Economic Models of (Algorithmic) Discrimination

Introdução

O objetivo do paper em questão é fazer uma ponte conceitual entre as diferentes formas que o conceito de discriminação algorítmica é tratada nas ciências sociais e exatas.
O autor começa o paper demonstrando a importância de algoritmos para as decisões tomadas atualmente na sociedade e como eles possuem alto impacto no bem-estar da população através da distribuição de recursos que eles impactam.

Ele explica que o cenário de pesquisa atual em discriminação algorítmica se divide entre duas vertentes distintas: computação e ciências sociais. E que elas focam no lado técnico ou aspectos sociais, éticos e legais, respectivamente. E apresenta os problemas dos trabalhos de cada área em relação a análise de discriminação algorítmica: enquanto o primeiro foca nos aspectos técnicos e deixa os impactos sociais de lado, o segundo consegue ser mais abrangente mas não consegue distinguir os impactos de diferentes algoritmos e seus diferentes tipos.

Com isto, ele explica a necessidade de uma forma de pesquisar algoritmos discriminatórios que seja sensível tanto aos aspectos técnicos quanto sociais. É aqui que ele propõe utilizar economia, dado seu histórico em utilizar métodos formais para explicar fenômenos sociais.

O autor utiliza um survey realizado por Romei e Ruggieri que atravessa as literaturas de ciências sociais, computação, direito e economia. E neste survey eles criam dois modelos formais de discriminação em economia:

1. O da teoria baseada em gosto ou “irracionais
2. E teorias estatísticas ou “racionais”

O paper desenvolve um modelo para cada caso.

Teoria baseada em gosto

Para explicar a teoria baseada em gosto, o autor utiliza o exemplo de um chefe que favorece um grupo a sobre um grupo b na hora da contratação se baseando em fatores que não são necessários para a contratação.

Ele explica que um algoritmo irá “adquirir” um preconceito se as variáveis alvo no dataset refletem o tratamento discriminatório. Nesta parte, ele explica de forma bem coerente através de uma função, como este preconceito existe e como pode ser calculado.
Utilizando um problema onde a única variável para determinar a qualidade de um profissional seja um coeficiente de produtividade. Um gerente discriminador poderia dar um coeficiente menor para

profissionais não-brancos N comparado à profissionais brancos W, mesmo que eles possuam, de fato, a mesma produtividade.
Desta forma, o coeficiente Y seria dado a partir da seguinte fórmula:

$$Y = XB - aZ + e$$

onde B e a são coeficientes positivos, X é a medida de produtividade e Z=1 indica que o profissional não é branco, e e é bias.

Logo, percebe-se que para o modelo possuir máxima acurácia, ele deve absorver esse bias discriminatório do gerente, e de certa forma, propagando este preconceito em suas predições.

Ele explica que uma possível consequência é que empresas que utilizem um algoritmo discriminatório fiquem em desvantagem por não conseguirem os funcionários mais produtivos.
Contudo, a teoria baseada em gosto faz duas uso de duas premissas que nem sempre são cumpridas:

- Pertencimento ao grupo é independente da variável-alvo (no exemplo, produtividade)
- Empregadores preferem grupos explicitamente

Em vários casos, ambas premissas são falsas.
Com estes argumentos, o autor argumenta de que algo mais sólido é necesśario como modelo.

Teoria estatística

Este teoria se baseia no fato de o algoritmo não possui acesso à todos os dados que sejam estatisticamente capaz de determinar a variável-alvo. Isto aborda discriminação como uma “solução para o problema de extração de sinais”.

Ela se difere da baseada em gosto pelo fato de que não há uma preferência a priori. E sim uma preferência como consequência da falta de dados que suportem outro cenário ou análise.

O autor dá o exemplo de um prisioneiro com variáveis baseadas em comportamento e outra baseadas em fatores demográficos. Onde a variável-alvo é a chance de reincidência do criminoso.
Ele explica que, por exemplo, se um bom prisioneiro em questões de comportamento pertencer à um grupo demográfico de pessoas que na base de treinamento utilizada forem todas reincidentes, ele provavelmente será considerado reincidente e terá benefícios cortados por isto, apesar de sabermos que o seu lugar de nascimento não determina a variável-alvo em questão.
Esse é um problema de “sub representação” já é reconhecido na comunidade, o autor passa a próxima seção inteira explicando um caso sobre o assunto que não é relevante para o entendimento do assunto.

Contudo, ao final da explicação, ele explica em um parágrafo exclusivo a diferença entre dois possíveis cenários que o algoritmo pode se tornar discriminatório por dois motivos: ele apresenta bias sobre uma população e o outro é no caso onde ele possui capacidades diferentes de classificação para os diferentes grupos, ou seja, ele é melhor classificador para uns do que para outros.

Na próxima seção o autor busca explorar o conceito de Active Learning Algorithms, ou seja, ao invés de classificar os elementos em “batches” e utilizar somente uma fonte de dados. O algoritmo em questão além de classificar um novo elemento, ele utiliza este elemento classificado num futuro próximo como aprendizado. E isto leva à vários problemas que aprendizagem em batch não possui, como por exemplo, se um grupo já era discriminado no dataset original, o algoritmo pode reforçar esse padrão nas suas predições e se retroalimentar aumentando ainda mais o preconceito já existente, tornando os padrões de um certo grupo desfavorecido a ainda mais desfavorecido.

Outro problema que ele aponta é o fato de que existem problemas onde as decisões dos algoritmos só serão sentidas ao longo do tempo, algo que pode ter um impacto duradouro na sociedade, como por exemplo, o ato de doar crédito, se um determinado grupo a possuir características que foram inicialmente negativas para dar créditos mas depois se tornem aptos e alguns dos seus membros comecem a pagar os empréstimos que coletaram, o algoritmo terá sido treinado em um grupo desfavorecido e existem uma distância temporal enorme até que uma quantidade significativa deste grupo se torne bons pagadores.

Conclusão

O paper explora muitos exemplos e utiliza de funções simples para demonstrar seus pontos, o que é interessante além de abordar um único tema durante o decorrer do texto, o que ajuda na compreensão e na análise do mesmo. Contudo, ao divagar entre active learning e batch learning, o autor deixa pouco espaço para uma discussão mais específica, dado que o primeiro é muito bem abordado enquanto o segundo, não.

Além de não ter explorado outras formas de discriminação citadas na literatura, como o próprio autor menciona na conclusão do texto.