Disentangled Representation with Causal Constraints for Counterfactual Fairness

Ziqi Xu[1(✉)], Jixue Liu[1], Debo Cheng[1], Jiuyong Li[1(✉)], Lin Liu[1], and Ke Wang[2]

[1] University of South Australia, Adelaide, Australia
{Ziqi.Xu,Debo.Cheng}@mymail.unisa.edu.au,
{Jixue.Liu,Jiuyong.Li,Lin.Liu}@unisa.edu.au
[2] Simon Fraser University, Burnaby, Canada
wangk@sfu.ca

Abstract. Much research has been devoted to the problem of learning fair representations; however, they do not explicitly state the relationship between latent representations. In many real-world applications, there may be causal relationships between latent representations. Furthermore, most fair representation learning methods focus on group-level fairness and are based on correlation, ignoring the causal relationships underlying the data. In this work, we theoretically demonstrate that using the structured representations enables downstream predictive models to achieve counterfactual fairness, and then we propose the Counterfactual Fairness Variational AutoEncoder (CF-VAE) to obtain structured representations with respect to domain knowledge. The experimental results show that the proposed method achieves better fairness and accuracy performance than the benchmark fairness methods.

Keywords: Counterfactual Fairness · Representation Learning · Variational AutoEncoder

1 Introduction

Machine learning algorithms have gradually penetrated into our life [23] and have been applied to decision-making for credit scoring [16], crime prediction [14] and loan assessment [5]. The fairness of these decisions and their impact on individuals or society have become an increasing concern. Some extreme unfair incidents have appeared in recent years. For example, COMPAS, a decision support model that estimates the risk of a defendant becoming a recidivist was found to predict higher risk for black people and lower risk for white people [1]; Facebook users receive a recommendation prompt when watching a video featuring blacks, asking them if they'd like to continue to watch videos about primates [21]. These incidents indicate that the machine learning models become a source of unfairness, which may lead to serious social problems. Since most models are trained with data, which will lead to unfair decisions due to discrimination in

H. Kashima et al. (Eds.): PAKDD 2023, LNAI 13935, pp. 471–482, 2023.
https://doi.org/10.1007/978-3-031-33374-3_37

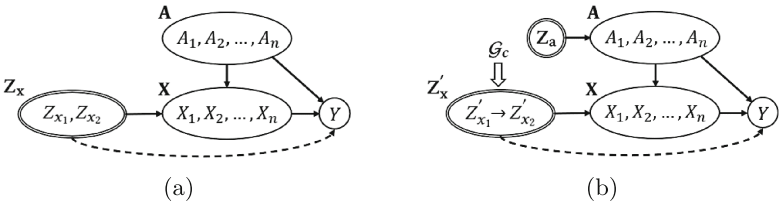


Fig. 1. (a) The process of existing works on learning fair representations to make predictions. (b) The process of our work. $\mathbf{A}$ is the set of sensitive attributes; $\mathbf{X}$ is the set of other observed attributes; $\mathbf{Z_a}$ is the representation of $\mathbf{A}$; Y is the target attribute; $\mathbf{Z_x}$ is the representation of $\mathbf{X}$; $\mathbf{Z'_x}$ is the structured representation of $\mathbf{X}$ with respect to the conceptual level causal graph $\mathcal{G}_c$. The dotted line denotes the prediction process.

the training data. Therefore, the key issue for solving unfair decisions becomes whether we can eliminate these discrimination embedded in the data through algorithms [23].

To obtain fair decisions, many methods [6,10,20,22,25,31] are proposed to learn fair representations through two competing goals: encoding data as much as possible, while eliminating any information that transfers through the sensitive attributes. To separate the information from sensitive attributes, various extensions of Variational Autoencoder (VAE) consider minimising the mutual information among latent representations [6,20,25]. For example, Creager et al. [6] introduced disentanglement loss into the VAE objective function to decompose observed attributes into sensitive latents and non-sensitive latents to achieve subgroup level fairness; Park et al. [25] improved the above methods and proposed the mutual attribute latent (MAL) to retain only beneficial information for fair predictions.

The existing methods [6,20] follow Fig. 1a to achieve fair predictions. Specifically, these methods learn fair representations $\mathbf{Z_x}$ without stating any relationships between Z_{x1} and Z_{x2}, which may not satisfy the domain knowledge. Let us consider an example where we aim to predict a person's salary using some observed attributes. Following the domain knowledge, we know that people's salary is determined by two semantic concepts, intelligence and career respectively. We also note that people's intelligence determines their career with high probability, which can be expressed as a conceptual level causal graph $\mathcal{G}_c$, i.e., $Intelligence \rightarrow Career$. Therefore, we need a method as shown in Fig. 1b that not only ensures the representation of observed attributes with no sensitive information but also retains causal relationships with respect to domain knowledge.

On the measurement of fairness, all fair representation learning methods use fairness metrics based on correlation, including the VAE-based methods [6,20,25]. It is well known that correlation does not imply causation. Recent studies [26,32] have shown that quantifying fairness based on correlation may produce higher deviations. Counterfactual fairness is a fundamental framework based on causation. With counterfactual fairness, a decision is fair towards an

individual if it is the same in the actual world and in the counterfactual world when the individual belongs to a different demographic group [17].

In this paper, we follow the counterfactual fairness and propose a VAE-based unsupervised fair representation learning method, namely Counterfactual Fairness Variational AutoEncoder (CF-VAE). We make the following contributions in this paper:

- We propose CF-VAE, a novel VAE-based unsupervised counterfactual fairness method. CF-VAE can learn structured representations with no sensitive information and retain causal relationships with respect to the conceptual level causal graph determined by domain knowledge.
- We theoretically demonstrate that the structured representations obtained by CF-VAE are suitable for training counterfactually fair predictive models.
- We evaluate the effectiveness of the CF-VAE method on real-world datasets. The experiments show that CF-VAE outperforms existing benchmark fairness methods in both accuracy and fairness.

2 Background

We use upper case letters to represent attributes and boldfaced upper case letters to denote the set of attributes. We use boldfaced lower case letters to represent the values of the set of attributes. The values of attributes are represented using lower case letters.

Let $\mathbf{A}$ be the set of sensitive attributes; $\mathbf{X}$ be the set of other observed attributes; $\mathbf{V}$ be the set of all observed attributes, i.e., $\mathbf{V} = \{\mathbf{A}, \mathbf{X}\}$; Y be the target attribute. We use $\widehat{Y}(\cdot)$ to represent the predictor. $\mathcal{G}_c$ is the conceptual level causal graph and represents domain knowledge. The nodes shown in $\mathcal{G}_c$ are "concepts", each of which represents a set of observed attributes that have similar meanings. Each "concept" has causal relationships with the other "concepts".

In this paper, a causal graph is used to represent a causal mechanism. In a causal graph, a directed edge, such as $V_j \rightarrow V_i$ denotes that V_j is a parent (i.e., direct cause) and we use pa_i to denote the set of parents of V_i. We follow Pearl's [26] notation and define a causal model as a triple $(\mathbf{U}, \mathbf{V}, \mathbf{F})$: $\mathbf{U}$ is a set of the latent background attributes, which are the factors not caused by any attributes in the set $\mathbf{V} = \{\mathbf{A}, \mathbf{X}\}$; $\mathbf{F}$ is a set of deterministic functions, $V_i = f_i(pa_i, U_{pa_i})$, such that $pa_i \subseteq \mathbf{V} \backslash \{V_i\}$ and $U_{pa_i} \subseteq \mathbf{U}$. Besides, some commonly used definitions in graphical causal modelling, such as faithfulness, d-separation and causal path can be found in [26,27].

With the causal model $(\mathbf{U}, \mathbf{V}, \mathbf{F})$, we have the following definition of counterfactual fairness:

Definition 1. (Counterfactual Fairness [17]). Predictor $\widehat{Y}(\cdot)$ is counterfactually fair if under any context $\mathbf{X} = \mathbf{x}$ and $\mathbf{A} = \mathbf{a}$, $P(\widehat{Y}_{\mathbf{A} \leftarrow \mathbf{a}}(\mathbf{U}) = y \mid \mathbf{X} = \mathbf{x}, \mathbf{A} = \mathbf{a}) = P(\widehat{Y}_{\mathbf{A} \leftarrow \bar{\mathbf{a}}}(\mathbf{U}) = y \mid \mathbf{X} = \mathbf{x}, \mathbf{A} = \mathbf{a})$, for all y and for any value $\bar{\mathbf{a}}$ attainable by $\mathbf{A}$.

Counterfactual fairness is considered to be related to individual fairness [17]. Individual fairness means that similar individuals should receive similar predicted outcomes. The concept of individual fairness when measuring the similarity of the individual is unknowable, which is similar to the unknowable distance between the real-world and the counterfactual world in counterfactual fairness [18].

3 Proposed Method

In this section, we first theoretically demonstrate that learning counterfactually fair representations are feasible. Then, we propose the Counterfactual Fairness Variational AutoEncoder (CF-VAE) to obtain the structured representations for predictors to achieve counterfactual fairness.

3.1 The Theory of Learning Counterfactually Fair Representations

We discuss what types of representations enable downstream predictive models to achieve counterfactual fairness. Following the work in [17], the implication of counterfactual fairness is described as follows:

Definition 2. (Implication of Counterfactual Fairness [17]**).** Let $\mathcal{G}$ be the causal graph of the given model $(\mathbf{U}, \mathbf{V}, \mathbf{F})$. If there exists $\mathbf{W}$ be any non-descendant of $\mathbf{A}$, then downstream predictor $\widehat{Y}(\mathbf{W})$ will be counterfactually fair.

We extend Definition 2 to the fair representation learning and present the following theorem.

Theorem 1. *Given the causal graph $\mathcal{G}$, $\mathbf{Z_a}$ is the representation of sensitive attributes $\mathbf{A}$, $\mathbf{Z}_\mathbf{x}^{'}$ is the structured representation of the other observed attributes $\mathbf{X}$ with respect to the conceptual level causal graph $\mathcal{G}_c$. We have $\widehat{Y}(\mathbf{Z}_\mathbf{x}^{'})$ satisfy counterfactual fairness.*

Proof. Given the causal graph $\mathcal{G}$ as shown in Fig. 2, there is not a parent node of $\mathbf{A}$ in $\mathbf{X}$, and there is not a child node of Y in $\mathbf{X}$. $\mathbf{X}$ contains four subsets: $\mathbf{X}_Y^\mathbf{A}$ is the subset of other observed attributes that are descendants of $\mathbf{A}$ and parents of Y; $\mathbf{X}_Y^\mathbf{N}$ is the subset of other observed attributes that are only parents of Y; $\mathbf{X}_\mathbf{N}^\mathbf{N}$ is the subset of other observed attributes that are no relationships with $\mathbf{A}$ and Y; $\mathbf{X}_\mathbf{N}^\mathbf{A}$ is the subset of other observed attributes that are only descendants of $\mathbf{A}$. After perfect representation learning, we obtain $\mathbf{Z_a}$ and $\mathbf{Z}_\mathbf{x}^{'}$.

We proof that $\mathbf{Z}_\mathbf{x}^{'}$ is not the descendant of $\mathbf{A}$ with the following two subsets. For the first subsets $\{\mathbf{X}_Y^\mathbf{A}, \mathbf{X}_Y^\mathbf{N}, \mathbf{X}_\mathbf{N}^\mathbf{A}\}$, there are seven paths between $\mathbf{A}$ and $\mathbf{Z}_\mathbf{x}^{'}$, including $\mathbf{A} \rightarrow \mathbf{X}_Y^\mathbf{A} \leftarrow \mathbf{Z}_\mathbf{x}^{'}$, $\mathbf{A} \rightarrow \mathbf{X}_Y^\mathbf{A} \rightarrow Y \leftarrow \mathbf{Z}_\mathbf{x}^{'}$, $\mathbf{A} \rightarrow \mathbf{X}_Y^\mathbf{A} \rightarrow Y \leftarrow \mathbf{X}_Y^\mathbf{N} \leftarrow \mathbf{Z}_\mathbf{x}^{'}$, $\mathbf{A} \rightarrow Y \leftarrow \mathbf{X}_Y^\mathbf{A} \leftarrow \mathbf{Z}_\mathbf{x}^{'}$, $\mathbf{A} \rightarrow Y \leftarrow \mathbf{Z}_\mathbf{x}^{'}$, $\mathbf{A} \rightarrow Y \leftarrow \mathbf{X}_Y^\mathbf{N} \leftarrow \mathbf{Z}_\mathbf{x}^{'}$ and $\mathbf{A} \rightarrow \mathbf{X}_\mathbf{N}^\mathbf{A} \leftarrow Y$. These seven paths are blocked by $\emptyset$ (i.e., $\mathbf{A}$ and $\mathbf{Z}_\mathbf{x}^{'}$ are d-separated by $\emptyset$), since each path contains a collider either $\mathbf{X}_Y^\mathbf{A}$ or Y or $\mathbf{X}_\mathbf{N}^\mathbf{A}$. For second subset $\mathbf{X}_\mathbf{N}^\mathbf{N}$, there is no path connecting $\mathbf{X}_\mathbf{N}^\mathbf{N}$ and Y. Hence, $\mathbf{Z}_\mathbf{x}^{'}$ is not the descendant of $\mathbf{A}$. Therefore, $\widehat{Y}(\mathbf{Z}_\mathbf{x}^{'})$ is counterfactually fair based on Definition 2. □

We use Fig. 2 to show whether the following predictors satisfy counterfactual fairness.

– $\widehat{Y}(\mathbf{A}, \mathbf{X})$: This model is unfair since it uses sensitive attributes to make prediction.

– $\widehat{Y}(\mathbf{X})$: This model satisfies fairness through awareness [8] but fails to achieve counterfactual fairness. Since it uses $\mathbf{X}_Y^{\mathbf{A}}$ and $\mathbf{X}_{\mathbf{N}}^{\mathbf{A}}$ which are the descendants of $\mathbf{A}$.

– $\widehat{Y}(\mathbf{Z_a}, \mathbf{Z}_{\mathbf{x}}^{'})$: This model is unfair because it uses sensitive attributes for prediction. The reason is that $\mathbf{Z_a}$ is the representation of $\mathbf{A}$, which should be consider as sensitive attributes either.

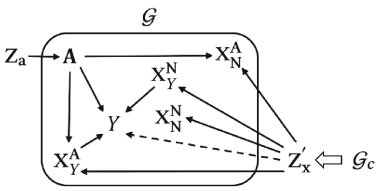


Fig. 2. $\mathcal{G}$ is the causal graph that represents the causal relationship between $\mathbf{A}$, $\mathbf{X} = \{\mathbf{X}_Y^{\mathbf{A}}, \mathbf{X}_Y^{\mathbf{N}}, \mathbf{X}_{\mathbf{N}}^{\mathbf{A}}, \mathbf{X}_{\mathbf{N}}^{\mathbf{N}}\}$ and Y. The dotted line represents the prediction process that uses $\mathbf{Z}_{\mathbf{x}}^{'}$.

– $\widehat{Y}(\mathbf{X}_Y^{\mathbf{N}}, \mathbf{X}_{\mathbf{N}}^{\mathbf{N}})$: This model satisfies counterfactual fairness since both $\mathbf{X}_Y^{\mathbf{N}}$ and $\mathbf{X}_{\mathbf{N}}^{\mathbf{N}}$ are non-descendants of $\mathbf{A}$. However, this predictor losses a lot of useful information that embeds in other observed attributes.

– $\widehat{Y}(\mathbf{Z}_{\mathbf{x}}^{'})$: This model is counterfactually fair based on Theorem 1 and achieves higher accuracy than $\widehat{Y}(\mathbf{X}_Y^{\mathbf{N}}, \mathbf{X}_{\mathbf{N}}^{\mathbf{N}})$ as shown in our experiments.

3.2 CF-VAE

We first discuss the causal constraints and then explain the loss function of CF-VAE in detail. The architecture of CF-VAE is shown in Fig. 3.

Learning Representations with Causal Constraints. We aim to retain causal relationships between "concepts" through a more easily accessible conceptual level causal graph $\mathcal{G}_c$ and embed these relationships in representations.

To formalise causal relationships, we consider n "concepts" in the dataset, which means $\mathbf{Z}_{\mathbf{x}}^{'}$ should have the same dimension as "concepts". The "concepts" in observations are causally structured by $\mathcal{G}_c$ with an adjacency matrix $\mathbf{C}$. For simplicity, in this paper, the causal constraints are exactly implemented by a linear structural equation model: $\mathbf{Z}_{\mathbf{x}}^{'} = (\mathbf{I} - \mathbf{C}^T)^{-1}\mathbf{Z}_{\mathbf{x}}$, where $\mathbf{I}$ is the identity matrix, $\mathbf{Z}_{\mathbf{x}}$ is obtained from the encoder, $\mathbf{Z}_{\mathbf{x}}^{'}$ is constructed from $\mathbf{Z}_{\mathbf{x}}$ and $\mathbf{C}$. $\mathbf{C}$ is obtained from $\mathcal{G}_c$ with respect to domain knowledge. The parameters in $\mathbf{C}$ indicate that there are corresponding edges, and the values of the parameters indicate the weight of the causal relationships. It is worth noting that if the parameter value is zero, it means that such an edge does not exist, i.e., no causal relationship between these two "concepts".

As mentioned above, $\mathbf{Z}_{\mathbf{x}}$ is obtained from the encoder, we cannot guarantee that each attribute inside is independent. To ensure the independence of each attribute in $\mathbf{Z}_{\mathbf{x}}$, we employ the total correction regularisation (TCR) in our loss function. TCR also encourages the correctness of structured $\mathbf{Z}_{\mathbf{x}}^{'}$ with respect to domain knowledge since it guarantees that there are no relationships between

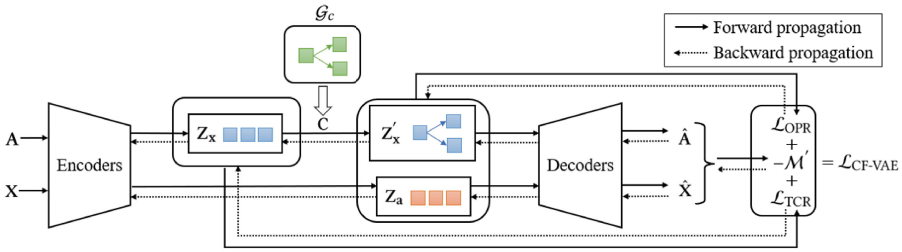


Fig. 3. The architecture of CF-VAE.

each attribute in $\mathbf{Z_x}$ before adding causal constraints. The TCR for our proposed CF-VAE is defined as, $\mathcal{L}_{\text{TCR}} = \gamma D_{KL}[q(\mathbf{Z_x})||\prod_{i=1}^{D_{\mathbf{Z_x}}} q(Z_{\mathbf{x}_i})]$, where γ is the weight value, $D_{\mathbf{Z_x}}$ is dimension of $\mathbf{Z_x}$.

Learning Strategy. We first explain the Evidence lower bound (ELBO) with causal constraints. Then, we add orthogonality promoting regularisation (OPR) to obtain the loss function of CF-VAE. Given the training samples, the parameters can be optimised by maximising the following ELBO:

$$\begin{aligned}\mathcal{M} = \ & \mathbb{E}_{q(\mathbf{Z_a}|\mathbf{A})}[\log p(\mathbf{A}|\mathbf{Z_a})] + \mathbb{E}_{q(\mathbf{Z'_x}|\mathbf{X})}[\log p(\mathbf{X}|\mathbf{Z'_x})] \\ & - D_{KL}[q(\mathbf{Z_a}|\mathbf{A})||p(\mathbf{Z_a})] - D_{KL}[q(\mathbf{Z'_x}|\mathbf{X})||p(\mathbf{Z'_x})],\end{aligned}$$

where $p(\mathbf{Z'_x}) = (\mathbf{I} - \mathbf{C}^T)^{-1}p(\mathbf{Z_x})$; $p(\mathbf{X}|\mathbf{Z'_x}) = \prod_{i=1}^{D_{\mathbf{X}}} p(X_i|\mathbf{Z'_x})$;
$q(\mathbf{Z'_x}|\mathbf{X}) = \prod_{i=1}^{D_{\mathbf{Z'_x}}} \mathcal{N}(\mu = \hat{\mu}_{Z'_{\mathbf{x}_i}}, \sigma^2 = \hat{\sigma}^2_{Z'_{\mathbf{x}_i}})$.

Then, we introduce orthogonality to encourage disentanglement between $\mathbf{Z_a}$ and $\mathbf{Z'_x}$. We employ orthogonality promoting regularisation based on the pairwise cosine similarity among latent representations: if the cosine similarity is close to zero, then the latent representations are closer to being orthogonal and independent [29]. The orthogonality promoting regularisation (OPR) for our proposed CF-VAE is defined as, $\mathcal{L}_{\text{OPR}} = \frac{1}{B}\sum_{i=1}^{B} \frac{\mathbf{Z_a}_i{}^T \mathbf{Z'_x}_i}{\|\mathbf{Z_a}_i\|_2 \ \|\mathbf{Z'_x}_i\|_2}$, where B denotes the batch size for neural network, $\|\cdot\|_2$ is the l_2 norm.

In conclusion, the loss function of our proposed CF-VAE is defined as:

$$\mathcal{L}_{\text{CF-VAE}} = -\mathcal{M} + \mathcal{L}_{\text{TCR}} + \mathcal{L}_{\text{OPR}}.$$

4 Experiments

In this section, we conduct extensive experiments to evaluate CF-VAE on real-world datasets. Before showing the detailed results, we first present the details of selected methods and the evaluation metrics. The code is available at https://github.com/IRON13/CF-VAE.

4.1 Framework Comparison

The proposed CF-VAE is considered as a pre-processing technique to address fairness issues. Hence, we compare CF-VAE with traditional and VAE-based pre-processing methods. For traditional methods, we select baselines including ReWeighting (RW) [13], Disparate Impart Remover (DIR) [9] and Optimized Preprocessing (OP) [2]. For VAE-based methods, we compare with VFAE [20] and FFVAE [6]. We also obtain the Full model for comparison, which uses all attributes in the dataset to make predictions.

We select several well-known predictive models to simulate the downstream prediction process. Linear Regression ($\mathrm{LR_R}$), Stochastic Gradient Descent Regression ($\mathrm{SGD_R}$) and Multi-layer Perceptron Regression ($\mathrm{MLP_R}$) are used for regression tasks; Logistic Regression ($\mathrm{LR_C}$), Stochastic Gradient Descent Classification ($\mathrm{SGD_C}$) and Multi-layer Perceptron Classification ($\mathrm{MLP_C}$) are used for classification tasks. For each predictive model, we run 10 times and record the mean and variance of the results for evaluation metrics.

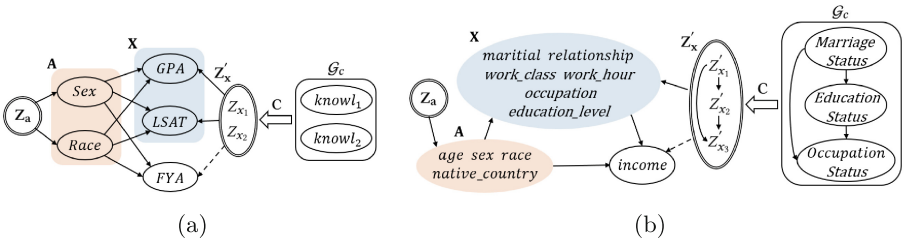
4.2 Evaluation Metrics

Fairness. There are no metrics to quantify counterfactual fairness since we can only obtain real-world data. Thus, we propose the situation test to measure fairness for different predictive models. In our experiment, we define unfairness score (UFS) to measure the result of the situation test. Specifically, the form of score differs for different predictive models. For regression tasks, we define $\mathrm{UFS_R} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{Y}_{\mathbf{A}\leftarrow\mathbf{a}}(\mathbf{Z}'_{\mathbf{x}_i}) - \widehat{Y}_{\mathbf{A}\leftarrow\bar{\mathbf{a}}}(\mathbf{Z}'_{\mathbf{x}_i})\right)^2}$; For classification tasks, we define $\mathrm{UFS_C} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{xor}\left(\widehat{Y}_{\mathbf{A}\leftarrow\mathbf{a}}(\mathbf{Z}'_{\mathbf{x}_i}), \widehat{Y}_{\mathbf{A}\leftarrow\bar{\mathbf{a}}}(\mathbf{Z}'_{\mathbf{x}_i})\right)$ (N is the number of samples for evaluation). The lower UFS value means that the predictive models achieve higher fairness performance.

Accuracy. We evaluate the performance on prediction with the following metrics. For regression tasks, we use Root Mean Square Error (RMSE) to compare the error between prediction results and target attributes' values. For classification tasks, we use accuracy to evaluate various predictive models.

4.3 Law School

The law school dataset comes from a survey [28] of admissions information from 163 law schools in the United States. It contains information of 21,790 law students, including their entrance exam scores (LSAT), their grade point average (GPA) collected prior to law school, and their first-year average grade (FYA). The school expects to predict if the applicants will have a high FYA. Gender and race are sensitive attributes in this dataset, and the school also wants to ensure that predictions are not affected by sensitive attributes. However, LSAT, GPA and FYA scores may be biased due to socio-environmental factors. We use the same $\mathcal{G}_c$ as shown in work [17] to model latent "concepts" of GPA and $LSAT$. The process of CF-VAE for the Law school dataset is shown in Fig. 4a.



(a) (b)

Fig. 4. (a) The process of CF-VAE for Law school dataset. (b) The process of CF-VAE for Adult dataset.

Table 1. The results for Law School dataset. The best fairness aware RMSE and the best UFS_R are shown in bold.

Model	Accuracy (RMSE) ↓			Fairness (UFS_R) ↓		
	LR_R	SGD_R	MLP_R	LR_R	SGD_R	MLP_R
Full	0.865(0.007)	0.867(0.007)	0.865(0.007)	0.660(0.019)	0.762(0.019)	0.760(0.045)
RW	0.955(0.013)	0.956(0.012)	0.953(0.012)	0.067(0.002)	0.067(0.001)	0.079(0.003)
DIR	0.943(0.009)	0.944(0.009)	0.941(0.010)	0.060(0.001)	0.060(0.001)	0.070(0.002)
OP	0.959(0.011)	0.960(0.011)	0.956(0.010)	0.047(0.001)	0.046(0.001)	0.055(0.003)
VFAE	0.932(0.007)	0.933(0.007)	0.934(0.007)	0.035(0.010)	0.074(0.017)	0.096(0.010)
FFVAE	0.933(0.005)	0.934(0.004)	0.935(0.005)	0.032(0.007)	0.060(0.022)	0.097(0.008)
CF-VAE	**0.931(0.006)**	**0.932(0.006)**	**0.932(0.006)**	**0.013(0.006)**	**0.025(0.011)**	**0.044(0.006)**

Results. As shown in Table 1, since the Full model uses sensitive attributes to make predictions, inverting sensitive attributes has the highest impact on the individual's prediction results, which means that the model is unfair. RW, DIR and OP achieves fair predictions by modifying the dataset compared to the Full model. Both VFAE and FFVAE disentangle the sensitive attributes with latent representations, so the influence of inverting the sensitive attributes on the prediction results is small. Our method achieves the lowest UFS_R, 0.013, 0.025, and 0.044 for LR_R, SGD_R, and MLP_R respectively, which means CF-VAE disentangle $\mathbf{Z}_x^{'}$ and $\mathbf{Z_a}$ more precisely.

For accuracy results, the Full model uses sensitive information to more accurately predict FYA and thus achieves the highest accuracy. The proposed CF-VAE achieves the best fairness aware accuracy in all predictive models than other methods.

4.4 Adult

The Adult dataset comes from the UCI repository [7] contains 14 attributes including race, age, education information, marital information as well as capital gain and loss for 48,842 individuals. We use the same $\mathcal{G}_c$ as shown in previous research [4,24] to model the latent "concepts". The adjacency matrix $\mathbf{C}$ is defined as: $\mathbf{C} = \begin{vmatrix} 0 & \lambda_{12} & \lambda_{13} \\ 0 & 0 & \lambda_{23} \\ 0 & 0 & 0 \end{vmatrix}$. Then, we construct $\mathbf{Z_x^{'}}$ from $\mathbf{Z_x}$ and $\mathbf{C}$ as follows:

Table 2. The results for Adult dataset. The best fairness aware accuracy and the best UFS_C are shown in bold.

Model	Accuracy ↑			Fairness (UFS_C) ↓		
	LR_C	SGD_C	MLP_C	LR_C	SGD_C	MLP_C
Full	0.802(0.002)	0.803(0.004)	0.831(0.004)	0.068(0.003)	0.060(0.018)	0.034(0.009)
RW	0.797(0.001)	0.792(0.002)	0.819(0.001)	0.038(0.001)	0.029(0.002)	0.052(0.001)
DIR	0.800(0.001)	0.793(0.003)	0.817(0.001)	0.035(0.001)	0.027(0.002)	0.046(0.001)
OP	0.780(0.002)	0.779(0.003)	0.783(0.002)	0.032(0.003)	0.030(0.004)	0.033(0.005)
VFAE	0.785(0.001)	0.781(0.003)	0.819(0.004)	0.062(0.002)	0.041(0.010)	0.025(0.003)
FFVAE	0.785(0.003)	0.782(0.001)	0.814(0.005)	0.062(0.001)	0.044(0.010)	0.032(0.010)
CF-VAE	**0.801(0.002)**	**0.794(0.004)**	**0.820(0.002)**	**0.031(0.002)**	**0.020(0.006)**	**0.024(0.004)**

$Z^{'}_{x_1} = Z_{x_1}$; $Z^{'}_{x_2} = \lambda_{12} Z_{x_1} + Z_{x_2}$; $Z^{'}_{x_3} = \lambda_{13} Z_{x_1} + \lambda_{23} Z_{x_2} + Z_{x_3}$. We set parameter $\{\lambda_{12} = 1, \lambda_{13} = 1, \lambda_{23} = 1\}$ to denote that edges within latent representations, i.e., $Z^{'}_{x_1} \rightarrow Z^{'}_{x_2}, Z^{'}_{x_1} \rightarrow Z^{'}_{x_3}, Z^{'}_{x_2} \rightarrow Z^{'}_{x_3}$. The process of CF-VAE is shown in Fig. 4b.

Results. The fairness results are shown in Table 2, the Full model achieves the worst UFS_C, since it use $\mathbf{A}$ to predict *income*. Both baseline fairness models and other VAE-based methods improve fairness to a certain extent. The proposed CF-VAE achieves the best UFS_C, only 3.1%, 2.0% and 2.4% of individuals' results are affected by sensitive attributes' values inversions in LR_C, SGD_C and MLP_C, respectively. Our method achieves better fairness performance than other methods, since it remains causal relationships in latent representations with respect to $\mathcal{G}_c$ and disentangles structured representations with sensitive attributes.

In order to achieve fairness, VFAE and FFVAE lose about 2% of their accuracy performance. RW, DIR and OP modify the dataset resulting in a loss of predictive performance. The proposed CF-VAE not only guarantees the fairness performance but also retains the causal relationships to improve accuracy. CF-VAE loses less information than other VAE-base methods and achieves the best fairness aware accuracy performance in all predictive models, i.e., 80.1%, 79.4% and 82.0% in LR_C, SGD_C and MLP_C, respectively.

4.5 Ablation Study

We follow the same procedure in [3] to generate synthetic datasets and conduct an ablation study to validate the contribution of each component in our method as shown in Table 3.

The Full model (Model i) uses all the observed attributes to train the predictors. The predictors achieve the best accuracy but the worst fairness performance. VFAE (Model ii) is the basic VAE-based unsupervised fair representation learning method. We set it to be the baseline. Model iii is CF-VAE without adding causal constraints, which achieves similar results as VFAE since both methods remove sensitive information from the learnt representations.

Table 3. The results of ablation study. The best fairness aware RMSE and the best UFS_R are shown in bold, and the runner-up results are underlined.

Model	Accuracy (RMSE) ↓			Fairness (UFS_R) ↓		
	LR_R	SGD_R	MLP_R	LR_R	SGD_R	MLP_R
i	0.078(0.001)	0.081(0.001)	0.081(0.001)	0.102(0.001)	0.098(0.001)	0.106(0.002)
ii	0.126(0.002)	0.126(0.002)	0.145(0.002)	0.006(0.001)	0.010(0.002)	0.104(0.005)
iii	0.125(0.001)	0.125(0.001)	0.145(0.001)	0.007(0.001)	0.011(0.003)	0.105(0.003)
iv	**0.109(0.001)**	0.111(0.001)	0.122(0.002)	0.003(0.001)	**0.004(0.002)**	0.071(0.002)
v	**0.109(0.001)**	**0.110(0.001)**	**0.121(0.001)**	**0.002(0.001)**	0.005(0.002)	**0.070(0.002)**

Then, we employ causal constraints and add TCR in the loss function as Model iv, which retains causal relationships in latent representations and improves both accuracy and fairness performance than previous models. Model v (a.k.a. CF-VAE) is to encourage $\mathbf{Z}_{\mathbf{x}}^{'}$ and $\mathbf{Z_a}$ are disentangled by adding OPR. As shown in Table 3, CF-VAE achieves the best accuracy performance and UFS_R among most predictive.

5 Related Works

The machine learning literature has increasingly focused on exploring how algorithms can protect marginalised populations from unfair treatment. An important research area is how to quantify fairness, which can be divided into two categories, the statistical framework and the causal framework.

In the statistical framework, Demographic parity was defined by [31], which is used to measure group-level fairness. Other similar metrics include equalised odds [11], predictive rate parity [30]. Dwork et al. [8] proposed a measurement to quantify individual-level fairness, that is, similar individuals should have similar treatments, and they use distance functions to measure how similar between individuals. In the causal framework, the (conditional) average causal effect is used to quantify fairness between groups [19]; Natural direct and natural indirect effects are used to quantify specific fairness [24,33]; When unfair causal paths are identified by domain knowledge, Chiappa [4] used the path-specific causal effects to quantify fairness on approved paths. For more related works, please refer to the literature review [23,32].

Our work is related to learning fair representations, which aims to encode data information into a lower space while removing sensitive information. VAE [15] and β-VAE [12] have inspired several studies in fair representation learning. Louizos et al. [20] first introduced VAE for learning fair representation to disentangle the sensitive information and non-sensitive information, they proposed a semi-supervised method to encourage disentanglement by using "Maximum Mean Discrepancy" (MMD). However, the organisations that collect the data cannot predict the downstream uses of the data and the models that might be used [10,31]. Due to this, many following up works [6,20] focus on unsupervised learning fair representation. But these works only focus on

correlation-based constraints to ensure fairness. Our approach combines counterfactual fairness and unsupervised fair representation learning to provide the proper representations. Furthermore, we innovatively embed domain knowledge into representations by adding causal constraints with respect to domain knowledge.

6 Conclusion

In this paper, we investigate unsupervised counterfactually fair representation learning and propose a novel method named CF-VAE which considers causal relationships with respect to domain knowledge. We theoretically demonstrate that the structured representations obtained by CF-VAE enable predictive models to achieve counterfactual fairness. Experimental results on real-world datasets show that CF-VAE achieves better accuracy and fairness performance on downstream predictive models than the benchmark fairness methods. Ablation study on synthetic datasets shows that causal constraints with total correction regularisation achieve better accuracy performance and orthogonality promoting regularisation encourages disentanglement with sensitive attributes.

Acknowledgements. This work has received partial support from the Australian Research Council Discovery Project (DP200101210) to J. Li, J. Liu and K. Wang, the discovery grant from the Natural Sciences and Engineering Research Council of Canada to K. Wang, and the University Presidents Scholarship (UPS) of the University of South Australia to Z. Xu.

References

1. Brennan, T., Dieterich, W., Ehret, B.: Evaluating the predictive validity of the compass risk and needs assessment system. Crim. Just. Behav. **36**(1), 21–40 (2009)
2. Calmon, F.P., Wei, D., Vinzamuri, B., Ramamurthy, K.N., Varshney, K.R.: Optimized pre-processing for discrimination prevention. In: NeurIPS, pp. 3992–4001 (2017)
3. Cheng, D., Li, J., Liu, L., Yu, K., Le, T.D., Liu, J.: Toward unique and unbiased causal effect estimation from data with hidden variables. IEEE Trans. Neural Netw. Learn. Syst., 1–13 (2022)
4. Chiappa, S.: Path-specific counterfactual fairness. In: AAAI, pp. 7801–7808 (2019)
5. Coşer, A., Maer-matei, M.M., Albu, C.: Predictive models for loan default risk assessment. Econ. Comput. Econ. Cybern. Stud. Res. **53**(2), 149–165 (2019)
6. Creager, E., et al.: Flexibly fair representation learning by disentanglement. In: ICML, pp. 1436–1445 (2019)
7. Dua, D., Graff, C.: UCI machine learning repository (2017)
8. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.S.: Fairness through awareness. In: ITCS, pp. 214–226 (2012)
9. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: SIGKDD, pp. 259–268 (2015)
10. Gitiaux, X., Rangwala, H.: Learning smooth and fair representations. In: AISTATS, pp. 253–261 (2021)

11. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: NeurIPS, pp. 3315–3323 (2016)
12. Higgins, I., et al.: beta-vae: learning basic visual concepts with a constrained variational framework. In: ICLR, pp. 1–22 (2017)
13. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. **33**(1), 1–33 (2012)
14. Kim, S., Joshi, P., Kalsi, P.S., Taheri, P.: Crime analysis through machine learning. In: IEMCON, pp. 415–420 (2018)
15. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR, pp. 1–14 (2014)
16. Kruppa, J., Schwarz, A., Arminger, G., Ziegler, A.: Consumer credit risk: individual probability estimates using machine learning. Expert Syst. Appl. **40**(13), 5125–5131 (2013)
17. Kusner, M.J., Loftus, J.R., Russell, C., Silva, R.: Counterfactual fairness. In: NeurIPS, pp. 4066–4076 (2017)
18. Lewis, D.: Counterfactuals. John Wiley & Sons, Hoboken (2013)
19. Li, J., Liu, J., Liu, L., Le, T.D., Ma, S., Han, Y.: Discrimination detection by causal effect estimation. In: IEEE BigData, pp. 1087–1094 (2017)
20. Louizos, C., Swersky, K., Li, Y., Welling, M., Zemel, R.S.: The variational fair autoencoder. In: ICLR, pp. 1–11 (2016)
21. Mac, R.: Facebook apologizes after ai puts 'primates' label on video of black men (2021). https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html
22. Madras, D., Creager, E., Pitassi, T., Zemel, R.S.: Fairness through causal awareness: learning causal latent-variable models for biased data. In: FAT*, pp. 349–358 (2019)
23. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6), 115:1–115:35 (2021)
24. Nabi, R., Shpitser, I.: Fair inference on outcomes. In: AAAI, pp. 1931–1940 (2018)
25. Park, S., Hwang, S., Kim, D., Byun, H.: Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment. In: AAAI, pp. 2403–2411 (2021)
26. Pearl, J.: Causality. Cambridge University Press, Cambridge (2009)
27. Spirtes, P., Glymour, C.N., Scheines, R., Heckerman, D.: Causation, Prediction, and Search. MIT press, Cambridge (2000)
28. Wightman, L.F.: Lsac national longitudinal bar passage study. lsac research report series (1998)
29. Xie, P., Wu, W., Zhu, Y., Xing, E.P.: Orthogonality-promoting distance metric learning: convex relaxation and theoretical analysis. In: ICML, pp. 5399–5408 (2018)
30. Zafar, M.B., Valera, I., Gomez-Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: WWW, pp. 1171–1180 (2017)
31. Zemel, R.S., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: ICML, pp. 325–333 (2013)
32. Zhang, L., Wu, X.: Anti-discrimination learning: a causal modeling-based framework. Int. J. Data Sci. Anal. **4**(1), 1–16 (2017)
33. Zhang, L., Wu, Y., Wu, X.: A causal framework for discovering and removing direct and indirect discrimination. In: IJCAI, pp. 3929–3935 (2017)