COSCFair: Ensuring Subgroup Fairness Through Fair Classification Framework

Begum Hattatoglu
Utrecht University
Utrecht, Netherlands
b.hattatoglu@students.uu.nl

Enas Khwaileh
Utrecht University
Utrecht, Netherlands
e.t.k.khwaileh@uu.nl

Abdulhakim Qahtan
Utrecht University
Utrecht, Netherlands
a.a.a.qahtan@uu.nl

Heysem Kaya
Utrecht University
Utrecht, Netherlands
h.kaya@uu.nl

Yannis Velegrakis
Utrecht University and University of Trento
Utrecht, Netherlands
i.velegrakis@uu.nl

Introduction: The usage of Machine Learning (ML) in a wide diversity of domains has affected everyone's daily life. For example, machine learning algorithms are used for decision making in business and government systems, in recommending systems, advertisements, hiring systems, and so on. Machine learning algorithms have become widespread because of their high performance compared to humans in such tasks.

Machine learning algorithms can handle big volumes of data for complex computational tasks in significantly shorter time compared to humans. Besides, people usually have subjective opinions and points of view, which can lead to bias in their decisions. Unfortunately, a large number of systems have been identified to show bias against specific groups of the society. For example: i) Amazon's algorithm for free same-day delivery made racially biased decisions while choosing which neighborhoods to provide this service; ii) the COMPAS recidivism estimation tool, which is used in many courts of the United States shows significant discrimination against African-American males by predicting a higher risk for recidivism compared to white male offenders [source]. According to the automatically predicted risk level of the defendants, courts can keep the defendants in custody until the trial and consider this risk score while deciding the verdict.

There are several reasons behind the bias in ML algorithms: i) the under-representation of a certain group of people in the training set of a dataset; ii) the historical bias or prejudice reflected in the decision variable (class label); iii) limited features in a dataset that could be less informative about the population; and iv) the existence of attributes that are directly related to the sensitive attributes, such as race and gender, even when these sensitive attributes are not used to train the algorithms could be considered as another reason. These potential problems in a dataset cause machine learning algorithms to keep the existing bias and reflect it in their decisions, or even sometimes exacerbates the existing bias.

However, in order to identify and prevent bias in machine learning, researchers have come up with several different fairness metrics around fairness-aware machine learning. To improve the algorithmic fairness according to the fairness metrics, different algorithmic approaches have been developed to eliminate the existing bias or mitigate it under a certain level. Unfortunately, there is no consensus on which fairness metrics and mitigation algorithms are the best to ensure fairness yet.

We propose COSCFair, a pre-processing framework that can handle datasets with multiple sensitive attributes by eliminating its class and group imbalance via an oversampling technique to mitigate the bias before training the classifiers. This way, the classifier will not carry on or exacerbate the existing bias in a dataset. By eliminating both class and group imbalance simultaneously and obtaining the same base rates for all subgroups in a dataset, the framework will be able to satisfy multiple fairness metrics in the literature, which cannot be satisfied otherwise. Our framework is, therefore, based on oversampling the under-represented subgroups in the dataset. We used two different oversampling techniques: I) the Synthetic Minority Over-sampling TEchnique (SMOTE) [1]; II) the Generative Adversarial Network (GANs) [2]. However, it is crucial to ensure that the generated samples are realistic and close to the original data. For this reason, when using SMOTE [1], we cluster the data before performing the oversampling to improve the quality of the synthetic data. Since the original data samples in each cluster has more similarity with each other, the SMOTE oversampling technique used on these clusters will yield better quality of synthetic samples.

GANs is a deep learning model that use two main machines: a) the generator which is responsible for generating new samples from the original dataset; and b) the discriminator that tries to classify the new generated samples as either real (close to the samples from the domain), or fake samples. When the discriminator classifies a generated sample as fake, the generator adjustes its weights to produce more realistic samples. In our work, we use the Conditional GAN (CTGAN), where both the generator and the discriminator are conditioned on a set of attributes. We use the GANs to generate samples from the under-represented groups such that the final number of instances from each group (combinations of the values of the sensitive attributes and the class label) are equal. Experimental results over different datasets that are widely used as benchmarks to evaluate algorithmic fairness show that our framework yields consistent improvements compared to a set of baseline methods.

REFERENCES

[1] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.
[2] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).