.74 ± 2.4	1.48 ± 1.9
FairGAT	66.29 ± 0.6	2.55 ± 0.5	**1.63** ± 0.9	67.81 ± 1.1	**0.71** ± 0.7	**1.23** ± 0.6	94.93 ± 0.1	7.39 ± 1.8	**1.02** ± 0.05

Statistical Parity $\Delta_{SP} := \left| P(\hat{c}_j = 1 \mid s_j = 0) - P(\hat{c}_j = 1 \mid s_j = 1) \right|$

Equal Opportunity $\Delta_{EO} := \left| P(\hat{c}_j = 1 \mid y_j = 1, s_j = 0) - P(\hat{c}_j = 1 \mid y_j = 1, s_j = 1) \right|$

Lower is better

Conclusion

- Improved fairness metrics with similar accuracy than the rest of models.
- Maintained complexity, thus, utility.
- More work needs to be done on GNN interpretability and theoretical bias analysis.