Justiça em Sistemas de Recomendação: Uma Análise de Técnicas de Regularização

Rodrigo O. Lima[1], **Giovanni Comarela**[3], **Fabiano Belém**[2], **Julio C. S. Reis**[1]

[1]Universidade Federal de Viçosa (UFV) – Brasil

[2]Universidade Federal de Minas Gerais (UFMG) – Brasil

[3]Universidade Federal do Espírito Santo (UFES) – Brasil

rodrigo.otavio@ufv.br, gc@inf.ufes.br, fmuniz@dcc.ufmg.br, jreis@ufv.br

Abstract. *Machine Learning strategies are increasingly used in decision-making processes in several areas of knowledge due to the abundance of data currently available. Although it was expected that such an application would solve the problem of human bias, it was noted that some models presented unfair behaviors in relation to groups historically discriminated by reflecting existing bias in the employed datasets. This problem has aroused great academic interest in recent years and several definitions, metrics and methodologies have been proposed to measure and ensure fairness in these contexts. One particular area is Recommendation Systems, where the objective is to recommend relevant items to users, and in some contexts it is not desirable for these recommendations to be associated with protected attributes of these users. This problem can be characterized as group fairness, in which groups of users are treated equally in the Recommendation System. In this work, we analyze the effectiveness of the use of group fairness regularization in a film recommendation system for men and women using two proposed metrics inspired by group fairness in classification. Empirical results show that this strategy improves results in relation to group fairness metrics and has a low impact on the final quality of the recommendation.*

Resumo. *Estratégias de aprendizagem de máquina são cada vez mais utilizadas em processos de tomada de decisões em diversas áreas de conhecimento devido à abundância de dados existentes atualmente. Embora esperava-se que tal aplicação solucionasse o problema da parcialidade inerente aos seres humanos, notou-se que alguns modelos apresentavam comportamentos enviesados em relação a grupos historicamente discriminados por refletirem preconceitos existentes nos conjuntos de dados utilizados. Este problema tem despertado grande interesse acadêmico nos últimos anos e diversas definições, métricas e metodologias têm sido propostas para garantir justiça nestes contextos. Uma área em particular é a de Sistemas de Recomendação, onde o objetivo é recomendar itens relevantes para os usuários e, em alguns contextos, não é desejável que estas recomendações estejam associadas a atributos protegidos destes usuários. Este problema pode ser caracterizado como justiça de grupo, na qual grupos de usuários são tratados igualmente no Sistema de Recomendação. Neste trabalho, analisamos a eficácia do uso de regularização com objetivo de justiça de grupo em um Sistema de Recomendação de filmes para homens e mulheres utilizando duas métricas propostas inspiradas em justiça de grupo em*

classificação. Os resultados empíricos mostram que esta estratégia melhora os resultados em relação às métricas e tem baixo impacto na qualidade final da recomendação.

1. Introdução

A popularização de soluções baseadas em estratégias de aprendizado de máquina se deve, em grande parte, à abundância de dados existentes atualmente[1]. Esta grande disponibilidade de dados permitiu o desenvolvimento de modelos precisos e generalizáveis, que são largamente utilizados em processos de tomada de decisões em diversas áreas de conhecimento (e.g., medicina, economia e direito). À medida que essa revolução ocorria, notou-se que alguns modelos apresentavam comportamentos enviesados em relação a grupos historicamente discriminados, principalmente por refletirem preconceitos históricos existentes nos conjuntos de dados utilizados por esses algoritmos[2]. Um exemplo deste comportamento é o sistema COMPAS, utilizado nos EUA, que previa a probabilidade de um detento cometer um novo crime. Após serem analisadas as previsões deste sistema para 7000 pessoas presas entre 2013 e 2014, e verificado quais crimes estas pessoas haviam cometido nos 2 anos seguintes, notou-se que réus negros que não cometeram crimes neste período receberam risco duas vezes maiores em relação a réus brancos[3].

Este problema tem despertado grande interesse acadêmico[4] nos últimos anos e diversas definições, métricas e metodologias têm sido propostas para garantir justiça[5] nestes modelos. Estas metodologias podem ser aplicadas no pré-processamento, durante o processamento, ou no pós-processamento destes modelos [Hajian et al. 2016], visando evitar discriminação nestas etapas. Uma abordagem de pré-processamento ingênua consiste em remover do conjunto de dados atributos protegidos (e.g., sexo, raça e idade), de forma que o modelo não consiga identificar e tratar parcialmente grupos de usuários. Porém, esta técnica é insuficiente devido às correlações existentes nestes conjuntos que permitem a identificação de atributos sensíveis pelo modelo [Kamishima et al. 2011]. Neste cenário, surge a necessidade de metologias mais sofisticadas para lidar com o problema e, consequentemente, o interesse em analisar a eficácia dessas propostas, com problemas de *classificação* ganhando maior atenção [Hardt et al. 2016, Dwork et al. 2012, Calders and Verwer 2010, L. Cardoso et al. 2019]. Outra área que ganhou interesse nos últimos anos é a de *Sistemas de Recomendação* [Brandão et al. 2013, Ko et al. 2022, Ahmadian et al. 2022].

Sistemas de Recomendação são utilizados para recomendar itens a usuários (e.g., filmes, livros, etc) de forma que os itens recomendados sejam de interesse do usuário. Dentro deste contexto, uma técnica em especial é a *Filtragem Colaborativa* [Ekstrand et al. 2011], onde as recomendações são baseadas no comportamento de usuários similares. Em Sistemas de Recomendação, justiça pode ser definida para itens ou para usuários. Em relação aos itens, existe o viés de popularidade [Celma and Cano 2008], na qual itens menos conhecidos são menos recomendados. Existe também viés em relação aos fornecedores dos itens, onde itens associados

[1] https://www.sciencedaily.com/releases/2013/05/130522085217.htm
[2] https://bit.ly/3HnYTnZ
[3] https://bit.ly/3mKp2nA
[4] https://facctconference.org/
[5] Tradução livre do termo *fairness* encontrado na literatura.

a grupos discriminados são menos recomendados que itens de grupos não discriminados [Ekstrand and Kluver 2021] (e.g. livros de autores mulheres serem menos recomendados que de autores homens). Em relação aos usuários, o conceito de justiça está associado a contextos onde não é desejável que as recomendações recebidas estejam associadas a atributos protegidos destes usuários (e.g. recomendar vagas de emprego conforme o sexo, ou recomendar moradias em bairros conforme a raça) [Burke et al. 2018]. Neste caso, o problema pode ser abordado como justiça individual, onde cada usuário é tratado de forma justa em relação aos outros, ou em justiça de grupo, onde grupos de usuários são tratados de forma justa. Neste trabalho, exploramos a justiça de grupo para usuários do sexo masculino e feminino, no contexto de um Sistema de Recomendação de filmes. Entendemos que nossos resultados podem ser generalizados para outros casos socialmente relevantes.

Mais precisamente, a partir de um Sistema de Recomendação de filmes utilizando Filtragem Colaborativa com decomposição de matrizes, analisamos a justiça de grupo antes e após aplicada regularização no treinamento com algum objetivo de justiça. Para tal, exploramos um conjunto de dados desbalanceado em relação ao gênero dos usuários e avaliamos duas métricas propostas no contexto de recomendação, *Equal Opportunity* e *Demographic Parity*, inspiradas em justiça de grupo para o problema de classificação. Com este experimento, queremos responder às seguintes perguntas de pesquisa (P):

P1: Qual a eficácia de se utilizar as regularizações propostas para garantir justiça de grupo em Sistemas de Recomendação?

P2: Qual o impacto na eficácia da recomendação após a aplicação destas regularizações?

De forma geral, os resultados mostram o potencial de aplicação de técnicas de regularização sem prejuízo significativo à qualidade da recomendação. Além disso, comparamos os resultados em níveis diferentes de esparsidade no conjunto de dados, e verificamos que os resultados se mantiveram. Estes resultados servem de embasamento para a construção de Sistemas Recomendação em cenários socialmente relevantes, onde não seja desejável existir discriminação em relação à atributos protegidos dos usuários, como sexo, cor e idade (e.g. recomendação de vagas de emprego, moradia ou cursos).

O restante do trabalho é organizado da seguinte forma: na Seção 2, revisamos trabalhos relevantes nesta área. Na Seção 3, discutimos o Sistema de Recomendação e o conjunto de dados utilizados. Na Seção 4, definimos métricas de justiça para Sistemas de Recomendação inspirada em problemas de classificação. Na Seção 5, definimos a metodologia para garantir justiça no Sistema de Recomendação a partir da regularização, e definimos os cenários do nosso experimento. Na Seção 6, apresentamos e discutimos os resultados dos experimentos propostos. Finalmente, a Seção 7 apresenta as considerações finais e propostas de trabalhos futuros.

2. Trabalhos Relacionados

É crescente o número de trabalhos relacionados que exploram aspectos de justiça em aprendizado de máquina. Nesta seção, nós apresentamos alguns desses trabalhos com foco em dois grupos principais de tarefas: *(i)* classificação e *(ii)* recomendação.

Classificação. Em [Dwork et al. 2012], justiça em classificação é abordada como um

problema de otimização e são analisados exemplos na qual se alcança uma paridade estatística entre os grupos, mas indivíduos são tratados injustamente no classificador. Este problema também é explorado em [Hardt et al. 2016], e são propostos objetivos na qual se alcança justiça tanto no nível individual quanto em grupo, sem perda de qualidade do classificador. Em [Kamishima et al. 2011], é discutida uma estratégia de regularização para reduzir discriminação, aplicada em um modelo de classificação utilizando Regressão Logística. Em [L. Cardoso et al. 2019], é proposto um *framework* para comparar diferentes modelos de classificação na redução da discriminação utilizando conjuntos de dados artificiais com viés ajustável.

Recomendação. Em [Burke et al. 2018] são apresentas definições de justiça em Sistemas de Recomendação, categorizando o problema em "*C-Fairness*", onde usuários do Sistema de Recomendação são tratados com justiça, e "*P-Fairness*", onde os itens são tratados com justiça. Além disso, é apresentada uma metodologia de justiça para Sistemas de Recomendação baseados em vizinhança, em que as vizinhanças dos usuários são balanceadas para evitar discriminação. Em [Ekstrand and Kluver 2021], a discriminação é explorada em um sistema de recomendação de livros, com viés em relação ao gênero dos autores, e uma abordagem de balanceamento no pré-processamento é analisada. Em [Yao and Huang 2017], são propostas novas métricas de justiça de grupo para diferentes objetivos, e estas métricas são utilizadas na regularização de um sistema de recomendação utilizando Filtragem Colaborativa. Em [Rastegarpanah et al. 2019] é proposto um método baseado em *Adversarial Machine Learning* para reduzir polarização e discriminação em Sistemas de Recomendação, onde um conjunto de dados antídoto é treinado para alcançar objetivos de justiça.

De forma geral, nosso trabalho é complementar aos anteriores, visto que nos inspiramos no arcabouço (ou *framework*) apresentado em [L. Cardoso et al. 2019] para comparar métodos de justiça em classificação e o estendemos para o contexto de recomendação, utilizando as noções de justiça apresentadas em [Hardt et al. 2016] adaptadas para este referido contexto. Para garantir justiça, utilizamos técnicas de regularização inspiradas em [Yao and Huang 2017] e ampliamos nossa análise para os efeitos da regularização nas predições das avaliações de cada grupo. Diferentemente dos trabalhos anteriores, comparamos o impacto da esparsidade do conjunto de dados nos resultados e acrescentamos uma análise da qualidade do *ranking* após a regularização.

3. Metodologia

Nesta seção é apresentada a metodologia proposta para a realização deste trabalho, incluindo um detalhamento do método de recomendação utilizado bem como uma descrição do conjunto de dados explorado.

3.1. Método de Recomendação

Neste trabalho, o método de recomendação utilizado é a *Filtragem Colaborativa* [Schafer et al. 2007]. Para recomendar itens a um usuário, a Filtragem Colaborativa analisa o comportamento de usuários semelhantes em relação aos itens do sistema. Esta relação entre usuários e itens pode ser obtida por *feedback* explícito ou implícito dos usuários. No *feedback* explícito, os usuários têm a oportunidade de avaliar os itens do sistema através de um mecanismo de notas. Pode ser binário (e.g., gostei ou não gostei),

discreto (e.g., notas de 1 a 5) ou até contínuo (e.g., nota entre 1 e 10). Já no *feedback* implícito, a interação do usuário com o item é considerada uma relação (e.g, visitou um produto e ouviu uma música). Portanto, não é possível definir com precisão se o usuário não gostou de um item. Em geral, dados implícitos são mais abundantes em relação aos explícitos, pois não dependem do usuário avaliar (diretamente) os itens do sistema.

Existem diversas técnicas para a Filtragem Colaborativa, incluindo baseadas em vizinhança [Desrosiers and Karypis 2011], decomposição de matrizes [Koren et al. 2009] e redes neurais [Salakhutdinov et al. 2007]. Neste trabalho, exploramos decomposição de matrizes, uma técnica bastante popular, que pode ser definida como se segue. Dada uma matriz $\mathbf{X} \in \mathbb{R}^{m \times n}$ de avaliações de m usuários a n itens, parcialmente observada, o objetivo do algoritmo é encontrar matrizes $\mathbf{U} \in \mathbb{R}^{m \times k}$ e $\mathbf{V} \in \mathbb{R}^{n \times k}$, $k << min(m, n)$, que resolva o seguinte problema de otimização:

$$\arg\min_{\mathbf{U},\mathbf{V}} \{\|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|_F^2 + \lambda(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2)\}, \quad (1)$$

onde $\mathbf{U}$ e $\mathbf{V}$ são os k fatores de usuários e itens, respectivamente, e $\|\cdot\|_F$ representa a norma de Frobenius. O primeiro termo da equação é o erro na estimativa, o qual deve ser computado apenas com relação às entradas conhecidas de $\mathbf{X}$, e o segundo termo corresponde a uma regularização, adicionada com o intuito de evitar sobreajuste (ou *overfitting*). Por fim, $\lambda > 0$ é um hiperparâmetro que controla a importância do termo de regularização no problema otimização. Uma vez que o problema de otimização é resolvido e as matrizes $\mathbf{U}$ e $\mathbf{V}$ são obtidas, as avaliações desconhecidas podem ser estimadas calculando $\hat{\mathbf{X}} = \mathbf{U}\mathbf{V}^T$.

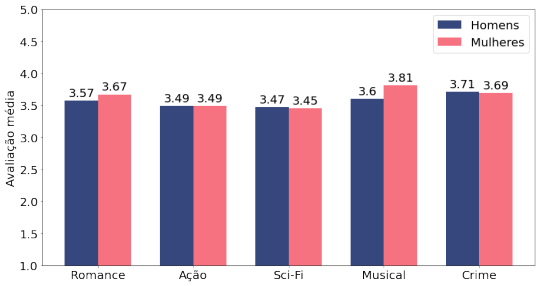
Embora existam técnicas mais sofisticadas de recomendação, tais como as baseadas em redes neurais profundas [Ahmadian et al. 2022]), as estratégias aqui propostas consistem em uma incorporação de novos objetivos da recomendação, que vão além da acurácia, que é o grande foco de trabalhos anteriores, mesmo aqueles que utilizam métodos mais sofisticados. Essa incorporação de objetivos é agnóstica em relação à estratégia de recomendação utilizada e esperam-se resultados similares independentemente dessa estratégia. Assim, o uso de técnicas mais sofisticadas de recomendação será abordado em trabalhos futuros.

3.2. Conjunto de Dados

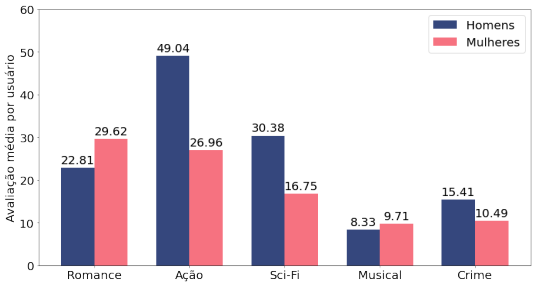
Para os experimentos, foi utilizado o conjunto de dados MovieLens-1M disponibilizado em [Harper and Konstan 2016]. O MovieLens-1M consiste em cerca de 1 milhão de avaliações de 6040 usuários para 3952 filmes, com uma nota entre 1 e 5. Além disso, o MovieLens-1M possui atributos demográficos dos usuários, como gênero e idade, o que permite analisar o comportamento do Sistema de Recomendação para diferentes grupos. Neste trabalho, utilizamos o gênero dos usuários para comparar as recomendações para dois grupos: homens e mulheres.

O conjunto de dados MovieLens-1M é significativamente desbalanceado em relação ao gênero dos usuários, com 71.71% de homens e 28.29% de mulheres. Além disso, 75.36% das avaliações são de homens, enquanto 24.64% são de mulheres. Considerando as categorias seguintes categorias de filme, Ação, Romance, Musical, Ficção Científica e Crime, é possível notar variações nas tendências dos grupos de usuários ao avaliar filmes (Figura 1(a)). Mulheres possuem uma média de avaliação maior que a dos

homens em Romance e Musical. Apesar de possuírem médias iguais para o gênero Ação, homens neste conjunto de dados tendem a avaliar mais filmes do gênero do que mulheres (Figura 1(b)). O mesmo ocorre para filmes de Ficção Científica.



(a) Avaliação média por gênero para cada categoria de filme



(b) Média de avaliações para cada categoria de filme por gênero

Figura 1. Visão geral do conjunto de dados MovieLens-1M.

Além de desbalanceado em relação ao gênero dos usuários, o conjunto de dados é bastante esparso: apenas 4,5% da matriz de avaliações é conhecida. Dessa forma, para analisarmos o impacto da esparsidade dos dados nos resultados, realizamos uma filtragem selecionando apenas filmes avaliados pelo menos 1000 vezes, e somente usuários que avaliaram no mínimo 50 desses filmes. A partir desta filtragem, obtivemos um novo conjunto de dados com 3571 usuários e 618 filmes, onde 23,9% das avaliações são conhecidas. Os dois conjuntos de dados, que no contexto deste trabalho nos referimos como "originais" e "filtrados", são utilizados em nossa análise.

4. Métricas

Para responder as perguntas de pesquisa (P) definidas na Seção 1, precisamos de métricas que sejam úteis e adequadas para avaliação de um Sistema de Recomendação. Assim, para medir a discriminação, utilizamos as noções de justiça apresentadas em [Hardt et al. 2016] para modelos de classificação, adaptando as propostas para modelos de recomendação.

4.1. *Demographic Parity*

Em classificação, *Demographic Parity* verifica se a proporção de usuários em cada grupo recebendo avaliações positivas é igual. Por comparar as predições de um modelo de classificação entre os grupos de usuários, esta restrição pode ser adaptada para o problema de recomendação comparando a média das avaliações previstas para cada grupo pelo sistema de recomendação. Mais formalmente, sejam $\hat{X}_g$ e $\hat{X}_{\neg g}$ os conjuntos das avaliações previstas pelo sistema para usuários que estão em um grupo g e para usuários que não estão em g, respectivamente. Assim, *Demographic Parity* para Sistemas de Recomendação é definida como:

$$R_d(g) = \left| \frac{1}{|\hat{X}_g|} \sum_{\hat{r} \in \hat{X}_g} \hat{r} - \frac{1}{|\hat{X}_{\neg g}|} \sum_{\hat{r} \in \hat{X}_{\neg g}} \hat{r} \right|. \quad (2)$$

Idealmente, a métrica deve apresentar valor zero. Note que por verificar apenas o que o sistema previu, esta métrica não considera a preferência real dos usuários nos grupos. Ela assume que um modelo justo é o modelo em que as predições são iguais para os dois grupos. Tal suposição pode ser prejudicial em alguns cenários, piorando a utilidade do sistema no nível individual [Dwork et al. 2012].

4.2. *Equal Opportunity*

No contexto de classificação, a métrica *Equal Opportunity* compara a proporção de verdadeiros positivos para cada grupo de usuários. Idealmente, a taxa de acertos deveria ser igual em todos os grupos. Note que a métrica, diferentemente de *Demographic Parity*, considera a classe real do conjunto de dados.

Em vez de comparar a proporção das predições do modelo de classificação, em um sistema de recomendação pode-se comparar a diferença dos erros médios do sistema entre os grupos. Assim, a métrica *Equal Opportunity* para Sistemas de Recomendação pode ser definida como:

$$R_e(g) = \left(\frac{1}{|X_g|} \|\mathbf{X}_g - \hat{\mathbf{X}}_g\|_F^2 - \frac{1}{|X_{\neg g}|} \|\mathbf{X}_{\neg g} - \hat{\mathbf{X}}_{\neg g}\|_F^2 \right)^2, \tag{3}$$

onde $\mathbf{X}_g$ e $\mathbf{X}_{\neg g}$ são as matrizes de avaliações reais considerando apenas os usuários no grupo g e fora do grupo g, respectivamente. De forma análoga, as matrizes $\hat{\mathbf{X}}_g$ e $\hat{\mathbf{X}}_{\neg g}$ representam as avaliações fornecidas pelo sistema para os usuários que estão em g e não estão em g, respectivamente. Enfatiza-se que o primeiro termo de $R_e(g)$ deve ser computado apenas sobre as entradas observadas de $\mathbf{X}_g$, e que o segundo termo de $R_e(g)$ deve ser computado apenas sobre as entradas observadas de $\mathbf{X}_{\neg g}$. Assim como a *Demographic Parity*, quanto mais próxima do valor zero, melhor é a *Equal Opportunity*.

Para ilustrar o cálculo das duas métricas de justiça, a Figura 2 mostra um exemplo simples contendo dois grupos de usuários A e B. Cada usuário recebe duas recomendações, para as quais são exibidos os valores estimados de *ratings*, bem como seus valores reais e os erros quadrados correspondentes. Na parte de baixo da figura há, para cada grupo, o valor médio dos *ratings* das recomendações recebidas pelos usuários e o erro quadrado médio (MSE), bem como o cálculo das duas métricas. Nota-se que as recomendações recebidas pelo Grupo A foram, em média, pior avaliadas que as recomendações recebidas pelo grupo B, levando a um valor relativamente alto de *Demographic Parity*. Além disso, o sistema de recomendação errou um pouco mais, em média, nas recomendações recebidas pelo Grupo A, levando a um valor de *Equal Opportunity* também diferente de zero. Logo, para ambas as métricas, o Grupo B foi mais favorecido que o Grupo A.

4.3. Qualidade da Recomendação

Para responder **P2** (Seção 1), precisamos medir a qualidade dos resultados fornecidos por um Sistema de Recomendação. Neste trabalho verificamos o erro de reconstrução e o *ranking* da recomendação, utilizando o RMSE (*Root Mean Squared Error*) e o NDCG (*Normalized Discounted Cumulative Gain*), métricas bastante utilizadas na literatura [Baeza-Yates et al. 1999]. Para obter o erro de reconstrução, o *RMSE* calcula a média do quadrado da diferença entre a avaliação prevista pelo modelo e a avaliação real,

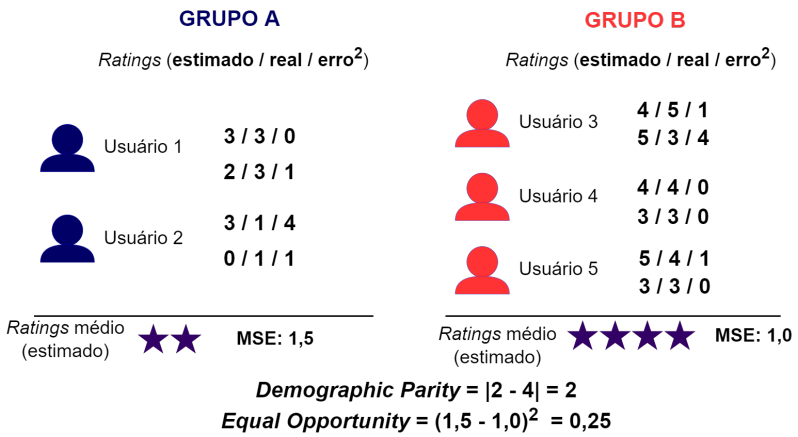


Figura 2. Exemplo de cálculo das métricas de justiça propostas.

e obtém a raiz quadrada dessa média. Com o *NDCG*, medimos a utilidade dos itens previstos para os usuários calculando a relevância dos itens mais recomendados em relação a cada usuário.

5. Configurações Experimentais

A partir de um Sistema de Recomendação, utilizando as métricas apresentadas nas seções anteriores, é possível aplicar técnicas para reduzir a discriminação em Sistemas de Recomendação e comparar os resultados. Uma abordagem comum é através da regularização [Kamishima et al. 2011]. Conforme discutido na Seção 3.1, a Filtragem Colaborativa utilizando decomposição de matrizes minimiza o erro de reconstrução durante o treinamento (Equação 1). Assim, podemos acrescentar a essa função objetivo outro termo de regularização que represente algum objetivo de justiça em Sistemas de Recomendação, de forma que o modelo minimize a discriminação durante o treinamento com o erro de reconstrução. Neste trabalho, utilizamos como termo de regularização as métricas de justiça *Demographic Parity* e *Equal Opportunity*, definidas para o problema de recomendação nas Seções 4.1 e 4.2, respectivamente.

Assim, podemos definir dois novos problemas de otimização, os quais são dados pelas seguintes equações:

$$\arg\min_{\mathbf{U},\mathbf{V}} \{\|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|_F^2 + \lambda(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) + \gamma R_d(g)\} \tag{4}$$

e

$$\arg\min_{\mathbf{U},\mathbf{V}} \{\|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|_F^2 + \lambda(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) + \gamma R_e(g)\}, \tag{5}$$

onde γ é um termo escalar para controlar a regularização e g indica o grupo de usuários de interesse. R_d e R_e foram definidos nas Equações (2) e (3), respectivamente.

Com esses novos problemas, estabelecemos três cenários experimentais (C):

C1: Sem regularização de justiça, utilizando a função objetivo da Equação (1).

C2: Regularização por *Demographic Parity*, como definido na Equação (4).

C3: Regularização por *Equal Opportunity*, como definido na Equação (5).

Para cada cenário mencionado acima, realizamos validação cruzada com 5 partições e, para calcular os gradientes dos objetivos (1), (4) e (5), utilizamos o algoritmo Adam [Kingma and Ba 2014], que combina taxa de aprendizagem adaptativa com momento, mantendo fixos λ=0.002 e γ = 1, e utilizamos k = 100 fatores de usuários e itens. Em cada iteração da validação cruzada, avaliamos o modelo utilizando as métricas de justiça *Demographic Parity* e *Equal Opportunity*, e as métricas de qualidade da recomendação $RMSE$ e $NDCG$, apresentadas na Seção 4. Os resultados apresentados são os valores médios sobre todas as partições.

6. Resultados

Nesta seção, apresentamos os resultados da comparação entre os cenários C1, C2 e C3 definidos na Seção 5. Como discutido na Seção 3.2, utilizamos as duas configurações do conjunto de dados MovieLens 1M (i.e., dados originais e filtrados) em cada cenário C.

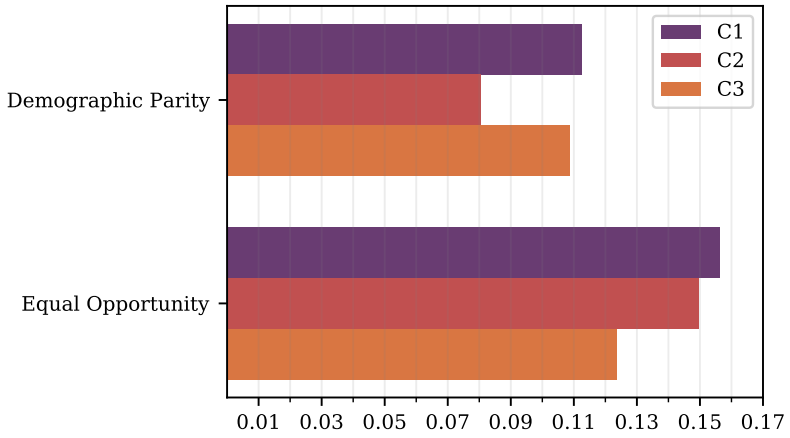
A Figura 3(a) exibe os resultados das métricas de justiça nos 3 cenários (i.e., C1, C2, e C3) considerando o conjunto de dados original (P1). Analisando a métrica *Demographic Parity*, verificamos que o valor melhora no cenário **C2** em comparação ao **C1**, ou seja, ao regularizar o modelo no treinamento de forma que ele iguale a média das avaliações previstas para homens e mulheres, obtivemos o melhor resultado para *Demographic Parity*. Interessantemente, houve também melhoria no cenário **C3**, embora com menor intensidade. Dessa forma, o objetivo de minimização da diferença do erro médio entre os grupos contribuiu indiretamente para melhorar também a diferença entre a média de *ratings* prevista para os dois grupos.

Ainda na Figura 3(a), são apresentados os resultados da métrica *Equal Opportunity* para os 3 cenários analisados. Podemos observar que o melhor resultado obtido é para o cenário **C3**, pois regularizamos utilizando a própria métrica. O cenário **C2** também melhora em relação ao **C1**: ao minimizar a diferença da média prevista para os grupos, percebemos uma redução na diferença do erro médio entre os dois grupos.

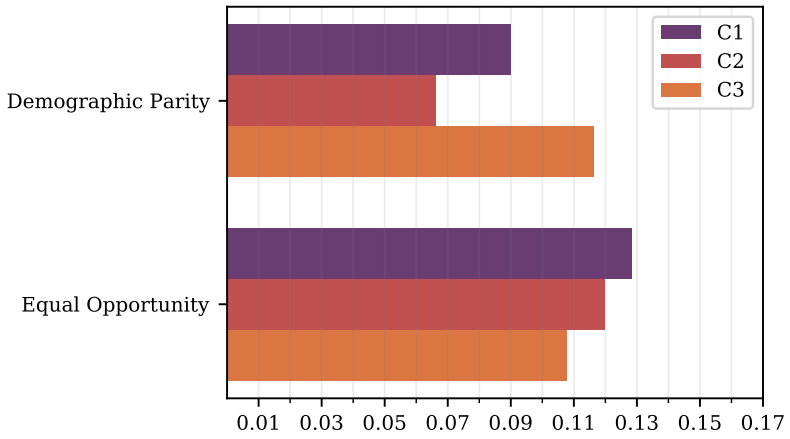
Na Figura 3(b), temos os resultados dos experimentos após a filtragem no conjunto de dados original, ou seja, utilizando um conjunto de dados menos esparso. Assim como nos dados originais, o melhor resultado nas duas métricas ocorre quando regularizamos pela própria métrica. Porém, ao regularizar o modelo de forma a igualar o erro médio entre os dois grupos (*Equal Opportunity*), o *Demographic Parity* piorou em relação ao modelo sem nenhuma regularização (**C1** *vs* **C3**). Especulamos que ao priorizar a preferência dos usuários de forma igual entre os grupos, aumentou a disparidade entre as médias previstas para os dois grupos, devido à não existência de uma paridade entre as avaliações dos dois grupos neste conjunto de dados.

Para melhor visualização deste caso particular, a Figura 4 exibe a diferença da média prevista para cada grupo nos 3 cenários analisados. Note que ao regularizar o modelo pela *Demographic Parity* no cenário **C2**, obtemos médias muito próximas para os dois grupos. Já ao regularizar pela *Equal Opportunity* no cenário **C3**, obtemos a maior diferença nas médias, o que pode justificar a piora ao avaliar a *Demographic Parity*.

Depois, nas Figuras 5 e 6 são apresentados os resultados de qualidade em cada cenário para os dados originais e filtrados, respectivamente (P2). Para o RMSE, valores menores são melhores, pois é uma medida de erro. Já o NDCG retorna resultados entre



(a) Conjunto de dados original



(b) Conjunto de dados filtrado

Figura 3. Resultados dos experimentos nos 3 cenários (C1, C2 e C3) utilizando cada um dos conjuntos de dados (i.e., original e filtrado). Note que o melhor resultado obtido é quando se regulariza pela métrica.

0 e 1 para a qualidade do *ranking*, e valores próximos de 1 são melhores. Pelos resultados, observamos que em ambos conjuntos de dados a regularização produz uma pequena redução na qualidade, tanto em termos de erro quanto de qualidade do *ranking*.

De forma geral, os resultados obtidos evidenciam que a regularização pelos objetivos de justiça apresentados melhoram os resultados das métricas sem impactos significativos na qualidade da recomendação, tanto quando avaliamos o erro de reconstrução quanto na qualidade do *ranking* do modelo, e os melhores resultados de justiça são obtidos ao regularizar pelo próprio objetivo de justiça desejado. Além disso, verificamos que este comportamento se mantém em conjuntos de dados com esparsidades diferentes.

7. Conclusões e Trabalhos Futuros

Neste trabalho, analisamos um Sistema de Recomendação de filmes utilizando Filtragem Colaborativa com decomposição de matrizes em termos de duas métricas de justiça de grupo: *Demographic Parity*, na qual deve existir paridade entre as avaliações previstas para os dois grupos de usuários, e *Equal Opportunity*, na qual o erro das predições deve

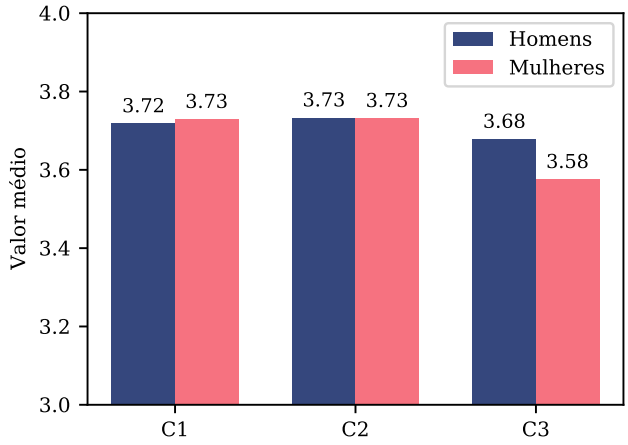


Figura 4. Médias das avaliações previstas para os grupos nos 3 cenários (i.e., C1, C2 e C3) utilizando o conjunto de dados filtrado. Ao regularizar pela *Demographic Parity*, as médias dos dois grupos ficam praticamente idênticas, enquanto ao regularizar pelo *Equal Opportunity*, obtém-se a maior diferença entre as médias.

ser igual entre os dois grupos. Além disso, comparamos os resultados dessas métricas ao utilizar treino com regularização a partir das próprias métricas, de forma que o modelo trate simultaneamente os objetivos justiça e minimização do erro, considerando diferentes cenários.

Nossos resultados demostraram que a regularização melhora os valores das métricas de justiça sem grandes impactos na qualidade final do Sistema de Recomendação, tanto em termos de erro quanto na qualidade do *ranking*, e esses resultados se mantiveram no conjunto de dados com diferentes níveis de esparsidade. Estes resultados podem ser úteis no desenvolvimento de Sistemas de Recomendação em contextos socialmente relevantes, onde justiça de grupo é desejável.

Em trabalhos futuros, pretendemos acrescentar novas coleções de dados em nossa análise, incluindo bases com recomendações implícitas. Desta forma, conseguimos expandir nossa análise para Sistemas de Recomendação com características diferentes. Pretendemos analisar também outras metodologias de justiça além da regularização, como algumas técnicas de pré-processamento e pós-processamento. Outras técnicas de Filtragem Colaborativa também serão exploradas, como as baseadas em vizinhança e técnicas mais recentes baseadas em aprendizado profundo *Deep Learning*.

Agradecimentos. Este trabalho foi parcialmente financiado pelo CNPQ e FAPEMIG.

Referências

Ahmadian, M., Ahmadi, M., and Ahmadian, S. (2022). A reliable deep representation learning to improve trust-aware recommendation systems. *Expert Systems with Applications*, 197:116697.

Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

Brandão, M. A., Moro, M. M., and Almeida, J. M. (2013). Análise de fatores impactantes na recomendação de colaborações acadêmicas utilizando projeto fatorial. In *Simpósio*

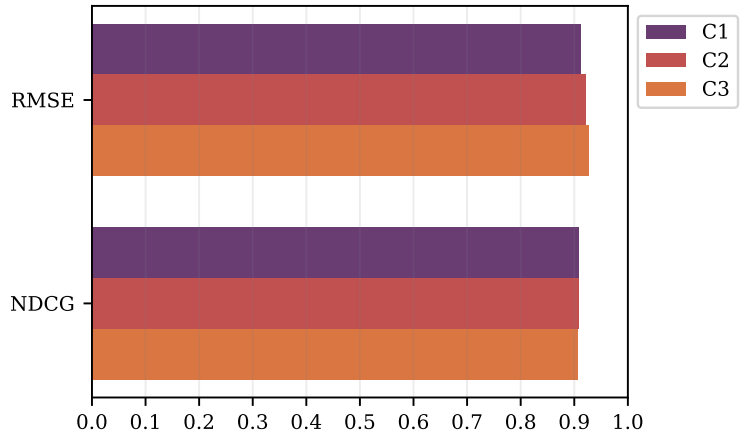


Figura 5. Qualidade da recomendação nos 3 cenários (i.e., C1, C2 e C3) utilizando o conjunto de dados original. O RMSE verifica o erro das previsões: quanto menor, melhor. O NDCG é um valor entre 0 e 1 que representa a qualidade do ranking. Verificamos uma piora desprezível tanto no RMSE quanto no NDCG nos cenários C2 e C3.

Brasileiro de Banco de Dados (SBBD), pages 1–5.

Burke, R., Sonboli, N., and Ordonez-Gauger, A. (2018). Balanced neighborhoods for multi-sided fairness in recommendation. In *Proc. of the Int'l Conference on Fairness, Accountability and Transparency (FAT)*, pages 202–214.

Calders, T. and Verwer, S. (2010). Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292.

Celma, Ò. and Cano, P. (2008). From hits to niches? or how popular artists can bias music recommendation and discovery. In *Proc. of the Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition (@KDD)*, pages 1–8.

Desrosiers, C. and Karypis, G. (2011). A comprehensive survey of neighborhood-based recommendation methods. *Recommender systems handbook*, pages 107–144.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proc. of the Innovations in Theoretical Computer Science Conference*, pages 214–226.

Ekstrand, M. D. and Kluver, D. (2021). Exploring author gender in book rating and recommendation. *User Modeling and User-Adapted Interaction*, pages 1–44.

Ekstrand, M. D., Riedl, J. T., Konstan, J. A., et al. (2011). Collaborative filtering recommender systems. *Foundations and Trends® in Human–Computer Interaction*, 4(2):81–173.

Hajian, S., Bonchi, F., and Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proc. of the ACM Int'l Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 2125–2126.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*.

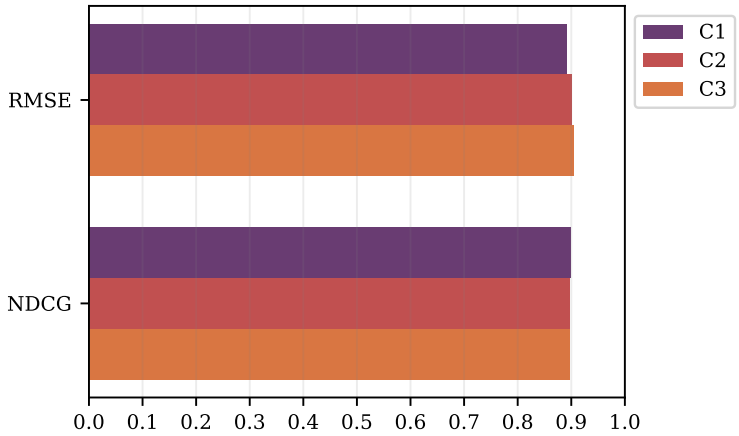


Figura 6. Qualidade da recomendação nos 3 cenários (C1, C2 e C3) utilizando o conjunto de dados filtrado. O RMSE verifica o erro das previsões: quanto menor, melhor. O NDCG é um valor entre 0 e 1 que representa a qualidade do ranking. Verificamos uma pequena piora no RMSE nos cenários C2 e C3. Já o NDCG se mantém nos 3 cenários.

Harper, F. M. and Konstan, J. A. (2016). The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TIIS)*, 5(4):19.

Kamishima, T., Akaho, S., and Sakuma, J. (2011). Fairness-aware learning through regularization approach. In *Proc. of the IEEE International Conference on Data Mining Workshops*, pages 643–650.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ko, H., Lee, S., Park, Y., and Choi, A. (2022). A survey of recommendation systems: Recommendation models, techniques, and application fields. *Electronics*, 11(1):141.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

L. Cardoso, R., Meira Jr., W., Almeida, V., and J. Zaki, M. (2019). A framework for benchmarking discrimination-aware models in machine learning. In *Proc. of the Int'l Conference on AI, Ethics, and Society (AIES)*, page 437–444.

Rastegarpanah, B., Gummadi, K. P., and Crovella, M. (2019). Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. In *Proc.of the ACM Int'l Conference on Web Search and Data Mining (WSDM)*, pages 231–239.

Salakhutdinov, R., Mnih, A., and Hinton, G. (2007). Restricted boltzmann machines for collaborative filtering. In *Proc. of the Int'l Conference on Machine learning*, pages 791–798.

Schafer, J. B., Frankowski, D., Herlocker, J., and Sen, S. (2007). Collaborative filtering recommender systems. In *The adaptive web*, pages 291–324. Springer.

Yao, S. and Huang, B. (2017). Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems*, pages 2921–2930.