iFair: Learning Individually Fair Representations for Algorithmic Decision Making

Preethi Lahoti\textsuperscript{1}  Krishna P. Gummadi\textsuperscript{2}  Gerhard Weikum\textsuperscript{1}

\textsuperscript{1} Max Planck Institute for Informatics

\textsuperscript{2} Max Planck Institute for Software Systems
Fairness in algorithm decision making

- Input data $X$ (user records)
- Learning algorithm (e.g., classification, recommendation)
- Users
  - Non-protected attributes (e.g., qualifications or education)
  - Protected attributes (e.g., race or gender)
- Decision making system (e.g., loans, visas, job interviews, parole request)
COMPAS – A commercial Risk Prediction Algorithm

Risk Score higher for African-American defendants who did not re-offend (ProPublica / Machine Bias)

Existing Group Fairness Notions:
Equalizing Group Level Statistics (e.g., Equality of Opportunity, Parity)

<table>
<thead>
<tr>
<th>Prediction Fails Differently for Black Defendants</th>
<th>WHITE</th>
<th>AFRICAN AMERICAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled Higher Risk, But Didn’t Re-Offend</td>
<td>23.5%</td>
<td>44.9%</td>
</tr>
<tr>
<td>Labeled Lower Risk, Yet Did Re-Offend</td>
<td>47.7%</td>
<td>28.0%</td>
</tr>
</tbody>
</table>

False Positive Rate
False Negative Rate
Group Fairness is Insufficient

Similarly qualified individuals receive dissimilar outcomes.

Group fairness notions don’t cover individual level unfairness.

Top K results satisfy group fairness (fraction of protected in top 10).

TABLE II: Top k results on www.xing.com (Jan 2017) for the job search query “Brand Strategist”.

<table>
<thead>
<tr>
<th>Search query</th>
<th>Work experience</th>
<th>Education experience</th>
<th>Profile views</th>
<th>Candidate</th>
<th>Xing ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brand Strategist</td>
<td>146</td>
<td>57</td>
<td>12992</td>
<td>male</td>
<td>1</td>
</tr>
<tr>
<td>Brand Strategist</td>
<td>327</td>
<td>0</td>
<td>4715</td>
<td>female</td>
<td>2</td>
</tr>
<tr>
<td>Brand Strategist</td>
<td>502</td>
<td>74</td>
<td>6978</td>
<td>male</td>
<td>3</td>
</tr>
<tr>
<td>Brand Strategist</td>
<td>444</td>
<td>56</td>
<td>1504</td>
<td>female</td>
<td>4</td>
</tr>
<tr>
<td>Brand Strategist</td>
<td>139</td>
<td>25</td>
<td>63</td>
<td>male</td>
<td>5</td>
</tr>
<tr>
<td>Brand Strategist</td>
<td>110</td>
<td>65</td>
<td>3479</td>
<td>female</td>
<td>6</td>
</tr>
<tr>
<td>Brand Strategist</td>
<td>12</td>
<td>73</td>
<td>846</td>
<td>male</td>
<td>7</td>
</tr>
<tr>
<td>Brand Strategist</td>
<td>99</td>
<td>41</td>
<td>3019</td>
<td>male</td>
<td>8</td>
</tr>
<tr>
<td>Brand Strategist</td>
<td>42</td>
<td>51</td>
<td>1359</td>
<td>female</td>
<td>9</td>
</tr>
<tr>
<td>Brand Strategist</td>
<td>220</td>
<td>102</td>
<td>17186</td>
<td>female</td>
<td>10</td>
</tr>
</tbody>
</table>
Motivation

Fairness Challenges
- Fairness as Group Property
- Fairness of Individuals Ignored

Generalizability Challenges
- Assume binary classification task
- Fairness concerns ubiquitous across ML tasks

Practical Challenges
- Access to protected attribute required
  - Infeasible due to privacy and legal concerns
  - Assumptions about protected attribute
    - Single Attribute | Binary Value | Fixed Value

Can we provide Individual Fairness while keeping utility high?

Can we be Application-Agnostic?

No explicit access to protected attribute?

Can we eliminate assumptions?
Learning Individually Fair Representations

• Remove protected attributes
  • Correlated attributes
• Identify and remove all correlations
  • Significant Loss in utility
• Utility vs Fairness trade-off

• Individual Fair Representation
  • Similar individuals should be treated similarly [Dwork et al 2012]
  • Given a fairness-aware distance metric
  • Make similar individuals indistinguishable in learned representation

Retain as much utility as possible while obfuscating information about protected attribute
Problem Formulation

Learn a Low-Rank Representation

• Unsupervised probabilistic soft clustering
• Learn k – representative prototypes that summarize data
• Prototypes: elements of the data space that represents a group of elements

Retain Utility

• Minimize reconstruction error

Individual Fairness

• Obfuscate Protected Information
• Make similar individuals indistinguishable in learned representation
• Similar individuals: according to some fairness-aware distance metric over features
Unsupervised Probabilistic Soft Clustering

Goal: Learn \( k \) prototypes \( v_k \) that summarize data

Probability that \( x_n \) maps to \( k \)-th prototype \( v_k \)

\[
U_{n,k} = Pr(Z = k \mid x_n) = \frac{\exp(-d(x_n,v_k))}{\sum_{j \in P} \exp(-d(x_n,v_j))}
\]

Low-rank representation: marginalize over \( k \) prototypes \( (X \cong UV^T =: \tilde{X}) \)

\[
\tilde{x}_n = \phi(x_n) = \sum_k U_{n,k} \cdot v_k
\]

Utility objective: minimize data loss \( L_x(X, \tilde{X}) \)
Fairness-aware Distance Metric

Distance Function: Minkowski $p$-metrics

$$d(x_i, x_j) = \left[ \sum_{m=1}^{M} \alpha^m (x_i^m - x_j^m)^p \right]^{1/p}$$

Goal: Learn feature weight vector $\alpha$

Fairness Intuition:
- Set $\alpha$ zero for protected attributes
- Learn near zero weights for correlated attributes

Fairness Objective: preserve fairness-aware distances

$$L_{fair}(X, \tilde{X}) = \sum_{i,j \in N} \left( d(x_i, x_j) - \tilde{d}(\tilde{x}_i, \tilde{x}_j) \right)^2$$

$d(.)$ applies to original data records and prototypes alike
$\tilde{d}(.)$ we use simple Euclidean distance
Optimization Problem

Learn a low-rank representation \( X \approx UV^T =: \tilde{X} \)

- Utility ~ Minimize data loss
- Individual Fairness ~ preserve fairness-aware distances

Objective:
\[
L = \lambda \cdot L_{util}(X, \tilde{X}) + \mu \cdot L_{fair}(X, \tilde{X})
\]

Gradient Descent Optimization:
- L-BFGS algorithm
- Model Parameters:
  - Learn k prototype vectors \( \nu_k \)
  - Learn feature weight vector \( \alpha \)
- Hyperparameters: \( \lambda, \mu \) and \( k \)
Experiments

• Does iFair reconcile utility vs individual fairness?

• Application-agnostic
  • Learn fair representation | apply out-of-the-box ML models
  • Classification task :
    • Compas - *recidivism risk* prediction | protected: *race*
    • Census income – *income* prediction | protected: *gender*
    • Credit scores – *credit risk* prediction | protected: *age*
  • Learning-to-rank task:
    • Xing – *candidate ranking* by qualification | protected: *gender*
    • Airbnb – *housing ranking* by rating/price | protected: *gender of host*

• Inspect group fairness from empirical perspective
Utility vs Individual Fairness

Classification task

Ranking task

Compas

Airbnb

AUC

MAP

yNN

Full Data
Masked Data
SVD
SVD-masked
LFR
FA*IR
iFair
Utility vs Individual Fairness

Classification task

Ranking task

Original Data

AUC vs yNN for Compas

MAP vs yNN for Airbnb

Legend:
- Full Data
- Masked Data
- SVD
- SVD-masked
- LFR
- FA*IR
- iFair
Utility vs Individual Fairness

Classification task

Ranking task

Compas

AUC

0.65

0.60

0.55

0.50

0.8

0.9

1.0

yNN

Airbnb

MAP

0.7

0.6

0.5

0.4

0.75

0.80

0.85

0.90

yNN

Naïve Baselines

Full Data

Masked Data

SVD

SVD-masked

LFR

FA*IR

iFair
Utility vs Individual Fairness

Classification task

Ranking task

State-of-the-art
Utility vs Individual Fairness

Classification task

Ranking task

\( y\text{NN} \) measures consistency of outcome of an individual with its \( p \)-nearest neighbors

AUC and MAP are measures of utility

Proposed Approach

16
Utility vs Individual Fairness

Classification task

Ranking task

Proposed Approach

iFair achieves the best utility-fairness tradeoff
Obfuscating Protected Information

Adversarial Classification task
Input: Learned fair representation
Task: predict protected attribute

iFair outperforms state-of-the-art in obfuscating protected information
Relation to Group fairness

Parity: $P(\hat{y} = 1|S = 0) = P(\hat{y} = 1|S = 0)$
EqOpp: $P(\hat{y} = 1|S = 0, Y = 1) = P(\hat{y} = 1|S = 0, Y = 1)$

Obfuscating protected information indirectly helps in group fairness
Conclusions

• iFair – unsupervised probabilistic clustering for individually fair representations
  • First approach to address individual fairness in learning-to-rank
  • Outperforms state-of-the art on classification
• Generic and versatile
  • application-agnostic (e.g., classification, learning-to-rank)
  • Makes no assumptions about protected attributes
• Pre-processing approach
  • fairness as a property of the dataset
  • Post-processing for group fairness can still be applied
• Visit poster # tonight for discussion
Additional slides
Empirical Observation on Group Fairness

- Measure = EqOpp
- Measure = Parity
- Measure = % Protected in top 10
Applying post-processing to iFair representations

![Graphs showing the effect of post-processing on Xing and Airbnb datasets.](image-url)