Análise Teórica do Impacto de Dados Faltantes em Atributos Sensíveis sobre a Métrica de Fairness *p%-rule*

Dimas Cassimiro Nascimento, Daliton da Silva, Luis Filipe Alves Pereira

[1]Universidade Federal do Agreste de Pernambuco (UFAPE)
Garanhuns – PE – Brasil

{dimas.cassimiro,daliton.silva,luis-filipe.pereira}@ufape.edu.br

Abstract. *Algorithmic fairness assessment has become a central topic in the analysis of automated decision-making systems. In this paper, we investigate how the use of imputation techniques to fill in missing sensitive data can affect the value of the* p%-rule *metric, depending on the degree of imputation error. We propose a formal mathematical model to quantify this impact, considering symmetric and bidirectional imputation errors. Additionally, we analyze the minimum number of manual corrections required to achieve a desired improvement in the fairness metric. The results provide a quantitative foundation for understanding the trade-offs between the cost of manual intervention and the quality of imputed data in algorithmic fairness analysis.*

Resumo. *A avaliação da justiça algorítmica tem se tornado um tópico central na análise de sistemas de decisão automatizados. Neste artigo, investigamos como o uso de técnicas de imputação para preencher dados sensíveis ausentes pode afetar o valor da métrica* p%-rule*, dependendo do grau de erro na imputação. Propomos um modelo matemático formal para quantificar esse impacto, considerando erros de imputação simétricos e bidirecionais. Além disso, determinamos a quantidade mínima de correções manuais necessárias para garantir uma melhoria desejada na métrica de fairness. Os resultados oferecem uma base quantitativa para entender os trade-offs entre custo de intervenção manual e qualidade dos dados imputados na análise de justiça algorítmica.*

1. Introdução

A equidade em modelos de decisão automatizada tem se consolidado como um tema central na pesquisa em Inteligência Artificial e Aprendizagem de Máquina (AM) [Agarwal and Mishra 2021], sobretudo devido ao impacto crescente desses sistemas em decisões sensíveis que afetam diretamente a vida das pessoas, como concessão de crédito, processos seletivos, alocação de benefícios e decisões judiciais [Barocas and Selbst 2016, Mehrabi et al. 2021, Mitchell 2021]. O crescente interesse acadêmico e social na avaliação de *fairness* tem motivado o desenvolvimento de métricas, algoritmos de mitigação e ferramentas específicas para esse fim. Justiça algorítmica em AM consiste na capacidade desses modelos em evitar predições enviesadas [Oliveira et al. 2024].

Um aspecto frequentemente negligenciado na avaliação de justiça algorítmica é a qualidade dos dados, particularmente a presença de valores faltantes no atributo sensível, que é aquele utilizado para distinguir grupos socialmente relevantes, como gênero, raça

ou etnia [Fernando et al. 2021]. A ausência de dados sensíveis ocorre por diferentes motivos: restrições legais, omissões voluntárias por parte dos indivíduos, erros operacionais ou até falhas estruturais no processo de coleta. Como consequência, a etapa de pré-processamento dos dados frequentemente recorre a técnicas de imputação para preencher esses valores ausentes.

No entanto, diversos estudos recentes destacam que a imputação de dados sensíveis pode introduzir distorções significativas nas métricas de *fairness*, comprometendo a validade das avaliações [Wang and Singh 2021, Fernando et al. 2021]. Embora a maior parte das técnicas de mitigação de viés parta do pressuposto de que os dados estão completos, na prática, conjuntos de dados do mundo real frequentemente apresentam valores ausentes não aleatórios (*Missing Not At Random* – MNAR), cuja distribuição está correlacionada aos próprios atributos sensíveis [Little and Rubin 2019]. Wang e Singh [Wang and Singh 2021] mostraram empiricamente que tanto os valores ausentes quanto o viés de seleção impactam de maneira não trivial as métricas estatísticas de justiça, como a *p%-rule*, frequentemente utilizada para avaliar disparidades entre grupos protegidos.

Motivado por esse contexto, este artigo propõe um modelo matemático formal que quantifica como erros de imputação no atributo sensível afetam diretamente a métrica *p%-rule*. Além disso, apresentamos um modelo analítico para calcular a quantidade mínima de correções manuais necessárias para garantir uma melhoria desejada na métrica de *fairness*, estabelecendo um trade-off quantitativo entre o custo da intervenção manual e a qualidade dos dados imputados. Nosso trabalho contribui para preencher uma lacuna na literatura ao oferecer ferramentas teóricas para avaliar o impacto de dados faltantes na análise de justiça algorítmica.

2. Definições

2.1. Conjunto de Registros e Valores Faltantes

Seja um conjunto de dados D contendo n registros. Cada registro é representado como um vetor de atributos (X, S, Y), onde: X é o conjunto de atributos não sensíveis; S é o atributo sensível, binário, assumindo valores $S \in \{0, 1\}$, que representam, respectivamente, o grupo não privilegiado e o grupo privilegiado; Y é o atributo alvo de classificação, binário, assumindo valores $Y \in \{+, -\}$, onde $Y = 1$ representa o desfecho positivo (e.g., aprovação de crédito ou aprovação em um processo de seleção).

2.2. Métrica de Fairness *p%-rule*

A métrica *p%-rule* [Feldman et al. 2015, Zafar et al. 2017], também conhecida como *disparate impact*, mede a equidade de um classificador com relação ao atributo sensível. Seja $P(\hat{Y} = +|S = s)$ a probabilidade de um indivíduo do grupo $S = s$ receber uma predição positiva, a métrica $p\%{-}rule$ é definida como:

$$p\%{-}rule = \min\left(\frac{P(\hat{Y} = +|S = 1)}{P(\hat{Y} = +|S = 0)}, \frac{P(\hat{Y} = +|S = 0)}{P(\hat{Y} = +|S = 1)}\right)$$

Valores baixos indicam maior disparidade entre os grupos, enquanto valores próximos de 1 (ou 100% quando expressos como percentual) indicam maior equidade.

2.3. Método de Imputação de Valores Faltantes

Seja $\mathcal{I}$ um método de imputação utilizado para preencher os valores faltantes no atributo sensível S dos registros em D. O método $\mathcal{I}$ gera, para cada registro com valor ausente de S, uma predição $\hat{S} \in \{0, 1\}$, que corresponde ao valor imputado para o atributo sensível.

Seja $\mathcal{M} \subset D$ o subconjunto de registros que possuem o valor do atributo S faltante, tal que $|\mathcal{M}| = m$. Cada registro $x_i \in \mathcal{M}$ possui um valor verdadeiro, embora desconhecido, $S_i \in \{0, 1\}$, e um valor imputado $\hat{S}_i = \mathcal{I}(x_i)$.

Definimos o erro de imputação e como a soma das taxas de erro nas duas direções: $e = e_0 + e_1$, tal que e_0 é a fração dos registros com verdadeiro $S = 0$ que foram erroneamente imputados como $S = 1$; e e_1 é a fração dos registros com verdadeiro $S = 1$ que foram erroneamente imputados como $S = 0$. Assumimos, para simplificação, que o erro é simétrico, i.e., $e_0 = e_1 = \frac{e}{2}$.

3. Problemas Investigados

Este artigo investiga os seguintes problemas:

1. **Impacto do Erro de Imputação na Métrica *p%-rule*:** Como a taxa de erro de imputação sobre um atributo sensível impacta a métrica de fairness *p%-rule*?
2. **Cálculo do Erro Máximo Tolerável:** Como determinar o erro máximo de imputação que permite que a métrica *p%-rule* aumente em $\alpha\%$?.
3. **Estimativa da Variação da *p%-rule*:** Como calcular a variação máxima da métrica *p%-rule* que pode ocorrer, dado uma taxa de erro de imputação e sobre os atributos sensíveis.
4. **Correções Manuais Necessárias:** Qual é a quantidade mínima de correções manuais k sobre registros imputados que é necessária para alcançar uma melhoria de $\alpha\%$ sobre a métrica *p%-rule*?

4. Análise do Impacto de Dados Faltantes na Métrica *p%-rule*

Seja um dataset com n registros, dos quais $m < n$ possuem o valor do atributo sensível S ausente. O atributo sensível assume dois valores binários: $S = 1$ (grupo privilegiado) e $S = 0$ (grupo não privilegiado). Sejam n_1 o número de registros com $S = 1$ e n_0 o número de registros com $S = 0$ no conjunto de dados onde o atributo sensível está disponível, tal que $n_1 + n_0 = n - m$. As proporções dos grupos no conjunto observado são:

$$p_1^{true} = \frac{n_1}{n_1 + n_0}, \quad p_0^{true} = \frac{n_0}{n_1 + n_0}$$

Seja uma técnica de imputação aplicada aos m registros faltantes, associada a uma taxa de erro de imputação igual a $e \in [0, 1]$. O erro é definido como a fração dos registros imputados cujo valor do atributo sensível foi atribuído incorretamente. Assumimos que o erro é **bidirecional e simétrico**, ou seja, metade dos erros corresponde a registros originalmente do grupo $S = 0$ imputados como $S = 1$, e a outra metade ao inverso.

O efeito da imputação dos valores faltantes nos tamanhos dos grupos é influenciado por dois fatores principais: (i) a quantidade de registros imputados corretamente

segundo as proporções reais dos grupos e (ii) o saldo líquido dos erros bidirecionais, resultante da troca de registros entre os grupos sensíveis.

Definindo $d = p_1^{true} - p_0^{true}$, os tamanhos dos grupos após o processo de imputação são calculados da seguinte forma: $n_1' = n_1 + m \cdot p_1^{true} \cdot (1-e) - m \cdot \frac{d}{2} \cdot e$ e $n_0' = n_0 + m \cdot p_0^{true} \cdot (1-e) + m \cdot \frac{d}{2} \cdot e$. Note que $n_i' = n_i$ + [imputações corretas para $S = i$]+[erros provenientes de $S = (1-i)$]−[erros produzidos para $S = (1-i)$]. Assim, o termo $m \cdot p_s^{true} \cdot (1-e)$ representa a quantidade de registros imputados corretamente para o grupo $S = s$, e o termo $m \cdot \frac{d}{2} \cdot e$ representa o saldo líquido dos erros bidirecionais, que favorece ou desfavorece um grupo dependendo da diferença de proporções d. Logo, a razão entre os grupos após o processo de imputação é dada por:

$$r(e) = \frac{n_0 + m \cdot p_0^{true} \cdot (1-e) + m \cdot \frac{d}{2} \cdot e}{n_1 + m \cdot p_1^{true} \cdot (1-e) - m \cdot \frac{d}{2} \cdot e}$$

4.1. Cálculo da métrica *p%-rule* após o processo de imputação

Sejam $a_1 = P(\hat{Y} = +|S = 1)$ e $a_0 = P(\hat{Y} = +|S = 0)$. A métrica *p%-rule* após imputação pode ser calculada como:

$$p_{new} = \min\left(\frac{a_1}{a_0} \cdot r(e), \frac{a_0}{a_1} \cdot \frac{1}{r(e)}\right)$$

Assumimos que o caso limitante no resultado da métrica ocorre no primeiro termo, i.e., $p_{new} = \frac{a_1}{a_0} \cdot r(e)$, almeja-se que: $p_{new} = p_{current} \cdot (1+\alpha)$, tal que $\alpha \in (0, 1]$ representa a porcentagem de aumento desejada na métrica *p%-rule*. Substituindo:

$$\frac{a_1}{a_0} \cdot r(e) = p_{current} \cdot (1+\alpha)$$

Isolando $r(e)$:

$$r(e) = \frac{a_0}{a_1} \cdot p_{current} \cdot (1+\alpha)$$

Substituindo a expressão de $r(e)$:

$$\frac{n_0 + m \cdot p_0^{true} \cdot (1-e) + m \cdot \frac{d}{2} \cdot e}{n_1 + m \cdot p_1^{true} \cdot (1-e) - m \cdot \frac{d}{2} \cdot e} = \frac{a_0}{a_1} \cdot p_{current} \cdot (1+\alpha)$$

Definindo:

$$C = \frac{a_0}{a_1} \cdot p_{current} \cdot (1+\alpha)$$

produz-se:

$$\Big(n_0 + m \cdot p_0^{true} \cdot (1-e)\Big) + m \cdot \frac{d}{2} \cdot e = C \cdot \Big(n_1 + m \cdot p_1^{true} \cdot (1-e)\Big) - C \cdot m \cdot \frac{d}{2} \cdot e$$

Agrupando os termos em e:

$$e \cdot m \cdot \frac{d}{2} \cdot (1+C) = C \cdot \Big(n_1 + m \cdot p_1^{true} \cdot (1-e)\Big) - \Big(n_0 + m \cdot p_0^{true} \cdot (1-e)\Big)$$

Portanto, a taxa de erro máxima e_{max} associada ao método de imputação para produzir $p_{new} = p_{current} \cdot (1+\alpha)$ pode ser calculada como:

$$e_{max} = \frac{2 \cdot (C \cdot (n_1 + m \cdot p_1^{true} \cdot (1-e)) - (n_0 + m \cdot p_0^{true} \cdot (1-e)))}{m \cdot d \cdot (1+C)} \tag{1}$$

4.2. Variação Máxima da *p%-rule* considerando um erro de imputação e

A métrica *p%-rule* após a aplicação do método de imputação é:

$$p_{new} = \frac{a_1}{a_0} \cdot r(e)$$

onde $r(e)$ é a razão calculada anteriormente. A variação máxima da métrica pode ser calculada como:

$$\Delta p\% = |p_{new} - p_{current}| \tag{2}$$

Se os grupos forem balanceados ($p_1^{true} = p_0^{true}$), então $d = 0$, e portanto:

$$r(e) = \frac{n_0 + m \cdot p_0^{true} \cdot (1-e)}{n_1 + m \cdot p_1^{true} \cdot (1-e)}$$

O que implica que o erro bidirecional não afeta a *p%-rule* nesse cenário. Por outro lado, quando há desbalanceamento ($d \neq 0$), o impacto de e cresce proporcionalmente a d.

4.3. Quantidade Mínima de Correções Manuais para Melhoria Percentual da Métrica *p%-rule*

Seja k o número de registros entre os m faltantes cujos valores do atributo sensível são corrigidos manualmente. O erro efetivo após k correções é reduzido proporcionalmente para:

$$e' = e \cdot \frac{m-k}{m}$$

Assim, a razão dos grupos sensíveis torna-se:

$$r(e') = \frac{n_0 + m \cdot p_0^{true} \cdot (1-e) + (m-k) \cdot \frac{d}{2} \cdot e}{n_1 + m \cdot p_1^{true} \cdot (1-e) - (m-k) \cdot \frac{d}{2} \cdot e}$$

Assumindo que almeja-se produzir:

$$\frac{a_1}{a_0} \cdot r(e') = p_{current} \cdot (1+\alpha)$$

e definindo $C = \frac{a_0}{a_1} \cdot p_{current} \cdot (1+\alpha)$, temos:

$$\Big(n_0 + m \cdot p_0^{true} \cdot (1-e)\Big) + (m-k) \cdot \frac{d}{2} \cdot e = C \cdot \Big(n_1 + m \cdot p_1^{true} \cdot (1-e)\Big) - C \cdot (m-k) \cdot \frac{d}{2} \cdot e$$

Isolando $(m-k)$:

$$(m-k) = \frac{2 \cdot (C \cdot (n_1 + m \cdot p_1^{true} \cdot (1-e)) - (n_0 + m \cdot p_0^{true} \cdot (1-e)))}{d \cdot e \cdot (1+C)}$$

Finalmente, isolando k:

$$k = m - \frac{2 \cdot (C \cdot (n_1 + m \cdot p_1^{true} \cdot (1-e)) - (n_0 + m \cdot p_0^{true} \cdot (1-e)))}{d \cdot e \cdot (1+C)} \tag{3}$$

Portanto, a Eq. (3) define o número mínimo k de correções manuais necessárias para permitir um aumento de $\alpha\%$ na métrica *p%-rule*, considerando a definição de erro de imputação bidirecional.

5. Conclusões e Trabalhos Futuros

Este artigo utilizou uma abordagem formal para analisar o impacto de valores faltantes em atributos sensíveis sobre a métrica de justiça algorítmica *p%-rule*. A análise desenvolvida demonstrou que o processo de imputação de valores ausentes no atributo sensível introduz incertezas que impactam diretamente a avaliação de fairness. Quantificamos como os erros de imputação, sob a suposição de erros simétricos e bidirecionais, afetam o valor da métrica *p%-rule*. Os modelos propostos oferecem uma base quantitativa robusta para apoiar profissionais e pesquisadores na compreensão dos trade-offs entre acurácia na imputação, custo de intervenções manuais e a confiabilidade das métricas de fairness calculadas em conjuntos de dados com informações sensíveis ausentes.

Como trabalhos futuros, pretende-se investigar o impacto de métodos de imputação probabilística, que incorporam a incerteza diretamente nos cálculos das métricas de fairness, bem como generalizar os modelos para atributos sensíveis com múltiplas categorias.

Referências

Agarwal, S. and Mishra, S. (2021). *Responsible AI*. Springer.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3):671–732.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268.

Fernando, M.-P., Cèsar, F., David, N., and José, H.-O. (2021). Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems*, 36(7):3217–3258.

Little, R. J. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Mitchell, S. e. a. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Communications of the ACM*, 64(5):58–66.

Oliveira, T. A., Oliveira, J. V., Farias, T. P., Cruz, E. W., Andrade, L. J., and Pita, R. (2024). Estudo experimental sobre justiça algorítmica aplicada em modelos de análise de crédito. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 29–36. SBC.

Wang, Y. and Singh, L. (2021). Analyzing the impact of missing values and selection bias on fairness. *International Journal of Data Science and Analytics*, 12(2):101–119.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pages 962–970. PMLR.