整数計画法に基づく学習済み決定木の公平性を考慮した編集法

Fairness-aware Edit of a Learned Decision Tree Using Integer Linear Programming

金森憲太朗 * 　　有村博紀
Kentaro Kanamori 　　Hiroki Arimura

北海道大学 大学院情報科学研究科
Graduate School of Inf. Sci. & Tech., Hokkaido University

Abstract: Fairness in machine learning is an emerging topic in recent years. In this paper, we propose a method for editing a given decision tree to be fair according to a specified discrimination criterion by modifying its leaf labels. Our approach can deal with the situation that a sensitive attribute or a discrimination criterion is given after learning a decision tree without learning again. We propose an integer linear programming (ILP) formulation for the problem, which can be solved exactly by any existing ILP solver, while the existing greedy approach can not guarantee the optimality of an obtained solution. By experiments, we confirm the effectiveness of our approach.

1 Introduction

Background and Motivation Recently, while machine learning models assist decision making in the actual world, problems other than their prediction accuracy, such as *interpretability* [6,21] and *fairness* [10,13], attract increasing attention. If their predictions are unexplainable, or discriminative, they are no longer usable in the actual world, even if they achieve high accuracy.

In this paper, we focus on *decision tree models* [8], and study the problems of editing decision trees so as to satisfy fairness constraints. More specifically, we consider the *relabeling problem* [17] that makes a given decision tree to be a fair model by modifying their leaf labels. By extending the framework proposed by Kamiran *et al.* [17], we formulate it as an *integer linear programming (ILP)* problem, which we can obtain an *optimal solution* by using powerful off-the-shelf ILP solvers such as CPLEX [1] and Gurobi [2]. Our approach has the following advantages by comparing with several existing methods for learning fair models [4,16,17,20,22,23]:

- **Interpretability:** Decision tree models are known as one of the interpretable machine learning models since their prediction based on a set of rules that human can understand easily [5,18,21].
- **Adaptivity:** Our approach can deal with the situation that the sensitive attribute corresponding to fairness constraints are given after obtaining a learned model, while existing methods almost need it for learning [4,16,17,20,22,23].
- **Optimality:** Our formulation can obtain a solution with the guarantee of its optimality by using existing efficient ILP solvers, while the existing greedy method can not [17].

Contribution In this paper, we make the following contributions.

1. We extend the framework proposed by Kamiran *et al.* [17] so as to deal with not only *demographic parity* [9] but also *equal opportunity* [15] as the fairness constraint.
2. We formulate the relabeling problem as a 0-1 ILP problem that we can obtain an optimal solution by using ILP solvers, and propose formulations for other constraints such as for the loss of accuracy.
3. By experiments on real datasets, we confirm the effectiveness of our proposed method by comparing the greedy method [17].

Related work *Fairness* in machine learning attracts increasing attention. According to Hajian *et al.* [13], existing methods for achieving fairness can be divided into three parts: the *pre-processing* [16,23], *in-processing* [4,17,20,22], and *post-processing* [15,17] approaches. The pre-processing approaches modify the given dataset to eliminate bias, which may lead learned models to be unfair [16,23]. The in-processing approaches learn models so as to satisfy the fairness constraints based on some criterion, such as demographic parity [9] and equal opportunity [15]. Several methods for learning fair models such as logistic regression [22], SVMs [20], and decision trees [4,17] were proposed. Our method relates to the post-processing approaches, which adjust already learned models to improve their fairness [15,17].

*E-mail: kanamori@ist.hokudai.ac.jp

In the context of interpretability, *non-greedy methods* for interpretable models are also emerging in recent years [6], and several methods for obtaining an optimal solution for the problem related to interpretability were proposed [5, 7, 11, 19]. Some of these methods formulate their problems as *integer programming (IP)* problems, which can obtain an optimal solution by efficient IP solvers. In particular, Bertsimas and Dunn formulated the problem of learning optimal classification trees as a *mixed integer programming (MIP)* problem [7], and very recently, Aghaei *et al.* extended their formulation so as to deal with regression problems and fairness constraints [4]. Note that, however, our problem is not learning fair decision trees but modifying already learned decision trees to be fair.

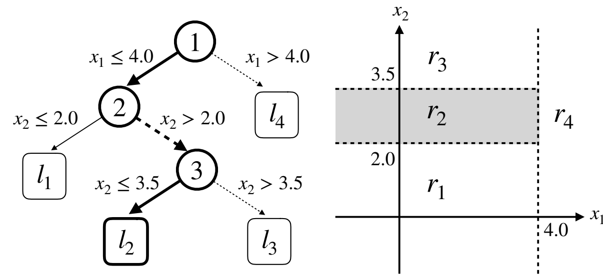

Figure 1: An illustration of a decision tree. The region $r_2 \subseteq \mathbb{R}^2$ corresponding to 2-nd leaf node is expressed by $r_2 = (-\infty, 4.0] \times (2.0, 3.5]$.

2 Preliminary

Notation

For a positive integer $n \in \mathbb{N}$, we denote by $[n] = \{1, \dots, n\}$. For a proposition ψ, $\mathbb{I}[\psi]$ denotes the indicator of ψ, i.e., $\mathbb{I}[\psi] = 1$ if ψ is true, and $\mathbb{I}[\psi] = 0$ if ψ is false.

In this paper, we consider a binary classification problem. Let a pair of an *input* and an *output* $(x, y) \in \mathbb{R}^D \times \{0, 1\}$ be an *example*, and $S = \{(x^{(j)}, y^{(j)})\}_{j=1}^N$ be a *dataset* with N examples. We call a function $h : \mathbb{R}^D \to \{0, 1\}$ a *prediction model*. The *accuracy* of h on S is defined by $acc(h \mid S) := \frac{1}{N} \sum_{j=1}^N \mathbb{I}\left[h(x^{(j)}) = y^{(j)}\right]$.

In addition, we consider a *sensitive attribute* $z \in \{0, 1\}$, such as gender and race. Let $z^{(j)}$ be the sensitive attribute value w.r.t. j-th example $(x^{(j)}, y^{(j)}) \in S$, and $Z = \{z^{(j)}\}_{j=1}^N$ be the set of its values w.r.t. S.

Decision trees

The *decision tree* [8] is a prediction model that consists of a set of prediction rules expressed by the binary tree structure. It makes the prediction according to the label of the leaf node that the input x reaches, and the corresponding leaf node is determined by traversing the tree from the root. Each internal node has a pair of parameters $(d, b) \in [D] \times \mathbb{R}$, where d is a feature index and b is a threshold value, and the input $x = (x_1, \dots, x_D) \in \mathbb{R}^D$ is directed to one of two child nodes depending on whether the statement $x_d \le b$ is true or not.

To formulate our problem, we use the *region-based expression of decision trees*. A decision tree expresses a partition of the input domain $\mathcal{X}$, which means that it partitions the input space into several disjoint regions [14,17,19]. Each leaf node corresponds to a region $r \subseteq \mathcal{X}$ that consists of the statements in internal nodes traversed from the root to the leaf node. Figure 1 illustrates an example of a decision tree and its partition. Then, we define the decision tree as follows.

Definition 1 (Decision trees) *A decision tree h is defined by a triplet $h := (K, L, R)$, where $K \in \mathbb{N}$ is the total number of leaf nodes in h, $L = \{l_k \in \{0, 1\}\}_{k=1}^K$ is a set of leaf (predictive) labels, and $R = \{r_k \subseteq \mathbb{R}^D\}_{k=1}^K$ is a partition of the input domain $\mathbb{R}^D$. The prediction model of the decision tree h is expressed as follows:*

$$h(x) = \sum_{k=1}^{K} l_k \mathbb{I}\left[x \in r_k\right].$$

Note that since R is a partition of $\mathbb{R}^D$, $\bigcup_{k=1}^K r_k = \mathbb{R}^D$ and $\forall k, l \in [K] : k \neq l \Rightarrow r_k \cap r_l = \emptyset$ hold.

Fairness

To evaluate the *fairness* of a model, several definitions of fairness has been proposed [9, 15]. We use two major criteria that measure the discrimination of the model.

One definition is *demographic parity (DP)* [9]. We say that a model h satisfies DP if $P(h(x) = 1 \mid z = 1) = P(h(x) = 1 \mid z = 0)$ where P is a probability on the joint distribution over $(z, h(x))$. Since we can not observe, we define DP score using the empirical probability on the given dataset S and the sensitive attribute Z instead.

Definition 2 (DP score) *DP score of a model h on a dataset S w.r.t. a sensitive attribute Z is defined by*

$$\begin{aligned} DP(h \mid S, Z) := &\ |\{(x, y) \in S_1 \mid h(x) = 1\}|/N_1 \\ &- |\{(x, y) \in S_0 \mid h(x) = 1\}|/N_0 \end{aligned}$$

where $S_z := \{(x^{(j)}, y^{(j)}) \in S \mid z^{(j)} = z\}$ and $N_z := |S_z|$ for $z \in \{0, 1\}$.

Another definition is *equal opportunity (EO)* [15]. We say that a model h satisfies EO if $P(h(x) = 1 \mid y = 1, z = 1) = P(h(x) = 1 \mid y = 1, z = 0)$ where P is a probability on the joint distribution over $(y, z, h(x))$. As with DP, we define EO score as follows.

Definition 3 (EO score) *EO score of a model h on a dataset S w.r.t. a sensitive attribute Z is defined by*

$$\begin{aligned} EO(h \mid S, Z) := &\ |\{(x, y) \in \tilde{S}_1 \mid h(x) = 1\}|/\tilde{N}_1 \\ &- |\{(x, y) \in \tilde{S}_0 \mid h(x) = 1\}|/\tilde{N}_0 \end{aligned}$$

where $\tilde{S}_z := \{(x^{(j)}, y^{(j)}) \in S \mid y^{(j)} = 1 \wedge z^{(j)} = z\}$ *and* $\tilde{N}_z := |\tilde{S}_z|$ *for* $z \in \{0,1\}$.

In this paper, we often denote DP and EO scores by $\delta(h \mid S, Z) \in [-1, 1]$ together, and call it *discrimination score*. Its absolute value approaches 1 as the model h tends to make the predictions unfairly for z, while it approaches 0 if the model makes the predictions fairly.

Problem Formulation

Here, we define our problem called *relabeling problem* [17]. We assume that a decision tree h is already given, and the goal is to reduce the discrimination score of h by changing several leaf labels in h.

Problem 1 (Relabeling problem) *Given a dataset S, a sensitive attribute Z, a decision tree $h = (K, L, R)$, and a threshold $t \in [0,1]$, relabeling problem is defined as follows:*

$$\min_{\hat{L} \in \{0,1\}^K} \Delta(\hat{h} \mid h, S) \quad \text{subject to} \quad \left|\delta(\hat{h} \mid S, Z)\right| \leq t,$$

where $\hat{h} = (K, \hat{L}, R)$ is a modified decision tree and $\Delta(\hat{h} \mid h, S) := acc(h \mid S) - acc(\hat{h} \mid S)$ is the loss of accuracy.

Kamiran *et al.* [17] proved that Problem 1 is polynomially equivalent to the knapsack preblem, and the NP-completeness of the problem. They also proposed a greedy approximation algorithm for Problem 1. In the next section, we formulate Problem 1 as an ILP problem, which can be solved exactly by ILP solvers.

3 Proposed Method

In this section, we propose an ILP formulation of Problem 1. In the following discussion, we assume that a dataset S, a sensitive attribute Z, a decision tree $h = (K, L, R)$, and a threshold $t \in [0,1]$ are given.

As with the framework proposed by [17], our method uses the following well-known facts for decision trees:

- $\bigcup_{k=1}^{K} S^{(k)} = S$,
- $\forall k, l \in [K] : k \neq l \Rightarrow S^{(k)} \cap S^{(l)} = \emptyset$,

where $S^{(k)} := \{(x, y) \in S \mid x \in r_k\}$. This implies that any $(x, y) \in S$ arrives at a unique leaf node in the decision tree. Hence, we can evaluate the accuracy and discrimination score of each leaf node independently, and the total values are expressed as the sum of these values w.r.t. each leaf node.

Let $\hat{h} = (K, \hat{L}, R)$ be a modified decision tree, where $\hat{L} = \{\hat{l}_k\}_{k=1}^{K}$. To formulate Problem 1 as an ILP problem, we introduce a K-dimensional binary vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K) \in \{0,1\}^K$, where $\eta_k = 1$ denotes k-th leaf label is changed, i.e., $\eta_k = 1 \iff l_k \neq \hat{l}_k$. Then, we formulate the loss of accuracy $\Delta(\hat{h} \mid h, S)$ and discrimination score $\delta(\hat{h} \mid S, Z)$ as linear functions over $\boldsymbol{\eta}$.

Objective function

We define an objective function as the loss of accuracy $\Delta(\hat{h} \mid h, S) = acc(h \mid S) - acc(\hat{h} \mid S)$. First, we show that the accuracy is expressed as the sum of the total number of examples whose labels are same with the predictive label in each leaf node.

Lemma 1 *For any decision tree h and dataset S,*

$$acc(h \mid S) = \frac{1}{N} \sum_{k=1}^{K} \left(l_k p_k + (1 - l_k) n_k\right),$$

where $p_k := |\{(x, y) \in S^{(k)} \mid y = 1\}|$ *and* $n_k := |\{(x, y) \in S^{(k)} \mid y = 0\}|$.

Note that $S^{(k)}$ is determined when h and S are given.

Secondly, we evaluate the loss of accuracy by changing leaf labels in the following lemma from Lemma 1.

Lemma 2 *For any h, S, and modified decision tree $\hat{h}$,*

$$\Delta(\hat{h} \mid h, S) = \frac{1}{N} \sum_{k=1}^{K} \left((l_k - \hat{l}_k) p_k + (\hat{l}_k - l_k) n_k \right).$$

Here, we denote by $\tilde{c}_k := (l_k - \hat{l}_k) p_k + (\hat{l}_k - l_k) n_k$, which indicates the difference of accuracy in k-th leaf node. Then, by considering whether the leaf label is changed or not, we can express $\tilde{c}_k$ as follows:

$$\tilde{c}_k = \begin{cases} (2l_k - 1)(p_k - n_k), & \text{if } l_k \neq \hat{l}_k, \\ 0, & \text{otherwise.} \end{cases}$$

Now, we formulate our objective function by using the K-dimensional vector $\boldsymbol{\eta} \in \{0,1\}^K$. Then, our objective function $f : \{0,1\}^K \to [-1, 1]$ is defined by

$$f(\boldsymbol{\eta}) := \frac{1}{N} \sum_{k=1}^{K} c_k \eta_k$$

where $c_k := (2l_k - 1)(p_k - n_k)$. We show that our objective $f(\boldsymbol{\eta})$ is equivalent to the loss of accuracy $\Delta(\hat{h} \mid h, S)$ in the following theorem.

Theorem 1 *For any modified decision tree $\hat{h}$ and the indicator vector $\boldsymbol{\eta} \in \{0,1\}^K$, $f(\boldsymbol{\eta}) = \Delta(\hat{h} \mid h, S)$ holds.*

Proof. We show that $\tilde{c}_k = c_k \eta_k$ for any $k \in [K]$. If $l_k = \hat{l}_k$, then $\tilde{c}_k = 0$. Since $\eta_k = 0$, $c_k \eta_k = \tilde{c}_k$ holds. Otherwise, if $l_k \neq \hat{l}_k$, then $\tilde{c}_k = (2l_k - 1)(p_k - n_k)$. Since $\eta_k = 1$ and $c_k = (2l_k - 1)(p_k - n_k)$, $c_k \eta_k = \tilde{c}_k$ holds. Therefore, $\tilde{c}_k = c_k \eta_k$ for any $k \in [K]$, which implies $f(\boldsymbol{\eta})$ is equivalent to $\Delta(\hat{h} \mid h, S)$. $\square$

Fairness constraint

Next, we formulate the fairness constraint $\left|\delta(\hat{h} \mid S, Z)\right| \leq t$. First, we formulate the discrimination score as the sum of values w.r.t. each leaf node in h. Note that DP and EO scores of h depend only on the examples $(x, y) \in S$ such that $h(x) = 1$. Hence, the discrimination score of a decision tree depends only on the leaf nodes whose predictive labels are 1. For convenience, we denote by $s_k := |\{(x^{(j)}, y^{(j)}) \in S^{(k)} \mid z^{(j)} = 1\}|$ and $\tilde{s}_k := |\{(x^{(j)}, y^{(j)}) \in S^{(k)} \mid y^{(j)} = 1 \wedge z^{(j)} = 1\}|$, respectively. Then, DP score of h is expressed as follows.

Lemma 3 *For any h, S and Z,*

$$DP(h \mid S, Z) = \sum_{k=1}^{K} \left(\frac{s_k}{N_1} - \frac{p_k + n_k - s_k}{N_0} \right) l_k.$$

Similar to this, EO score of h is expressed as follows:

Lemma 4 *For any h, S and Z,*

$$EO(h \mid S, Z) = \sum_{k=1}^{K} \left(\frac{\tilde{s}_k}{\tilde{N}_1} - \frac{p_k - \tilde{s}_k}{\tilde{N}_0} \right) l_k.$$

Secondly, we formulate the discrimination score of $\hat{h}$. The statement $\hat{l}_k = 1$ is equivalent to the following two conditions for the original label l_k and η_k:

1. l_k is 0 and changed to 1, i.e., $\eta_k = 1$, or
2. l_k is 1 and not changed, i.e., $\eta_k = 0$.

Hence, we can formulate these relationships as follows:

$$\begin{aligned} \hat{l}_k &= (1 - l_k)\eta_k + l_k(1 - \eta_k) \\ &= (1 - 2l_k)\eta_k + l_k. \end{aligned}$$

By using this, we show that the discrimination score of $\hat{h}$ can be expressed using $\boldsymbol{\eta}$ in the following theorems.

Theorem 2 *For a modified decision tree $\hat{h}$ and $\boldsymbol{\eta} \in \{0,1\}^K$, $DP(\hat{h} \mid S, Z) = \sum_{k=1}^{K} d_k \eta_k + DP(h \mid S, Z)$ holds, where $d_k := (1 - 2l_k) \left(\frac{s_k}{N_1} - \frac{p_k + n_k - s_k}{N_0} \right)$.*

Proof. We denote by $\tilde{d}_k := \left(\frac{s_k}{N_1} - \frac{p_k + n_k - s_k}{N_0} \right)$. Then, from Lemma 3 and $\hat{l}_k = (1 - 2l_k)\eta_k + l_k$, we have

$$\begin{aligned} DP(\hat{h} \mid S, Z) &= \sum_{k=1}^{K} \tilde{d}_k \hat{l}_k \\ &= \sum_{k=1}^{K} \tilde{d}_k (1 - 2l_k)\eta_k + \sum_{k=1}^{K} \tilde{d}_k l_k \\ &= \sum_{k=1}^{K} d_k \eta_k + DP(h \mid S, Z). \end{aligned}$$

□

EO score also can be expressed using $\boldsymbol{\eta}$.

Theorem 3 *For a modified decision tree $\hat{h}$ and $\boldsymbol{\eta} \in \{0,1\}^K$, $EO(\hat{h} \mid S, Z) = \sum_{k=1}^{K} d_k \eta_k + EO(h \mid S, Z)$ holds, where $d_k := (1 - 2l_k) \left(\frac{\tilde{s}_k}{\tilde{N}_1} - \frac{p_k - \tilde{s}_k}{\tilde{N}_0} \right)$.*

From Theorem 2 and Theorem 3, the fairness constraint $\left|\delta(\hat{h} \mid S, Z)\right| \leq t$ is equivalent to

$$-t - \delta(h \mid S, Z) \leq \sum_{k=1}^{K} d_k \eta_k \leq t - \delta(h \mid S, Z).$$

Note that the constant value d_k is determined depending on using either DP or EO score.

Overall formulation

Combining the above discussions, now we can express Problem 1 as follows.

$$\begin{aligned} \min_{\boldsymbol{\eta} \in \{0,1\}^K} \quad & f(\boldsymbol{\eta}) = \frac{1}{N} \sum_{k=1}^{K} c_k \eta_k \\ \text{subject to} \quad & \sum_{k=1}^{K} d_k \eta_k \leq t - \delta(h \mid S, Z), \\ & -\sum_{k=1}^{K} d_k \eta_k \leq t + \delta(h \mid S, Z), \end{aligned} \tag{1}$$

where $c_k = (2l_k - 1)(p_k - n_k)$, $d_k = (1 - 2l_k) \left(\frac{s_k}{N_1} - \frac{p_k + n_k - s_k}{N_0} \right)$ if using DP score, and $d_k = (1 - 2l_k) \left(\frac{\tilde{s}_k}{\tilde{N}_1} - \frac{p_k - \tilde{s}_k}{\tilde{N}_0} \right)$ if using EO score. Note that c_k, d_k, and $\delta(h \mid S, Z)$ are constant values determined automatically when S, Z, and h are given.

Since all the objective and constraints are liner functions over $\boldsymbol{\eta}$, (1) is a *0-1 ILP problem*. ILPs have been extensively studied and we can obtain an *optimal solution* by efficient ILP solvers such as CPLEX [1] and Gurobi [2], while the greedy method [17] can not guarantee its optimality. However, from the NP-completeness of Problem 1, it may computationally expensive than the greedy method.

Formulation of other constraints We show that our framework can deal with some variants of Problem 1 by modifying (1) while keeping its linearity.

First, we consider the variation of Problem 1 that added the constraint for the total number of changing leaf labels, which is defined as follows for given two thresholds $t_{disc} \in [0, 1]$ and $t_{edit} \in [K]$:

$$\begin{aligned} \min_{\hat{L} \in \{0,1\}^K} \quad & \Delta(\hat{h} \mid h, S) \\ \text{subject to} \quad & \left|\delta(\hat{h} \mid S, Z)\right| \leq t_{disc}, \\ & \sum_{k=1}^{K} \mathbb{I}\left[l_k \neq \hat{l}_k\right] \leq t_{edit}. \end{aligned}$$

Recall that $\eta_k = 1$ denotes k-th leaf label is changed, i.e., $\eta_k = 1 \iff l_k \neq \hat{l}_k$. Then, this problem can be expressed by adding the following constraint into (1):

$$\sum_{k=1}^{K} \eta_k \leq t_{edit}.$$

Secondly, we consider the minimization problem of the discrimination score with a constraint for the loss of accuracy w.r.t. a given threshold $t_{acc} \in [0, 1]$, which is defined as follows:

$$\min_{\hat{L} \in \{0,1\}^K} \left| \delta(\hat{h} \mid S, Z) \right| \quad \text{subject to} \quad \Delta(\hat{h} \mid h, S) \leq t_{acc}.$$

Then, we introduce a variable $\epsilon \geq 0$ for expressing $|\delta(\hat{h} \mid S, Z)|$, and this problem can be expressed as follows:

$$\begin{aligned}
\min_{\boldsymbol{\eta} \in \{0,1\}^K, \epsilon \geq 0} \quad & \epsilon \\
\text{subject to} \quad & \frac{1}{N} \sum_{k=1}^{K} c_k \eta_k \leq t_{acc}, \\
& -\epsilon - \sum_{k=1}^{K} d_k \eta_k \leq \delta(h \mid S, Z), \\
& -\epsilon + \sum_{k=1}^{K} d_k \eta_k \leq -\delta(h \mid S, Z),
\end{aligned}$$

where the last two constrains are essential for ϵ to express the absolute value of $\delta(\hat{h} \mid S, Z)$.

4 Experiments

In this section, we evaluate the proposed method by experiments on real datasets, and compare it with the greedy method proposed by Kamiran *et al.* [17].

Experimental setup We used three real datasets. These details are summarized as follows:

- **COMPAS [3]** ($N = 6172, D = 9$): $y^{(j)}$ indicates whether j-th person recidivates within two years. We use the attribute "African_American" as $z^{(j)}$.
- **Adult [12, 22]** ($N = 32561, D = 58$): $y^{(j)}$ indicates whether j-th person's income exceeds 50K USD. We use the attribute "sex" as $z^{(j)}$.
- **Wine [12,20]** ($N = 6497, D = 11$): $y^{(j)}$ indicates whether j-th wine is rated as a 6 or above out of 10 ranks. $z^{(j)}$ indicates whether it is a white wine or not (i.e., a red wine).

We got decision trees by using CART algorithm [8] for each dataset, and solved Problem 1 by our proposed method (**ILP**) and the **greedy** method proposed by [17] for each decision tree. In our experiments, we sampled 70% examples from each dataset randomly, and report the average statistics over 5 samples. The details learned decision trees are summarized in Table 1.

Table 1: Details of learned decision trees. "Acc.", "DP", "EO", and K_{ave} denote the average accuracy, DP score, EO score, and total number of leaf nodes.

name	K_{ave}	**Acc.**	**DP**	**EO**
COMPAS	29.0	0.679 ± 0.003	0.232	0.235
Adult	164.4	0.859 ± 0.001	0.182	0.068
Wine	39.4	0.762 ± 0.003	0.219	0.138

We used the threshold value $t = 0.01$ for all datasets. All codes were implemented in Python 3.6 with scikit-learn and CPLEX Python API [1]. All experiments were conducted on 64-bit macOS Sierra 10.12.6 with Intel Core i5 2.90GHz CPU and 8GB Memory.

Comparison results Table 2 shows the experimental results for DP and EO scores. For all datasets, our method maintained slightly higher accuracy than the greedy method. This implies that the greedy method sometimes failed to obtain an optimal solution. In addition, discrimination scores that our method attained were close to the given threshold value $t = 0.01$. On the other hands, these of the greedy method decreased overly, which may cause the loss of accuracy. Note that the running times of our ILP method were longer than these of the greedy method as expected in section 3.

It is noteworthy that the total numbers of leaf nodes changed by our method were smaller than these by the greedy method. This implies that our method satisfied the given fairness constraint with less loss of accuracy and fewer edit operations for the model. From a model reliability perspective, it is desirable for users that only a small part of the given learned model is changed to satisfy the fairness constraint [21].

5 Conclusion and Discussion

We proposed an ILP formulation of the relabeling problem, which makes a given decision tree fair by modifying their leaf predictive labels. Our approach can handle both demographic parity and equal opportunity as a fairness constraint, and we can obtain an optimal solution by using any existing ILP solver. By experiments on real datasets, we confirmed the effectiveness of our methods by comparing with the existing greedy method.

In this paper, we focused only on an edit operation of leaf labels. As feature work, we will try to develop more flexible edit operations for satisfying fairness constraints, such as modifying the prediction rules in each internal node or the structure of the tree itself.

Acknowledgements This work was partially supported by JSPS KAKENHI(S) 15H05711.

Table 2: Experimental results averaged over 5 trials. "Acc.", "DP", "EO", and "edit" denote the average accuracy with its standard deviation, DP score, EO score, and total number of changed leaf nodes of the modified decision tree on each dataset, respectively.

		DP score				EO score			
name	**method**	**Acc.**	**DP**	**edit**	**time[ms]**	**Acc.**	**EO**	**edit**	**time[ms]**
COMPAS	greedy	0.569 ± 0.009	0.006	10.0	0.260	0.596 ± 0.018	0.001	9.8	0.305
	ILP	**0.581** ± 0.005	0.009	9.2	6.143	**0.605** ± 0.007	0.009	8.8	8.354
Adult	greedy	0.812 ± 0.003	0.005	27.2	2.074	0.847 ± 0.003	0.005	12.2	1.250
	ILP	**0.814** ± 0.002	0.010	26.6	20.267	**0.849** ± 0.002	0.010	11.6	22.137
Wine	greedy	0.745 ± 0.003	0.005	4.0	0.164	0.721 ± 0.012	0.001	7.4	0.451
	ILP	**0.746** ± 0.003	0.004	3.6	12.790	**0.752** ± 0.003	0.007	3.0	6.437

References

[1] CPLEX Optimizer — IBM. `https://www.ibm.com/analytics/cplex-optimizer`.

[2] Gurobi Optimization - The State-of-the-Art Mathematical Programming Solver. `http://www.gurobi.com/`.

[3] J. Adebayo. FairML: Auditing black-box predictive models. `https://github.com/adebayoj/fairml`, 2018.

[4] S. Aghaei, M. J. Azizi, and P. Vayanos. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proc. AAAI 2019, Hawaii*, 2019 (to appear).

[5] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin. Learning certifiably optimal rule lists. In *Proc. ACM KDD 2017, Halifax*, pages 35–44, 2017.

[6] D. Bertsimas. Interpretable AI. AAAI 2019, invited talk, Hawaii, 2019 (to appear).

[7] D. Bertsimas and J. Dunn. Optimal classification trees. *Mach. Learn.*, pages 1039–1082, 2017.

[8] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees.* CRC press, 1984.

[9] T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.*, pages 277–292, 2010.

[10] K. Crawford. The trouble with bias. NIPS 2017, invited talk, Long Beach, 2017.

[11] Z. Cui, W. Chen, Y. He, and Y. Chen. Optimal action extraction for random forests and boosted trees. In *Proc. ACM KDD 2015, Sydney*, pages 179–188, 2015.

[12] D. Dheeru and E. Karra Taniskidou. UCI machine learning repository, 2017.

[13] S. Hajian, F. Bonchi, and C. Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proc. ACM KDD 2016, San Francisco*, pages 2125–2126, 2016.

[14] S. Hara and K. Hayashi. Making tree ensembles interpretable: A bayesian model selection approach. In *Proc. AISTATS 2018, Canary Islands*, pages 77–85, 2018.

[15] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *NIPS 2016, Barcelona*, pages 3315–3323.

[16] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.*, 33(1):1–33, Oct 2012.

[17] F. Kamiran, T. Calders, and M. Pechenizkiy. Discrimination aware decision tree learning. In *IEEE ICDM 2010*, pages 869–874, 2010.

[18] H. Lakkaraju, S. H. Bach, and J. Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proc. ACM KDD 2016, San Francisco*, pages 1675–1684, 2016.

[19] N. Meinshausen. Node harvest. *Ann. Appl. Stat.*, 4:2049–2072, 2010.

[20] M. Olfat and A. Aswani. Spectral algorithms for computing fair support vector machines. In *Proc. AISTATS 2018, Canary Islands*, pages 1933–1942, 2018.

[21] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proc. ACM KDD 2016, San Francisco*, pages 1135–1144, 2016.

[22] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In *Proc. AISTATS 2017, Fort Lauderdale*, pages 962–970, 2017.

[23] I. liobaite, F. Kamiran, and T. Calders. Handling conditional discrimination. In *IEEE ICDM 2011*, pages 992–1001, Dec 2011.