Convex Fairness Constrained Model Using Causal Effect Estimators (因果効果推定量を用いた 公平性制約について)

数理情報学専攻 48-186203 小倉 輝
指導教員 武田 朗子 教授

1 序論

近年、多くの機械学習における公平性に関する研究が行われている。機械学習の実システムへの応用の際、学習に用いるデータがバイアスを含んでいたり、学習に用いるモデルやアルゴリズムによっては学習したモデルが特定の特徴を持つデータに対して差別的な予測を行うことがある。バイアスを含んでいる可能性のあるデータセットから、精度の低下を抑えつつ性別や人種といったセンシティブ特徴と呼ばれる特徴について差別的でないモデルを学習することが公平性を考慮した機械学習の目標の一つである。本研究ではセンシティブ特徴が 2 値 (男/女, 白人/黒人など) である設定を考える。公平性考慮型機械学習の応用先には与信審査や再犯予測などがある。

機械学習における差別には、センシティブ特徴が予測に直接与える影響 (直接差別) とセンシティブ特徴と相関のある特徴を経由して生じる間接的な影響 (間接差別) がある。間接差別の中には説明可能な差異という差異がある。これはセンシティブ特徴と相関があるものの、予測へ影響しても差別的でないと考えられる特徴 (説明可能な特徴) を生じる差異である。例えば、与信審査を行う機械学習モデルが人種に関して差別的かを考えるとき、予測値に生じる人種間の差のうち年収の差による差異は許容し、差別ではないと考える。図 1 にこれらの関係図を示す。

予測の公平さを評価する指標に、センシティブ特徴ごとに分けた各集団の予測平均の差である Mean Difference (MD) [1] があるが、MD は説明可能な差異も差別に含めて定量化するため、MD=0 を制約にするモデルでは逆差別を引き起こす可能性がある。今回の研究では、MD のうち説明可能な差異を除いた以下の差異を差別とし、この差別を防ぐような手法を考える。

$$\text{Discrim.} = \text{MD - 説明可能な差異.} \tag{1}$$

説明可能な差異を残しつつ差別を除去する既存手法に Multi MD[1] があるが、幾つかの状況下で性能が悪化する。本研究では既存手法が正しく動作しない状況でも性能の良いモデル FairCEEs を提案し、理論的かつ実験的に性能を検証した。

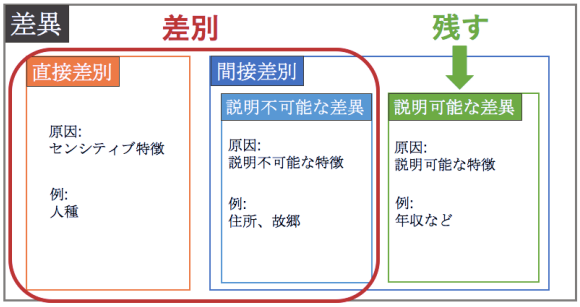


図 1. 直接差別、間接差別、説明可能な差異 [1] の関係図。

2 提案手法

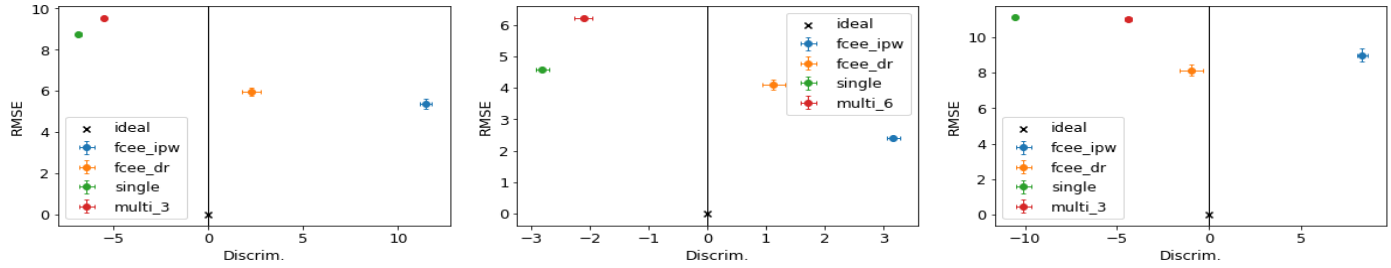
本研究では、因果効果 [2] を差別の定量化と関連付けることで提案手法を設計した。因果効果は介入 (新薬の投与や広告の実施など) が結果に与える影響を定量化することに用いられる。因果効果の推定では介入と結果の両方に相関のある特徴である共変量の影響を取り除くことで推定を行う。本研究では介入をセンシティブ特徴、共変量を説明可能な特徴とすることで因果効果推定を差別の定量化に用いる。具体的には、因果効果推定量である IPW [3]、DR 推定量 [4] を用いて差別を定量化し、これを制約に用いることで説明可能な差異を残しつつ差別を除去している。

Definition 2.1 (IPW 推定量 [3])**.** z_i を i 番目の入力の傾向スコア *[5]*, $\hat{y}_i$ を i 番目のデータの予測値とする。*IPW* 推定量は以下のように定義される。

$$IPW = \frac{\sum_{i=1}^{N} \frac{s_i}{z_i}\hat{y}_i}{\sum_{i=1}^{N} \frac{s_i}{z_i}} - \frac{\sum_{i=1}^{N} \frac{1-s_i}{1-z_i}\hat{y}_i}{\sum_{i=1}^{N} \frac{1-s_i}{1-z_i}}. \tag{2}$$

IPW 推定量を制約に用いた FairCEE-IPW を以下のように定式化する。($\mathcal{L}$ は損失関数である)

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \mathcal{L}(\boldsymbol{w}) \\ \text{s.t.} \quad & \boldsymbol{h}^\top \boldsymbol{X} \boldsymbol{w} = 0, \end{aligned} \tag{3}$$



CASE 1　　CASE 2　　CASE 3

図 2. 人工データでの提案手法と既存手法の比較。横軸は Discrim. (1) を表し 縦軸は RMSE を表す。既存手法の適用が困難な 3 つの CASE において、FairCEE-DR の RMSE, Discrim. ともに既存手法である Multi MD よりも小さく、精度と公平性の両面で既存手法よりも優れていることがわかる。

ただし

$$\boldsymbol{h} = \frac{\boldsymbol{a}}{\mathbf{1}^\top \boldsymbol{a}} - \frac{\boldsymbol{b}}{\mathbf{1}^\top \boldsymbol{b}}, \boldsymbol{a} = (\frac{s_1}{z_1}, \frac{s_2}{z_2}, ..., \frac{s_N}{z_N})^\top, \quad (4)$$

$$\boldsymbol{b} = (\frac{1-s_1}{1-z_1}, \frac{1-s_2}{1-z_2}, ..., \frac{1-s_N}{1-z_N})^\top. \quad (5)$$

DR 推定量の定義は以下である。

Definition 2.2 (DR 推定量 [4])**.** z_i を i 番目の入力の傾向スコア *[5]*, $\hat{y}_i$ を i 番目のデータの予測値とし、g_i^+, g_i^- を i 番目のデータの潜在結果変数 *[2]* の推定値とする。*DR* 推定量は以下のように定義される。

$$DR = \frac{1}{N}\sum_{i=1}^{N}[(\frac{s_i}{z_i} - \frac{1-s_i}{1-z_i})\hat{y}_i + (1-\frac{s_i}{z_i})g_i^+ - (1-\frac{1-s_i}{1-z_i})g_i^-].$$

DR 推定量を制約に用いたモデル FairCEE-DR を以下のように定式化する。

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \mathcal{L}(\boldsymbol{w}) \\ \text{s.t.} \quad & (\boldsymbol{a}-\boldsymbol{b})^\top \boldsymbol{X}\boldsymbol{w} + (\mathbf{1}-\boldsymbol{a})^\top \boldsymbol{g}^+ - (\mathbf{1}-\boldsymbol{b})^\top \boldsymbol{g}^- = 0 \end{aligned} \quad (6)$$

ただし

$$\boldsymbol{g}^+ = (g_1^+, g_2^+, ..., g_N^+)^\top, \quad \boldsymbol{g}^- = (g_1^-, g_2^-, ..., g_N^-)^\top.$$

3　理論保証

本研究では、FairCEE-IPW について以下の定理を示した。

Theorem 3.1 (FairCEE-IPW Loss)**.** 目的関数を最小二乗誤差とする。$\mathcal{L}_{md}^*$ を *MD=0* 制約での目的関数の最適値、$\mathcal{L}_{ipw}^*$ を *FairCEE-IPW (3)* の目的関数の最適値とする。

$$0 \leq IPW \leq MD$$

の仮定が成り立つとき、以下の不等式が成り立つ。

$$\mathcal{L}_{ipw}^* \leq \mathcal{L}_{md}^*.$$

IPW 推定量を用いて説明可能な差異を考慮して差別を定量化する FairCEE-IPW の訓練誤差が、考慮しない MD=0 制約よりも小さくなることを示している。これは逆差別を防ぐことで精度の悪化を緩和していると解釈できる。なお本研究では既存手法 [1] についても類似した定理を示すことができた。

4　結論

本研究では、差別の定量化に因果効果を用いるアイデアのもとに、因果効果推定量である IPW, DR 推定量を制約に用いることで、説明可能な差異を考慮した公平な機械学習モデル FairCEEs を提案した。また目的関数に $\mathcal{L}(\boldsymbol{w})$ に二乗損失を用いた場合に、ある仮定のもとで FairCEE-IPW (3) の最適解が説明可能な差異を考慮しない MD=0 制約を用いた場合よりも訓練誤差を小さくすることを理論的に示した。最後に、既存手法の適用が困難な状況下において提案手法が既存手法よりも優れていることを数値実験によって示した。

参考文献

[1] Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. Controlling attribute effect in linear regression. *2013 IEEE 13th International Conference on Data Mining*, pp. 71–80, 2013.

[2] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, Vol. 66, pp. 688–701, 10 1974.

[3] Donald. B. Rubin. The use of propensity scores in applied Bayesian inference. In *Bayesian Statistics 2*, pp. 463–472. North-Holland/Elsevier (Amsterdam; New York), 1985.

[4] Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, Vol. 61, pp. 692–972, 2005.

[5] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, Vol. 70, No. 1, pp. 41–55, 04 1983.