Fairness-Aware Unsupervised Feature Selection

Xiaoying Xing[1], Hongfu Liu[2], Chen Chen[3], Jundong Li[3]
[1]Tsinghua University, Beijing, China 100084
[2]Brandeis University, Waltham, MA, USA 02453
[3]University of Virginia, Charlottesville, VA, USA 22904
{xingxy0505, chenannie45}@gmail.com, hongfuliu@brandeis.edu, jundong@virginia.edu

ABSTRACT

Feature selection is a prevalent data preprocessing paradigm for various learning tasks. Due to the expensive cost of acquiring supervision information, unsupervised feature selection sparks great interests recently. However, existing unsupervised feature selection algorithms do not have fairness considerations and suffer from a high risk of amplifying discrimination by selecting features that are over associated with protected attributes such as gender, race, and ethnicity. In this paper, we make an initial investigation of the fairness-aware unsupervised feature selection problem and develop a principled framework, which leverages kernel alignment to find a subset of high-quality features that can best preserve the information in the original feature space while being minimally correlated with protected attributes. Specifically, different from the mainstream in-processing debiasing methods, our proposed framework can be regarded as a model-agnostic debiasing strategy that eliminates biases and discrimination before downstream learning algorithms are involved. Experimental results on real-world datasets demonstrate that our framework achieves a good trade-off between feature utility and promoting feature fairness.

CCS CONCEPTS

• **Computing methodologies** → **Unsupervised learning**; **Feature selection**.

KEYWORDS

Fairness, Feature Selection, Unsupervised Learning

ACM Reference Format:
Xiaoying Xing[1], Hongfu Liu[2], Chen Chen[3], Jundong Li[3]. 2021. Fairness-Aware Unsupervised Feature Selection. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3459637.3482106

1 INTRODUCTION

Feature selection is an effective data preprocessing strategy for various learning tasks [10, 15]. As it gives learning models better readability and interpretability by maintaining the physical meanings of original features, it is often preferred in high-stake applications [12, 20, 25]. Traditional feature selection algorithms can be mainly categorized as supervised and unsupervised methods [15]. As supervision information is often costly to amass, unsupervised methods have attracted increasing attention. However, most of the existing algorithms do not have fairness considerations and may exhibit discriminatory actions toward specific groups by over associating protected attributes (e.g., gender, race) [5, 8, 22]. Though it is intuitive to manually remove the protected attributes to avoid direct discrimination, some non-protected attributes that are highly correlated with the protected attributes may still cause unintentional discrimination problems (e.g., residential zip code may indicate the race information because of the residential areas) [13, 29].

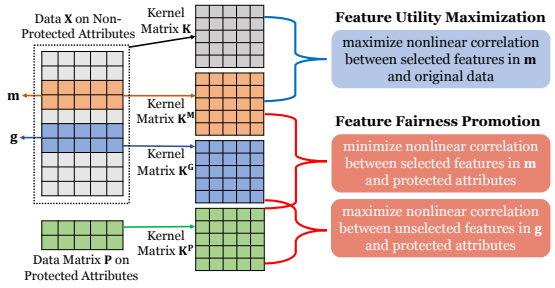


Figure 1: An illustration of the proposed fairness-aware unsupervised feature selection framework FUFS.

In this paper, we make an initial investigation of the fairness issues of unsupervised feature selection and develop a general model-agnostic debiasing solution. Our efforts have the potential to alleviate unwanted biases before applying downstream learning algorithms and are complementary to the mainstream in-processing algorithmic fairness research [22]. However, the problem is non-trivial with the following challenges. (1) Feature selection should achieve a good trade-off between fairness and feature utility. However, without label information, we are in short of effective evaluation criteria to quantify these two targets simultaneously. (2) Due to the trade-off between utility and fairness, it is difficult to achieve the maximums of both. It is necessary to explicitly exclude the features which have strong correlations with protected attributes.

To tackle the challenges above, we propose a novel Fairness-aware Unsupervised Feature Selection (FUFS) framework (as shown in Fig. 1). To ensure that the selected features do not cause much utility loss, we select features that can maximally preserve the original information. Additionally, we impose fairness constraints to enforce the protected attributes being minimally correlated with the selected features while over associating with a small number of unselected features. All the considerations are modeled in a joint optimization framework. The major contributions of our work

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482106

are as follows: (1) We address a crucial and newly emerging problem, fairness-aware unsupervised feature selection. (2) We propose a novel FUFS framework, which selects high-quality features by preserving original information and obeying the fairness considerations. (3) We formulate two desiderata of fairness-aware unsupervised feature selection (i.e., utility maximization and fairness promotion) as an optimization problem with a principled solution. (4) We validate the selected features by utility and fairness measurements and corroborate the superiority of our proposed framework.

2 THE PROPOSED FRAMEWORK - FUFS

In this work, we assume there are n data instances, the matrix $\mathbf{P} \in \mathbb{R}^{p\times n}$ denotes the set of p protected attributes for instances (e.g., age, gender, and race), and the matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$ denotes the set of d non-protected attributes (often we have $p \ll d$).

Problem Definition (Fairness-Aware Unsupervised Feature Selection). *Given the input data $\mathbf{X} \in \mathbb{R}^{d\times n}$ and $\mathbf{P} \in \mathbb{R}^{p\times n}$ with d non-protected attributes and p protected attributes, the problem aims to select a subset of k features among d non-protected attributes ($k \ll d$) which can maximally preserve the information in the original feature space while being minimally correlated with the protected attributes.*

2.1 Maximizing Feature Utility

In an unsupervised scenario, we need to seek alternative evaluation criteria to assess the importance of features. To ensure that the selected features can well capture the information embedded in the original feature space, we would like to maximize the correlation between the selected features and the original ones. However, since the original features could be high-dimensional, complex nonlinear correlations could exist between these two feature spaces. Hence, we measure their nonlinear correlation with kernel alignment [7].

Suppose the vector $\mathbf{m} \in \{0,1\}^d$ is the feature selection indicator vector such that $\mathbf{1}^\top\mathbf{m} = k$, where $m_i = 1$ if the i-th feature is selected, otherwise $m_i = 0$. The data matrix on the selected features can be obtained as $\mathbf{M} = \mathrm{diag}(\mathbf{m})\mathbf{X}$. Then we define a kernel κ which implicitly computes the similarity between instances in a high-dimensional reproducing kernel Hilbert space (RKHS) [1], such that $\mathbf{K}_{ij} = \kappa(\mathbf{X}_{*i}, \mathbf{X}_{*j})$ and $\mathbf{K}^{\mathbf{M}}_{ij} = \kappa(\mathbf{M}_{*i}, \mathbf{M}_{*j})$. In practice, we can choose polynomial kernel or RBF kernel. Denoting the centering matrix as $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$, these two kernel matrices after centering can be denoted as $\mathbf{K}_c = \mathbf{HKH}$ and $\mathbf{K}^{\mathbf{M}}_c = \mathbf{HK}^{\mathbf{M}}\mathbf{H}$, respectively. Then we can characterize the inherent nonlinear correlation between these two feature spaces with the centered kernel alignment:

$$\rho\left(\mathbf{K}, \mathbf{K}^{\mathbf{M}}\right) = \mathrm{Tr}(\mathbf{K}_c\mathbf{K}^{\mathbf{M}}_c) = \mathrm{Tr}(\mathbf{HKHHK}^{\mathbf{M}}\mathbf{H}). \tag{1}$$

With the observation that $\mathbf{HH} = \mathbf{H}$ and $\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$ (where $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$), we can further simplify $\rho\left(\mathbf{K}, \mathbf{K}^{\mathbf{M}}\right)$ as $\mathrm{Tr}(\mathbf{HKHK}^{\mathbf{M}})$. Our goal expects that the selected features in $\mathbf{m}$ can maximally preserve the information embedded in the original feature space.

2.2 Promoting Feature Fairness

Maximizing Eq. (1) alone does not address the fairness considerations as the selected features may be associated with the protected attributes in $\mathbf{P}$. Thus, we further impose fairness constraints to make the selected features in $\mathbf{M}$ not well aligned with the protected attributes $\mathbf{P}$. To achieve this goal, suppose $\mathbf{K}^{\mathbf{P}} \in \mathbb{R}^{n\times n}$ is the kernel matrix of $\mathbf{P}$, we can also leverage centered kernel alignment to minimize the nonlinear correlation between $\mathbf{M}$ and $\mathbf{P}$ in kernel space:

$$\rho\left(\mathbf{K}^{\mathbf{M}}, \mathbf{K}^{\mathbf{P}}\right) = \mathrm{Tr}(\mathbf{HK}^{\mathbf{M}}\mathbf{HK}^{\mathbf{P}}). \tag{2}$$

To further enforce that the sensitive information is eliminated in the selected features, a small number of unselected features should exhibit high correlation with the protected attributes. Hence, we further define a decomposition indicator $\mathbf{g} \in \{0,1\}^d$ to indicate the index of non-protected attributes that are highly correlated with $\mathbf{P}$, where $\mathbf{1}^\top\mathbf{g} = l$, and l denotes the number of sensitive features. Ideally, the nonzero indices of $\mathbf{g}$ should not overlap with those of $\mathbf{m}$. Hence, the data matrix $\mathbf{G}$ corresponding to $\mathbf{g}$ can be obtained as $\mathbf{G} = \mathrm{diag}(\mathbf{g})(\mathbf{I} - \mathrm{diag}(\mathbf{m}))\mathbf{X}$. Assume the corresponding kernel matrix is $\mathbf{K}^{\mathbf{G}} \in \mathbb{R}^{n\times n}$, then the centered kernel alignment can also be utilized to maximize the nonlinear correlation between $\mathbf{G}$ and $\mathbf{P}$:

$$\rho\left(\mathbf{K}^{\mathbf{G}}, \mathbf{K}^{\mathbf{P}}\right) = \mathrm{Tr}(\mathbf{HK}^{\mathbf{G}}\mathbf{HK}^{\mathbf{P}}). \tag{3}$$

2.3 Objective Function of FUFS

Combining the two desiderata of fairness-aware unsupervised feature selection, we obtain a joint constrained optimization problem:

$$\min_{\mathbf{m},\mathbf{g}} -\mathrm{Tr}\left(\mathbf{HKHK}^{\mathbf{M}}\right) + \alpha\,\mathrm{Tr}\left(\mathbf{HK}^{\mathbf{M}}\mathbf{HK}^{\mathbf{P}}\right) - \alpha\,\mathrm{Tr}\left(\mathbf{HK}^{\mathbf{G}}\mathbf{HK}^{\mathbf{P}}\right) \tag{4}$$
$$\text{s.t. } \mathbf{m}, \mathbf{g} \in \{0,1\}^d,\ \mathbf{1}^\top\mathbf{m} = k,\ \mathbf{1}^\top\mathbf{g} = l,$$

where α is a hyperparameter that can control how strong we would like to enforce the fairness of unsupervised feature selection.

The optimization problem in Eq. (5) is not joint convex w.r.t. $\mathbf{m}$ and $\mathbf{g}$ simultaneously. Although we can employ alternating optimization scheme for a local optimum, the whole optimization still remains difficult since $\mathbf{m}$ and $\mathbf{g}$ are discrete. To address this issue, we relax the discrete constraints to a real-valued vector in the range of $[0,1]$. We rewrite the optimization problem as follows:

$$\min_{\mathbf{m},\mathbf{g}} \mathcal{L} = -\mathrm{Tr}\left(\mathbf{HKHK}^{\mathbf{M}}\right) + \alpha\,\mathrm{Tr}\left(\mathbf{HK}^{\mathbf{M}}\mathbf{HK}^{\mathbf{P}}\right) - \alpha\,\mathrm{Tr}\left(\mathbf{HK}^{\mathbf{G}}\mathbf{HK}^{\mathbf{P}}\right)$$
$$+ \beta(\|\mathbf{m}\|_1 + \|\mathbf{g}\|_1) \ \text{ s.t. } \mathbf{m}, \mathbf{g} \in [0,1]^d, \tag{5}$$

where the ℓ_1-norm is introduced for the sparsity of model parameters $\mathbf{m}$ and $\mathbf{g}$. The hyperparameter β is used to control the number of selected features that are relevant and do not correlate with protected attributes and the number of unselected features that are highly correlated with protected attributes, respectively.

Updating m and g. We update two model parameters $\mathbf{m}$ and $\mathbf{g}$ alternatively until the objective function converges to a local optimum. The update rules are as follows:

$$m_i \leftarrow P[m_i - \eta\,\partial\mathcal{L}/\partial m_i],\ g_i \leftarrow P[g_i - \eta\,\partial\mathcal{L}/\partial g_i], \tag{6}$$

where $P[x]$ is a box projection operator which projects x into a bounded range. Specifically, since we relax the constraints of m_i and g_i in the range of $[0,1]$, we have $P[x] = 0$ if $x < 0$, $P[x] = 1$ if $x > 1$, and otherwise $P[x] = x$. η is the learning rate.

3 EXPERIMENTAL EVALUATIONS

3.1 Experimental Setup

Datasets. We perform experiments on four public available datasets. (1) Crime[1] combines census data, law enforcement data, and crime data of US communities. We define the percentage of population for

[1] http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized

African American as a protected attribute. We define two clusters by the number of crimes and the cutoff threshold is 0.15 crimes per 100K population. We have 2,215 communities described by 147 different attributes. (2) Adolescent[2] comes from a longitudinal study of adolescents. The attributes are personal information of the interviewees and their answers to a questionnaire. Bio-sex is regarded as the protected attribute and we define two clusters by whether their Picture Vocabulary test score is more than 65. In total, it contains 6,504 instances and 2,793 attributes. (3) Google+[3] comes from Google+, which contains user features and social relations within multiple social circles. Each instance refers to a user and attributes are obtained from personal information of users. Gender is regarded as the protected attribute. We have two clusters defined by the social circles that the users belong to without overlapping. The dataset consists of 2,437 users and 1,695 features. (4) Toxicity[4] is obtained from a Toxic Comment Classification Challenge, where each comment is considered as an instance. We apply a tokenizer to transform text data to numerical values. The identity label 'female' is regarded as the protected attribute. The features are from identity labels and comment texts. There are two clusters defined by whether the comment is regarded toxic or not. We collect a subset of 200 instances with 4,253 features.

Evaluation Criteria. For unsupervised feature selection, clustering performance is often used as an evaluation metric [15]. Specifically, we use *Clustering Accuracy (ACC)* and *Normalized Mutual Information (NMI)*, and higher values often imply better feature utility. Meanwhile, we use the widely adopted metrics *Balance* [18] and define a new fairness metric *Proportion* to quantify fairness—the selected features are considered fairer with higher value of *Balance* and lower value of *Proportion*. They are defined as follows:

$$Balance = \min_i \frac{\min_g |C_i \cap X_g|}{|C_i|}, \ Proportion = \sum_i \frac{\max_g |C_i \cap X_g|}{|C_i|}, \quad (7)$$

where C_i and X_g denote the i-th cluster and the g-th protected subgroup regarding the sensitive attribute.

Competitive Methods and Implementation. We compare our proposed framework FUFS with the following unsupervised feature selection methods that are widely used: (1) **LapScore** [11]; (2) **MCFS** [3]; (3) **UDFS** [28]; (4) **NDFS** [19]; (5) **REFS** [16].

We follow the original papers to specify the hyperparameters for the baselines. For FUFS, we set $\alpha = 1, \beta = 0.1$ on Crime and Google+ while $\alpha = 0.01, \beta = 10$ on Adolescent and Toxicity. The original distribution of the protected groups in Crime and Google+ is more unbalanced so a larger value of α is necessary to eliminate discrimination. Whereas Adolescent and Toxicity have more features and a larger value of β is necessary for sparsity. Besides, we specify the kernel function as the RBF kernel. We first apply each method to select the top-k ranked features and employ K-means clustering on the selected features. Since the results of K-means depend on initialization, we repeat K-means 50 times and report the average results. Choosing the optimal number of selected features is still an open problem, thus we follow conventional settings [15] to vary the number of selected features as {10%, 15%, ..., 40%} of the original features and report the best results.

[2]https://www.thearda.com/
[3]http://snap.stanford.edu/data/ego-Gplus.html
[4]https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/

Table 1: Results on Crime w.r.t. validity and fairness.

Method	*ACC*	*NMI*	*Balance*	*Proportion*
LapScore	0.644 (35)	0.024 (35)	0.192 (10)	1.492 (35)
NDFS	0.627 (20)	0.021 (30)	0.201 (10)	1.527 (15)
UDFS	0.728 (30)	0.150 (30)	0.107 (40)	1.456 (25)
REFS	0.774 (15)	0.082 (15)	0.208 (25)	1.552 (40)
MCFS	0.683 (25)	0.101 (20)	0.182 (20)	1.511 (25)
FUFS (ours)	0.758 (15)	0.141 (35)	0.204 (35)	1.446 (10)

Table 2: Results on Adoles. w.r.t. validity and fairness.

Method	*ACC*	*NMI*	*Balance*	*Proportion*
LapScore	0.555 (10)	0.006 (10)	0.379 (10)	1.163 (10)
NDFS	0.554 (15)	0.006 (15)	0.380 (35)	1.184 (35)
UDFS	0.556 (10)	0.007 (10)	0.359 (15)	1.184 (15)
REFS	0.544 (10)	0.004 (10)	0.380 (10)	1.184 (10)
MCFS	0.562 (10)	0.010 (10)	0.380 (15)	1.184 (15)
FUFS (ours)	0.553 (35)	0.013 (35)	0.407 (10)	1.148 (10)

Table 3: Results on Google+ w.r.t. validity and fairness.

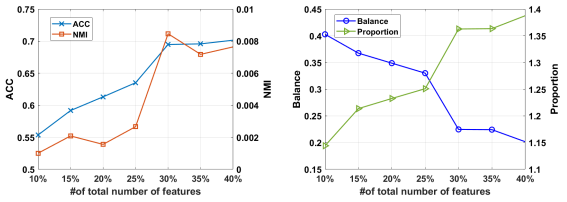
Method	*ACC*	*NMI*	*Balance*	*Proportion*
LapScore	0.723 (40)	0.114 (15)	0.004 (40)	1.865 (10)
NDFS	0.724 (40)	0.113 (40)	0.000 (10)	1.885 (15)
UDFS	0.723 (30)	0.115 (20)	0.000 (15)	1.881 (10)
REFS	0.724 (35)	0.114 (20)	0.004 (20)	1.886 (15)
MCFS	0.719 (40)	0.109 (15)	0.228 (10)	1.412 (35)
FUFS (ours)	0.721 (10)	0.164 (15)	0.301 (10)	1.308 (10)

Table 4: Results on Toxicity w.r.t. validity and fairness.

Method	*ACC*	*NMI*	*Balance*	*Proportion*
LapScore	0.803 (30)	0.012 (30)	0.009 (40)	1.568 (10)
NDFS	0.675 (40)	0.007 (40)	0.240 (20)	1.327 (15)
UDFS	0.663 (40)	0.006 (30)	0.284 (10)	1.309 (10)
REFS	0.674 (40)	0.007 (35)	0.334 (40)	1.579 (40)
MCFS	0.650 (35)	0.006 (35)	0.285 (10)	1.391 (15)
FUFS (ours)	0.701 (40)	0.008 (25)	0.409 (15)	1.136 (15)

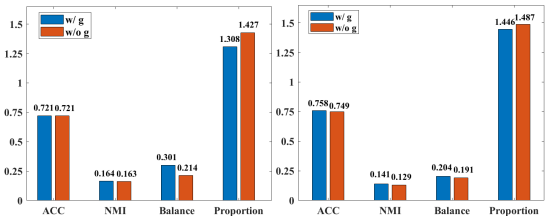
3.2 Performance Evaluation

The experimental results are shown in Tables 1-4. The number in parentheses denotes the percentage of features when the best performance is achieved. Values in red cell indicates the best result, and blue cell indicates the second best one. We make the following observations: (1) FUFS significantly outperforms the baseline methods in terms of *Balance* and *Proportion* with the best performance in almost all cases and the second best performance in terms of *Balance* on Crime. Existing unsupervised feature selection methods often do not have the fairness considerations and deliver the unfair results, while our proposed FUFS framework can obtain the most balanced clustering results across different protected subgroups. (2) FUFS achieves a good balance between feature utility and feature fairness. While achieving a good performance w.r.t. different fairness metrics, the clustering performance on the selected features is not jeopardized. For example, on Crime and Toxicity, FUFS achieves the second best performance in terms of *ACC* and *NMI* while on Adolescent and Google+, FUFS achieves the best *NMI* values and does not have obvious difference w.r.t. *ACC* compared with the best baseline method. (3) The proposed FUFS framework can achieve great performance in terms of fairness with a small number of features. Specifically, on Adolescent and Google+, FUFS achieves the best results in terms of *Balance* and *Proportion* compared with the baseline methods with merely 10% of the total number of features. On Toxicity, FUFS achieves the best results of fairness with 15% of the total number of features.



(a) Feature utility (b) Feature Fairness

Figure 2: Utility and fairness performance variation on Toxicity w.r.t. different numbers of selected features.



(a) Google+ (b) Crime

Figure 3: Clustering results w/ and w/o indicator vector g.

Table 5: Fairness results (w.r.t. *Balance* and *Proportion*) comparison based on the top-ranked features in m and g.

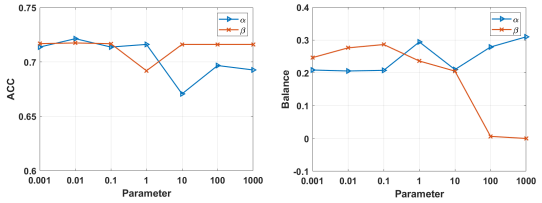
	Top-k ranked features in **m**		Top-k ranked features in **g**	
Dataset	*Balance*	*Proportion*	*Balance*	*Proportion*
Crime	0.204 (35)	1.446 (10)	0.188 (40)	1.468 (15)
Adolescent	0.407 (10)	1.148 (10)	0.308 (40)	1.179 (30)
Google+	0.301 (10)	1.308 (10)	0.000 (20)	1.863 (20)
Toxicity	0.409 (15)	1.136 (15)	0.063 (15)	1.629 (40)

3.3 In-Depth Exploration of FUFS

Effects of the Number of Selected Features. Choosing an optimal number of features is still an open problem, thus we vary the number of selected features as {10%, 15%, 20%, 25%, 30%, 35%, 40%} of the total feature number and investigate how the feature utility and fairness performance change. We only show the results on Toxicity (Fig. 2) as we have similar observations on other datasets. As we can see, *ACC* and *NMI* first increase and then keep stable when the number of selected features increases. Meanwhile, the fairness performance is the best when only 10% of features are selected (lower values of *Proportion* denotes fairer results). The fairness performance gradually decreases when the number of selected features increases, since more features correlated with sensitive features could be included in the selected feature subset.

Effects of the Decomposition Indicator Vector g. In order to investigate the effect of the vector **g**, we remove it from our framework and compare its performance with the original FUFS. The results on Crime and Google+ shown in Fig. 3 imply that the introduction of **g** improves both the utility and fairness performance. We also compare the fairness performance based on the top-ranked features in **m** and **g** as shown in Table 5. The number in parentheses denotes the percentage of features when the best performance is achieved. Obviously the clustering results based on the top-ranked features in the vector **m** are fairer than those in **g**. It shows the effectiveness of introducing the decomposition indicator vector **g**.

Parameter Study. The framework has two important hyperparameters α and β. We first fix $\beta = 0.1$ and vary α among {0.001, 0.01, 0.1, 1, 10, 100, 1000}. Next, We fix $\alpha = 1$ and vary β among {0.001, 0.01, 0.1, 1, 10, 100, 1000}. Due to space limit, we only show the results



(a) *ACC* (b) *Balance*

Figure 4: Performance variation on Google+ w.r.t. different parameter settings. X-axis is not in a linear scale.

on Google+ in terms of *ACC* and *Balance*, as shown in Fig. 4. It should be noted that the X-axis is plotted in a log scale, we do not expect to see a smooth curve. The results imply that the clustering performance is relatively stable when $\alpha = 1, \beta \in [0.001, 0.1]$ or $\alpha \in [0.001, 0.1], \beta = 0.1$. When the parameter α increases, the algorithm becomes more partial to the fairness consideration with decreasing *ACC* and increasing *Balance*. Besides, the fairness performance decreases a lot if β is specified as a very large value.

4 RELATED WORK

Unsupervised Feature Selection. Unsupervised methods often rely on alternative evaluation criteria based on characteristics of data. Specifically, similarity based methods [11, 31] select features that can best preserve the local manifold structure of data. Some methods aim to select features that can best reconstruct [16, 30] or maximally preserve the original information [27]. Many studies learn the pseudo label from data by exploiting local/global discriminative information and select features to predict these pseudo labels with $\ell_{2,1}$-norm based regression [17, 19]. Recently, data reconstruction [9, 16, 30] emerged as a new criterion to evaluate feature relevance, which evaluates the capability of features in approximating the original data through data reconstruction.

Fairness of Unsupervised Learning Methods. Here we review some related fairness topics in terms of clustering and representation learning. The initial work [4] defines fair variants of classical clustering problems such as k-center and k-median and proposes the concepts of fairlets and fairlet decomposition, which is further extended to k-means++ algorithm by [24]. Other related works focus on scalable fair clustering [2], fair spectral clustering [14], and deep fair clustering [18, 26]. Another family of work aims to learn fair representations. Fair PCA [6] is a two-step algorithm for dimension reduction. Fair Autoencoders [21, 23] encourage independence between sensitive and latent factors of variation for representation learning. Extended work [6] learns general-purpose flexible fair representations regarding multiple sensitive attributes.

5 CONCLUSION

In this paper, we addressed a novel problem of fairness-aware unsupervised feature selection and developed a principled framework FUFS. FUFS leverages the technique of kernel alignment to select high-quality features that achieve a good balance between improving downstream learning tasks and eliminating sensitive information that is highly correlated with protected attributes. These two desiderata were modeled together in a joint optimization framework. Experimental evaluations on real-world datasets demonstrated the superiority of the proposed FUFS framework in terms of feature utility and feature fairness.

REFERENCES

[1] Nachman Aronszajn. 1950. Theory of reproducing kernels. *Transactions of the American mathematical society* (1950).

[2] Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. 2019. Scalable fair clustering. *arXiv preprint arXiv:1902.03519* (2019).

[3] Deng Cai, Chiyuan Zhang, and Xiaofei He. 2010. Unsupervised feature selection for multi-cluster data. In *KDD*.

[4] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. 2017. Fair clustering through fairlets. In *NIPS*.

[5] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).

[6] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. 2019. Flexibly fair representation learning by disentanglement. *arXiv preprint arXiv:1906.02589* (2019).

[7] Nello Cristianini, Jaz Kandola, Andre Elisseeff, and John Shawe-Taylor. 2006. On kernel target alignment. In *Innovations in Machine Learning*.

[8] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. 2020. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems* (2020).

[9] A. K. Farahat, A. Ghodsi, and M. S. Kamel. 2011. An Efficient Greedy Method for Unsupervised Feature Selection. In *2011 IEEE 11th International Conference on Data Mining*. 161–170. https://doi.org/10.1109/ICDM.2011.22

[10] Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* (2003).

[11] Xiaofei He, Deng Cai, and Partha Niyogi. 2006. Laplacian score for feature selection. In *NeurIPS*.

[12] H Hannah Inbarani, Ahmad Taher Azar, and G Jothi. 2014. Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis. *Computer Methods and Programs in Biomedicine* (2014).

[13] Nathan Kallus, Xiaojie Mao, and Angela Zhou. 2019. Assessing algorithmic fairness with unobserved protected class using data combination. *arXiv preprint arXiv:1906.00285* (2019).

[14] Matthäus Kleindessner, Samira Samadi, Pranjal Awasthi, and Jamie Morgenstern. 2019. Guarantees for spectral clustering with fairness constraints. *arXiv preprint arXiv:1901.08668* (2019).

[15] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. 2017. Feature selection: a data perspective. *Comput. Surveys* (2017).

[16] Jundong Li, Jiliang Tang, and Huan Liu. 2017. Reconstruction-based Unsupervised Feature Selection: An Embedded Approach. In *IJCAI*.

[17] Jundong Li, Liang Wu, Harsh Dani, and Huan Liu. 2018. Unsupervised Personalized Feature Selection. In *AAAI*.

[18] Peizhao Li, Han Zhao, and Hongfu Liu. 2020. Deep Fair Clustering for Visual Learning. In *CVPR*.

[19] Zechao Li, Yi Yang, Jing Liu, Xiaofang Zhou, and Hanqing Lu. 2012. Unsupervised feature selection using nonnegative spectral analysis. In *AAAI*.

[20] Deron Liang, Chih-Fong Tsai, and Hsin-Ting Wu. 2015. The effect of feature selection on financial distress prediction. *Knowledge-Based Systems* (2015).

[21] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2015. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830* (2015).

[22] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[23] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. 2018. Invariant representations without adversarial training. In *NeurIPS*.

[24] Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. 2018. Fair coresets and streaming algorithms for fair k-means clustering. *arXiv preprint arXiv:1812.10854* (2018).

[25] Drishty Sobnath, Tobiasz Kaduk, Ikram Ur Rehman, and Olufemi Isiaq. 2020. Feature selection for UK disabled students' engagement post higher education: a machine learning approach for a predictive employment model. *IEEE Access* (2020).

[26] Hanyu Song, Peizhao Li, and Hongfu Liu. 2021. Deep Clustering based Fair Outlier Detection. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[27] Xiaokai Wei, Bokai Cao, and Philip S Yu. 2016. Nonlinear joint unsupervised feature selection. In *SDM*.

[28] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. 2011. $\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI*.

[29] Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. A causal framework for discovering and removing direct and indirect discrimination. *arXiv preprint arXiv:1611.07509* (2016).

[30] Zhou Zhao, Xiaofei He, Deng Cai, Lijun Zhang, Wilfred Ng, and Yueting Zhuang. 2015. Graph regularized feature selection with data reconstruction. *IEEE Transactions on Knowledge and Data Engineering* (2015).

[31] Zheng Zhao and Huan Liu. 2007. Spectral feature selection for supervised and unsupervised learning. In *ICML*.