Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer

論文情報

- David Madras, Toniann Pitassi, Richard Zemel (University of Toronto, Vector Institute)
- 32nd Conference on Neural Information Processing Systems (NeurIPS 2018).

はじめに

- ローンの審査や裁判の判断などで自動決定機が使われてきている また重要な決定の時ではこの判断に対して人間が監査する．機械学習モデルは精度としてはSOTAかも知れないが，本当の効果は，人間がどう判断するかの過程から来るものではないのか？

- 意思決定者(裁判官など)と自動決定機は同じ目標を持っていなければならない．

- 意思決定者と自動決定機らを同じ目標にすることを解決する手法として，Rejection learningがある．自動決定機たちに予測を行わないという選択肢を与える．自動決定機が予測しなかった時は，意思決定者の判断にする．しかしこれでは，もし意思決定者が苦手な分野だったときに自動決定機の判断が好ましい時に対応できない．

- 貢献は次のようになる

    - 適応的rejection learningの定式化をおこなった．
    - 延期モデルが従来のrejection learningのモデルよりいいことを実験と理論を示した．
    - 延期することで精度と公平性を向上させ，ユーザ(意思決定者)と協力して公平で責任のある意思決定を行えることを可能にした．

準備

2.1 A Joint Decision-Making Framework

Figure 1: A larger decision system containing an automated model. When the model predicts, the system outputs the model's prediction; when the model says PASS, the system outputs the decision-maker's (DM's) prediction. Standard rejection learning considers the model stage, in isolation, as the system output, while learning-to-defer optimizes the model over the system output.

システム全体としては，二つの構成要素から成り立っている．1つ目は，学習させていくモデル(Model).2つ目は意思決定をしてくモデル(Decision makers, DM). 最終的な目標はModelのパラメータを得ることである．DMには人間のようなブラックボックスなモデルを想定してる．出力を出す流れとしては，まずModelがPASSするかどうかを決める，PASSした時はDMに問い合わせて出力を得る．PASSしない時はModelの出力をそのまま使う．

ここまでのモデルを定式化すると次のようになる.

$$P_{defer}(Y|X,Z) = \prod_i [P_M(Y_i=1|X_i)^{Y_i}(1-P_M(Y_i=1|X_i))^{1-Y_i}]^{(1-s_i|X_i)} \\ [P_D(Y_i=1|X_i,Z_i)^{Y_i}(1-P_D(Y_i=1|X_i,Z_i))^{1-Y_i}]^{(s_i|X_i)} \quad (1)$$

Xは入力データ，Yは出力，_Mは，Modelの出力，_Dは, DMのモデルの出力, sはPASSするかどうかのパラメタ(s=1のときPASSする．)

Model, DMの出力や最終的な出力，sの決定方法は次のようになる.

$$\hat{Y}_M = f(X) = P_M(Y=1|X) \in [0,1]; \quad \hat{Y}_D = h(X,Z) = P_D(Y=1|X,Z) \in [0,1] \\ \hat{Y} = (1-s)\hat{Y}_M + s\hat{Y}_D \in [0,1]; \quad s = g(X) \in \{0,1\} \quad (2)$$

この時の目標は,よいY_M,Sを学習していき，よい出力を得ることであるので，(1)を最大化する，つまり次の式を最小化していく.

$$\mathcal{L}_{defer}(Y,\hat{Y}_M,\hat{Y}_D,s) = -\log P_{defer}(Y|X,Z) = -\sum_i [(1-s_i)\ell(Y_i,\hat{Y}_{M,i}) + s_i\ell(Y_i,\hat{Y}_{D,i})] \quad (3)$$

fに微分可能な関数を持ってくると，勾配法で学習することができる．hはblack-boxを想定しているため，観測できない．sを決める関数gは，あとで出てくるModelとDMの混合率とベルヌーイ分布で決めたり，Modelの出力によって変えたりなどある.

ここでは，l(Y,p)はクロスエントロピー．またこのL_deferを最小化していくことをlearning to deferということにする.

提案手法

我々の手法は mixture-of-experts modelに影響を受けて作られている．我々の手法が異なる点として，DMが専門家として機能している.

$$\mathcal{L}_{defer}(Y,\hat{Y}_M,\hat{Y}_D,\pi;\theta) = \mathbb{E}_{s\sim Ber(\pi)}\mathcal{L}(Y,\hat{Y}_M,\hat{Y}_D,s;\theta) \\ = \sum_i \mathbb{E}_{s_i\sim Ber(\pi_i)}[(1-s_i)\ell(Y_i,\hat{Y}_{M,i};\theta) + s_i\ell(Y_i,\hat{Y}_{D,i})] \quad (9)$$

PASSするかどうかのパラメータsは，混合率πからベルヌーイ分布によって決定していく.

公平性を達成するために，正則化した公平性損失関数を用いる．式(9)に正則項を入れて精度と公平性のバランスを取る．αはバランスを取るための係数.

$$\mathcal{L}_{defer}(Y,\hat{Y}_M,\hat{Y}_D,\pi;\theta) = \mathbb{E}_{s\sim Ber(\pi)}[\mathcal{L}(Y,\hat{Y}_M,\hat{Y}_D,s;\theta) + \alpha_{fair}\mathcal{R}(Y,\hat{Y}_M,\hat{Y}_D,s) \quad (11)$$

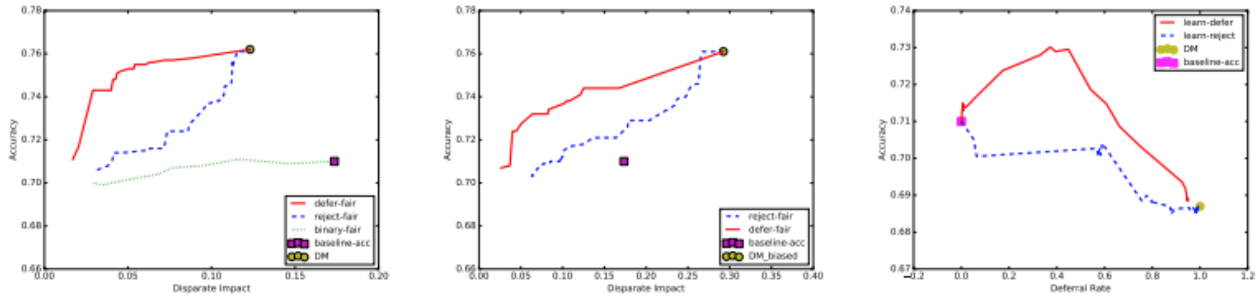
実験

3つのシナリオで実験を行う．DMに持ってくる人の性質でそれぞれ実験を行う.

1. High-accuracy DM: 公平性を無視して，精度が高いDM.
2. Highly-biased DM: 不公平さが強いDM.　たくさんのバイアスが残っているDM.
3. Inconsistent DM: 公平性を無視している.

使うデータセットはCOMPASとHeritage Healthの2つで行う．訓練するモデルとDMはそれぞれ二層の全結合ニューラルネットワークを用いて行う.

実験結果は下のようになった.



(a) COMPAS, High-Accuracy DM (b) COMPAS, Highly-Biased DM (c) COMPAS, Inconsistent DM

Figure 3: Comparing learning-to-defer, rejection learning and binary models. dataset only; Health dataset results in Appendix A. Each figure is a different DM scenario. In Figs. 3a and 3b, X-axis is fairness (lower is better); in Fig. 3c, X-axis is deferral rate. Y-axis is accuracy for all figures. Square is a baseline binary classifier, trained only to optimize accuracy; dashed line is fair rejection model; solid line is fair deferring model. Yellow circle is DM alone. In Fig. 3a, green dotted line is a binary model also optimizing fairness. Figs. 3a and 3b are hyperparameter sweep over $\gamma_{reject/defer}/\alpha_{fair}$; Fig. 3c sweeps $\gamma_{reject/defer}$ only, with $\alpha_{fair} = 0$ (for $\alpha_{fair} \geq 0$, see Appendix G).

横軸はDisparate Impactで，公平性の指標であり，小さい方が良い．縦軸はAccuracy.

High-Accuracy DM

Fig3-aについて，DMは，入力データ以外の情報(その人の経験，知識，外部リソース)を用いて判断を行うため，精度は高い．公平性と精度はトレードオフであるが，deferモデルでは，精度を落とさずに公平性を向上させることができている.

Highly-Biased DM

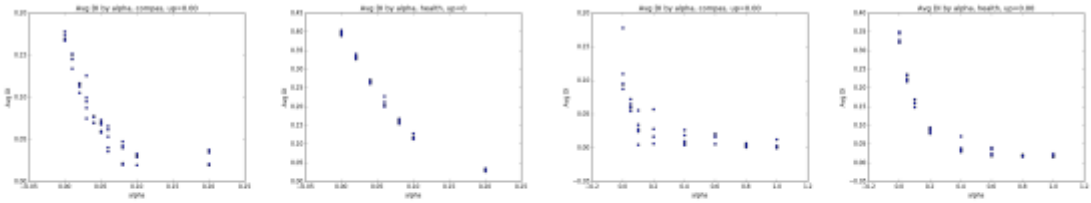
Fig3-bについて，DMがバイアスが多い状態であったとき，パラメータを調整してくことで，DMのDisparete Impactを減少させることができた.

Inconsistent DM

Fig3-cについて，横軸は, DMとModelの二つの出力の混合率の比であり，1のときDMをつかい，0のときDMを使わない．2つのモデルを混合するときに精度がいい出力を得ることができた.

正則化項でのFairnessの検証

実験結果は次のようになった.

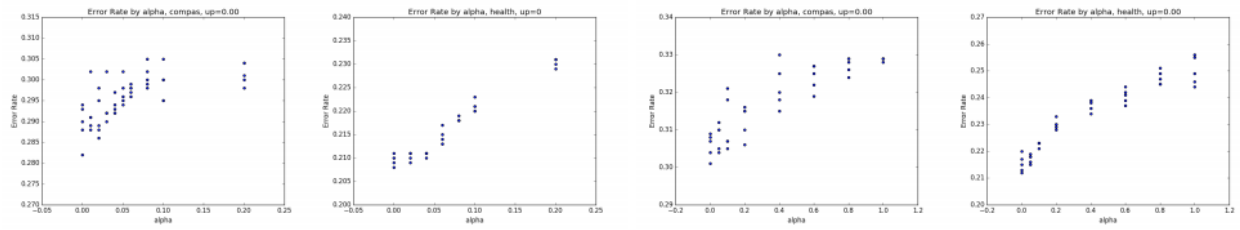


(a) COMPAS, MLP (b) Health, MLP (c) COMPAS, BNN (d) Health, BNN

Figure 7: Relationship of DI to α, the coefficient on the DI regularizer, 5 runs for each value of α. Two datasets, COMPAS and Health. Two learning algorithms, MLP and Bayesian weight uncertainty.

横軸が,正則化項の係数の値．縦軸が公平性の指標DIの値(小さいほどいい．) 係数の値を大きくするにつれて，DIの指標が小さくなっているので公平性が向上していると言える.

正則化項の係数とerror rateの関係は次のようになった.



(a) COMPAS, MLP (b) Health, MLP (c) COMPAS, BNN (d) Health, BNN

Figure 8: Relationship of error rate to α, the coefficient on the DI regularizer, 5 runs for each value of α. Two datasets, COMPAS and Health. Two learning algorithms, MLP and Bayesian weight uncertainty.

係数の値が大きくなるにつれ，誤差率は大きくなった.

結論

意思決定のプロセスの中に，延期する行動を入れた手法を提案した．大きなシステムの中に延期するという行動を与えるアルゴリズムを提案し，どれくらい公平になるかを示した.

所感

- 直接的に公平性が向上した理由がよくわからなかった，実験で公平性があがったのは, ただ正則化したおかげな気がする.
- この論文の手法は，能動学習にちかいけど，学習データを増やすことはしてないからちょっと違う？
- DMにクラウドソーシングを入れて公平なモデルを獲得できるなら使えそうな気がする．．．