The Workshops of the Thirty-Second
AAAI Conference on Artificial Intelligence

Fairness-Aware Relational Learning and Inference

Golnoosh Farnadi,[†*] Behrouz Babaki,[§*] Lise Getoor[†]
[†]University of California, Santa Cruz, USA,
[§]KU Leuven, Belgium

1 Introduction

AI and machine learning have become essential tools in automatic decision-making. When used in processes such as employment, education, advertising, loan approval, criminal risk assessment, and policing, these tools can have significant influence on the lives of individuals. This potential influence has raised concerns about algorithmic discrimination and bias. Recently, a number of methods to address these concerns have been proposed. These methods deal with bias through unawareness/blindness, awareness/Lipschitz property (Dwork et al. 2012), demographic parity/disparate impact (Feldman et al. 2015), preference-based (Zafar et al. 2017), and equality of opportunity (Hardt, Price, and Srebro 2016).

The existing studies on fairness in machine learning assume the attribute-value data format. In this setting, the goal is to design algorithms that make fair predictions across two groups that are defined in terms of an attribute value such as age, gender, race, religion, etc. These two groups are called *protected* and *unprotected*.

Since many forms of bias occur in a social context, leveraging relational information is essential. In this paper, we extend the concepts of fairness-aware inference and learning to the relational setting. Instead of defining the protected group in terms of a single attribute value, we introduce new definitions to include the relational context in which discrimination may occur. This extension leads to a richer notion of fairness, which can capture complex dependenices that are present in real-world scenarios. For example in an organization, the existing biases in evaluation reports produced by employees can influence the promotion decisions. These biases can be resulting from the complex relational network of opinions of employees about each other.

We first introduce and formulate fairness in relational setting, then we propose 1) fairness-aware constrained conditional inference subject to common data-oriented fairness measures and 2) fairness-aware learning by incorporating decision-oriented fairness measures.

[*]Equal contributors.

2 Fairness in Relational Settings

In this section, we formalize relational fairness using first-order logic: An atom is an expression of the form $p(a_1, a_2, \ldots, a_n)$ where p is a *predicate symbol*, and each argument $a_1, a_2, \ldots, a_n$ is either a constant or a variable. The finite set of all possible substitutions of a variable to a constant for a particular variable a is called its *domain* D_a. If all variables in $p(a_1, a_2, \ldots, a_n)$ are substituted by some constant from their respective domain, then we call the resulting atom a *ground atom*. A formula is defined by induction: every atom is a formula. If α and β are formulae, then $\alpha \vee \beta$, $\alpha \wedge \beta$, $\neg\alpha$, $\exists x\, \alpha$, $\forall x\, \alpha$ are formulae. An interpretation I is a mapping that associates a truth value $I(P)$ to each ground atom P.

We denote formula F which has only one free variable v (i.e. other variables in F are quantified) by $F[v]$. The *population* defined by $F[v]$ is the set of substitutions of v for which $F[v]$ holds. A *discriminative pattern* is a pair $DP[v] \equiv (F_1[v], F_2[v])$, where $F_1[v]$ and $F_2[v]$ are formulae.

Example 1. *Consider a hypothetical scenario in an organization in which young female workers who have older male supervisors have lower chances of promotion than their male counterparts*[1]*. In this scenario, the discrimination pattern is:*

$$DP[v] := \big(\mathit{Female}(v), \mathit{Young}(v) \wedge \big(\exists u, \neg\mathit{Young}(u) \wedge \neg\mathit{Female}(u) \wedge \mathit{Supervise}(u, v)\big)\big).$$

Given an interpretation I, the *protected group*

$$PG \equiv \{v : F_1[v] \wedge F_2[v]\}$$

is defined as the set of all instances hold for variable v for which $F_1[v] \wedge F_2[v]$ is true under interpretation I, that is, $I(F_1[v] \wedge F_2[v]) = 1$. Similarly, the *unprotected group*

$$UG \equiv \{v : \neg F_1[v] \wedge F_2[v]\}$$

is defined as the set of all instances holds for variable v for which $I(\neg F_1[v] \wedge F_2[v]) = 1$. A *decision atom* $d(v)$ is an

[1]Of course, many other patterns may be possible: female bosses may promote female subordinates and discriminate against male workers, or male bosses may promote female employees. Our goal is to provide a general framework which is able to describe arbitrarily complex discrimination patterns.

atom containing exactly one variable v that specifies a decision affecting the protected group which is defined either by law or end-user.

Example 2. *The protected group of the discrimination pattern specified in Example 1 is*

$$PG := \{v : \textit{Female}(v) \wedge \textit{Young}(v) \wedge (\exists u, \neg\textit{Young}(u) \wedge \neg\textit{Female}(u) \wedge \textit{Supervise}(u, v))\}$$

and the unprotected group is

$$UG := \{v : \neg\textit{Female}(v) \wedge \textit{Young}(v) \wedge (\exists u, \neg\textit{Young}(u) \wedge \neg\textit{Female}(u) \wedge \textit{Supervise}(u, v))\}.$$

The decision atom $d[v] := Promotion(v)$ indicates the promotion decision.

3 Fairness-aware Inference

To formulate fairness in the relational setting, we propose fairness-aware constrained conditional inference subject to common data-oriented fairness measures. We first introduce these fairness measures and then re-define them using the notation introduced in Section 2. Let a and c denote the counts of denial (i.e., negative decisions) for protected and unprotected groups, and n_1 and n_2 denote their sizes, respectively. Let $p_1 = {}^{a}/_{n_1}$ be the proportion of benefit denied for the protected group, and $p_2 = {}^{c}/_{n_2}$ be the proportion of benefit denied for the unprotected group. Using p_1 and p_2 we can define three well-known fairness measures as follows:

1. **Risk difference**: $RD = p_1 - p_2$, also known as absolute risk reduction. The UK uses RD as it's legal definition of fairness measure.
2. **Risk Ratio**: $RR = {}^{p_1}/_{p_2}$, also known as relative risk. The EU court of justice has given more emphasis on the RR as a measure of fairness.
3. **Relative Chance**: $RC = {}^{1-p_1}/_{1-p_2}$ also known as selection rate. The US laws and courts mainly refer to the RC as a measure of fairness. For further information we refer to (Pedreschi, Ruggieri, and Turini 2012).

Notice that RR is the ratio of benefit denial between the protected and unprotected groups, while RC is the ratio of benefit granting.

Now, we can formulate these measures using the formalism defined in Section 2. Given the decision atom $d(v)$ and discriminative pattern $DP(F_1[v], F_2[v])$, the counts of denial for both protected and unprotected groups are computed by the following equations:

$$a \equiv \sum_{v \in D_v} I\big(\neg d(v) \wedge F_1[v] \wedge F_2[v]\big)$$
$$c \equiv \sum_{v \in D_v} I\big(\neg d(v) \wedge \neg F_1[v] \wedge F_2[v]\big)$$
$$n_1 \equiv \sum_{v \in D_v} I\big(F_1[v] \wedge F_2[v]\big)$$
$$n_2 \equiv \sum_{v \in D_v} I\big(\neg F_1[v] \wedge F_2[v]\big)$$

Using these counts, the fairness measures can be computed as: $RD \equiv {}^{\mathbf{a}}/_{n_1} - {}^{\mathbf{c}}/_{n_2}$, $RR \equiv \frac{{}^{\mathbf{a}}/_{n_1}}{{}^{\mathbf{c}}/_{n_2}}$, and $RC \equiv \frac{1-{}^{\mathbf{a}}/_{n_1}}{1-{}^{\mathbf{c}}/_{n_2}}$. Finally, we introduce the notion of δ-fairness.

Definition 1 (δ-fairness)**.** *If a fairness measure for a decision making process falls within some δ-window, then the process is δ-*fair*. Given $0 \leq \delta \leq 1$, the δ-windows for measures RD/RR/RC are defined as:*

$$-\delta \leq RD \leq \delta$$
$$1 - \delta \leq RR \leq 1 + \delta$$
$$1 - \delta \leq RC \leq 1 + \delta$$

The standard MAP inference aims at finding values that maximize the conditional probability of unknowns. Once a decision is made according to these values, one can use the fairness measure to quantify the degree of discrimination. To develop fairness-aware inference, we propose to incorporate fairness in MAP inference by adding the δ-fairness constraints to the underlying optimization problem of MAP inference.

Consider risk difference, *RD*, where $RD \equiv \frac{\mathbf{a}}{n_1} - \frac{\mathbf{c}}{n_2}$. The δ-fairness constraint $-\delta \leq RD \leq \delta$ can be encoded as the following constraints:

$$n_2\mathbf{a} - n_1\mathbf{c} - n_1n_2\delta \leq 0$$
$$n_2\mathbf{a} - n_1\mathbf{c} + n_1n_2\delta \geq 0$$

Similarly, from $RR \equiv \frac{\mathbf{a}/n_1}{\mathbf{c}/n_2}$ and the δ-fairness constraint $1 - \delta \leq RR \leq 1 + \delta$ we obtain:

$$n_2\mathbf{a} - (1+\delta)n_1\mathbf{c} \leq 0$$
$$n_2\mathbf{a} - (1-\delta)n_1\mathbf{c} \geq 0$$

And finally, $RC \equiv \frac{1-\mathbf{a}/n_1}{1-\mathbf{c}/n_2}$ and the δ-fairness constraint $1 - \delta \leq RC \leq 1 + \delta$ gives:

$$-n_2\mathbf{a} + (1+\delta)n_1\mathbf{c} - \delta n_1n_2 \leq 0$$
$$-n_2\mathbf{a} + (1-\delta)n_1\mathbf{c} + \delta n_1n_2 \geq 0$$

4 Fairness-aware parameter learning

In this section, we first review five measures of fairness from literature. The difference between these measures and the ones introduced earlier is that the latter are based on the *decision* made by an algorithm. To explain these measures, assume that symbols tp, fn, fp, and tn denote true positive, false negative, false positive, and true negative rate, respectively. Each of the following measures assume that a decision is fair if the values of some quantity among the protected and unprotected group are the same:

1. **Overall accuracy equality**: equal values for ${}^{(tp+tn)}/_{(tp+fn+fp+tn)}$. This measure is not commonly used because it does not distinguish between the accuracy for success and failure.

2. **Demographic parity**: equal marginal distributions of the predicted classes $^{(tp+fp)}/_{(tp+fn+fp+tn)}$ or $^{(fn+tn)}/_{(tp+fn+fp+tn)}$ in both groups. This measure has been criticized as it can lead to highly undesirable decisions (Dwork et al. 2012).
3. **Equality of opportunity**: equal values for $^{tp}/_{(tp+fn)}$ or $^{tn}/_{(fp+tn)}$.
4. **Conditional use accuracy equality**: equal values for $^{tp}/_{(tp+fp)}$ or $^{tn}/_{(fn+tn)}$.
5. **Treatment equality**: equal ratio of false negatives and false positives (i.e., $^{fp}/_{fn}$ or $^{fn}/_{fp}$) in both groups.

To incorporate these measures in learning the parameters of a decriminative relational model with joint probability distribution $P(y|x)$, we first introduce their logical counterparts. Let $\hat{y}_j$ and y_j denote the actual and predicted truth values for n atoms of interest. We extend the definitions of tp, fn, fp and tn as:

$$tp = \sum_{j=1}^{n} I(\hat{y}_j \wedge y_j)$$
$$fn = \sum_{j=1}^{n} I(\hat{y}_j \wedge \neg y_j)$$
$$fp = \sum_{j=1}^{n} I(\neg\hat{y}_j \wedge y_j)$$
$$tn = \sum_{j=1}^{n} I(\neg\hat{y}_j \wedge \neg y_j)$$

Fair parameter learning is an optimization problem with two possibly conflicting goals: 1) to achieve high prediction power according to the data, and 2) ensuring fair predictions. The first goal can be for example translated into a high likelihood of the data.

In order to achieve the second goal, we add a term to the objective function that reflects the degree of fairness of predictions according to known fair truth values for a subset of variables. More specifically, let $\mathbf{y}_{PG}$ and $\mathbf{y}_{UG}$ denote the fair truth values of the target predicate for subsets of protected and unprotected groups, respectively. Given a fairness measure M (which can be one of the measures defined above), we aim at decreasing the value of $|M(\mathbf{y}_{PG}) - M(\mathbf{y}_{UG})|$. Combining these two goals leads to the following objective function for fairness-aware parameter learning:

$$\max_{W} \left\{ \log P_W(\mathbf{y}|\mathbf{x}) - \gamma \cdot \mathbb{E}_W\Big[|M(\mathbf{y}_{PG}) - M(\mathbf{y}_{UG})|\Big] \right\}$$

where γ is a positive constant that determines the relative importance of the two components of the objective function.

5 Conclusion

In this paper, we introduce the notion of fairness in relational setting. We extend MAP inference with fairness-aware constrained conditional inference subject to common data-oriented fairness measures. In addition, we propose a fairness-aware learning algorithm that incorporates decision-oriented fairness measures to ensure fairness in learning. We believe that extending fairness to the relational setting facilitates defining complex discrimination patterns. Many applications in social network analysis, personalized advertising, education science, and computational social science can benefit from this extension.

References

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. S. 2012. Fairness through awareness. In *ITCS*, 214–226. ACM.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *KDD*, 259–268. ACM.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *NIPS*, 3315–3323.

Pedreschi, D.; Ruggieri, S.; and Turini, F. 2012. A study of top-k measures for discrimination discovery. In *SAC*, 126–131. ACM.

Zafar, M. B.; Valera, I.; Gomez-Rodriguez, M.; Gummadi, K. P.; and Weller, A. 2017. From parity to preference-based notions of fairness in classification. *CoRR* abs/1707.00010.