Fair Representation for Safe Artificial Intelligence via Adversarial Learning of Unbiased Information Bottleneck

February 07, 2020

Jin-Young Kim and Sung-Bae Cho
Yonsei University, Seoul, South Korea
http://sclab.yonsei.ac.kr
Outline

- Motivation
- Related works
- Proposed method
- Experiments
- Discussion
Background

• Algorithmic bias
  – Increased difference btw. predicted labels according to protected features
    (Koene, 2017; Mehrabi, 2019)
  – $|p_{data}(y|a = 1, x) - p_{data}(y|a = 0, x)| < |f(y|a = 1, x) - f(y|a = 0, x)|$
    • $x$: data ; $y$: label ; $a$: protected feature
  – Biased AI model $\rightarrow$ adverse in specific group $\rightarrow$ untrustworthy

• Approaches to solve it (Calmon et al., 2017)
  – Pre-processing
    • Training data (Hajian, 2013; Calmon et al., 2017)
  – In-processing
    • Learning algorithm (Kamishima, 2011; Fish, 2016; Zafar, 2016)
  – Post-processing
    • Ensuing decisions themselves (Hardt et al., 2016)
Pre-processing Method to Reduce Bias

• Summary of related works in pre-processing approach

<table>
<thead>
<tr>
<th>Category</th>
<th>Authors</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data generation</td>
<td>Iosifidis et al., 2018</td>
<td>• Elimination of bias itself</td>
<td>• Distortion of original data’s distribution</td>
</tr>
<tr>
<td></td>
<td>Zhao et al., 2018</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Xu et al., 2019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fair representation</td>
<td>Louizos, 2015</td>
<td>• Effective unbiased feature extraction</td>
<td>• Loss of data’s features</td>
</tr>
<tr>
<td></td>
<td>Amini et al., 2019</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Celis and Rao, 2019</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• Limitation and solution
  – Inconsistency btw. processes
    • 1. Reduce bias in generated data (or fair representation)
    • 2. Maintain features in them
  – Separated representation
    • z: contains all of features of data
    • r: z without information on protected features
Diagram Pre-processing Method

- **Data generation** (Iosifidis, 2013; Zhao, 2018; Xu, 2019)

  ![Diagram](image)

- **Fair representation** (Louizos, 2015; Amini, 2019; Celis, 2019)

  ![Diagram](image)

- **Problems**

  ![Diagram](image)

  Contains all features  Inconsistency  Remove protected feature
Proposed Fair Representation

- \( \max I(Z, X) - \beta I(Z, A) \)
  - \( Z \): data representation ; \( X \): data ; \( A \): p.f.
  - \( \max I(Z, X) \): maintain features of data
  - \( \max[-I(Z, A)] \): remove info. on p.f. in representation
  - Intractable formula

- Variational mutual information maximization
  - Separated latent space
    - \( z \): contains all of features of data
    - \( r \): \( z \) without information on protected features
  - Fair representation \( r \)

\[
I(Z, X) = \mathbb{E}_{z \sim q(z|x)} \left[ \mathbb{E}_{x' \sim P(x|z)} \left[ \log P(x'|z) \right] \right] + H(x)
\]
\[
\geq \mathbb{E}_{z \sim q(z|x)} \left[ \mathbb{E}_{x' \sim P(x|z)} \left[ \log Q(x'|z) \right] \right] + H(x)
\]
\[
= \mathbb{E}_{x' \sim P(x'), z \sim q(z|x)} \left[ \log Q(x'|z) \right] + H(x)
\]

\[
I(Z, A)
\]
\[
= \int p(a|z)p(z) \log \frac{p(z|a)}{p(z)}
\]
\[
= \int p(a|z)p(z) \log p(a|z) - \int p(z) \log p(z)
\]
\[
\leq \int p(a|z)p(z) \log p(a|z) - \int p(z) \log m(z)
\]
\[
= \int p(a|z)p(z) \log \frac{p(z|a)}{m(z)}
\]
\[
I\left( Z, D(t(Z)) \right) \leq I(Z, t(Z)) \leq \int t(r|z)p(z) \log \frac{t(r|z)}{s(r)}
\]
Training Scheme

- Overall architecture

![Diagram of training scheme]

- Phase 1
  - Train discriminator
  - To classify actual protected attr. $a$ from latent space well

- Phase 2
  - Train encoder and decoder
  - To disturb classifying protected attr. well from latent space
  - With random protected attr. $\tilde{a}$
• Overall architecture

• Phase 1
  – Train discriminator
  – To classify actual protected attr. \( a \) from latent space

• Phase 2
  – Train encoder and decoder
  – To disturb classifying protected attr. well from latent space
  – With random protected attr. \( \tilde{a} \)
Training Scheme (2)

- Overall architecture

![Diagram of overall architecture]

- Phase 1
  - Train discriminator
  - To classify actual protected attr. \( \alpha \) from latent space well

- Phase 2
  - Train encoder and decoder
  - To disturb classifying protected attr. from latent space
  - With random protected attr. \( \tilde{\alpha} \)

Proposed Method
Training Scheme

- Overall architecture

```
 x ----| Encoder | z ----| Feature extractor | Classifier | y
  |        r |        |            |            |
  | Decoder |

\[ U(0,1) \]
\[ a \quad \bar{a} \]
\[ \hat{a} \quad 0 \quad 1 \quad pf_1 \]
\[ 1 \quad 1 \quad pf_2 \]
```

- Training results
  - \( z \): contains data \( x \)'s features to be reconstructed as \( \hat{x} \)
  - \( r \): exploits \( z \)'s information except protected feature \( a \)
    - Bias is removed by adversarial learning of \( t \) and discriminator
### Experimental Settings

- **Dataset**

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of instance</th>
<th>X</th>
<th>S</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCI Adults</td>
<td>48,842</td>
<td>User profile</td>
<td>Gender</td>
<td>Income (&gt; 50K, ≤ 50K)</td>
</tr>
<tr>
<td>(Kohavi, 1996)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UCI Census</td>
<td>299,285</td>
<td>User profile</td>
<td>Gender</td>
<td>Income (&gt; 50K, ≤ 50K)</td>
</tr>
<tr>
<td>(Lane, 2000)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COMPAS</td>
<td>18,610</td>
<td>Criminal info.</td>
<td>Race</td>
<td>Risk score (&lt; 5, ≥ 5)</td>
</tr>
<tr>
<td>(Larson, 2016)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Metrics** (Dwork, 2012; Hardt, 2016; Rachel, 2018; Mehrabi, 2019)
  - **Equalized Opportunity**
    - Difference of TPR by protected features
    - Are both privileged and unprivileged groups provided for their right?
  - **Equalized Odds**
    - Sum of differences of TPR and FPR by protected features
    - Do privileged and unprivileged groups receive the same benefits?
## Results for Reduced Bias

<table>
<thead>
<tr>
<th>Equalized opportunity</th>
<th></th>
<th>Equalized odds</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Adult dataset</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>0.0851</td>
<td>0.1987</td>
<td>0.0576</td>
</tr>
<tr>
<td>Std. dev.</td>
<td>0.0132</td>
<td>0.0328</td>
<td>0.0274</td>
</tr>
<tr>
<td>p-value</td>
<td><strong>1.20×10^{-11}</strong></td>
<td><strong>6.12×10^{-9}</strong></td>
<td><strong>2.02×10^{-4}</strong></td>
</tr>
<tr>
<td><strong>Census dataset</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>0.1961</td>
<td>0.2448</td>
<td>0.1884</td>
</tr>
<tr>
<td>Std. dev.</td>
<td>0.0190</td>
<td>0.0770</td>
<td>0.0600</td>
</tr>
<tr>
<td>p-value</td>
<td><strong>5.83×10^{-2}</strong></td>
<td><strong>1.09×10^{-2}</strong></td>
<td><strong>9.91×10^{-2}</strong></td>
</tr>
<tr>
<td><strong>Compas dataset</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>0.2891</td>
<td>0.3038</td>
<td>0.1184</td>
</tr>
<tr>
<td>Std. dev.</td>
<td>0.0239</td>
<td>0.0250</td>
<td>0.0959</td>
</tr>
<tr>
<td>p-value</td>
<td><strong>1.40×10^{-7}</strong></td>
<td><strong>3.55×10^{-9}</strong></td>
<td><strong>2.01×10^{-2}</strong></td>
</tr>
</tbody>
</table>
Bias-Accuracy Tradeoff

- Compared to VFAE, **5% less accuracy** and **93% less discrimination**
- Compared to Edwards’ model, **3% less accuracy** and **83% less discrimination**
Discussion

• Summary
  – Propose a fair representation learning method to reduce discrepancy results across different protected features
  – Address the limitation of previous studies and solve it by separating a latent space
  – Verify the proposed model with various datasets and metrics
  – Minimize the tradeoff btw. accuracy and discrimination

• Future works
  – Reduce a bias on multiple protected features
  – Control a degree of discrimination in data representation
  – Construct demo system to visualize the process of reducing bias