公正配慮型学習
正則化によるアプローチ

神嶌 敏弘，赤穂 昭太郎
産業技術総合研究所
http://www.kamishima.net/

公正/差別配慮型学習

データマイニングが個人の生活にも影響する決定に利用される
与信，採用，昇進…

公正 / 差別配慮型学習

Fairness / Discrimination-Aware Learning

人種・性別など社会的公正さの観点から望ましくない要因が
決定に関係しないように配慮してデータ学習・分析を行う

本発表の寄与

- マイニングにおける公正さの要因について考察
- 公正さに配慮した分類ができるように既存の分類手法を改良
- 実験で有効性を確認

マイニングにおける不公正

[Calders 10]

アメリカの国勢調査データ：年収が5万ドル以上かどうかを識別

	男性	女性
高収入	3,256	590
低収入	7,604	4,831

少ない

高収入クラスで女性は少数派

- 高収入の男性は，高収入の女性の5.5倍
- 男性の30%は高収入だが，女性は11%だけ

オッカムの剃刀：単純な仮説がよい

↓

低頻度のパターンは無視されやすく，少数派は不利に扱われやすい

CVスコア

[Calders 10]

CVスコア (Calders-Verwer discrimination score)

Pr[Y=高収入 | S=男性] - Pr[Y=高収入 | S=女性]

目的変数 Y が有利な値になる条件付き確率の
要配慮特徴 S が多数派の場合から，少数派の場合の値を引いた値

- 元の正解データについて計算すると
  - ➡ 0.19 となりこれをベースラインに
- 要配慮特徴 S と配慮不要特徴 X の両方を使って単純ベイズで分類
  - ➡ 0.34 まで増加してしまい，不公正な扱いがみられる
- 要配慮特徴 S を使わなくても
  - ➡ 0.28 に改善されるが，ベースラインよりと比べると不公正

要配慮特徴は無視するだけでは不十分で，積極的是正策が必要

マイニングにおける不公正の要因

- **先入観 (Prejudice)**

  要配慮特徴が決定に影響すること

- **過小評価 (Underestimation)**

  配慮すべき値を持つ事例が，経験分布より不利に扱われる

- **負の遺産 (Negative Legacy)**

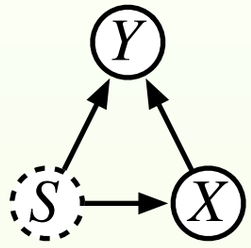
  決定を行う学習器の訓練データ自体に，不公正な決定に基づく教示データが含まれている

先入観 (Prejudice)

先入観：要配慮特徴が決定に影響すること

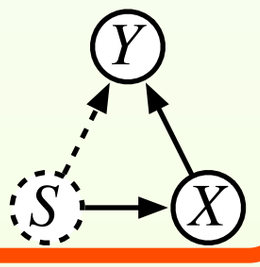
直接先入観 (Direct Prejudice)

- 要配慮特徴が直接的な決定への影響
- 要配慮特徴を取り除いてモデル化して解消



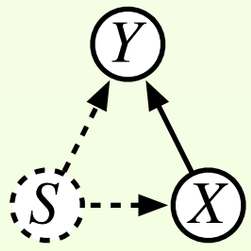
間接先入観 (Indirect Prejudice)

- 要配慮特徴の直接的・間接的な決定への影響
- 要配慮特徴と目的変数が独立になるようにして解消



潜在先入観 (Potential Prejudice)

- 要配慮特徴の配慮不要特も含めた影響
- 要配慮特徴と目的変数・配慮不要特徴も独立にして解消



過小評価 (Underestimation)

過小評価：配慮すべき値を持つ事例が，経験分布より不利に扱われる

データが無限にあれば：経験分布と学習結果は近づく

↓

データ数が十分でなければ：経験分布と学習結果はずれることも

↓

- 漸近的な一致性という概念は数学的には合理的といえる
- データが有限だからということは，少数派に不利な決定をしてしまうことの，社会的に合理的な理由とはいえないだろう
- データとの不一致は，たとえ分析が意図的なものでないとしても，少数派にとっての心証は悪い

プランニングのAnytimeアルゴリズムの考えが利用できる可能性

負の遺産 (Negative Legacy)

負の遺産：決定を行う学習器の訓練データ自体に，不公正な決定に基づく教示データが含まれている

訓練データ自体が不公正なのでそれを補正するための情報が必要

公正な決定だけを含んでいることが判明しているデータが
一部でもあれば…

転移学習，特に標本選択バイアスの問題と関連

Logistic回帰

Logistic回帰：代表的な識別モデルによる分類手法

↑

間接先入観を削除：要配慮特徴と目的変数を独立にする

配慮不要特徴の
サンプル

目的変数の
サンプル

配慮不要特徴の
重みパラメータ

要配慮特徴の
サンプル

要配慮特徴の
重みパラメータ

正則化係数2

$$-\ln \Pr(\{(y, \mathbf{x})\}; \mathbf{w}) + \eta \mathrm{R}(\{(y, \mathbf{x}, \mathbf{s})\}, [\mathbf{w}, \mathbf{v}]) + \frac{\lambda}{2} \|[\mathbf{w}, \mathbf{v}]\|_2{}^2$$

対数尤度

予稿と異なり
尤度項には要配
慮特徴は除外

正則化係数1

大きいと
公正重視

先入観削除用の正則化項

小さな値ほど要配慮特徴
と目的変数は独立になる

普通の正則化項

L2正則化を採用
過学習の抑制

Prejudice Remover

要配慮特徴と目的変数を独立にするため,
これらの変数間の相互情報量を小さくする制約を加える

直接的な実装は…

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

X の値域によっては，勾配の計算にMCMCなどが必要になって大変

$$\sum_{Y\in\{0,1\}} \sum_{(\mathbf{x},\mathbf{s})} \Pr[y|\mathbf{x},\mathbf{s};\mathbf{w},\mathbf{v}] \ln \frac{\Pr[y|\bar{\mathbf{x}}(\mathbf{s}),\mathbf{s};\mathbf{w},\mathbf{v}]}{\Pr[y|\bar{\mathbf{x}},\bar{\mathbf{s}};\mathbf{w},\mathbf{v}]}$$

Pr[X,S]をサンプル分布で置換

特徴量が平均値の時の値で代用

Unfairness Hater

最も差別的な分類器からできるだけ離れた
分類器が選ばれるように制約を加える

- **最も差別的な分類器**：要配慮特徴のみで決定してしまう分類器
- **分類器から離れた**：KLダイバージェンスが大きい

事前に学習した
最も差別的な分類器

これから学習する
目標の分類器

$$-\sum_{X,S} \mathrm{D_{KL}}(\Pr[Y|S;\mathbf{\Psi}^*] \| \Pr[Y|X,S;[\mathbf{w},\mathbf{v}]])$$

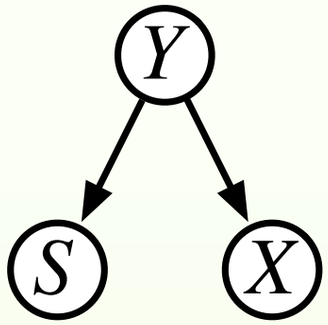
サンプル分布への置換と定数項の削除

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

Calders-Verwer 2単純ベイズ

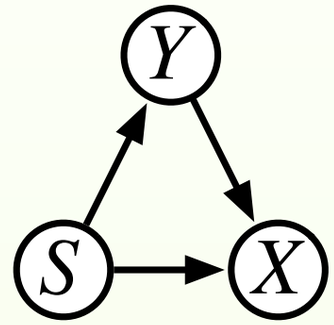
[Calders 10]

通常の単純ベイズ



- 要配慮・配慮不要特徴のどちらも，目的変数が与えられたとき全て条件付き独立

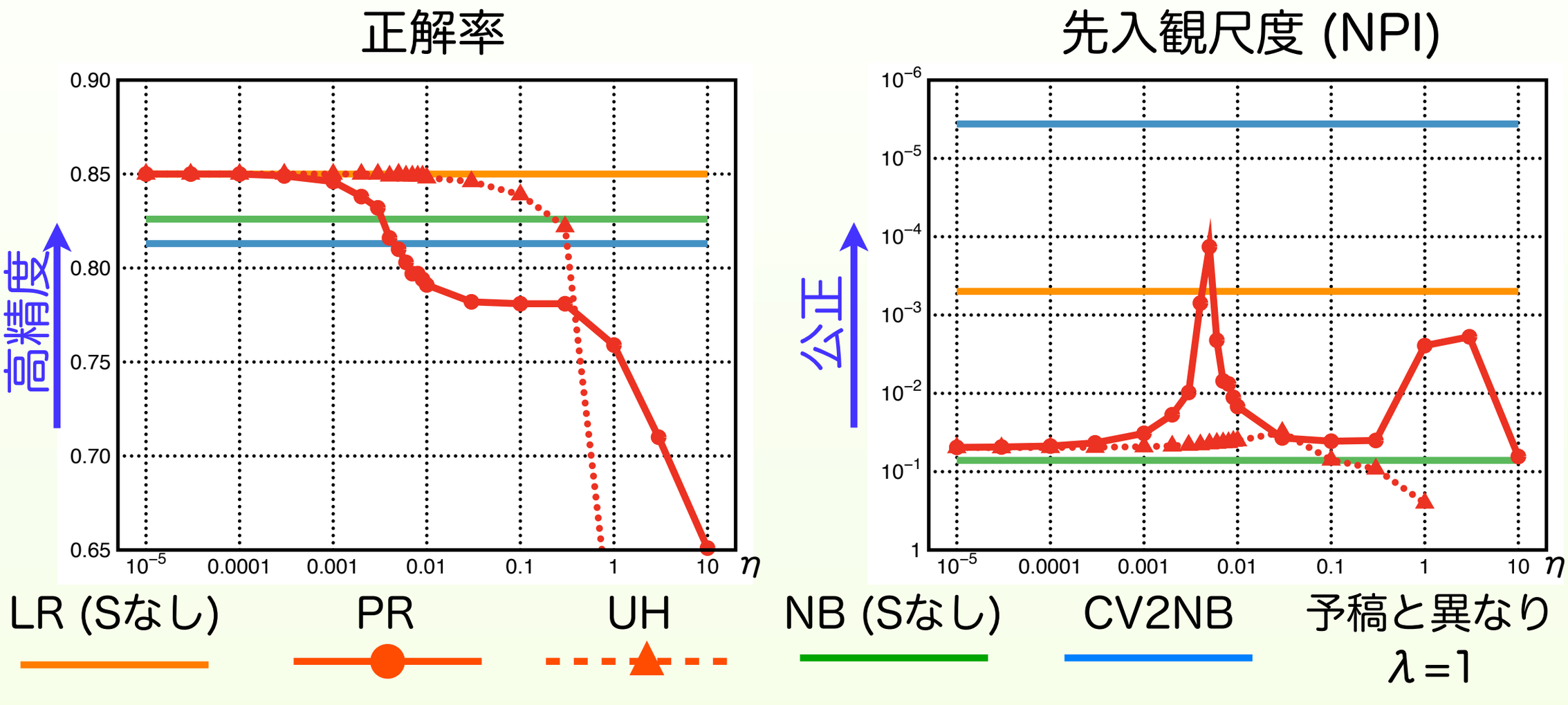
CV2NB



- 配慮不要特徴は，要配慮特徴と目的変数が与えられたとき全て条件付き独立
- 目的変数は要配慮特徴に依存

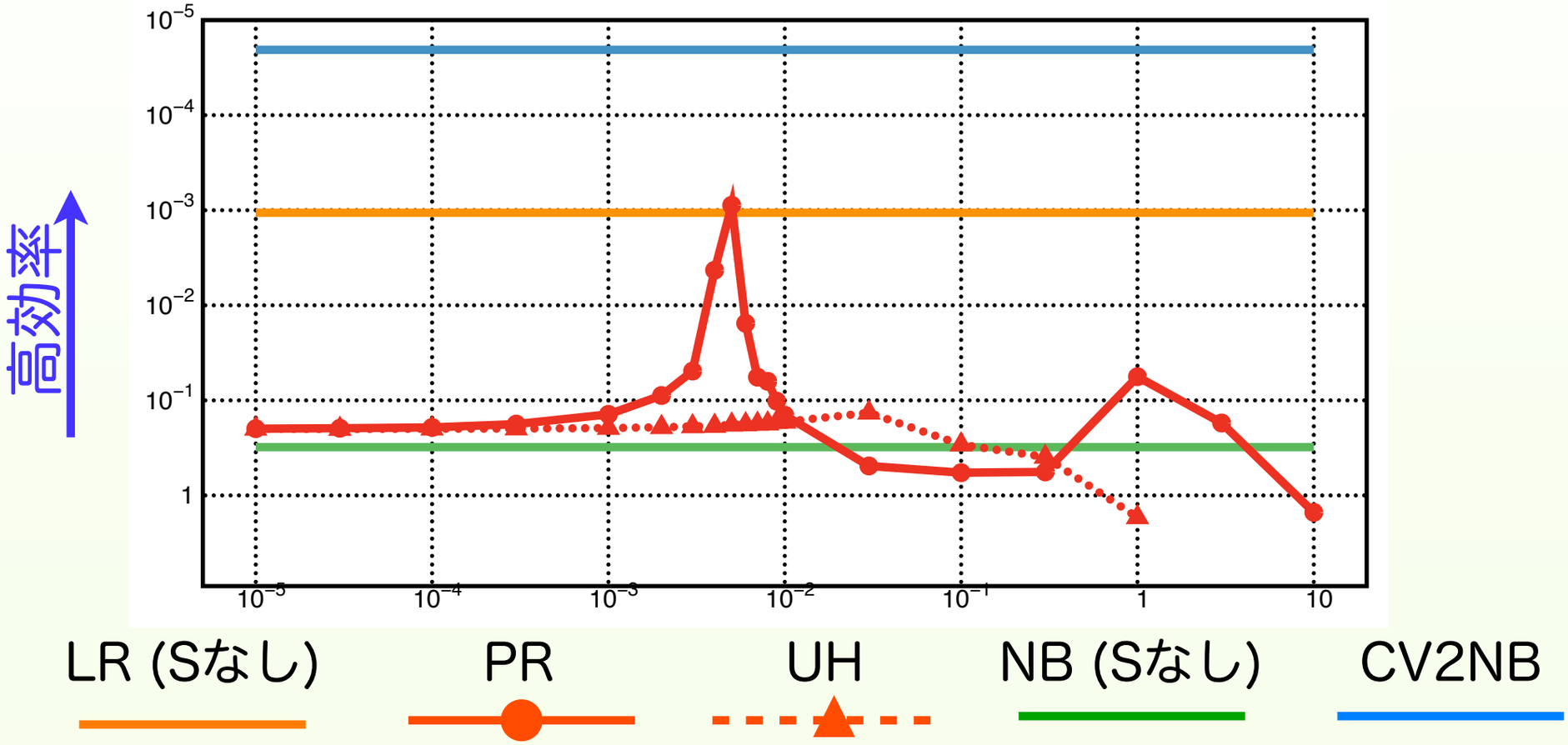
- 通常の単純ベイズと同様に，サンプルの分布からパラメータを決定
- さらに，サンプルでのCVスコアを最小化するように，Y と S の同時分布を修正

実験結果：正解率と先入観尺度



- ηを大きくすると正解率は低下する傾向，先入観の傾向は不明瞭
- CV2NBはNBより，PRとUHはLRに対して先入観の改善が見られた
- PRの方が，独立性の制約を直接的に定式化しており，UHより有効
- ηが大きいとき，特にUHは数値的に不安定

実験結果：精度・公正のトレードオフ



- PI/MI で精度と公正さのトレードオフを評価
  - PI：要配慮特徴と目的変数の相互情報量 → 小さいと公正
  - MI：予測クラスと標本クラスの相互情報量 → 大きいと高精度
- PRは若干LRを上回るが，UHは効率も悪い
- CV2NBの効率は非常によい

関連研究

[Pedreschi 08, Pedreschi 09]

差別的な相関ルールの列挙

例：クレジットの可否を決めるルール

(a) city=NYC => class=bad -- conf:(0.25)

(b) race=African, city=NYC ==> class=bad -- conf:(0.75)

a保護：要配慮なアイテムを加えたとき不利な決定をされる確信度が a 倍以上になると差別的とみなす

conf(b) / conf(a) = 3 なので a が 3 未満なら a保護 ではない

(c) neighborhood=10451, city=NYC ==> class=bad -- conf:(0.95)

(d) neighborhood=10451, city=NYC ==> race==African -- conf:(0.80)

(e) race=African, neighborhood=10451, city=NYC ==> class=bad

(c)には要配慮なアイテムは含まれていないが，(d)のように人種と住所に相関が強いと(e)のような不公正なルールができてしまう

まとめ

公正 / 差別配慮型学習

- 人種・性別など社会的公正さの観点から望ましくない要因が決定に関係しないように配慮したデータ学習・分析

本発表の寄与

- マイニングにおける公正さの要因について考察し三つの要因を挙げた
- 公正さに配慮した分類ができるような正則化項を2種類提案
- 実験により，公正さを改善できることを確かめた

社会責任的マイニング (Socially Responsible Mining)

- プライバシ保護データマイニング，敵対的学習，公正/差別配慮型マイニングなど，社会での利用を考慮したマイニング手法