symmetry

Article

Fair Outlier Detection Based on Adversarial Representation Learning

Shu Li [1,*], Jiong Yu [2,*], Xusheng Du [2], Yi Lu [1] and Rui Qiu [1]

[1] School of Software, Xinjiang University, Urumqi 830046, China; outman@stu.xju.edu.cn (Y.L.); qiurui@stu.xju.edu.cn (R.Q.)
[2] School of Information Science and Engineering, Xinjiang University, Urumqi 830046, China; duxusheng@stu.xju.edu.cn
* Correspondence: lishu@stu.xju.edu.cn (S.L.); yujiong@xju.edu.cn (J.Y.)

Abstract: Outlier detection aims to identify rare, minority objects in a dataset that are significantly different from the majority. When a minority group (defined by sensitive attributes, such as gender, race, age, etc.) does not represent the target group for outlier detection, outlier detection methods are likely to propagate statistical biases in the data and generate unfair results. Our work focuses on studying the fairness of outlier detection. We characterize the properties of fair outlier detection and propose an appropriate outlier detection method that combines adversarial representation learning and the LOF algorithm (AFLOF). Unlike the FairLOF method that adds fairness constraints to the LOF algorithm, AFLOF uses adversarial networks to learn the optimal representation of the original data while hiding the sensitive attribute in the data. We introduce a dynamic weighting module that assigns lower weight values to data objects with higher local outlier factors to eliminate the influence of outliers on representation learning. Lastly, we conduct comparative experiments on six publicly available datasets. The results demonstrate that compared to the density-based LOF method and the recently proposed FairLOF method, our proposed AFLOF method has a significant advantage in both the outlier detection performance and fairness.

Keywords: fair outlier detection; algorithmic fairness; adversarially fair representation learning; local outlier factors; symmetric structure

check for updates

Citation: Li, S.; Yu, J.; Du, X.; Lu, Y.; Qiu, R. Fair Outlier Detection Based on Adversarial Representation Learning. *Symmetry* **2022**, *14*, 347. https://doi.org/10.3390/sym14020347

Academic Editor: José Carlos R. Alcantud

Received: 3 January 2022
Accepted: 3 February 2022
Published: 9 February 2022

1. Introduction

With the development of machine learning technology, more and more decision-making problems have been replaced by algorithms. Machine learning is a data-driven approach to automated decision-making that has a high potential to introduce or even perpetuate discriminatory issues already present in the data [1]. Existing research results suggest that algorithms trained using unbalanced datasets may reflect or even reinforce the social biases present in the data, such as the bias of facial analysis algorithms against skin color [2], Word2Vec algorithms against gender [3], and advertising recommendation systems against gender [4]. Research work in fairness machine learning aims to eliminate potential discrimination of algorithms. In recent years, most work in fairness machine learning has focused on supervised learning, especially on classification problems [5,6]. The latest research work has also been on fairness research in unsupervised directions, such as clustering algorithm [7] and recommendation systems [8].

The primary task of eliminating prejudice and realizing algorithmic fairness is to define the concept of fairness. The definitions of fairness are broadly classified as follows: individual fairness [9] (similar individuals have the same treatment); group fairness [10] (treatment of different groups equally); subgroup fairness [11] (combination of individual and group fairness according to fairness constraints); counterfactual fairness [12] (complementary to group equity, based on causality). Unbalanced datasets and biased algorithms are the main reasons that affect the fairness of machine learning, and achieving fairness constraints from the algorithm level is the key to achieving fairness. The latest developments in algorithm fairness mainly focus on fair representation learning and adversarial techniques. Just representation learning generates intermediate representations of original data and deletes sensitive attribute information in the data while retaining task-related information [13]. The adversarial network joint training was first proposed by Goodfellow [14]. According to the characteristics of the GAN network, Zhang [15] and Madras [16] proposed fair representation models that eliminate discrimination through adversarial learning.

Despite the rapid development of fairness research in machine learning, relatively little research on fairness has been done in the field of outlier detection. Outlier detection is widely used in critical areas such as fraud detection, intrusion detection, public safety, and medical supervision. The nature of the task and the applied scenarios dictate the necessity to introduce fairness into outlier detection. For example, in credit risk assessment, people determined to be anomalous by outlier detection systems will have their bank accounts frozen; those classified as abnormal will be detained or imprisoned in crime detection work. To a certain extent, being misclassified as an anomaly can be detrimental to both the individual and the organization. Suppose the outlier detection model cannot correctly distinguish between social minorities (defined by sensitive attributes, such as gender, race, etc.) and statistical minorities (outliers). In that case, more members of social minorities will be incorrectly labeled as outliers, which will further strengthen stereotypes in human society and even cause social conflicts. Therefore, it is crucial to ensure that sensitive attributes do not influence outlier detection. However, the fairness of outlier detection has not received any attention until 2020, and so far, there has not been a comprehensive solution.

Fair Outlier detection methods usually aim at group fairness. Davidson [17] discussed fairness in outlier detection algorithms and proposed a framework based on combinatorial optimization problems for detecting fairness in outlier detection methods. The FairLOF algorithm proposed by Deepak and Abraham [18] in 2020 introduced the concept of fairness into the outlier detection method for the first time. It improved the fairness of the density-based LOF algorithm through three heuristic principles. However, the FairLOF algorithm mainly acts on the original feature space, and the redundant attribute information in the data will affect outlier detection. In addition, the FairLOF algorithm uses the principle of statistical parity [19] to perform fairness processing on its detection results, but only Considering statistical parity does not guarantee fairness or even weakens outlier detection performance [20]. Aiming at the insufficiency of the FairLOF algorithm in outlier detection performance and fairness, this paper proposes an outlier detection method based on adversarial fair representation learning. Our approach uses the symmetric model structure, and it hides sensitive attributes in the data through confrontational training and improves fairness while ensuring the effectiveness of outlier detection. We generalize our contributions in this paper as follows:

- We discuss the fairness of outlier detection and characteristic the four properties of fair outlier detection. Further, we propose three metrics for measuring the fairness of outlier detection from three different perspectives (statistical parity, equality of opportunity, and conditional use accuracy equality).
- We combine the density-based LOF method with fair representation learning to optimize the effectiveness and group fairness of outlier detection by learning fairness representations of the original data through adversarial training.
- We use local outlier factors to represent the outlier scores of data and assign lower weight values to data with higher outlier scores by adjusting the dynamic weights to mitigate the impact of outliers on representation learning.
- We conduct several experiments on six public datasets from different real-world domains. The results demonstrate the significant advantages of our proposed AFLOF method over the LOF method and the FairLOF method in terms of fairness and performance.

2. Materials and Methods

2.1. Datasets

Our experiments use the public datasets commonly used in algorithmic fairness research from six different fields in the real world. Table 1 lists the primary characteristics of each dataset. To make the experiments more illustrative, we will appropriately downsample some groups in the dataset and keep the percentage of outlier points in each sensitive attribute subgroup at 5%. Next, we will introduce each dataset in detail.

Table 1. Characteristics of six datasets.

Dataset	Size	Number of Attributes	Sentitive Attribute	Outlier Definition
Weight	1500	17	gender	insufficient weight
Drug	1190	32	gender	used within last week
Crime	4000	23	gender	multiple crimes within two years
Credit	6000	25	age	delinquent
Student	630	33	gender	final score less than 7
Adult	9400	15	race	income more than 50 K

Weight [21]. This dataset records individuals' dietary habits and physical conditions in Mexico, Peru, and Colombia. Overall, 77% of the data is generated using a combination of Weka tools and SMOTE filters, and 23% is collected directly from users through a web-based platform. Gender is a sensitive attribute of this dataset, where the ratio of males to females is 2:1.

Drug [22]. This dataset records the interviewee's drug use, including attributes such as gender, education level, and drug use. The sensitive attribute is gender, where the ratio of males to females is 3:1.

Crime [23]. This dataset is the data evaluated by Florida using the COMPAS risk assessment tool, including information such as gender, number of crimes, and arrest status. The sensitive attribute is gender, where the ratio of males to females is 4:1.

Credit [24]. This dataset belongs to the financial domain, and the research object is credit card customers in Taiwan. It records the payment status, credit data, historical bills, and other records of credit card customers. The sensitive attribute is age, and the ratio of people over 25 years old and under 25 years old is 5:1.

Student [25]. This dataset is close to the student performance of two Portuguese schools, including information such as name, gender, grades, family status, and social status. The sensitive attribute is gender, where the ratio of females to males is 6:1.

Adult [26]. This dataset is extracted from the 1994 Census database and records individual income levels and education, occupation, household, etc. Race (White, Black, Asian-Pac-Islander, Amer-Indian-Eskimo, Other) is a sensitive attribute in this dataset, and the number of sensitive subgroups is in the ratio of 30:10:5:1:1.

2.2. Preliminary

2.2.1. Notion of Fair Outlier Detection

Given a dataset of size N, the task of outlier detection is to find a small fraction of data points from dataset $X = \{x_i\}_{i=1}^{N}$ that are considered to be anomalous. Each data point x_i is associated with a sensitive attribute $S = \{s_i\}_{i=1}^{N}, s_i \in R$ where $R = \{r_i\}_{i=1}^{M}$ represents the sensitive attribute subgroup. We denote the dataset processed by the feature extractor as $D = \{d_i\}_{i=1}^{N}$ and use it as the input of the LOF algorithm. The local outlier factors $L = \{l_i\}_{i=1}^{N}$ obtained by the LOF algorithm will be used to represent the outlier scores of the data. Further, we denote $L = \{l_i\}_{i=1}^{N} Y = \{y_i\}_{i=1}^{N}, y_i \in \{0,1\}$ as the true labels of the data and represent the detector-assigned labels to the data as $O = \{o_i\}_{i=1}^{N}, o_i \in \{0,1\}$. Table 2 lists the frequently used symbols and related descriptions in the paper.

Table 2. Symbols and descriptions.

Symbols	Descriptions
N	size of the dataset
M	number of the sensitive attribute categories
X	original dataset
D	dataset processed by the feature extractor
S	sensitive attribute
R	categories of the sensitive attribute
L	local outlier factors
Y	true labels of the data
O	detector-assigned labels to the data

The purpose of unsupervised fair outlier detection is to ensure the fairness of the detection results while maintaining the detection performance. Having presented the problem setup and symbols, we characterize the properties of fair outlier detection intending to achieve group fairness as follows:

1. Effective detection. The primary task of outlier detection methods is to ensure detection performance. It makes sense to consider fairness only if the outlier detection methods can accurately and efficiently identify outliers. The fair outlier detection model needs to meet the following condition (see Equation (1)) to ensure effectiveness.

$$P(Y = 1 \mid O = 1) > P(Y = 1) \tag{1}$$

2. Statistical parity [19]. Statistical parity means that outlier detection is independent of sensitive attributes; that is, outlier detection performance among sensitive attribute subgroups should be consistent. The fair outlier detection model needs to meet the following condition (see Equation (2)) to ensure statistical parity.

$$P(O = 1 \mid S = r_i) = P\big(O = 1 \mid S = r_j\big), \forall r_i, r_j \in R \tag{2}$$

3. Equality of opportunity [20]. Equality of opportunity requires fairness in the target group. In outlier detection, equal opportunity means that outliers should be given higher scores and flagged regardless of the sensitive genus subgroup they belong to. The fair outlier detection model needs to meet the following condition (see Equation (3)) to ensure equality of opportunity.

$$P(O = 1 \mid Y = 1, S = r_i) = P\big(O = 1 \mid Y = 1, S = r_j\big), \forall i, j \in [1, M] \tag{3}$$

4. Conditional use accuracy equality [27]. In outlier detection, conditional use accuracy equality implies that the probability of true positive and true negative rates among sensitive attribute subgroups should be the same. The fair outlier detection model needs to meet the following condition (see Equation (4)) to ensure conditional use accuracy equality.

$$P(Y = 1 \mid O = 1, S = r_i) = P\big(Y = 1 \mid O = 1, S = r_j\big), \forall i, j \in [1, M] \tag{4}$$

2.2.2. Evaluation Metrics

Table 3 lists the confusion matrix of outlier detection. According to the above problem definition of fairness outlier detection, we will evaluate the model from the following two aspects: outlier detection performance and fairness performance.

Table 3. Confusion matrix.

True Label	Predicted Label	
	Positive	**Negative**
Positive	TP (True Positive)	FN (False Negative)
Negative	FP (False Positive)	TN (True Negative)

- Outlier detection performance. The Receiver Operating Characteristic Curve (ROC) is used to measure the classification performance [28]. The x-axis of the ROC curve is the false positive rate (TP/(TP + FN)), and the y-axis is the true positive rate (FP/(FP + TN)). Area Under Receiver Operating Characteristic Curve (AUC) is the area under the ROC curve, where a higher value means better outlier detection performance. AUC represents the diagnostic ability of outlier detection at each scoring threshold. Therefore, we will use the AUC score to measure the outlier detection performance. AUC is defined as follows (see Equation (5)).

$$\mathrm{AUC} = \frac{1}{2}\sum_{i=1}^{n}(x{-}axis_{i+1} - x{-}axis_i) \times (y{-}axis_{i+1} + y{-}axis_i) \tag{5}$$

- Fairness performance.
  1. We will measure whether the outlier detection algorithm achieves statistical parity by comparing the difference in detection effectiveness on each sensitive attribute subgroup. Specifically, we calculate the AUC score for every sensitive attribute subgroup and assign Fair Statistical Parity (FSP) with the value of the most significant gap among AUC scores. We represent the number of sensitive attribute subgroups as M and denote FSP as follows (see Equation (6)).

$$\mathrm{FSP} = \max\left(\mathrm{AUC}_i - \mathrm{AUC}_j\right), i, j \in M \tag{6}$$

  2. We will measure the equality of opportunity of the outlier detection methods by comparing the distribution of sensitive attribute subgroups over the outlier candidates and the entire dataset. We use P to represent the sensitive attribute subgroup distribution in the dataset and Q to describe the distribution of sensitive attribute subgroup in the top 5% outlier candidates. To be specific, we sort the data in descending order by local outlier factors and use relative entropy (also known as Kullback Leibler (KL) divergence) to calculate the difference between P and Q. Then, we assign Fair Equality of Opportunity the most apparent distribution difference value. The definition of FEO is as follows (see Equation (7)).

$$\mathrm{FEO} = \max(\mathrm{KL}(\mathrm{Q}_m \| \mathrm{P}_m)), m \in M \tag{7}$$

  3. The Matthews Correlation Coefficient (MCC) [29] is a metric used to measure binary classification performance, considering true positive, true negative, false positive, and false negative. MCC applies to unbalanced datasets and is one of the most appropriate evaluation metrics when considering the misclassification of detection results [30]. We will measure whether the algorithm achieves the conditional use accuracy equality by comparing MCC scores among sensitive attribute subgroups. To be specific, we calculate the MCC score for every sensitive attribute subgroup and assign Fair Matthews Correlation Coefficient (FMCC) with the value of the most significant gap among MCC scores. We denote FMCC as follows (see Equation (8)).

$$\mathrm{FMCC} = \max\left(\left|\mathrm{MCC}_i - \mathrm{MCC}_j\right|\right), i, j \in M \tag{8}$$

The MCC is calculated based on the confusion matrix, which is calculated as follows (see Equation (9)).

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{9}$$

FSP, FEO, and FMCC provide a more comprehensive measure of the fairness performance of outlier detection algorithms from three different perspectives. All three metrics are negative measurements, where a smaller value suggests higher fairness.

2.3. Methods

2.3.1. Model Overview

In this study, the model has three main modules: a feature extractor, a sensitive attribute discriminator, and an outlier detector, where the feature extractor is an autoencoder (AE) structure. We visualize the architecture of our proposed AFLOF method in Figure 1. The model's input is a dataset containing sensitive attributes, and the output is outlier scores of the data object. The training process of the model consists of three parts: minimizing the reconstruction loss to train the feature extractor, maximizing the classification loss to train the discriminator, and guiding subsequent iterative learning by dynamically adjusting weight factors based on outlier scores. The rest of the section will detail the model's main components and training methods.

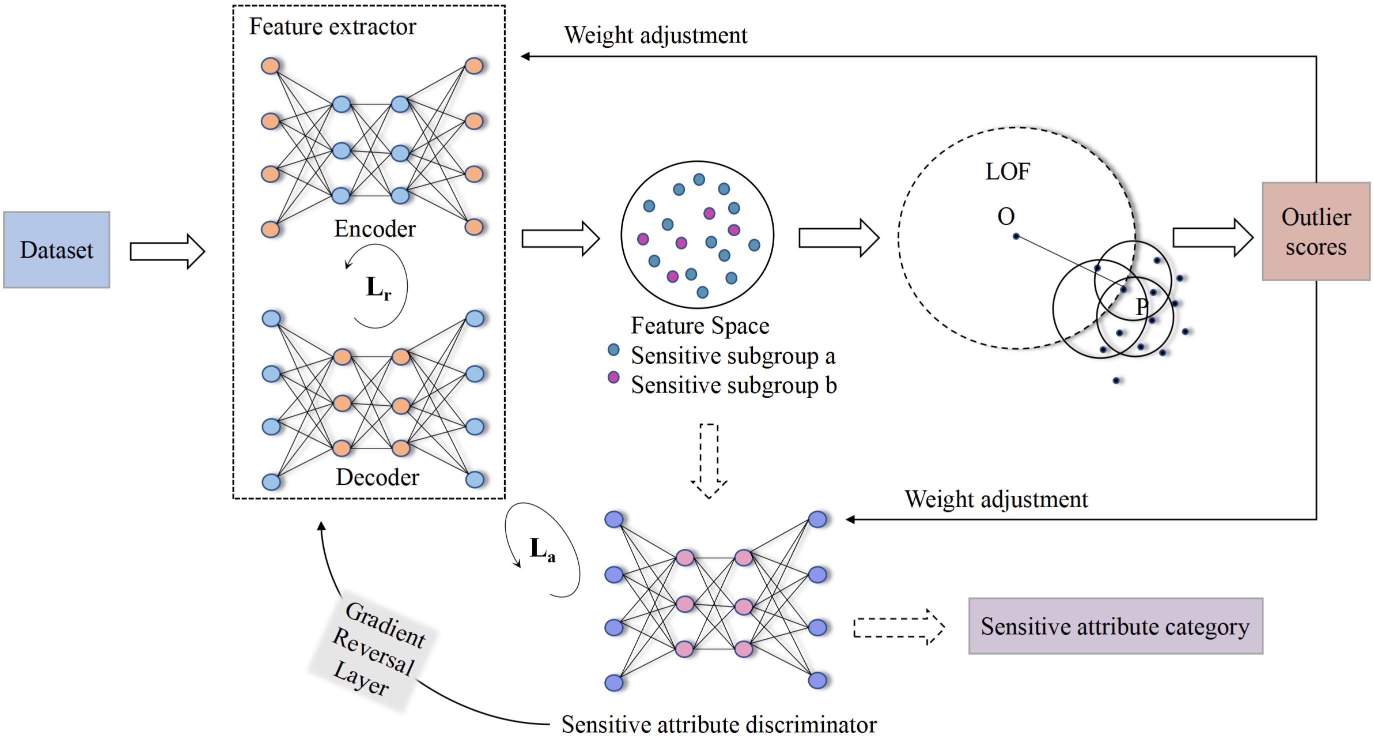


Figure 1. Architecture of the fair outlier detection method.

2.3.2. The Feature Extractor and the Sensitive Attribute Discriminator

The feature extractor and the sensitive attribute discriminator are deep learning artifacts in the overall model, where there is symmetry between the encoder and the decoder in the extractor. The function of the feature extractor is to extract features from native data and remove redundant information, thus improving the identification effects of the outlier detector. The sensitive attribute discriminator distinguishes whether the features extracted by the feature extractor contain sensitive attributes. Since the outlier detector is not a deep learning component, fairness factors should be incorporated in the feature extraction stage to make the results of outlier detection fairer; that is, the extracted features should not contain sensitive attributes that may lead to unfairness as far as possible. Thus, in the process of model optimization, we will minimize the reconstruction loss to train the feature extractor and maximize the classification loss to train the discriminator. The adversarial training between the extractor and discriminator hides the sensitive attribute of the data in the feature space and ensures the fairness of outlier detection. In addition, to achieve adversarial training between the feature extractor and the sensitive attribute discriminator, we introduce a gradient reversal layer [31]. It acts as an identity transform in the forward propagation and inverts the gradient direction automatically during the backward propagation to update the parameters.

2.3.3. The Outlier Detector

Due to the LOF algorithm can quantify the degree of abnormality for each data object without knowing the distribution of the dataset, we choose it as the outlier detector of our model. Our fair outlier detection model aims to improve the detection performance and fairness of the LOF algorithm. The LOF algorithm is the basis of the fair outlier detection algorithm proposed in this paper, which identifies outliers by comparing the density of each data with its neighbors and assigns each data object a local outlier factor characterizing its degree of outliers. We will use local outlier factors as outlier scores of the data object. The computational process of the LOF algorithm is divided into four stages to obtain the local outlier factors of the data step by step.

1. K-distance. Let $d(x_i, p)$ be the distance between x_i and p. $N_k(x_i)$ represents the set of k nearest neighbors of x_i. For any data object $x_i \in X$, the k_{distance} of x_i is defined as the furthest distance from $p \in N_k(x_i)$ to x_i. The k_{distance} of a data object x_i is defined as follows (see Equation (10)).

$$k_{\text{distance}}(x_i) = \max\{d(x_i, p) \mid p \in N_k(x_i)\} \tag{10}$$

2. Reachability Distance. If a data object x_i is within the k neighborhood of p, the reachability distance between x_i and p is the k_{distance} of p; otherwise, the reachability distance between x_i and p is the real distance of x and p. The reachability distance of a data object x_i is defined as follows (see Equation (11)).

$$\text{reach}_{\text{ distance }}(x_i, p) = \max\{k_{\text{distance }}(p), d(x_i, p)\} \tag{11}$$

3. Local Reachability Density. For any data object $x_i \in X$, the local reachability density is the inverse of the average reachability distance from all data within the k neighborhood of x_i to x_i. The lower the local reachability density, the more likely it is to be an outlier. The local reachability density of a data object x_i is defined as follows (see Equation (12)).

$$\text{lrd}(x_i) = 1/\left(\frac{\sum_{p \in N_k(x_i)} \text{ reach}_{\text{ distance }}(x_i, p)}{|N_k(x_i)|}\right) \tag{12}$$

4. Local Outlier Factor. The local outlier factor is the average ratio of the local reachability density of all data in the k neighborhood of x_i to the local reachability density of x_i, which indicates the degree of outliers of the data object x_i. The larger the value, the more likely it is to be an outlier. The local outlier factor of a data object x_i is defined as follows (see Equation (13)).

$$\text{lof}(x_i) = \left(\frac{\sum_{p \in N_k(x_i)} \text{lrd}(p)}{|N_k(x_i)|}\right) / \text{ lrd}(x_i) \tag{13}$$

2.3.4. Dynamic Weights

To ensure the effectiveness of outlier detection in adversarially fair representation learning. We introduce the dynamic weights to our model. We use local outlier factors obtained by the LOF algorithm to represent the outlier scores of data objects in the feature space. We decrease the negative influences of outliers on representation learning by dynamically adjusting the weights after each iteration. Specifically, we assign higher weights to data objects with lower outlier scores and lower weights to data objects with higher outlier scores. We further emphasize outliers through lower weight values, which mitigate the influence of outliers on feature representation learning and enhance the fair representation of outliers in adversarial representation learning. To calculate the weights, we will use the softmax function to normalize the local outlier factor of each data. We represent the local outlier factor of a data object x_i as l_i, the specific calculation of the dynamic weights is as follows (see Equation (14)).

$$w_i = \frac{e^{-l_i}}{\sum_j^N e^{-l_j}} \tag{14}$$

2.3.5. Adversarially Fair Representation Learning

The key to traditional adversarial learning is to distinguish whether an image is a fake image generated by a generator. This study aims to distinguish whether the extracted features contain sensitive attributes that may lead to unfair results. Our proposed AFLOF method aims to obtain a fair representation for describing the data by adversarial learning. A fair representation is obtained when the data after representation learning is independent of sensitive attributes. We represent the encoder, sensitive attribute discriminator, and decoder as h, d, and g. Specifically, h is used to map the original data X into a feature space, denoted as $h : X \to D$. d is used to identify the sensitive attribute classes of the data after feature extraction, denoted as $d : D \to S$. g is used to reconstruct the sample data in the feature space, denoted as $g : D \to \bar{X}$. According to the above components, we set the training goal for outlier detection based on adversarial fair representation learning as the following two loss functions: reconstruction loss function L_r and classification loss function L_a.

Our model instantiates an autoencoder, which consists of two parts: encoder, denoted as $f = h(X)$, and decoder, denoted as $f = g(h(X))$. The encoder maps the original data to a feature space, and the decoder reconstructs the data objects in the feature space. The data reconstructed by the decoder is different from the original data. Therefore, we set a loss function to measure the reconstruction loss of the decoder and then adjust the parameters by backpropagation of the loss until convergence, to minimize the reconstruction error. We use the loss function of the L2 norm, also known as mean square error, as the reconstruction loss function. The principle of the L2 norm is to apply a penalty to the loss function in the optimization phase, making the representation more sparse and preventing the overfitting problem. In particular, we weighted the encoder representation to reduce the impact of outliers on the update of the feature extractor parameters. The expression of L_r is shown below (see Equation (15)).

$$L_r = \sum_{i=1}^{N} w_{\mathrm{i}} \times (x_{\mathrm{i}} - g(h(x_{\mathrm{i}})))^2 \tag{15}$$

The goal of the model is to learn a fair representation of the original data. For this purpose, we use a sensitive attribute discriminator, which serves to identify the sensitive attribute categories of the data in the feature space. Specifically, we use the cross-entropy loss function as the loss function of the sensitive attribute discriminator. The function is used to evaluate the gap between the sensitive attribute categories of the data predicted by the sensitive attribute discriminator and their actual sensitive attribute categories. We perform a weighted representation for the sensitive attribute discriminator to enhance the fair representation of outliers by adversarial representation learning. The classification loss function is shown below (see Equation (16)).

$$L_a = \sum_{i=1}^{N} w_i \times \left(s_i \times \log\left(\frac{1}{d(h(x_\mathrm{i}))} \right) \right) \tag{16}$$

Our AFLOF model employs an adversarial network with a minimum–maximum strategy to improve fairness while ensuring the effect of outlier detection. We apply different training strategies to the encoder and the sensitive attribute discriminator in the model training stage. We realize the fair representation learning of the model by minimizing the reconstruction loss of the encoder and maximizing the classification loss of the sensitive attribute discriminator. In summary, we address the overall objective function of the model as follows (see Equation (17)):

$$min_{h,g} max_d E_{X,S}[L(h,g,d)] \tag{17}$$

The combined objective function expressed as $L(h,g,d) = \alpha L_r + \beta L_a$. The hyperparameters α and β are used to balance the performance and fairness of outlier detection. We show the processing of the AFLOF algorithm using pseudo-code, as shown in Algorithm 1.

Algorithm 1 AFLOF

```
Input: X: dataset, S: sensitive attribute, T: training epochs, h: encoder, d: discriminator, g: decoder.
Output: L: oulier scores
Initialize parameters for $h$,$d$ and $g$;
Train the encoder network $h$ and discriminator $d$ via minimizing $L_r$ in Equation (15) and maximizing $L_a$ in Equation (16) for $E$ epochs.
for epoch from 1 to $E$ do
    Calculate the local outlier factor $L$ of each data object by Equation (13);
    Calculate the dynamic weight $W$ of each data object by Equation (14);
    Calculate the reconstruction loss $L_r$ of the feature extractor;
    Calculate the classification loss $L_a$ of the sensitive attribute discriminator;
    Back-propagate the $loss$;
    update $h$,$d$ and $g$;
end for
return oulier scores
```

2.4. Implementation

We implement LOF, FairLOF, and our proposed outlier detection model (AFLOF) based on adversarially fair representation learning in PyTorch. To make the experiment more contrastive, we set the k neighborhood size of LOF, FairLOF, and AFLOF to be uniformly 5, following the setting in the FairLOF literature [18]. Our model uses the fully connected network to implement the encoder, decoder, and sensitive attribute discriminator. The dimension of feature space is 12. Specifically, the network structure of the encoder is N-500-2000-500-12, that of the decoder is 12-2000-500-500-N, and that of the sensitive attribute discriminator is 12-500-500-2000-M. The initial learning rate of the encoder and decoder is 0.001, and the initial learning rate of the discriminator is 0.0001. To achieve better training performance, we adjust the learning rate dynamically, reducing the learning rate to the original 0.1 every 30 epochs. To speed up the training process, we set the number of iterations to 90 and the batch size to 64 for datasets smaller than 5000, otherwise to 40 and 256, respectively. The hyperparameters α and β are set to 8 and 20. We conduct experiments on ten random seeds and take the average of ten experimental results as the final result. Our experiments are run in an environment where the processor is Intel(R) Xeon(R) Gold 5117 CPU @ 2.00 GHz, and the graphics card is Nvidia Tesla V100-PCIe-16GB 256 GB RAM.

3. Results

3.1. The Unfairness of LOF Algorithm

This section will discuss the unfairness of the density-based LOF algorithm. We explore whether the LOF algorithm can produce fair detection results in two cases. One is the balanced dataset, where the size of sensitive attribute subgroups is consistent. And the other is the unbalanced dataset, where the size of sensitive attribute subgroups in the dataset is inconsistent. Specifically, we divide each dataset into two types: balanced and unbalanced. The unbalanced data set is our original dataset. The balanced data set is generated by reducing the number of majority groups in the original dataset and keeping the proportion of outliers (5%) among sensitive attribute subgroups. Table 4 lists the relevant information of the data used in the experiment and displays the size of the two types of data and the percentage of sensitive attribute subgroups within them. We perform outlier detection under these two datasets separately using the LOF algorithm and measure the fairness of the LOF algorithm based on three fairness metrics (FSP, FEO, FMCC) for outlier detection. All three metrics are negative measurements, where the smaller value suggests the algorithm is fairer. We conducted experiments on six datasets. Figure 2 shows the fairness performance of the LOF algorithm on two types of data in each dataset. Observing these figures, we can see that the LOF algorithm demonstrates significant unfairness in the outlier detection for balanced and unbalanced data.

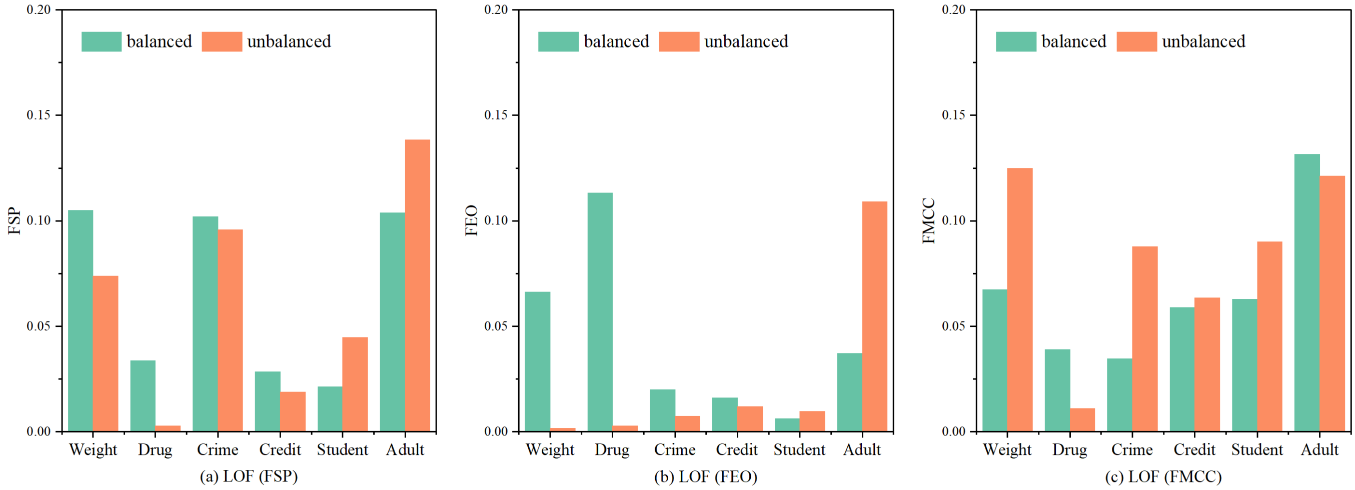


Figure 2. The fairness performance of the LOF algorithm on balanced datasets (green bars) and unbalanced datasets (orange bars).

Table 4. Characteristics of balanced datasets and unbalanced datasets used in the experiment.

	Weight	Drug	Crime	Credit	Student	Adult
Balanced	1000 (1:1)	600 (1:1)	2000 (1:1)	2000 (1:1)	180 (1:1)	1000 (1:1)
Unbalanced	1500 (1:2)	1190 (1:3)	4000 (1:4)	6000 (1:5)	630 (1:6)	9400 (1:1:5:10:30)

FSP determines whether the LOF algorithm relies on sensitive attributes when performing outlier detection by comparing the detection effectiveness on sensitive attribute subgroups. Figure 2a shows the fairness of the outlier detection algorithm from statistical parity perspective. We measure whether the LOF algorithm achieves statistical parity by comparing the difference in outlier detection effectiveness between each sensitive subgroup. From Figure 2a, we can see that in the four datasets of weight, drug, crime, and credit, the dependence of the LOF algorithm on the sensitive attribute is more apparent in the balanced dataset. As the gap in the number of sensitive attribute subgroups increases, the LOF algorithm becomes increasingly dependent on the sensitive attribute and more and more unfair when performing outlier detection on unbalanced datasets.

FEO compares the distribution of sensitive attribute subgroups between the entire dataset and the top 5% outlier candidates detected by the LOF algorithm. Figure 2b shows the fairness of the outlier detection algorithm from the equality of opportunity perspective, where equal chance in outlier detection means that outliers should be assigned higher scores and flagged regardless of the class of sensitive attribute subgroups they belong to. We measure whether the LOF algorithm achieves equality of opportunity by comparing the distribution of sensitive subgroups among outlier candidates. Figure 2b indicates that the distribution of sensitive attribute subgroups in balanced datasets varies significantly over the entire dataset and the outlier candidates, especially in the datasets of Drug and Weight. Specifically, the LOF algorithm shows more significant unfairness on the balanced dataset under the metric of the equality of opportunity.

FMCC measures the fairness of the LOF algorithm by comparing the accuracy of detection results among sensitive attribute subgroups. Figure 2c measures the fairness of the outlier detection algorithm from the perspective of conditional use accuracy equality. In outlier detection, we measure whether the LOF algorithm achieves conditional accuracy equality by comparing the differences in MCC scores of the sensitive attribute subgroups. Observing Figure 2c, we can see that the fairness performance of the LOF algorithm is poor on both data types, especially on the unbalanced dataset, where the difference in misclassification rates among sensitive attribute subgroups is more prominent. Comprehensive analysis of the above results, we understand that it is necessary to introduce fairness awareness into outlier detection to mitigate the unfairness of the LOF algorithm.

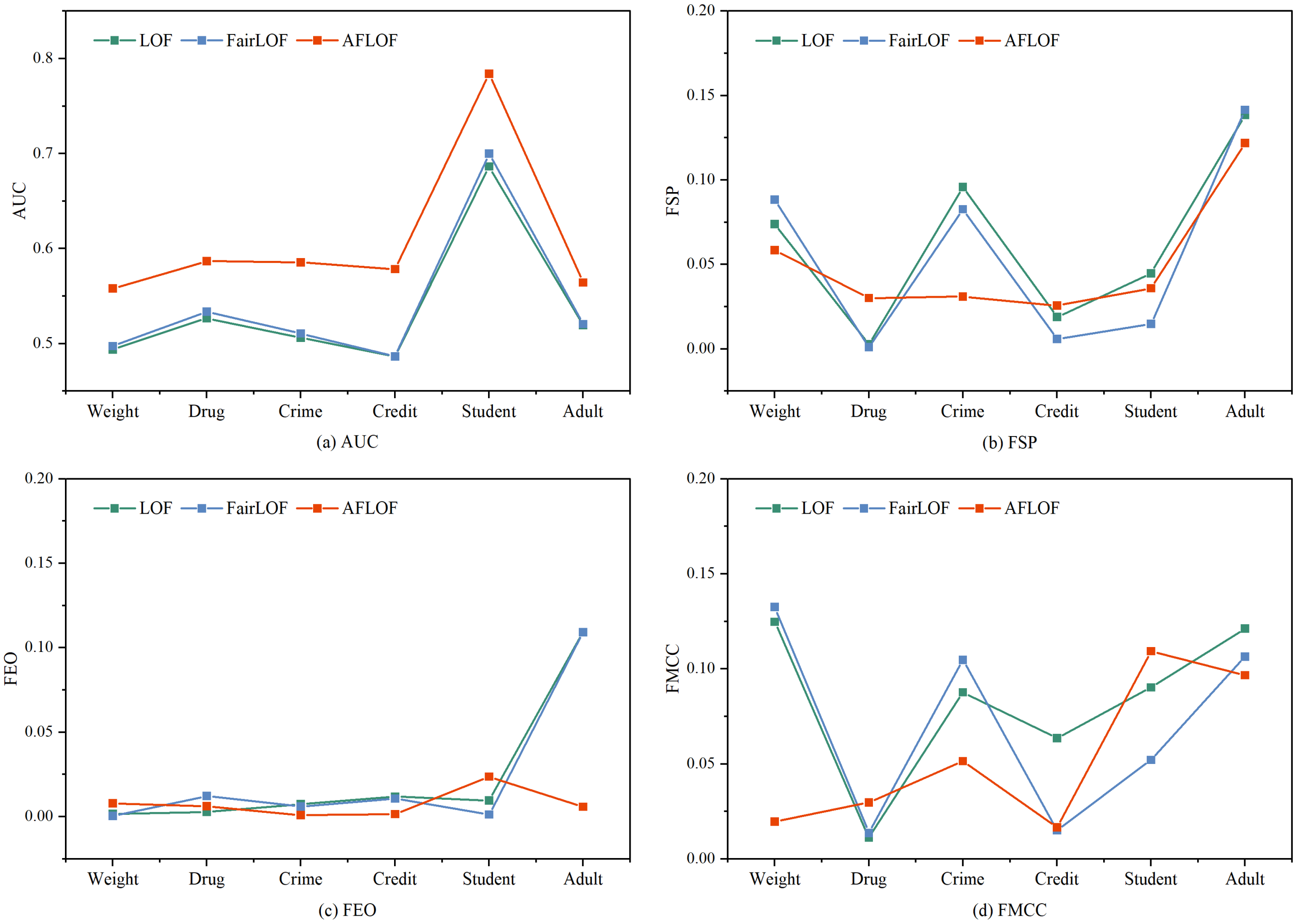
3.2. Evaluation

In this section, we will evaluate our proposed AFLOF method's outlier detection performance and fairness performance and make a comparison with LOF and FairLOF. We evaluate the performance for all three methods on six unbalanced datasets. Figure 3 shows the experimental results. Figure 3a shows the outlier detection performance. Note that the AFLOF method achieves better outlier detection results with a significant improvement in the AUC score. Figure 3b–d shows the fairness performance of outlier detection methods from three different aspects. We can see that the AFLOF method has advantages over the LOF method and the FairLOF method.

For outlier detection performance, AFLOF has the highest AUC scores on all datasets. Figure 3a shows that the difference in outlier detection performance between LOF and the FairLOF is minimal. The outlier detection performance of AFLOF is far superior to the other two algorithms. Specifically, in the adult dataset, AFLOF improve the AUC scores by 4% compared to LOF and FairLOF. In the datasets of credit and crime, the AUC scores of AFLOF improve by nearly 10% compared to the other two algorithms.

For fairness performance, we can see that AFLOF performs well compared to LOF and FairLOF. Figure 3b shows the dependence of LOF, FairLOF, and AFLOF on sensitive attributes. We see that LOF and FairLOF perform better on the drug, credit, and student datasets. Still, AFLOF does not compare poorly with them either. Moreover, AFLOF performs well in the three datasets of weight, crime, and adult, especially in the crime dataset, reducing the detection effectiveness gap among sensitive attribute subgroups by nearly double compared to the other two methods. Figure 3c shows the distribution difference of sensitive attribute subgroups between the top 5% outlier candidates and the whole dataset. We can see that the performance of the three methods is excellent, and AFLOF performs better overall than LOF and FairLOF and generates a fairer subgroup distribution. Figure 3d shows the detection accuracy differences among the three methods for the sensitive genus subgroups. We can see that the AFLOF method achieves lower detection accuracy differences on the four datasets of weight, crime, credit, and adult.

Analyzing the above results together, we can see that our proposed AFLOF has apparent advantages over the LOF algorithm and the FairLOF algorithm, both in terms of outlier detection performance and fairness.



(a) AUC

(b) FSP

(c) FEO

(d) FMCC

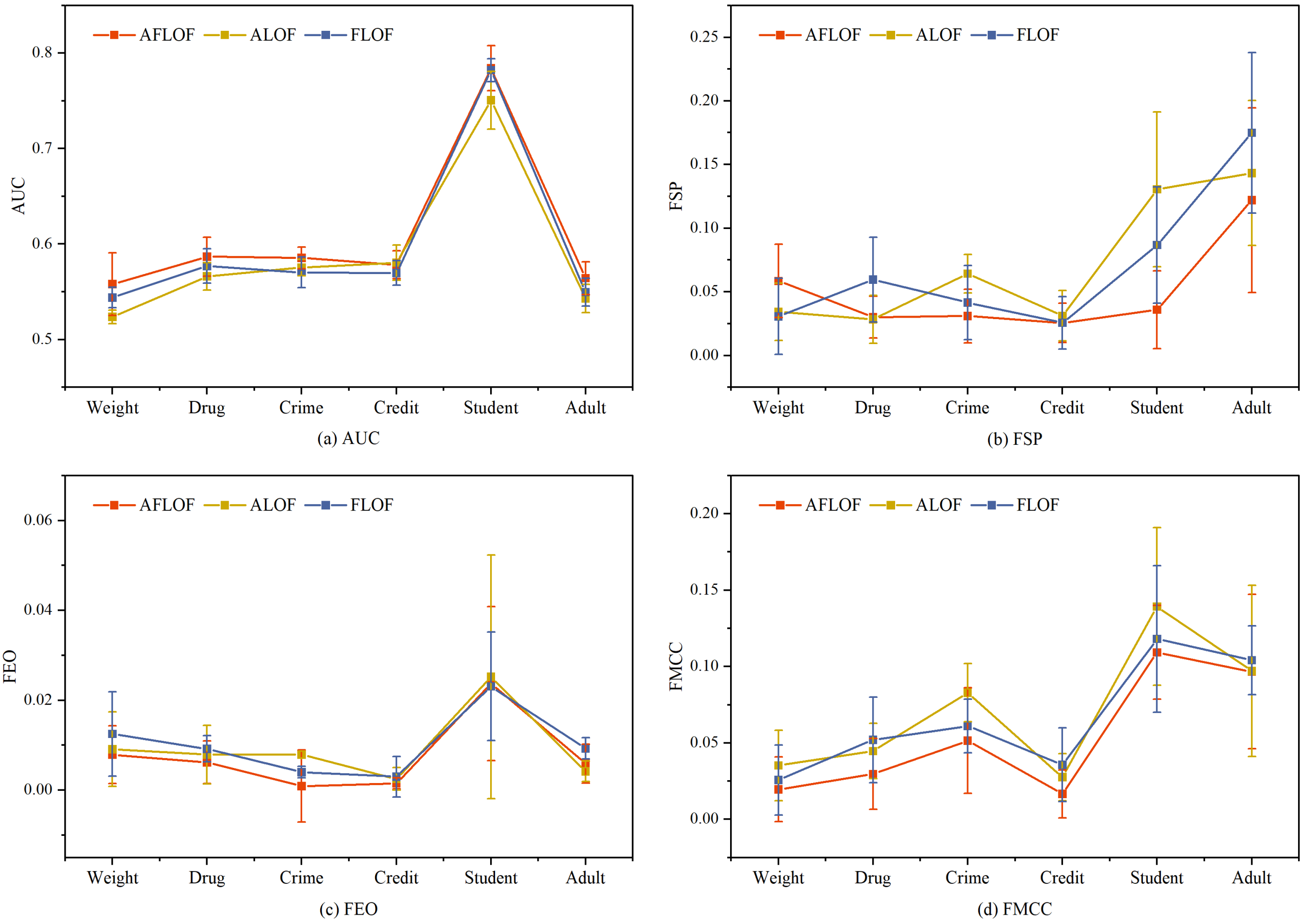
Figure 3. Comparison of LOF, FairLOF, and AFLOF on outlier detection performance and fairness performance.

3.3. Ablation Study

In this section, we will evaluate the effect of various components in the design of AFLOF. The concept of ablation research first emerged in psychology to study the impact of partial removal of an animal's brain on its behavior. Literature [32] first introduced ablation research to machine learning. Ablation research is used to learn networks by removing parts and studying their performance. Ablation studies are crucial for deep learning research and help us explore the causal relationships between experimental methods in the simplest way possible.

Simply put, we can think of an ablation study as a controlled variables approach, where a control group is set up, and its effect on the final results is demonstrated by adding or removing a module. This section will conduct an ablation study on our proposed algorithm to make the experiment more illustrative. Our method is based on the LOF algorithm by introducing a fair adversarial representation module and a dynamic weight module. We will remove these two modules separately to examine their effects on the experimental results.

Based on the AFLOF method, we get an ALOF model by removing the sensitive attribute discriminator and a FLOF model by removing the dynamic weight module. We conducted repeated experiments on ten random seeds. Figure 4 shows the mean and standard deviation of the three methods on six datasets. Observing Figure 4, we can see that AFLOF significantly outperforms ALOF and FLOF in both outlier detection performance and fairness performance. The removal of the adversarial training and the dynamic weighting module will impact outlier detection performance and fairness performance, which indicates the reasonability of AFLOF's superior performance on both detection effectiveness and group fairness.



(a) AUC

(b) FSP

(c) FEO

(d) FMCC

Figure 4. Comparison of AFLOF, ALOF and FLOF on outlier detection performance and fairness performance. ALOF (yellow lines) and FLOF (blue lines) is the variants of AFLOF (orange lines).

3.4. Trade-Off between Fairness Performance and Outlier Detection Performance

In this section, we will investigate the trade-off between detection performance and fairness performance of the fair outlier detection method based on adversarial representation learning (AFLOF). The ultimate goal of achieving algorithmic fairness is to incorporate fairness constraints to remove bias without affecting the primary task performance of the original machine learning model. In the field of outlier detection, fair outlier detection methods should achieve good fairness without impacting outlier detection performance compared to ordinary outlier detection methods.

In our AFLOF model, α and β are hyperparameters used to balance the outlier detection performance and fairness performance, both of them are set to 8 and 20, respectively. To prove the rationality of the hyperparameter settings, we will measure the detection performance and fairness of outlier detection through the control variable method. We conducted experiments on six datasets. The outlier detection performance metric for outlier detection is AUC, and the fairness performance metric is FSP, which is used to evaluate the difference in detection performance among sensitive attribute subgroups. Figure 5 shows how the model's detection performance and fairness performance vary with parameters. We analyzed hyperparameters α and β. Figure 5 (above) shows the trend of AUC on the six datasets, while α ranging from (1, 2, 4, 6, 8, 10, 15, 30) and β fixed to 20. Figure 5 (below) shows the trend of FSP on the six datasets, while β ranging from (1, 5, 10, 15, 20, 30, 40, 50) and α fixed to 8. As shown in Figure 5, when $\beta = 20$, the AUC scores of six datasets almost reach the highest value where α takes the value of 8; when $\alpha = 8$, the lowest values of FSP for all six datasets are concentrated where β takes the value of 20. In summary, the fairness and detection performance of the model reaches the best balance point at $\alpha = 8$ and $\beta = 20$.

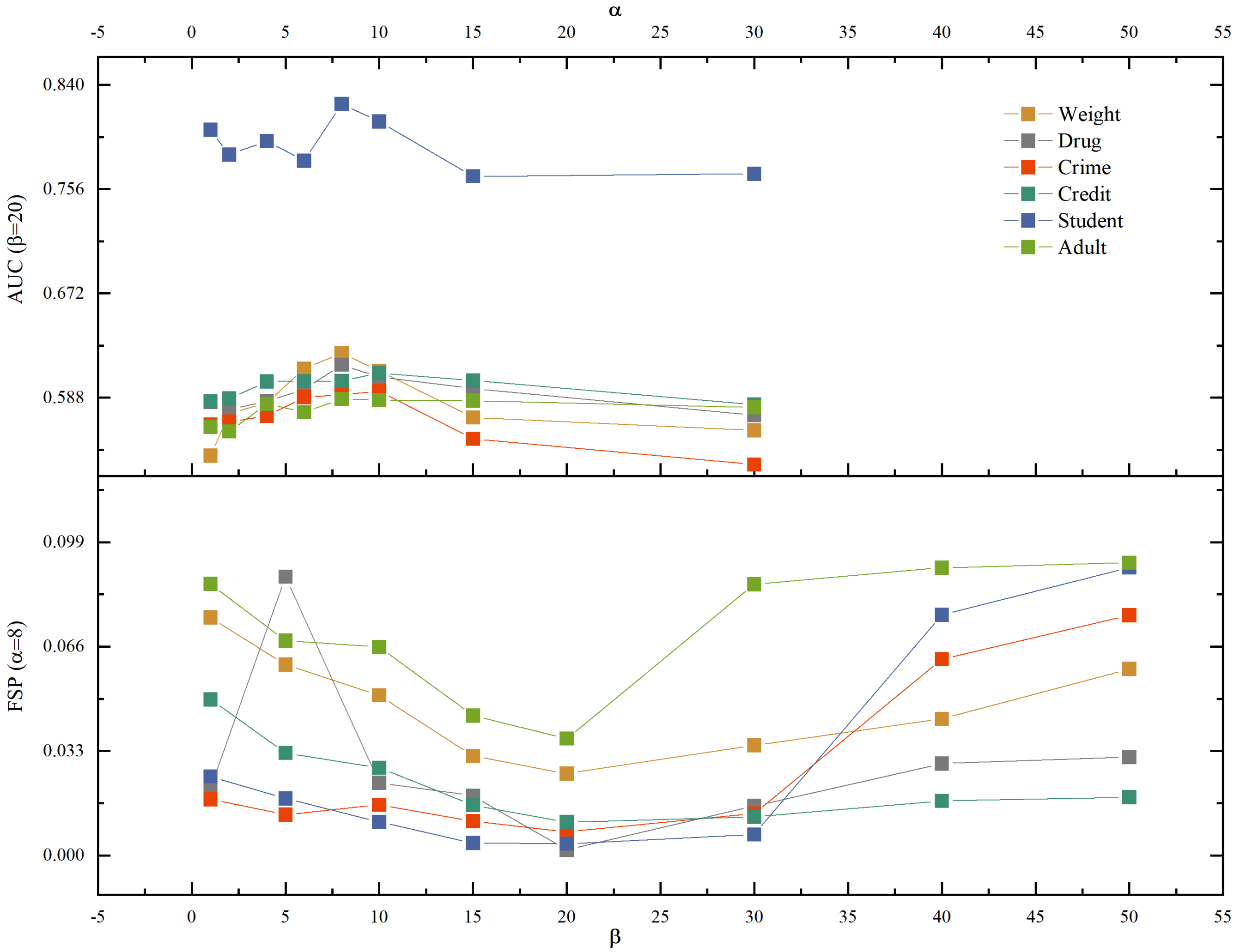


Figure 5. The tradeoff between outlier detection performance and fairness performance.

4. Discussion

Following the previous works [16,18,33], we introduce fair representations to the LOF algorithm and propose a fair outlier detection method (AFLOF) to tackle the fair problem in outlier detection. In this paper, we characterize the properties of fair outlier detection and propose three methods to measure the fairness of outlier detection methods. Subsequently, we investigate the unfairness in the LOF algorithm and conduct empirical research on six publicly available datasets to demonstrate the effectiveness of our AFLOF algorithm in terms of outlier detection performance and fairness. Further, we conduct ablation experiments to verify the effectiveness of the methods used in our model. Finally, we explore the trade-off between outlier detection performance and fairness, and analyze the hyperparameter settings' rationality to balance fairness performance and outlier detection performance.

There are three main techniques to achieve fairness: pre-processing, in-processing, and post-processing [13]. The FairLOF [18] algorithm improves the density-based LOF algorithm in three aspects: domain diversity, prior distribution, and attributes asymmetry. It introduces fairness constraints into the LOF algorithm and eliminates the sensitive attributes in the outlier detection results to a certain extent. However, FairLOF belongs to the post-processing method to achieve fairness technology, and one of its problems is that it has certain limitations on both detection effect and fairness. Our proposed AFLOF algorithm employs fair representation learning, which uses adversarial networks to learn to represent fairness to the original data and achieves a balance between fairness and detection accuracy in the data processing.

The reasons for the superior performance of AFLOF in outlier detection mainly come from the following two aspects:

1. Our model eliminates redundant attribute information by mapping the original data to a feature space, which facilitates the detection of outliers by the LOF method.
2. The outliers are further emphasized by the dynamic assignment of weights, which mitigates the negative impact of outliers on feature representation learning and enables the feature extractor to learn a better representation of the original data.

The reasons for the excellent performance of AFLOF in fairness performance mainly come from the following two aspects:

1. The adversarial training between the encoder and the sensitive attribute discriminator enables the model to learn the optimized representation of the original data while hiding the sensitive attributes in the data.
2. The dynamic assignment of weights further emphasizes outliers and enhances the fair representation of outliers in adversarial representation learning.

Based on the above analysis, we can see that, compared to LOF and FairLOF, our proposed fair outlier detection method (AFLOF) improves group fairness and outlier detection performance by adversarial fair representation learning. However, our paper is limited to studying a single multi-valued sensitive attribute. In the future, we will consider combining more advanced outlier detection methods to study multiple sensitive attributes to solve the more complex fairness problem of outlier detection.

Author Contributions: Conceptualization, S.L. and J.Y.; methodology, S.L. and Y.L.; software, S.L.; validation, J.Y., S.L. and R.Q.; formal analysis, S.L.; investigation, S.L.; resources, Y.L.; data curation, X.D.; writing—original draft preparation, S.L.; writing—review and editing, X.D., Y.L. and R.Q.; visualization, S.L.; supervision, J.Y.; project administration, J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The National Natural Science Foundation of China under Grant 61862060, Grant 61462079 and Grant 61562086.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in https://archive.ics.uci.edu/ml/datasets.php (accessed on 10 September 2021). Details on each dataset can be found in our paper.

Acknowledgments: The authors would like to thank editors and referees for their precious remarks and comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Barocas, S.; Selbst, A.D. Big data's disparate impact. *Calif. Law Rev.* **2016**, *104*, 671. [CrossRef]
2. Bacchini, F.; Lorusso, L. Race, again: How face recognition technology reinforces racial discrimination. *J. Inf. Commun. Ethics Soc.* **2019**, *17*, 321–335. [CrossRef]
3. Bolukbasi, T.; Chang, K.W.; Zou, J.Y.; Saligrama, V.; Kalai, A.T. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Proceedings of the 30th Annual Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 4349–4357.
4. Schnabel, T.; Swaminathan, A.; Singh, A.; Chandak, N.; Joachims, T. Recommendations as treatments: Debiasing learning and evaluation. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1670–1679.

5. Huang, L.; Vishnoi, N. Stable and fair classification. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 2879–2890.
6. Zafar, M.B.; Valera, I.; Rogriguez, M.G.; Gummadi, K.P. Fairness constraints: Mechanisms for fair classification. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistic, Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 962–970.
7. Li, P.; Zhao, H.; Liu, H. Deep fair clustering for visual learning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 9067–9076.
8. Lambrecht, A.; Tucker, C. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Manag. Sci.* **2019**, *65*, 2966–2981. [CrossRef]
9. Kang, J.; He, J.; Maciejewski, R.; Tong, H. InFoRM: Individual fairness on graph mining. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York, NY, USA, 6–10 July 2020; pp. 379–389.
10. Li, B.; Li, L.; Sun, A.; Wang, C.; Wang, Y. Approximate group fairness for clustering. In Proceedings of the 38th International Conference on Machine Learning, Long Beach, CA, USA, 18–24 July 2021; pp. 6381–6391.
11. Kearns, M.; Neel, S.; Roth, A.; Wu, Z.S. An empirical study of rich subgroup fairness for machine learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 100–109.
12. Chiappa, S. Path-specific counterfactual fairness. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 7801–7808.
13. Beutel, A.; Chen, J.; Doshi, T.; Qian, H.; Wei, L.; Wu, Y.; Heldt, L.; Zhao, Z.; Hong, L.; Chi, E.H.; et al. Fairness in recommendation ranking through pairwise comparisons. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2212–2220.
14. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the 28th Annual Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014.
15. Zhang, B.H.; Lemoine, B.; Mitchell, M. Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 2–3 February 2018; pp. 335–340.
16. Madras, D.; Creager, E.; Pitassi, T.; Zemel, R. Learning adversarially fair and transferable representations. In Proceedings of the 35th International Conference on Machine Learning, Stockholmsmässan, Stockholm, Sweden, 10–15 July 2018; pp. 3384–3393.
17. Davidson, I.; Ravi, S.S. A framework for determining the fairness of outlier detection. In Proceedings of the 24th European Conference on Artificial Intelligence, Online and Santiago de Compostela, Spain, 29 August–8 September 2020; pp. 2465–2472.
18. Deepak, P.; Abraham, S.S. Fair outlier detection. In Proceedings of the 21st International Conference on Web Information Systems Engineering, Leiden, South Holland, Nederland, 20–24 October 2020; pp. 447–462.
19. Garg, P.; Villasenor, J.; Foggo, V. Fairness metrics: A comparative analysis. In Proceedings of the 2020 IEEE International Conference on Big Data, Atlanta, GA, USA, 10–13 December 2020; pp. 3662–3666.
20. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. In Proceedings of the 30th Annual Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 3315–3323.
21. Palechor, F.M.; de la Hoz Manotas, A. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data Brief* **2019**, *25*, 104344. [CrossRef] [PubMed]
22. Fehrman, E.; Muhammad, A.K.; Mirkes, E.M.; Egan, V.; Gorban, A.N. The five factor model of personality and evaluation of drug consumption risk. In Data Science; Springer International Publishing: Cham, Switzerland, 2017; pp. 231–242.
23. Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine Bias. Risk Assessments in Criminal Sentencing. 2020. Available online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed on 10 September 2021).
24. Yeh, I.C.; Lien, C. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Syst. Appl.* **2009**, *36*, 2473–2480. [CrossRef]
25. Cortez, P.; Silva, A. Using data mining to predict secondary school student performance. In Proceedings of the 5th Future Business Technology Conference, Porto, Portugal, 9–13 April 2008; pp. 5–12.
26. Dua, D.; Graff, C. UCI Machine Learning Repository. 2017. Available online: http://archive.ics.uci.edu/ml (accessed on 10 September 2021).
27. Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociol. Methods Res.* **2021**, *50*, 3–44. [CrossRef]
28. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
29. Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS ONE* **2017**, *12*, e0177678. [CrossRef] [PubMed]
30. Xu, J.; Zhang, Y.; Miao, D. Three-way confusion matrix for classification: A measure driven view. *Inf. Sci.* **2020**, *507*, 772–794. [CrossRef]
31. Ganin, Y.; Lempitsky, V. Unsupervised domain adaptation by backpropagation. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 7–9 July 2015; pp. 1180–1189.

32. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the 29th Annual Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
33. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: Identifying Density-Based Local Outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 16–18 May 2000; pp. 93–104.