Fairness Learning in Semi-Supervised Setting

Jie Zhang
University of Utah

UUCS-18-008

School of Computing
University of Utah
Salt Lake City, UT 84112 USA

10 December 2018

Abstract

More and more often nowadays, decision systems that rule many aspects of our lives are trained by machine learning algorithms. The trained decision systems are biased in situations where only partial and biased data are available for training. Those biased decision systems will make biased decisions about people and negatively affect people's lives.

In this work, we study the possibility of involving unbiased unlabeled data to learn an accurate and fair classifier. To this end, two fair learning algorithms are proposed. Those two algorithms employ a iterative label swapping procedure to discover more fair and accurate labelings of test points closest to decision boundary. The algorithms then use the new labelings to refit a fairer classifier. The closeness of test points to decision boundary is measured empirically by a threshold. Such a threshold is also a key element through which we can quantify a known trade-off between accuracy and fairness in the fair learning setting. With access to this threshold, a user of proposed fair learning algorithms can fine-tune and make informed decisions when weighing accuracy against fairness. Our experiments show classifiers learned with proposed fair algorithms give fairer predictions than classifiers learned with pure semi-supervised algorithms and supervised algorithm.