電子情報通信学会研究報告, IBISML 2014-28

社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

公正配慮型分類器の公正性に関する分析

神嶌　敏弘†　　赤穂昭太郎†　　麻生　英樹†　　佐久間　淳††

† 産業技術総合研究所
〒 305–8568 茨城県つくば市梅園 1–1–1 産総研つくば中央第 2
†† 筑波大学，〒 305-8577 茨城県つくば市天王台 1–1–1
E-mail: †mail@kamishima.net, ††{s.akaho,h.asoh}@aist.go.jp, †††jun@cs.tsukuba.ac.jp

あらまし　特定の情報の影響を排除するという公正性を保つ公正配慮型分類器において，非常に高い公正性を達成できる Calders と Verwer の 2 単純ベイズ法の理論解析を行う．その原因が分類決定則とモデルバイアスの影響であることを示し，この結果に基づいて，明示的な理論基盤をもつように既存手法を改良し，拡張する．
キーワード　公正配慮型データマイニング, 差別配慮型データマイニング, 単純ベイズ

Analysing the Fairness of Fairness-aware Classifiers

Toshihiro KAMISHIMA†, Shotaro AKAHO†, Hideki ASOH†, and Jun SAKUMA††

† National Institute of Advanced Industrial Science and Technology (AIST),
AIST Tsukuba Central 2, Umezono 1-1-1, Tsukuba, Ibaraki, 305-8568 Japan
†† University of Tsukuba, 1-1-1 Tennodai, Tsukuba, 305-8577 Japan
E-mail: †mail@kamishima.net, ††{s.akaho,h.asoh}@aist.go.jp, †††jun@cs.tsukuba.ac.jp

Abstract　Calders and Verwer's two-naive-Bayes is one of fairness-aware classifiers, which classify objects while excluding the influence of a specific information. We analyze why this classifier achieves very high level of the fairness, and show that this is due to a decision rules and a model bias. Based on these findings, we develop methods that are grounded on rigid theory and are applicable to wider types of classifiers.
Key words　fairness-aware data mining, discrimination-aware data mining, naive Bayes classifier

1. はじめに

公正配慮型データマイニングの目的は，公正性，差別，中立性，および独立性などの潜在的な問題を考慮しつつデータを分析することである．社会的な差別を回避することは，このマイニング技術の代表的な適用事例である．与信，保険料率設定，就職などといった個人の生活にとって重要な決定に，データマイニング技術はますます使われるようになっている．貸付履歴に統計的予測技術を適用して行う与信の決定などは，その例であり，もしこれらの決定が，性別，宗教，人種，民族，ハンディキャップ，政治信条などの個人のセンシティブな情報に基づいたものであれば，それは社会的・法的に不公正であると考えられる．Pedreschi らによる不公正な決定を検出する公正配慮型データマイニングの提案以降，いくつかのマイニングタスクが提案されている．

本論文では，公正配慮型データマイニングのタスクの一つである公正配慮型分類問題について論じる．これは分析結果の公正性を考慮する分類器を設計することを目的とするもので，ここでは，Calders と Verwer の 2 単純ベイズ法 (CV2NB 法) に注目する．この CV2NB 分類器は，他の公正配慮型分類器と比べて高い公正性を達成している．しかし，この方法はやや発見的な後処理によって公正性を強化していることから，背後の統計モデルが不明瞭になっているので，高い公正性を達成を達成できる理由は不明確であった．

本研究の最初の寄与は，この CV2NB 法の性能が優れている理由を明確にすることである．そのために，簡潔な比較モデルを導入し，このモデルの性能が CV2NB モデルより悪い理由を分析する．この比較モデルは，生成モデルによる分類を公正なものにする変換である仮説的公正分解を単純ベイズに適用したものである．他の公正配慮型分類器と同様に，この比較モデルも CV2NB 法より性能が悪いことを実験的に確認する．

その後，この比較モデルの性能を悪化させる二つの原因を示す．第 1 の原因は，真の分布と推定分布の乖離を引き起こすモデルバイアスで，この乖離が公正性を悪化させる．第 2 の原因は，クラスラベルは確定的決定則で確定的に選ばれるにもかかわらず，分類モデルでは確率的にラベルを選択することを仮定

しているという不一致である．

本研究の第 2 の寄与は，CV2NB 法で生成されるモデルを模倣したとみなせるモデル化手法を開発したことである．この，実公正分解と呼ぶ手法ではは，上述のモデルバイアスや確定的決定則によって生じる乖離を修正する．この修正により，仮説的なクラスラベルとではなく，実際のラベルと，センシティブな特徴との関連を断つことができる．この方法が，CV2NB 法と同等であることを実験的にも示す．

本研究の第 3 の寄与は，実公正分解を，生成モデルによる分類器以外の，識別モデルや識別関数による分類器にも適用できるように拡張したことである．この拡張手法により，公正な決定を出力するように任意の種類の分類器を修正できる．

本論文の構成は以下のとおりである．2. 節は公正配慮分類のタスクと手法を簡潔に紹介する．3. 節で，仮説公正分解と，そのベンチマークデータに対する実験結果を示したあと，4. 節で，その性能が劣る原因を分析する．5. 節では，これらの問題を解消した実公正分解法を開発し，この手法の有効性を実験的に示す．6. 節では，生成モデル以外の，識別モデルや決定関数に基づく分類器にも適用できるように，この手法を拡張する．最後の 7. 節はまとめである．

2. 公正配慮型分類

本節では公正配慮型分類の概要を述べる．記法と問題設定に続き，形式的な公正性の概念を導入する．その後，各種の公正配慮型分類手法，特に Calders と Verwer の 2 単純ベイズ法について述べる．

2.1 表記と問題設定

公正配慮型データマイニングの目的は，公正性の潜在的問題に配慮しつつデータを分析することである．この公正配慮型データマイニングのタスクの一つである，**公正配慮型分類** (fairness-aware classification) は，公正性，差別，中立性，独立性などの問題を考慮しつつデータを分類する．この公正配慮型分類では Y，$\mathbf{X}$，および S の 3 種類の変数を用いる．確率変数 S と $\mathbf{X}$ は，それぞれセンシティブと非センシティブ特徴を表す．センシティブ特徴は，公正性を保証すべき情報を表す．例えば，与信決定の場合では，社会的・法的見地に基づいて指定された性別，人種，宗教などに相当し，与信の判定はこれらの特徴に関して公正でなければならない．一方の，非センシティブ特徴は，センシティブ特徴以外の全ての特徴である．確率変数 Y は，分類対象のクラスを表現するクラス変数である．

本論文では，確率変数をさらに制限する．クラス変数 Y は二値クラスを表し，その定義域は $\{0,1\}$ とする．クラス 1 と 0 はそれぞれ，ローンの請求に対する可と不可といった，有利と不利な結果を表す．S も二値に制限し，その定義域は $\{0,1\}$ である．センシティブ特徴の値がそれぞれ 1 と 0 である分類対象を，それぞれ非保護状態と保護状態にあるという．保護対象は，社会的に不公正な待遇から保護されるべき個人や対象を表す．ある分類対象集合のうち，保護状態にある分類対象のグループを保護グループ，残りの対象全てを非保護グループと呼ぶ．$\mathbf{X}$ は，K 個の確率変数 $X^{(1)},\dots,X^{(K)}$ で構成され，各変数は離散でも連続でもよい．

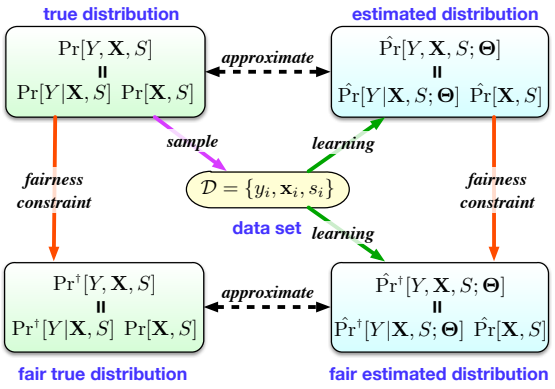

図 1　分布の表記

Fig. 1　Notations of distributions

各分類対象を，真の分布 $\Pr[\mathbf{X},S]$ から生成された実現値の対 $(\mathbf{x},s)$ で表す．この対象を分類するクラスの実現値 y は，真の条件付き分布 $\Pr[Y|\mathbf{X}{=}\mathbf{x},S{=}s]$ から生成する．この真の分布 $\Pr[Y|\mathbf{X},S]$ は，センシティブ特徴に依存した潜在的に不公正な決定をしうることに注意されたい．これらの真の分布自体は分からないが，真の同時分布 $\Pr[Y,\mathbf{X},S]=\Pr[Y|\mathbf{X},S]\Pr[\mathbf{X},S]$ から得られた標本は観測できる．この手続きを N 回繰り返してデータ集合 $\mathcal{D}=\{(y_i,\mathbf{x}_i,s_i)\}$, $i=1,\dots,N$ を得る．$\mathcal{D}[cond]$ は，$\mathcal{D}$ 中で条件 $cond$ を満たす全てのデータで構成される集合を表すものとする．あるパラメトリックなモデルの族 $\hat{\Pr}[Y|\mathbf{X},S;\boldsymbol{\Theta}]$ と，訓練データ集合 $\mathcal{D}$ に対し，真の分布を最もよく近似するであろう分布を表すようにパラメータ $\boldsymbol{\Theta}$ を求めることが，標準的なあてはめ問題の目的である．

ここで，対応するセンシティブ特徴値に関して公正なクラスの値を生成する公正な真の分布 $\Pr^\dagger[Y|\mathbf{X},S]$ が存在すると仮定する．この分布を，真の分布 $\Pr[Y|\mathbf{X},S]$ に，2.2 節で述べるある事前に定めた公正性制約を強制することで得る．実世界での決定は公正性制約を満たさない可能性があるので，真の分布とは異なり，この公正な真の分布からは標本を得ることさえもできない．それゆえ，公正な真の分布からの標本の代わりに，真の分布からの標本を訓練データとして用いる．この訓練データと，公正な真の分布が満たすべき公正性制約を満たしている公正なパラメトリックモデルの族 $\hat{\Pr}^\dagger[Y|\mathbf{X},S;\boldsymbol{\Theta}]$ に対して，公正な真の分布を最もよく近似できるような公正な推定推定分布となるようにパラメータを最適化することが公正配慮型分類の目的である．以上の分布の表記については図 1 にまとめた．

2.2 分類における公正性

ここではデータマイニングにおける公正性の形式的定義についてまとめる．公正性制約は，形式的には，ある公正性指標が満たすべき不等式である．公正性指標は，観測・推定された $(Y,\mathbf{X},S)$ 上の分布に基づいて公正性の度合いを測る．多くの種類の公正性指標が提案されてきた：拡張リフト [1]，CV スコア [2]，相互情報量 [3], [4]，χ^2 統計量 [5], [6]，η 中立性 [7]，および統計的一致性と Lipschitz 条件の組み合わせ [8], [9]．もしこれらの公正性指標が，ある指定した値よりも悪ければ，その

ときの決定は不公正であるとみなす．

ほとんど全ての公正性指標は，クラス変数 Y とセンシティブ特徴 S 間の統計的独立性と基本的に関係がある．ここで，単にセンシティブ特徴を計算過程から排除するだけでは，センシティブ特徴の間接的な影響のため，不適切な決定を避けるには不十分であることは重要である．非センシティブ特徴ベクトル中のある変数 X がセンシティブ特徴と強く相関している場合を考えよう．例えば，特定の人種がある地域にまとまって住んでいると，センシティブ特徴 race が，address などの非センシティブ特徴と相関することになる．この場合，センシティブ特徴を使わなくても，クラス変数は間接的にセンシティブ特徴の影響を受ける **red-lining 効果**と呼ばれる現象が生じる．形式的には，Y と S が条件なしに独立 $Y \perp\!\!\!\perp S$ ではなく，条件付き独立 $Y \perp\!\!\!\perp S \mid \mathbf{X}$ である場合に red-lining 効果は生じる．なお，$A \perp\!\!\!\perp B$ は確率変数 A と B の（条件なし）独立性を，$A \perp\!\!\!\perp B \mid C$ は，確率変数 C が与えられたときの A と B の独立性をそれぞれ表す．

2.3 公正配慮型分類の手法

ここでは公正配慮型分類用の手法を俯瞰する．

2.3.1 Calders と Verwer の 2 単純ベイズ法

Calders と Verwer の 2 単純ベイズ法 (Calders and Verwer's two-naive-Bayes method; CV2NB 法) [2] を紹介し，その理論背景を論じる．この手法の生成モデルは次式である：

$$\hat{\Pr}[Y, \mathbf{X}, S] = \hat{\Pr}[Y \mid S]\,\hat{\Pr}[S] \prod_k \hat{\Pr}\left[X^{(k)} \mid Y, S\right] \tag{1}$$

標準の単純ベイズモデルでは，各 $X^{(k)}$ は Y のみに依存しているのに対し，CV2NB モデルでは Y の他に S にも依存している．なお，センシティブ特徴の値に応じてあたかも二つの単純ベイズモデルが学習されるのでこの方法は 2 単純ベイズ法と呼ばれている．公正に分類するために，図 2 の後処理アルゴリズムによって，同時分布 $\hat{\Pr}[Y, S] = \hat{\Pr}[Y \mid S]\,\hat{\Pr}[S]$ を修正する．このアルゴリズムの停止後に，モデルパラメータ $\hat{\Pr}^{\dagger}[y, s]$ は $N(y, s)$ から導出できる．

元のモデルを二つの条件，(1) 分類での公正性，および (2) クラス分布の保存を満たすように，この後処理アルゴリズムは設計されている．第一に，公正性条件を満たすために，この後処理は Calders-Verwer's discrimination score (CV スコア) を公正性指標として利用する．この CV スコアは，保護分類対象が有利な決定を受ける確率から，非保護分類対象が有利な決定を受ける確率を引いたものである：

$$\hat{\Pr}[Y{=}1 \mid S{=}1] - \hat{\Pr}[Y{=}1 \mid S{=}0] \tag{2}$$

CV スコアが増えると，非保護グループの個人は有利な決定をより頻繁に受けるとともに，保護グループのメンバーの個人が有利な決定をあまり頻繁に受けなくなる．Y と S が共に二値変数である場合には，CV スコアが 0 であることが Y と S の独立性を含意することは容易に示せる．後処理の 6–7 行と 9–10 行は得られる分布に対する CV スコアを 0 に近づけるような設計になっている．具体的には，6 行で有利に扱われる保護グループの個人を増やすと共に，7 行で不利に扱われる個人を減らしている．9–10 行も，同様に非保護グループの個人数を調整している．アルゴリズムのメインループは，CV スコアが 0 に近づいたときに 16 行で終了するので，そのときに得られる分布 $\hat{\Pr}[Y, S]$ は，Y と S の独立性条件を満たす．

```
1:  procedure CV2NB Post-Process(N(Y, S))
2:      disc ← a CV score of the predicted classes by the current model
3:      while disc > 0 do
4:          numpos ← the number of positively classified samples by the current model
5:          if numpos < the number of positive samples in D then
6:              N(Y=1, S=0) ← N(Y=1, S=0) + ΔN(Y=0, S=1)
7:              N(Y=0, S=0) ← N(Y=0, S=0) − ΔN(Y=0, S=1)
8:          else
9:              N(Y=0, S=1) ← N(Y=0, S=1) + ΔN(Y=1, S=0)
10:             N(Y=1, S=1) ← N(Y=1, S=1) − ΔN(Y=1, S=0)
11:         end if
12:         if Any entry of N(Y, S) is negative then
13:             cancel the previous update of N(Y, S) and abort
14:         end if
15:         Recalculate P̂r[Y|S] and a CV score, disc, based on updated N(Y, S)
16:     end while
17: end procedure
```

図 2　CV2NB モデル用後処理アルゴリズム

Fig. 2　A post-processing algorithm for a CV2NB model

NOTE: Δ は小さな正数のパラメータで，原著と同じ 0.01 に設定した．$N(y, s)$ は訓練データ中で $Y{=}y \wedge S{=}s$ の条件を満たすデータ数．なお，元のアルゴリズムは停止しない場合があるため，$N(Y, S)$ の非負性を保証するよう 12–14 行を追加している．

第 2 の条件については，クラス分布を元の分布に近くなるように，すなわち $\hat{\Pr}^{\dagger}[Y] \approx \hat{\Pr}[Y]$ となるように 5 行で修正している．しかし，Y の周辺分布は，3 行の終了条件では考慮されていないので，得られる Y の分布が常に標本分布と一致するわけではない．

2.3.2 棄却オプションベース分類

Kamiran らは，公正性制約を満たすようにクラス事後分布からクラスラベルを決定する理論について論じた [10]．標準的な分類では，$\hat{\Pr}[Y{=}1 \mid \mathbf{X}] \geq \hat{\Pr}[Y{=}0 \mid \mathbf{X}]$ の不等式をクラス事後確率が満たすときに，対象をクラス 1 に分類する．この条件は $\hat{\Pr}[Y{=}1 \mid \mathbf{X}] \geq 0.5$ と等価であるが，この 0.5 を決定しきい値と呼ぶ．

棄却オプションベース分類（Reject Option based Classification; ROC 法）と呼ぶ提案手法は，公正な分類をするように，この決定しきい値を変更する．保護グループの個人に対しては，より頻繁に有利な決定がなされるように，このしきい値を減らすとともに，非保護グループの個人に対しては，このしきい値を増やす．この手法では，公正性の制約を満たすと同時に，予測精度をあまり下げないようにするため，分類結果の確信度の低い決定境界付近にある対象のクラスラベルを変更する．形式的には，しきい値パラメータ $0.5 \leq t < 1$ を導入し，$\hat{\Pr}[Y{=}1 \mid \mathbf{X}, S{=}0] \geq 1 - t$ であるような，$S{=}0$ の対象はクラス 1 に分類する．逆に，$S{=}1$ の対象は，$\hat{\Pr}[Y{=}1 \mid \mathbf{X}, S{=}1] \geq t$ の場合にクラス 1 に分類する．

著者らはこの決定則と，誤分類コストを最小化するように対象を分類するコスト考慮型学習 [11] との関係を指摘している．誤分類コストとは，推定クラスと真のクラスが異なったときに

与える罰則コストのことである．標準的な分類では，真のクラスが 1 であるとき（0 であるとき）にそれを 0 と誤って（1 と誤って）分類するときのコストは 1 である．これを ROC 則の場合で考察する．S=0 である保護対象では，真のクラスが 0 のものを誤分類するコストは 1 のままだが，真のクラスが 1 のもののコストは $t/(1-t)$ に増やす．これは，もし有利な決定を受けるべき保護対象が不利に扱われた場合には，より大きな誤分類コストを課していると共に，不利な決定をうけるべき場合についてはそのままにするということである．非保護グループの個人の扱いは逆で，真のクラスが 1 の誤分類コストは $t/(1-t)$ に増やすが，0 ならばそのままである．すなわち，不利な決定を受けるべき非保護対象が有利に扱われるときに，より大きな誤分類コストを課す．

その他に公正配慮型分類を扱った研究としては [4], [7], [9], [12] などがある．

3. 仮説公正分解

前節で述べた公正配慮型分類手法の中でも，公正性に関しては CV2NB 法は非常に優れていた．しかし，なぜ CV2NB 法が優れているのかは，発見的な後処理によって公正性を強化しており，どのような統計的モデルが獲得されているのかが不明瞭であるため，よくわからなかった．

この原因を特定するため，CV2NB モデルと類似した生成モデルを導入する．これは，分類の生成モデルに公正性制約を強制する仮説公正分解を適用して得られる．他の公正配慮型分類手法と同様に，このモデルも CV2NB 法より不公正な決定しかできないことを実験的に確かめる．このように公正性に関して性能が劣る原因は，モデルバイアスと確定的な決定則であることを次節で示す．

3.1 仮説公正分解の手法

まず，クラス変数と特徴との同時分布 $\hat{\Pr}[Y, \mathbf{X}, S]$ をモデル化するための仮説公正分解の手法について述べる．2.3.1 節の CV2NB 法の後処理は，分類の公正性とクラス分布の保存の二つの制約を満たすようになっていた．そこで，これらの制約を満たすような同時分布のモデルを考える．最初の公正性条件に注目すると，この条件は，形式的には推定分布が $Y \perp\!\!\!\perp S$ の条件を満たす，すなわち $\hat{\Pr}^\dagger[Y, S] = \hat{\Pr}^\dagger[Y]\hat{\Pr}^\dagger[S]$ が成立することである．この制約を分類の生成モデルに組み込んで次式を得る：

$$\begin{aligned}\hat{\Pr}^\dagger[Y, \mathbf{X}, S] &= \hat{\Pr}^\dagger[Y, S]\hat{\Pr}^\dagger[\mathbf{X}|Y, S] \\ &= \hat{\Pr}^\dagger[Y]\hat{\Pr}^\dagger[S]\hat{\Pr}^\dagger[\mathbf{X}|Y, S] \end{aligned} \tag{3}$$

生成モデルで Y と S を無関係にするこの手法を公正分解と呼ぶことにする．この公正分解は仮説空間上の分布に適用するので，特に**仮説公正分解** (Hypothtical Fair-Factorization) と呼び，実際の分布について分解する 5. 節の方法と区別する．

次に，この仮説公正分解を単純ベイズモデルに適用した HFFNB モデル (Hypothetical Fair-Factorization Naive Bayes) について述べる．式 (1) の CV2NB モデルのように，HFFNB でも，Y と S が与えられたとき，各非センシティブ特徴 $X^{(k)}, k = 1, \ldots, K$ は条件付き独立と仮定する．HFFNB モデルでは，さらに，Y と S の間の独立性も仮定する．すなわち公正分解を適用する．その結果，次の HFFNB モデルを得る：

$$\hat{\Pr}^\dagger[Y, \mathbf{X}, S] = \hat{\Pr}^\dagger[Y]\hat{\Pr}^\dagger[S]\prod_k \hat{\Pr}^\dagger\left[X^{(k)}|Y, S\right] \tag{4}$$

Y と S が共に二値変数であるとき，このモデルの最尤推定量は訓練データ集合から容易に導出できる．そして，$\hat{\Pr}^\dagger[Y]$, $\hat{\Pr}^\dagger[S]$, および $\hat{\Pr}^\dagger\left[X^{(k)}|Y, S\right], k = 1, \ldots, K$ は個別に当てはめることができる．$\hat{\Pr}^\dagger[Y = 1]$ は $|D[Y = 1]|/|D|$ で求めることができ，他のパラメータも訓練データ中の事例数の比を求めるだけで同様に計算できる．なお，後の実験では 0 頻度問題を回避するためラプラス平滑化を適用した．

次に，第 2 の条件である，クラス分布の保存に移る．2.3.1 節で述べたように，CV2NB 法の後処理はクラス分布を保存するように設計されてはいるが，常に一致するようにはなっていない．しかし，HFFNB 法では，式 (4) の第 1 因子である $\hat{\Pr}^\dagger[Y]$ が Y の周辺分布に一致することは，HFFNB モデルから S と $\mathbf{X}$ を積分消去することで容易に示せる．よって，Y の周辺分布は，Y の訓練データ D 上の標本分布に一致する．まとめると，CV2NB 法の後処理により満たそうとする二つの条件を，この HFFNB モデルも満たす．

ここで，HFFNB モデルと Kamiran らの ROC 決定則との関連について論じておく．まず，Elkan の文献 [11]. の定理 2 について述べる．この定理によれば，クラス 1 の事前確率が b' で，決定しきい値が p' あるベイズ分類器に対し，事前確率を b に変えたとき，二つの分類器が同じ決定をするようにするように定めた決定しきい値を p とする．このとき，これらの関係は次式となる．

$$p' = \frac{b'p(1-b)}{b - pb + b'p - bb'} \tag{5}$$

HFFNB モデルの場合，公正分解によって，事前確率を $b' = \hat{\Pr}[Y|S]$ から $b = \hat{\Pr}[Y]$ に変えている．HFFNB モデルの決定しきい値が $p = 1/2$ であるとき，もとの分類器で等価にな決定をする分類器の決定しきい値は次式となる．

$$p' = \frac{\hat{\Pr}[Y|S]\,(1 - \hat{\Pr}[Y])}{\hat{\Pr}[Y] + \hat{\Pr}[Y|S] - 2\,\hat{\Pr}[Y]\,\hat{\Pr}[Y|S]} \tag{6}$$

このことから，HFFNB モデルは，元の分類器の決定しきい値を変化させたものと等価であることが分かる．この意味で，HFFNB 法は ROC アプローチの一種とみなせる．

3.2 実　　験

ここでは，HFFNB 法と CV2NB 法の性能を二つのベンチマークデータを用いて比較し，HFFNB 法が CV2NB 法よりも劣ることを確認する．

実験に用いたベンチマークデータ[1]は文献 [13] 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 [14] で配布されている．このデータを 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

(注1)：https://sites.google.com/site/conditionaldiscrimination/

表 1　HFFNB 法と，CV2NB 法および二つのベースライン手法との比較
Table 1　Comparison of our HFFNB method with the CV2NB method and two baselines

	Adult data			Dutch data		
Methods	Acc	CVS	NMI	Acc	CVS	NMI
HFFNB	0.828	0.129	1.52×10^{-2}	0.810	0.312	7.17×10^{-2}
CV2NB	0.828	−0.003	6.89×10^{-6}	0.761	−0.003	8.79×10^{-6}
NB	0.829	0.345	1.16×10^{-1}	0.816	0.365	9.86×10^{-2}
NBns	0.836	0.278	7.62×10^{-2}	0.789	0.162	1.90×10^{-2}

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 [4] で用いた評価指標を求めた．公正配慮型分類器の性能評価のため，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，どれだけ厳密に公正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，予測精度と公正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．公正性の評価には 2 種類の指標を用いた．一つ目は式 (2) の 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，$\hat{Y}$ 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，より公正な決定がなされたことにになる．

比較する手法は 4 種類である．そのうち二つは公正配慮型分類器の HFFNB と CV2NB であり，残り二つは標準的な単純ベイズを用いたベースライン手法である．一つ目のベースラインは，センシティブ特徴と非センシティブ特徴の両方を用いた単純ベイズ分類器で，NB と記す．二つ目のベースラインは，非センシティブ特徴のみを用いた単純ベイズ分類器で，NBns と記す．これらの 4 種類の手法（HFFNB，CV2NB，NB，および NBns）を，2 種類のベンチマークデータ（Adult と Dutch）に適用し，3 種類の評価指標（Acc，CVS，および NMI）を計算した．

実験結果を表 1 に示す．まず二つのベースライン手法 NB と NBns に注目すると，どちらのデータでも NB より NBns の方がより公正な決定をしていることが，CVS と NMI の両方の指標から分かる．これは，モデルからセンシティブな特徴を排除したことでより公正な決定ができることを示している．しかし，CVS と NMI のどちらの指標も 0 よりかなり大きく，単にセンシティブ特徴を取り除くだけでは red-lining 効果のため完全には公正なモデルは学習できなかったことが分かる．

次に HFFNB 法と二つのベースライン手法を比較する．Adult データでは，HFFNB 法の予測精度はベースライン手法より悪い．しかし，Dutch データでは，HFFNB の Acc は NB よりは悪いのに対し，NBns に対しては良かった．公正配慮型のモデルでは，公正性を改善するために，予測精度は一般に下がってしまう．二つの公正性指標をみると，HFFNB 法はどちらのベースライン手法より Adult データではより公正な決定をしたが，Dutch データではできなかった．残念ながら，HFFNB 法では，Dutch データに対して十分に公正なモデルを獲得できなかった．

最後に，HFFNB 法と CV2NB 法とを比較する．予測精度に関しては，HFFNB 法は CV2NB 法よりわずかに良かった．しかし，CV2NB 法は，CVS と NMI のどちらの指標も 0 に非常に近く，ほぼ完全に公正なモデルを獲得できたのに対し，HFFNB 法では十分に公正なモデルを獲得できなかった．

4. なぜ HFFNB 法は失敗したのか？

前節の実験結果のように，HFFNB は，明示的に Y と S の独立性制約を組み込んでいるにもかかわらず，公正なモデルの学習に失敗した．これには二つの原因があると考える．一つ目は，モデルバイアスによって，推定分布が真の分布と乖離してしまうため，学習した分類器の公正性が悪化する．二つ目は，確定的なベイズ決定則により実際のクラスラベルは確定的に選ばれるのに対し，HFFNB モデルでは，それが確率的に決定されることを仮定していることの影響である．

4.1 モデルバイアス

まずモデルバイアスがどのように公正性を悪化させるかを示す．生成モデルに寄る分類では，推定分布 $\hat{\Pr}[Y|\mathbf{X},S]$ に基づいてクラスラベルを予測する．一方，分類対象は真の分布 $\Pr[\mathbf{X},S]$ に従って生成される．推定分布はモデルの部分空間上になければならないが，この制限は真の分布には当てはまらないため，推定分布は真の分布とは一般には異なる．例えば，HFFNB モデルでは，非センシティブ特徴 $X^{(k)}, k=1,\ldots,K$ は，Y と S が与えられたとき互いに条件付き独立と仮定しているが，この仮定は真の分布に対しては一般には成立しない．そのため，$(Y,\mathbf{X},S)$ 上の同時分布は，仮説公正分解した生成モデルとはかけ離れたものとなる：

$$\hat{\Pr}[Y|\mathbf{X},S]\Pr[\mathbf{X},S] \neq \hat{\Pr}[Y]\hat{\Pr}[S]\hat{\Pr}[\mathbf{X}|Y,S] \tag{7}$$

よって，推定同時分布 $\hat{\Pr}[Y|\mathbf{X},S]\Pr[\mathbf{X},S]$ から $\mathbf{X}$ を積分消去して同時分布 $\hat{\Pr}[Y,S]$ を得たとき，この得られた $\hat{\Pr}[Y,S]$ は真の分布とは異なるので，公正性条件 $Y \perp\!\!\!\perp S$ を満たさない．

4.2 確定的決定則

次に確定的な決定則によるクラスラベルの選択の影響について論じる．実際のクラスラベルの分布が生成モデルから導出される分布と等しければ，独立性条件 $Y \perp\!\!\!\perp S$ は満たされる．しかし，実際のラベル y^* は次のベイズ決定則によって確定的に選ばれるので，この条件は成立しない．

$$y^* = \arg\max_y \hat{\Pr}[Y=y|\mathbf{X}=x,S=s]. \tag{8}$$

次に，生成モデルから導出される分布は，ベイズ決定則で選ばれる実際のラベルの分布とどれくらい異なっているかを調べる．このために，二値クラス変数 Y と一つの二値特徴変数 X を含む簡潔なモデルを考える．クラスの事前分布は一様とする，

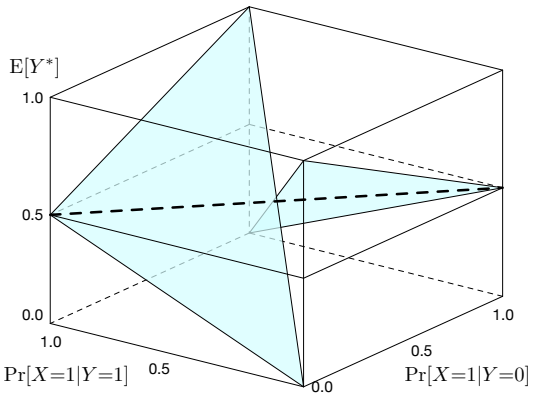

図 3　実際のラベルの期待値 E[Y^*] の変化
Fig. 3　The changes of the expectation of actual labels, E[Y^*]

すなわち $\hat{\Pr}[Y{=}1] = 0.5$ であるとする．他に二つのパラメータ（$\hat{\Pr}[X{=}1|Y{=}0]$ と $\hat{\Pr}[X{=}1|Y{=}1]$）が X と Y の同時分布を表現するために必要となる．このとき，Y がこのモデルから導出される分布に従うなら，その期待値 E[Y] は 0.5 の定数である．さらに，式 (8) の決定則で選ばれる実際のラベルを表す変数 Y^* を考える．二つのパラメータ $\hat{\Pr}[X{=}1|Y{=}0]$ と $\hat{\Pr}[X{=}1|Y{=}1]$ を変化させたときの実際のラベルの期待値 E[Y^*] の変化を図 3 に示す．驚くべきことに，E[[Y] = E[Y^*] の条件が成立するのは，図 3 中の太破線で示した $\hat{\Pr}[X{=}1|Y{=}0] + \hat{\Pr}[X{=}1|Y{=}1] = 1$ が満たされる場合だけである．その結果，二つの変数 Y と Y^* のはほとんど全ての点で乖離し，この乖離のために公正性が保たれなくなる．

5. 実公正分解

前節では，HFFNB 法の性能が低い原因を二つ示した．ここでは，これらの二つの原因を取り除いた公正分解の手法を提案しする．この新しいモデルを，実公正分解単純ベイズ法（AFFNB 法）と呼ぶ．AFFNB 法の性能が CV2NB 法と同等であることを示すことにより，HFFNB 法の性能低下の原因が前節の二つの因子であったことを示す．

5.1 実公正分解単純ベイスモデル

前節での考察を元に，実際の分布を公正分解する，**実公正分解法** (Actual Fair-Factorization method) を提案する．仮説公正分解法では，仮説空間中の分布 $\hat{\Pr}[Y, \mathbf{X}, S]$ で，クラス変数とセンシティブ特徴とを無関係にしていた．しかし，4. 節で述べた二つの原因により，実際の分布はこの仮説の分布とは異なってしまう．第 1 のの原因はモデルバイアスで，第 2 の原因は確定的な決定則を適用したことである．第 1 の原因を修正するため，推定した分布 $\hat{\Pr}^\dagger[\mathbf{X}, S]$ の代わりに，入力の真の分布 $\Pr[\mathbf{X}, S]$ を用いる．第 2 の原因に対処するため，仮説のクラスラベルを含む分布 $\hat{\Pr}^\dagger[Y, \mathbf{X}, S]$ の代わりに，実際のクラスラベルを含む分布 $\hat{\Pr}^\dagger[Y^*, \mathbf{X}, S]$ を考える．ここで，CV2NB 法の後処理も実際のクラスラベルの分布を対象としてることを強調しておきたい．図 2 の後処理の 15 行では，CV スコアを実際に分類した標本の数に基づいて求めている．

決定則の影響で生じる乖離を修正するために，CV2NB 法の後処理で扱う二つの条件，すなわち分類の公正性とクラス分布の保存の条件を，仮説的なラベルではなく，実際のラベルに対して実公正分解では満たすようにする．第 1 の公正性の条件 $Y^* \perp\!\!\!\perp S$ は次式で定式化できる：

$$\hat{\Pr}^\dagger[Y^* = 1|S = s] = \hat{\Pr}^\dagger[Y^* = 1], \text{ for } s \in \{0, 1\} \tag{9}$$

第 2 の分布の保存条件は，実ラベルの分布と標本ラベルの分布の等価性，すなわち $\hat{\Pr}^\dagger[Y^*{=}1] = |D[Y{=}1]|/N$ の条件とみなせる．この条件と式 (9) を併せると，目的の条件は次式となる：

$$\hat{\Pr}^\dagger\left[Y^* = 1|S = s\right] = |D[Y = 1]|/N, \text{ for } s \in \{0, 1\} \tag{10}$$

この条件を満たすため，パラメトリックなモデル $\hat{\Pr}^\dagger[Y^*|S = s; \boldsymbol{\Theta}]$ を導入し，このパラメータを $s \in \{0, 1\}$ について次の最適化問題を解くことで求める．

$$\min_{\boldsymbol{\Theta}} \left(\hat{\Pr}^\dagger\left[Y^* = 1|S = s; \boldsymbol{\Theta}\right] - \frac{|D[Y = 1]|}{N} \right)^2 \tag{11}$$

次に式 (11) のパラメトリックモデル $\hat{\Pr}^\dagger[Y^* = 1|S = s; \boldsymbol{\Theta}]$ について考える．提案生成モデルでは，まず，仮説ラベル Y を式 (1) の CV2NB モデルから生成する．この生成モデルでは，後処理前の CV2NB 法で得られた値にパラメータの値を固定する．ここで，実際のラベル Y^* は，この仮説ラベル Y とセンシティブ特徴 S には依存するが，非センシティブ特徴 $\mathbf{X}$ とは独立であると仮定する．すると，実ラベルと特徴との同時分布は次式となる：

$$\hat{\Pr}^\dagger\left[Y^*{=}1, S{=}s, \mathbf{X}; \boldsymbol{\Theta}\right] = \sum_Y \hat{\Pr}^\dagger\left[Y^*{=}1|Y, S{=}s\right] \hat{\Pr}[Y|S{=}s] \hat{\Pr}[S{=}s] \prod_k \hat{\Pr}\left[X^{(k)}|Y, S{=}s\right]$$

$\sum_Y \hat{\Pr}^\dagger[Y^*|Y, S{=}s] \hat{\Pr}[Y|S{=}s]$ を q_s と置き換えて次式を得る：

$$\hat{\Pr}^\dagger\left[Y^*{=}1, S{=}s, \mathbf{X}; \boldsymbol{\Theta}\right] = q_s \hat{\Pr}[S{=}s] \prod_k \hat{\Pr}\left[X^{(k)}|Y, S{=}s\right] \tag{12}$$

すでにパラメータ $\hat{\Pr}[S{=}s]$ と $\hat{\Pr}[X^{(k)}|Y, S{=}s]$ は固定しているので，求めるべき残りのパラメータは $\boldsymbol{\Theta} = \left\{q_s | s \in \{0, 1\}\right\}$ だけとなる．

この式 (12) のモデルを用いて，式 (11) の最適化問題を解くには $\hat{\Pr}^\dagger[Y^*{=}1|S{=}s]$, $s \in \{0, 1\}$ を計算する必要がある．これらは式 (12) を $\mathbf{X}$ で周辺化し，$\hat{\Pr}[S]$ で割ることで得ることができる．

$$\hat{\Pr}^\dagger\left[Y^*{=}1|S{=}s\right] = \sum_{\mathbf{X}} \hat{\Pr}^\dagger\left[Y^*{=}1|\mathbf{X}, S{=}s\right] \Pr[\mathbf{X}|S{=}s] \tag{13}$$

ここで，推定分布 $\hat{\Pr}^\dagger[\mathbf{X}|S{=}s]$ ではなく，真の分布 $\Pr[\mathbf{X}|S{=}s]$ を使うことが，モデルバイアスの影響を避けるためには重要である．この真の分布による周辺化は，データ集合 $D[S{=}s]$ 上の標本平均で近似できる：

表 2　AFFNB 法と，HFFNB 法および CV2NB 法との比較
Table 2　Comparison of our AFFNB method with HFFNB and CV2NB methods

	Adult data			Dutch data		
Methods	Acc	CVS	NMI	Acc	CVS	NMI
AFFNB	0.828	−0.002	5.43×10^{-6}	0.761	−0.002	2.68×10^{-6}
HFFNB	0.828	0.129	1.52×10^{-2}	0.810	0.312	7.17×10^{-2}
CV2NB	0.828	−0.003	6.89×10^{-6}	0.761	−0.003	8.79×10^{-6}

$$\frac{1}{|D[S{=}s]|}\sum_{(\mathbf{x})\in D[S=s]}\hat{\Pr}^{\dagger}\left[Y^{*}{=}1|\mathbf{X}{=}\mathbf{x},S{=}s\right] \tag{14}$$

ただし，$\hat{\Pr}^{\dagger}[Y^{*}{=}1|\mathbf{X}{=}\mathbf{x},S{=}s]$ は，あるデータ $(\mathbf{x},s)$ が与えられたときに実際のラベルが 1 となる確率である．この確率は，式 (8) の決定則によってラベルが確定的に割り当てられるため，0 または 1 のいずれかの値しかとることはなく，1 になるのは次の条件が満たされる場合である：

$$\hat{\Pr}^{\dagger}\left[Y^{*}{=}1|\mathbf{X}{=}\mathbf{x},S{=}s\right]\geq\hat{\Pr}^{\dagger}\left[Y^{*}{=}0|\mathbf{X}{=}\mathbf{x},S{=}s\right] \tag{15}$$

モデル (12) を用いると，この条件は次式と等価になる：

$$\frac{q_s\,\hat{\Pr}^{\dagger}[S{=}s]\,\hat{\Pr}^{\dagger}[\mathbf{X}{=}\mathbf{x}|Y^{*}{=}1,S{=}s]}{\hat{\Pr}^{\dagger}[\mathbf{X}{=}\mathbf{x},S{=}s]}\geq \frac{(1-q_s)\,\hat{\Pr}^{\dagger}[S{=}s]\,\hat{\Pr}^{\dagger}[\mathbf{X}{=}\mathbf{x}|Y^{*}{=}0,S{=}s]}{\hat{\Pr}^{\dagger}[\mathbf{X}{=}\mathbf{x},S{=}s]}$$

$$q_s\geq\frac{\hat{\Pr}^{\dagger}[\mathbf{X}{=}\mathbf{x}|Y^{*}{=}0,S{=}s]}{\sum_{y\in\{0,1\}}\hat{\Pr}^{\dagger}[\mathbf{X}{=}\mathbf{x}|Y^{*}{=}y,S{=}s]}. \tag{16}$$

これと式 (14) を併せると，$\hat{\Pr}^{\dagger}[Y^{*}{=}1|S{=}s]$ を得る：

$$\hat{\Pr}^{\dagger}\left[Y^{*}{=}1|S{=}s\right]=\frac{1}{|D[S{=}s]|}\sum_{(\mathbf{x}_i)\in D[S=s]}\mathrm{I}[\mathbf{x}_i,s] \tag{17}$$

ただし，$\mathrm{I}[\mathbf{x},s]$ は，式 (16) の不等式が成立するときに 1 をとり，そうでなければ 0 をとる指示関数である．

式 (17) を用いて式 (11) の最適化問題を解いてパラメータ q_s の値を定める．式 (17) 中の離散変換によりこの式は微分できないので，この問題は数値最適化手法により最適化する．実験では，SciPy ライブラリ [15] の Brent 法により最適化した．訓練データそれぞれについて式 (16) の左辺を $O(N)$ 時間で計算したあと，q_s は $O(N\log N)$ 時間で最適化できる．よって，AFFNB 法の全体の計算量は $O(N\log N)$ となる．一方，CV2NB 法の場合では，図 2 の 15 行の $disc$ を計算するために，訓練データ全体を分類し直す必要があるため，後処理アルゴリズムの各反復には $O(N)$ の時間が必要になる．よって，CV2NB 法の反復数が $O(\log N)$ より多ければ，CV2NB よりも AFFNB 法の方が高速になる．我々の実装では，AFFNB 法は CV2NB 法よりかなり高速であった．

5.2　実験結果

この新しい AFFNB 法を，HFFNB 法や CV2NB 法と比較する．実験条件は 3.2 節で述べたものと同じである．実験結果を表 2 に示す．AFFNB 法の性能は，HFFNB 法と比べて格段に改善された．さらに，AFFNB 法は，予測精度と公正性の両面において CV2NB 法と同等の性能を示した．これは，CV2NB 法が仮説分布上ではなく，AFFNB 法と同様に，実際の分布上で公正分解するように設計されていることを示唆している．よって，CV2NB 法と AFFNB 法は，4. 節で述べたモデルバイアスと決定則の影響を受けない．以上のことから，AFFNB モデルは，CV2NB 法で生成される統計モデルを模擬的に表したものとみなせる．

加えて，AFFNB 法には CV2NB 法にはない有用な性質がある．2.3.1 節で述べたように CV2NB 法はクラス分布を保存しないことがあるが，この条件を明示的に強制する AFFNB 法ではクラス分布は保存される．この性質は，入学試験などの場合には，入学者数は公正な決定をしても変化しないため有用である．

6.　生成モデル以外の分類器への拡張

最後に，実公正分解の手法をより広範囲に適用できるように拡張する．分類器は 3 種類の型に分類できる [16, section 1.5.4]：生成モデル，識別モデル，そして識別関数．しかし，実公正分解の手法は生成モデルによる分類器にしか適用できないので，これを他の 2 種類の型の分類器にも適用できるように拡張する．なお，2.3.2 節の ROC も広い範囲に適用できる手法だが，識別関数の分類器には適用できない．

分類器の決定は，識別関数 $f(\mathbf{x})$ の符号に依存ずる．ロジスティック回帰のような識別モデルでは，クラスの事後確率を直接的に表現し，その予測クラスを次式の識別関数の符号に基づいて選択する：

$$f(\mathbf{x})=\hat{\Pr}[Y{=}1|\mathbf{X}=\mathbf{x}]-\hat{\Pr}[Y{=}0|\mathbf{X}=\mathbf{x}] \tag{18}$$

他に，サポートベクトルマシンのような，各入力値をクラスラベルに直接的に写像する識別関数による分類器がある．この分類器も識別関数 $f(\mathbf{x})$ の符号に基づいて，その予測クラスを選択する．

では，各データの予測クラスを，対応する関数 $f(\mathbf{x})$ に基づいて選択する二つの型の分類器に実公正分解を適用する．まず，訓練データをそのセンシティブ特徴の値に基づいて二つに分割し，各データ集合から二つの決定関数 $f_s(\mathbf{x})$, $s\in\{0,1\}$ を学習する．そして，バイアスパラメータ b_s, $s\in\{0,1\}$ を導入する．分割した訓練集合 $D[S{=}s]$, $s\in\{0,1\}$ それぞれについて，次式の公正識別関数によって正クラスに分類される事例の割合が，全体の訓練集合 D 中の正事例の数の比と等しくなるように，バイアスパラメータ b_s の値を決定する．

$$f_s^{\dagger}(\mathbf{x})=f_s(\mathbf{x})+b_s,\ \text{for}\ s\in\{0,1\} \tag{19}$$

実際のクラスとセンシティブ特徴を無関係にすることがこの手続きの目的であり，前節の式 (10) の条件を満たすことに対応する．

ここで，この枠組みは前節で述べた生成モデルに基づく分類器にも対応できることを述べておきたい．不等式 (12) の両辺の対数をとったあと，右辺から左辺を引くと次の公正識別関数が得られる：

表 3　実公正分解を適用した線形 SVM とロジスティック回帰の正解率と公正性指標

Table 3　The accuracy and fairness indeces of a linear SVM and logistic regression with an actual fair-factorization technique

	Adult data			Dutch data		
Methods	Acc	CVS	NMI	Acc	CVS	NMI
AFFLR	0.833	0.002	2.80×10^{-6}	0.774	−0.001	6.65×10^{-7}
LRns	0.863	0.163	4.29×10^{-2}	0.819	0.171	2.20×10^{-2}
AFFSVM	0.833	0.002	2.80×10^{-6}	0.774	−0.001	4.19×10^{-7}
SVMns	0.863	0.163	4.29×10^{-2}	0.818	0.158	1.89×10^{-2}
AFFNB	0.828	−0.002	5.43×10^{-6}	0.761	−0.002	2.68×10^{-6}
CV2NB	0.828	−0.003	6.89×10^{-6}	0.761	−0.003	8.79×10^{-6}

$$f_s^\dagger(\mathbf{x}) = \left[\log\frac{\hat{\Pr}^\dagger[S{=}s]\,\hat{\Pr}^\dagger[\mathbf{X}{=}\mathbf{x}|Y^*{=}1, S{=}s]}{\hat{\Pr}^\dagger[\mathbf{X}{=}\mathbf{x}, S{=}s]} - \log\frac{\hat{\Pr}^\dagger[S{=}s]\,\hat{\Pr}^\dagger[\mathbf{X}{=}\mathbf{x}|Y^*{=}0, S{=}s]}{\hat{\Pr}^\dagger[\mathbf{X}{=}\mathbf{x}, S{=}s]}\right] + \Big[\log q_s - \log(1-q_s)\Big]$$

カギ括弧内の第 1 項と第 2 項は，それぞれ式 (19) の $f_s(\mathbf{x})$ と b_s に対応していることが分かる．

上記の拡張した実公正分解を，識別モデルのロジスティック回帰と，識別関数の線形 SVM でテストした．実験条件は 3.2 節と同じである．ロジスティック回帰と SVM は `scikit-learn` [17] の実装を用いた．実験結果を表 3 に示す．LRns と SVMns と記した行には，それぞれ非センシティブ特徴のみを用いてロジスティック回帰と線形 SVM を適用した結果を示した．AFFLR と AFFSVM と記した行には，それぞれ実公正分解をロジスティック回帰と線形 SVM に適用した結果を示した．

LRns と AFFLR を比較すると，予測精度を犠牲にすることで，公正性を劇的に改善している．同様の現象が SVMns と AFFSVM との間にも見られる．これらのことから，拡張した実公正分解の手法も，分類の公正性を改善するのに有効であるといえる．

次に，AFFNB 法と，AFFLR 法や AFFSVM 法とを比較する．どの分類器においても，実公正分解を適用することで，ほぼ完全な水準の公正性が達成できている．予測精度に関しては，これらのデータでは AFFLR 法と AFFSVM 法はともに，AFFNB 法より若干よい．拡張公正分解はどの型の分類器にも適用できるので，分類の公正性を保ちつつ最も予測精度のよい分類器を利用者は選んで用いることができる．

7. ま と め

本論文では，最初に，公正配慮型分類器についてまとめ，その中で CV2NB 法が，他の手法よりより高い水準の公正性をなぜ達成できるかを論じた．仮説公正分解を適用した単純ベイズモデルとの比較によって，CV2NB 法が優れた性能を示す原因がモデルバイアスと決定則の影響であることを示した．この知見に基づいて，実公正分解を開発した．これは CV2NB 法で生成されるモデルと模擬的に等価と考えられるモデルである．最後に，この実公正分解を生成モデル分類器以外の，識別モデルや決定関数による分類器にも適用出来るように拡張した．

現在の実公正分解には非常に強い制限があある，すなわち，実ラベルは仮説ラベルとセンシティブ特徴のみに依存し，非センシティブ特徴には依存しないという仮定である．この制限を緩めることができれば，予測精度と公正性の間でよりよいトレードオフを実現できる分類器を開発できるだろう．

謝辞　研究の詳細な情報を提供してくれた Sicco Verwer 氏，およびベンチマークデータを提供している Žliobaitė 氏に感謝する．本研究は JSPS 科研費 16700157，21500154，24500194，25540094 の助成を受けたものである．

文　献

[1] D. Pedreschi, S. Ruggieri, and F. Turini, “Discrimination-aware data mining,” Proc. of the 14th ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining, pp.560–568, 2008.

[2] T. Calders and S. Verwer, “Three naive bayes approaches for discrimination-free classification,” Data Mining and Knowledge Discovery, vol.21, pp.277–292, 2010.

[3] D. Gondek and T. Hofmann, “Non-redundant data clustering,” Proc. of the 4th IEEE Int’l Conf. on Data Mining, pp.75–82, 2004.

[4] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, “Fairness-aware classifier with prejudice remover regularizer,” Proc. of the ECML PKDD 2012, Part II, pp.35–50, 2012. [LNCS 7524].

[5] B. Berendt and S. Preibusch, “Exploring discrimination: A user-centric evaluation of discrimination-aware data mining,” Proc. of the IEEE Int’l Workshop on Discrimination and Privacy-Aware Data Mining, pp.344–351, 2012.

[6] L. Sweeney, “Discrimination in online ad delivery,” Communications of the ACM, vol.56, no.5, pp.44–54, 2013.

[7] K. Fukuchi, J. Sakuma, and T. Kamishima, “Prediction with model-based neutrality,” Proc. of the ECML PKDD 2013, Part II, pp.499–514, 2013. [LNCS 8189].

[8] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, “Fairness through awareness,” Proc. of the 3rd Innovations in Theoretical Computer Science Conf., pp.214–226, 2012.

[9] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, “Learning fair representations,” Proc. of the 30th Int’l Conf. on Machine Learning, pp.●●–●●, 2013.

[10] F. Kamiran, A. Karim, and X. Zhang, “Decision theory for discrimination-aware classification,” Proc. of the 12th IEEE Int’l Conf. on Data Mining, pp.924–929, 2012.

[11] C. Elkan, “The foundations of cost-sensitive learning,” Proc. of the 17th Int’l Joint Conf. on Artificial Intelligence, pp.973–978, 2001.

[12] F. Kamiran, T. Calders, and M. Pechenizkiy, “Discrimination aware decision tree learning,” Proc. of the 10th IEEE Int’l Conf. on Data Mining, pp.869–874, 2010.

[13] I. Žliobaitė, F. Kamiran, and T. Calders, “Handling conditional discrimination,” Proc. of the 11th IEEE Int’l Conf. on Data Mining, pp.●●–●●, 2011.

[14] A. Frank and A. Asuncion, “UCI machine learning repository,” University of California, Irvine, School of Information and Computer Sciences, 2010. ⟨`http://archive.ics.uci.edu/ml`⟩.

[15] E. Jones, T. Oliphant, P. Peterson, et al., “SciPy: Open source scientific tools for Python,” 2000-. ⟨`http://www.scipy.org/`⟩.

[16] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[17] F. Pedregosa, et al., “Scikit-learn: Machine learning in python,” Journal of Machine Learning Research, vol.12, pp.2825–2830, 2011. ⟨`http://scikit-learn.org`⟩.