Sino-US English Teaching, May 2025, Vol. 22, No. 5, 183-187
doi:10.17265/1539-8072/2025.05.006

Research on the Framework of Bias Detection and Elimination in Artificial Intelligence Algorithms

Haoxuan Lyu
The Chinese University of Hong Kong, Hong Kong, China

The excessive use of artificial intelligence (AI) algorithms has caused the problem of errors in AI algorithms, which has challenged the fairness of decision-making, and has intensified people's inequality. Therefore, it is necessary to conduct in-depth research and propose corresponding error detection and error elimination methods. This paper first proposes the root causes and threats of bias in AI algorithms, then summarizes the existing bias detection and error elimination methods, and proposes a bias processing framework in three-level dimensions of data, models, and conclusions, aiming to provide a framework for a comprehensive solution to errors in algorithms. At the same time, it also summarizes the problems and challenges in existing research and makes a prospect for future research trends. It is hoped that it will be helpful for us to build fairer AI.

Keywords: artificial intelligence (AI), algorithm bias, bias detection, bias elimination, fairness, framework research

Introduction

With the rapid development of artificial intelligence (AI) technology, it has become more and more widely used in various industries, even in the medical field, financial field, legal field, etc. However, because the data analysis decision process of AI algorithms is likely to be affected by bias, it can result in injustice and even false results. The emergence of such algorithmic bias not only harms the fairness of the algorithm, but is also likely to intensify the gap in society. Therefore, finding and removing biases in algorithms is one of the most important issues in the AI field at present. The goal of the bias detection and elimination framework proposed in this paper is to use a series of technical means to detect possible biases in the algorithm learning process and provide corresponding solutions to maintain the professionalism and clarity of the algorithm. This article will discuss the current research status of the bias detection and elimination framework, the technical difficulties, and the future development trends faced.

Analysis of Bias Sources in AI Algorithms

The deviations that occur in AI algorithms can come from data, models, and results. Differences in data are the most common situation. AI requires a large amount of information input to complete the corresponding machine learning (ML) work, and it may cause bias in the process of data collection, sorting, and labeling. For example, some datasets exist in insufficient and unrepresentative samples, which makes some of the learned rules not applicable to all populations. In addition, the deviation of the model is triggered by certain specific assumptions in the formulation and training stage of the algorithm or by algorithm selection. For example, if an

Haoxuan Lyu, Master's student, Faculty of Business Administration, The Chinese University of Hong Kong, Hong Kong, China.

algorithm focuses too much on a particular feature, it will lead to some unfair biases that the data has. In the end, the deviation in the results cannot be ignored, because the results of the algorithm may also be generated from the deviation, resulting in the final judgment quality. The root causes of these biases are not only the distortion of the model, but also the potential to lead to more serious social problems, such as discrimination and imbalance.

Prejudice Detection Method

Before addressing bias, it should be analyzed accurately. The detection methods of bias mainly include analyzing data, establishing models, and evaluating results. These methods are used to diagnose bias from different angles, providing a reference for the subsequent solution to bias.

Data-Based Bias Detection

Data bias, as a potential source of algorithmic bias, is the bias we must first identify. By analyzing the performance of each feature in the dataset, we can find the biases that exist in the dataset. For example, the number of data samples for certain specific groups may be small, which makes the model unable to learn relevant knowledge from them, which leads to the algorithm's discriminatory bias against the group. To determine whether there is data bias, some of the statistical methods we most often use, tools that compare the performance gaps in each category in the data set, are used to find bias. If the number of samples in one category is much lower than the number of samples in another category, or if a certain feature has a serious deviation in its distribution, it can be judged that there is data bias. At the same time, labelers also have the possibility of data bias in the labeling process. For example, labelers may unconsciously bias certain categories of data, resulting in an imbalance in their labeling results. Therefore, it is necessary to use data to conduct inspection and analysis, observe the data through graphical tools, statistical means, etc., to detect biases in the data as early as possible, and provide guidance for subsequent operations.

Model-Based Bias Detection

Model deviation detection is mainly used to detect whether the performance of the algorithm model in different groups meets their preferences, but this can usually only be used to use some traditional model evaluation indicators, such as precision, accuracy, and recall observing whether there are differences in the performance of different groups of models using the same algorithm. To this end, the general method of detecting model bias is to compare the model performance of each population, especially some populations with characteristics affected by sensitive attributes (such as age, gender, race, etc.). For example, the classification accuracy of the classification model in the male group is 90%, while the accuracy rate in the female group is only 70%. It can be preliminarily believed that there are certain deviations in the model. At this time, we introduced some fairness indicators, such as equality of opportunity, prediction consistency, etc., to measure the significant differences in the model performance of each group among the population. Once, there is a significant difference in model performance between different populations of the model, it is considered that the model has a bias. In addition, another method of using the model is to detect whether the learning results of the model intentionally further amplify the deviations of certain specific groups through the model learning function or the gradient update pattern. This approach allows for in-depth insight into the model process and thus finds hidden sources of bias.

Results-Based Bias Detection

The result-oriented bias identification method is mainly adopted, that is, to evaluate whether the decisions made by the model cause unfair differences in different groups. Specifically, the decision results of each group

(based on gender, race, socioeconomic status, etc.) can be monitored to determine whether the model can ensure equilibrium among groups. If a model causes unfair results to a certain group, it means that the group will be treated unfairly by unfairly allocating resources, referees, and hiring.

Prejudice Elimination Framework

If potential bias against ML is not discovered and processed in a timely manner, it will lead to unfair behavior of the algorithm in actual application. Therefore, the treatment of bias is one of the important concerns of AI ethics and fairness, and justice. For bias processing, we can use some methods in the data layer and the modeling layer, respectively. Each method has its methods and means. Next, we will introduce the bias elimination of the data layer and the modeling layer.

Data-Level Bias Elimination

One reason that affects bias removal is difficult to achieve is data bias, which is generally caused by inequality or bias that exists in the data set. To solve such problems, we can optimize and adjust the data from the data dimension, strive to reflect the situation of various people, and eliminate the discriminatory components.

Data recalibration method. Data re-marking is mainly to eliminate bias in the data dimension and usually refers to the means to eliminate the influence of bias by modifying the data set labels or characteristics. The main idea is that if the labels in the training data are injustice or biased, we will be able to make the labels of the dataset more just by redefining the labels or adjusting some characteristics. For example, in classification problems, some samples of special groups may cause misjudgment of labels due to the bias of human judges. We can use human wisdom to correct labels or use ML algorithms to repair labels in the dataset, so that labels in the dataset can objectively and accurately reflect the actual situation. After applying data re-labeling, the adverse effects caused by the inherent unfair labeling of the data can be reduced, and to a large extent, the bias of the algorithm during training can be alleviated. At the same time, data re-annotation can also eliminate data bias by increasing the importance of a category of samples and reducing their impact on model training.

Data balance and compensation. Data balancing and compensation are other ways to mitigate bias at the data level, i.e., balancing problems caused by insufficient sample or imbalance of distribution among special groups. In practical applications, samples from a certain group often appear excessively, while samples from another group are ignored or underestimated. For data balance, there are usually two methods: One is over-sampling, increasing the number of a few classes to obtain a balanced data set; two is undersampling, reducing the number of the majority class to obtain a balanced data set. This method will avoid the prediction of the outcome bias caused by the vast majority of tendencies during the learning process of the model.

There are ways in which data can be compensated over the entire data set to represent these people in the overall data. For example, additional remedial measures (such as interpolation or synthesis) can be used to increase the relative weight of its training set in cases where the proportion of a sample in a certain category is relatively low in the data set to increase the relative weight of its training set to make the data more consistent and balanced. Then, these methods train the learner to learn fairer rules for these different categories of people to ensure that realistic models are not biased.

Model-Level Bias Elimination

In addition to data adjustments available at the data level, there are also many ways to mitigate bias at the model level. Mainly, the imposition of fairness constraints and regularization methods on the model during the

training process can effectively reduce the bias of the model to group differences, improve the accuracy of predictions, and take into account fairness.

Fairness constraint algorithm. The fairness constraint algorithm introduces a method to introduce attention to fairness in the process of constructing the model. The goal is to optimize the model and improve the prediction accuracy while ensuring that the model reaches a certain level of fairness criteria. This compromise method is often expressed in the form of mathematical formulas, so as to achieve the purpose of controlling the performance differences between groups in the model. Fairness constraints often come in many forms, such as the balance of predictions, the equality of opportunities, and the balance of results. For example, for the algorithm of a job search platform, fairness constraints may include "no matter the gender and race of the applicants, the model needs to allocate opportunities to them with the same probability". Commonly used fairness constraints include "group fairness" and "individual fairness", which correspond to the fairness and equality of each group, respectively.

To achieve this goal, the practice of modifying the model loss function or optimization method can be adopted, so that the model should not only focus on reducing the prediction error rate, but also consider fairness evaluation criteria, to avoid discrimination caused by excessive attention to a specific group during the training process. However, compromises are needed to ensure fairness. After all, too strict requirements may lead to poor performance of the model. Therefore, how to take into account fair needs while maintaining a good performance level within a reasonable range of choices is the problem we have to face in designing fair constraint algorithms.

Reinforcement learning and regularization methods. Other model-level bias removal methods rely on reinforcement learning (RL) and regularization methods. Reinforcement learning can train the model's decision-making behavior through reward and punishment methods to ensure fairness among groups. For example, it can formulate a reward and punishment rule to guide the model to make fair decisions and punish unfair decisions, so that the model's decision-making behavior continues to evolve and tends to be fair. Reinforcement learning can play a good role in eliminating bias, and there are many application scenarios. For example, in some scenarios that require real-time decision-making (such as recommendation systems, autonomous driving, etc.), the model can be guided to consider fairness issues in the decision-making process through appropriate reward functions.

Conclusion

In summary, in order to ensure the fairness of the AI system, bias detection and mitigation are needed. This paper sorts out some existing major technologies, including bias discovery techniques, such as finding bias in data, finding bias in models, and finding bias in results, as well as bias removal solutions, such as relabeling, adjusting algorithms, and post-correction. Although the existing bias removal system has developed to a certain extent, there are still many problems, such as how to deal with the complexity of algorithmic bias, improve the time efficiency in the mitigation process, and ensure its interpretability. In the future, we should pay more attention to the enhancement of the permeability and practicality of bias detection and mitigation technology, and further understand its transparency and interpretability, so that AI technology can further develop in a fairer way.

References

Chen, W. K., Liu, H. F., & Wu, S. S. (2023). Framework for the elimination of AI algorithms based on fairness constraints. *Computer Science and Exploration,* 12, 102-108.

Wang, C. X., Zhang, P. F., & Liu, K. (2023). Research on algorithm bias detection and correction methods based on machine learning. *Computer Applications and Software, 40*(11), 48-52.

Zhao, Z. Y., & Liu, Y. X. (2024). Bias and elimination strategies in deep learning models. *Artificial Intelligence and Big Data,* 3, 25-30.