Fair Semi-supervised Representation Learning for Tabular Data Classification

(Discussion Paper)

Shuyi Yang[1,2], Mattia Cerrato[3], Dino Ienco[4], Ruggero G. Pensa[1,*] and Roberto Esposito[1]

[1]*University of Turin, Italy*
[2]*Intesa Sanpaolo, Turin, Italy*
[3]*Johannes Gutenberg-Universität Mainz, Germany*
[4]*INRAE, UMR TETIS, Montpellier, France*

Abstract

Semi-supervised learning has shown its potential in many real-world applications where only few labeled examples are available. However, when some fairness constraints need to be satisfied, semi-supervised classification models often struggle as they are required to cope with the lack of sufficient information for predicting the target variable while forgetting its relationships with any sensitive and potentially discriminatory attribute. To address this issue, we propose a fair semi-supervised representation learning architecture that leads to fair and accurate classification results even in very challenging scenarios with few labeled (but biased) instances. We show experimentally that our model can be easily adopted in very general settings, as the learned representations may be employed to train any supervised classifier. Moreover, when applied to several real-world datasets, our method is competitive with state-of-the-art fair semi-supervised approaches.

Keywords

semi-supervised autoencoder, fairness, deep neural networks

1. Introduction

In an ideal scenario, modern supervised machine learning algorithms are able to get the most from all available training data instances so to accomplish the task at hand, be it classification, regression or ranking. Unfortunately, in real-world applications, this is almost never the case due to several reasons, among the others, the necessity to access huge amounts of labeled instances to train supervised algorithms. Labels often require cost-intensive collection procedures and huge efforts from human experts, especially in challenging domains such as medical and financial ones. Semi-supervised learning precisely addresses this issue by considering, together with a small amount of labeled information, unlabeled instances during the learning process, leveraging the so-called *smoothness* and *cluster* assumptions: if two data instances are close to each other

SEBD 2023: 31st Symposium on Advanced Database System, July 02–05, 2023, Galzignano Terme, Padua, Italy
*Corresponding author.
shuyi.yang@unito.it (S. Yang); mcerrato@uni-mainz.de (M. Cerrato); dino.ienco@inrae.fr (D. Ienco); ruggero.pensa@unito.it (R. G. Pensa); roberto.esposito@unito.it (R. Esposito)
0000-0002-8736-3132 (D. Ienco); 0000-0001-5145-3438 (R. G. Pensa); 0000-0001-5366-292X (R. Esposito)

or belong to the same cluster in the input distribution, then they are likely to belong to the same class [1, 2]. If the few available labels are of good quality, and clusters are well separated, unlabeled instances contribute to improve the accuracy significantly. Nonetheless, the labels might contain biases against certain groups. This might be an effect of historical explicit discriminations which may be reflected in a human expert's beliefs, data scarcity or even biases in the data generation/measuring process itself [3]. Beyond ethical issues, fairness in machine learning models is becoming an increasingly pressing concern at a practical level as regulators and the general public become more aware of the potential for automatic discrimination.

If the lack of labeled training instances and fairness are complex problems individually, avoiding biases in a semi-supervised learning scenario is even more challenging. In a worst-case scenario, the few available labeled instances could be all or almost all associated to unfair sources, thus leading to very biased results or preventing any debiasing process. On the other hand, unlabeled instances do not carry any explicit bias and could be useful for driving the learning algorithm towards a fairer model. Despite its clear potential, fair semi-supervised learning has not been deeply investigated. The few existing approaches are based on preprocessing strategies that seek to extract fair training datasets by leveraging unlabeled instances [4, 5]. In short, these strategies train the model on a "fair subset" of the original data [4], although it is also possible to perform pseudo-labeling over the remaining data [5]. These techniques bear some resemblance to well-known preprocessing strategies in fully-supervised fair classification [6]. However, to the best of our knowledge, no representation learning method specifically designed for semi-supervised learning with fairness constraints has been proposed so far.

Representation learning allows one to automatically construct a new feature space that better captures the different factors of variation behind the data [7]. Such new representation can then be used to feed any machine learning algorithms, including supervised and unsupervised ones. Autoencoders are among the most popular representation learning methods and both fair [8] and semi-supervised [9] versions of them have been proposed. Louizos *et al.*'s Variational Fair Autoencoder (VFAE) [10] could be employed in semi-supervised settings, in principle, however, it has only been tested in the fully supervised ones.

This paper is an extended abstract of [11], where we propose a fair semi-supervised autoencoder that leads to fair and accurate classification results even in very challenging scenarios with few labeled (but biased) instances. The classic auto-encoding architecture [12] is enhanced with two components. One is trained to classify instances and employs the available labeled training instances. The second is a debiasing component that removes as much information as possible about the sensitive attribute, in an adversarial fashion. Additionally, our model is inductive and, as such, it can be used to classify unseen examples as well. We name our contribution *FairSwiRL*, which stands for **Fair S**emi-supervised classification **wi**th **R**epresentation **L**earning.

Through an extensive experimental validation on synthetic and real world datasets, we show that the representations learned by *FairSwiRL* as the training data for different classifiers leads to reasonably accurate models while respecting the fairness constraint. Moreover, our method compares favorably to other state-of-the-art fair semi-supervised classification approaches.

2. Problem Setting

In this section, we describe the problem of semi-supervised fair classification. In this scenario, one seeks to learn a classifier by using both labeled instances and unlabeled ones. Moreover, we would also like to satisfy a fairness constraint with respect to a given sensitive attribute, i.e. a feature representing an individual's membership in an historically underprivileged group. The rationale here is to avoid potentially discriminatory decisions by the learned classifier [3].

We denote with $(\mathbf{X}_l, \mathbf{s}_l, \mathbf{y}_l)$ the features, the sensitive attributes, and the target variables of labeled instances, with $(\mathbf{X}_u, \mathbf{s}_u)$ the features and the sensitive attributes of unlabeled instances. In semi-supervised fair classification, we seek to learn a classifier which is able to leverage both $(\mathbf{X}_l, \mathbf{s}_l, \mathbf{y}_l)$ and $(\mathbf{X}_u, \mathbf{s}_u)$ such that the predictions of target variable $\mathbf{y}_t$ computed on an unseen test set $(\mathbf{X}_t, \mathbf{s}_t)$ are accurate and satisfy some fairness constraints. In the following, capital non-bold letters will be used to denote random variables (e.g., X, Y, S will denote the stochastic variables associated with examples, labels and sensitive attributes).

As a fairness constraint, we here consider independence, or *statistical parity* (SP [13, 3]). Thus, we require that the probability of assigning a positive outcome to an individual is independent of the sensitive information S. Formally, we require that $P(\hat{Y} = 1 \mid S = 0) = P(\hat{Y} = 1 \mid S = 1)$, where $\hat{Y}$ is the stochastic variable associated with the prediction of the model. As a way to quantify how far we are from the statistical parity, we consider the statistical absolute difference (SAD) measure, [14]:

$$\text{SAD} = \left| \mathbb{E}[\hat{Y} \mid S = 0] - \mathbb{E}[\hat{Y} \mid S = 1] \right|. \tag{1}$$

The lower the SAD, the better it is, with statistical parity at SAD = 0. We note here that removing the sensitive attributes $\mathbf{s}_l$ and $\mathbf{s}_u$ is usually insufficient to achieve statistical parity as some information about S may be present in the remaining variables $\mathbf{X}_l$ and $\mathbf{X}_u$ or the labels $\mathbf{y}_l$. Thus, *FairSwiRL* seeks to optimize the SAD metric by learning a debiased representation of the original data - i.e. a new representation of the data X in which all information about S has been removed. After the debiasing, any classifier trained on the latent representation will be able to achieve low SAD values without being specifically optimized for this metric. Our proposal is a semi-supervised representation learning method which is able to leverage the unlabeled examples and obtain a less biased representation of the data. We describe our contribution in detail in the next section.

3. Fair Semi-supervised Representations for Classification

In our problem setting, label scarcity is paired with fairness constraints. To face these issues, we design an inductive and fair semi-supervised model which leverages representation learning techniques. We employ an auto-encoding architecture [12] which is able to leverage both labeled $\mathbf{X}_l$ and unlabeled data $\mathbf{X}_u$. This architecture maps the original data $\mathbf{X} = \{\mathbf{X}_l \cup \mathbf{X}_u\}$ into a compact representation $\mathbf{z}$ via a series of fully-connected layers, a process which is commonly referred to as encoding. In the following we will refer to this section of our model as the encoder $E_{\theta_e}(\mathbf{x})$, where θ_e are the learnable parameters for the fully connected layers, and the learned latent representation as $\mathbf{z}$. The dimension of this representation is a hyperparameter for the algorithm and may be set up to be lower than $\mathbf{x}$, therefore compressing information. Another

series of fully connected layers, a decoder $D_{\theta_d}(\mathbf{z})$, then maps back the latent representation into an approximation $\hat{\mathbf{x}}$ of the original data. This architecture may be learned via gradient descent over a *reconstruction loss* $\mathcal{L}_{\text{rec}}$ which is defined as follows:

$$\mathcal{L}_{\text{rec}}(E_{\theta_e}, D_{\theta_d}) = \sum_{\mathbf{x}_i \in (\mathbf{X}_u \cup \mathbf{X}_l)} \|\mathbf{x}_i - D_{\theta_d}(E_{\theta_e}(\mathbf{x}_i))\|^2.$$

In the semi-supervised setting there is also the additional opportunity to exploit the limited amount of class information provided by the labeled examples $\mathbf{x}_l \in \mathbf{X}_l$. Exploiting this is paramount to obtain representations that are also useful for classification. Therefore, we employ an auxiliary network $C_{\theta_c}(\mathbf{z}_l)$ and train it on the representations $\mathbf{z}_l = E_{\theta_e}(\mathbf{x}_l)$ for which label data are available. As is commonly done in classification with neural networks, we exploit the cross entropy loss to drive the training of this component of the network:

$$\mathcal{L}_{\text{cla}}(E_{\theta_e}, C_{\theta_c}) = \sum_{\mathbf{x}_i \in \mathbf{X}_l} \left(-\sum_{j=1}^{|\mathcal{Y}|} y_{i,j} \cdot log(C_{\theta_c}(E_{\theta_e}(\mathbf{x}_i)))_j \right),$$

where the notation $y_{i,j}$ assumes the one-hot encoding of the class j for the labeled example $\mathbf{x}_i \in \mathbf{X}_l$, and $\mathcal{Y}$ is the set of possible labels (numbered from 1 to $|\mathcal{Y}|$). Lastly, we employ a component which is able to remove information about the sensitive attribute $\mathbf{s}$ from the obtained representations $\mathbf{z}$. This is possible by training another auxiliary classifier which predicts the sensitive attribute from the representation, which we will refer to in the following as F_{θ_f}. Once again, this may be trained via cross-entropy, albeit over both labeled and unlabeled examples, as we assume that sensitive information is available for all data samples:

$$\mathcal{L}_{\text{fair}}(E_{\theta_e}, F_{\theta_f}) = \sum_{\mathbf{x}_i \in \mathbf{X}_l \cup \mathbf{X}_u} \left(-\sum_{j=1}^{|\mathcal{S}|} s_{i,j} \cdot log(F_{\theta_f}(E_{\theta_e}(\mathbf{x}_i)))_j \right),$$

where $s_{i,j}$ is the j-th component of the one-hot-encoded $\mathbf{s}$ vector and $\mathcal{S}$ is the set of possible sensible values. Formally, the overall training objective for our method is as follows:

$$\mathcal{L}_{\text{tot}}(\theta_e, \theta_d, \theta_c, \theta_f) = w_{\text{cla}}\mathcal{L}_{\text{cla}}(\theta_e, \theta_c) + w_{\text{rec}}\mathcal{L}_{\text{rec}}(\theta_e, \theta_d) - w_{\text{fair}}\mathcal{L}_{\text{fair}}(\theta_e, \theta_f),$$

where w_{fair}, w_{cla}, w_{rec} are hyperparameters which may be picked to control the fairness/classification/reconstruction trade-off. The networks are pitted against one another in an *adversarial* fashion. This implies setting up a min-max game where networks E_{θ_e}, D_{θ_d} and C_{θ_c} are employed to respectively minimize the reconstruction and classification losses; the network F_{θ_f}, on the other hand, should have maximal loss, i.e., it should be impossible to reconstruct information about the sensitive attribute s from the learned representations z. This leads to the following multi-objective optimization problem:

$$\hat{\theta}_e, \hat{\theta}_d, \hat{\theta}_c, \hat{\theta}_f = \arg \left\{ \min_{\theta_e, \theta_d, \theta_c} \left[w_{\text{cla}}\mathcal{L}_{\text{cla}}(\theta_e, \theta_c) + w_{\text{rec}}\mathcal{L}_{\text{rec}}(\theta_e, \theta_d) - w_{\text{fair}} \min_{\theta_f} \mathcal{L}_{\text{fair}}(\theta_e, \theta_f) \right] \right\}.$$

The equilibrium point in the above problem can be found via *gradient reversal* [15], a procedure where the gradient information from a sub-network is multiplied by -1 when backpropagating

Table 1
Datasets used during our experiments

Dataset	instances	original features	post-processed features	sensitive attribute	target variable
ADULT	48 842	14	107	sex	income
BANK	45 211	17	61	previous campaign	subscription
CARD	30 000	23	23	education	default
COMPAS	6 127	53	18	ethnicity	criminal recidivism

into the main architecture. Specifically, we invert the gradient from F_{θ_f} when updating the parameters in our encoder E_{θ_e}.

In summary, the proposed network (*FairSwiRL*) is a fairness focused extension of the semi-supervised autoencoder. One core property of *FairSwiRL* is that it leverages representation learning to obtain feature vectors which are both useful and fair. The obtained representations may then be used for further downstream tasks with no restriction on the employed model, allowing a practitioner to use the model that best fits the domain knowledge on the task or any business requirements. We show the flexibility of our approach in Section 4.1, where we report experimental results for different classifiers trained on *FairSwiRL*'s representations.

4. Experiments

Our experiments are aimed at evaluating the representations learned by *FairSwiRL*. To this purpose, we experiment on four real-world datasets that have been extensively employed for fair classification and fair representation learning [8, 10] (see Table 1 for summary statistics).

4.1. *FairSwiRL* in combination with different supervised classifiers

Here, we compare different classifiers in combination with *FairSwiRL*, namely: random forest (*FairSwiRL*+RF), k-nearest neighbors (*FairSwiRL*+KNN), logistic regression (*FairSwiRL*+LR), support vector machines (*FairSwiRL*+SVC) and neural network (*FairSwiRL*+NN).

We now define the data splits and the evaluation metric we will employ in this section and in the rest of the paper. Let n_l, n_u, n_v, and n_t be the number of labeled, unlabeled, validation and test examples. We start with the following configuration: $n_l = 100$, $n_u = 10000$, $n_v = 100$, $n_t = 10000$ (in case of the COMPAS dataset $n_l = 100$, $n_u = 1900$, $n_v = 100$, $n_t = 1900$). We use the validation examples to find a good configuration of the hyperparameters and then, by using the same hyperparameters, we increase the number of labeled instances n_l from 100 to 2000. For each combination of (n_l, n_u, n_v, n_t) we repeat the experiments ten times by sampling different datasets from the original data, and compute the average performance metrics. We stress that the number of available examples for a given experimental run is computed in absolute terms, not relative. This lets us compare the performance of the methodologies across the same number of test examples, no matter how many labeled examples are available. To measure the fairness level we employ 1-SAD (see Equation 1) while for the predictive performance we compute the Matthews Correlation Coefficient (MCC [16, 17]).

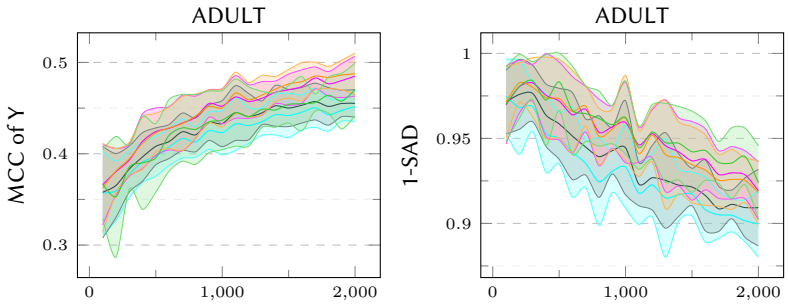

Figure 1: Performances of *FairSwiRL* for increasing number of labeled instances and in combination with different classifiers: RF, KNN, LR, SVC, NN. Lines represent the mean of the given metric over 10 repetitions, shaded area correspond to $\pm$ one standard deviation. The x-axis represents the number of labeled instances used during the training process. Best viewed in color.

In Figure 1, we report only the results for ADULT , but the behaviors of different combinations of *FairSwiRL*+classifier are similar across the datasets. We note that the trends of the different classifiers are the same in predicting the target variable and in being fair. These results show that the latent representations induced by *FairSwiRL* can be used by different classifiers and, as the number of labeled examples increases, the performances on the target variable tend to increase. While the 1-SAD value (higher is better) slightly suffers from the bias introduced by the additional examples, we note that it remains very close to optimal values (> 0.9) nonetheless.

In the next section, we will compare *FairSwiRL* with competing approaches. In order to enable a fair comparison, we choose the worst combination (*FairSwiRL*+RF, according to the previous experiment) and keep it fixed in all the experiments presented in this work.

4.2. *FairSwiRL*+RF compared to competitors

In this experiment, we test the effectiveness of *FairSwiRL* on different datasets and against different competitors. The experiment setting is the same as in Section 4.1, but we choose only the worst performing combination (*FairSwiRL*+RF) as our candidate combination. In addition to *FairSwiRL*+RF, we include the following competitors: **FESF**, an implementation of Fairness-Enhanced Sampling Framework [4]; **FairSSL**, an implementation of the algorithm presented by Chakraborty et al. [5] with Label Spreading [18] as the pseudo-labeling algorithm; **VFAE** an implementation of the Variational Fair Autoencoder [10] used to get the latent representation on which a random forest is then trained for the classification task, as in *FairSwiRL*+RF.

The results are reported in Figure 2. The plots report on the x-axis the performance metric (MCC) and on the y-axis the fairness metric (1-SAD). We vary the number of labeled examples and run the experiments ten times for each configuration. Each point in the plot represents one experiment, shapes vary according to the algorithm used and colors vary according to the number of labeled examples in the dataset. The best possible point in each plot is at coordinates (1,1), but this is usually unattainable. The gray dashed line has slope -1 and, as such, points on that line have the same trade-off between accuracy and fairness. The lines showed in each plot pass through the point closest to (1,1) under the L_1 metric. These points are, thus, the

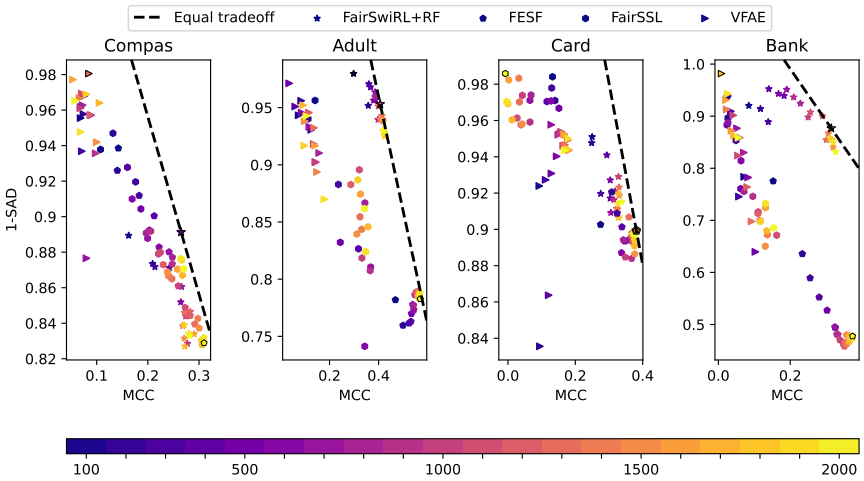

Figure 2: A comparison of *FairSwiRL*+RF to the other competitors. The plots report on the x-axis the performance metric MCC and on the y-axis the fairness metric 1-SAD (higher is better for both). Each point is the average of 10 repeated runs with the same configuration but different samples. The colors represent the number of labeled instances. Best viewed in color.

best performers under the assumption that fairness and accuracy are equally important. We can see that, with the exception of the plot concerning the COMPAS dataset, the points (★) representing *FairSwiRL*+RF are always in the upper half of the plots. Higher values of 1-SAD mean that the debiasing component of *FairSwiRL* is working as expected.

The comparisons with FairSSL (⬢), FESF (⬟) and VFAE (▶) are also favorable. Except for the CARD dataset, *FairSwiRL* lies on the optimal tradeoff line. In CARD, where the best results are attained by FESF, *FairSwiRL* has a better fairness, but the lower MCC leads the FESF model to prevail in terms of the linear trade-off we are assuming here. This is a typical case of accuracy-fairness dilemma: higher 1-SAD implies also lower predictive power when the sensitive attribute and target variable are correlated. On the COMPAS dataset we have a mixed situation, while the best points are attained by *FairSwiRL*, we can see that for some experiments (specifically, those with fewer labeled examples) it attains worse performances than the competitors. Overall, we would not judge this experiment as a clear win for *FairSwiRL*, but we still maintain that it is a competitive approach also in this case.

As far as more general trends are concerned, we observe that more labeled instances (warmer colors in Figure 2) lead all methodologies to more accurate, but less fair results. This result, in our view, justifies further future employment of semi-supervised techniques in fair classification: a small amount of labeled data does not impact fairness negatively.

Beyond the linear tradeoff discussed above, we also experiment in an hypothetical context in which fairness is paramount and performance may be pursued only when fairness is already guaranteed. To model this situation, we repeated the experiments recording the discounted MCC metric: $\text{DisMCC} = \text{MCC}_y \cdot e^{-\alpha \text{SAD}}$, where MCC_y is the MCC computed on the target variable. It is worth noting that, in this metric, the fairness performances, as measured by the SAD statistic, are weighted exponentially. Figure 3 plots the average rankings of the competing

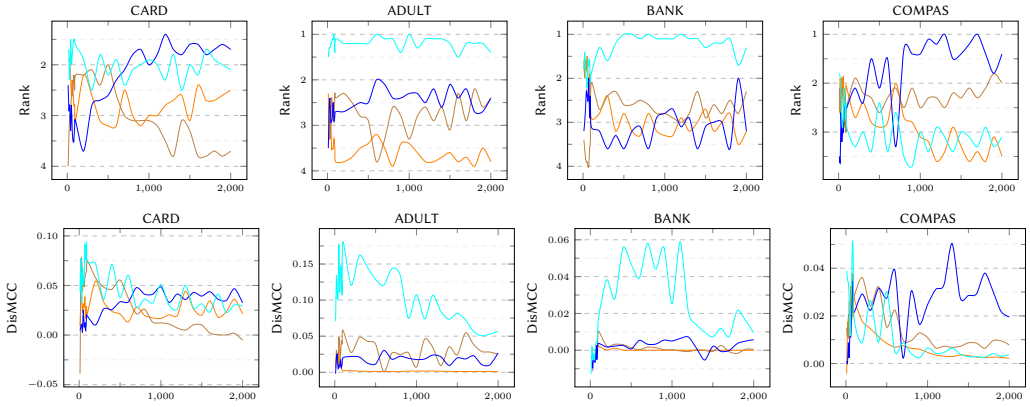

Figure 3: Performances of *FairSwiRL* and competitors for increasing number (100-2000) of labeled instances: *FairSwiRL*+RF, FairSSL, FESF, VFAE. Lines represent the average rank (on the top) and the average DisMCC (on the bottom) over 10 repetitions. Best viewed in color.

approaches for increasing number of labeled examples. Rankings are evaluated according to the value of DisMCC with $\alpha = 30$. We note that lower rankings, which are better, are displayed higher in the picture. The actual values of DisMCC obtained in the corresponding experiment are displayed in the right column (higher values are better). In SYNTHETIC the PD+RF method dominates, as expected, because it represents the theoretical upper-bound, unreachable in a real setting since the data generation process is usually unknown. However, the second best candidate is *FairSwiRL*+RF. In CARD *FairSwiRL*+RF reaches the best performance only sometimes but if compared to VFAE and FairSSL it has a more stable trajectory when the number of labeled instances changes. *FairSwiRL* is overall the strongest performer on both ADULT and BANK. In COMPAS we observe worse performances than the competitors, while the other fair representation learning strategy we tested (VFAE) is the strongest performer. Overall, even in a context where the fairness is exponentially weighted, *FairSwiRL*+RF performs well on average.

5. Conclusion

We have proposed a neural network for representation learning that addresses two challenging issues simultaneously: the lack of sufficient labeled examples in the training data, and the presence of sensitive attributes potentially leading to unfair decisions. We have shown that unlabeled examples help the learning algorithm to cope with both problems, leading to fair and accurate semi-supervised classification of unseen examples. The experiments have shown the effectiveness of our approach, even in comparison with state-of-the-art fair semi-supervised methods which employ preprocessing strategies. We have also performed a full comparison with another fair representation learning strategy (VFAE) [10] which had so far never been tested in the SSL setting. Our experiments show that fair representation learning approaches are able to outperform feature preprocessing strategies in the semi-supervised setting and such a result transfers across different tradeoffs for fairness vs. accuracy.

References

[1] O. Chapelle, B. Schölkopf, A. Zien, Introduction to semi-supervised learning, in: Semi-Supervised Learning, The MIT Press, 2006, pp. 1–12.

[2] J. E. van Engelen, H. H. Hoos, A survey on semi-supervised learning, Mach. Learn. 109 (2020) 373–440.

[3] S. Barocas, M. Hardt, A. Narayanan, Fairness and machine learning., URL: http://www.fairmlbook.org (2019).

[4] T. Zhang, T. Zhu, J. Li, M. Han, W. Zhou, P. S. Yu, Fairness in Semi-Supervised Learning: Unlabeled Data Help to Reduce Discrimination, IEEE Trans. Knowl. Data Eng. 34 (2022) 1763–1774.

[5] J. Chakraborty, H. Tu, S. Majumder, T. Menzies, Can we achieve fairness using semi-supervised learning?, CoRR abs/2111.02038 (2021).

[6] F. Kamiran, T. Calders, Classifying without discriminating, in: Proceesings of IEEE-IC4 2009, 2009.

[7] Y. Bengio, A. C. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 1798–1828.

[8] D. Madras, E. Creager, T. Pitassi, R. S. Zemel, Learning adversarially fair and transferable representations, in: Proceedings of ICML 2018, 2018, pp. 3381–3390.

[9] A. Gogna, A. Majumdar, Semi supervised autoencoder, in: Proceedings of ICONIP 2016, 2016, pp. 82–89.

[10] C. Louizos, K. Swersky, Y. Li, M. Welling, R. S. Zemel, The variational fair autoencoder, in: Proceedings of ICLR 2016, 2016.

[11] S. Yang, M. Cerrato, D. Ienco, R. G. Pensa, R. Esposito, FairSwiRL : Fair Semi-supervised Classification with Representation Learning, Mach. Learn. (2023) 1–26. URL: https://doi.org/10.1007/s10994-023-06342-9. doi:10.1007/s10994-023-06342-9.

[12] G. E. Hinton, R. S. Zemel, Autoencoders, minimum description length and helmholtz free energy, in: Proceesings of NIPS 1993, Morgan Kaufmann, 1993, pp. 3–10.

[13] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, The zoo of fairness metrics in machine learning, CoRR abs/2106.00467 (2021).

[14] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, Y. Zhang, AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, CoRR abs/1810.01943 (2018).

[15] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. S. Lempitsky, Domain-adversarial training of neural networks, J. Mach. Learn. Res. 17 (2016) 59:1–59:35.

[16] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Bioinform. 16 (2000) 412–424.

[17] D. Chicco, G. Jurman, The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation, BMC genomics 21 (2020) 1–13.

[18] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, B. Schölkopf, Learning with local and global consistency, in: Proceesings of NIPS 2003, MIT Press, 2003, pp. 321–328.