

Método busca reduzir vieses em IA médica ao combinar desempenho e múltiplas métricas de equidade

Objetivo é reduzir os desequilíbrios de gênero, etnia-raça e faixa etária nos diagnósticos médicos mediados por aprendizado de máquina

José Tadeu Arantes | Agência FAPESP – Modelos de inteligência artificial são usados para apoiar tarefas como triagem e predição de risco em saúde. Mas, quando treinados com bases de dados que não representam de forma equilibrada diferentes grupos populacionais, podem apresentar vieses que comprometem a equidade dos resultados em relação a gênero, grupos étnico-raciais e faixas etárias.

Um estudo realizado no Instituto de Ciência e Tecnologia da Universidade Federal de São Paulo (ICT-Unifesp), campus de São José dos Campos, com parcerias internacionais propõe uma estratégia de otimização que incorpora múltiplas restrições relativas à equidade durante o treinamento do modelo, com o objetivo de reduzir disparidades sem prejudicar o desempenho global. Artigo sobre o trabalho foi publicado no periódico Applied Soft Computing.

“Os algoritmos de aprendizado de máquina precisam de dados de treinamento. E, em muitos cenários reais, esses dados são desbalanceados em relação a gênero, grupos étnico-raciais e faixas etárias. Como o algoritmo aprende melhor com os dados mais abundantes, seus resultados tendem a ser menos precisos para mulheres, negros ou indígenas e pessoas com idades fora do padrão médio. Isso pode ser crítico no campo da saúde, porque estamos lidando com diagnósticos e vidas humanas”, diz Lilian Berton, professora do ICT-Unifesp e coordenadora da pesquisa. Ela atua com aplicações de aprendizado de máquina e inteligência artificial em saúde, redes sociais, dados climáticos, finanças e indústria 4.0.

O estudo insere-se na área de fairness, termo ainda sem tradução consolidada em português. “A tradução mais usada e mais aceita no Brasil é ‘equidade’. Tem pessoas usando as expressões ‘equidade algorítmica’ ou ‘justiça algorítmica’. Para alcançar essa meta, estamos propondo uma otimização multiobjetivo com mais de uma restrição em relação à equidade”, afirma Berton.

A maioria dos modelos “clássicos” busca maximizar acertos globais, como a acurácia. Mas isso não filtra os vieses. Por outro lado, as técnicas de mitigação de vieses frequentemente deterioram o desempenho médio. Por isso, encontrar o ponto ótimo era um problema aberto na literatura. O estudo buscou preencher essa lacuna. “Nossa contribuição foi tratar o ajuste como um problema multiobjetivo: o treinamento passa a equilibrar, ao mesmo tempo, desempenho e mais de uma métrica de equidade. É como se eu tivesse, por exemplo, três objetivos ao mesmo tempo: o desempenho, a métrica de equidade 1 e a métrica de equidade 2”, explica a pesquisadora.

Em aprendizado de máquina, as taxas básicas são: verdadeiros positivos (VP), no caso, pessoas doentes corretamente identificadas como doentes; falsos positivos (FP), pessoas saudáveis classificadas como doentes; verdadeiros negativos (VN), pessoas saudáveis identificadas como saudáveis; e falsos negativos (FN), pessoas doentes classificadas como saudáveis. Cada uma dessas taxas define uma métrica distinta.

“O algoritmo pode atribuir mais falsos positivos para um recorte demográfico do que para outro. Isso é um viés. Pode também atribuir mais falsos negativos para o mesmo ou para outros recortes. Isso é outro viés. Ambos os comportamentos estão relacionados com vieses. Mas são acessados por meio de métricas distintas. Nosso estudo, como foi dito, conjugou três objetivos: o desempenho e pelo menos duas métricas de equidade diferentes”, reitera a cientista.

O método foi avaliado em três conjuntos públicos de diagnósticos: arritmia, com 452 pacientes; diabetes, com 1.635 pacientes; e hemorragia intracraniana, com 526 pacientes. Foram aferidos, conforme o caso, atributos de gênero, etnia e idade.

“Nos experimentos iniciais sem restrições relativas à equidade, o conjunto arritmia apresentou uma diferença de paridade demográfica de 0,4543 entre grupos definidos por gênero. Ao aplicar o modelo com restrições simultâneas de gênero e idade, essa diferença foi reduzida para valores inferiores a 0,1, o que corresponde a uma redução aproximada de 78% na disparidade. No caso de diabetes, a disparidade inicial associada à etnia era de 0,0110. Com a introdução de múltiplas restrições, o valor caiu para menos de 0,005, representando uma redução superior a 54% no viés racial, sem degradação observável nas métricas de desempenho do classificador. No conjunto hemorragia intracraniana, a abordagem também apresentou melhorias estatisticamente significativas, sugerindo robustez em cenários de dados limitados”, relata Berton.

Para verificar se esses ganhos não eram fruto de flutuações aleatórias, os autores aplicaram o teste não paramétrico de Wilcoxon às diferenças observadas entre os modelos. Trata-se de um recurso, proposto pelo estatístico norte-americano Frank Wilcoxon (1892-1965), que verifica se a disparidade observada entre dois conjuntos de resultados é real ou apenas produto do acaso. Em praticamente todos os cenários analisados, os valores indicaram confiança estatística superior a 99% de que as melhorias em justiça e equilíbrio resultaram efetivamente da metodologia de múltiplas restrições.

Segundo Berton, os experimentos foram executados de acordo com o custo computacional: “A gente, às vezes, usa nuvens gratuitas, como, por exemplo, do Google Colab, que oferece acesso a recursos computacionais especialmente adequados para aprendizado de máquina, ciência de dados e educação. Em outros casos, quando necessário, recorreremos a clusters de alto desempenho, como o Santos Dumont, aqui no Brasil”.

Os resultados indicam que tratar a equidade algorítmica como um problema de otimização interseccional permite avançar além de soluções pontuais, frequentemente restritas a um único atributo sensível. A metodologia, conjugando desempenho e equidade, oferece maior confiabilidade e robustez para decisões clínicas apoiadas por algoritmos, enfatiza a pesquisadora. O estudo, no entanto, limita-se a modelos lineares de regressão logística e a dados tabulares. A extensão da abordagem para modelos não lineares, como redes neurais profundas, e para outros tipos de dados, como imagens médicas ou séries temporais, permanece como desafio para trabalhos futuros.

O projeto contou com apoio da FAPESP, por meio de auxílio à pesquisa e bolsa de pós-doutorado concedida a Maira Blumer Fatoretto, primeira autora do estudo.

O artigo *Optimizing fairness and utility in healthcare machine learning models* pode ser lido em: [sciencedirect.com/science/article/abs/pii/S1568494625007379](https://www.sciencedirect.com/science/article/abs/pii/S1568494625007379).

<https://agencia.fapesp.br/metodo-busca-reduzir-vieses-em-ia-medica-ao-combinar-desempenho-e-multiplas-metricas-de-equidade/57003>

Veículo: Online -> Agência de Notícias -> Agência de Notícias - Agência FAPESP