Test-Fairness Deep Learning with Influence Score

Dr. Jacky Chung-Hao Wu

Institute of Statistics, National Yang Ming Chiao Tung University, Taiwan

Abstract

The prevalence of artificial intelligence has led to many significant problems that required new perspectives. Among them, the hidden bias in AI systems is a serious problem that should be carefully addressed. The source of bias usually comes from data collection focusing on heterogeneous and imbalanced groups. The data with bias will often generate the model with inhomogeneous discrimination that has diverse prediction performances for different groups. Therefore, it is crucial to eliminate the phenomenon of inhomogeneous discrimination and improve the fairness. To mitigate the inhomogeneous discrimination, we propose a new feature selection method based on deep learning and statistics to enhance the model fairness and preserve the prediction performance simultaneously. Integrating with deep learning models, we adopt the influence score (I-score), which develops the statistical methodology that can detect the interaction patterns between multiple features in the proposed method. The features related to the bias information will be detected and excluded in the fair model. We call this method the fair I-score method. The fair I-score method will explore the features unassociated with the discriminatory factors so that the resulting prediction performance is homogeneous for distinct groups. We conduct the empirical study on skin lesion datasets and show that the fair I-score method can produce a model that can correctly classify the types of skin lesions by eliminating the bias information inherent in the diverse groups. This report is based on the joint work with Professor Shaw-Hwa Lo in Columbia University, Professor Inchi Hu in George Mason University, Professor Henry Horng-Shing Lu and the related members in National Yang Ming Chiao Tung University.

Keywords: deep learning; fairness; influence score.