Função de custo para aprendizado robusto à injustiça

Ygor Canalli

Orientador: Geraldo Zimbrão da Silva

01 de setembro de 2021

PESC/COPPE/UFRJ

Table of contents

Função correção de injustiça

Correção de função de custo

É possível criar robustez a ruído alterando a função de custo[1,2].

- Estima-se uma matriz de transição T capaz de descrever o ruído (NAR) de acordo com a classe
- A correção *forward* de uma função de custo $\ell()$ é definida como

$$\ell^{\rightarrow}(P(\widetilde{Y}|X)) = \ell(T^T P(\widetilde{Y}|X))$$

- A correção *backward* por sua vez é definida como

$$\ell^{\leftarrow}(P(\widetilde{Y}|X)) = T^{-1}\,\ell(P(\widetilde{Y}|X))$$

---

[1] Giorgio Patrini et al. *Loss Factorization, Weakly Supervised Learning and Label Noise Robustness*. Tech. rep. 2016. URL: http://proceedings.mlr.press/v48/patrini16.pdf.

[2] Giorgio Patrini et al. "Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach". In: July 2017, pp. 2233–2241. DOI: 10.1109/CVPR.2017.240.

Função correção de injustiça

Sejam

- X a matriz de atributos
- $\widetilde{Y}$ a matriz de classes previstas em formato *one hot encoding*
- S um atributo sensível em X
- $s_i \in S$ o conjunto dos grupos delimitados em S de tamanho $|S|$
- X_{S_i} um vetor binário indicando se cada instância em X pertence ao grupo s_i
- T_i a matriz de transição correspondente ao grupo $s_i \in S$
- ℓ uma função de custo

A função $\ell_S()$ de correção de injustiça sobre o atributo S é dada por

$$\ell_S(P(\widetilde{Y}|X)) = \sum_{i=1}^{|S|} \ell(T_i^T P(\widetilde{Y}|X))X_{S_i}$$

Estado da arte

Estado da arte

- Otimização multiobjetivo focando no *tradeoff*[3]
- Algoritmo personalizado de treinamento
- 3 Parâmetros de pesos para os objetivos

$$r = AUC - \alpha \cdot ADS - \beta \cdot AEOD - \gamma \cdot AOD$$

- Avaliação de quatro datasets, um deles inédito* no contexto de fairness
  - Adult income (gênero)
  - German credit (gênero e idade)
  - Hospital readmissions* (gênero)
  - Hospital expenditures (raça)

[3] Andrija Petrović et al. "Fair classification via Monte Carlo policy gradient method". In: *Engineering Applications of Artificial Intelligence* 104 (2021), p. 104398. ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2021.104398. URL: https://www.sciencedirect.com/science/article/pii/S0952197621002463.

Estado da arte - Income Dataset

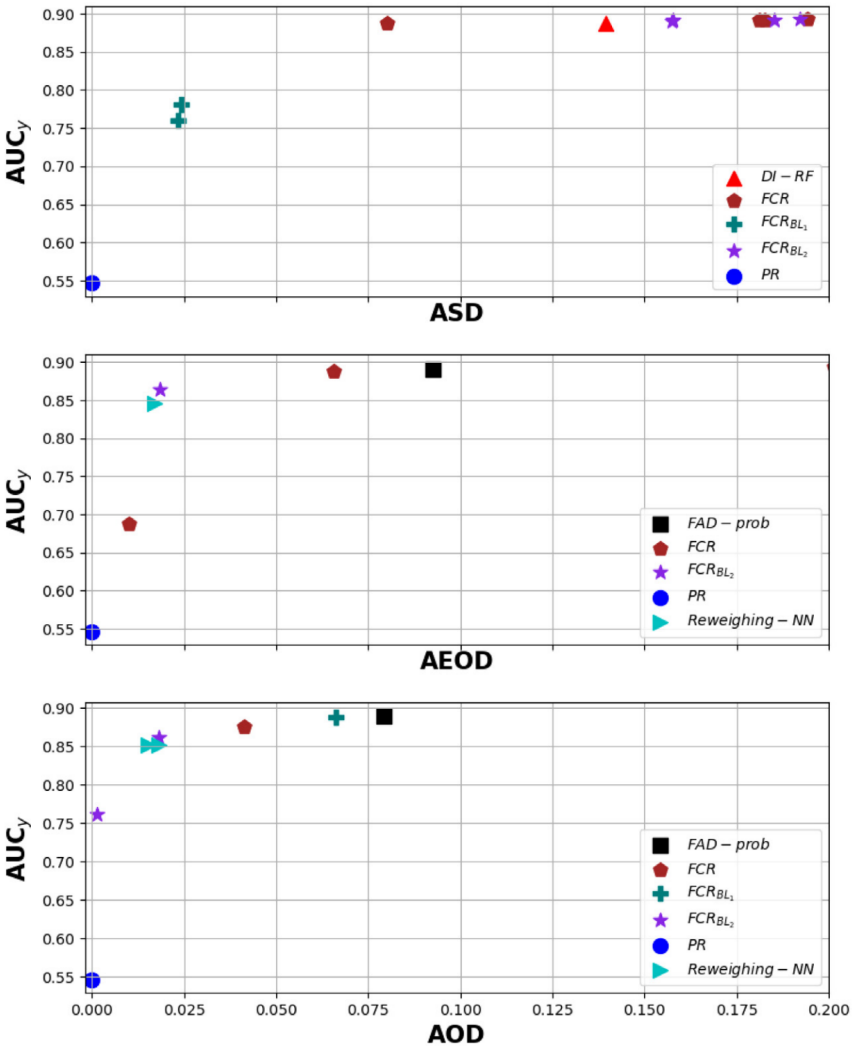

Fig. 4. Classification performance and fairness of the compared models as measured by **AUC$_y$** and **AOD**, **ASD** or **AEOD** on the *Adult income* dataset.

Estado da arte - German Dataset (sex)

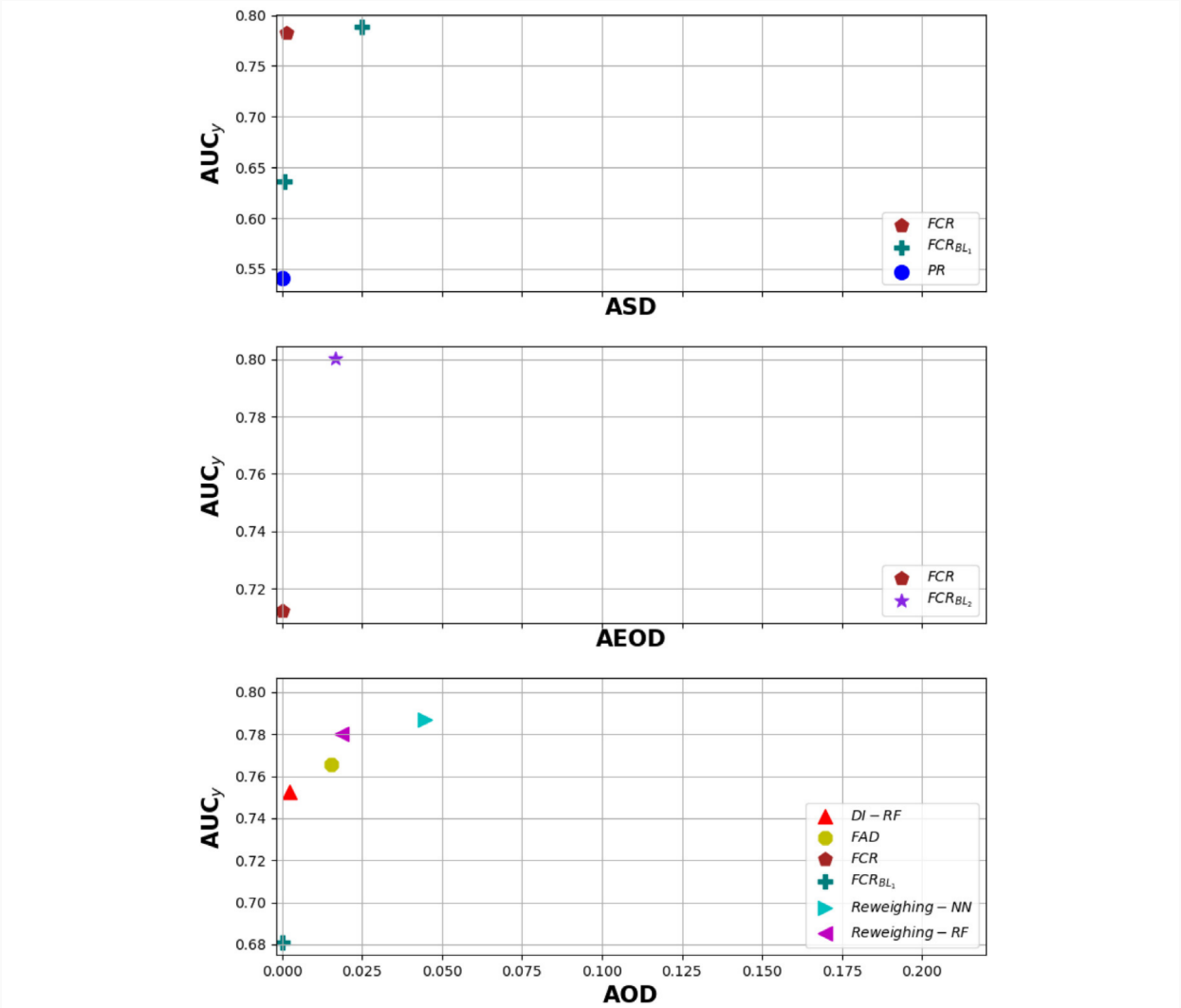

Fig. 7. Classification performance and fairness of the compared models as measured by **AUC$_y$** and **AOD**, **ASD** or **AEOD** on the *German sex* dataset.

Estado da arte - German Dataset (age)

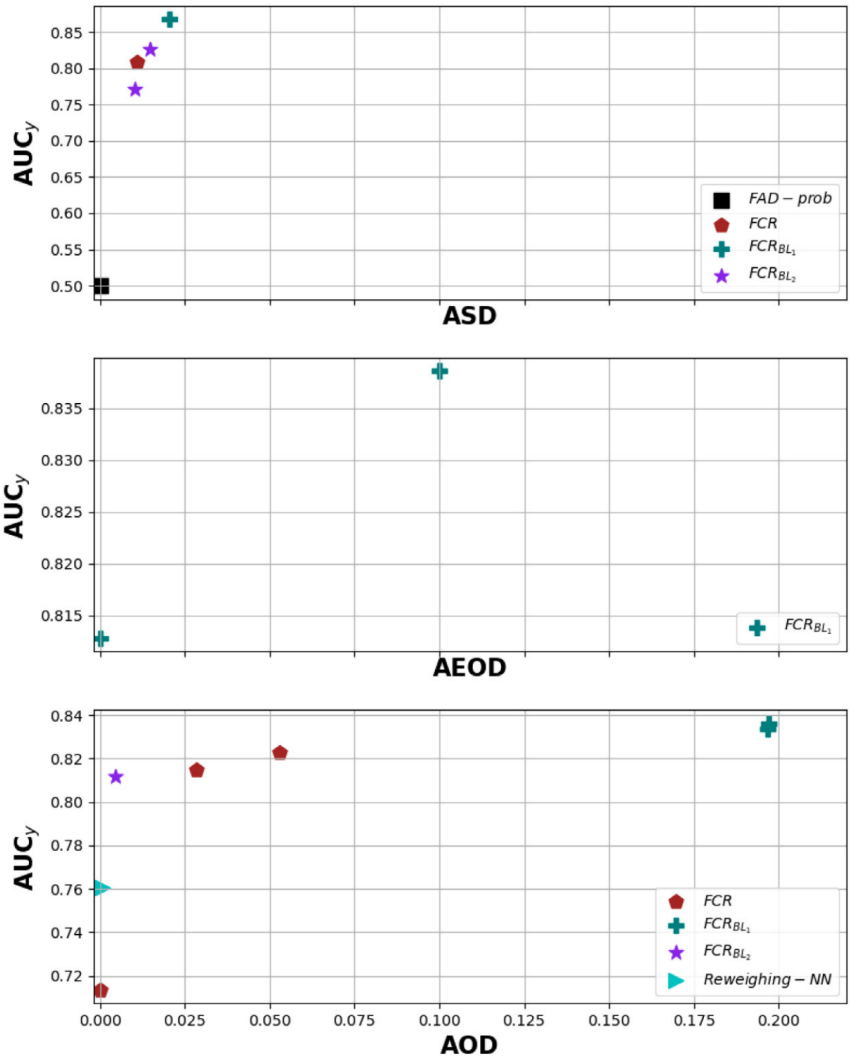

Fig. 6. Classification performance and fairness of the compared models as measured by **AUC$_v$** and **AOD**, **ASD** or **AEOD** on the *German age* dataset.

Metodologia

Datasets

- Experimentos nos datasets *census income*
  - Classificação binária
  - Gênero como atributo sensível
    - Protegido: Feminino
    - Privilegiado: Masculino
- Experimentos nos datasets *german credit*
  - Classificação binária
  - Gênero como atributo sensível
    - Protegido: Feminino
    - Privilegiado: Masculino
  - Idade como atributo sensível
    - Protegido: <=25 anos
    - Privilegiado: >25 anos

Metodologia

- Uma matriz de transição para cada grupo
- Testar diferentes combinações de matrizes de transição
- Medir AUC
  - vantagem sobre acurácia em classes desbalanceadas
- Medir métricas de injustiça
  - absolute statistical parity difference (ASD)
  - absolute equal opportunity difference (AEOD)
  - average odds difference (AOD)

Parâmetros de na matriz de transição

- Para cada matriz de transição basta definir os valores da diagonal secundária, de forma que cada linha some 1.
- A taxa r é o fator de incerteza da previsão positiva, com efeito prático de rebaixar a classificação
- A taxa p é o fator de incerteza da previsão negativa, com efeito prático de promover a classificação

$$T = \begin{bmatrix} 1-r & r \\ p & 1-p \end{bmatrix}$$

Busca por parâmetros

- Tomamos uma matriz de transição T_{priv} para a grupo privilegiado e T_{prot} para o protegido
- Buscamos combinações r_{priv}, p_{priv}, r_{prot} e p_{prot} com algoritmos de busca
- Minimizar a métrica de injustiça
- Eliminação de soluções com desempenho abaixo de limiar estabelecido

$$T_{priv} = \begin{bmatrix} 1 - r_{priv} & r_{priv} \\ p_{priv} & 1 - p_{priv} \end{bmatrix}, \quad T_{prot} = \begin{bmatrix} 1 - r_{prot} & r_{prot} \\ p_{prot} & 1 - p_{prot} \end{bmatrix}$$

Busca por parâmetros com algoritmo genético

- Genes: $[r_{priv}, p_{priv}, r_{prot}, p_{prot}]$
- Espaço de busca: $0.0, 0.01, 0.02, \ldots, 0.98, 0.99, 1.0$
- Aptidão: métrica de injustiça no conjunto de teste
  - Arquitetura

$$\begin{aligned}\text{Input} &\rightarrow \text{Dropout}(0.2)\\ &\rightarrow \text{Dense}(32) \rightarrow \text{ReLU}\\ &\rightarrow \text{Dense}(64) \rightarrow \text{ReLU}\\ &\rightarrow \text{Dense}(32) \rightarrow \text{ReLU}\\ &\rightarrow \text{Dense}(2) \rightarrow \text{Softmax} \rightarrow \text{Output}\end{aligned}$$

  - Divisão: Treino/Validação/Teste
  - 10 épocas
  - Se $AUC < limiar$: aptidão é 0
  - Se $AUC >= limiar$: aptidão é $1/metrica$

Busca por parâmetros com algoritmo genético

- População por geração: 8
- Selecionados por geração: 4
- Técnica de seleção: steady state
- Primeira geração: aleatória
- Taxa de mutação: 10%
- Elitismo: 2 melhores soluções
- Gerações: 50

Resultados

Resultados - Income Dataset

r_{priv}	p_{priv}	r_{prot}	p_{prot}	AOD	std	AUC	std
0.66	0.08	0.07	0.15	0.00068	0	0.85514	0
0.64	0	0	0	0.00637	0.00339	0.8002	0.01124
0.59	0	0	0.17	0.0074	0.00569	0.81947	0.0059
0.56	0.23	0.07	0.69	0.00761	0.00472	0.86854	0.00325
0.64	0.3	0.07	0.69	0.00764	0.00287	0.84599	0.00781
0.59	0	0	0.29	0.00981	0.00689	0.8154	0.00633
0.59	0	0	0.1	0.01057	0.0055	0.82392	0.00519
0.59	0.08	0.13	0.55	0.01162	0.00428	0.88348	0.00334
0.29	0.33	0	0.89	0.01246	0.00778	0.85321	0.00246
0.59	0	0	0.05	0.01281	0.00409	0.81965	0.00568

Table 1: Melhores resultados de AOD no Income Dataset

Estado da arte - Income Dataset

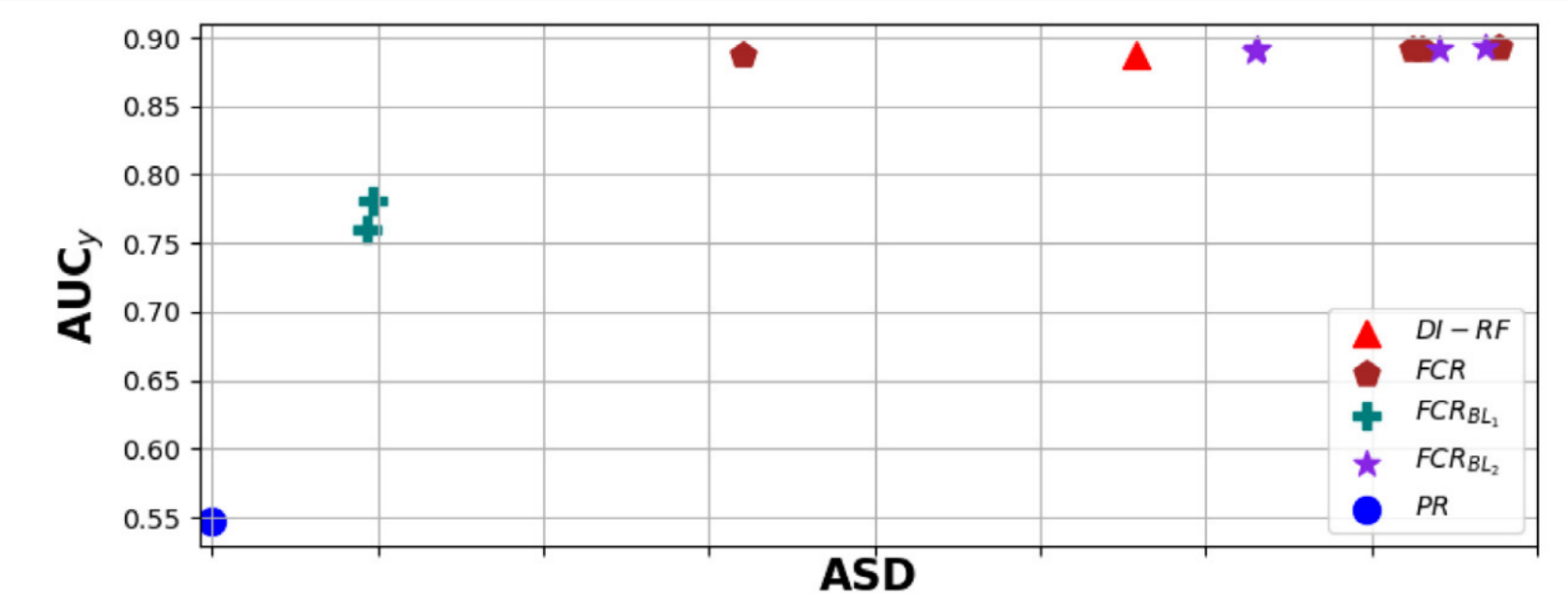

Figure 1: Estado da arte do AOD no Income Dataset

Resultados - German Dataset (sex)

r_{priv}	p_{priv}	r_{prot}	p_{prot}	AOD	std	AUC	std
0.18	0.24	0.13	0.28	0.01471	0.00368	0.73212	0.00454
0.29	0.39	0.14	0.28	0.019	0.01777	0.71992	0.0012
0.29	0.24	0.13	0.18	0.02037	0.02138	0.73478	0.01464
0.41	0.3	0.14	0.28	0.02512	0.00797	0.74253	0.00299
0.07	0.24	0.67	0.49	0.02819	0.02696	0.70617	0.01351
0.48	0.3	0.41	0.49	0.03094	0.01952	0.72411	0.01419
0.24	0.24	0.13	0.28	0.03248	0.00674	0.74468	0.00371
0.07	0.24	0.42	0.49	0.03248	0.03658	0.72339	0.01724
0.29	0.3	0.14	0.28	0.03309	0.0049	0.7259	0.00144
0.47	0.39	0.25	0.28	0.03411	0.02618	0.72216	0.00889

Table 2: Melhores resultados de AOD no German Dataset (sex)

Estado da arte - German Dataset (sex)

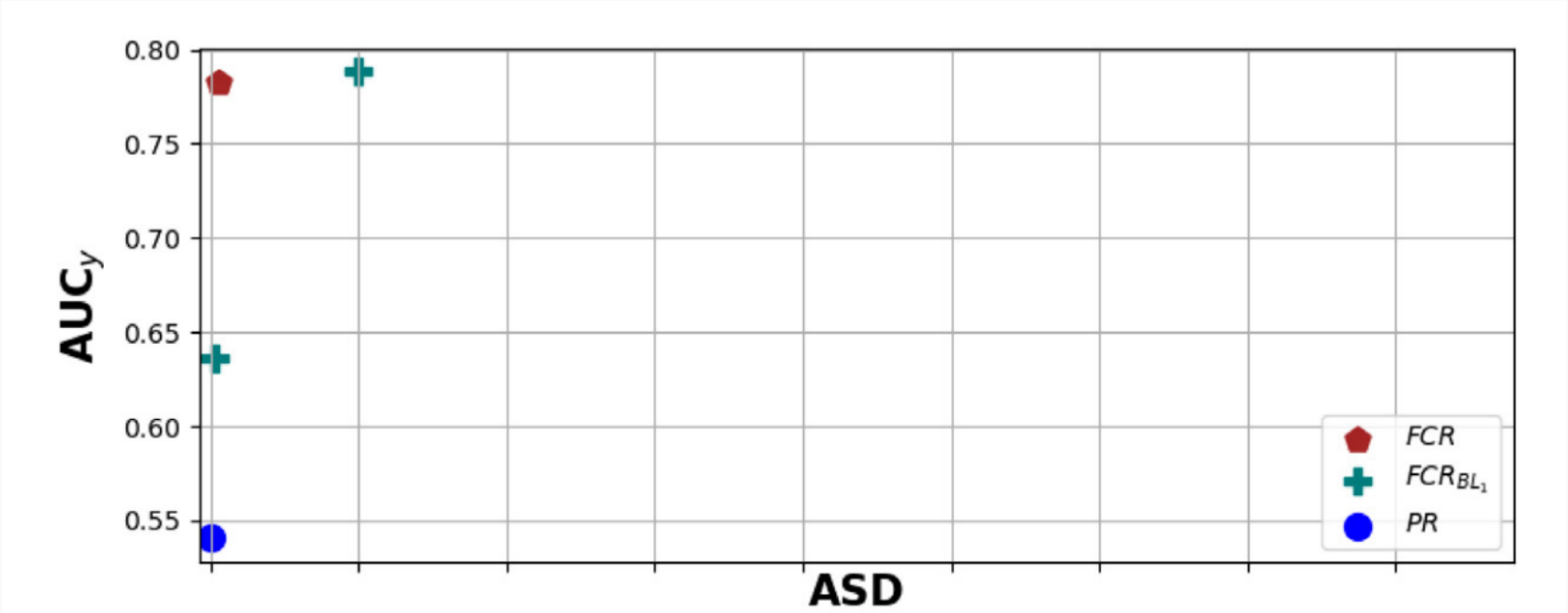

Figure 2: Estado da arte do AOD no German Dataset (sex)

Resultados - German Dataset (age)

r_{priv}	p_{priv}	r_{prot}	p_{prot}	AOD	std	AUC	std
0.45	0.42	0.06	0.35	0.0214	0.01399	0.71143	0.01173
0.46	0.06	0.25	0.18	0.02419	0.02047	0.73643	0.00072
0.45	0.42	0.62	0.35	0.02481	0.02481	0.70043	0.01399
0.32	0.3	0.13	0.18	0.02492	0.01498	0.72865	0.00751
0.15	0.23	0.2	0.19	0.02605	0.01923	0.74133	0.00993
0.36	0.37	0.13	0.3	0.0273	0.02012	0.72626	0.00966
0.22	0.21	0.17	0.18	0.02776	0.00912	0.7314	0.0061
0.32	0.33	0.13	0.18	0.02931	0.02528	0.7277	0.01151
0.36	0.06	0.13	0.15	0.0304	0.01985	0.74384	0.00813
0.17	0.23	0.06	0.19	0.0304	0.00124	0.73116	0.01913

Table 3: Melhores resultados de AOD no German Dataset (age)

Estado da arte - German Dataset (age)

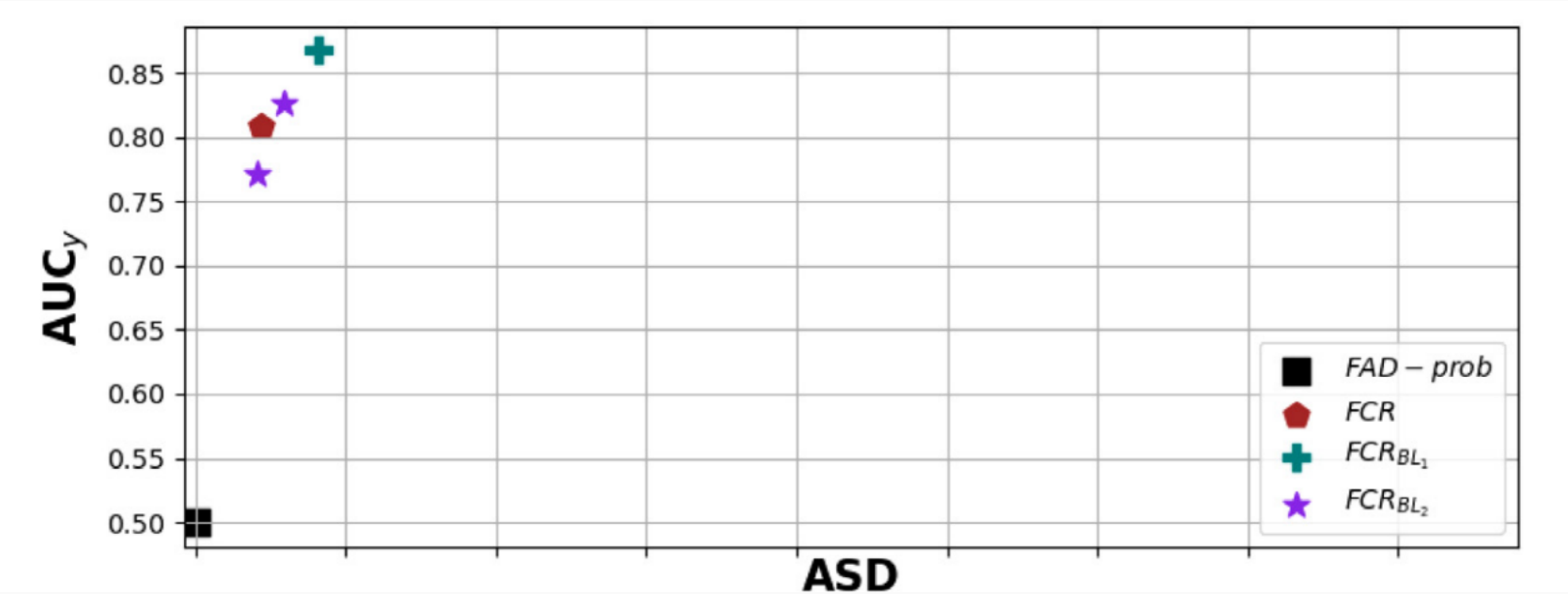

Figure 3: Estado da arte do AOD no German Dataset (age)