Uma Abordagem em Etapa de Processamento para Redução do Viés de Popularidade

Rodrigo Ferrari de Souza
Universidade de São Paulo
São Carlos-SP, Brasil
rodrigofsouza@usp.br

Marcelo Garcia Manzato
Universidade de São Paulo
São Carlos-SP, Brasil
mmanzato@icmc.usp.br

ABSTRACT

Recommendation systems are designed to provide personalized suggestions to each user to enhance user experience and satisfaction across various applications. However, despite their widespread adoption and benefits, such as increased user retention and profits, certain challenges persist, particularly popularity bias, which impacts the quality of recommendations. This bias introduces inconsistencies among user groups, resulting in issues such as lack of calibration, unfairness, and filter bubbles. To address these challenges, several studies have proposed calibration strategies to improve the quality of recommendations and achieve consistency among user groups, focusing on mitigating popularity bias. However, integrating these approaches into a unified model remains a challenge. This study proposes an innovative approach combining popularity-based personalized calibration with the Bayesian Personalized Ranking (BPR) method in the processing step. Our approach aims to provide consistent and fair recommendations while leveraging the efficiency gains of the BPR method. Experimental results on different datasets demonstrate the effectiveness of our modified approach in achieving comparable or superior results to state-of-the-art methods in terms of ranking, popularity, and fairness metrics.

KEYWORDS

Sistemas de Recomendação, Viés de Popularidade, Justiça, Calibração

1 INTRODUÇÃO

Os sistemas de recomendação são projetados para gerar sugestões personalizadas para cada usuário com o objetivo de melhorar sua experiência e satisfação em diversas aplicações. Consequentemente, estes sistemas são cada vez mais predominantes nos diversos contextos atuais, incluindo comércio eletrônico, vídeos, música e muito mais. Como resultado de sua ampla adoção, os sistemas de recomendação tornaram-se um tópico altamente relevante tanto na indústria quanto na academia [10].

Assim, embora estes sistemas produzam benefícios como maior retenção de usuários e lucro por meio da venda de itens recomendados, certos problemas persistem. Notavelmente, a presença de viés de popularidade afeta significativamente a usabilidade e a qualidade das recomendações, levando a possíveis consequências, como falta de calibração e injustiça [5]. Nesse contexto, o viés de popularidade introduz inconsistência e injustiça nas recomendações entre diferentes grupos de indivíduos, fazendo com que aqueles que preferem itens de nicho recebam sugestões de itens populares e vice-versa [1, 10].

Estas questões nas recomendações podem ter ainda mais consequências, tais como bolhas de filtro, que recentemente ganharam atenção considerável em sistemas como as redes sociais. Nestes casos, usuários são expostos exclusivamente a recomendações alinhadas com seus interesses e crenças, sem receber conteúdos com ideias opostas. Consequentemente, nota-se atualmente um aumento considerável na polarização social, levando a tensões significativas em países democráticos [13].

A Figura 1 exemplifica o problema em questão apresentando três possíveis cenários de recomendações geradas para o usuário. Para isso, foi considerada a popularidade dos itens como a característica de preferência para o perfil do usuário, sendo dividida em três tipos: itens populares, itens diversos e itens de nicho.

O cenário (A) mostra uma distribuição onde o tipo de popularidade preferida (itens populares) supera o menos preferido (itens de nicho) nas recomendações. O cenário (B) mostra que uma preferência por itens diversos não está representada nas recomendações. Já o cenário (C) apresenta um conjunto de recomendações calibrado, retornando uma lista coerente e justa com as preferências do usuário. Os principais algoritmos de recomendação criam rankings semelhantes a (A) e (B), enquanto para recomendar itens semelhantes a (C), os algoritmos mais utilizados precisam passar por uma etapa extra de calibração.

A literatura existente indica que vários estudos tentam calibrar sistemas de recomendação para melhorar a qualidade das sugestões e alcançar consistência entre diferentes grupos de usuários [8, 9, 17, 24, 27]. Neste contexto, alguns trabalhos se concentram na implementação de técnicas de calibração para mitigar o viés de popularidade [7, 21, 25, 29, 31]. Conforme descrito em [19], os trabalhos que propõem a calibração de recomendações podem empregar esta estratégia em três momentos distintos:

- **Pré-processamento:** Normalmente, a calibração nesta etapa considera que inconsistências surgem dos dados usados para treinar o modelo de recomendação. Portanto, visa ajustar ou mesmo desconsiderar determinados aspectos dos dados de treinamento. Por outro lado, esses ajustes podem levar a perda de informações importantes dos dados a serem utilizados.
- **Pós-processamento:** A calibração neste último estágio altera a saída gerada pelo algoritmo de recomendação, permitindo que ele resolva os vieses existentes na recomendação produzida pelo modelo. Porém, esta técnica pode comprometer a precisão do sistema, modificando os resultados iniciais.

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2024). Juiz de Fora, Brazil. Porto Alegre: Brazilian Computer Society, 2024.

ISSN 2966-2753

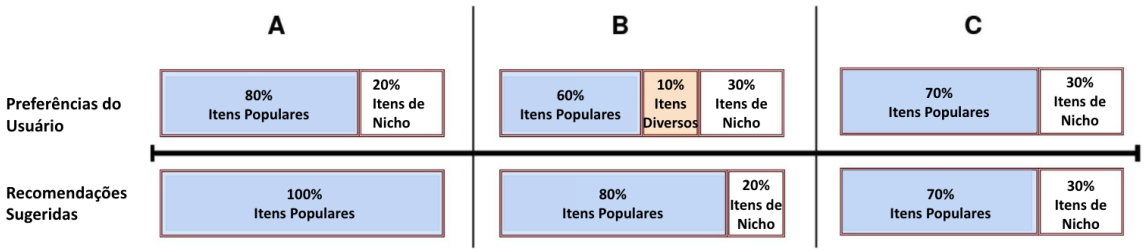

Figura 1: Representação de três possíveis cenários de geração de recomendações.

- **Em processamento:** A calibração nesta etapa envolve a modificação ou introdução de novos algoritmos com o objetivo de reduzir vieses no modelo durante o treinamento.

Dessa forma, a estratégia estudada neste trabalho foi a estratégia de calibração em etapa de processamento. Esse tipo de estratégia visa modificar algoritmos existentes ou introduzir novos algoritmos que resultem em classificações e recomendações justas, por exemplo, removendo preconceitos e discriminação durante o processo de treinamento do modelo. Normalmente, tais métodos visam aprender um modelo sem vieses, ao mesmo tempo que consideram a justiça, incorporando mudanças na função objetivo de um algoritmo por um termo de justiça ou impondo restrições de justiça [19].

A abordagem BPR *(Bayesian Personalized Ranking for Implicit Feedback)* [20] é uma técnica LTR (*Learning to Rank*) do tipo *pairwise* que procura posicionar itens relevantes no topo da lista de recomendação. Para isso, são feitas comparações entre pares de itens – um conhecido e outro desconhecido pelo usuário – de modo a maximizar a diferença de suas respectivas representações. Apesar de não lidar com injustiça e vieses, o BPR possui uma flexibilidade em sua construção que permite utilização com outros modelos de recomendação, e também extensões para que outras condições (como vieses e injustiça) sejam impostas durante seu treinamento [3]. Porém, conforme relatado na seção de trabalhos relacionados, há uma deficiência de trabalhos em etapa de processamento que sejam capazes de lidar com diferentes aspectos de justiça e vieses.

Assim, a proposta deste trabalho é combinar uma forma de calibração personalizada baseada na popularidade dos itens com o método BPR [20] em etapa de processamento. O objetivo dessa combinação é trazer um sistema eficiente que traga recomendações coerentes com as preferências dos usuários, reduza o viés de popularidade e se aproveite do mecanismo de otimização de ranking das recomendações de acordo com a relevância dos itens.

A estrutura deste trabalho é a seguinte: na Seção 2, discutimos trabalhos relacionados e comparamos as abordagens existentes com o nosso trabalho. A Seção 3 descreve a estrutura para nossas proposta de calibração. A Seção 4 detalha a metodologia de avaliação do sistema proposto. A Seção 5 discute os resultados obtidos. Finalmente, na Seção 6, concluímos nosso estudo, apresentando algumas direções futuras para pesquisas.

2 TRABALHOS RELACIONADOS

A literatura recente apresenta diversas propostas destinadas a calibrar recomendações para alinhá-las de forma mais consistente com os perfis dos usuários. Conforme [19], é possível aplicar a calibração em três diferentes etapas, que serão detalhadas a seguir.

2.1 Etapa de Pré-Processamento

Essa etapa consiste em alterar os dados a serem utilizados pelo sistema antes das recomendações serem geradas. Possui vantagens como: ajuda a melhorar a qualidade dos dados removendo as inconsistências antes de serem utilizados e pode melhorar a eficiência do tempo de processamento e adaptar os dados às necessidades do algoritmo. Apesar disso, esses ajustes podem aumentar o tempo de preparação do sistema e levar a perda de informações importantes dos dados a serem utilizados.

O trabalho [14] apresenta uma abordagem de omissão de atributos para tentar remover o viés nos dados. Além disso, o mesmo trabalho mostra uma estratégia de alteração dos rótulos dos itens para remover os vieses. Por outro lado, essas técnicas podem ser insuficientes para garantir justiça nas recomendações, além de existir a possibilidade de tais métodos reduzirem a acurácia do sistema.

Uma outra estratégia é apresentada no trabalho [22], onde é proposto um algoritmo baseado em seleção de amostras para um treinamento justo e robusto. Para tanto, é formulado um problema de otimização combinatória para a seleção imparcial de amostras na presença de problemas nos dados de treinamento. Um dos principais riscos dessa seleção de amostras é a introdução de injustiças, caso o procedimento de amostragem não seja cuidadosamente projetado, já que se a seleção for tendenciosa para determinados grupos de dados, o modelo treinado também poderá apresentar vieses. Um outro problema é o custo computacional dessa seleção de amostras e a complexidade para realizar essa seleção dependendo do contexto dos dados.

2.2 Etapa de Pós-Processamento

A etapa de pós-processamento consiste em calibrar o sistema após o modelo ter gerado as recomendações e tem as seguintes vantagens: é independente do algoritmo de recomendação, pois pode ser feita uma reclassificação da lista gerada por qualquer outro modelo; também é mais eficiente, porque não afeta o tempo de processamento do algoritmo que gera as recomendações. Todavia, essa etapa pode reduzir a precisão do sistema já que altera a lista calibrada gerada pelo modelo de recomendação.

Trabalhos como [24] e [9] fazem ajustes com base nos interesses dos usuários nos gêneros de itens para alcançar um sistema mais consistente. Apesar do foco na consistência, esses trabalhos não abordam a questão do viés de popularidade.

Para abordar o viés de popularidade, alguns trabalhos implementam estratégias de calibração nessa etapa, como o artigo [23] que apresenta uma proposta de calibração baseada em chaveamento. Esta abordagem opta pela calibração com base na popularidade dos itens ou nos gêneros dos itens. Por outro lado, conforme mencionado anteriormente, a precisão do sistema pode ser afetada, já que há uma reordenação na lista de itens sugeridos após a etapa de recomendação.

2.3 Etapa de Processamento

Na etapa de processamento, a calibração ocorre junto com o treinamento e geração das recomendações, e pode levar a um melhor desempenho do algoritmo em termos de precisão. Além disso, também permite aplicar diferentes técnicas de mitigação da injustiça durante o treinamento do algoritmo. Entretanto, esses ajustes podem aumentar a complexidade e o tempo de treinamento do sistema.

As abordagens existentes na literatura normalmente usam aprendizado de máquina para construir modelos de classificação. Em geral, esses modelos categorizam listas não vistas de forma semelhante à classificação dos dados de treinamento. O objetivo geral é ter um modelo que minimize uma função de perda que capture a distância entre o que foi aprendido e a classificação de entrada [19].

O trabalho [30] segue essa ideia adicionando termos de regularização, que expressam medidas de injustiça que o modelo deve minimizar além da minimização da função de perda original. Outra abordagem apresentada é vista em [16], que utiliza a estratégia de rede neural artificial denominada *variational autoencoders* (VAE). Nessa abordagem a filtragem colaborativa é feita juntamente com parâmetros de regularização que melhoram a representação dos dados implementados pelo modelo.

O trabalho [3] propõe uma abordagem para reduzir o viés de popularidade em sistemas de recomendação. O método apresentado conecta as perspectivas do usuário e do item para minimizar a correlação entre a popularidade de um item e sua relevância para um usuário específico. Isso é feito por meio de um modelo probabilístico que aprende os fatores latentes dos usuários e dos itens. O algoritmo atualiza os fatores do usuário com base na diferença entre a avaliação do usuário para dois itens (popular e menos popular) e a probabilidade prevista de o usuário preferir o item menos popular. Devido à similaridade com a proposta deste trabalho, o método foi usado como um dos *baselines* desta pesquisa.

Existem estratégias alternativas para gerar recomendações mais consistentes. A proposta [12] emprega grafos para mitigar injustiças no sistema, considerando o gênero do usuário. O trabalho [6] implementa uma abordagem pareada considerando a justiça entre grupos de itens, ajustando a lista de recomendações durante o treinamento. A abordagem [17] sugere a utilização de grafos e redes neurais para atribuir pesos aos itens recomendados, equilibrando-os de acordo com as preferências do usuário.

Embora essas abordagens produzam resultados interessantes, os estudos não abordam o viés de popularidade de forma a trazer recomendações que atendam os interesses dos usuários por esse aspecto, havendo uma lacuna na área com relação a sistemas que tragam recomendações coerentes e que lidem com o viés de popularidade em etapa de processamento. Além disso, o trabalho [11] destaca a importância dos vieses e como eles afetam os sistemas, o que faz ser necessário selecionar métodos adequados para lidar com o viés presente no sistema.

Dessa forma, a estratégia deste trabalho é combinar uma forma de calibração personalizada baseada na popularidade dos itens com o método BPR [20] em etapa de processamento. O objetivo dessa combinação é trazer um sistema eficiente que traga recomendações coerentes, reduza o viés de popularidade e seja capaz de fornecer recomendações relevantes de acordo com o perfil de cada usuário. As próximas seções descrevem a implementação dessa combinação e os resultados dessa abordagem.

3 ESTRATÉGIA DE CALIBRAÇÃO

Supondo que há um conjunto de itens $I = \{i_1, i_2, ..., i_{|I|}\}$, um conjunto de usuários $U = \{u_1, u_2, ..., u_{|U|}\}$ e um conjunto de itens candidatos para cada usuário $CI_u = \{i_1, i_2, ..., i_N\}$, onde N é o número de itens sugeridos pelo sistema de recomendação. Além disso, existem as informações dos usuários sobre as preferências de popularidade. A tarefa é explorar essas preferências para gerar uma lista de recomendações que aumente a justiça em relação a popularidade dos itens.

Para tanto, propõe-se uma abordagem de calibração em etapa de processamento. Na prática, o método utiliza medidas de divergência na etapa de geração de recomendações para realizar uma calibração de acordo com diferentes níveis de popularidade de interesse do usuário. Como resultado, os usuários recebem uma lista de recomendações próxima ao seu perfil de interesse em termos de popularidade. Essa calibração é incorporada ao BPR. A Figura 2 apresenta a estrutura de calibração de popularidade, cujos detalhes são descritos a seguir.

Figura 2: Estrutura de calibração proposta. A calibração por popularidade é aplicada de forma combinada ao treinamento do BPR, resultando em uma lista calibrada de recomendações de acordo com as preferências do usuário sobre popularidade e gêneros.

3.1 Divisão de Popularidade

A calibração da lista de recomendações com base na popularidade dos itens já consumidos pelo usuário é feita por meio de uma divisão de popularidade para agrupar os itens com base na quantidade

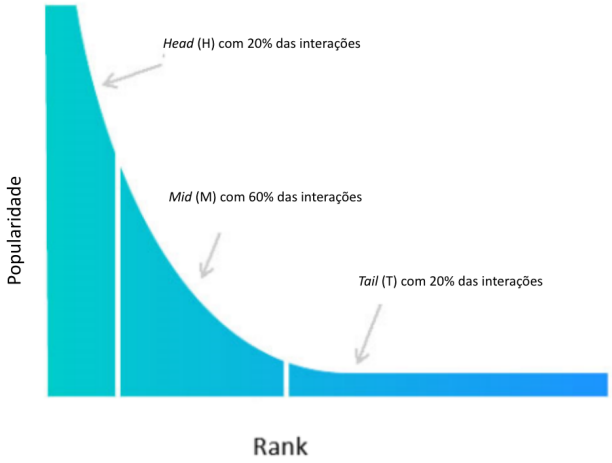

Figura 3: Curva representando a divisão dos itens em grupos de popularidade.

de interações. A divisão de popularidade, introduzida em [2], é baseada no conceito de cauda longa dos sistemas de recomendação, conforme pode ser visualizado na Figura 3. A curva foi dividida em três partes. O **Head (H)**, com itens representando 20% do total de interações. A **Tail (T)** com itens que somam menos de 20% das interações, e o grupo **Mid (M)**, que contém itens que não são nem **Head (H)** nem **Tail (T)**. Vale ressaltar que esta divisão por percentual foi escolhida com base no princípio de Pareto [18].

3.2 Calibração

A calibração por popularidade foi uma adaptação da fórmula proposta por [24]. Seu trabalho pressupõe que os itens podem ter mais de um gênero, o que não é válido no contexto de popularidade, onde um item possui apenas um nível de popularidade. Então, ao invés disso, foram calculadas as somas dos pesos de cada tipo de popularidade sobre a soma de todos os pesos.

Assim, $x(t|u)$ é definido como a distribuição alvo baseada na popularidade dos itens com os quais o usuário interagiu no passado. Na Equação 1 os pesos r_{ui} são definidos como a classificação explícita ou implícita que o usuário u deu ao item i:

$$x(t|u) = \frac{\sum_{i \in I_u} r_{ui} \cdot x(t|i)}{\sum_{i \in I_u} r_{ui}} \tag{1}$$

onde I_u é o conjunto de itens interagidos pelo usuário u, e $x(t|i)$ é definido como 1 se o item i estiver na categoria de popularidade t. Então, para lidar com a distribuição de lista recomendada, a Equação 2 define $y(t|u)$ como:

$$y(t|u) = \frac{\sum_{i \in R_u^*} w_p(u,i) \cdot x(t|i)}{\sum_{i \in R_u^*} w_p(u,i)} \tag{2}$$

Neste caso, usamos os pesos $w_p(u,i)$ como a posição de classificação do item i na lista reordenada recomendada R_u^* para o usuário u.

Várias métricas avaliam a imparcialidade em sistemas de recomendação [26]. Porém, nesse caso, utiliza-se a medida de divergência Kullback-Leibler pelas mesmas razões apontadas por [24] e exploradas por [9]. O Kullback-Leibler quantifica a desigualdade no intervalo $[0, \infty]$, onde 0 significa que ambas as distribuições são quase iguais e valores mais altos indicam injustiça.

Adicionalmente, é adotada a regularização proposta por [24], que definiu $\alpha = 0.01$ como uma variável de regularização para evitar divisão por zero quando $y(t|u)$ vai para zero. Embora existam outras métricas de divergência, como Hellinger e Person Qui-Square, propostas por [4] e exploradas por [9], foi utilizada apenas a Kullback-Leibler devido à sua simplicidade:

$$D_{KL}(x\|y) = \sum_t x(t|u) \cdot log \frac{x(t|u)}{(1-\alpha) \cdot y(t|u) + \alpha \cdot x(t|u)} \tag{3}$$

A divergência de Kullback-Leibler é uma medida que quantifica a diferença entre duas distribuições de probabilidade, neste caso, entre a distribuição observada $x(t|u)$ e a distribuição de referência $y(t|u)$. No contexto da calibração por popularidade, $x(t|u)$ representa a distribuição empírica dos itens observados pelo usuário u, enquanto $y(t|u)$ representa uma distribuição de referência desejada, que é baseada na popularidade dos itens na base de dados.

3.3 O Método BPR

O BPR [20] é uma abordagem eficaz para recomendação de itens em sistemas de recomendação baseados em feedback implícito. Ao modelar as preferências dos usuários por meio de características latentes e otimizar a função de perda, o modelo é capaz de aprender efetivamente as preferências dos usuários e gerar recomendações personalizadas, levando em consideração a ordem de preferência dos itens.

Mantendo a notação utilizada na seção anterior, as letras de indexação especial distinguem usuários e itens: um usuário é indicado como u e um item é referido como i, j; r_{ui} refere-se ao feedback explícito ou implícito de um usuário u para um item i. No primeiro caso, é um número inteiro fornecido pelo usuário indicando o quanto ele gostou do conteúdo; no segundo caso, é apenas um booleano mostrando se o usuário consumiu ou visitou o conteúdo ou não. A predição do sistema sobre a preferência do usuário u para o item i é representada por $\hat{r}_{ui}$, que é um valor de ponto flutuante estimado pelo algoritmo de recomendação. O conjunto de pares (u,i) para os quais r_{ui} é conhecido é representado por $K = \{(u,i)|r_{ui}\}$.

Em um modelo de fatorização tradicional, cada usuário u é associado a um vetor de fatores $p_u \in \mathbb{R}^f$ e cada item i com um vetor de fatores $q_i \in \mathbb{R}^f$. Uma regra de previsão seria:

$$\hat{r}_{ui} = p_u^T q_i \quad (1) \tag{4}$$

Conjuntos adicionais são $N(u)$, que indica o conjunto de itens para os quais o usuário u forneceu um feedback implícito, e $\overline{N}(u)$, que indica o conjunto de itens desconhecidos para o usuário u. Uma característica importante desse tipo de feedback é que apenas as observações positivas são conhecidas; os pares usuário-item não observados são interpretados como feedback negativo.

O trabalho [20] discute um problema que surge quando um modelo de recomendação de itens é treinado apenas com esses dados

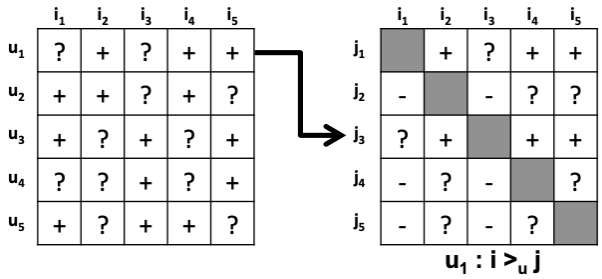

Figura 4: O quadro à esquerda representa os dados observados. A abordagem cria uma relação par de itens específica para o usuário $i \succ_u j$ entre dois itens. No lado direito da tabela, o sinal de mais indica que o usuário u está mais interessado no item i do que no item j; o sinal de menos indica que o usuário prefere o item j ao i; o ponto de interrogação indica que não se pode inferir nenhuma conclusão entre os itens.

positivos/negativos. Como as entradas observadas são positivas e as restantes são negativas, o modelo será ajustado para fornecer apenas pontuações positivas para os itens observados. Os elementos restantes, incluindo aqueles que podem ser de interesse para o usuário, serão classificados pelo modelo como pontuações negativas e a classificação não poderá ser otimizada, pois as previsões estarão em torno de zero.

Os autores propuseram um método genérico para aprender o comportamento do usuário para classificação personalizada [20]. Em vez de treinar o modelo usando apenas os pares usuário-item, eles também consideraram a ordem relativa entre um par de itens, de acordo com as preferências do usuário. Se um item i foi visualizado pelo usuário u e j não ($i \in N(u)$ e $j \in \overline{N}(u)$), então i é preferido a j. A Figura 4 mostra um exemplo do método. Quando i e j são desconhecidos para o usuário, ou equivalentemente, ambos são conhecidos, nenhuma conclusão sobre sua importância relativa para o usuário pode ser inferida.

Para estimar se um usuário prefere um item a outro, [20] propuseram uma análise Bayesiana usando uma função de probabilidade $prob(i \succ_u j|u, \Theta)$ e a probabilidade anterior para o parâmetro do modelo $prob(\Theta)$. O critério final de otimização, BPR-Opt, é definido como:

$$BPR\text{-}Opt := \sum_{(u,i,j) \in S_K} \ln \sigma(\hat{s}_{uij}) - \Lambda_\Theta \|\Theta\|^2$$

onde $\hat{s}_{uij} := \hat{r}_{ui} - \hat{r}_{uj}$ e S_K é o conjunto de triplas (u, i, j) onde i está em $N(u)$ e j não está. O símbolo Θ representa os parâmetros do modelo, Λ_Θ é o conjunto de constantes de regularização, e σ é a função logística definida como $\sigma(x) = \frac{1}{1+e^{-x}}$.

Os autores também propuseram uma variação na técnica de descida de gradiente estocástico, denominada LearnBPR, que amostra aleatoriamente de S_K para ajustar Θ. O Algoritmo 1 mostra uma visão geral do método de aprendizagem, onde α é a taxa de aprendizado.

No presente estudo, definimos a abordagem BPR para considerar a regra de predição $\hat{r}_{ui}$ do modelo de fatorização simples definido na Equação 4. Portanto, aplicar a Equação 4 em $\hat{s}_{uij}$ resulta em $\Theta = \{p_u, q_i, q_j\}$, que devem ser aprendidos. Calculamos as derivadas parciais em relação a $\hat{s}_{uij}$:

Algorithm 1: Aprendizado via LearnBPR.

Input: D_K
Output: Parâmetros ajustados Θ
1 Inicializar Θ com valores aleatórios
2 **for** *cont = 1,...,#Iterações* **do**
3 obtenha (u, i, j) a partir de S_K
4 $\hat{s}_{uij} \leftarrow \hat{r}_{ui} - \hat{r}_{uj}$
5 $\Theta \leftarrow \Theta + \alpha \left(\frac{e^{-\hat{s}_{uij}}}{1+e^{-\hat{s}_{uij}}} \cdot \frac{\partial}{\partial \Theta} \hat{s}_{uij} - \Lambda_\Theta \Theta \right)$
6 **end**

$$\frac{\partial}{\partial \Theta} \hat{s}_{uij} = \begin{cases} q_i - q_j & \text{quando } \Theta = p_u \\ p_u & \text{quando } \Theta = q_i \\ -p_u & \text{quando } \Theta = q_j \\ 0 & \text{caso contrário} \end{cases}$$

Esses gradientes são então usados para atualizar os fatores de usuário e item em direção ao mínimo da função de perda, iterativamente, até que a convergência seja alcançada ou um número fixo de iterações seja concluído. Desse modo, o SGD permite ajustar os fatores de usuário e item de forma a maximizar a diferença entre as pontuações dos itens positivos e negativos, resultando em recomendações mais precisas e personalizadas.

3.4 BPR com Calibração por Popularidade

Com o objetivo de combinar um modelo em etapa de processamento com uma forma de calibração de popularidade para trazer recomendações que reduzam o viés de popularidade no sistema, a estratégia adotada foi alterar o algoritmo de aprendizado *LearnBPR* (Algoritmo 1). Assim, pode-se combinar o BPR com a calibração por popularidade, acrescentando no algoritmo a divergência de Kullback-Leibler implementada na calibração de popularidade.

Essa combinação pode possibilitar uma maior justiça nas recomendações em termos de popularidade, já que esse aspecto seria levado em conta na função de perda do BPR. A alteração é realizada na linha 5 do Algoritmo 1 somente quando $\Theta = p_u$:

$$\begin{aligned} p_u \leftarrow p_u + \alpha \Bigg(& \frac{e^{-\hat{s}_{uij}}}{1 + e^{-\hat{s}_{uij}}} . (q_i - q_j) \\ & + \lambda \left(1 - \frac{D_{KL}(x \| y)}{D_{KL\,void}} \right) - \Lambda_{p_u} p_u \Bigg) \end{aligned} \tag{5}$$

onde λ é utilizado como coeficiente do impacto que a divergência terá no sistema, e $D_{KL\,void}$ é definido como:

$$D_{KL\,void} = \sum_t x(t|u) \cdot log \frac{x(t|u)}{\alpha \cdot x(t|u)} \tag{6}$$

A razão para dividir a divergência $D_{KL}(x, y)$ por $D_{KL\,void}$ é normalizar o valor da divergência, deixando o valor ajustado para uma escala específica. Essa normalização pode ser útil para realizar a calibração por popularidade entre diferentes usuários ou grupos,

independentemente do número total de itens ou da escala de popularidade na base de dados, possibilitando a aplicação em diferentes contextos.

Ao considerar não apenas as preferências individuais dos usuários, mas também a popularidade relativa dos itens, a abordagem modificada pode levar a recomendações mais relevantes e personalizadas. Isso pode resultar em uma melhor experiência do usuário e maior satisfação com o sistema de recomendação.

4 AVALIAÇÃO

A execução do experimento foi realizada três vezes em dois conjuntos de dados do domínio de filmes para garantir a confiabilidade dos resultados. A repetição dos testes ajuda a mitigar o impacto de variações aleatórias, assegurando que a média dos valores obtidos seja representativa do desempenho do modelo. O t-test de Student foi escolhido para análise por ser amplamente utilizado na comparação de médias entre dois grupos, especialmente com amostras pequenas ou moderadas [15], onde as variáveis seguem uma distribuição aproximadamente normal. Como as bases de dados utilizadas possuem dados contínuos e distribuídos de forma aproximada à normal, este teste é adequado para a análise estatística. A Tabela 1 resume as informações dos conjuntos de dados utilizados.

- **Yahoo Movies**[1]: Este conjunto de dados é uma classificação de filmes do usuário, onde o usuário atribui notas de um a cinco aos filmes que assistiu. Na etapa de pré-processamento, foram removidos apenas filmes sem gênero nos metadados. Em vez de binarizar a classificação como feito por [24], foi utilizado o feedback explícito como o peso r_{ui} na Equação 1.
- **MovieLens-20M**[2]: Neste conjunto de dados, semelhante a [24] e em contraste com o conjunto de dados do Yahoo Movies, foi feita a binarização das classificações retendo as interações onde a classificação era superior a 4. Além disso, devido a limitações de hardware, o tamanho do conjunto de dados foi reduzido, removendo filmes com menos de dez interações e usuários com menos de 180 filmes.

Tabela 1: Estatísticas dos conjuntos de dados após realização do pré-processamento.

Conjunto de dados	# Usuários	# Interações	# Itens
Yahoo Movies	7,642	211,231	11,916
MovieLens 20M	12,603	3,984,599	10,417

O experimento foi executado três vezes em cada conjunto de dados para obter a média dos valores gerados pelas métricas e garantir a estabilidade dos resultados. Os conjuntos de dados de teste e treinamento foram escolhidos dividindo aleatoriamente o conjunto de dados em 70/30% de interações, seguindo respectivamente [2, 9]. O desempenho da abordagem foi comparado com os seguintes trabalhos do estado da arte:

(1) **BPR**: Proposta em [20], é um algoritmo de recomendação projetado para lidar com dados de feedback implícito, onde as interações entre usuários e itens são representadas como preferências binárias. Para os três conjuntos de dados, foi aplicado o $batch = 1024$.
(2) ***PairWise***: Proposto por [3], este método atua como uma etapa de processamento para redução do impacto do viés de popularidade. Para o conjunto de dados do Yahoo Movies, foram aplicados $epoch = 100$, $batch = 1024$ e escolhido o melhor α variando no intervalo $[0, 1]$. Para o conjunto de dados MovieLens, foram utilizados $batch = 2048$ e $epoch = 20$. A implementação seguiu aquela feita pelos autores[3].

Métricas. Em nossos experimentos, avaliamos os efeitos da calibração em termos de precisão, justiça e viés de popularidade, conforme detalhado a seguir:

(1) **Precisão e Qualidade**: usamos as métricas *Mean Reciprocal Rank* (MRR) e *Mean Average Precision* (MAP) para medir a qualidade da classificação do item na lista reclassificada. MAP e MRR variam no intervalo $[0, 1]$ onde **valores mais altos são melhores**.
(2) **Justiça**: utilizamos uma métrica proposta por [9], denominada *Mean Rank Miscalibration* (MRMC), que cobre o intervalo $[0, 1]$, onde **valores mais baixos são melhores**. Inicialmente, ela foi usada para calcular a justiça em gêneros na lista de recomendações, mas neste trabalho ela foi adaptada para calcular o erro de calibração de popularidade. Embora nossa proposta visa reduzir a injustiça em termos de popularidade, nós também medimos a justiça de gêneros neste trabalho. Para isso, usamos a média harmônica F1 entre MRMC de gêneros e popularidade, onde **valores mais altos são melhores**:

$$F1 = 2\frac{(1 - MRMC\ Genero) * (1 - MRMC\ Pop)}{(1 - MRMC\ Genero) + (1 - MRMC\ Pop)} \tag{7}$$

(3) **Viés de popularidade**: usamos as métricas de cobertura de cauda longa (LTC) [2] e popularidade média do grupo (ΔGAP) [2] para medir o viés de popularidade. A métrica LTC indica a fração de itens que os usuários recebem nas listas de recomendação e varia no intervalo $[0, 1]$, onde 0 significa que todos os itens recomendados são os mais populares e 1 significa que todos os itens recomendados a um usuário estão nas categorias menos populares. Assim, **quanto mais próximo de 1, mais diversificado será o conteúdo recomendado** [2]. O ΔGAP varia no intervalo $[-1, 1]$, onde valores negativos significam que as recomendações são menos populares do que o esperado segundo as preferências dos usuários, e valores positivos significam que as recomendações são mais populares do que o esperado. Também adotamos três divisões de grupos de usuários, com base em [2] para o ΔGAP: **BlockBuster (BB)** cujo consumo dos usuários é de pelo menos 50% dos itens mais populares, **Nicho (N)** onde o consumo dos usuários é de pelo menos 50% dos itens de menor popularidade e **Diverso (D)** cujas preferências dos usuários divergem dos outros dois grupos. Finalmente, como os valores ótimos de ΔGAP devem ser próximos de zero, propomos neste artigo a utilização do *Root Mean Squared Error* (RMSE) entre os três grupos de usuários, onde **valores mais baixos são melhores**:

[1] https://webscope.sandbox.yahoo.com/
[2] https://grouplens.org/datasets/movielens/20m/
[3] https://github.com/biasinrecsys/wsdm2021

$$RMSE = \frac{\sqrt{\Delta GAP_{BB}^2 + \Delta GAP_{N}^2 + \Delta GAP_{D}^2}}{3} \quad (8)$$

5 RESULTADOS

5.1 Yahoo Movies

A Tabela 2 apresenta os resultados obtidos para o conjunto de dados Yahoo Movies. Analisando apenas a **precisão** dos modelos pela métrica MAP, notamos que a abordagem *PairWise* [3] atingiu o maior valor de MAP. No entanto, esta conquista significa que os itens não são muito diversos entre si, como mostram os seus resultados relativos a LTC, F1 e RMSE.

Em relação à **justiça dos gêneros** através do MRMC de gêneros, a Tabela 2 indica que a proposta de calibração combinada com o BPR produziu o melhor resultado, indicando que foi capaz de fornecer itens mais próximos do perfil em termos de gênero. O mesmo foi verificado em relação à **justiça de popularidade**, com a proposta tendo o melhor resultado do MRMC Pop.

Em termos de **cobertura de cauda longa**, a tabela indica que o modelo mais eficaz para recomendar itens diversos foi o BPR. O *PairWise* com pontuações mais altas no MAP obteve valores mais baixos para o LTC. Em relação à métrica **F1**, é possível observar que a proposta conseguiu alcançar o melhor resultado, indicando que a abordagem de calibração foi capaz de calibrar recomendações de acordo com gêneros e popularidade. Este aspecto é ainda validado ao analisar a métrica **RMSE**, onde a mesma abordagem obteve menor erro com a calibração, indicando que ela aborda os pontos de justiça mencionados e reduz o viés de popularidade do sistema.

Os resultados relatados na Tabela 2 mostram que a abordagem de calibração foi capaz de equilibrar recomendações de acordo com gêneros e popularidade, em oposição aos outros trabalhos, que são mais adequados para um único aspecto, como precisão, gêneros ou popularidade. Além disso, os resultados mostram a importância de adotar métricas além da precisão na análise de algoritmos de recomendação. Reconhece-se a alta precisão do *PairWise*, conforme indicado pela métrica MAP. No entanto, os usuários que preferem itens de nicho, diversos e impopulares são afetados por recomendações injustas e tendenciosas produzidas por essas abordagens.

5.2 MovieLens 20M

A Tabela 2 apresenta os resultados obtidos para o conjunto de dados MovieLens 20M. Analisando a **precisão**, assim como na base de dados anterior, o *PairWise* [3] superou as outras abordagens. No entanto, os resultados também indicam que estas abordagens devolvem recomendações injustas em termos de gênero e popularidade, e carecem de diversidade.

Em relação à **justiça dos gêneros** e à **justiça de popularidade**, a abordagem de calibração proposta obteve os melhores resultados, fato confirmado pela métrica F1. Em relação à **cobertura de cauda longa**, o BPR obteve o melhor resultado entre todas as abordagens. Além disso, o *PairWise* [3] alcançou um valor baixo para essa métrica, apesar de ter uma alta precisão.

Com relação ao F1, pode-se observar que a proposta obteve os melhores valores, destacando seu alto desempenho em termos de justiça nos gêneros e popularidade. Ademais, a proposta também obteve o melhor resultado em **RMSE**, indicando que o sistema reduziu com sucesso o viés de popularidade para diferentes grupos de usuários.

A Tabela 2 reporta resultados semelhantes aos do conjunto de dados Yahoo Movies, indicando que a proposta melhorou a justiça dos gêneros e a popularidade em ambos os conjuntos de dados. Embora a abordagem de calibração proposta não tenha alcançado alta precisão, obteve o menor erro de calibração de gênero e de popularidade, o que significa que o modelo fornece recomendações que respeitam o perfil do usuário tanto no gênero quanto no consumo de popularidade.

6 CONCLUSÃO

O objetivo do BPR é aprender representações latentes para usuários e itens que capturem suas preferências individuais. O procedimento de aprendizado do BPR envolve a otimização de uma função de perda que visa maximizar a ordenação correta dos pares de itens positivos e negativos para cada usuário. Isso é feito por meio de gradiente descendente estocástico, onde os gradientes da função de perda são calculados para atualizar os vetores latentes dos usuários e dos itens.

A modificação proposta, que combina o BPR com a calibração por popularidade, visa melhorar a justiça nas recomendações, considerando não apenas as preferências individuais dos usuários, mas também a popularidade relativa dos itens. Isso é alcançado incorporando a divergência de Kullback-Leibler na função de perda do BPR, levando a recomendações mais relevantes e personalizadas. Os experimentos realizados em dois conjuntos de dados mostram que a abordagem modificada obtém resultados comparáveis ou melhores em relação aos métodos do estado da arte, tanto em métricas de classificação quanto em métricas de popularidade e justiça.

No entanto, é importante ressaltar que a abordagem proposta ainda pode ser aprimorada em vários aspectos. Por exemplo, a escolha dos parâmetros do modelo, como o tamanho do lote e o número de épocas, pode afetar significativamente o desempenho do sistema. Além disso, a implementação de técnicas adicionais de regularização ou otimização pode ajudar a evitar o sobreajuste e melhorar a convergência do modelo. Há também a possibilidade de combinar técnicas de redução do viés de popularidade com os sistemas conversacionais [28]. Futuras pesquisas podem explorar essas direções para desenvolver ainda mais a abordagem proposta e melhorar sua eficácia em uma variedade de cenários de recomendação.

AGRADECIMENTOS

Os autores gostariam de agradecer o apoio financeiro da FAPESP, processo número 2022/07016-9, e CNPq.

REFERÊNCIAS

[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2020. The connection between popularity bias, calibration, and fairness in recommendation. In *Fourteenth ACM conference on recommender systems*. Association for Computing Machinery, New York, NY, USA, 726–731. https://doi.org/10.1145/3383313.3418487

[2] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward C. Malthouse. 2021. User-centered Evaluation of Popularity Bias in Recommender Systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2021, Utrecht, The Netherlands, June, 21-25, 2021*, Judith Masthoff, Eelco Herder, Nava Tintarev, and Marko Tkalcic (Eds.). ACM, 119–129. https://doi.org/10.1145/3450613.3456821

Tabela 2: Comparação da abordagem proposta com os outros trabalhos nos conjuntos de dados Yahoo Movies e MovieLens 20M. O símbolo ▲ significa que a proposta teve um ganho significativo com relação aos outros trabalhos, com um *p-value* < 0.05 usando o *t-test* de Student; o símbolo • significa que não houve um ganho ou perda significativo; e o símbolo ▼ indica que o outro trabalho é estatisticamente melhor que a proposta. Cada par de símbolos se refere ao BPR e ao *PairWise*, respectivamente.

Yahoo Movies										
Algoritmo	**LTC**	**MRMC Gêneros**	**MRMC Pop.**	**F1 Score**	**MRR**	**MAP**	ΔGAP_{BB}	ΔGAP_{N}	ΔGAP_{D}	**RMSE**
BPR	0.409	0.629	0.687	0.340	0.002	0.001	-0.991	-0.881	-0.978	0.549
PairWise	0.140	0.696	0.661	0.321	0.012	0.038	-0.680	3.105	0.043	1.060
BPR Modificado	0.317 ▼ ▲	0.589	0.496	0.444 ▲ ▲	0.012 ▲ •	0.004 ▲ ▼	-0.934	-0.142	-0.835	0.420 ▲ ▲
MovieLens 20M										
Algoritmo	**LTC**	**MRMC Gêneros**	**MRMC Pop.**	**F1 Score**	**MRR**	**MAP**	ΔGAP_{BB}	ΔGAP_{N}	ΔGAP_{D}	**RMSE**
BPR	0.513	0.459	0.409	0.565	0.001	0.001	-0.912	-0.340	-0.790	0.419
PairWise	0.110	0.554	0.501	0.452	0.776	0.583	-0.997	-0.997	-0.996	0.575
BPR Modificado	0.464 ▼ ▲	0.453	0.330	0.596 ▲ ▲	0.002 ▲ ▼	0.001 • ▼	-0.865	-0.060	-0.693	0.370 ▲ ▲

[3] Ludovico Boratto, Gianni Fenu, and Mirko Marras. 2021. Connecting user and item perspectives in popularity debiasing for collaborative recommendation. *Information Processing & Management* 58, 1 (2021), 102387.

[4] Sung-Hyuk Cha. 2007. Comprehensive survey on distance/similarity measures between probability density functions. *City* 1, 2 (2007), 1.

[5] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems* 41, 3 (2023), 1–39.

[6] Xiao Chen, Wenqi Fan, Jingfan Chen, Haochen Liu, Zitao Liu, Zhaoxiang Zhang, and Qing Li. 2023. Fairly adaptive negative sampling for recommendations. In *Proceedings of the ACM Web Conference 2023*. 3723–3733.

[7] Zhihong Chen, Jiawei Wu, Chenliang Li, Jingxu Chen, Rong Xiao, and Binqiang Zhao. 2022. Co-training disentangled domain adaptation network for leveraging popularity bias in recommenders. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 60–69.

[8] Diego Corrêa da Silva and Frederico Araújo Durão. 2023. Introducing a framework and a decision protocol to calibrated recommender systems. *Applied Intelligence* (2023), 1–29.

[9] Diego Corrêa da Silva, Marcelo Garcia Manzato, and Frederico Araújo Durão. 2021. Exploiting personalized calibration and metrics for fairness recommendation. *Expert Systems with Applications* 181 (2021), 115112.

[10] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2023. Fairness in recommender systems: research landscape and future directions. *User Modeling and User-Adapted Interaction* (2023), 1–50.

[11] Michael Färber, Melissa Coutinho, and Shuzhou Yuan. 2023. Biases in scholarly recommender systems: impact, prevalence, and mitigation. *Scientometrics* 128, 5 (2023), 2703–2736.

[12] Alireza Gharahighehi, Celine Vens, and Konstantinos Pliakos. 2021. Fair multi-stakeholder news recommender system with hypergraph ranking. *Information Processing & Management* 58, 5 (2021), 102663.

[13] Ruben Interian, Ruslán G. Marzo, Isela Mendoza, and Celso C Ribeiro. 2023. Network polarization, filter bubbles, and echo chambers: an annotated review of measures and reduction methods. *International Transactions in Operational Research* 30, 6 (2023), 3122–3158.

[14] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems* 33, 1 (2012), 1–33.

[15] Tae Kyun Kim. 2015. T test as a parametric statistic. *Korean journal of anesthesiology* 68, 6 (2015), 540–546.

[16] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*. 689–698.

[17] Haifeng Liu, Nan Zhao, Xiaokun Zhang, Hongfei Lin, Liang Yang, Bo Xu, Yuan Lin, and Wenqi Fan. 2022. Dual constraints and adversarial learning for fair recommenders. *Knowledge-Based Systems* 239 (2022), 108058.

[18] MEJ Newman. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* 46, 5 (sep 2005), 323–351. https://doi.org/10.1080/00107510500052444

[19] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2022. Fairness in rankings and recommendations: an overview. *The VLDB Journal* (2022), 1–28.

[20] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[21] Wondo Rhee, Sung Min Cho, and Bongwon Suh. 2022. Countering Popularity Bias by Regularizing Score Differences. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 145–155.

[22] Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. 2021. Sample selection for fair and robust training. *Advances in Neural Information Processing Systems* 34 (2021), 815–827.

[23] Andre Sacilotti., Rodrigo Souza., and Marcelo G. Manzato. 2023. Counteracting Popularity-Bias and Improving Diversity Through Calibrated Recommendations. In *Proceedings of the 25th International Conference on Enterprise Information Systems - Volume 1: ICEIS*. INSTICC, SciTePress, 709–720. https://doi.org/10.5220/0011846000003467

[24] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM conference on recommender systems*. 154–162.

[25] Amit Sultan, Avi Segal, Guy Shani, and Ya'akov Gal. 2022. Addressing Popularity Bias in Citizen Science. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*. 17–23.

[26] Sahil Verma, Ruoyuan Gao, and Chirag Shah. 2020. Facets of fairness in search and recommendation. In *Bias and Social Aspects in Search and Recommendation: First International Workshop, BIAS 2020, Lisbon, Portugal, April 14, Proceedings 1*. Springer, 1–11.

[27] Lili Wang, Sunit Mistry, Abdulkadir Abdulahi Hasan, Abdiaziz Omar Hassan, Yousuf Islam, and Frimpong Atta Junior Osei. 2023. Implementation of a Collaborative Recommendation System Based on Multi-Clustering. *Mathematics* 11, 6 (2023), 1346.

[28] Xi Wang, Hossein A Rahmani, Jiqun Liu, and Emine Yilmaz. 2023. Improving Conversational Recommendation Systems via Bias Analysis and Language-Model-Enhanced Data Augmentation. *arXiv preprint arXiv:2310.16738* (2023).

[29] Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. 2021. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1791–1800.

[30] Meike Zehlike and Carlos Castillo. 2020. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of the web conference 2020*. 2849–2855.

[31] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 11–20.