Fairness-aware User Classification in Power Grids

Ruijie Du
University of California, Irvine
Electrical Engineering and Computer Science
ruijied@uci.edu

Yanning Shen*
University of California, Irvine
Electrical Engineering and Computer Science
yannings@uci.edu

***Abstract*—With the development of intelligent measurement systems, power grids improved the reliability and efficiency according to the vast amount of collected information. Machine learning techniques are increasingly used in smart grids since they are efficient to deal with the huge amount of collected data and extract valuable information. The availability of large-scale data enables the employment of machine learning methods in various tasks in power grids. However, large-scale deployment of machine learning model relies on how trustworthy the model is. While sole pursuit of overall learning performance, may leads to unfair results. Specifically, the model may unintentionally discriminate different subgroups. Machine learning models for smart grids also have fairness concerns. Power consuming users and buildings with different power consumption patterns may be treated with different conditions. To mitigate the unfairness, we propose accuracy parity , equal opportunity and predictive equality regularizers, which can be used for different classification tasks in power grids to mitigate the corresponding performance discrepancy. Experiments on user classification using loading data show that the regularizers are effective at avoiding disparate mistreatment and sometimes can benefit the overall performance with fine tuning weights.**

I. Introduction

With the development of intelligent measurements systems and smart metering technology, migrating to an electronically controlled grid has improved the reliability and efficiency. The evolution of power grids has led to more efficient data collection process. Machine learning techniques provide an efficient way to analyze the massive amount of data and extract valuable information. By analyzing the measurements, more information can be obtained more accurately: the status of the network, the actual detailed load patterns, etc. Machine learning functionalities bring huge benefits to the grids and being used in various tasks, including predictions of consumption[1, 2], fault detection[3], etc.

As decision-making increasingly relies on machine learning and data, the issue of fairness is receiving increasing attention. In classical machine learning, when considering a classification task, the objective is to minimize a loss function(e.g., cross entropy) that reflects the errors. This approach is unable to control the distribution of errors across different subgroups. In recent years, research has pointed out plenty of evidences that decision making by machine learning models may unintentionally discriminate different subgroups and causes unfairness, especially in media and social studies. In settings such as loan approvals[4] or college admissions[5], fairness must be carefully taken into account in order to ensure the absence of discrimination. More and more fairness-aware machine learning solutions have been proposed[6, 7, 8, 9]. However, the issue of potential bias has not been considered in power grids.

While development of smart metering technology provides more convenient data collection, it also leads to potential bias since more detailed and private data may be revealed. And the prevalence of smart meters and sensors varies widely among users or locations. Machine Learning models trained from such datasets may henceforth inherit the bias in the collected data. Hence, the goal of the present work is to introduce a novel fairness-aware framework to eliminate potential bias in power grid user classification.

The paper is organized as follow: Section II introduces the problem of high load indication, and proposes the proposed fairness enhancing framework. Section III focuses on fairness-aware user type identification task. Numerical tests are presented in Section IV to evaluate the performance of the proposed schemes. And conclusions are summarized in Section V.

II. Fairness-aware high-load indication

High-load indication task refers to the task of predicting whether the load of the next time period is high-load state or not based on current and previous load information, see e.g., [10]. Knowing the loading status or behavioral categories of power consumers can better model the behavior forecast which is an important task for load balancing. The input information is the load information of T time frames, including the load information from now to $T-1$ time frames ago. And the output is the binary indicator $y \in \{0,1\}$ of high-load status at time $T+1$, where $y=1$ indicate high-load status.

Specifically, such task can be viewed as a binary classification problem, where the objective is to minimize a loss function that reflects the errors made by the classifier F. Specifically, the loss function considered in the present work is the cross-entropy loss

$$\mathcal{L}(\theta;\mathcal{D}) \tag{1}$$
$$= -\frac{1}{|\mathcal{D}|}\sum_{\mathbf{x}\in\mathcal{D}}(y\log(F(\mathbf{x};\theta))+(1-y)\log(1-F(\mathbf{x};\theta)))$$

where $\mathbf{x} \in \mathbb{R}^n$ is the input feature of a data sample and $y \in \{0,1\}$ is the ground truth label corresponding to $\mathbf{x}$. In

*Work in this paper was partially supported by Google Research Scholar Award (PI: Y. Shen).

high-load indication task, $\mathbf{x}$ represents time-series load data and y represents the load status. θ denotes the parameters of the classifier that will be trained to minimize the loss function. $F(\mathbf{x};\theta)$ is the output of the classifier which represents the predicted probability of $\mathbf{x}$ being class 1. Because the classification is binary, $1-F(\mathbf{x};\theta)$ is the predicted probability of $\mathbf{x}$ being class 0. Such loss is widely used for classification tasks including but not limited to anomaly detection [11], quality prediction[12], fault detection [3], or load indication[10][13].

A. Accuracy parity regularizer

Existing frameworks for load indication in power grids usually focus on how to design the classifier F with lower error rate, but they are usually bias-oblivious. This will lead to potential bias in the results. The classification of loading status is based on their consumption patterns which potentially have fairness issues due to the sensitive attributes, e.g. building sizes, geographical locations, user types.

Let $z \in \{1,0\}$ denote the sensitive attribute, which can denote, e.g., building size or location. Let $d(\mathbf{x})$ represent the signed distance between the feature vectors of samples and the classifier decision boundary [7]. The covariance between the sensitive attributes z and $d(\mathbf{x})$ can be written as

$$\begin{aligned}\mathrm{Cov}(z,d(\mathbf{x})) &= E[(z-\bar{z})g(y,\mathbf{x})] - E[(z-\bar{z})]\bar{g}(y,\mathbf{x}) \\ &\approx \frac{1}{|\mathcal{D}|}\sum_{(\mathbf{x},y,z)\in\mathcal{D}}(z-\bar{z})g(y,\mathbf{x}) \end{aligned} \tag{2}$$

where $\bar{z}$ and $\bar{g}(y,\mathbf{x})$ denote the average of the value z and $g(y,\mathbf{x})$ for all data samples in $\mathcal{D}$, $g(y,\mathbf{x}) := \max(0,(\frac{1}{2}-y)d(\mathbf{x}))$, $E[(z-\bar{z})]\bar{g}(y,\mathbf{x}) = 0$ since $E[(z-\bar{z})] = 0$. In neural networks for binary classification, the final decision is based on the output of last linear layer, denoted as $f(\mathbf{x})$. The final output of the classifier $F(\mathbf{x})$ is obtained as the output of the last logistic activation function with input $f(\mathbf{x})$. Hence, $\hat{y}=1$ if $f(\mathbf{x})>0$, otherwise $y=0$. The decision boundary is simply the hyperplane that $f(\mathbf{x})=0$. Hence $d(\mathbf{x}) = f(\mathbf{x}) - 0 = f(\mathbf{x})$ and $g(y,\mathbf{x}) := \max(0,(\frac{1}{2}-y)f(\mathbf{x}))$.

Splitting the sum in (2) into two terms with respect to the sensitive attribute z, we obtain

$$\begin{aligned}&\sum_{\mathcal{D}}(z-\bar{z})g(y,\mathbf{x}) \\ =&\sum_{z=0}(0-\bar{z})g(y,\mathbf{x}) + \sum_{z=1}(1-\bar{z})g(y,x)\end{aligned} \tag{3}$$

It can be readily observed that if the prediction matches the true label $\hat{y}=y$, we have $f(\mathbf{x})>0$ for $y=1$ and $f(\mathbf{x})<0$ for $y=0$. Hence, $(\frac{1}{2}-y)f(\mathbf{x})<0$. Due to the maximum operation, $g(y,\mathbf{x})=0$. Similarly, $g(y,\mathbf{x})>0$ if $\hat{y}\neq y$. Meanwhile, the term $(z-\bar{z})$ takes different signs for different sensitive groups: for the subgroup with $z=0$, $(0-\bar{z})g(y,\mathbf{x})\leq 0$; for the subgroup with $z=1$, $(1-\bar{z})g(y,\mathbf{x})\geq 0$. Hence, the addition of the two terms in (3) characterizes the difference of the mismatch accumulation between two subgroups.

If a decision boundary satisfies accuracy parity (AP) in the sense that $P(\hat{y}\neq y|z=0) = P(\hat{y}\neq y|z=1)$, the covariance will be close to zero, $\mathrm{Cov}(z,d(\mathbf{x}))\approx 0$.Therefore, the accuracy parity regularizer can be introduced as as

$$\begin{aligned}\mathcal{R}_{\mathrm{AP}} &= \left(\frac{1}{|\mathcal{D}|}\sum(z-\bar{z})g(y,\mathbf{x})\right)^2 \\ &= \left(\frac{1}{|\mathcal{D}|}\sum(z-\bar{z})\max(0,(\frac{1}{2}-y)f(\mathbf{x}))\right)^2 \\ &= \Bigg(\frac{1}{|\mathcal{D}|}\sum_{z=0}\max(0,(0-\bar{z})(y-\frac{1}{2})f(\mathbf{x})) \\ &\quad -\frac{1}{|\mathcal{D}|}\sum_{z=1}\max(0,(1-\bar{z})(\frac{1}{2}-y)f(\mathbf{x}))\Bigg)^2\end{aligned} \tag{4}$$

The penalty comes from the difference of the mismatch accumulation between the sensitive groups. If the model is more fair, the performance of two sensitive groups will be more similar and the penalty will be smaller. Hence, the problem can be formulated as :

$$\min_{\theta} \quad \mathcal{L}(\theta;\mathcal{D}) + \alpha\mathcal{R}_{\mathrm{AP}} \tag{5}$$

where $\alpha>0$ is a hyperparameter used to tradeoff between training accuracy and fairness in terms of accuracy parity.

B. Equal opportunity and predictive equality regularizers

In addition to accuracy parity criterion presented in the previous subsection, in certain scenarios, we may be interested in different fairness criteria. For example, in anomaly detection, where costly immediate actions may be taken towards the high-load states, more emphasis needs to be put on the classes that are predicted positive. In this case, fairness criteria such as equal opportunity or predictive equality need to be incorporated. Specifically, equal opportunity $\Delta_{\mathrm{EO}} := |P(\hat{y}=1|y=1,z=0) - P(\hat{y}=1|y=1,z=1)|$ characterizes the difference of true positive rate $TRP := P(\hat{y}=1|y=1)$, while predictive equality $\Delta_{\mathrm{PE}} := |P(\hat{y}=1|y=0,z=0) - P(\hat{y}=1|y=0,z=1)|$ measures the discrepancy between the false positive rate $FPR := P(\hat{y}=1|y=0)$ see e.g., [8]. Based on these two criteria, we can obtain the corresponding regularizers as

$$\mathcal{R}_{\mathrm{PE}} = \left(\frac{\sum_{z=0,y=0}f(\mathbf{x})}{N_{z=0,y=0}} - \frac{\sum_{z=1,y=0}f(\mathbf{x})}{N_{z=1,y=0}}\right)^2 \tag{6}$$

$$\mathcal{R}_{\mathrm{EO}} = (\frac{\sum_{z=0,y=1}f(\mathbf{x})}{N_{z=0,y=1}} - \frac{\sum_{z=1,y=1}f(\mathbf{x})}{N_{z=1,y=1}})^2 \tag{7}$$

Similarly, $f(\mathbf{x})$ is the output of the last linear layer. β_1 and β_2 are the weights of regularization terms which trade off between fairness and accuracy. Adding the equal opportunity and predictive equality regularizers, the problem can be formulated as:

$$\min_{\theta}\mathcal{L}(\theta;\mathcal{D}) + \beta_1\mathcal{R}_{\mathrm{PE}} + \beta_2\mathcal{R}_{\mathrm{EO}} \tag{8}$$

III. Fairness-aware user type identification

There exist large numbers of users in power grids, which fall into different classes, e.g., building types, safety or quality status. Hence, user type identification plays an important

role in many practical problems, including but not limited to accurate pricing, abnormal behavior detection. Accurate identification of the consumer types can also help estimate their future consumption which can further help the power companies balance the load of power grids and manage the demand and supply. However, existing user type classification frameworks mainly focus on imputing classification accuracy, see e.g.,[14]. While the user consumption patterns can be used for classifying the user types, it may also lead to potential bias due to underlying correlation between users' consumption patterns and their sensitive attributes such as locations, building sizes.

Since user type identification typically faces with customers from more than one category, the corresponding cross-entropy loss can be written as

$$\mathcal{L}(\theta;\mathcal{D}) = -\frac{1}{|\mathcal{D}|}\sum_{i}^{K}\sum_{\mathbf{x}\in\mathcal{D}} \mathbb{1}_{y=i} log(F_i(\mathbf{x};\theta)) \tag{9}$$

where $F(\mathbf{x};\theta) \in \mathbb{R}^K$ and $y \in \{1,\ldots,K\}$, with K denoting the number of classes. $\mathbb{1}_{y=i}$ is the identify function, which returns value 1 if $y = i$, meaning the input data belongs to class i. $F_i(\mathbf{x};\theta)$ denotes the ith entry of $F(\mathbf{x};\theta)$ which provides the predicted probability of $\mathbf{x}$ being class i. In user type identification task, $\mathbf{x}$ represents the load time-series data and y represents user types. Sensitive attribute z denotes the geographical location of the user. In this case the output of last fully connected layer is a vector of size K, $f(\mathbf{x}) \in \mathbb{R}^K$. Hence the distance to the decision boundary for class i, $1 \le i \le K$, is the ith entry of $f(\mathbf{x})$, denoted as $f_i(\mathbf{x})$.

A. Accuracy parity regularizer

Faced with multiple user classes, the accuracy parity regularizer also needs to be designed for multiple user classes. Specifically, the accuracy parity regularizer can be written as follow:

$$\begin{aligned}\mathcal{R}_{\mathrm{AP}} &= \left(\frac{1}{|D|}\sum(z-\bar{z})g(y,\mathbf{x})\right)^2 \\ &= \left[\frac{1}{|\mathcal{D}|}\sum(z-\bar{z})\max\left(0,(f_{\hat{y}(\mathbf{x})}-f_y(\mathbf{x}))\right)\right]^2.\end{aligned} \tag{10}$$

Since the final decision is based on the maximum value of $f(\mathbf{x})$, $f_{\hat{y}}(\mathbf{x})$, denotes the $\hat{y}$th entry of $f(\mathbf{x})$, is greater than any other value in the vector, $f_{\hat{y}}(\mathbf{x}) \ge f_i(\mathbf{x}), \forall i \in \{1,\ldots,K\}$. The term contributes to the sum only when the predicted label mismatches the true label. Due to the fact that $f_{\hat{y}}(\mathbf{x}) - f_y(\mathbf{x}) \ge 0$, and the equality holds if $\hat{y} = y$, meaning the predicted label correct. Therefore, $\max(0,(f_{\hat{y}}(\mathbf{x}) - f_y(\mathbf{x}))) = f_{\hat{y}}(\mathbf{x}) - f_y(\mathbf{x})$, and the max operation can be removed. The problem can be formulated as (5) with loss function in (9) and the regularization $\mathcal{R}_{\mathrm{AP}}$ in (10).

B. Equal opportunity and predictive equality regularziers

For equal opportunity and predictive equality regularizers, we treat every class as a one versus $K-1$ binary classification problem and aggregate the regularization term for K classes. For each class, the two penalizers are similar to (6) and (7) except that the distance to the decision boundary is $f_i(\mathbf{x})$ instead of $f(\mathbf{x})$. Then the EO and PE regularizer for multiple classes can be written shown as:

$$\mathcal{R}_{\mathrm{PE,i}} = \left(\frac{\sum_{z=0,y\neq i} f_i(\mathbf{x})}{N_{z=0,y\neq i}} - \frac{\sum_{z=1,y\neq i} f_i(\mathbf{x})}{N_{z=1,y\neq i}}\right)^2$$

$$\mathcal{R}_{\mathrm{EO,i}} = \left(\frac{\sum_{z=0,y=i} f_i(\mathbf{x})}{N_{z=0,y=i}} - \frac{\sum_{z=1,y=i} f_i(\mathbf{x})}{N_{z=1,y=i}}\right)^2$$

$$\mathcal{R}_{\mathrm{PE}} = \frac{\beta_1}{K}\sum_{i=0,\ldots,K-1} \mathcal{R}_{\mathrm{PE,i}} \tag{11}$$

$$\mathcal{R}_{\mathrm{EO}} = \frac{\beta_2}{K}\sum_{i=0,\ldots,K-1} \mathcal{R}_{\mathrm{EO,i}}. \tag{12}$$

Upon adding the EO and PE regularizers, the problem can be formulated as (8) by employing the loss in (9) and the regularization $\mathcal{R}_{\mathrm{EO}}$, $\mathcal{R}_{\mathrm{PE}}$ in (11), (12).

IV. Experiments

In this section, experimental results for high-load indication and user type identification are presented to evaluate the proposed fairness-aware framework. Specifically, details of data preprocessing and experimental settings will be clarified.

A. Dataset

The data was obtained from the database "Commercial and Residential Hourly Load Profiles for TMY3 Locations in the United States" [15]. It consists of hourly collected load profile data for 16 different commercial building types and residential buildings. The commercial buildings data is based on the DOE commercial reference building models and the residential buildings data is based on the Building America House Simulation Protocols. Specifically, the data consists 7 types(classes) of commercial buildings in several cities in the US. The input feature $\mathbf{x}$ is the load information of a building.

B. Fariness-aware measurements

In order to evaluate fairness performance, the equal opportunity(EO), predictive equality(PE) and accuracy parity(AP) measures are used, with $\Delta_{EO} := |P(\hat{y}=1|y=1,z=1) - P(\hat{y}=1|y=1,z=0)|$, $\Delta_{PE} := |P(\hat{y}=1|y=0,z=1) - P(\hat{y}=1|y=0,z=0)|$. Similarly, AP measures the difference between two subgroup's accuracy: $\Delta_{AP} = |P(\hat{y}=y|z=0) - P(\hat{y}=y|z=1)|$. For the user type identification task, although the number of classes increases, the Δ_{AP} stays the same. In order to measure the EO fairness among multiple classes, taking every class as a binary classification case and weighted aggregate the difference between two subgroups' TPR as EO measurement: $\Delta_{EO} = \sum_i^K w_i \Delta_{EO,i}$ where $\Delta_{EO,i} = |TPR_{y=i,z=0} - TPR_{y=i,z=1}|$ and $TPR_i = P(\hat{y}=i,y=i)/P(y=i)$. Similarly, the PE measurement for multiple classes is defined as $\Delta_{PE} := \sum_i^K w_i \Delta_{PE,i}$ where $\Delta_{PE,i} = |FPR_{y=i,z=0} - FPR_{y=i,z=1}|$ and $FPR_i = P(\hat{y}=i,y\neq i)/P(y\neq i)$. The weight of each class, $w_i := \frac{N_i}{N}$, is based on the distribution of each class across the dataset, N_i is the number of samples in class i and N is the total number of samples.

C. High load indication

The task here is predicting next hour load status of the building based on its history loading pattern. We use the long short-term memory(LSTM) network as the first layer of the model. LSTM is an artificial recurrent neural network (RNN) architecture which can provides an internal memory for the networks. The output of the LSTM goes through 3 fully connect layers with ReLU activation function in hidden layers and Sigmoid in the last layer to get the final result. $\mathbf{x}$ is the load information of a building in the last 12 hours. y indicates the high load status of water heat load in next time hour: $y = 1$ if it is high load, $y = 0$ otherwise.

Table I and Table II list the results where building type and building location are treated as sensitive attributes respectively. Specifically, In Table I, $z = 0$ represents large building; $z = 1$ represents small building. In Table II, z denotes which state the building is in: $z = 1$ indicates the building is in New York, $z = 0$ for buildings in California. All algorithms were run with training and test ratio of $4 : 1$ where a random set of periodic loading sequences is sampled from the original dataset. Experimental results were averaged over 4 random runs.

Table I and Table II list the results evaluated on testing set in terms of accuracy(Acc), Δ_{PE}, Δ_{EO} and Δ_{AP}. In each row, the number in the first column refers to the equation used for training. (1) represents the vanilla logistic classifier. (5) and (8) are trained with the corresponding regularizers. It can be observed from Table I and II that the proposed regularizers improve the fairness without degrading the classification performance.

TABLE I
PERFORMANCE OF HIGH LOAD INDICATION WITH BUILDING TYPE AS THE SENSITIVE ATTRIBUTE.

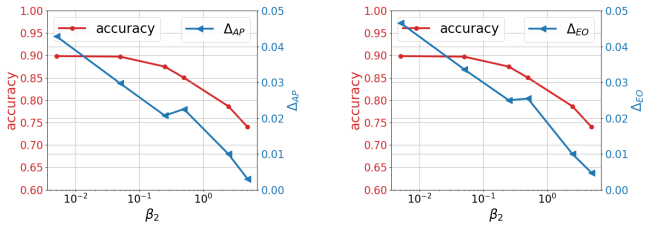
	Acc(%)	$\Delta_{PE}(e^{-2})$	$\Delta_{EO}(e^{-2})$	$\Delta_{AP}(e^{-2})$
(1)	87.04±0.5	5.6±1.2	8.3±3.3	1.25±0.7
(5)	87.6±1.2	4.7±1.5	9.8±5.1	0.45 ±0.4
(8)	87.8±0.5	5.1 ±2.1	7.7 ±4.0	1.3±0.6

TABLE II
PERFORMANCE OF HIGH LOAD INDICATION WITH BUILDING LOCATION AS THE SENSITIVE ATTRIBUTE.

	Acc(%)	$\Delta_{PE}(e^{-2})$	$\Delta_{EO}(e^{-2})$	$\Delta_{AP}(e^{-2})$
(1)	87.04±0.5	6.03±1.2	8.6±3.2	3.5±1.4
(5)	87.12±1.5	5.79±1.9	8.4±2.5	2.2±1.3
(8)	86.76±1.1	5.3±1.5	5.5±3.7	2.3±1.0

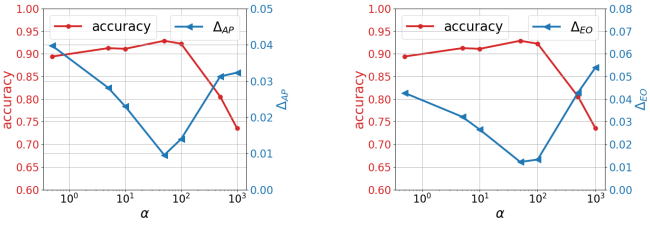
D. User type identification

The user type identification task is using the loading pattern of the buildings' previous T hours to determine the type of the building. Correct user type identification can assist the smart power system for multiple tasks, such as the power management, demand prediction[14]. The neural network contains 4 fully connected layers with ReLU activation function in hidden layers and Softmax in the last output layer.



(a) Accuracy and Δ_{AP} vs β_2 (b) Accuracy and Δ_{EO} vs β_2

Fig. 1. Impact of equal opportunity and predictive equality regularizers. β_2 represents the weight of $\mathcal{R}_{\text{EO}}$, $\beta_1 = 0.8\beta_2$.



(a) Accuracy and Δ_{AP} vs α (b) Accuracy and Δ_{EO} vs α

Fig. 2. Impact of accuracy parity regularizer. α represents the weight of $\mathcal{R}_{\text{AP}}$.

In Table III, z denotes which state the building is in: $z = 1$ means that the building is in New York, $z = 0$ means that the building is in California. In this task, $\mathbf{x}$ is the load information of a building in the last 8 hours and $y \in \{0, \ldots, 6\}$ represents 7 types of buildings.

TABLE III
PERFORMANCE OF BUILDING CLASSIFICATION WITH BUILDING LOCATION AS THE SENSITIVE ATTRIBUTE.

	Acc(%)	$\Delta_{PE}(e^{-3})$	$\Delta_{EO}(e^{-2})$	$\Delta_{AP}(e^{-2})$
(9)	88.97±3.4	9.4 ±3.1	4.55±1.9	4.18±1.3
(5)	92.9±2.4	3.1±0.6	1.22±0.5	0.95±0.51
(8)	89.72±3.4	7.0±3.0	3.36±1.8	2.97±1.3

The vanilla logistic classifier in (9) results more unfairness compared with the regularized model. In Table III, the regularizers ensure the fairness during the training without performance loss. The AP regularizer in (5) even hugely improve the overall performance.

E. Impact of fairness regularizers on the performance

Though fine-tuning the weights of two regularizers could let the model ensure the fairness while having little impacts on final performance, increasing the weights which strengthens the fair constraints will probably hurt the overall performance. In this section, we will tune the weights of regularizers and study the effect of fairness regularizers on the accuracy performance.

The impact of EO regularizer shows as Fig. 1(a) and Fig. 1(b). While the weight of EO regularizer increases, the fairness measurements(Δ_{AP} and Δ_{EO}) decrease and the accuracy also decreases.

The impact of AP regularizer shows as Fig. 2(a) and Fig. 2(b). The fairness measures firstly decreases with increasing weight while the accuracy performance is not hurt at all. However, when the weight is larger than 50, the accuracy decreases and two fairness measurements both increase.

The AP regularizer has less impact than EO regularizer when the accuracy is high due to the following possible reason: the penalty of AP regularizer of a subgroup only considers the cases that are mis-classifed; based on the expression of (7), the EO regularizer takes all cases into consideration no matter the prediction is correct or not. When the model has very high accuracy, the penalty from AP regularizer is small which makes its weight less sensitive than EO regularizer.

V. Conclusion

Wide variety of data arises over smart power grids, which enables the use of machine learning models for user classification tasks such as high-load indication and user type identification. Due to the abundant access to personal data, potential unfairness is a critical concern in such a complex large system, with possible bias inherited in data collection and data processing. In order to achieve more fair results in decision making, management and resource allocation in power systems, existing machine learning schemes need to introduce additional bias-mitigating schemes before they can be readily applied to the grids. To this end, the present paper first examined and showcased the existing bias in directly applying Neural Network models, then two types of regularizers were introduced to promote the fairness in user classification tasks. The proposed regularizers could indeed improve the fairness without significant performance loss. The proposed regularizers could be readily used in other tasks not limited to user classification, e.g.,abnormality detection, theft detection.

References

[1] K. Tornai, A. Olah, and M. Lőrincz, "Forecast based classification for power consumption data," in *International Conference on Intelligent Green Building and Smart Grid*, 2016, pp. 1–5.

[2] F. Rahimi and A. Ipakchi, "Demand response as a market resource under the smart grid paradigm," *IEEE Trans. on Smart Grid*, vol. 1, no. 1, pp. 82–88, 2010.

[3] A. E. Labrador Rivas and T. Abrão, "Faults in smart grid systems: Monitoring, detection and classification," *Electric Power Systems Research*, vol. 189, p. 106602, 2020.

[4] E. L. Lee, J.-K. Lou, W.-M. Chen, Y.-C. Chen, S.-D. Lin, Y.-S. Chiang, and K.-T. Chen, "Fairness-aware loan recommendation for microfinance services," in *Proceedings of the International Conference on Social Computing*, 2014, p. 1–4.

[5] S. Bird, K. Kenthapadi, E. Kiciman, and M. Mitchell, "Fairness-aware machine learning: Practical challenges and lessons learned," in *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2019, p. 834–835.

[6] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness Constraints: Mechanisms for Fair Classification," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, vol. 54. PMLR, 2017, pp. 962–970.

[7] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the International Conference on World Wide Web*, 2017, p. 1171–1180.

[8] Y. Bechavod and K. Ligett, "Penalizing unfairness in binary classification," 2018. [Online]. Available: https://arxiv.org/abs/1707.00044

[9] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proceedings of the International Conference on Machine Learning*, vol. 28, no. 3. PMLR, 2013, pp. 325–333.

[10] M. Macedo, J. Galo, L. de Almeida, and A. de C. Lima, "Demand side management using artificial neural networks in a smart grid environment," *Renewable and Sustainable Energy Reviews*, vol. 41, pp. 128–133, 2015.

[11] L. Zhang, X. Shen, F. Zhang, M. Ren, B. Ge, and B. Li, "Anomaly detection for power grid based on time series model," in *IEEE International Conference on Computational Science and Engineering and IEEE International Conference on Embedded and Ubiquitous Computing*, 2019, pp. 188–192.

[12] T. Vantuch, S. Mišák, and J. Stuchlý, "Power quality prediction designed as binary classification in ac coupling off-grid system," in *IEEE International Conference on Environment and Electrical Engineering*, 2016, pp. 1–6.

[13] A. Reinhardt, P. Baumann, D. Burgstahler, M. Hollick, H. Chonov, M. Werner, and R. Steinmetz, "On the accuracy of appliance identification based on distributed load metering data," in *Sustainable Internet and ICT for Sustainability*, 2012, pp. 1–9.

[14] K. Tornai, A. Oláh, R. Drenyovszki, L. Kovács, I. Pinté, and J. Levendovszky, "Recurrent neural network based user classification for smart grids," in *IEEE Power Energy Society Innovative Smart Grid Technologies Conference*, 2017, pp. 1–5.

[15] (2014) Commercial and residential hourly load profiles for all tmy3 locations in the united states. National Renewable Energy Laboratory. [Online]. Available: https://data.openei.org/submissions/153