1E1-OS-24a-1

公平配慮型データマイニング技術の進展

Advances in the Technologies of Fairness-aware Data Mining

神嶌 敏弘 *1
Toshihiro Kamishima

*1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, such technologies are being used for determinations that seriously affect individuals' lives, e.g., credit scoring. Fairness-aware data mining is a technology to make such determinations fair in terms of sensitive features, such as race, gender, religion, and so on. We survey the progress of this technology after our last survey at the JSAI2013.

1. はじめに

本稿では，公平配慮型データマイニング (fairness-aware data mining) 呼ばれる分析手法を概観する．公平配慮型データマイニングとは，公平性，差別，中立性，独立性などの潜在的な問題を考慮にいれたデータ分析のことである．データマイニング技術は与信，採用，保険などの個人に生活に大きく影響を与える決定に使われるようになってきている．このとき，社会的・法的な公平さに配慮した，すなわち，性別や人種などに基づく先入観や差別のない判断をするための手段として研究が始まった．2013 年の人工知能学会全国大会にて，この分野について報告したが，今回はその後の進展について紹介する．

2. 節では公平性の規準について，3. 節では，公平配慮型データマイニングのタスクについて述べる 4. 節では，近年の関連会議の動向について述べる

2. データマイニングにおける公平性

特徴と目的変数についていくつかの記号を定義したあと，形式的な公平性の規準を示す．

2.1 準備

確率変数 S と $\mathbf{X}$ は，それぞれセンシティブ特徴 (sensitive feature) と非センシティブ情報特徴（もしくは，単に特徴；non-sensitive feature）を表す．公平配慮型データマイニングでは，センシティブ特徴の表す性質に対して公平性を保証しつつ分析する．例えば，与信，採用，保険などの決定について扱うとき，社会的公平性の観点からその関与を排除すべき対象者の性別や人種といった個人属性情報を，このセンシティブ情報とする．なお，このセンシティブ特徴に何を設定するかは，データマイニングで扱うタスクと，法や規制などの社会的環境を考慮して与えるものとする．S は，連続変数でも離散変数でもよいが，既存の研究では主に定義域が $\{0, 1\}$ である二値変数の場合が扱われている．値 1 と 0 をとるときを，それぞれそれぞれ非保護状態と保護状態にあるといい，あるデータ集合中で，保護状態ある事例の集合を保護グループ，それ以外の事例集合を非保護グループという．一方の非センシティブ特徴 $\mathbf{X}$ は，対象を表す特徴の中で，上記のセンシティブ特徴以外の全てを含む特徴ベクトルである．確率変数 Y は目的変数で，分析者はこの変数の表す内容に関心がある．冒頭で挙げた例では，タスクでは与信・採用・保険などの決定を表す．公平配慮型タスクの場合では，Y は，与信などで有利な決定をする場合を正クラス 1 で，不利な場合を負クラス 0 で表す二値変数となる場合が主に研究されている．

連絡先: http://www.kamishima.net/

2.2 公平性の規準

いくつかの公平性規準を列挙する．最も単純なものはセンシティブ特長を取り除いてモデルを訓練する，すなわち目的変数は S とは独立で，$\mathbf{X}$ のみに依存する，$\Pr[Y|\mathbf{X}] = \Pr[Y|S, \mathbf{X}]$. この条件は条件付き独立性 $Y \perp\!\!\!\perp S \mid \mathbf{X}$ に該当する．この条件を満たさない場合は，直接差別 (direct discrimination) や差別的な取り扱い (disparate treatment) と呼ばれている [Pedreschi 08, Feldman 15, Zafar 16].

しかし，$\mathbf{X}$ の中に S と独立ではない変数は多数存在しうる．よく知られた事例として，銀行がローンの可否を定めるとき，アフリカ系住民が多い地区を不利に扱う運用をしていた red-lining がある．このような結果として不公平になっている場合を，間接差別 (indirect discrimination) や差別的効果 (disparate impact) という [Feldman 15]．この不公平を解消するには，条件なしの独立性 $Y \perp\!\!\!\perp S$ が成立する必要がある．

特定の条件・文脈が成立するときにのみ差別的かどうかを扱う場合がある．例えば，ローンの可否を決めるとき，アフリカ系全般では不公平な扱いはないが，ある特定の市に限れば不公平がある場合である．このような場合は文脈依存独立性 (context-specific independence) $Y \perp\!\!\!\perp S \mid X = x$ で表せる．この種の公平性規準は，相関ルールを対象に扱われている [Pedreschi 08, Hajian 13, Hajian 14].

以上の公平性規準は，$S = 0$ となる個人全体のグループについて平均的に成立する条件で，グループ公平性 (group fairness) と呼ばれている．この条件を厳密にし，どの個人についても公平にするのを個人公平性 (individual fairness) という．グループ公平性に特徴の写像についてリプシッツ性条件を加えるものや [Dwork 12]，公平性指標の平均ではなく最大値を制約するもの [Fukuchi 13] がある．

条件付き独立性 $Y \perp\!\!\!\perp S$ が成立しなくても，依存している条件 $Y \perp\!\!\!\perp S \mid \mathbf{X}^{(E)}$ によっては公平である場合がある．ここで，$\mathbf{X}^{(E)}$ は非センシティブ特徴 $\mathbf{X}$ の部分集合である．$\mathbf{X}^{(E)}$ は，たとえ間接的に目的変数に影響を与えたとしても，専門家や分析者が問題ないと判断した要因を表し，説明可能特徴 (explainable feature) や法的根拠のある属性 (legally-grounded attribute) という [Žliobaitė 11, Calders 13]．例えば，女性の入試の合格率が低かったとしても，それが合格率の低い医学部を受ける比率が女性の方が高いためであったとしたら不公平とはいえず，このと

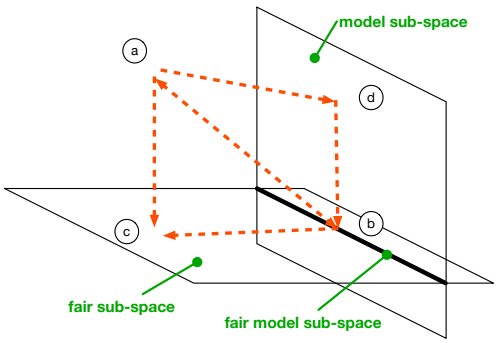

図 1: 不公平防止手法の分類

き，受験する学部が $\mathbf{X}^{(E)}$ に該当する．また，因果推論の文脈では $\mathbf{X}^{(E)}$ は交絡因子や合流点に該当し，傾向スコアを導入して扱うことができる [Calders 13].

その他，過去の決定に対して誤り率が偏らないようにという均等機会 (equal opportunity) は形式的には条件付き独立性 $Y \perp\!\!\!\perp S \mid Y'$ で表される [Hardt 16]．ただし，Y'，訓練データなどで観測された過去の判断結果で，場合によっては偏見のある判断を含む．独立性を弱めて無相関性とする場合 [Zafar 15]，バンディットで有意に差がある場合以外は腕を公平に扱う [Joseph 16] といった規準がある．

3. 公平配慮型データマイニングのタスク

公平配慮型データマイニングの分析タスクは，大きく不公平発見 (unfairness discovery) と不公平防止 (unfairness prevention) に分類できる [Ruggieri 10]．不公平発見では，分析結果に不公平なものが含まれているか，また含まれているとすればその結果を抽出する．不公平防止とは，不公平な分析結果が生じないようにしつつ，クラス分類や回帰といった分析を行う手法である．

不公平発見タスクには，最初にデータマイニングで公平性の問題を扱った不公平な相関ルールを発見する問題がある [Pedreschi 08]．与えられたデータ集合から頻出パターンマイニングで結論部が目的変数となる相関ルールを抽出したとき，目的変数の決定が不公平なものを検出する．監査 (auditing) は，ある変数の目的変数に与える影響の度合いを測る [Adler 16]．因果推論などでは変数の依存関係などのモデルが明かだが，この監査では特徴ベクトルを入力として与えその結果のみを観測できるブラックボックス分類器を仮定する．

不公平防止タスクが扱う分析クラスは，クラス分類 [Calders 10, Kamishima 12b, Kamiran 12a, Hajian 13, Zemel 13, Feldman 15, Kamiran 10, Fukuchi 14, Zafar 15, Zafar 16]，回帰 [Fukuchi 13, Calders 13]，推薦 [Kamishima 12a, Kamishima 13, Kamishima 16] などがある．さらに処理の方法に基づいて，前処理型 (pre-process)，中処理型 (in-process)，および後処理型 (post-process) に分けられる [Ruggieri 10]．図 1 は $(Y, \mathbf{X}, S)$ 上の分布を表す．垂直な平面で表したものは確率分布のモデル分布の族を表すモデル部分空間，水平な平面は 2. 章の規準を満たす公平部分空間である．公平性制約を満たさない可能性のある ⓐの分布から得た標本・訓練データから，公平性を満たすモデル分布の中で最も近似誤差の小さな ⓑを見つけることが，不公平防止タスクの目標である．

前処理型では，ⓐの訓練データを公平性を満たしつつ歪みが最小な ⓒに写像し，その後，通常の分類器などを使って最終モデル ⓑを見つける [Kamiran 12a, Hajian 13, Zemel 13, Feldman 15]．任意の分類器を利用できる利点があるが，分類器についてなんら仮定を導入せずに公平部分空間への適切な写像を決めるのには困難が伴う．

中処理型の手法は，ⓐの訓練データから，最終モデルを ⓑの目標モデルを直接獲得する [Kamishima 12b, Kamishima 12a, Kamishima 13, Kamishima 16, Fukuchi 13, Kamiran 10, Fukuchi 14, Zafar 15, Zafar 16, Calders 10, Kamiran 12b, Kamiran 13, Hardt 16]．この手法は制約が少ないので，潜在的に最もよい公平性と性能のトレードオフを達成できる可能性がある．しかし，目的関数の設計やその最適化には技術的な困難が伴う．

後処理型では，通常の予測器を使って学習して ⓒのモデルを獲得し，その後公平性制約を満たすようにそのモデルを修正して最終モデル ⓑを得る [Calders 10, Kamiran 12b, Kamiran 13, Hardt 16]．この方法では，公平な予測結果は非センシティブ特徴には依存せず，通常の予測モデル ⓒの予測結果とセンシティブ特徴にのみ依存する紛失性 (oblivious) という仮定 [Hardt 16] が必要になる．しかし，この仮定により公平な予測結果の設計は非常に簡潔になる．

4. 近年の公平性関連会議の動向

最後にデータマイニング・機械学習における公平性に関連した会議の動向についてまとめておく．2013 年の人工知能学会全国大会で公平配慮型データマイニングを紹介したときは，ICDM2012 併設のワークショップ「Discrimination and Privacy-Aware Data Mining」しか公平性を扱うワークショップは開催されていなかった．その後，NIPS2013 や ICML2014 では新たに「Fairness, Accountability, and Transparency in Machine Learning」のワークショップが続けて開催され，公平配慮型データマイニングに加えて機械学習の説明可能性や透明性を含めた議論が始まった．2016 年にはさらに議論は活発になり NIPS2016 ではシンポジウム「Machine Learning and the Law」が，ICDM2016 では新たなワークショップ「Privacy and Discrimination in Data Mining」が企画された．

その他，KDD2016 ではチュートリアル「Algorithmic Bias: from Discrimination Discovery to Fairness-aware Data Mining」[Hajian 16] が開催された．資料も公開されており，この分野を俯瞰するのによいだろう．今年の KDD2017 では差分プライバシで著名な Dwork による基調講演で，この公平性の問題が扱われる予定とのことである．このように，研究コミュニティも活発化しており，さらなる研究の進展が期待される．

謝辞：本研究は JSPS 科研費 24500194，15K00327，23240043，および 16H02864 の助成を受けた．

参考文献

[Adler 16] Adler, P., Falk, C., Friedler, S., Rybeck, G., Schedegger, C., Smith, B., and Venkatasubramanian, S.: Auditing Black-box Models for Indirect Influence, in *Proc. of the 16th IEEE Int'l Conf. on Data Mining*, pp. 1–10 (2016)

[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)

[Calders 13] Calders, T., Karim, A., Kamiran, F., Ali, W., and Zhang, X.: Controlling Attribute Effect in Linear Regression,

in *Proc. of the 13th IEEE Int'l Conf. on Data Mining*, pp. 71–80 (2013)

[Dwork 12] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R.: Fairness Through Awareness, in *Proc. of the 3rd Innovations in Theoretical Computer Science Conf.*, pp. 214–226 (2012)

[Feldman 15] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S.: Certifying and Removing Disparate Impact, in *Proc. of the 21st ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 259–268 (2015)

[Fukuchi 13] Fukuchi, K., Sakuma, J., and Kamishima, T.: Prediction with Model-based Neutrality, in *Proc. of the ECML PKDD 2013, Part II*, pp. 499–514 (2013), [LNCS 8189]

[Fukuchi 14] Fukuchi, K. and Sakuma, J.: Neutralized Empirical Risk Minimization with Generalization Neutrality Bound, in *Proc. of the ECML PKDD 2014, Part I*, pp. 418–433 (2014), [LNCS 8724]

[Hajian 13] Hajian, S. and Domingo-Ferrer, J.: A Methodology for Direct and Indirect Discrimination Prevention in Data Mining, *IEEE Trans. on Knowledge and Data Engineering*, Vol. 25, No. 7, pp. 1445–1459 (2013)

[Hajian 14] Hajian, S., Domingo-Ferrer, J., and Farràs, O.: Generalization-based Privacy Preservation and Discrimination Prevention in Data Publishing and Mining, *Data Mining and Knowledge Discovery* (2014)

[Hajian 16] Hajian, S., Bonchi, F., and Castillo, C.: Algorithmic Bias: from Discrimination Discovery to Fairness-Aware Data Mining, The 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, Tutorial (2016)

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

[Joseph 16] Joseph, M., Kearns, M., Morgenstern, J., and Roth, A.: Fairness in Learning: Classic and Contextual Bandits, in *Advances in Neural Information Processing Systems 29* (2016)

[Kamiran 10] Kamiran, F., Calders, T., and Pechenizkiy, M.: Discrimination Aware Decision Tree Learning, in *Proc. of the 10th IEEE Int'l Conf. on Data Mining*, pp. 869–874 (2010)

[Kamiran 12a] Kamiran, F. and Calders, T.: Data Preprocessing Techniques for Classification without Discrimination, *Knowledge and Information Systems*, Vol. 33, pp. 1–33 (2012)

[Kamiran 12b] Kamiran, F., Karim, A., and Zhang, X.: Decision Theory for Discrimination-aware Classification, in *Proc. of the 12th IEEE Int'l Conf. on Data Mining*, pp. 924–929 (2012)

[Kamiran 13] Kamiran, F., Žliobaitė, I., and Calders, T.: Quantifying Explainable Discrimination and Removing Illegal Discrimination in Automated Decision Making, *Knowledge and Information Systems*, Vol. 35, pp. 613–644 (2013)

[Kamishima 12a] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J.: Enhancement of the Neutrality in Recommendation, in *The 2nd Workshop on Human Decision Making in Recommender Systems* (2012)

[Kamishima 12b] 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 13] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J.: Efficiency Improvement of Neutrality-enhanced Recommendation, in *The 3rd Workshop on Human Decision Making in Recommender Systems* (2013)

[Kamishima 16] Kamishima, T., Akaho, S., Asoh, H., and Sato, I.: Model-Based Approaches for Independence-Enhanced Recommendation, in *Proc. of the IEEE 16th Int'l Conf. on Data Mining Workshops*, pp. 860–867 (2016)

[Pedreschi 08] Pedreschi, D., Ruggieri, S., and Turini, F.: Discrimination-aware Data Mining, in *Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 560–568 (2008)

[Ruggieri 10] Ruggieri, S., Pedreschi, D., and Turini, F.: Data Mining for Discrimination Discovery, *ACM Transactions on Knowledge Discovery from Data*, Vol. 4, No. 2 (2010)

[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)

[Zafar 15] Zafar, M. B., Martinez, I. V., Rodriguez, M. G., and Gummadi, K.: Fairness Constraints: A Mechanism for Fair Classification, in *ICML2015 Workshop: ss, Accountability, and Transparency in Machine Learning* (2015)

[Zafar 16] Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P.: Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment, `arXiv:1610.08452 [stat.ML]` (2016)

[Zemel 13] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C.: Learning Fair Representations, in *Proc. of the 30th Int'l Conf. on Machine Learning* (2013)