Abolfazl Asudeh
Fall 2020
9/29/2020

InDeX Lab
Innovative Data eXploration Laboratory

CS 594: Preprocess Interventions to Achieve Fairness

Interventions to achieve responsible scoring

- Pre-process Techniques
- In-process Techniques (Scoring Algorithm Modification)
- Post-process techniques

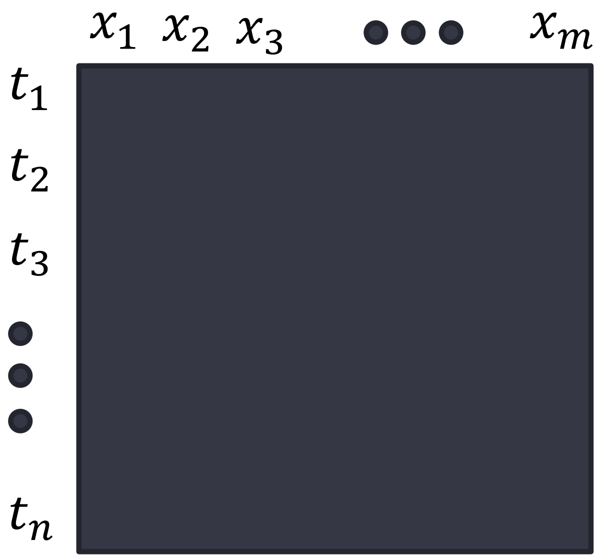
[*] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In FAT*, 2019.

Pre-processing and Data Investigation

Reminder: Bias in rows v.s. columns

- Bias in rows: Not enough representative tuples from minority (sub)groups

- Bias in columns: Features are biased (correlated) with sensitive attributes



Prelim. thoughts?

Data preprocessing techniques for classification without discrimination

Faisal Kamiran and Toon Calders

Knowledge and Information Systems 33.1 (2012): 1-33

- Preprocessing techniques for discrimination-free evaluation
    1. Suppression of Sensitive Attribute
    2. Massaging the dataset
    3. Reweighting
    4. Resampling
- Binary target variable, one binary sensitive attribute

Suppression of Sensitive Attribute

- To remove the attributes that highly correlate with the sensitive attribute.

Massaging the dataset

- Change the label of some tuples in the training data, in order to minimize the discrimination.
- Considers a subset of data from the minority group as promotion candidates:
  - Change the labels of promotion candidates from – to +
- a subset of data from the majority group as demotion candidates:
  - Change the labels of demotion candidate from + to –
- Which labels to select?
    - Learn a classifier; rank the tuples based on their probability of having positive labels
    - Select the top-k of minority (for promotion) and the bottom-k of majority (for demotion)

Notes

Reweighting

- Instead of changing the labels, each tuple in the training data is assigned with a weight

- This works for all the methods for which tuple weights can be used as frequency counts

1. For each of the group-value combinations, it computes the probability if independence would hold.
2. The weight of a group is ratio b/w its probability under independence and it actual probability in the dataset

Reweighting, Example

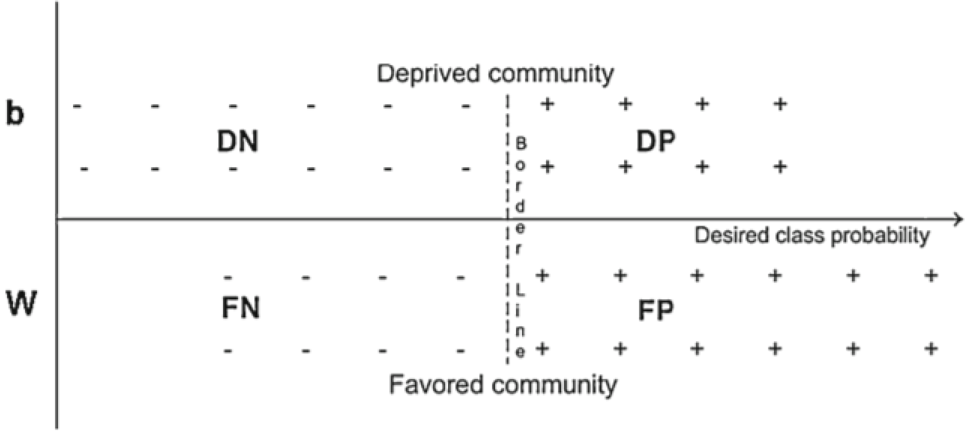
Compute the weight for (female,+)

Sex	Ethnicity	Highest degree	Job type	Class
M	Native	H. school	Board	+
M	Native	Univ.	Board	+
M	Native	H. school	Board	+
M	Non-nat.	H. school	Healthcare	+
M	Non-nat.	Univ.	Healthcare	–
F	Non-nat.	Univ.	Education	–
F	Native	H. school	Education	–
F	Native	None	Healthcare	+
F	Non-nat.	Univ.	Education	–
F	Native	H. school	Board	+

Resampling

- Calculate the sample size for each of the group-value combination.
  - e.g.: {male reject, male accept, female reject, female accept}

Sample size	DP	DN	FP	FN
Actual	8	12	12	8
Expected	10	10	10	10



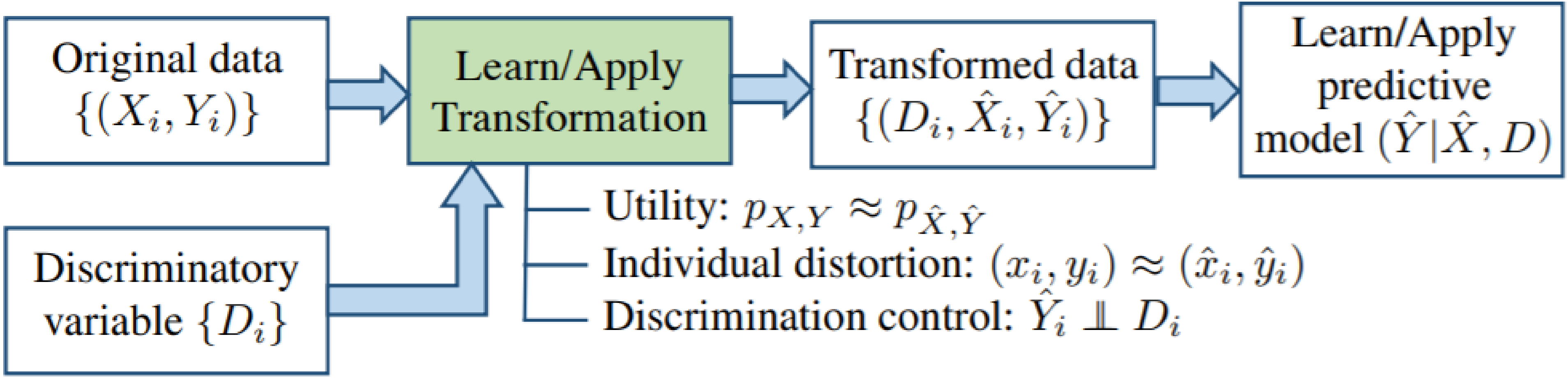
Optimized pre-processing for discrimination prevention

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney

Advances in Neural Information Processing Systems. 2017.

- A probabilistic formulation of data pre-processing to reduce discrimination
- Convex optimization to learn a data transformation that:

1. Control discrimination
2. Limit the distortion in individual data samples
3. Preserve utility



Original data $\{(X_i, Y_i)\}$
Learn/Apply Transformation
Transformed data $\{(D_i, \hat{X}_i, \hat{Y}_i)\}$
Learn/Apply predictive model $(\hat{Y}|\hat{X}, D)$
Discriminatory variable $\{D_i\}$
Utility: $p_{X,Y} \approx p_{\hat{X},\hat{Y}}$
Individual distortion: $(x_i, y_i) \approx (\hat{x}_i, \hat{y}_i)$
Discrimination control: $\hat{Y}_i \perp\!\!\!\perp D_i$

Certifying and removing disparate impact

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian

KDD 2015

- The goal is to certify and remove disparate impact by modifying each attribute such that:

    1. predictability of sensitive attribute using the input data is impossible (minimized)

    2. predictability of class label is preserved

Disparate Impact

- Consider an attribute X, a single binary sensitive attribute S, and a binary classifier f

- f has disparate impact of t, if:

$$\frac{P(f(X) = 1|S = 0)}{P(f(X) = 1|S = 1)} \leq t$$

- That is, the probability that a member of a protected class being classified as 1 (accept) is at most t times (e.g. t=80% -- the 80% rule) less than a member of unprotected class.

Certifying disparate impact

- The main idea is that a classifier $f(X)$ **does not have disparate impact, if the sensitive attribute *S* is not predictable by *X*.**

- → We can check the data without knowing the label attribute or the even the algorithm

Certifying Disparate Impact

- **Balanced Error Rate (BER): consider a classifier $g: X \rightarrow S$**

$$BER(g(X), S) = \frac{P(g(X) = 0|S = 1) + P(g(X) = 1|S = 0)}{2}$$

- **ϵ-Predictability: The data is ϵ-predictable if there exists $g: X \rightarrow S$ such that $BER(g(X), S) \leq \epsilon$**

Theorem: If a dataset D admits a classifier f with disparate impact of 0.8, then D is $(0.5 - \frac{B}{8})$-predictable, where $B = P(F(X) = 1|S = 0)$

$$BER(f(X), S) = \frac{P(f(X) = 0|S = 1) + P(f(X) = 1|S = 0)}{2}$$

$$= \frac{1 - P\left(f(X) = 1\middle|S = 1\right) + B}{2}$$

$$\leq \frac{1 - P\left(f(X) = 1\middle|S = 0\right)/0.8 + B}{2}$$

$$= \frac{1 - B/0.8 + B}{2} = \frac{1}{2} - \frac{B}{8}$$

Removing Disparate Impact

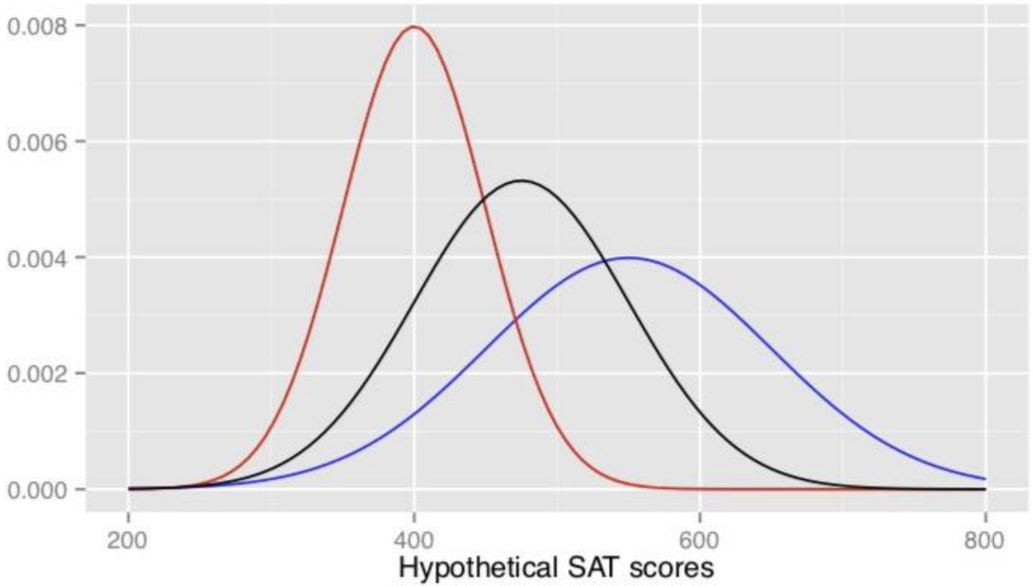
- It is easy to remove the data disparate-impact free: Just set all values of X’=0
- This, however, removes the power of data to predict class labels
- We want to transform X to X’ such that prediction power of data is preserved:
    - we want to transform X in a way that the rankings within demographic groups is preserved (but not necessarily across groups).

Removing Disparate Impact

- Let p_x^s be the percentage of tuples at group $S = s$ with value <u>at most</u> $X = x$

- for each tuple (x_i, s_i):
  - Calculate $p_{x_i}^{s_i}$
  - Find x_i^{-1} such that $p_{x_i^{-1}}^{(1-S_i)} = p_{x_i}^{s_i}$
  - Repair $\overline{x_i}$ as median (x_i, x_i^{-1})

Removing Disparate Impact



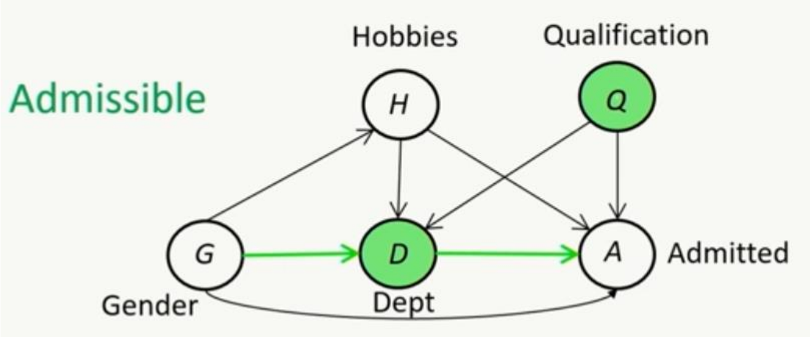
Interventional Fairness: Causal Database Repair for Algorithmic Fairness

Babak Salimi, Luke Rodriguez, Bill Howe, Dan Suciu

SIGMOD 2019

- Repair the pre-existing human bias before using the data for learning
- Proposes the causal notion of fairness and reduces the problem to dataset repair

- User specify admissible variables K, only allow causal influence through K
- Admissible variables are socially not discriminative



- An application is fair if the protected attribute does not affect the outcome for any possible configuration of admissible variables

- Given admissible variables, derive a set of conditional independence constraints that imply interventional fairness.
- Model as a database repair problem
- Classifiers trained on repaired data:
  - Provably fair by interventional fairness
  - Empirically fair by other metrics

Assessing and Remedying Coverage for a Given Dataset

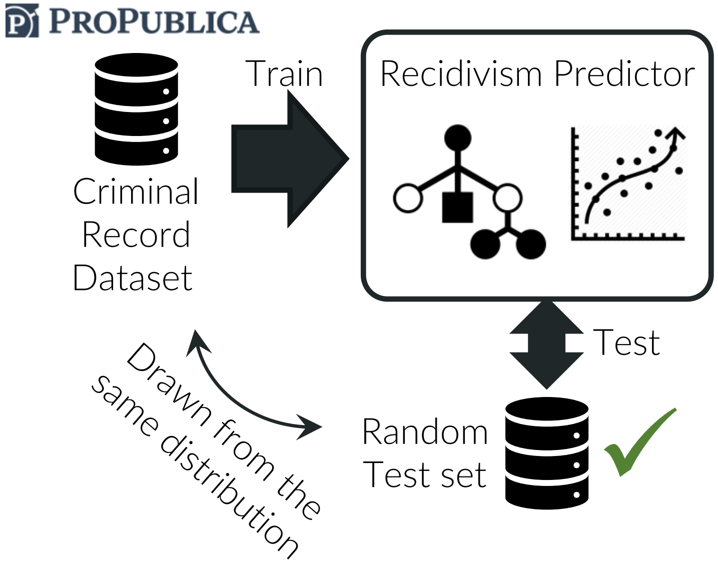
A. Asudeh, Z. Jin, H. V. Jagadish

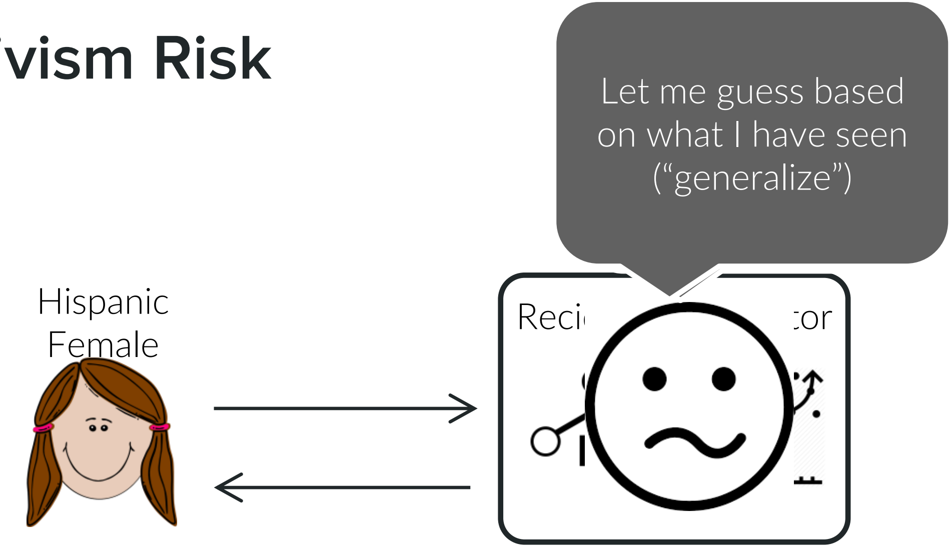
ICDE 2019

Coverage

- To make sure the dataset has “enough” representatives from the minority subgroups

Example: predicting the recidivism Risk



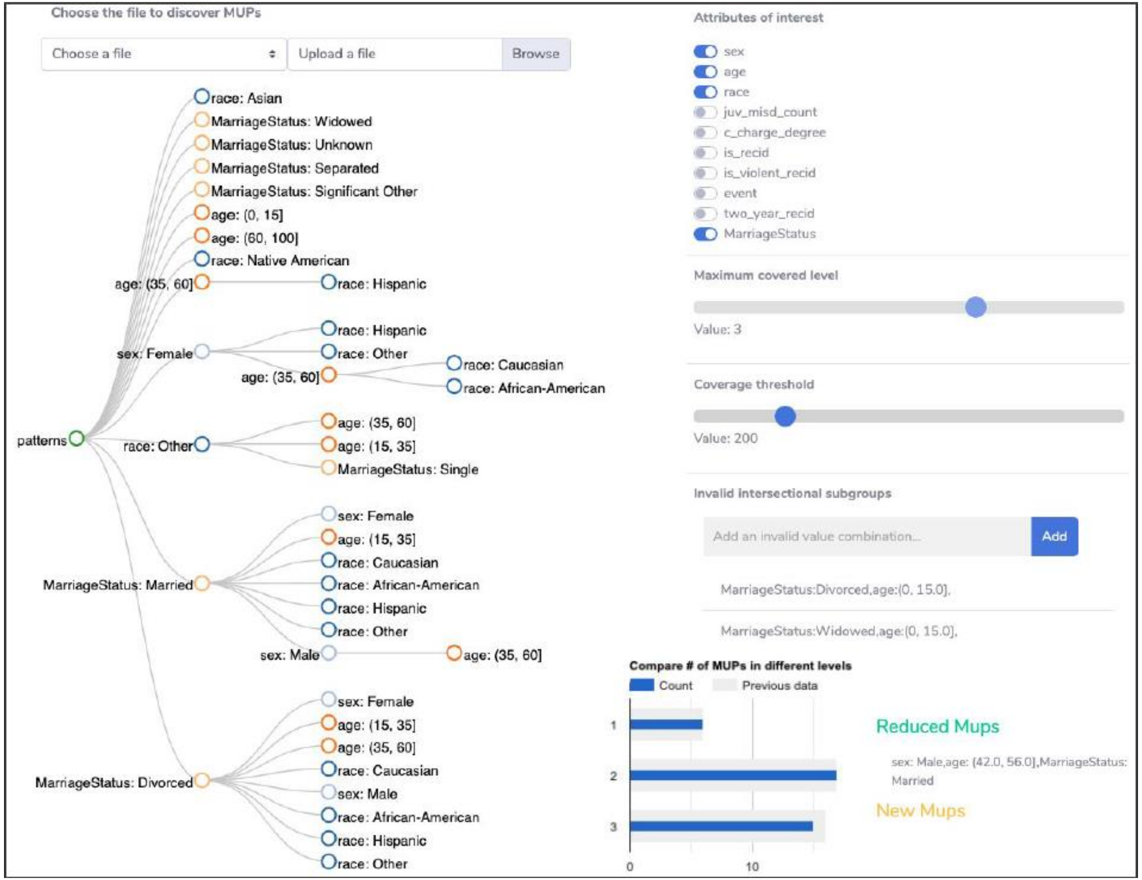


(Lucky): Similar “behavior” →

(Unlucky): Diff. “behavior” →

- Identifying lack of coverage:
    - Challenge: Combinatorial attributes space → #P-hard problem
    - Transform the problem to a DAG traversal; practically efficient algorithms
- Coverage Enhancement:
    - What are the min. records to collect, in order to remove lack of coverage
    - A set cover instance with exponential size input

MithraCoverage



[*] Z Jin, M Xu, C Sun, A Asudeh, and H. V. Jagadish. MithraCoverage: A System for Investigating Population Bias for Intersectional Fairness. In SIGMOD 2020.