New machine learning model could remove bias from social network connections

March 12 2021

Credit: CC0 Public Domain

Did you ever wonder how social networking applications like Facebook and LinkedIn make recommendations on the people you should friend or pages you should follow?

Behind the scenes are machine learning models that classify nodes based on the data they contain about users—for example, their level of education, location or political affiliation. The models then use these classifications to recommend people and pages to each user. But there is significant bias in the recommendations made by these models—known as graph neural networks (GNNs)—as they rely on user features that are highly related to sensitive attributes such as gender or skin color.

Recognizing that the majority of users are reluctant to publicize their sensitive attributes, researchers at the Penn State College of Information

Sciences and Technology have developed a novel framework which estimates sensitive attributes to help GNNs make fair recommendations.

The team found that their [model](#), called FairGNN, maintains high performance on node [classification](#) using limited, user-supplied sensitive information, while at the same time reducing bias.

"It has been widely reported that people tend to build relationships with those sharing the same sensitive attributes such as ages and regions," said Enyan Dai, doctoral candidate in informatics and lead author on the research paper. "There are some existing machine learning models that aim to eliminate bias, but they require people's sensitive attributes to make them fair and accurate. We are proposing to apply another model based on the very few sensitive attributes that we have (and instead look at other provided information) which could provide us very good insight to give fair predictions toward sensitive attributes such as your gender and skin color."

The researchers trained their model with two real-world datasets: user profiles on Pokec, a popular social network in Slovakia, similar to Facebook and Twitter; and a dataset of approximately 400 NBA basketball players. In the Pokec dataset, they treated the region in which each user was from as the sensitive attribute, and set the classification task to predict the working field of the users. In the NBA data, they identified players as those in the U.S. and those overseas, using location as the sensitive attribute with the classification task to predict whether the salary of each player is over the median.

They then used the same datasets to test their model with other state-of-the-art methods for fair classification. First, they evaluated FairGNN in terms of fairness and classification performance. Then, they performed "ablation studies"—which remove certain components of the model to test the significance of each component to the overall system—to further

strengthen the model. They then tested whether FairGNN is effective when different amounts of sensitive attributes are provided in the training set.

"Our experiment shows that the classification performance doesn't decrease," said Suhang Wang, assistant professor of information sciences and technology and principal investigator on the project. "But in terms of fairness, we can make the model much more fair."

According to the researchers, their framework could make an impact for other real-world use cases.

"Our findings could be useful in applications, such as job applicant rankings, crime detection or in financial loan applications," said Wang. "But those are domains where we don't want to introduce bias. So we want to give accurate predictions while maintaining fairness."

Added Dai, "[If] this fair machine learning model could be introduced in these applications, we will have more fair data and this problem would be gradually dissolved."

Dai and Wang presented their work this week at the virtual ACM International Conference on Web Search and Data Mining.

Provided by Pennsylvania State University