1D3-4

公正配慮型学習 — 正則化によるアプローチ

Fairness-Aware Learning — A Regularization Approach

神嶌 敏弘　　赤穂 昭太郎
Toshihiro Kamishima　　Shotaro Akaho

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

With the spread of data mining technologies and the accumulation of social data, datamining techniques are being used for determinations that seriously affect people's lives, e.g., credit scoring. Such determinations must be unbiased and nondiscriminatory in sensitive features, such as race, gender, religion, and so on. Several researchers have recently begun to tackle this problem of analysis techniques that are aware of social fairness or discrimination. In this paper, we propose a new regularization approach that is applicable to a wider variety of analysis techniques.

1. はじめに

本論文では，データマイニングにおける社会的公正さや差別性について論じ，公正さの尺度と，より公正な決定がなされるようにロジスティック回帰を修正する方法を提案する．

膨大な個人データが集積され，またデータマイニングが容易に利用できる環境が整備されるに従い，与信，保険，採用などの重要な決定にもデータマイニング技術が利用されるようになり始めている．このとき，社会的・法的な公正さに配慮した，すなわち，人種，性別，信仰などに基づく先入観や差別のない判断がなされなければならない．また，差別以外にも，ある種の特徴・属性の取り扱いに注意すべき要因として，データ提供者との契約がある．例えば，顧客の個人情報を推薦システムで利用する目的で収集したデータは他の目的には利用できない．しかし，このデータを利用した推薦システムで商品を提示すると，提示した商品の履歴データは間接的に影響を受ける．すると，この履歴データの推薦以外の目的への利用には配慮が必要になる場合も考えられる．

社会的公正・差別に配慮した分析技術については研究が始まっている [Calders 10, Pedreschi 08, Pedreschi 09]．これらの研究では，配慮が必要な特徴を排除するだけでは，それらの特徴の間接的な影響を排除できないこと報告されている．例えば，特徴『人種』を利用せずに与信の識別を行っても，ある人種が特定の地区に集まっていると，特徴『住所』を利用することで差別的な決定がなされてしまう．これは red-lining 効果 [Calders 10] や間接差別 (indirect discrimination) [Pedreschi 08] などと呼ばれている．

これまでに，非差別的な決定をするように修正した単純ベイズ [Calders 10] や，差別的な相関ルールを抽出する手法 [Pedreschi 08, Pedreschi 09] が提案されてきた．本研究では，データ分析での公正さの定式化，公正さを高める新たなデータ分析手法の開発，および分析手法の有効性・効率性の評価手法の提案の三つの点について述べる．第一のデータ分析における公正さについては，既存研究でも論じられてきた．Pedreschi らは法令に基づいて，特定の集団が与信を受けられる確率の，全体での確率に対する比などを考慮している [Pedreschi 08]．本研究では，こうした不公正な決定の要因として，先入観，過小評価，および負の遺産の 3 点について論じる．ここでいう先入観とは配慮の必要な特徴の直接的・間接的な決定への影響，過小評価とはデータの経験分布と予測との乖離，負の遺産とは過去の差別的な決定の影響のことである．これらの要因を評価するための尺度も提案する．

第二に，先入観の要因に注目し，この先入観を減らす機械学習手法を開発する．この手法は，学習器に対する正則化項として定式化されているため，正則化項を利用できる多様な分析手法に適用可能である．提案手法には prejudice remover と unfairness hater があり，前者は配慮の必要な特徴と決定との間の独立性を高め，後者は最も不公正な分類器から乖離しようとする．これらの正則化項は簡潔さと実行効率にも配慮している．この正則化項をロジスティック回帰に導入した．

最後に，提案手法の有効性と効率を，2 単純ベイズ法 [Calders 10] との比較実験により検証する．

既存研究では，差別配慮型 (discrimination-aware) と呼ばれているが，ここでは公正配慮型学習 (fairness-aware learning) と呼ぶ理由について述べておく．差別解消以外にも，データ提供者との契約に配慮するなど他の目的にもこの技術適用できるからであり，英語の discrimination という語は，機械学習の文脈では『判別』の意味で使われ誤解が生じやすいためでもある．

以下，2. 節では公正さの概念について，3. 節では公正さを強化する手法について述べる．4. 節は 2 単純ベイズ法との比較実験であり，5. 節はまとめである．

連絡先: http://www.kamishima.net/

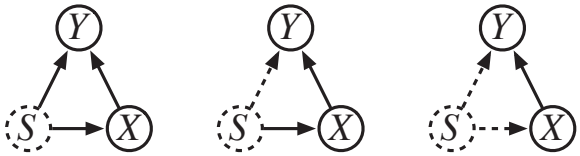
2. マイニングにおける公正さ

公正配慮型の困難さの例を紹介したのち，不公正さの要因とこれらを定量評価する尺度について述べる．

2.1 公正配慮型学習の難しさ

ここでは，公正配慮型学習の難しさを示した文献 [Calders 10] の例を紹介する．収入が 5 万ドルかを識別する問題を扱う．識別について配慮が必要な特徴 S は性別を示し，Male と Female の値をとる．他に，High か Low の値をとり，収入を示す目的変数 Y と，特に配慮を必要としないその他の特徴 X とがある．データ集合中では High–Male の同時頻度は High–Female のそれの 5.5 倍である．Male データの約 30%は High クラスだが，Female データでは 11% だけなので，Female データは Low に分類されやすく不公正が生じやすい．

差別の度合いを測る尺度（ここでは著者名にちなみ CV スコアと呼ぶ）として，特徴 S が配慮を要しない値のときに正になる条件付き確率から，配慮を要するときに正になる確率を



(a) 直接先入観　(b) 間接先入観　(c) 潜在先入観

図 1: 3 種類の先入観の削除

引いた値を用いる：

$$\Pr[Y{=}\mathsf{High}|S{=}\mathsf{Male}] - \Pr[Y{=}\mathsf{High}|S{=}\mathsf{Female}].$$

元のデータでのデータの頻度から計算すると CV スコアは 0.19 だが，特徴 S も用いた単純ベイズで予測したクラスについて CV スコアを求めると CV スコアは 0.34 まで増加する．これは少数派の Female–High が不公正に扱われていることを示唆する．この現象は多くの識別器が採用するオッカムの剃刀原理により主に生じる．すなわち，希で特殊なパターンは，観測データの一般化の過程で棄却されることが多く，無視されて不公正になりやすい．さらに，特徴 S を除いた単純ベイズで予測した結果でも CVS は 0.28 と，S と相関する他の特徴の影響により不公正なままである．この現象は red-lining effect [Calders 10] や間接差別 (indirect discrimination) [Pedreschi 08] などと呼ばれる．よって，特徴 S を単に削除するだけでは不十分で，積極的差別是正策 (affirmative action) が必要となる．

2.2 マイニングにおける不公正の 3 要因

マイニングでの不公正さの要因として，ここでは **先入観**，**過小評価**，および **負の遺産** の 3 点について考察する．まず，前提として教師あり学習で，目的変数 Y の値を特徴量から予測する識別器を獲得する問題を扱う．特徴には，社会的要因で配慮が必要な **要配慮特徴 (sensitive feature)** S と，そうではない **配慮不要特徴 (non-sensitive feature)** X とに分けられる．なお，これらの変数が離散である場合について述べるが，和を積分に置き換えれば連続の場合にも拡張可能である．

2.2.1 先入観 (Prejudice)

ここでの先入観とは要配慮特徴が決定に影響することであり，さらに図 1 のように三つに分類する．の直接先入観 (direct prejudice) とは決定が明示的に要配慮特徴に基づいて行われることで，文献 [Pedreschi 08] の直接差別の概念に対応する．図 1(a) のように要配慮特徴を削除すれば，明示的に影響しないようにできるため，直接先入観は排除できる．

2.1 節で述べたように要配慮特徴を削除しても，red-lining 効果が残る．この効果を除くため，直接的な効果と共に間接的な効果である間接先入観を扱う必要がある．間接先入観は，図 1(b) のように，目的変数 Y を要配慮変数 S から独立にして，S の値が Y の決定に無関係になるようにして削除する．この考えに基づき，間接先入観を変数 Y と S 間の依存性と定め，ここでは Y と S の間の相互情報量で定量化する．

$$\mathrm{PI} \equiv \mathrm{I}(Y;\, S) = \sum_{Y,S} \Pr[Y,S] \ln \frac{\Pr[Y,S]}{\Pr[Y]\Pr[S]} \qquad (1)$$

相互情報量の値が大きいと依存性は強くなるので，この量を（間接）先入観尺度 (prejudice index; PI) と呼ぶ．さらに，$[0,1]$ の範囲に対称なまま正規化した次の尺度を正規化先入観尺度 (normalized prejudice index; NPI) と定義する．

$$\mathrm{NPI} \equiv \frac{\mathrm{I}(Y;\, S)}{\sqrt{\mathrm{H}(Y)\mathrm{H}(S)}} \qquad (2)$$

ただし，$\mathrm{H}(X)$ はエントロピー関数である．

目的変数 Y と要配慮特徴 S とを独立にしても，配慮不要特徴 X はまだ S に依存している可能性がある．この依存性は結果には影響してはいないが，情報の利用自体を制限するようなルールや法には反しているとも考えられる．そこで，この依存性を潜在先入観 (potential prejudice) と呼び，X と S の間の相互情報量によって定量化する．この先入観は，図 1(c) のように，X と S の間も独立にすれば削除できる．

2.2.2 過小評価 (Underestimation)

過小評価とは，2.1 節の例のように，配慮すべき値をもつ事例が，データから経験的に得られる分布よりも不利な決定をされる場合とする．間接先入観を削除していて，分類器に一致性があり，無限個のデータから学習できれば，この問題は生じない．しかし，有限個のデータでは，データの経験分布と，分類器が学習した分布には乖離が生じてしまう．極限における一致の考えは数学的には妥当だろうが，十分な事例がない状態で不利な扱いを受ける人にとっては社会的には妥当といえないだろう．また，こうした乖離から，意図的に不利な扱いをしている疑念をもたれる場合もあるだろう．そこで，この過小評価の度合いは，データの経験分布と，モデルによる分布との距離で差で測る．形式的には，過小評価尺度 (underestimation index; UEI) を，Y と S の同時分布に関するサンプル分布とモデル分布の差の Hellinger 距離で定義する．

$$\mathrm{UEI} = \left(\frac{1}{2} \sum_{Y,S} \left(\sqrt{\tilde{\Pr}[Y,S]} - \sqrt{\hat{\Pr}[Y,S]} \right)^2 \right)^{1/2} \qquad (3)$$

ただし，$\tilde{\Pr}[\cdot]$ と $\hat{\Pr}[\cdot]$ は，それぞれサンプルとモデル分布の確率質量関数である．UEI の値域は $[0,1]$ である．

2.2.3 負の遺産 (Negative Legacy)

負の遺産とは，訓練データにおいて不公平なラベル付けやアノテーションが行われている問題をさす．この問題は，文献 [Calders 10] の潜在変数モデルの部分で暗黙的に示されている．ラベル付けが不公平であるかどうかはラベルの値だけからは判別できないので，この問題の修正には，他の情報が必要だろう．例えば，小規模でよいので公正なラベル付けがなされたデータ集合や，不公正なラベル付けがなされている事例の割合などの情報である．この問題は転移学習 [神嶌 10]，特に標本選択バイアス [Zadrozny 04] と密接に関連すると考える．

潜在先入観の排除を要求するルールや法律は希だろう．実験では UEI 尺度の評価は行うが，過小評価の問題を積極的に解消することは本論文では扱わない．負の遺産については，上記のように追加情報がなければ解決は難しいだろう．よって，以後は間接先入観の削除に注力する．

3. 先入観削除手法

ここでは prejudice remover と unfairness hater の二種類の間接先入観の削除手法を提案する．これらは学習器に制約を与える正則化項として定式化されているので，正則化が利用できる多様な方法で利用可能である．

3.1 全般的な枠組み

まず全般的な枠組を，ロジスティック回帰による分類の場合について述べる．Y，X，および S はそれぞれ，クラス，配慮不要特徴，要配慮特徴に相当する確率変数である．X と S が離散の場合について記すが，和を積分に置き換えれば連続の場合でも同様の議論が成立する．訓練データは，これらの

実現値の組の集合 $\mathcal{D} = \{(y, \mathbf{x}, \mathbf{s})\}$ である．$\mathbf{\Theta}$ をパラメータとして，特徴が与えられたときのクラスの条件確率のモデルを $\Pr[Y|X, S; \mathbf{\Theta}]$ とする．データ $\mathcal{D}$ に対する対数尤度 $\ell(\mathcal{D}; \mathbf{\Theta})$ を用いた損失関数を最適化することでパラメータ $\mathbf{\Theta}$ を決める．

ここでは 2 種類の正則化項を用いる．一つは過学習を避けるための通常の正則化項で，標準的な L2 正則化 $\|\mathbf{\Theta}\|_2^2$ を用いる．もう一つの正則化項 R は，より公正な識別がなされるようにするもので，詳細は後に述べる．負の対数尤度関数にこれらの正則化項を加えた次式が最小化すべき目的関数となる．

$$-\ell(\mathcal{D}; \mathbf{\Theta}) + \eta \mathrm{R}(\mathcal{D}, \mathbf{\Theta}) + \frac{\lambda}{2} \|\mathbf{\Theta}\|_2{}^2, \tag{4}$$

ただし，λ と η は正実数の正則化パラメータである．

以下は，ロジスティック回帰の場合について述べる．目的変数 Y は二値 $\{0, 1\}$ に限定し，配慮不要特徴 X と要配慮特徴 S はそれぞれ実ベクトル $\mathbf{x}$ と $\mathbf{s}$ をとる．$\mathbf{x}$ と $\mathbf{s}$ に対する重みベクトル $\mathbf{w}$ と $\mathbf{v}$ がパラメータ $\mathbf{\Theta}$ に相当する．条件付き確率 $\Pr[Y|X, S; \mathbf{\Theta}]$ の対数を，$\mathrm{L}(y, [\mathbf{x}, \mathbf{s}]; [\mathbf{w}, \mathbf{v}])$ でモデル化する．ただし，$[\mathbf{a}, \mathbf{b}]$ は二つの縦ベクトルを縦に連結したもので，関数 $\mathrm{L}(\cdot)$ は，シグモイド関数 $\sigma(\cdot)$ を用いて次式で定義される．

$$\mathrm{L}(y, \mathbf{x}; \boldsymbol{\theta}) = y\sigma(\boldsymbol{\theta}^\top \mathbf{x}) + (1 - y)(1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x})) \tag{5}$$

定数項を特徴ベクトル $\mathbf{x}$ に含めておくものとする．対数尤度は $\ell(\mathcal{D}, \mathbf{\Theta}) = \sum_{\mathcal{D}} \mathrm{L}(y, [\mathbf{x}, \mathbf{s}]; [\mathbf{w}, \mathbf{v}])$ となる．それでは，R に相当する 2 種類の正則化項の説明に移る．

3.2 Prejudice Remover

一つ目の正則化項 **prejudice remover** (PR と略記) $\mathrm{R_{PR}}$ は，先入観尺度 PI の直接的な最小化を試みる．Y と S の相互情報量 $\mathrm{I}(Y; S)$ である PI の計算には次の Y と S の同時分布の計算が必要となる．

$$\Pr[Y, S] = \sum_X \Pr[X, S] \Pr[Y|X, S; \mathbf{\Theta}]$$

この同時分布は，$\Pr[X, S]$ を $\mathcal{D}$ 上の経験分布で置き換え MCMC などのサンプリング手法を使えば厳密な計算が可能である．しかし，X や S の値域によっては計算量が多いので，以下の簡潔で高速なアプローチを採用する．

ここで，$\mathbf{\Theta}$ は，X と S に関連する部分それぞれ分割できるものと仮定する．この仮定は，ロジスティック回帰を含む一般化線形モデルでは，X と S のそれぞれについての重みとして分離できるため成立する．式 (1) の先入観尺度は次式となる．

$$\mathrm{PI} = \sum_{Y,X,S} \Pr[Y|X, S] \Pr[X, S] \ln \frac{\Pr[Y, S]}{\Pr[S] \Pr[Y]}$$

$\Pr[Y|X, S]$ をモデル $\Pr[Y|X, S; \mathbf{\Theta}]$ で，$\sum_{X,S} \Pr[X, S]$ を標本上の和で置き換えると次式を得る．

$$\sum_{(\mathbf{x},\mathbf{s})\in\mathcal{D}} \sum_{Y\in\{0,1\}} \Pr[Y|\mathbf{x}, \mathbf{s}; \mathbf{\Theta}] \ln \frac{\Pr[Y, \mathbf{s}]}{\Pr[\mathbf{s}] \Pr[Y]}$$

この式の対数の中身を，$\Pr[Y|\mathbf{s}]/\Pr[Y]$ と書き換えて求める．$\Pr[Y|S]$ と $\Pr[Y]$ は $\Pr[Y|X, S; \mathbf{\Theta}] \Pr[X, S]$ の周辺化で計算できるが，上記のようにその計算コストは大きい．そこで，次の単純なモデルを用いる．

$$\Pr[y|\mathbf{s}; \mathbf{\Theta}] = \mathrm{L}(y, [\bar{\mathbf{x}}(\mathbf{s}), \mathbf{s}]; [\mathbf{w}, \mathbf{v}])$$
$$\Pr[y; \mathbf{\Theta}] = \mathrm{L}(y, [\bar{\mathbf{x}}, \bar{\mathbf{s}}]; [\mathbf{w}, \mathbf{v}])$$

ただし，$\bar{\mathbf{x}}(\mathbf{s})$ は，$\mathcal{D}$ 中で，その要配慮特徴 S が $\mathbf{s}$ であるデータの配慮不要特徴の平均値であり，$\bar{\mathbf{x}}$ と $\bar{\mathbf{s}}$ はそれぞれ X と S の標本平均である．すなわち，周辺化する代わりに，平均特徴ベクトルのときの確率を利用する．すると，次式を得る．

$$\mathrm{R_{PR}}(\mathcal{D}, \mathbf{\Theta}) = \sum_{(\mathbf{x},\mathbf{s})} \sum_{Y} \Pr[y|\mathbf{x}, \mathbf{s}; \mathbf{\Theta}] \ln \frac{\Pr[y|\bar{\mathbf{x}}(\mathbf{s}), \mathbf{s}; \mathbf{\Theta}]}{\Pr[y|\bar{\mathbf{x}}, \bar{\mathbf{s}}; \mathbf{\Theta}]} \tag{6}$$

この正則化項は，要配慮特徴により強く依存してクラスの値が決定される場合に大きくなるので，要配慮特徴が最終結果に与える影響を小さくする効果がある．ロジスティック回帰に導入すると，最小化すべき目的関数は次式となる．

$$-\Big(\sum_{(y,\mathbf{x},\mathbf{s})} \mathrm{L}(y, [\mathbf{x}, \mathbf{s}]; [\mathbf{w}, \mathbf{v}])\Big) + \eta \mathrm{R_{PR}}(\mathcal{D}, \mathbf{\Theta}) + \frac{\lambda}{2} \|[\mathbf{w}, \mathbf{v}]\|_2{}^2 \tag{7}$$

prejudice remover は，目標である先入観尺度を直接的に最小化する自然なものである．しかし，この正則化項はパラメータ $\mathbf{\Theta}$ を X と S に対応する部分に分割できない場合には利用できないとか，計算コスト削減のために計算が近似である短所がある．そこで，もう一つの正則化項も開発した．

3.3 Unfairness Hater

二つ目の正則化項は，最も不公正な予測器から乖離するようにするヒューリスティックに基づく．不公正な予測器とは要配慮な特徴のみを参照して決定を行うもので，形式的には $\Pr[Y|S; \mathbf{\Psi}]$（$\mathbf{\Psi}$ はパラメータ）でモデル化される．不公正予測器は事前に $\mathcal{D}$ 上の最尤推定でパラメータを求めておき，そのときのパラメータを $\mathbf{\Psi}^*$ と記す．正則化項は，不公正予測器 $\Pr[Y|S; \mathbf{\Psi}^*]$ から，目標モデル $\Pr[Y|X, S; \mathbf{\Theta}]$ までの KL ダイバージェンスを大きくするように設計する．このダイバージェンスは次式である．

$$\sum_{(\mathbf{x},\mathbf{s})\in\mathcal{D}} \sum_{Y} \Pr[y|\mathbf{s}; \mathbf{\Psi}^*] \Big(\ln \Pr[y|\mathbf{s}; \mathbf{\Psi}^*] - \ln \Pr[y|\mathbf{x}, \mathbf{s}; \mathbf{\Theta}] \Big)$$

第 1 項は $\mathbf{\Theta}$ の最小化に関して定数なので無視できる．最終目的関数を最小化するので，このダイバージェンスの負をとると，次の **unfairness hater**（UH と略記）を得る．

$$\mathrm{R_{UH}}(\mathcal{D}, \mathbf{\Theta}) = \sum_{(\mathbf{x},\mathbf{s})\in\mathcal{D}} \sum_{Y} \Pr[y|\mathbf{s}; \mathbf{\Psi}^*] \ln \Pr[y|\mathbf{x}, \mathbf{s}; \mathbf{\Theta}] \tag{8}$$

この正則化項は，不公正予測器が明確で強い決定をするような訓練事例の重みを小さくするような働きをする．ロジスティック回帰に導入した場合に，最小化すべき目的関数は次式となる．

$$-\Big(\sum_{(y,\mathbf{x},\mathbf{s})} \mathrm{L}(y, [\mathbf{x}, \mathbf{s}]; [\mathbf{w}, \mathbf{v}])\Big) + \eta \mathrm{R_{UH}}(\mathcal{D}, \mathbf{\Theta}) + \frac{\lambda}{2} \|[\mathbf{w}, \mathbf{v}]\|_2{}^2 \tag{9}$$

4. 比較実験

Calders と Verwer の 2 単純ベイズ法との比較実験を行う．

4.1 手法と実験データ

文献 [Calders 10] で良い性能を示した Calders と Verwer の 2 単純ベイズ法 (two-naive-Bayes method)（CV2NB と略す）を簡単に紹介する．この方法は，二値分類問題で，1 個の二値の要配慮特徴を想定している．この手法の生成モデルは次式．

$$\Pr[Y, \mathbf{X}, S] = \Pr[S] \Pr[Y|S] \prod_i \Pr[X_i|Y, S]. \tag{10}$$

1 Calculate a CV score, $disc$, of the predicted classes by the current model.
2 `while` $disc > 0$
3 $numpos$ is the number of positive samples classified by the current model.
4 `if` $numpos <$ the number of positive samples in $\mathcal{D}$ `then`
5 $N(Y{=}1, S{=}0) \leftarrow N(Y{=}1, S{=}0) + \Delta N(Y{=}0, S{=}1)$
6 $N(Y{=}0, S{=}0) \leftarrow N(Y{=}0, S{=}0) - \Delta N(Y{=}0, S{=}1)$
7 `else`
8 $N(Y{=}0, S{=}1) \leftarrow N(Y{=}0, S{=}1) + \Delta N(Y{=}1, S{=}0)$
9 $N(Y{=}1, S{=}1) \leftarrow N(Y{=}1, S{=}1) - \Delta N(Y{=}1, S{=}0)$
10 Recalculate $\Pr[Y|S]$ and a CV score, $disc$ based on updated $N(Y, S)$

図 2: naive Bayes modification algorithm

注釈：$N(Y{=}y, S{=}s)$ は訓練事例集合 $\mathcal{D}$ 中でクラスが y で，要配慮特徴が s である事例の数．実験では Δ は原論文の 0.01 を採用．

表 1: 実験結果

method	Acc	NMI	NPI	UEI	CVS	PI / MI
LR	0.850	0.265	5.31E-02	0.0447	0.188	2.16E-01
LRns	0.850	0.264	4.50E-04	0.0407	-0.031	9.66E-04
PR η=0.01	0.851	0.267	5.08E-02	0.0398	0.187	2.06E-01
PR η=1	0.843	0.243	1.78E-02	0.0548	0.110	7.88E-02
PR η=2	0.756	0.149	1.84E-01	0.1230	0.444	1.33E+00
PR η=10	0.588	0.037	6.02E-02	0.2697	-0.297	1.77E+00
UH η=0.01	0.851	0.267	5.16E-02	0.0375	0.190	2.08E-01
UH η=1	0.696	0.205	1.11E-01	0.2116	0.400	5.84E-01
NB	0.822	0.246	1.12E-01	0.0679	0.332	4.90E-01
NBns	0.826	0.249	7.17E-02	0.0427	0.267	3.11E-01
CV2NB	0.813	0.191	3.64E-06	0.0819	-0.002	2.05E-05

このモデルでは，不公正な決定がなされる可能性を，目的変数 Y が要配慮特徴 S に依存させることでモデル化している．通常の単純ベイズでは，Y が与えられたときに特徴は互いに条件付き独立だが，CV2NB では Y と S が与えられたときに条件付き独立になっている．S の二値それぞれについて通常の単純ベイズが学習されているとも解釈できるため 2 単純ベイズと呼ぶ．学習は訓練事例中のデータの頻度を求めればよい．分布 $\Pr[S]\Pr[Y|S]$ を，CV スコアが正になるよう図 2 のアルゴリズムで修正し，先入観を削減する．この CV2NB を，提案手法 PR や UH と比較する．参考のため全特徴を用いたロジスティック回帰と単純ベイズ LR と NB，及びこれらの要配慮特徴を用いないバージョン LRns と NBns の結果も示す．

実験データは文献 [Calders 10] で用いられたもので，2.1 節の例のデータである．このデータでは，単に要配慮特徴を削除しても間接先入観を排除できない．UCI Repository [Frank 10] の Adult/Census Income の 16281 個データを含む．目的変数は収入が 5 万ドル以上かどうかの二値であり，要配慮特徴は性別で二値ある．配慮不要特徴は 13 個あり，文献の手続きにより離散化した．実験では，単純ベイズは確率が 0 にならないように Dirichlet 事前分布を採用した．ロジスティック回帰では 0/1 ベクトルで離散特徴を扱い，影響の少なかった L2 正則化のパラメータ λ は 0.1 に固定した．5 分割交差確認で評価値を得た．

4.2 実験結果

表 1 に，正解率 Acc，2. 節の NPI，UEI，CV スコアを示す．MI は，予測ラベルと標本ラベルの間の相互情報量で，予測ラベルから真のラベルについて知ることのできる情報量の推定値である．この情報量の単位あたりに犠牲にする先入観の量を PI / MI として求めた．これは，予測精度の向上と先入観の削除とのトレードオフの評価となり，小さいほど効率が良い．

まず提案手法のパラメータ η の影響について述べる．この η は，$\mathrm{R_{PR}}$ や $\mathrm{R_{UH}}$ の係数であり，大きいほど予測精度よりも先入観の削除が重視される．どちらも，η がある程度までは単調に先入観の削除が重視されているが，ある値以上になると単調には減らないかった．数値計算の不安定性などの理由は考えられるが，詳細な原因は調査中である．PR と UH を比較すると，直接的に先入観を削除する PR が良い結果を得た．特に，UH で $\eta > 1$ の場合に，特徴の値にかかわらず正か負のいずれかだけに分類するような分類器が得られてしまった．これは，KL ダイバージェンスは有界でないため，$\mathrm{R_{UH}}$ の影響がある程度以上になると，精度に関する尤度項が完全に無視されてしまうためである．先入観を削除しない分類器と比較すると，LR, NB，NBns に対しては，先入観削除効果が見られた．しかし，LRns はほとんど red-lining 効果がないため，提案手法の効果は見られなかった．まとめると，いずの提案手法も，先入観の削除効果はあるが，PR の方が性能面ですぐれ，また調整も容易であった．

η=1 での PR と CV2NB とを比較すると，生成モデルより識別モデルの利点により精度面では PR が良かったが，PI / MI の先入観の削除効率では CV2NB が優れていた．

5. まとめ

本論文の寄与は以下の通りである：第一に，マイニングにおける不公正の概念の要因として先入観，過小評価，負の遺産の三つを提案し，その定量化手法などについて論じた．第二に，先入観を削減するための手法として，prejudice remover と unfairness hater の 2 種類の正則化項を開発した．最後に，既存手法との比較実験により，提案手法の有効性や効率を検証した．公正配慮型学習研究は初期の段階でまだ多くの課題があり，公正さの概念についての議論や，公正さを保つ分析手法の開発を今後も進める必要がある．

データマイニング技術の社会での利用は日々拡大しているが，残念なことに，しばしば人々の生活に悪影響を及ぼす場合もある [Boyd 10]．一方で，データ分析技術は，社会資本の効率的な運用や，各種のリスク管理に欠かせないものとなり，その重要性は増してゆくだろう．今後は，人々の生活への影響に配慮した，プライバシ保護マイニングや公正・差別配慮マイニングのような社会責任的マイニング (socially responsible mining) の概念が重要になるであろう．

謝辞: 実験用ソフトを提供いただいた Sicco Verwer 先生に感謝する．本研究は科研費 21500154 の助成を受けた．

参考文献

[Boyd 10] Boyd, D.: Privacy and Publicity in the Context of Big Data, in *Keynote Talk of The 19th Int'l Conf. on World Wide Web* (2010)

[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, `http://archive.ics.uci.edu/ml` (2010), University of California, Irvine, School of Information and Computer Sciences

[神嶌 10] 神嶌 敏弘：転移学習, 人工知能学会誌, Vol. 25, No. 4, pp. 572–580 (2010)

[Pedreschi 08] Pedreschi, D., Ruggieri, S., and Turini, F.: Discrimination-aware Data Mining, in *Proc. of The 14th Int'l Conf. on Knowledge Discovery and Data Mining* (2008)

[Pedreschi 09] Pedreschi, D., Ruggieri, S., and Turini, F.: Measuring Discrimination in Socially-Sensitive Decision Records, in *Proc. of the SIAM Int'l Conf. on Data Mining* (2009)

[Zadrozny 04] Zadrozny, B.: Learning and Evaluating Classifiers under Sample Selection Bias, in *Proc. of The 21st Int'l Conf. on Machine Learning*, pp. 903–910 (2004)