Improving Individual Fairness in Fuzzy Classifiers via Evolutionary Multi-Objective Optimization

Takeru Konishi
Graduate School of Informatics
Osaka Metropolitan University
Sakai, Japan
sc23197z@st.omu.ac.jp

Naoki Masuyama
Graduate School of Informatics
Osaka Metropolitan University
Sakai, Japan
masuyama@omu.ac.jp

Yusuke Nojima
Graduate School of Informatics
Osaka Metropolitan University
Sakai, Japan
nojima@omu.ac.jp

***Abstract*—To address the ethical and societal risks of artificial intelligence, responsible artificial intelligence that ensures not only high accuracy but also transparency and fairness is increasingly demanded. Inherently interpretable models are valuable in scenarios where both transparency and fairness are important, as the model's internal mechanisms can be inspected to support bias diagnosis and mitigation. Fuzzy systems are representative inherently interpretable models that enable flexible decisions considering real-world uncertainties. Multi-objective fuzzy genetics-based machine learning generates a diverse set of fuzzy classifiers that consider trade-offs among multiple objectives using an evolutionary multi-objective optimization algorithm. However, previous studies on fairness in fuzzy classifiers have focused exclusively on group fairness and have not considered individual fairness. In this study, we propose a multi-objective fuzzy genetics-based machine learning that simultaneously optimizes accuracy and individual fairness, and investigate how incorporating individual fairness as an objective affects the resulting set of fuzzy classifiers. Experimental results show the effectiveness of this approach and provide empirical insights into the relationships among accuracy, group fairness, and individual fairness, contributing to fair AI design. Source code is available at https://github.com/TakeruKonishi/MoFGBML_Fairness.**

***Index Terms*—Algorithmic Fairness, Individual Fairness, Fuzzy Classifier, Evolutionary Multi-objective Optimization**

I. Introduction

As the social implementation of artificial intelligence (AI) accelerates, it has become a practical concern that AI decisions can entail ethical and social risks in high-risk domains such as employment, lending, and healthcare [1]. Examples include unfair hiring practices, unjust credit decisions, and misdiagnoses. Accordingly, there is a growing demand for responsible AI that ensures not only high accuracy but also transparency and fairness. Transparency enables humans to understand and review decisions, while fairness helps avoid unjust disadvantages to specific individuals or groups.

To ensure transparency, it is important that users can understand both the rationale behind a model's decisions and the internal process by which the model arrives at its decisions [2]. These two perspectives are commonly discussed under the terms explainability and interpretability, respectively. In particular, inherently interpretable models, such as linear models, decision trees, and fuzzy systems, are useful in domains where accountability is strongly required.

In addition to transparency, fairness is essential for the trustworthy use of AI. There are concerns that biases in training data and learning algorithms may lead AI to make decisions that unfairly disadvantage particular individuals or groups. Such unfairness can reinforce existing prejudice and exacerbate social inequalities. Ensuring fairness is therefore critical in high-risk domains. Fairness is often discussed from two perspectives [3]. Group fairness aims to reduce statistical disparities in outcomes or error rates across specific groups. Individual fairness, in contrast, aims to avoid unfair discrimination at the level of individual decisions. Importantly, even when group fairness is satisfied, unfair treatment can occur in individual cases. Therefore, the individual-fairness perspective is also crucial in high-risk domains.

In recent years, considering fairness in highly transparent AI has attracted much attention [4], [5]. When the rationale and internal processes leading to the model's decisions are understandable, model designers can more readily identify where bias may arise and why, and can implement appropriate mitigation strategies. For users, transparency can also strengthen the perception that decisions are fair, which can in turn increase their trust in AI. In particular, considering fairness in inherently interpretable models is valuable in scenarios where both transparency and fairness are important, as the model's internal mechanisms can be inspected to support bias diagnosis and mitigation [6].

Fuzzy classifiers (i.e., fuzzy systems for classification tasks) make decisions based on fuzzy rules whose antecedent parts are fuzzy sets associated with linguistic labels [7]. This property makes fuzzy classifiers inherently interpretable and transparent, and enables them to make flexible decisions considering real-world uncertainties. Evolutionary Fuzzy Systems (EFSs) are frameworks for designing and optimizing fuzzy systems using evolutionary computation and have been actively applied to fuzzy classifier design [8]. Multi-objective Fuzzy Genetics-Based Machine Learning (MoFGBML) [9] is one of the most representative multi-objective EFSs. It directly optimizes a set of fuzzy rules using an Evolutionary

This work was supported by JST BOOST, Japan Grant Number JPMJBS2401. The authors used an AI system (OpenAI ChatGPT, version GPT-5.2) to assist in the refinement of English expression and the clarification of several explanations in this manuscript. All technical content, experimental design, and conclusions were created and verified by the authors.

Multi-objective Optimization Algorithm (EMOA), generating a diverse set of fuzzy classifiers that consider trade-offs among multiple objectives.

To the best of our knowledge, [10] and [11] are the only studies addressing fairness in fuzzy classifiers. [10] proposed MoFGBML that simultaneously optimizes accuracy and fairness, and [11] provided a detailed analysis of the resulting set of fuzzy classifiers. However, these studies focus exclusively on group fairness and do not consider individual fairness. In this study, we propose MoFGBML that simultaneously optimizes accuracy and individual fairness, and investigate how incorporating individual fairness as an objective affects the resulting set of fuzzy classifiers.

The contributions of this paper are summarized as follows:

(I) We present MoFGBML that incorporates individual fairness as an objective, and provides a framework for generating a set of fuzzy classifiers simultaneously optimized for accuracy and individual fairness. To isolate the effect of incorporating individual fairness as an objective, we compare the proposed approach with the fuzzy classifier obtained by accuracy-oriented single-objective optimization, and show, on multiple real-world datasets, that our approach can consistently improve individual fairness while achieving comparable or higher accuracy.

(II) Because fuzzy classifiers are inherently interpretable, their internal mechanisms can be analyzed. We analyze the set of fuzzy classifiers obtained by an individual-fairness-aware MoFGBML in terms of their effects on group fairness and attribute usage rates, and derive insights into the relationships among accuracy, group fairness, and individual fairness. To the best of our knowledge, most studies on fairness have mainly focused on model outputs. In contrast, few studies examine fairness in relation to the model's internal mechanisms. Therefore, our findings provide empirical insights that should offer valuable guidelines for fair AI design.

The rest of this paper is organized as follows. First, Section II gives the background. Section III proposes MoFGBML that simultaneously optimizes accuracy and individual fairness. Section IV discusses the experimental results on four real-world datasets that are well-known in the field of fairness. Finally, Section V concludes this paper.

II. Background

In this section, we explain the theory and methodology behind this study. We first provide an overview of fairness in machine learning, and then introduce fuzzy classifiers and MoFGBML.

A. Fairness in Machine Learning

To address biases inherent in training data and learning algorithms, a variety of fairness definitions have been proposed [3]. In general, fairness can be discussed from two perspectives: group fairness, which aims to reduce statistical disparities in outcomes or error rates across specific groups, and individual fairness, which aims to avoid unfair discrimination at the level of individual decisions [3]. Multiple definitions exist for both group fairness and individual fairness, and satisfying them simultaneously is difficult or impossible except in unrealistic scenarios. Therefore, when evaluating fairness, it is essential to choose a definition that aligns with the target context and application domain.

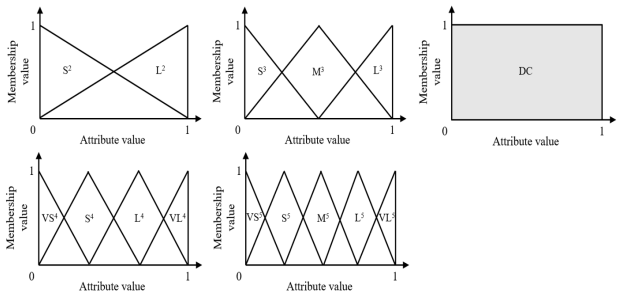


Fig. 1: Triangular membership functions and "*don't care*".

Bias mitigation methods based on fairness definitions are broadly categorized into three types according to where they intervene in the machine learning pipeline: pre-processing, in-processing, and post-processing [3]. Pre-processing approaches modify training data to reduce bias. In-processing approaches modify the learning algorithm to reduce bias. The multi-objective optimization-based method used in this study falls into this category. Post-processing approaches adjust the outputs of trained models to reduce bias.

To the best of our knowledge, fairness in fuzzy classifiers was first addressed in [10] and [11]. However, these studies focus exclusively on group fairness and do not consider individual fairness. Therefore, in this study, we propose a fuzzy classifier design method considering individual fairness and investigate how incorporating individual fairness as an objective affects the resulting set of fuzzy classifiers.

B. Fuzzy Classifier

This subsection provides an overview of the fuzzy classifier.

Consider a pattern classification problem given n-dimensional M-class data with m patterns $\mathbf{x}_p = (x_{p1}, \ldots, x_{pn})$ where x_{pi} denotes the i-th attribute value $(i = 1, 2, \ldots, n)$ in the p-th training pattern $(p = 1, 2, \ldots, m)$. A fuzzy classifier S consists of a set of fuzzy if-then rules of the following form [7].

If x_{p1} is A_{q1} and $\ldots$ and x_{pn} is A_{qn} then Class C_q with CF_q,

where q is the rule index and $\mathbf{A}_q = (A_{q1}, \ldots, A_{qn})$ is the antecedent fuzzy set. In this study, we use 14 triangular-type membership functions from 2 to 5 partitions and a "*don't care*" condition which sets the membership value to 1 regardless of the attribute value shown in Fig. 1. C_q is the consequent class and CF_q is the rule weight. These are uniquely calculated from the antecedent part of the fuzzy rule and the training data [7]. Thus, a single solution (i.e., a fuzzy classifier) is represented by a concatenated set of the antecedent conditions. The coding scheme is the same as [9].

Algorithm 1: NSGA-II-based MoFGBML

Input: Training data: D, Number of generations: G, Population size: N_{pop}.
Output: Fuzzy classifiers: P.

Initialize population P by heuristic rule generation using D.
Evaluate each solution in P using D.
Perform non-dominated sorting on P.
For each rank, calculate crowding distance for each solution.
for $g \leftarrow 1$ **to** G **do**
  $Q \leftarrow \emptyset$.
  for $n \leftarrow 1$ **to** N_{pop} **do**
    if rand$[0, 1) < 0.5$ **then**
      $q \leftarrow$ Offspring generated by Pittsburgh-style genetic operation.
    else
      $q \leftarrow$ Offspring generated by Michigan-style genetic operation.
    Evaluate q using D.
    $Q \leftarrow Q \cup q$.
  $R \leftarrow P \cup Q$.
  Perform non-dominated sorting on R.
  For each rank, calculate crowding distance for each solution.
  $P \leftarrow$ select the best N_{pop} solutions from R based on rank and crowding distance.
return P

In this study, we use a single-winner rule strategy [7] to infer the class label of an input pattern $\mathbf{x}$. In the single-winner rule strategy, for each rule, we calculate the product of the compatibility grade with $\mathbf{x}$ and the rule weight, and output the consequent class of the rule that maximizes this value (i.e., the single-winner rule) as the classification result. Therefore, each classification decision can be explained by the single-winner rule.

C. MoFGBML

MoFGBML [9] is one of the most representative multi-objective EFSs, and a variety of extensions and applications have been reported [10]–[12]. It directly optimizes a set of fuzzy rules using EMOA, generating a diverse set of fuzzy classifiers that consider trade-offs among multiple objectives. In this study, we use the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [13] as the EMOA. NSGA-II approximates the Pareto front by combining non-dominated sorting for solution ranking with the crowding distance for diversity preservation. Thanks to its strong balance between convergence and diversity, NSGA-II is one of the most widely used in multi-objective optimization algorithms.

The procedure for designing fuzzy classifiers using NSGA-II-based MoFGBML is shown in Algorithm 1. In this procedure, each fuzzy classifier corresponds to a candidate solution. In initial population generation, a heuristic rule generation method is used to generate the fuzzy rules that compose a fuzzy classifier. Genetics-based machine learning is roughly divided into two types: the Pittsburgh approach, in which the classifier is represented as a solution, and the Michigan approach, in which the rule is represented as a solution [14]. In this study, we adopt a hybrid approach that uses both the Pittsburgh and Michigan approaches [9]. The Pittsburgh approach is used as the main framework, and the Michigan approach is used for local search.

Note that the MoFGBML used in this study has been slightly modified from the original MoFGBML [9] to improve search performance and simplify the algorithm. The code of the MoFGBML used in this study is available in the authors' GitHub repository[1].

III. Individual-Fairness-aware MoFGBML

In this section, we propose MoFGBML that simultaneously optimizes accuracy and individual fairness.

Many bias mitigation methods in machine learning focus on binary classification tasks. The main reasons are that most fairness-critical applications are binary classification tasks (e.g., hiring vs. not hiring, loan approval vs. not loan approval) and that fairness criteria are often more straightforward to define and evaluate for binary outcomes. Accordingly, we also consider binary classification tasks in this study.

Below, we introduce the accuracy measure and the individual fairness measure for binary classification tasks, and then define the multi-objective optimization problem used in this study based on these measures.

A. Measures

1) Accuracy Measure: In this study, because the datasets considered include imbalanced datasets, we use the geometric mean (G-mean) of the true positive and true negative rates as an accuracy measure for the binary classification task.

G-mean is useful for evaluating accuracy on imbalanced datasets because it accounts for classification performance on both the positive and negative classes. G-mean is defined as follows, where larger values indicate higher accuracy.

$$\text{G-mean} = \sqrt{P[\hat{y} = 1 \mid y = 1] \cdot P[\hat{y} = 0 \mid y = 0]}, \quad (1)$$

where P is the conditional probability, y is the true label, and $\hat{y}$ is the predicted label. Since this study targets a minimization problem, $1 -$ G-mean is used as the objective function.

2) Individual Fairness Measure: Regarding individual fairness, multiple frameworks have been proposed, including definitions based on task-specific similarity [15] and counterfactual fairness based on causal models [16]. In addition, Individual Discrimination (ID) [17] has been proposed as a practical measure of unfairness. ID checks, at the individual level, whether a model's prediction changes for a counterfactual input in which only the sensitive attribute is altered, and

[1] https://github.com/TakeruKonishi/MoFGBML_Fairness

uses the frequency of such changes as the unfairness score. Here, sensitive attributes are attributes that should not lead to unfair treatment, such as gender or race.

In this study, we use ID as the individual-fairness measure. ID directly measures the sensitivity of a model's prediction to changes in the sensitive attribute, and minimizing ID corresponds to suppressing discriminatory treatment at the individual level based on the sensitive attribute. However, ID is defined using counterfactual inputs in which only the sensitive attribute is altered, and it does not guarantee causal counterfactual fairness itself. In this study, we use ID as a practical measure because it can directly measure the sensitivity of a model's prediction to changes in the sensitive attribute without requiring additional assumptions, such as a task-specific similarity measure or a causal model.

Let $\mathcal{D} = \{(\mathbf{x}_i, a_i, y_i)\}_{i=1}^n$ denote a dataset, where $\mathbf{x}_i$ is the non-sensitive attribute vector, a_i is the sensitive attribute, and y_i is the true label. Let $h(\mathbf{x}, a) \in \{0, 1\}$ be a classifier. When the sensitive attribute is binary, $a \in \{0, 1\}$, we construct a counterfactual input $(\mathbf{x}_i, a_i^{\mathrm{cf}})$ by flipping only the sensitive attribute, as defined below.

$$a_i^{\mathrm{cf}} = 1 - a_i. \tag{2}$$

In this case, ID is defined as follows, where smaller values indicate higher fairness.

$$\mathrm{ID} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left[h(\mathbf{x}_i, a_i) \neq h(\mathbf{x}_i, a_i^{\mathrm{cf}})\right]. \tag{3}$$

Here, $\mathbb{I}[\cdot]$ denotes the indicator function, which returns 1 if the condition is true and 0 otherwise.

B. Multi-objective Optimization Problem

A problem that simultaneously optimizes multiple objectives is called a Multi-objective Optimization Problem (MOP). In this study, we define the following MOP for fuzzy classifier design using the accuracy measure described in Section III-A1 and the individual fairness measure described in Section III-A2.

$$\text{MOP: Minimize } f_1 \text{ and minimize } f_2, \tag{4}$$

where f_1 is $1 -$ G-mean and f_2 is ID.

IV. Computational Experiments

In the computational experiments, we applied MoFGBML to the MOP defined in Section III-B and investigated how incorporating individual fairness as an objective affects the resulting set of fuzzy classifiers.

A. Datasets

Table I shows the real-world datasets used in the computational experiments. In this study, we used four real-world datasets that are well-known in the field of fairness. These datasets are freely available from the GitHub repository[2]. We performed multiple runs with different random seeds. In each run, the training and test data were randomly sampled at a split ratio of 9:1.

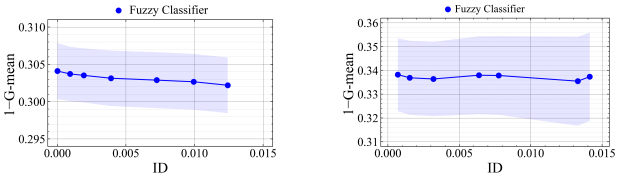
[2]https://github.com/algofairness/fairness-comparison/tree/master/fairness/data

TABLE I: Real-world Datasets

Dataset	Number of Patterns	Number of Attributes	Number of Classes	Sensitive Attribute	Imbalance Rate (+:-)
Adult	32,561	14	2	race	1:3.15
German	1,000	20	2	age	2.33:1
ProPublica -Recidivism	7,214	12	2	race	1:1.22
ProPublica -Violent-Recidivism	4,743	12	2	race	1:5.12

TABLE II: Parameter Setting for MoFGBML

Parameters	Settings
Number of generations	1,000
Population size	50
Number of offspring	50
Number of evaluated solutions	50,000
Selection probability between Pittsburgh and Michigan operations	0.5
Crossover probability in Pittsburgh operation	0.9
Mutation probability in Pittsburgh operation	$1/N(S)$ ($N(S)$: number of rules)
Crossover probability in Michigan operation	0.9
Mutation probability in Michigan operation	$1/n$ (n: number of attributes)
"*don't care*" probability	$(n-3)/n$



(a) Training data (b) Test data

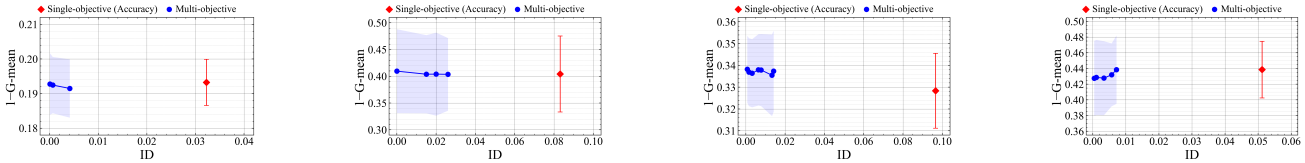
Fig. 2: Relationship between accuracy and ID on the training and test data for the set of fuzzy classifiers obtained by MoFGBML on ProPublica-Recidivism.

B. Experimental Settings

We performed 10 runs with different random seeds. Table II shows the parameters used in the computational experiments. The parameters were set in accordance with [10] and [11]. The termination condition is 1,000 generations.

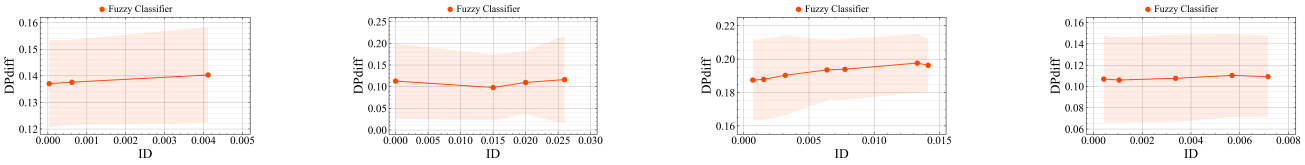
C. Experimental Results

1) Comparison with Accuracy-oriented Single-objective Optimization: Fig. 2 shows the relationship between accuracy and ID on the training and test data for the set of fuzzy classifiers obtained by MoFGBML on ProPublica-Recidivism. For the training-data results shown in Fig. 2a, solid dots connected by solid lines represent the percentile-wise mean curve. This curve is obtained by summarizing the trade-off curve from each run at predefined percentiles and averaging the corresponding values across runs. The construction



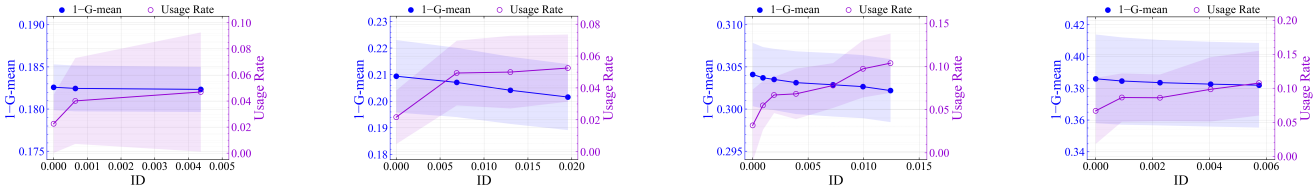
(a) Adult (b) German (c) ProPub (d) ProP. V.

Fig. 3: Relationship between accuracy and ID on the test data for the set of fuzzy classifiers obtained by MoFGBML and for the fuzzy classifier obtained by accuracy-oriented single-objective optimization.



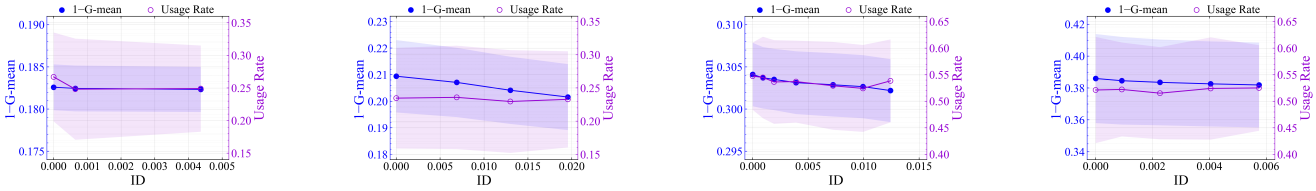
(a) Adult (b) German (c) ProPub (d) ProP. V.

Fig. 4: Relationship between DPdiff and ID on the test data for the set of fuzzy classifiers obtained by MoFGBML.



(a) Adult (b) German (c) ProPub (d) ProP. V.

Fig. 5: Relationship between ID on the training data and the usage rate of the sensitive attribute.



(a) Adult (b) German (c) ProPub (d) ProP. V.

Fig. 6: Relationship between ID on the training data and the usage rate of the most correlated attribute.

procedure follows [11]. Note that the percentile-wise mean curve is a visual summary representing the central tendency of the non-dominated solutions obtained across runs, and it is not the mean of the non-dominated solutions in a strict sense. For the test-data results shown in Fig. 2b, we construct a summary curve by averaging, across runs, the test-data values corresponding to the same percentiles used in Fig. 2a. The shaded area indicates the mean $\pm$ standard deviation of accuracy.

From Fig. 2, no clear trade-off between accuracy and ID is observed for either the training or test data. A similar tendency was observed across all other datasets. These results indicate that solutions exist that reduce ID while maintaining accuracy, suggesting that accuracy improvements may not necessarily be accompanied by greater dependence on the sensitive attribute.

Fig. 3 shows the relationship between accuracy and ID on the test data for the set of fuzzy classifiers obtained by MoFGBML and for the fuzzy classifier obtained by accuracy-oriented single-objective optimization within the same framework (i.e., FGBML that minimizes only $1 - $ G-mean). Solid dots connected by solid lines are obtained by the same procedure as in Fig. 2b.

From Fig. 3, MoFGBML consistently reduces ID while achieving comparable or higher accuracy than accuracy-oriented single-objective optimization. It is also noteworthy that the accuracy slightly improves on Adult and ProPublica-Violent-Recidivism. This suggests that using ID as an objective may act as a form of regularization, avoiding excessive dependence on the sensitive attribute and consequently improving generalization performance.

2) Analysis of the Effects on Group Fairness and Attribute Usage Rates: Fig. 4 shows the relationship between group fairness and individual fairness on the test data for the set of fuzzy classifiers obtained by MoFGBML. In this study, we use the Demographic Parity difference (DPdiff) as the group fairness measure. DPdiff is one of the most representative group fairness measures that quantifies the disparity in positive prediction rates between groups. It is defined as follows, where

P is the conditional probability, $\hat{y}$ is the predicted label, and a is the sensitive attribute. Smaller values indicate higher fairness.

$$\mathrm{DPdiff} = |P[\hat{y} = 1 \mid a = 0] - P[\hat{y} = 1 \mid a = 1]|. \quad (5)$$

Solid dots connected by solid lines are obtained by the same procedure as in Fig. 2b. The shaded area indicates the mean $\pm$ standard deviation of DPdiff.

From Fig. 4, DPdiff does not change substantially even when ID changes. This observation supports the view that individual fairness and group fairness capture different aspects, and that reducing individual-level bias does not necessarily lead to a corresponding reduction in group-level bias, even for fuzzy classifiers.

Next, we analyze attribute usage rates. We define the usage rate of an attribute as the ratio of the number of rules that include a specific attribute in the antecedent part to the total number of rules included in the model. In addition, [11] identified the non-sensitive attribute most strongly correlated with the sensitive attribute for each dataset (Adult: native-country; German: housing; ProPublica-Recidivism: c-charge-desc; ProPublica-Violent-Recidivism: c-charge-desc).

Fig. 5 shows the relationship between ID on the training data and the usage rate of the sensitive attribute. Fig. 6 shows the same relationship for the non-sensitive attribute most strongly correlated with the sensitive attribute. Solid dots connected by solid lines are obtained by the same procedure as in Fig. 2b. For clarity, the percentile-wise mean curve showing the relationship between accuracy and ID on the training data is also shown in the same figure. The shaded area represents the mean $\pm$ standard deviation of the usage rate.

From Fig. 5, the usage rate of the sensitive attribute decreases as ID is reduced. In contrast, Fig. 6 shows that the usage rate of the non-sensitive attribute most strongly correlated with the sensitive attribute does not change substantially even when ID changes. These results suggest that optimizing ID can suppress direct discrimination based on the sensitive attribute, while it may not necessarily mitigate indirect discrimination mediated by strongly correlated proxy attributes. This interpretation is consistent with the observation in Fig. 4 that DPdiff does not improve substantially. Overall, these results provide a structural view, through attribute-usage analysis of fuzzy classifiers, of how ID optimization reduces reliance on the sensitive attribute while leaving proxy-based effects largely unchanged. In practice, it is important to clarify the notion of fairness of interest and to select appropriate measures and intervention methods accordingly.

V. Conclusion

In this paper, we proposed MoFGBML that simultaneously optimizes accuracy and individual fairness, and investigated how incorporating individual fairness as an objective affects the resulting set of fuzzy classifiers. Experiments on multiple real-world datasets showed that MoFGBML consistently improves individual fairness while achieving comparable or higher accuracy than the fuzzy classifier obtained by an accuracy-oriented single-objective optimization. We also analyzed the effects of individual-fairness optimization on group fairness and attribute usage rates, and discussed the relationships among accuracy, group fairness, and individual fairness from the perspective of the model's internal mechanisms.

Future work includes applying other individual-fairness measures and extending MoFGBML to simultaneously optimize accuracy, fairness, and interpretability.

References

[1] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[2] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain, "Interpreting black-box models: A review on explainable artificial intelligence," *Cognitive Computation*, vol. 16, no. 1, pp. 45–74, 2024.

[3] S. Caton and C. Haas, "Fairness in machine learning: A survey," *ACM Computing Surveys*, vol. 56, no. 7, pp. 1–38, 2024.

[4] E. Balkir, S. Kiritchenko, I. Nejadgholi, and K. C. Fraser, "Challenges in applying explainability methods to improve the fairness of NLP models," in *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pp. 80–92, 2022.

[5] P. Haghighat, D. Gándara, L. Kang, and H. Anahideh, "Fair multivariate adaptive regression splines for ensuring equity and transparency," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 22076–22086, 2024.

[6] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[7] H. Ishibuchi, T. Nakashima, and M. Nii, *Classification and Modeling with Linguistic Information Granules: Advanced Approaches to Linguistic Data Mining*. Springer Science & Business Media, 2004.

[8] A. Fernandez, F. Herrera, O. Cordon, M. J. Del Jesus, and F. Marcelloni, "Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to?," *IEEE Computational Intelligence Magazine*, vol. 14, no. 1, pp. 69–81, 2019.

[9] H. Ishibuchi and Y. Nojima, "Analysis of interpretability-accuracy trade-off of fuzzy systems by multiobjective fuzzy genetics-based machine learning," *International Journal of Approximate Reasoning*, vol. 44, no. 1, pp. 4–31, 2007.

[10] T. Konishi, N. Masuyama, J. Casillas, and Y. Nojima, "Fairness-aware classifier design via multi-objective fuzzy genetics-based machine learning," in *2024 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–8, IEEE, 2024.

[11] T. Konishi, N. Masuyama, J. Casillas, and Y. Nojima, "Fairness via fuzzy systems: Analysis of accuracy-fairness trade-off by multi-objective fuzzy genetics-based machine learning," *IEEE Transactions on Fuzzy Systems*, vol. 34, no. 2, pp. 650–664, 2026.

[12] H. Ishibuchi, Y. Nakashima, and Y. Nojima, "Performance evaluation of evolutionary multiobjective optimization algorithms for multiobjective fuzzy genetics-based machine learning," *Soft Computing*, vol. 15, pp. 2415–2434, 2011.

[13] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[14] A. Fernández, S. García, J. Luengo, E. Bernadó-Mansilla, and F. Herrera, "Genetics-based machine learning for rule induction: State of the art, taxonomy, and comparative study," *IEEE Transactions on Evolutionary Computation*, vol. 14, no. 6, pp. 913–941, 2010.

[15] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012.

[16] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[17] M. T. Islam, A. Fariha, A. Meliou, and B. Salimi, "Through the data management lens: Experimental analysis and evaluation of fair classification," in *Proceedings of the 2022 International Conference on Management of Data*, pp. 232–246, 2022.