The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

公平ロジスティック回帰での確定的決定則の影響

Influence of Deterministic Decision Rules in Fair Logistic Regression

神嶌 敏弘 *1 Toshihiro Kamishima　赤穂 昭太郎 *1 Shotaro Akaho　麻生 英樹 *1 Hideki Asoh　佐久間 淳 *2 Jun Sakuma

*1産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

*2筑波大学／理化学研究所 革新知能統合研究センター
University of Tsukuba; and RIKEN Center for Advanced Intelligence Project

The goal of fairness-aware classification is to categorize data while taking into account potential issues of fairness. For example, when applying data mining technologies to university admissions, admission criteria must be fair with regard to sensitive features, such as gender or race. We developed logistic regression satisfying such a fairness constraint. In this paper, we show the trade-off between prediction accuracy and fairness can be drastically improved by explicitly considering the influence of a deterministic decision rule.

1. はじめに

公平性配慮型分類とは，採用の可否などの判定の過程から，性別などの公平性の観点から影響してはならない情報を除外するような制約下で行うクラス分類問題である．我々は，今までにロジスティック回帰について正則化項を加える方法を提案していた [Kamishima 12]．ここでは，確定的な決定則の影響を明示的に考慮する実独立性と呼ぶ公平性の規準を満たすことで予測精度と公平性のより良いトレードオフを実現できることを示す．さらに，この実独立性を満たすロジスティック回帰法の高速化についても予備実験を行った．

2. 章では公平配慮型分類問題の形式的定義と実独立性の概念を示したのち，この問題に対処できるように修正したロジスティック回帰について述べる．3. 章では実独理性を達成することでより小さな予測精度の低下でより公平な分類が可能であることを実験的に検証する．4. 章は高速化についての予備的検討で，5. 章はまとめである．

2. 公平ロジスティック回帰

公平配慮型分類問題を定義したのち，モデルベース独立性と実独立性の概念を示す．最後に，これらの独立性を達成する公平ロジスティック回帰モデルについて述べる．

2.1 表記と問題設定

潜在的な公平性の問題に配慮しつつデータを分類するのが公平配慮型分類の目的である．確率変数 S と $\mathbf{X}$ は，それぞれセンシティブ特徴と非センシティブ特徴を表す．公平性を保証すべき情報はこのセンシティブ特徴で表し，それ以外が非センシティブ特徴である．例えば，採用の可否を決めるときには，法的に配慮すべき求職者の性別や人種の情報をセンシティブ特徴とする．ここではセンシティブ特徴は二値のスカラー変数で，非センシティブ特徴は m 次元の実数値ベクトルとする．クラス変数 Y は，採用の可否といった分類クラスを表し，ここでは二値分類を扱う．さらに，真の分布でのクラスの事後分布を近似したものを予測モデルとする．この予測モデルから確率的に生成されたクラスラベルを $\hat{Y}$ で表し，真の分布から生成されたラベル Y とは区別しておく．ここで，実際の予測ラベルは，確率的には生成されず，誤分類リスクを最小化するように次の決定則によって確定的に生成される．

$$\tilde{y} = \arg\max_{\hat{y}} \Pr[\hat{Y}{=}\hat{y}|\mathbf{X}{=}x, S{=}s] \tag{1}$$

この実際の予測ラベルを変数 $\tilde{Y}$ で表す．

次に，分類での 3 種類の公平性を紹介する．一つ目は，$\hat{Y} \perp\!\!\!\perp S \mid \mathbf{X}$ の条件付き独立性で，センシティブ特徴を単純に予測モデルから削除した場合に相当する．このようにセンシティブ特徴を予測モデルから削除しても，センシティブ特徴と相関のある他の変数からの間接的な影響のために不公平な決定がなされる場合がある．これを red-lining 効果という [Calders 10]．二つ目の条件である，予測クラスとセンシティブ特徴の条件なし独立性 $\hat{Y} \perp\!\!\!\perp S$ は，この red-lining 効果をを回避するのに有効である．この公平性では，訓練データ中のラベル情報は潜在的に不公平に判断に基いていると仮定している．このデータ中のラベル情報は公平であると仮定しているのが三つ目の公平性である [Hardt 16, Zafar 17]．この公平性は予測誤差がセンシティブ特徴にはよらないというもので，観測されたラベルが与えられたときの予測クラスとセンシティブ特徴の独立性 $\hat{Y} \perp\!\!\!\perp S \mid Y$ として形式的には定義される．これらの規準のうち，本稿では $\hat{Y} \perp\!\!\!\perp S$ を扱う．

公平性に配慮した分類問題の前に，標準的な分類問題について述べる．真の分布から得られた実現値の対 $(\mathbf{x}, s)$ で各対象は表される．この対象のクラスの実現値 y は真の分布 $\Pr[Y|\mathbf{X}{=}\mathbf{x}, S{=}s]$ から生成されるものとする．なお，この真の分布は，センシティブ特徴に依存した潜在的に不公平なラベルを生成することがあることに注意されたい．この真の分布自体を知ることはできないが，この真の分布から得られたデータは観測できる．これらのデータを集めたものが（訓練）データ集合 $\mathcal{D} = \{(y_i, \mathbf{x}_i, s_i)\}, i = 1, \ldots, n$ である．さらに，センシティブ特徴の値が s であるデータを集めた $\mathcal{D}$ の部分集合を $\mathcal{D}_s$ と記す．モデル分布の族 $\Pr[\hat{Y}, \mathbf{X}, S]$ も与えられたとき，この中から真の分布を最もよく近似するものを特定することが，標準的な分類タスクの目標となる．

では，公平性配慮型分類問題に移る．本論文では，予測クラスとセンシティブ特徴の条件なし独立 $\hat{Y} \perp\!\!\!\perp S$ が公平性の規準

連絡先: ホームページ http://www.kamishima.net

である場合を扱う．この場合では，訓練データ中のラベルは潜在的に不公平で，公平性に配慮した真のラベルの分布は観測できないだけでなく，そこからデータをサンプリングすることですらできない．それゆえ，公平なラベルは公平性規準を満たしているとの仮定を導入する．潜在的に不公平な訓練データ集合，モデル分布の族，および公平性規準が与えられたとき，モデル分布の族中で公平性規準を満たす分布の中から，真の分布を最も良く近似する公平モデル分布を見つけることが公平性配慮型分類問題の目的である．公平性制約の影響で予測に利用可能な情報は一般的に減少するため，予測精度と公平性はトレードオフ関係にある．

2.2 モデルベース独立性と実独立性

ここでは，モデルベース独立性と実独立性の概念を導入する [Kamishima 18]．モデルベース独立性では，モデル分布族の中の分布から直接的にクラスラベルは生成される．一方で，実独立性では，モデルバイアスと決定則を考慮した分布からクラスラベルは生成される．モデルベース独立性ではなく実独立性を満たすことで，予測精度と公平性のよりよいトレードオフを実現できることを 3. 章の実験で示す．

識別モデル [Bishop 08, 1.5.4 節] であるロジスティック回帰をここでは対象としているので，識別モデルの場合での 2 種類の独立性を紹介する．まず，予測モデルからクラスラベルが確率的に生成される場合であるモデルベース独立性から始める．形式的には，この独立性を次式で定義する．

$$\hat{Y} \perp\!\!\!\perp S, \text{ where } (\hat{Y}, S) \sim \Pr[\hat{Y}, S] \tag{2}$$

条件付き分布 $\Pr[\hat{Y}|\mathbf{X}, S]$ を直接的にモデル化するのが識別モデルである．この識別モデルに対しては，$\mathbf{X}$ 上の期待値を標本平均によって近似することで，分布 $\Pr[\hat{Y}, S]$ を得る．

$$\Pr[\hat{y}, s] \approx \frac{|\mathcal{D}_s|}{n} \frac{1}{|\mathcal{D}_s|} \sum_{\mathbf{x} \in \mathcal{D}_s} \Pr[\hat{y}|\mathbf{x}, s] = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{D}_s} \Pr[\hat{y}|\mathbf{x}, s] \tag{3}$$

なお，ここで標本平均を用いて真の分布を近似しているので，モデルバイアスは除去されており，決定則の影響のみが残っている．

もう一つの実独立性は，予測クラスとセンシティブ特徴の間の独立性である点についてはモデルベースの独立性と同じである．しかし，実独立性では，クラスラベルはモデル分布から生成されるのではなく，決定則の影響をも考慮した分布から生成される．実独立性の形式的定義は次式である．

$$\tilde{Y} \perp\!\!\!\perp S, \text{ where } (\tilde{Y}, S) \sim \Pr[\tilde{Y}, S] \tag{4}$$

予測クラス $\tilde{Y}$ は，確率的に生成されるモデルベース独立性の場合とは異なり，式 (1) の決定則で確定的に生成される．モデルベースの場合の式 (3) と同様に，$\Pr[\tilde{Y}|\mathbf{X}, S]$ の標本平均をとることで分布 $\Pr[\tilde{Y}, S]$ を得る．

$$\Pr[\tilde{y}, s] = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{D}_s} \Pr[\tilde{y}|\mathbf{x}, s] \tag{5}$$

確定的にラベルを生成する分布 $\Pr[\tilde{Y}|\mathbf{X}, S]$ では，各実現値を生成する確率質量が 0 または 1 のいずれかの値になる．

$$\begin{cases} \Pr[\tilde{Y}{=}1|\mathbf{x}, s] = \begin{cases} 1, & \text{if } \Pr[\hat{Y}{=}1|\mathbf{x}, s] \geq \Pr[\hat{Y}{=}0|\mathbf{x}, s] \\ 0, & \text{otherwise} \end{cases} \\ \Pr[\tilde{Y}{=}0|\mathbf{x}, s] = 1 - \Pr[\tilde{Y}{=}1|\mathbf{x}, s] \end{cases} \tag{6}$$

ただし，$\Pr[\hat{Y}|\mathbf{X}, S]$ は元の識別モデルである．なお，この式の $\Pr[\tilde{Y}|\mathbf{x}, s]$ は $\Pr[\hat{Y}{=}1|\mathbf{x}, s] - \Pr[\hat{Y}{=}0|\mathbf{x}, s]$ にステップ関数を適用したものに相当する．

以上のように，モデルベース独立性と実独立性の二つの公平性制約は，クラスラベルを生成する分布でモデルバイアスや決定則の影響を考慮しているかどうかが異なる．

2.3 実独立な公平ロジスティック回帰

本論文では，偏見除去正則化項 (prejudice remover regularizer) 付きロジスティック回帰 [Kamishima 12]（PR モデルと略す）と呼ぶ公平分類モデルについて扱う．モデルベース独立性と実独立性それぞれの制約を満たす二つの PR モデルを述べる．このモデルの目的関数は，ロジスティック回帰の目的関数に公平性を強化するための制約項を加えたものである．通常のロジスティック回帰の予測モデルは次式である．

$$\Pr[\hat{y}|\mathbf{x}; \mathbf{w}] = y \ \mathrm{sig}(\mathbf{x}^\top \mathbf{w}) + (1 - y)(1 - \mathrm{sig}(\mathbf{x}^\top \mathbf{w})) \tag{7}$$

ただし，$\mathrm{sig}(\cdot)$ はシグモイド関数であり，$\mathbf{w}$ は重みベクトルである．一般性を失うことなく，バイアス項を扱うため入力 $\mathbf{x}$ の最初の要素 $x^{(1)}$ は定数 1 であると仮定しておく．

このモデルを公平性を扱えるように修正する．センシティブ特徴に予測モデルが依存するようにするために，センシティブ特徴のそれぞれの値ごとにロジスティック回帰モデルを作る．

$$\Pr[\hat{y}|\mathbf{x}, s] = \Pr[\hat{y}|\mathbf{x}; \mathbf{w}^{(s)}]$$

重みパラメータ $\mathbf{w}^{(s)}$, $s \in \{0, 1\}$ はセンシティブ特徴の各値ごとに必要となる．PR モデルでは，過学習を避けるための L_2 正則化項 $\|\mathbf{\Theta}\|_2^2$ と，公平性を強化する偏見除去正則化項 $\mathrm{R}_{\mathsf{PR}}(Y, S)$ の 2 種類の正則化項を採用する．負の対数尤度関数にこれら二つの正則化項を加えたものが PR モデルの目的関数である．

$$\mathrm{loss}(\{\mathbf{w}^{(s)}\}; \mathcal{D}) = - \sum_s \mathcal{L}(\mathcal{D}_s) + \eta \, \mathrm{R}_{\mathsf{PR}}(Y, S) + \frac{\lambda}{2} \sum_s \|\mathbf{w}^{(s)}\|_2^2 \tag{8}$$

ただし，λ と η は正の正則化パラメータで，$\mathcal{L}(\cdot)$ は対数尤度関数である．

文献 [Kamishima 12] のモデルベース独立性を満たす PR 法の場合，$\hat{Y}$ と S の非独立性を測るためこれらの変数の相互情報量を用いた．

$$\mathrm{R}_{\mathsf{PR\text{-}MI}}(Y, S) = n \sum_{\hat{Y}, S} \Pr[\hat{Y}, S] \ln \frac{\Pr[\hat{Y}, S]}{\Pr[\hat{Y}] \Pr[S]} \tag{9}$$

なお，n 倍してあるのは，尤度項とオーダーを揃えるためである．式中の $\Pr[\hat{Y}, S]$ は式 (3) から導出でき，この分布 $\Pr[\hat{Y}, S]$ から他の分布 $Pr[\hat{Y}]$ と $\Pr[S]$ も導くことができる．この正則化項は解析的に微分可能なので，目的関数 (8) は効率的な勾配降下型の手法で最適化できる．このモデルを PR-MI と略記する．

この偏見削除正則化項を実独立性を満たすように修正する．そこで，式 (9) の $\Pr[\hat{Y}, S]$ を $\Pr[\tilde{Y}, S]$ と置き換えて，次式を得る．

$$\mathrm{R}_{\mathsf{PR\text{-}AI}}(Y, S) = n \sum_{\tilde{Y}, S} \Pr[\tilde{Y}, S] \ln \frac{\Pr[\tilde{Y}, S]}{\Pr[\tilde{Y}] \Pr[S]} \tag{10}$$

同時分布 $\Pr[\tilde{Y}, S]$ は，式 (5) と (6) から導出できる．このモデルを PR-AI と略記する．わずかな修正ではあるが，これにより

表 1: 通常のロジスティック回帰（LR）と公平ロジスティック回帰（PR-MI 法と PR-AI 法）の比較

Methods	Adult dataset			Dutch dataset		
	Acc	CVS	NMI	Acc	CVS	NMI
LR	0.862	0.170	4.36×10^{-02}	0.819	0.171	2.20×10^{-02}
PR-MI	0.822	0.055	1.81×10^{-02}	0.792	0.162	2.30×10^{-02}
PR-AI	0.825	0.008	6.03×10^{-05}	0.715	0.001	1.77×10^{-06}

3. 章のように公平性を大きく改善できる．しかし残念ながらこの偏見削除正則化項 $\mathrm{R}_{\text{PR-AI}}(\tilde{Y}, S)$ は，式 (6) に不連続な変換があるため微分できない．そのため，この目的関数を最適化するには勾配がなくても適用できる最適化手法を用いる必要がある．しかし，こうした手法では，パラメータ数を $|\boldsymbol{\Theta}|$ として，目的関数を $O(|\boldsymbol{\Theta}|^2)$ 回評価する（[Bishop 08] の 5.2.3 節などを参照）しなくてはならない．これは，勾配を用いる最適化手法の評価回数 $O(|\boldsymbol{\Theta}|)$ よりも多いため，このモデルの最適化は一般に非効率的である．

3. 公平ロジスティック回帰の性能評価

モデルベース独立性ではなく，実独立性を満たすようにすることで公平ロジスティック回帰の予測精度と公平性の間のトレードオフが改善されるかを検証する．

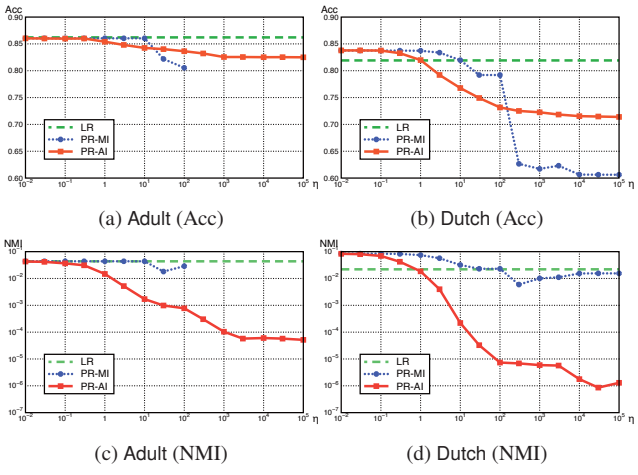
3.1 実験条件

実験に用いたベンチマークデータ[1]は文献 [Žliobaitė 11] で用いられたものである．一つ目は adult データ（別名 census income データ）であり，元データは URI レポジトリ [Frank 10] で配布されている．このデータ集合を Adult で参照する．クラス変数は個人の収入が高いかどうかの二値であり，センシティブ特徴は個人の性別である．データ数は 15,696 個，非センシティブな特徴数は 12 個で，どの特徴も離散である．二つ目は Dutch census で，これを Dutch で参照する．クラス変数は個人の職業が高収入のものか，そうでないかを表し，センシティブ特徴は個人の性別である．データ数は 60,420 個，非センシティブ特徴数は 10 個で，どの特徴も離散である．

5 分割の交差確認を行い，文献 [Kamishima 12] で用いた評価指標を求めた．公平ロジスティック回帰の性能評価のため，どれだけ正しくクラスラベルを予測できたかだけでなく，どれだけ厳密に公平性制約を満たすことができたかも評価する必要がある．なぜなら，予測精度と公平性はトレードオフの関係にあるからである．予測精度の評価には，正しくラベル付けできた標本の割合である正解率 (Acc) を用いた．正解率が高いほど，より正確にクラスが予測できている．公平性の評価には 2 種類の指標を用いた．一つ目は，S=1 で正ラベルになる割合から S=0 での正ラベルの割合を引いた CV スコア (CVS) で，0 に近づくほどクラス変数はセンシティブ特徴と独立になる．二つ目は，正規化相互情報量 (NMI) で，$\tilde{Y}$ と S の相互情報量を $[0, 1]$ の範囲になるように正規化したものである．NMI が小さくなると，より公平な決定がなされたことにになる．

3.2 実験結果

実験結果を表 1 に示す．LR は，センシティブ情報を取り除いた通常のロジスティック回帰である．なお，予測精度と公平性のトレードオフを調整するパラメータ η は，PR-MI では 3×10^1 に，PR-AI では 1×10^4 に設定した．



(a) Adult (Acc) (b) Dutch (Acc)

(c) Adult (NMI) (d) Dutch (NMI)

図 1: η に伴う予測精度 Acc と公平性 NMI の変化

NOTE: 横軸はパラメータ η，縦軸はキャプションに示した統計量である．緑の破線，青の丸付き点線，および赤の四角付き実線はそれぞれ LR，PR-MI，および PR-AI の結果である．Acc は大きいほどより正確な，MNI は小さいほどより公平な決定ができていることを示す．

まず，通常のロジスティック回帰（LR）と公平ロジスティック回帰（PR-MI と PR-AI）とを比較する．LR では，公平性指標 CVS と NMI に注目すると十分な公平性は達成できていない．このように単にセンシティブ情報をモデルから取り除くだけでは red-lining 効果のため公平な決定はできないことが分かる．それに対し，Dutch での PR-MI の場合を除き，公平ロジスティック回帰は通常のロジスティック回帰より公平な決定をしている．一方で，センシティブ特徴に含まれる情報を予測に使わないようにしているため，予測精度は低下している．

次に，モデルベース独立性の代わりに，実独立性を達成することの利点を検証する．公平性に関しては PR-AI が PR-MI よりどちらの指標でも非常に改善されている．表では PR-AI の予測精度は，Dutch では PR-MI より悪いが，Adult では良い．しかし，Dutch の場合でも，公平性が表の PR-MI と同等である η=3 の状況では PR-AI の予測精度は 0.792 であり，PR-MI と同等である．以上のことから，実独立性を達成することで，同等の公平性ではより予測精度の高い分類器が得られているといえる．

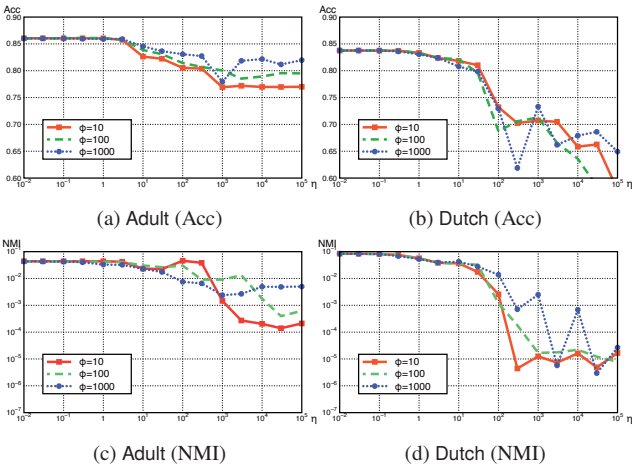
予測精度と公平性のトレードオフについてさらに検証する．釣り合いを調整するパラメータ η を変えたときの予測精度 Acc と公平性 NMI の変化を図 1 に示す．PR-MI では η が一定以上になると極端に予測精度が低下してしまい，それ以上は公平性を強化できなくなる問題生じるが，PR-AI ではそのような現象は見られず，η に応じて公平性は向上させることができる．

以上の実験結果をまとめておく．公平ロジスティック回帰は通常のロジスティック回帰より公平な決定ができるが，それに伴って予測精度はやや低下する．モデルベース独立性の代わりに実独立性を達成することで，同水準の公平性でより正確な予測が実現できる．

4. PR-AI モデルの最適化手法の改良

実験の結果，モデルベース独立性の代わりに実独立性を達成することで，より良い予測精度と公平性のトレードオフが実現できることが分かった．しかし，実独立性を達成する PR-AI

*1 https://sites.google.com/site/conditionaldiscrimination/



(a) Adult (Acc) (b) Dutch (Acc)

(c) Adult (NMI) (d) Dutch (NMI)

図 2: 平滑化を用いた緩和手法での予測精度 Acc と公平性 NMI の変化

NOTE: 横軸はパラメータ η，縦軸はキャプションに示した統計量である．赤の四角付き実線，緑の破線，および青の丸付き点線は，それぞれ ϕ が 10，100，および 1000 である場合の結果である．

モデルは 2.3 節の最後で述べたように，目的関数が微分できないため，効率的に最適化できない問題がある．ここでは，この目的関数を平滑な関数で近似して微分可能にすることで効率的に最適化する手法について予備的検討を行う．

4.1 平滑化した近似目的関数

前述のように，PR-AI モデルの目的関数（式 (8)）には勾配を用いた最適化手法を適用できない．これは，式 (10) には，不連続なステップ関数を含む式 (6) があるため，この目的関数は $\Pr[\hat{Y}{=}1|\mathbf{x}, s] = \Pr[\hat{Y}{=}0|\mathbf{x}, s]$ なる点で不連続になり，微分できないことが理由である．この問題を避けるため，この式 (6) のステップ関数をシグモイド関数で置き換えて平滑化する．しかし，この置き換えは $\Pr[\tilde{Y}|\mathbf{X}{=}x, S{=}s]$ を $\Pr[\hat{Y}|\mathbf{X}{=}x, S{=}s]$ に置き換えることと等価であり，$\mathrm{R}_{\text{PR-MI}}(Y, S)$ と等しい偏見除去正則化項になってしまい，明らかに無意味である．

そこで，$\Pr[\tilde{Y}|\mathbf{X}{=}x, S{=}s]$ をモデル化するときに，よりよくステップ関数を近似できるように，より急激に変化する関数を用いる．ここで，ϕ を大きな正定数とすれば，シグモイド関数 $\mathrm{sig}(\phi x)$ はより急激に変化するようになる．この修正したシグモイド関数を用いて，式 (6) を近似する．

$$\Pr[\tilde{Y}|\mathbf{X}{=}x, S{=}s] \approx$$
$$y\ \mathrm{sig}(\phi \mathbf{x}^\top \mathbf{w}^{(s)}) + (1-y)(1-\mathrm{sig}(\phi \mathbf{x}^\top \mathbf{w}^{(s)})) \quad (11)$$

ϕ が正の無限大であれば，式 (11) は (6) に等しくなるため，ϕ が大きな値である方が望ましい．一方で，大きすぎると計算中にオーバーフローを生じて計算できない．よって，ϕ はオーバーフローとならない程度に大きな値にする必要があり，この ϕ この調整が微妙である点が，この平滑化を用いた近似手法の短所である．一方で，目的関数は微分可能であるため，効率的な最適化手法を適用できる．

4.2 緩和手法の効果の検証実験

図 2 は，平滑化による緩和手法の予測精度と公平性の変化である．この図から得られる結果をまとめる．Adult では PR-MI と比べてあまり改善はされていないが，Dutch ではより公平な予測が実現できている．一方で，PR-AI と比べると，計算は速かったが，どちらのデータ集合でも同等の公平性では予測精度は悪く，これらの間のトレードオフは悪い．さらに，パラメータ ϕ の値に対して結果は大きく変動するためこれをうまく調整する必要があることや，特に ϕ が大きいときに指標の η に対する変化も不安定という問題点もある．効率的に実独立性を達成できる分類器を学習するには，ϕ を適応的に調整する手段が必要になるだろう．

5. まとめ

本稿では，公平配慮型分類問題でのモデルベース独立性と実独立性の概念を示し，モデルベース独立性の代わりに実独立性を達成することで予測精度と公平性よりよいトレードオフを実現できることを実験的に確認した．今後は，予備的検討を行った高速化について研究を進める予定である．

謝辞：本研究は JSPS 科研費 JP24500194，JP15K00327，および JP16H02864 の助成を受けた．

参考文献

[Bishop 08] Bishop, C. M.: パターン認識と機械学習 — ベイズ理論による統計的予測, 上下, シュプリンガー・ジャパン (2007–2008), [監訳：元田 浩 他；翻訳：神嶌 敏弘 他]

[Calders 10] Calders, T. and Verwer, S.: Three naive Bayes Approaches for Discrimination-free Classification, *Data Mining and Knowledge Discovery*, Vol. 21, pp. 277–292 (2010)

[Frank 10] Frank, A. and Asuncion, A.: UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences (2010), ⟨`http://archive.ics.uci.edu/ml`⟩

[Hardt 16] Hardt, M., Price, E., and Srebro, N.: Equality of Opportunity in Supervised Learning, in *Advances in Neural Information Processing Systems 29* (2016)

[Kamishima 12] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J.: Fairness-aware Classifier with Prejudice Remover Regularizer, in *Proc. of the ECML PKDD 2012, Part II*, pp. 35–50 (2012), [LNCS 7524]

[Kamishima 18] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J.: Model-based and Actual Independence for Fairness-aware Classification, *Data Mining and Knowledge Discovery*, Vol. 32, pp. 258–286 (2018)

[Zafar 17] Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P.: Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment, in *Proc. of the 26th Int'l Conf. on World Wide Web*, pp. 1171–1180 (2017)

[Žliobaitė 11] Žliobaitė, I., Kamiran, F., and Calders, T.: Handling Conditional Discrimination, in *Proc. of the 11th IEEE Int'l Conf. on Data Mining* (2011)