DNN模型减少非公平性代表性算法整理

TABLE 2. Representative algorithms for mitigating unfairness in DNN models.

Class	Preprocessing	In-processing	Postprocessing
Discrimination via Input	Sensitive features removal Sensitive features replacement Reweighing[17] Optimized pre-processing[19]	Attribution regularization[14,15] Reduction game[16] Prejudice remover[18]	Calibrated distribution[6] Calibrated equalized odds[9]
Discrimination via Representation	Balanced dataset collection	Adversarial training[2,5] Adversarial fairness desideratum[20] Semantic constraints[21] Distance metrics[22,23]	Troubling neurons turn OFF
Prediction Quality Disparity	Diverse dataset collection[24] Synthetic data generation[26]	Transfer learning[25] Multitask learning[27]	

Preprocessing, in-processing, and postprocessing correspond to three stages of deep learning pipeline: dataset construction, model training, and model inference.

在Fairness_in_Deep_Learning_A_Computational_Perspective这篇综述里，给出了截至2021年8月发表的具有代表性的去除模型非公平性的算法，由上面这张Table2的表格展示。

此类算法按照算法进行时关注的阶段被归为了三类，分别是发生在预处理阶段、训练时、后处理阶段的三大类算法。

根据文章给出的引文信息，对这些算法的原文进行了搜查和算法适应的场景分析，并且初步挖掘了算法文章中是否给出了具有利用价值的源码资源等。

序号	论文题目	bias来源范畴	去偏阶段	资源&类型&备注	pdf链接
17	Data preprocessing techniques for classification without discrimination	input	预处理	数据集，算法实现 源码link：https://sites.google.com/site/faisalkamiran/	https://link.springer.com/content/pdf/10.1007/s10115-011-0463-8.pdf
19	Optimized pre-processing for discrimination prevention	input	预处理	datasetCOMPAS：https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis 未见显著算法实现本身的源码资源	https://proceedings.neurips.cc/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf
6	Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints	input	后处理	代码和数据：https://github.com/uclanlp/reducingbias	https://arxiv.org/pdf/1707.09457.pdf
15	Incorporating Priors with Feature Attribution on Text Classification	input	训练时	dataset：https://git.io/fjVEo 4.3提到了CNN 未见显著的源码资源	https://arxiv.org/pdf/1906.08286.pdf
14	Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations	input	训练时	code:https://github.com/dtak/rrr code参考的高品质code：https://github.com/HIPS/autograd https://github.com/marcotcr/lime	https://arxiv.org/pdf/1703.03717.pdf
16	A Reductions Approach to Fair Classification	input	训练时	code exponentiated-gradient reduction：https://github.com/fairlearn/fairlearn	http://proceedings.mlr.press/v80/agarwal18a/agarwal18a.pdf

序号	论文题目	bias来源范畴	去偏阶段	资源&类型&备注	pdf链接
18	Fairness-aware classifier with prejudice remover regularizer	input	训练时	数据集：e Adult / Census Income 未见显著代码资源	https://link.springer.com/content/pdf/10.1007%2F978-3-642-33486-3.pdf 60-75页
9	Equality of opportunity in supervised learning	input	后处理	未见显著代码资源	https://arxiv.org/pdf/1610.02413.pdf
2	Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations	representation	训练时	未见显著代码资源	https://arxiv.org/pdf/1811.08489.pdf
5	Adversarial removal of demographic attributes from text data	representation	训练时	code 和data acquisitionhttps://github.com/yanaiela/demog-text-removal	https://arxiv.org/pdf/1808.06640.pdf
20	Learning adversarially fair and transferable representations	representation	训练时	dataset： https://archive.ics.uci.edu/ml/datasets/adult https://www.kaggle.com/c/hhp	http://proceedings.mlr.press/v80/madras18a/madras18a.pdf
21	Discovering fair representations in the data domain	representation	训练时	data：http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html https://www.research.ibm.com/artificial-intelligence/trusted-ai/diversity-in-faces/ https://archive.ics.uci.edu/ml/datasets/adult tensorflow code实现：https://github.com/predictive-analytics-lab/Data-Domain-Fairness	https://openaccess.thecvf.com/content_CVPR_2019/papers/Quadrianto_Discovering_Fair_Representations_in_the_Data_Domain_CVPR_2019_paper.pdf

序号	论文题目	bias来源范畴	去偏阶段	资源&类型&备注	pdf链接
22	The variational fair autoencoder	representation	训练时	未见显著代码资源	https://arxiv.org/pdf/1511.00830.pdf
23	Wasserstein Fair Classification	representation	训练时	参考代码：https://github.com/PythonOT/POT data来自uci： http://archive.ics.uci.edu/ml 未见实现本身的源码资源	http://proceedings.mlr.press/v115/jiang20a/jiang20a.pdf
24	Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016	prediction quality	预处理	dataset：https://www.zooniverse.org/projects/pszmt1/faces-of-the-world/ HOIP etc. 涉及CNN，图片	https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7789583
25	Inclusivefacenet: Improving face attribute detection with race and gender diversity	prediction quality	训练时	dataset（face of the World）：http://chalearnlap.cvc.uab.es/challenge/13/track/20/description/ 参考代码：Prediction race from face for movie data. https://github.com/usc-sail/mica-race-from-face/wiki 未见本身实现给出的代码地址 涉及CNN，图片	https://arxiv.org/pdf/1712.00193.pdf

序号	论文题目	bias来源范畴	去偏阶段	资源&类型&备注	pdf链接
26	Age progression/ regression by conditional adversarial autoencoder	prediction quality	预处理	主页https://zzutk.github.io/Face-Aging-CAAE/ code：https://bitbucket.org/aicip/face-aging-caae/src/master/ https://github.com/ZZUTK/Face-Aging-CAAE 在线Face Transformer (FT) demo. http://cherry.dcs.aber.ac.uk/transformer/	https://openaccess.thecvf.com/content_cvpr_2017/papers/Zhang_Age_ProgressionRegression_by_CVPR_2017_paper.
27	Mitigating bias in gender, age and ethnicity classification: A multi-task convolution neural network approach	prediction quality	训练时	已废弃仓库（404）https://github.com/davidsa-ndberg/facenet dataset：https://sites.google.com/site/eccvbefa2018/	https://openaccess.thecvf.com/content_ECCVW_2018/papers/11129/Das_Mitigating_Bias_in_Gender_Age_and_Ethnicity_Classification_a_Multi-Task_ECCVW_2018_paper.pdf

以上囊括了Table2展示出的算法的引用以及相关资源整理，便于后续实验挑选适合的算法进行。综述中给出了评估的算法包括了17，19，18，9四篇给出的，结果由下表展示，后续实验可以参照综述这篇文章的做法和思路。

TABLE 3. Mitigation comparison between five methods for discrimination via input.

	Adult Income				COMPAS			
Model/Data	Acc	Parity	Opty	Odds	Acc	Parity	Opty	Odds
Dataset_bias	n/a	0.386	n/a	n/a	n/a	0.747	n/a	n/a
DNN_original	0.836	0.347	-0.094	-0.089	0.658	0.741	−0.160	−0.136
Reweighting[17]	0.832	0.654	−0.106	−0.090	0.652	0.788	−0.186	−0.149
Optimized_pre[19]	0.778	0.573	−0.107	-0.088	0.665	0.959	−0.018	−0.024
Prejudice_rem[18]	0.817	0.961	0.005	0.039	0.635	0.937	0.008	0.062
Calibrated_odds[9]	0.804	0.546	0.148	0.052	0.639	0.819	0.036	0.150

For accuracy and demographic parity, the close to 1 the better. For equality of opportunity and equality of odds, the close to 0 the better