3L1-OS-06a-1

中立性・公正性に配慮したデータ分析

Data Analysis Techniques Concerning the Neutrality and Fairness

神嶌 敏弘
Toshihiro Kamishima

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. For example, credit scoring is frequently determined based on the records of past credit data together with statistical prediction techniques. Needless to say, such determinations must be fair regarding sensitive features such as race, gender, religion, and so on. In this paper, after demonstrating the applications of these techniques, we explore the techniques concerning fairness and relation with privacy-preserving techniques.

1. はじめに

本稿では，公正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．公正配慮型データマイニングとは，公正性，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．いくつかの応用分野を紹介したあとで，今までに提案された概念や手法の概略を紹介し，プライバシ保護技術との関連について論じる．当初，差別配慮型データマイニング (discrimination-aware data mining) という名称で提案された [Pedreschi 08] が，ここでは公正配慮型データマイニングと呼ぶ．その理由は，差別解消以外の目的にもこの技術は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』や『識別』の意味で使われ誤解が生じやすいためでもある．

2. 節では公正配慮型データマイニングの応用分野を挙げる．3. 節では，公正配慮型データマイニングの手法を俯瞰する．4. 節では，プライバシ保護技術との関連を論じる．最後の 5. 節はまとめである．

2. 応用分野

社会的差別，公正性，中立性，独立性に対して配慮したマイニング手法の応用タスクを例示する．

2.1 差別的要因に配慮した決定

社会的差別を回避する配慮をしたマイニング手法の提案 [Pedreschi 08] は，公正配慮型データマイニングの最初の適用分野である．個人属性情報，金融取引履歴，通信履歴，税務記録など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が進んでいる．それに伴って，与信，採用，保険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っている．このとき，社会的・法的な公正さに配慮した，すなわち，性別，信仰，人種，ハンディキャップ，政治的信条などに基づく先入観や差別のない判断がなされるようなマイニング手法が利用されるべきである．

単純に考えると，これらの差別的な情報や特徴を分析で利用しなければ十分に思えるが，それではこうした情報や特徴の間接的な影響を排除できないことが報告されている．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間接的になされてしまう．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．利用者の配慮が必要な情報が多くの情報が複合的に関連している場合においても，差別的な決定を回避できるようなマイニング手法が必要となる．

2.2 情報中立推薦

公正配慮型データマイニングは，Pariser が主張するフィルターバブル問題に対し，利用者が指定した視点に関して中立性を保つ目的でも利用できる．フィルターバブル (Filter Bubble) [Pariser 11, パリサー 12] とは，推薦を含めた個人化技術によって，利用者が接する情報の話題の範囲が狭められたり，偏ったりすることが，利用者が知らないうちに行われるという問題である．この問題に関する TED Talk [Pariser] などで，Facebook の友人推薦で推薦される利用者の政治的立場の偏りや，2011 年のエジプトの政変という重大な事件であっても個人の嗜好によっては Google 検索の検索結果から除外される事例を示している．このフィルターバブルによって生じると主張されている問題は，利用者が多様な情報に接する機会が少なくなることと，各人がそれぞれ異なる限られた情報にのみ接していて，社会の中で供給される情報が減ってしまうことの二つに要約できるだろう．

この問題に対し，利用者や他の主体が指定した観点に対して中立性を保証するのが情報中立推薦 (Information-neutral Recommendation) [神嶌 12d, Kamishima 12b] である．Pariser の Facebook の場合を例にとれば，推薦される友人が保守派か革新派かという特定の観点については中立性を保証するが，他の観点，例えば出身地などについては，個人の嗜好を反映した偏りを許すような推薦を行う．この情報中立推薦では，利用者が指摘した情報に対する中立性に配慮して，利用者の嗜好の予測を行うため，公正配慮型のマイニング手法が必要となる．

2.3 無関心な情報の排除

公正配慮型データマイニングが提案されるより前ではあるが，無関心な情報を排除するクラスタリング手法の適用事例を紹介する．Coordinated Conditional Information Bottleneck (CCIB) [Gondek 04] は，特定の補助情報 (side information) とは独立な分割を獲得するクラスタリング手法である．この論文では，顔画像のクラスタリング問題に適用している．単純にクラスタリングを適用すると，顔だけの画像と，肩から上の画像の二種類

連絡先: http://www.kamishima.net/

のクラスタに分割されてしまうが，この結果は分析者にとっては無関心なものだったとしよう．そこで，この分割結果を補助情報とし，この分割をはできるだけ無関係なクラスタを CCIB によって獲得したところ，男女を分けるクラスタが得られた．このように，分析者には関心のない情報とは無関係な分析結果を得る目的でも公正配慮型データマイニングは利用できる．

2.4 その他の応用問題

プライバシーポリシーによって制限される個人情報や，金融取引においてその利用が制限されるインサイダー情報など法や規制によって制限された情報の利用を排除することも，公正配慮型データマイニングの応用分野の一つである．

情報中立推薦では，推薦システムの利用者が中立な推薦を受けるものであったが，検索や推薦のサービス提供者が，商品や情報の提供者を公正に扱う目的にも利用可能である．Google は自身のサービスと競合するサービスを不公平に扱ったかどうかについて FTC の調査を受けた [Forden 12]．このような問題に対し，サービスや商品の提供者という情報に関して中立性を保つ検索や推薦を行う目的にも，公正配慮型データマイニング技術は利用可能だろう．

3. 公正配慮型データマイニング

ここでは，公正配慮型データマイニングの問題について，変数の分類，公正性・中立性，そしてタスクの分類について順に述べる．

3.1 変数の分類

確率変数 S と $\mathbf{X}$ は，それぞれ要配慮特徴 (sensitive feature) と配慮不要特徴 (non-sensitive feature) を表す．公正配慮型データマイニングでは，要配慮特徴の表す性質に対して公正性を保証しつつ分析する．2.1 節の差別配慮型タスクでは，社会的・法的な公正さを反映した性別・信仰・人種を表す要配慮特徴に，2.2 節の情報中立推薦では，要配慮特徴が示す情報に対して中立性を保証するフィルタリングをする．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値域が $\{0, 1\}$ の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．一方の配慮不要特徴は，対象を表す特徴の中で，上記の要配慮特徴以外の全てである．さらに，配慮不要特徴を二種類に分類する場合もある．特徴のセマンティクスに基づいて，たとえ配慮不要特徴が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配慮不要特徴を説明可能 (explainable) であるという．これは文献 [Žliobaitė 11] の定義に従うもので，文献 [Luong 11] での法的根拠のある属性 (legally-grounded attribute) に該当する．この説明可能な配慮不要特徴を集めた確率変数を $\mathbf{X}^{(E)}$ で表し，それ以外の説明不可能な配慮不要特徴を集めた確率変数を $\mathbf{X}^{(U)}$ と表記する．

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．目的変数は，差別配慮型タスクでは与信・採用・保険などの決定を表し，情報中立推薦では評価スコアや適合不適合の決定を表し，そして非冗長クラスタリングの例では潜在的なクラスタを表現する．差別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．文献 [神嶌 12d, Kamishima 12b] の情報中立推薦の実装では，評価スコアを表す Y は実数変数である．

3.2 公正性・中立性

次に，これらの変数間の関係に基づいた公正性や中立性について述べる．今までにいくつかの公正性や中立性の規準が提案されているが，それらは目的変数 Y と要配慮特徴 S の間の独立性に基づいている [Kamishima 12a]．この独立性の条件の違いにより公正性は直接的なものと間接的なものとに分けられる．

目的変数が分布 $\Pr[Y|\mathbf{X}, S]$ によって決定されているとき，目的変数 Y は要配慮特徴 S に依存しているので明らかに不公正である．そこで，この不公正を解消するため，目的変数の生成モデルから S を取り去って $\Pr[Y|\mathbf{X}, S] = \Pr[Y|\mathbf{X}]$ が成立するとする．この状態を直接的公正性が達成されている，もしくは直接的差別が解消されているという．このとき，変数 Y，$\mathbf{X}$，および S の関係は次式で表される．

$$\begin{aligned}\Pr[Y, \mathbf{X}, S] &= \Pr[Y|\mathbf{X}, S]\Pr[S|\mathbf{X}]\Pr[\mathbf{X}] \\ &= \Pr[Y|\mathbf{X}]\Pr[S|\mathbf{X}]\Pr[\mathbf{X}]\end{aligned}$$

この関係は，$\mathbf{X}$ が与えられたときに Y と S は条件付き独立である，すなわち $Y \perp\!\!\!\perp S \,|\, \mathbf{X}$ であることを示している．

もう一方の間接的公正性が達成されている，もしくは間接的差別が解消されている状態とは，条件なしに Y と S が独立であること，すなわち $Y \perp\!\!\!\perp S \,|\, \emptyset = Y \perp\!\!\!\perp S$ である．Y の生成モデルに直接的に公正な場合でも，間接的には差別的な場合がありうる．例として，Y，X，および S が全て実数のスカラー変数であり，真のモデルは次式を満たす場合を考える．

$$Y = X + \varepsilon_Y \quad \text{and} \quad S = X + \varepsilon_S$$

ただし，ε_Y と ε_S 互いに独立で，平均 0 の確率変数とする．$\Pr[Y, X, S] = \Pr[Y|X]\Pr[S|X]\Pr[X]$ より，これらの変数は $Y \perp\!\!\!\perp S \,|\, X$ の条件を満たすが，$Y \perp\!\!\!\perp S$ の条件は満たさない．もし，ε_Y と ε_S の分散が小さければ，Y と S の相関は非常に高くなる．このとき，モデルが直接的には公正でも，分類結果は明らかに要配慮特徴に依存してしまう．よって，このモデルは直接的に公正だが，間接的には不公正で，これが sec:discrimination-aware 節の red-lining 効果に該当する．こうした間接的不公正を取り除くには，決定モデルが $Y \perp\!\!\!\perp S$ を満たすようにしなければならない．

さらに，これらの独立性を測る尺度にもいろいろなものが提案されており，拡張リフト (extended lift; `elift`) [Pedreschi 08]，CV スコア (Calders-Verwer's discrimination score; CV score)[Calders 10]，相互情報量に基づく先入観尺度 (prejudice index; PI)[Kamishima 12c]，差分公正性 (differential fairness)[Dwork 11]，統計的独立性検定の p 値 [Pedreschi 09]，χ 二乗統計量 [Berendt 12] などがある．これらの尺度の期待値や上限の大小により公正性の度合いを定量化している．さらに，間接公正性において，説明可能な配慮不要特徴が存在する場合には，それら条件部に残ることを許す場合もある．こうした設定は条件付き差別 (conditional discrimination)[Žliobaitė 11] や situation testing[Luong 11] として研究が行われている．

3.3 問題の種類

前述の公正性を用いて，不公正の検出，公正配慮型予測，および公正配慮型データ公開などの問題が扱われている．不公正の検出とは，与えられたデータベースの中の分布や，データベースから抽出した相関ルールの中から差別的な決定を検出するものである．公正配慮型予測とは，公正性の規準を満たすような制約の下で，目的変数の値を予測する関数・ルールを獲得

する問題である．最後の，公正配慮型データ公開とは，要配慮特徴の影響が配慮不要特徴に及ばないように配慮不要特徴を変換してデータを公開することで，そのデータを用いて分析をしても不公正な判断が生じないようにする問題である．

4. プライバシ保護技術との関連

最後に，公正配慮型データマイニング技術とプライバシ保護データマイニング技術との関連について述べる．プライバシ保護技術の目的は，個人の特定を抑制するものと，属性情報の秘匿とに大きく分けられる．前者は世の中の誰であるかを特定出来ないようにするもので，後者は誰かということは分かるがその性質や行動を表す情報を秘匿する．個人の特定を抑制する技術としては，k 匿名性・l 多様性，差分プライバシ，画像のモザイク処理，統計的個票開示などが挙げられる．一方で，ランダム化や秘密関数計算などは属性情報を秘匿する技術である．

公正性は目的変数 Y と要配慮特徴 S の間の独立性を扱うが，これは属性情報の秘匿と関連している．この独立性が成り立つことは，Y と S の相互情報量が小さいことを示すが，これは，Y の値が知られたときでも，S の情報を秘匿したことに相当し，属性情報 S を秘匿したことに該当している．

この意味では，公正配慮型データマイニングはプライバシ保護データマイニングの一部といえる．しかし，2.2 節の情報中立推薦や，2.3 節の無関心な情報の排除といった応用問題においては，要配慮特徴 S の情報を秘匿することが目的ではなく，プライバシ保護とはあまり関係がない．このように，公正配慮型データマイニングとプライバシ保護データマイニングは深い関連はあるが，異なる問題を扱っているといえよう．

5. まとめ

以上，公正配慮型データマイニングの応用分野，手法，他分野との関連について論じた．なお，本公正配慮型データマイニングについてまとめた文書を次の URL で公開しているので参考にされたい．

http://www.kamishima.net/archive/fadmdoc.pdf
http://www.kamishima.net/archive/fadm.pdf

謝辞

本研究は JSPS 科研費 16700157，21500154，23240043，および 24500194 の助成を受けた．

参考文献

[Berendt 12] Berendt, B. and Preibusch, S.: Exploring Discrimination: A User-Centric Evaluation of Discrimination-Aware Data Mining, in *Proc. of the IEEE Int'l Workshop on Discrimination and Privacy-Aware Data Mining*, pp. 344–351 (2012)

[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)

[Dwork 11] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R.: Fairness Through Awareness, arXiv:1104.3913 [cs.CC] (2011)

[Forden 12] Forden, S.: Google Said to Face Ultimatum From FTC in Antitrust Talks, Bloomberg (2012), ⟨http://bloom.bg/PPNEaS⟩

[Gondek 04] Gondek, D. and Hofmann, T.: Non-Redundant Data Clustering, in *Proc. of the 4th IEEE Int'l Conf. on Data Mining*, pp. 75–82 (2004)

[Kamishima 12a] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J.: Considerations on Fairness-aware Data Mining, in *Proc. of the IEEE Int'l Workshop on Discrimination and Privacy-Aware Data Mining*, pp. 378–385 (2012)

[Kamishima 12b] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J.: Enhancement of the Neutrality in Recommendation, in *Proc. of the 2nd Workshop on Human Decision Making in Recommender Systems*, pp. 8–14 (2012)

[Kamishima 12c] 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]

[神嶌 12d] 神嶌 敏弘, 赤穂 昭太郎, 麻生 英樹, 佐久間 淳：情報中立推薦システム, 人工知能学会全国大会（第 26 回）論文集, 3E1-R-6-1 (2012)

[Luong 11] Luong, B. T., Ruggieri, S., and Turini, F.: k-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention, in *Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 502–510 (2011)

[Pariser] Pariser, E.: The Filter Bubble: ⟨http://www.thefilterbubble.com/⟩

[Pariser 11] Pariser, E.: *The Filter Bubble: What The Internet Is Hiding From You*, Viking (2011)

[パリサー 12] パリサー イーライ, 井口 耕二：閉じこもるインターネット — グーグル・パーソナライズ・民主主義, 早川書房 (2012)

[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* (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*, pp. 581–592 (2009)

[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)