Gabriel Lucas Silva Machado

Métricas para *fairness* em ciência de dados

Belo Horizonte

2018/1

Gabriel Lucas Silva Machado

Métricas para *fairness* em ciência de dados

Proposta de Pesquisa Científica para o trabalho da disciplina Projeto Orientado em Computação I do curso de Bacharelado em Ciência da Computação da UFMG

Univerisdade Federal de Minas Gerais - UFMG

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Orientador: Mário Sérgio Alvim

Belo Horizonte

2018/1

Sumário

1 Introdução

A grande popularidade da *internet* e o amplo avanço tecnológico nos últimos anos causaram a viabilidade e necessidade de se ter um *smartphone* e um computador pessoal no mundo todo. A acessibilidade a essas tecnologias promoveu o crescimento do volume de dados coletados e consequentemente causou o início de uma nova era, intitulada *big data*. (WAMBA et al., 2015)

A importância do dilúvio de dados coletados mundialmente levou à ascensão da ciência de dados, que nada mais é que o estudo de como analisar, extrair e interpretar informações referentes a um conjunto de dados. Os resultados vindo dessas técnicas acabaram sendo utilizados para a criação de ferramentas de recomendação ou até mesmo de tomada de decisões. (TECHTARGET, 2017)

Graças à ciência de dados muitos problemas de natureza computacional difícil puderam ser superados utilizando técnicas de mineração de dados e aprendizado de máquina. Devido aos típicos bons resultados alcançados por essas técnicas em uma grande diversidade de problemas, estas começaram a ser usadas em aplicações que afetam diretamente a vida das pessoas, como por exemplo *credit score*, avaliação de CV, entre outros. Como grande parte dessas técnicas funcionam explorando padrões históricos nos dados, quando isso é aplicado usando dados de pessoas é possível que o algoritmo aprenda a tomar decisões baseadas em qualquer característica delas, o que pode vir a se tornar uma mera discriminação. (ŽLIOBAITė, 2017)

Tendo esse problema em mente, a proposta para esse trabalho consiste em uma organização de conhecimento sobre como é o estado da arte referente às técnicas utilizadas para detecção e prevenção de ocorrências de discriminação nos algoritmos de classificação (*fairness*). Especificamente, o trabalho consistirá no estudo individual das principais técnicas, com o propósito de entender como funcionam, em quais situações e maneiras cada uma delas pode ser utilizada e suas vantagens e desvantagens.

2 Referencial Teórico

No contexto de algoritmos de classificação, a discriminação é definida por tratamentos que desfavorecem alguém por alguma característica "protegida", isto é, raça, gênero, deficiência, idade, entre outras. Essas discriminações podem ocorrer em duas formas distintas denominadas direta e indireta. (ŽLIOBAITė, 2017) A direta ocorre exclusivamente por causa de atributos protegidos, ou seja, o algoritmo toma decisões baseando-se diretamente nessas características. Já a indireta ocorre através de atributos não protegidos que são correlacionados com características protegidas. (TRIPATHI, 2014)

A solução para o problema da discriminação direta é trivial, pois basta retirar as características protegidas dos dados, de tal forma que o algoritmo não poderá usá-las ao tomar uma decisão. Apesar disso, a exclusão das características protegidas não é uma prática usual, pois ao fazer isso torna-se impossível detectar e corrigir a discriminação indireta.

Para detectar e quantificar a discriminação indireta é necessário o uso de métricas que avaliam comparativamente grupos favorecidos e desfavorecidos e detectam quando as decisões do algoritmo foram prejudicadas por um atributo correlacionado a uma característica protegida.

Tendo em mente o significado contextual da discriminação direta e indireta, é possível definir outros conceitos que serão pertinentes no trabalho. O primeiro desses conceitos consistem em tomadores de decisões justos, que nada mais são que algoritmos em que as escolhas feitas independem das características protegidas, isso é, toda e qualquer decisão tomada é garantidamente livre de disciminação direta e indireta. Em contrapartida, um tomador de decisão é dito como injusto quando em pelo menos uma tomada de decisão existe algum tipo de discriminação. Bons exemplos de algoritmos injustos são: um classificador de CV que rotula indivíduos extremamente semelhantes (se diferem somente na raça, por exemplo) em grupos distintos (discriminação direta), ou até mesmo um *credit scorer* cuja pontuação rejeita qualquer pessoa que more em um determinado bairro, sendo este tipicamente habitado por pessoas negras (discriminação indireta).

O estudo das métricas para detecção de discriminação indireta consiste em um dos objetivos dessa pesquisa. Por se tratar de um trabalho de organização de conhecimento, é natural encontrar alguma semelhança com outros trabalhos na literatura, e isso ocorre com o *survey "Measuring discrimination in algorithmic decision making"* (ŽLIOBAITė, 2017), que faz um breve resumo sobre o assunto desse trabalho. Apesar disso, este projeto se diferencia ao entrar em detalhes não contidos no *survey*, como por exemplo as vantagens e desvantagens de cada métrica.

3 Metodologia

A metodologia proposta para a realização deste trabalho pode ser dividida em 3 fases:

1. Leitura de textos e artigos científicos sobre métricas específicas: nessa etapa será feita uma pesquisa e leitura de textos e artigos específicos sobre as métricas contidas no *survey "Measuring discrimination in algorithmic decision making"* (ŽLIOBAITė, 2017). Ao final dessa fase é esperado obter as informações necessárias sobre cada métrica (como funciona, em quais situações e maneiras ela pode ser utilizada e suas vantagens e desvantagens). .

2. Organização do conhecimento adquirido: nesse estágio será definido como será a organização estrutural do artigo. Além disso, serão escritos resumos sobre as informações adquiridas na fase 1, que serão usados posteriormente para a formulação do artigo.

3. Escrita do artigo científico: nessa parte será feito a escrita da monografia no formato de um artigo.

4 Resultados Esperados

Como dito anteriormente, esse projeto consiste em uma organização de conhecimento referente às métricas para fairness em ciência dos dados.

Ao final desse trabalho, que engloba a disciplina POC1, espera-se obter um artigo científico no qual conterá uma análise individual sobre as principais métricas do estado da arte para medir fairness. Nesse estudo, as principais informaçoes que se pretende obter são como a métrica funciona, quando deve ser usada, como utilizá-la no contexto de algoritmos de classificação e suas vantagens e desvantagens.

5 Etapas e Cronogramas

Como especificado na metodologia, o trabalho se dividirá em 3 etapas, sendo estas:

- A: Leitura de textos e artigos científicos sobre métricas específicas;
- B: Organização do conhecimento adquirido;
- C: Escrita do artigo científico.

Levando em consideração o calendário proposto para a disciplina e os itens acima, a tabela 1 indica o planejamento de como o trabalho será realizado.

Tabela 1: Cronograma

Semana	**Abril**	**Maio**	**Junho**
1	A	A	C
2	A	A	C
3	A	B	C
4	A	B	C

Referências

TECHTARGET. *Data Science*. 2017. Disponível em: <http://searchcio.techtarget.com/definition/data-science>. Citado na página 3.

TRIPATHI, K. Analysis of direct and indirect discrimination discovery and prevetion algorithms in data mining. 2014. Citado na página 4.

WAMBA, S. et al. How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, v. 165, p. 234–246, 2015. Citado na página 3.

ŽLIOBAITė, I. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, v. 31, p. 1060–1089, 2017. Citado 3 vezes nas páginas 3, 4 e 5.