Interventional Fairness: Causal Database Repair for Algorithmic Fairness

Salimi et. al. SIGMOD 2019

Presented by:

Shishir Adhikari

Outline

- Overview
- Background
- Definitions and Examples
- Testing and Enforcing Justifiable Fairness
- Repairing Training Data
- Generalizability and Scalability
- Evaluation metrics
- Experimental Results
- Discussion

Overview: Research Area

- Combination of ideas from multiple research areas
- Fairness notions in terms of causality
- Relational between conditional independence and database theory



Overview: Approach and Fairness Notion



Overview: Main Takeaways

- Need of data repair
 - Discrimination due to pre-existing bias in data
- New (testable) notion of fairness based on causality
 - Existing associational and causal fairness notions under/over estimate discrimination
- Association of protected attribute with admissible variables is (socially) acceptable
 - Outcome should be independent of "Inadmissible" variables given "admissible" variables
- Use CI constraint for database repairing
 - Reduce to Multi-valued Dependency (MVD) problem
- New evaluation metric for measuring discrimination
 - Ratio of Observable Discrimination (ROD)



Source: Lecture Slides

Background: Structural Causal Model (SCM)

- Models how nature assigns values to the features of interest
 - **V** (Endogeneous Variables): Variables of interest (for causal relationship).
 - **U**(Exogenous Variables): Variables external to the model. Disturbances or noise.
 - **f** (Function): The function that assign values to each variable in **V**.
- A variable is defined by the function of its **direct causes** and **unknown disturbances**.

 $V_c = f_c(\mathbf{V^{(c)}},\mathbf{U^{(c)}})$

- Represented as a Causal DAG G(V,E)
- A variable (V) has incoming edges from its direct causes and unknown disturbances (U).
- U is not explicitly shown in G
- Compact representation of joint probability distribution (like Bayesian Network)

$$P(V_1, V_2, \dots, V_n) = \prod_{i=1}^n P(V_i | pa(V_i))$$

- Conditional independencies by d-separation in G (from G to P)
- Faithfulness: All the conditional independencies in the data are entailed by d-separation conditions (from P to G)

Background: SCM and Causal Hierarchy

- Observational Question
 - \circ What if we **see** A? P(y|A)
- Interventional Question
 - What if we **do** A? P(y|do(A))
- Counterfactual Question
 - What if **we did things differently**? $P(y_{A'} | A)$

$$P(o,g,h,d) = P(g)P(h|g)P(d|g,h)P(o|d,h)$$

Observed outcome for cs department

$$P(o|D=cs) = \sum_{g,h} P(g)P(h|g)P(D=cs|g,h)P(o|D=cs,h)$$

Outcome when you set department to cs

$$P(o|do(D=cs)) = \sum_{g,h} P(g)P(h|g)P(o|D=cs,h)$$

What would be the outcome had the applicant been a female? $P(O_{G=female}|G=male, O=accept)$



Background: Algorithmic Fairness

• Associational Fairness

Fairness Metric	Description
Demographic Parity (DP) [5, 13, 47]	SшO
Conditional Statistical parity [10]	$S \perp O \mathbf{A}$
Equalized Odds (EO) [19, 57]	$S \perp O Y$
Predictive Parity (PP)[9, 9, 19, 47]	$S \perp \!\!\!\perp Y O$

- Causal Fairness
 - Counterfactual Fairness
 - Individual level fairness, hard to estimate
 - Proxy Fairness
 - Fails to capture some discrimination
 - Path-specific Fairness
 - Hard to identify path-specific effects



Example: 2.3

- P(Q=1) =P(Q=0)= P(D='A') = P(D='B') = 1/2
- $f(G, 'A',Q) = G \land Q, f(G, 'B',Q) = (1-G) \land Q$

Qualified male/female

 $\bullet P(O \text{PLAX}db \text{PCP}) = P(O = 1 | do(G = 0))$

 $P(O = 1 | do(G = 1)) = \sum_{d,q} P(q) P(d) P(O = 1 | q, d, G = 1)$

Definition: Interventional Fairness

Definition 3.1 (**K**-*fair*). Fix a set of attributes $\mathbf{K} \subseteq \mathbf{V} - \{S, O\}$. We say that an algorithm $\mathcal{A} : Dom(\mathbf{X}) \to Dom(O)$ is **K**-fair w.r.t. a protected attribute *S* if, for any context $\mathbf{K} = \mathbf{k}$ and every outcome O = o, the following holds:

Pr(O = o|do(S = 0), do(K = k)) = Pr(O = o|do(S = 1), do(K = k))(7)

An algorithm is **interventionally fair** if it is **K**-fair for every set **K**.

Example fails to satisfy **K**-fairness when **K** = {D}.

$$P(O=1|do(G=1), do(D='A'))
e P(O=1|do(G=0), do(D='A'))$$



Example: 2.3

- P(Q=1) =P(Q=0)= P(D='A') = P(D='B') = 1/2
- f(G, 'A',Q) = G ∧ Q, f(G, 'B',Q) = (1-G)
 ∧ Q
 - Qualified male/female

Definition: Fairness Application and Justifiable Fairness

- Interventional Fairness fails to satisfy all cases of individual fairness
 - P(o | do(G=0)) = P(o | do(G=1)) is K-fair with K={}
- Too restrictive
 - \circ ~ No path from S to O as sufficient condition

Definition 3.3 (Fairness application). A fairness application over a domain V is a tuple (\mathcal{A} , S, A, I), where $\mathcal{A} : Dom(X) \rightarrow$ Dom(O) is an algorithm mappying input variables $X \subseteq V$ to an outcome $O \in V$, $S \in V$ is the protected attribute, and $A \cup I = V - \{S, O\}$ is a partition of the variables into admissible and inadmissible.

Definition 3.4 (Justifiable fairness). A fairness application $(\mathcal{A}, S, \mathbf{A}, \mathbf{I})$ is justifiability fair if it is **K**-fair w.r.t. all supersets $\mathbf{K} \supseteq \mathbf{A}$.



Example: 3.2

•
$$P(U_0=1)=P(U_0=0)=\frac{1}{2}$$

- f(G,0)= G
- f(G,1) = 1-G

Definition: Justifiable Fairness with Causal DAG

THEOREM 3.5. If all directed paths from S to O go through an admissible attribute in **A**, then the algorithm is justifiably fair. If the probability distribution is faithful to the causal DAG, then the converse also holds.



Question: Which one is (justifiably) fair when A = {D}?

So Far...

- Overview
- Background
- Definitions and Examples
- Testing and Enforcing Justifiable Fairness
- Repairing Training Data
- Generalizability and Scalability
- Evaluation metrics
- Experimental Results
- Discussion

Justifiable Fairness with Conditional Independence

• Avoid knowledge of causal model

In terms of Outcome (O)

THEOREM 3.7. A sufficient condition for a fairness application $(\mathcal{A}, S, \mathbf{A}, \mathbf{I})$ to be justifiably fair is $MB(O) \subseteq \mathbf{A}$.

 A Markov Boundary (MB) of a variable includes its parents, children, and spouses (other parents of children)

In terms of Labels (Y)

COROLLARY 3.8. Fix a training data D, Pr, where $Y \in V$ is the training label, and A, I are admissible and inadmissible attributes. Then any reasonable classifier trained on a set of variables $X \subseteq V$ is justifiably fair w.r.t. a protected attribute S, if either:

(a) Pr satisfies the CI $(Y \perp X \cap I | X \cap A)$, or

(b) $\mathbf{X} \supseteq \mathbf{A}$ and \Pr satisfies the saturated $CI(Y \perp \!\!\!\perp I | \mathbf{A})$.

Building Fair Classifiers

Implications of Corollary 3.8:

- (a) Use only **A** for training
 - Decreases utility

COROLLARY 3.8. Fix a training data D, Pr, where $Y \in V$ is the training label, and A, I are admissible and inadmissible attributes. Then any reasonable classifier trained on a set of variables $X \subseteq V$ is justifiably fair w.r.t. a protected attribute S, if either:

- (a) Pr satisfies the CI $(Y \perp\!\!\!\perp X \cap I | X \cap A)$, or
- (b) $\mathbf{X} \supseteq \mathbf{A}$ and \Pr satisfies the saturated $CI(Y \perp | \mathbf{A})$.

- (b) Repairing training data
 - Use CI condition as integrity constraint for training data D
 - Minimal insertions and deletions of tuples in D to obtain D' satisfying Cl
 - *Reduces to MVD problem in database theory*
 - In terms of causal DAG, it corresponds to modifying underlying causal model so that there is no directed path from S->I->Y or S->Y, without the knowledge of causal model

Minimal Data Repair for MVD and Cl

Given a partition, **V** = **X** U **Y** U **Z**, we say D satisfies multi-valued dependency(MVD)

 $\mathbf{Z} \twoheadrightarrow \mathbf{X}$ if $D = \Pi_{\mathbf{XZ}}(D) \bowtie \Pi_{\mathbf{ZY}}(D)$.

For D, the projections are:

Z	Х	Z	Y
С	а	С	а
С	b	С	b
d	b	d	b

D.	V	v	7	Dr	D_1 :	Х	Y	Ζ	[
D.	Λ	1			t_1	a	а	С	D_2 :	Х	Y	Ζ
t_1	a	а	С	3/8	t_2	a	b	C	t_1	a	a	C
t_2	a	b	С	2/8	t_	h	0	0	<i>v</i> ₁	0	h	0
t_3	b	a	С	2/8	13	U	u	C	12	u	U	C
t ₄	h	h	d	1/8	t_4	b	b	С	t_4	b	b	d
•4	U	U	u	1/0	t_5	b	b	d				

Figure 5: A simple database repair: *D* does not satisfy the MVD $Z \rightarrow X$. In D_1 , we inserted the tuple (b, b, c) to satisfy the MVD, and in D_2 we deleted the tuple (b, a, c) to satisfy the MVD.

Minimal Data Repair for MVD and Cl

- Given a partition, **V** = **X** U **Y** U **Z**, we say D satisfies multi-valued dependency(MVD)
- $\mathbf{Z} \twoheadrightarrow \mathbf{X}$ if $D = \Pi_{\mathbf{XZ}}(D) \bowtie \Pi_{\mathbf{ZY}}(D)$.
- A uniform Pr satisfies a saturated CI (**X**;**Y**|**Z**) iff its support D satisfies the MVD $Z \rightarrow X$
- Hard to have uniform Pr in training data
- Workaround: CI (KX;Y|Z) implies CI (X;Y|Z), where {K} is fresh variable not in V

					D	V	17	7		17	V	37	77	D	17	V	17	7
D:	X	Y	7.	Pr	B:	X	Y	L	D_B :	K	Х	Ŷ	L	D'_B :	K	X	Y	L
2.		-	-			a	a	C		1	a	a	С		1	a	a	С
t_1	a	a	С	3/8		a	a	С		2	a	a	С		2	а	a	С
t_2	a	b	С	2/8		а	а	С		3	a	а	С		1	а	b	С
to	h	a	C	2/8		a	b	С		1	a	b	С		2	a	b	С
13	U	u	L	2/0		a	b	С		2	a	b	С		1	b	a	С
t_4	b	b	d	1/8		b	a	с		1	b	a	с		1	b	b	С
				1		b	a	С		2	b	a	С		1	b	b	d

a	b	С		2	a	b	С		1	b	a	С
b	a	С		1	b	a	С		1	b	b	С
b	a	С		2	b	a	С		1	b	b	d
b	b	d		1	b	b	d	D'.	V	v	7	Dr'
								D.	A	1	C	2/7
									a	h	c	2/7
									h	0	C	1/7
									6	u b	C	1/7
									0	0	C	1/7
									D	b	a	1//

Reducing Minimal Repair to 3SAT

- Given a partition, $V = X \cup Y \cup Z$, we say D satisfies multi-valued dependency(MVD) $Z \rightarrow X$ if $D = \prod_{XZ}(D) \bowtie \prod_{ZY}(D)$.
- Then, the database D' obtained after minimal repair is subset of $D^* \stackrel{\text{def}}{=} \Pi_{XZ}(D) \bowtie \Pi_{ZY}(D)$
- Lineage expressions (Hard clauses): boolean conditions that doesn't allow MVD condition and its negation

$$\begin{array}{c} \bigcirc \qquad \Phi_{\varphi} = \qquad (X_{t_1} \land X_{t_4} \land \neg X_{t_2}) \lor (X_{t_2} \land X_{t_3} \land \neg X_{t_1}) \lor \\ (X_{t_3} \land X_{t_2} \land \neg X_{t_4}) \lor (X_{t_4} \land X_{t_1} \land \neg X_{t_3}) \end{array}$$

Hence,

$$\neg \Phi_{\varphi} = (\neg X_{t_1} \lor \neg X_{t_4} \lor X_{t_2}) \land (\neg X_{t_2} \lor \neg X_{t_3} \lor X_{t_1}) \land (\neg X_{t_3} \lor \neg X_{t_2} \lor X_{t_4}) \land (\neg X_{t_4} \lor \neg X_{t_1} \lor X_{t_3})$$

D.	V	V	7	Dr	D_1 :	Х	Y	Ζ
D.	Λ	1	L	11	t_1	a	a	С
t_1	a	a	С	3/8	t_2	a	b	C
t_2	a	b	С	2/8	+	h	0	0
t3	b	a	С	2/8	13	U	a	C
t.	h	h	d	1/8	t_4	b	b	
<i>ι</i> 4	U	U	u	1/0	t_5	b	b	đ

- Errata: t4 should be t5 in above example.
- Membership in D vs {D* D} (Soft clauses): X_{t1} , X_{t2} , X_{t3} , X_{t4} , ~ X_{t5}

Repair via Matrix Factorization: NMF to Cl

First, we review the problem of non-negative rank-one matrix factorization. Given a matrix $\mathbf{M} \in \mathbb{R}^{n \times m}$, the problem of *rank-one nonnegative matrix factorization (NMF)* is the minimization problem: $\operatorname{argmin}_{\mathbf{U} \in \mathbb{R}^{n \times 1}_{+}, \mathbf{V} \in \mathbb{R}^{1 \times m}_{+}} \|\mathbf{M} - \mathbf{UV}\|_{F}$, where \mathbb{R}_{+} stands for non-negative real numbers and $\|.\|_{F}$ is the Euclidean norm of a matrix.⁵

We express the connection between our repair problem and the NMF problem using contingency matrices. Given three disjoint subsets of attributes X, Y, $Z \subseteq V$, let m = |Dom(X)|, n = |Dom(Y)|, k = |Dom(Z)| and $B_z = \sigma_{Z=z}(B)$. A *multiway-contingency matrix* over X, Y and Z consists of $k \ n \times m$ matrices $M_{X,Y,Z}^B = \{M_{X,Y}^{B_z} | z \in Dom(Z)\}$ where, $M_{X,Y}^{B_z}(ij) = \sum_{t \in B} 1_{t[XY]=ij}$. Intuitively, $M_{X,Y}^{B_z}(ij)$ represents the joint frequency of X and Y in a subset of bag with Z = z.

PROPOSITION 4.2. Let *B* be a bag and Pr be the empirical distribution associated to *B*. It holds that $(X \perp\!\!\!\perp Y|_{Pr}Z)$ iff each contingency matrix $\mathbf{M} \in \mathbf{M}_{X|Y|Z}^B$ is of rank-one.

B:	Х	Y	Ζ	B	2 :	Х	Y	Ζ
	a	a	С			a	a	С
	a	a	С			a	a	С
	a	a	С			a	b	С
	a	b	С			a	b	С
	a	b	С			b	a	С
	b	а	С			b	h	C
	b	a	c d			b	b	d
	U	U	u			0	U	и
D		[3	2]	ת	,	Γ0	0	1
$M^{\scriptscriptstyle B_{Z=c}}_{\scriptscriptstyle {f vv}}$	=		-	$, M_{v}^{B}$	$Z^{Z=d}_{V} =$	= Ŭ	U	
ΛΪ		$\lfloor 2 \rfloor$	0	΄ Λ	Ŷ	$\lfloor 0$	1	
		_	_			_		-

$$M_{XY}^{B_{Z=c}'}=egin{bmatrix}2&2\1&1\end{bmatrix},M_{XY}^{B_{Z=d}'}=egin{bmatrix}0&0\0&1\end{bmatrix}$$

Repair via Matrix Factorization: Algorithm

Algorithm 2: Repair using Matrix Factorization.

Input: A bag *B* with attributes $\mathbf{V} = \mathbf{X}\mathbf{Y}\mathbf{Z}$ a CI statment $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z})$. Output: *B'* a repair of *B* 1 for $\mathbf{z} \in Dom(\mathbf{Z})$ do 2 $\begin{bmatrix} M_{\mathbf{X}}^{B'}, M_{\mathbf{Y}}^{B'} \leftarrow \mathbf{Factorize}(M_{\mathbf{X},\mathbf{Y}}^{B_{\mathbf{z}}}) \\ M_{\mathbf{X},\mathbf{Y}}^{B'\mathbf{z}} \leftarrow \frac{1}{|B_{\mathbf{Z}}|} M_{\mathbf{X}}^{B'\mathbf{T}} M_{\mathbf{Y}}^{B'} \end{bmatrix}$ 4 return *B'* associated with $\mathbf{M}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}^{B'\mathbf{z}} = \{M_{\mathbf{X},\mathbf{Y}}^{B'\mathbf{z}}\}$

Setting 1 (MF): Factorize is implemented by off-the-shelf NMF algorithm

Setting 2 (Independent Coupling): **Factorize** is implemented by simple factorization with marginal frequencies of X and Y in B_z

Generalizability and Scalability

• Generalizability

LEMMA 5.1. If the repaired data satisfies $(Y \perp S, \mathbf{I}|_{\Pr_{B'}} \mathbf{A})$ and the unseen test data satisfies $\Pr_T(s, \mathbf{i}|\mathbf{a}) = \Pr_{B'}(s, \mathbf{i}|\mathbf{a})$, then the unseen test data also satisfies $(Y \perp S, \mathbf{I}|_{\Pr_T} \mathbf{A})$

- Performance on unseen test data
- Assumption: Training and test data from same distribution
- Asymptotically, $Pr_{T}(s,i|a) = Pr_{B}(s,i|a)$. Implies, $Pr_{B}(s,i|a) = Pr_{B'}(s,i|a)$ should be satisfied.
- Independent Coupling (IC) approach satisfies by construction
- Other approaches only approximate the condition
- Scalability
 - NP-Complete Problem
 - Depends on MaxSAT solvers and MF approximators
 - Can be highly parallelizable as the problem can be partitioned for domain of Z for CI (X;Y | Z)

So Far...

- Overview
- Background
- Definitions and Examples
- Testing and Enforcing Justifiable Fairness
- Repairing Training Data
- Generalizability and Scalability
- Evaluation metrics
- Experimental Results
- Discussion

Evaluation Metrics

Utility Metric: Accuracy

Bias Metrics: Shown in table.

Metric	Description and Definition
POD	Ratio of Observation Discrimination:
ROD	(See Sec.6.1)
תת	Demographic Parity:
DP	Pr(O = 1 S = 1) - Pr(O = 1 S = 0)
TPB	True Positive Rate Balance:
	Pr(O = 1 S = 1, Y = 1) - Pr(O = 1 S = 0, Y = 1)
TND	True Negative Rate Balance:
IND	Pr(O = 0 S = 1, Y = 0) - Pr(O = 0 S = 0, Y = 0)
	Conditional Statistical Parity:
CDP	$\mathbb{E}_{\mathbf{a}}[Pr(O=1 S=1,\mathbf{a}) - Pr(O=1 S=0,\mathbf{a})]$
OTDP	Conditional TPRB:
CIPB	$\mathbb{E}_{\mathbf{a}}[Pr(O=1 S=1, Y=1, \mathbf{a}) - Pr(O=1 S=0, Y=1, \mathbf{a})]$
OTNE	Conditional TNRB:
CINB	$\mathbb{E}_{\mathbf{a}}[\Pr(O=0 S=1, Y=0, \mathbf{a}) - \Pr(O=0 S=0, Y=0, \mathbf{a})]$

Definition 6.1. Given a fairness application (\mathcal{A} , S, A, I), let $A_b = MB(O) - I$. We quantify the *ratio of observational discrimination (ROD)* of \mathcal{A} against S in a context $A_b = a_b$ as $\delta(S; O|a_b) \stackrel{\text{def}}{=} \frac{\Pr(O=1|S=0, a_b)\Pr(O=0|S=1, a_b)}{\Pr(O=0|S=0, a_b)\Pr(O=1|S=1, a_b)}$.

Intuitively, ROD calculates effect of membership in a protected group on the odds of the positive outcome of algorithm for subjects that are similar on $A_b = a_b$

Experiment: Setup

- ML classifiers.
 - Linear Regression (LR)
 - Random Forest (RF)
 - Multilayer perceptron (MLP)
- Approaches of database repair
 - Original (No Repair)
 - Dropping Inadmissible
 - IC: Independent Coupling
 - MF: Matrix Factorization
 - MS(Hard): MaxSAT using all clauses of the lineage expression
 - MS(Soft): MaxSAT using *a small fraction* of clauses
- Datasets
 - Adult, Adult-binned, COMPAS, COMPAS-binned
- Baselines
 - Feldman et. al., 2015
 - Calmon et. al., 2017

Result: Bias vs Utility



Utility(Accuracy) 99.0 99.0 Original Dropped IC MF 0.64 MS(Hard) MS(Soft) 0.63 0.2 0.0 0.3 0.0 0.1 0.0 0.3 0.1 0.3 Bias(ROD) Bias(ROD) Bias(ROD)

Classifier: MLP

Classifier: LR

Classifier: RF

0.68

0.67

Figure 9: Performance of CAPUCHIN on COMPAS data.

Adult Data: S=Gender, Y=Income Class, I={Marital Status}

COMPAS Data: S=Race, Y= if individual is a recidivist, A={number of prior convictions, severity of charge degree, age}

Result: Comparison with Baselines

Data dependent performance

Experiments on binned data to make comparable with baselines

CAPUCINE balances bias vs utility (generalizes to unseen test data)

Baselines were designed for DP but work with ROD metric as well



Figure 13: Comparison with other methods on Binned Adult data.



Figure 14: Comparison with other methods on Binned COMPAS.

Result: Comparison with other Fairness Metrics



Figure 8: Bias reduction performance of CAPUCHIN for MLP classifier.

Result: Bias, Insertions/Deletions, Parallelization



Figure 12: Speed up achieved (a) by partitioning and parallel processing on 128 cores; (b) by partitioning on a single core.



Figure 10: Comparison of different repair methods.

Any Questions? Class Discussion

• Strengths

• Limitations

• Future Research Direction

