Evaluating graph fairness in transductive learning

Fernanda Ribeiro[*1,2] FERNANDA.RIBEIRO@UQ.EDU.AU
[1] *School of Psychology, The University of Queensland, Australia*
[2] *Queensland Brain Institute, The University of Queensland, Australia*

Valentina Shumovskaia[*3] VALENTINA.SHUMOVSKAIA@EPFL.CH
[3] *School of Engineering, École Polytechnique Fédérale de Lausanne, Switzerland*

Thomas Davies[4] T.O.M.DAVIES@SOTON.AC.UK
[4] *Electronics and Computer Science, University of Southampton, UK*

Ira Ktena[5] IRAKTENA@DEEPMIND.COM
[5] *DeepMind, UK*

Abstract

Recent work on neuroimaging has demonstrated significant benefits of using population graphs to capture non-imaging information in the prediction of neurodegenerative and neurodevelopmental disorders. These non-imaging attributes may not only contain demographic information about the individuals, e.g. age or sex, but also the acquisition site, as imaging protocols might significantly differ across sites in large-scale studies. In addition, recent studies have highlighted the need to investigate potential biases in the classifiers devised using large-scale datasets, which might be imbalanced in terms of one or more sensitive attributes. This can be exacerbated when employing these attributes in a population graph to explicitly introduce inductive biases to the machine learning model and lead to disparate predictive performance across sub-populations. In this work, we explore the impact of stratification strategies and graph structures on the fairness of a semi-supervised classifier that relies on a population graph for the prediction of autism-spectrum disorder.

Keywords: Connectomics, graph representation learning, fairness, autism-spectrum disorder

1. Introduction

Issues related to fairness in healthcare decision-making have been the focus of intense scholarly debate (Seyyed-Kalantari et al., 2020; Wiens et al., 2019). Even though computer-aided diagnosis systems have integrated significant advances to assist clinicians in various tasks, these systems have rarely been scrutinised enough for their potential for discrimination against certain population subgroups. Disparate treatment concerns can arise solely due to the composition of the training data (Larrazabal et al., 2020), meaning that certain population subgroups might be underrepresented or completely missing during training of a machine learning system. In healthcare applications, the disease prevalence may vary across population subgroups (Werling and Geschwind, 2013), e.g. autism spectrum disorders (ASD) are more prevalent in males compared to females. Simultaneously, the clinical presentation of a disease might be completely different across subgroups. In ASD, in particular, differences have been established between neurodiverse males and females in terms of the interactions between key functional brain networks (Alaerts et al., 2016).

[*] Contributed equally

Despite progress in other domains, fairness issues are still under-explored for approaches that operate in irregular domains in a transductive setting. These have been shown to lead to significant performance improvements in neuroimaging tasks, like ASD and Alzheimer's disease prediction (Parisot et al., 2018), by employing semi-supervised learning on population graphs that leverage demographic or other auxiliary information. Such studies only report overall performance metrics, i.e. prediction accuracy and area under the receiver-operating characteristic (AUC-ROC) curve. Therefore, there is a limited understanding of whether these methods and training strategies inadvertently improve predictive performance in one subgroup of the population at the expense of another.

2. Methods and Results

Several fairness metrics have been proposed in the recent literature, but given the growing recognition that not all conditions can be simultaneously satisfied, we focus on the TPR gap (or equality of odds) as, arguably, the most relevant to the application scenario. We use the *ABIDE database* described in (Di Martino et al., 2014) – a consortium of several international acquisition sites comprising functional neuroimaging and phenotypic data from 871 participants, 403 neurodiverse and 468 neurotypical. The number of individuals participating in each acquisition site varies significantly. This database presents a particularly challenging setting in which we can explore the propensity of GNN models to be biased against underrepresented populations.

As defined in (Parisot et al., 2018), the phenotypic population graph is constructed by weighting the connectome similarity matrix with a phenotypic graph that captures the agreement of pairs of participants in terms of phenotypic features. We further adopt the graph convolutional model used in this work to predict ASD diagnosis with transductive learning. We consider four different graph structures to understand the impact of the population graph on the fairness of the target predictions: **(1)** a weighted graph based on the subjects' *sex* alone, **(2)** the *acquisition site* alone, **(3)** *both* sex and acquisition site, and **(4)** a *complete* graph that does not leverage phenotypic information.

In prior work, k-fold stratified cross-validation was used to evaluate the performance of the proposed method. However, stratification based on diagnosis can lead to a significantly imbalanced training set with respect to the sensitive attribute of interest, given that for certain sites no female participants were recruited. As previous studies have shown (Larrazabal et al., 2020; Puyol-Antón et al., 2021), the composition of the training data can significantly impact the bias of the devised classifier. Hence, the training data bias with respect to the sensitive attribute can be further accentuated by a stratification based solely on diagnosis, due to the demographic shift between acquisition sites. To test for the robustness of our GNN model to distribution shifts, we consider four different stratification strategies: **(1)** based on the *target variable* – diagnosis, **(2)** based on *diagnosis* and the *sensitive attribute* (i.e., sex), **(3)** based on *diagnosis* and the *acquisition site*, and **(4)** based on the *sensitive attribute* and the *acquisition site*. Figure 1 shows that stratification strategies did not significantly impact TPR differences, but graph structures did.

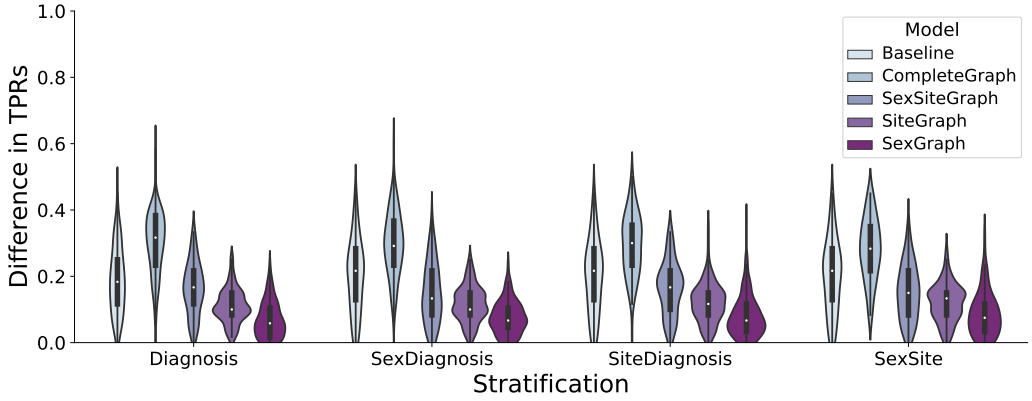

Figure 1: Absolute difference in true positive rates between males and females in the test set across graph structures and stratification strategies for a fixed held-out set.

References

K Alaerts, SP Swinnen, and N Wenderoth. Sex differences in autism: a resting-state fmri investigation of functional brain connectivity in males and females. *Social cognitive and affective neuroscience*, 11(6):1002–1016, 2016.

A Di Martino, CG Yan, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6): 659–667, 2014.

AJ Larrazabal, N Nieto, V Peterson, DH Milone, and E Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *PNAS*, 117(23):12592–12594, 2020.

S Parisot, SI Ktena, et al. Disease prediction using graph convolutional networks: application to autism spectrum disorder and alzheimer's disease. *Medical image analysis*, 48: 117–130, 2018.

E Puyol-Antón, B Ruijsink, et al. Fairness in cardiac MR image analysis: An investigation of bias due to data imbalance in deep learning based segmentation. In *MICCAI*, pages 413–423. Springer, 2021.

L Seyyed-Kalantari, G Liu, M McDermott, IY Chen, and M Ghassemi. CheXclusion: Fairness gaps in deep chest X-ray classifiers. In *BIOCOMPUTING 2021*, pages 232–243. World Scientific, 2020.

Donna M Werling and Daniel H Geschwind. Sex differences in autism spectrum disorders. *Current opinion in neurology*, 26(2):146, 2013.

J Wiens, S Saria, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340, 2019.