TADA 2021
Assignment 3 - Fairly Causal

[redacted]

June 20, 2021

1 Introduction

In this report, we discuss the use of causal graphs on detecting discrimination as proposed by Bonchi et al. [1]. We contrast their approach with statistical parity, one of the canonical notions of fairness in machine learning. While Bonchi et al.'s SBCN showed some encouraging experimental results, we mainly look at its potential weaknesses, which can be thought of as directions for future research. Lastly, we go beyond classification problems and examine how fairness can be ensured in the ranking of objects.

2 Statistical Parity

Statistical parity, or Independence, is one of the most popular criteria for fairness. It assumes that everything is fair if the distribution of outcome is the same for all groups. For a binary classification problem, if we denote the protected group by S and its complement (the non-protected group) by S^C, then statistical parity tests for:

$$p(+|S) \approx p(+|S^C)$$

where $p(+|S)$ is the probability an instance is classified as positive given that it is a member of S. While statistical parity is desired in some cases, Dwork et al. [2] made a reasonable argument that it alone is not sufficient. Specifically, statistical parity only ensures group fairness while individuals might still be treated unfairly. There are various scenarios in which statistical parity fails.

First, statistical parity still holds if wrong subsets of the groups are selected (for being classified as positive) as long as the sizes of those subsets are proportional to their corresponding groups, which reduces the utility of prediction. Second, statistical parity cannot prevent self-fulfilling prophecy, the act of (deliberately or not) choosing unqualified candidates from a group, which makes the discrimination even stronger for that group in the future. Third, it cannot prevent violation of privacy, in which case a subgroup (or an individual) of a group can be identified with confidence as a result of a malicious manipulation of exposure.

3 The Suppes-Bayes causal networks

3.1 Definitions

Due to those severe problems of statistical parity, research has been conducted to find alternative methods. One of those attempts was from Bonchi et al. [1]. They proposed a method to construct Suppes-Bayes causal networks (SBCN) from observational data and then identify discrimination as causal paths on these networks (or graphs).

Given a database, the causal graph is constructed by forming nodes and directed edges. A node represents an assignment <*attribute*=*value*>. So a categorical attribute with c categories should result in c nodes in the graph. All numerical attributes are expected to be binned and then treated as categorical afterward. A directed edge represents a causal effect from its source to its sink. The weight (or confidence) of an edge is a positive number which implies the strength of its causal effect.

At first, each attribute is manually assigned with a temporal indicator of its relative chronological order. For example, if the database is about people, then the temporal order of the attribute *nationality* should be smaller than *income*. Edges are only allowed to go from a node with a lower or equal temporal order of attribute to a node with a higher or equal temporal order of attribute. Furthermore, for an edge from u to v, its weight is set to $p(v|u) - p(v|\neg u)$ (note that if $p(v|u) \leq p(v|\neg u)$, we remove this edge from consideration). Over all possible subsets of the edges that satisfy the stated requirements, the best subset is chosen based on the regularized maximum likelihood with the BIC-score. The discrimination is then computed as the proportion of random walks that go from nodes that represent group memberships to nodes that represent outcomes.

3.2 Strengths

Bonchi et al. made an interesting argument that discrimination is causal, and thus it makes sense to treat it as a causal problem by modeling its causal graph. The use of the temporal order allows researchers to include domain knowledge into the process.

If input correctly, it can help eliminate lots of obvious non-causal edges while still maintaining the potential causal relationships.

Constructing the SBCN costs $O(sm)$ time and $O(m^2)$ space complexity (where s is the number of data points and m is the number of attribute-value pairs), which is quite reasonable for many datasets. The time needed to compute the discrimination score for a pair of ground and outcome is then $O(tm)$ (where t is the number of random-walk simulations), which is also practical in most cases.

Moreover, the authors claim that the SBCN can detect different types of discrimination, including group discrimination, indirect discrimination, genuine requirements, and individual/subgroup discrimination. Given that most of the existing approaches only address one (or two) of the mentioned problems, the SBCN's ability to deal with all of them confirms its versatility.

Lastly, let us emphasize that under some conditions, the graph-reconstruction algorithm is sound. That is, if (1) there is a correct (but unknown) DAG that can represent the causal effects of the attributes, (2) each node that has more than one cause has conjunctive parents, (3) all relevant attributes are present in the data, (4) the temporal order is correct and complete, (5) the sample size $s \to \infty$, (6) the data is uniformly randomly corrupted, then the resulting causal graph is the correct one.

3.3 Weaknesses

We can see that the list of conditions for the guarantee of soundness is quite long. First, it is almost always unrealistic that the assumption of conjunctive parents holds. In many cases, not all causes have to happen before the observation of their common effect. For example, getting promoted and reading a funny meme both make us happy, but we do not need both of them in order to be happy.

Second, the assumption about the completeness of the data is also impractical. In real life, it is always hard to ensure that all relevant information is included in the dataset. Especially, the absence of confounders can largely affect the correctness of the resulting causal graph. For example, let us denote by A the outside temperature, by B the air conditioner, and by C the room temperature. Note that the air conditioner automatically adjusts with the outside temperature, it attempts to keep the room temperature the same no matter how the outside temperature changes. The actual causal relationships are: A causes both B and C, B causes C. Now, if we model the causal graph without A, we would make a false conclusion that B and C are independent (because the room temperature remains the same when the intensity of the air conditioner changes). Moreover, although B is not a confounder of A and C, the absence of B in the graph (which contains A and C) also leads to a false conclusion that A and C are independent.

Third, even though we are allowed to include our domain knowledge into the SBCN in terms of the temporal order, the proposed method does not allow to skip temporal ordering the attributes that we are not sure of. In practice, there might be cases in which we do not have enough knowledge to deduce the ordering of some attributes. This, together with the fact that this ordering is very important for the reconstruction of the SBCN, result in a big risk factor. While in fact, A happens before B, if we mistakenly order A after B, not only may the direction of causation between them is reversed but also are many other consequences to other nodes.

Forth, the relationships between attributes with equal temporal order are complicated. For two nodes a and b that have the same ordering value, it is not always easy to make the decision of whether to conclude that a causes b or b causes a using only observational data.

Fifth, it does not have any means to handle numerical attributes explicitly but requires them to be transformed to categorical somehow, most likely through binning. Since this transformation is lossy, there is a high chance that some useful information is not preserved. In addition, the number of values per attribute (either that attribute is truly categorical or a transformed version of a numerical attribute) has to be relatively small compared to the sample size for the statistical inference to be reliable. More precisely, each pair of <*attribute*=*value*> needs a significant amount of support. Big size is also a desired (but not necessarily enough) property for the assumption of uniformly randomly corrupted data.

Sixth, the weight of the edge going from any node u to v, denoted by $W(v, u)$ and computed by $W(v, u) = p(v|u) - p(v|\neg u)$ only considers the pairwise dependency between u and v, but not any joint dependency in which u and another node u' can together cause v but neither of u and u' can cause v alone. Consider the XOR function as an example, even though the two input values determine the output, none of them can affect the output on its own. The SBCN will eliminate both edges from the inputs to the output, resulting in a false negative.

Seventh, as hill-climbing is a greedy method, using it to optimize for the subset of edges in the SBCN makes no guarantee that the resulting graph is optimal in general. Furthermore, it also does not give a lower bound on how good its result is compared to the optimal one.

Eighth, the choice of the distribution to measure the likelihood of the data given the graph (i.e. $LL(D'|G')$ where D' is derived from the original data and G' is the current graph) needs to be chosen carefully. Only when an appropriate distribution is used should we expect a good resulting causal graph.

Ninth, although the authors claim that individual and subgroup discrimination can be detected via SBCN with the personalized PageRank algorithm, this approach suffers from the problem of joint causality. The personalized PageRank basically simulates and estimates which outcome is more likely to be reached from any of the set of starting nodes. It does not consider how multiple starting nodes jointly determine the outcome, which is desirable for detecting individual discrimination.

Tenth, it is worth it to note that this method only detects discrimination but does not ensure fairness on its own. Also, the SBCN does not output all discrimination in the data but only allows us to query if any user-inputted group/subgroup/individual is discriminated or not.

3.4 In contrast to statistical parity

For simplicity, let us assume a binary classification problem with a protected group (S) and its complement (S^C) as similar to the above. To test for group fairness with the SBCN, we simulate walks from the nodes representing S and S^C and count how many times they reach the nodes represent positive and negative outcomes. Call by t the number of simulations for each of S and S^C, S_+ and S^C_+ the number of simulations starting from S and S^C that reach the positive outcome before the negative outcome, respectively. Typically, $\frac{S_+}{t} \approx \frac{S^C_+}{t}$ implies group fairness. As a corner case, we also consider it fair if there is no path to go from either S or S^C to either positive or negative outcomes.

The problem is that this approach suffers from the same problem as statistical parity: despise the use of a causal graph, it only tests for whether the same proportion of all groups reaches each outcome, rather than ensuring every individual is treated right. Thus, the critiques from Dwork et al. [2] can also be applied here. The three concerns about reduced utility, self-fulfilling prophecy, and privacy violation remain even if the SBCN claims the classification is fair for the protected group.

To illustrate, let us take Dwork's example:

> "Suppose in the culture of S the most talented students are steered toward science and engineering and the less talented are steered toward finance, while in the culture of S c the situation is reversed: the most talented are steered toward finance and those with less talent are steered toward engineering. An organization ignorant of the culture of S and seeking the most talented people may select for "economics," arguably choosing the wrong subset of S..." [2]

In this scenario, the resulting SBCN will learn from the (flawed) decisions that pursuing finance and being talented are strongly associated with disregard to group membership, and thus it is fair. However, in fact, it is unfair for members of S who are talented and were steered toward science and engineering.

4 Fairness in ranking problems

In this section, we reason about the use of statistical parity and SBCN to ensure fairness on ranking problems. Let us assume that we only have these 2 tools at hand.

In some scenarios, like the simplified problem of selecting students for a seminar course, because the number of seats for each seminar is basically fixed with some small variation, it makes sense to rank the applicants by suitability and then select the top k according to the desired number of positions. For this type of ranking problem, in which order is just a proxy for classification (applicants are accepted or rejected), just applying statistical parity and/or SBCN on the classification outcomes may be an easy but sufficient way.

For other (actual) ranking problems, it is not obvious how these 2 methods can work out. One potential solution may be to distribute the instances into bins by their ranking outcome and measure fairness in each of those bins. The bins may be disjoint (e.g. the first bin contains instances in top 1%, the second bin contains instances from below top 1% up to top 2%, and so on) or not (e.g. the first bin contains instances in top 1%, the second bin contains instances in top 2%, and so on). The bin size may be either relative (e.g. top 1%, top 2%) or absolute (e.g. top 10, top 100) and different bins may have different sizes (e.g. the first bin contains the first 1%, the second bin contains the first 5%, the third bin contains the first 20%). The resulting conclusion on fairness is then aggregated from all bins, with bins of the higher rankings may have more weights than bins of lower rankings, depending on the specific problem. Note that both statistical purity and SBCN can be used to measure fairness with this approach.

References

[1] F. Bonchi, S. Hajian, B. Mishra, and D. Ramazzotti. Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics*, 3(1):1–21, 2017.

[2] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.