Explorando Justiça em Sistemas de Recomendação

Calibragem Ponderada, Balanceamentos e Métricas

Diego Corrêa da Silva
diego.correa@ufba.br
Universidade Federal da Bahia
Salvador, Bahia, Brasil

Frederico Araújo Durão
fdurao@ufba.br
Universidade Federal da Bahia
Salvador, Bahia, Brasil

Abstract

Sistemas de Recomendação são ferramentas utilizadas para sugerir itens, que sejam de interesse dos usuários. Estes sistemas baseiam-se no histórico de preferências do usuário para gerar uma lista com recomendações, que possuam maior similaridade ou relevância com o perfil do usuário. A recomendação baseada em similaridade/relevância pode causar efeitos colaterais na lista como: superespecialização das recomendações em um determinado núcleo de itens, pouca diversidade de categorias e desbalanceamento de gêneros. Assim, esta dissertação tem como objetivo explorar a calibragem, que é um meio para produzir recomendações que sejam relevantes aos usuários e ao mesmo tempo considerar todas as áreas de suas preferências. Para isto, foram propostas formas de realizar o balanceamento entre a relevância e a calibragem, assim como um modelo conceitual de sistema calibrado e um protocolo de decisão. Os resultados indicam que a calibragem produz efeitos positivos tanto para a precisão quanto para a justiça.

Keywords: Calibragem, Justiça, Métricas, Personalização, Recomendação

1 Introdução

Recentes estudos da literatura de Sistemas de Recomendação abordam o tema de justiça em recomendações, focando em um aspecto em particular chamado de Calibragem [2, 5, 8, 10], que busca gerar recomendações concisas com os gêneros/categorias que compõem as preferências do usuário. A calibragem utiliza a distribuição dos gêneros nos itens das preferências do usuário como alvo para criar uma lista de recomendação que respeite a distribuição dos gêneros. Quando um sistema de recomendação não segue a proporção das preferências do usuário diz-se que ele está descalibrado. Descalibragem, por si só, não quer dizer que o sistema é injusto, mas sim que não está devidamente personalizado, ou seja, não está calibrado. Entretanto, se usuários ou grupos de usuários experimentam diferentes níveis de descalibragem, isso pode indicar um tratamento injusto dos usuários.

In: IV Concurso de Teses e Dissertações (CTD 2022), Curitiba, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2022.

ISSN 2596-1683

Um exemplo motivacional é, supondo que o perfil de um usuário seja composto por músicas 60% de Rock, 20% de Samba e 20% de Mangue-Beat, assim temos que o gênero de maior preferência é o Rock. Ademais, supomos que a aplicação possua em sua base de dados músicas dos mais diferentes gêneros, mas que em sua maioria sejam pertencentes ao Rock. Assim, temos uma tendência das músicas mais populares serem do Rock e em acréscimo temos que a própria preferência do usuário é composta por maioria de Rock. Isto causa um efeito bolha [7], em que apenas um gênero é popular e constantemente recomendado.

O problema principal que esta pesquisa investigou foi a utilização da calibragem para promover justiça na lista de recomendação, tomando como base a ausência de mecanismos que personalizem e avaliem o grau de relevância e calibragem que a lista possui com as preferências do usuário. O objetivo do trabalho foi desenvolver um modelo de sistema e um protocolo de decisão para recomendação calibradas. Protocolo este que visa indicar entre um conjunto de sistemas calibrados propostos qual é o recomendado a ser implementado. Os 702 sistemas calibrados testados foram baseados em filtragem colaborativa. Cada sistema buscou produzir algum grau de justiça à distribuição dos gêneros que compõem as preferências de cada usuário, recomendando-os com listas de itens calibradas a estes gêneros. Diante do problema e dos objetivos apresentados, sete **Q**uestões de **P**esquisa (**QP**) guiaram a condução deste trabalho e são apresentadas e respondidas na Seção 6. As contribuições deste estudo são:

- Comparação entre nove algoritmos recomendadores;
- Duas formulações para encontrar o peso personalizado do balanceamento: Contagem de gêneros e Variância;
- Uso de duas novas medidas de divergência: Hellinger e Pearson Chi Square;
- Uma nova formulação de balanceamento (Logarítmico);
- Duas novas métricas calibragem;
- Dois coeficientes decisórios;
- Um modelo de sistema de recomendação calibrado;
- Um protocolo de decisão para escolha do melhor sistema calibrado para o domínio.

Os resultados demonstram que a pesquisa avançou com melhorias sobre o tema e contribuiu com a comunidade científica de Sistemas de Recomendação e Calibragem ao explorar um tema recente e pouco desenvolvido.

2 Modelo de Sistema Calibrado

Esta pesquisa visa produzir um modelo de sistema de recomendação calibrado e a partir deste modelo gerar diversos sistemas de recomendação. O modelo proposto é dividido em três etapas e 12 componentes [4]. A primeira etapa é a de **pré-processamento** que é composta por três componentes: limpeza, filtragem e modelagem. Focada em preparar os dados para o sistema calibrado. A etapa de **processamento** é composta pelo componente do algoritmo de recomendação. A fim de predizer o valor da similaridade/relevância do item com o usuário. A etapa de **pós-processamento** é o foco principal dos sistemas calibrados sendo composto por oito componentes apresentados nas seções seguir.

2.1 Componente de Distribuição dos Gêneros

Para encontrar a **distribuição alvo**, que é baseada no modelo do usuário, e a **distribuição realizada**, que é baseada na lista de recomendação, são utilizadas as funções de extração da distribuição [1, 2, 5, 8–10].

- **p(g|i):** função para identificar o valor da importância do gênero g no item i, que é dada pela probabilidade de escolher um gênero em um item ($\frac{1}{|genresIn(i)|}$).
- **p(g|u):** função que extrai a **distribuição alvo** do perfil do usuário.
- **q(g|u):** função que extrai a **distribuição realizada** a partir da lista de recomendação.

2.2 Componente de Medida de Justiça

A medida de divergência calcula o quanto distante a distribuição realizada esta da distribuição alvo. Nesta pesquisa foram implementadas três medidas de divergência como medida de justiça.

- **Kullback-Leibler:** A medida é proposta e implementada pelo estado-da-arte [1, 2, 5, 8, 10].
- **Hellinger:** Como proposta desta pesquisa, a medida de divergência Hellinger foi implementada [5].
- **Pearson Chi-Square:** Como proposta desta pesquisa, a Pearson Chi-Square (χ^2) [3] foi implementada.

2.3 Componente de Medida de Relevância

Os algoritmos recomendadores normalmente predizem um valor de relevância/similaridade do item com o usuário. Os itens das listas de recomendação possuem este valor de relevância. Assim, um somatório é aplicado nestes valores, retornando o valor único da relevância desta lista. Esta implementação também foi investigada pelo estado-da-arte [5, 8, 10].

2.4 Componente do Peso do Balanceamento

As formulações do balanceamento apresentadas pelo estado-da-arte [5, 8, 10] utilizam-se de um peso constante λ para balancear a relevância do ranque com as medidas de divergência. Esta dissertação propõe como contribuição duas formulações para encontrar o valor de λ de forma personalizada (Variância e Contagem de Gêneros). Nenhum dos trabalhos do estado-da-arte explora formulações para o peso do balanceamento λ de forma personalizada.

- **Variância** - *VAR***:** Como proposta desta pesquisa, a variância encontra a dispersão dos gêneros no perfil do usuário.
- **Contagem de Gêneros** - *CGR***:** Nesta proposta, conta-se quantos gêneros o usuário possui e divide-se pelo número total de gêneros no sistema.

2.5 Componente de Balanceamento

Neste trabalho são utilizadas duas formulações de balanceamento para calibragem. A abordagem *Linear* proposta por Steck [10] e a *Logarítmica* proposta deste estudo.

- **Linear** - *LIN***:** Steck [10] apresenta um balanceamento linear. A função é dividida em duas partes: a primeira é a multiplicação do $1 - \lambda_u$ pelo valor da relevância do ranque, a segunda é a multiplicação do λ_u pelo valor da divergência. Assim, quanto maior a divergência, maior é a subtração no valor do ranque.
- **Logarítmica** - *LOG***:** Uma forma de considerar os gêneros, os itens e o usuário no balanceamento é adicionar o uso do viés. Assim, como contribuição, esta dissertação propõe um novo balanceamento. Este é dado de forma logarítmica para suavizar a curva, além de contar com a adição de um viés do usuário. A inserção do viés permite ao pós-processamento em calibragem adaptar a lista de recomendação de acordo com os itens inseridos na lista.

2.6 Componente do Algoritmo de Seleção

Inspirado no estado-da-arte [5, 8, 10], o Surrogate Submodular foi implementado como algoritmo de seleção de itens.

3 Protocolo de Decisão

A implementação de todas as propostas geram diversos sistemas de recomendação calibrados. Assim, um protocolo de decisão foi desenhado. Para atingir a decisão foram propostas: duas métricas, dois coeficientes e a decisão.

3.1 Métricas

O estado-da-arte propõe a métrica *Miscalibration* [1, 2, 8, 10] como uma forma de avaliar o grau de descalibragem na lista de recomendação. Entretanto, esta métrica não considera a ordem do ranque. Assim, nossa pesquisa propõe duas novas métricas de avaliação para o contexto da calibragem.

- **MACE:** A Mean Average Absolute Error (MACE) considera a diferença absoluta da calibragem de cada posição da lista de recomendação com o perfil do usuário. Apresentando um valor absoluto resumindo toda posição do ranque de todos os usuários.

- **MRMC:** A Mean Rank MisCalibration (MRMC) considera a descalibragem entre as distribuições, observando cada posição da lista de recomendação. O cálculo é realizado utilizando a medida de divergência implementada pelo sistema.

3.2 Coeficientes

Os coeficientes são propostas desta pesquisa e utilizam as duas métricas de calibragem e o Mean Average Precision (MAP).

- **CCE:** O Coefficient of Calibration Error (CCE) é o valor encontrado pelo MACE divido pelo valor do MAP.
- **CMC:** O Coefficient of MisCalibration (CMC) é o valor encontrado pelo MRMC divido pelo valor do MAP.

3.3 Decisão

A decisão do melhor sistema calibrado a partir de um conjunto de sistemas, é dado pela soma do CCE mais o CMC. O sistema que obter o menor valor é considerado o melhor para ser implementado no domínio.

4 Configuração do Experimento

Similar a estudos do estado-da-arte [8, 10], duas bases de dados públicas foram utilizadas a Movielens 20M e One Million Songs (OMS). Neste experimento, de cada modelo do usuário foram retirados aleatoriamente 70% dos dados para treinamento e 30% para teste.

Para obter uma maior compreensão da questão de justiça, foram utilizados nove algoritmos recomendadores popularmente conhecidos [6]: 1) o K Nearest Neighbors (KNN) baseado no usuário (User-KNN), 2) o Item-KNN que é baseado no item, 3) Slope One, 4) Non-negative Matrix Factorization (NMF), 5) Singular Value Decomposition (SVD), 6) Singular Value Decomposition Plus Plus (SVD++), 7) Co-Clustering, 8) Melhor Nota e 9) Popularidade. Os hiper-parâmetros dos algoritmos foram otimizados através do método *Grid Search*.

Foram obtidas a partir das propostas $2 \cdot 9 \cdot 13 \cdot 3 \cdot 1 \cdot 2 \cdot 1 = 1404$ combinações de resultados: 2 bases de dados, 9 algoritmos recomendadores, 13 pesos do balanceamento, 3 medidas de justiça, 1 distribuição, 2 balanceamentos e 1 algoritmo de seleção. Cada sistema calibrado foi executado 3 vezes para cada base de dados, particionando aleatoriamente os dados em 70% para treinamento e 30% para teste. Uma média das execuções representa o resultado final. As listas de recomendação foram geradas e avaliadas com os tamanhos de $[1; 10] \in N$.

Nesta pesquisa foram aplicadas quatro métricas, dois coeficientes e um protocolo de decisão, que são: MAP, Mean Reciprocal Rank (MRR), MACE, MRMC, CCE e CMC com o protocolo de decisão.

5 Discussão dos resultados

Para o Movielens, a métrica MAP indica que a formulação do balanceamento logarítmico, proposto nesta dissertação, propicia um incremento nos desempenhos dos recomendadores SVD++, SVD e Popularidade. É possível afirmar também que a proposta de utilizar a medida de divergência χ^2 produz efeitos positivos nos desempenhos dos recomendadores. As propostas do peso do balanceamento produzem melhorias no desempenho ou, quando não produzem, reproduzem o desempenho e comportamento dos pesos atribuídos pelo especialista do sistema. Para o OMS, a métrica MAP indica que ser popular é mais influente do que ser ouvido muitas vezes individualmente pelos usuários. Similar ao *Movielens*, os pesos personalizados do balanceamento, proposta desta pesquisa, produzem melhorias no recomendador ou reproduzem o comportamento e desempenho dos valores constantes, assim como o uso da medida χ^2 também produz ou reproduz efeitos positivos similares às outras medidas usadas pelo estado-da-arte. Para ambas as bases, o pós-processamento produziu efeitos positivos no desempenho, quando avaliado pela métrica MAP.

Os resultados obtidos a partir da avaliação com o MRR seguem o mesmo comportamento do MAP para ambas as bases de dados e para todos os algoritmos. A partir das observações sobre as duas métricas de ranqueamento pode-se concluir que, ao criar a lista de recomendação, o pós-processamento traz o primeiro item mais relevante para mais perto do topo da lista, influenciando positivamente, assim, todo o processo de preenchimento das próximas posições.

Os resultados obtidos analisados a partir da métrica de avaliação proposta nesta dissertação, chamada de MACE, indica que os resultados de menor erro absoluto na calibragem dependem da base de dados e sua composição. Para o Movielens os melhores resultados são obtidos pela formulação do balanceamento logarítmico. Para o OMS o balanceamento linear obtém os melhores desempenhos. Os comportamentos observados em outras métricas de avaliação são capturados pela MACE também, como é possível verificar que o peso do balanceamento totalmente focado em justiça causa um efeito imediato na base do OMS. Esse comportamento pode ser atribuído à constituição das músicas, que possuem no mínimo um e no máximo dois gêneros, sendo que os filmes do Movielens possuem de um até cinco gêneros.

O MRMC indica a possibilidade de reduzir a descalibragem em qualquer algoritmo recomendador, demonstrando que a utilização do pós-processamento produz um efeito positivo na lista. É possível observar que a medida de divergência usada é diretamente influente no resultado obtido. A medida χ^2, proposta desta dissertação, obtém os melhores resultados.

A partir das análises obtidas com os cruzamentos das métricas e a decisão do protocolo é possível afirmar que, dependendo da base de dados utilizada, a melhor combinação de sistema muda. Para a base de dados do Movielens

é recomendável utilizar a combinação de sistema: SVD++, a formulação logarítmica do balanceamento, a medida de divergência Pearson Chi Square e o peso personalizado *VAR*. Para a base de dados do OMS é recomendável utilizar a combinação de sistema com: Item-KNN, a formulação linear do balanceamento, a medida de divergência χ^2 e o peso personalizado *CGR* ou *VAR*.

6 Respostas para as questões de pesquisa

As sete **Q**uestões de **P**esquisa são:

QP1: *Como encontrar o melhor sistema de recomendação calibrado?* Esta dissertação demonstrou que é possível encontrar o melhor sistema calibrado dentre um conjunto de possibilidades, utilizando o protocolo de decisão.

QP2: *A base de dados utilizadas influencia no comportamento ou desempenho do sistema calibrado?* Para cada base de dados uma combinação de sistema calibrado diferente obteve o melhor desempenho. Para o Movielens foi o SVD++, *LOG* e χ^2 (12,14). Para o OMS foi o Item-KNN, *LIN* e χ^2 (91,75).

QP3: *É possível padronizar os sistemas calibrados de modo que auxilie o desenvolvimento?* Pode-se afirmar que o modelo de sistema calibrado proposto é uma padronização, i. e., um quadro pré-moldado que pode ser alterado gerando diferentes combinações de sistema.

QP4: *Quais são os efeitos que a calibragem produz nas listas de recomendação de cada algoritmo recomendador?* Os resultados indicam que a depender do recomendador utilizado o efeito produzido pela calibragem na lista de recomendação muda. Alguns algoritmos obtiveram um acréscimo no MAP e outros, decréscimo, assim como para as outras métricas. Com isso, é possível afirmar que cada recomendador produzirá diferentes resultados e efeitos em conjunto com a calibragem.

QP5: *Utilizar pesos personalizados do balanceamento obtém melhorias ou mantém desempenho quando comparados com pesos constantes?* É possível observar que o desempenho da combinação de sistema é influenciado pelo peso utilizado. Entretanto, se a formulação da combinação de sistema tende a reduzir a precisão, os pesos utilizados apenas afetarão o resultado em maior ou menor escala, mas manterão o comportamento da combinação.

QP6: *A medida de divergência utilizada influência nas listas de recomendação?* Os resultados indicam que a medida influencia na lista de recomendação, gerando diferentes tipos de lista. Estas que por sua vez, trazem maior ou menor precisão, assim como produzem acréscimo ou decréscimo no desempenho da calibragem. Entretanto, para ambas as bases de dados a medida de divergência com o melhor desempenho foi o Pearson Chi Square, obtendo todos os nove melhores desempenhos na Movielens e 7 dos melhores na OMS.

QP7: *Ao considerar o viés do usuário na formulação do balanceamento é possível obter melhoria no desempenho?* Os resultados demonstraram que ao considerar o viés do usuário é possível obter melhora no desempenho, tanto da precisão como da calibragem. A depender da base de dados utilizada o viés do usuário pode ser positivo.

7 Conclusão

Este trabalho teve como objetivo desenvolver um modelo sistema de recomendação baseado em filtragem colaborativa que busca ser justo com as preferências dos usuários. A partir do modelo, foram implementados 1.404 sistemas diferentes, que foram avaliados por quatro métricas, dois coeficientes e um protocolo de decisão. As propostas desta pesquisa, apresentaram um incremento no desempenho dos sistemas. As métricas propostas demonstram a sua efetividade em capturar os comportamentos do sistema e metrificar os resultados.

Acknowledgments

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - fomentou esta pesquisa sob o Código de Financiamento 88887.502736/2020-00. O Laboratório Nacional de Computação Científica (LNCC/MCTI, Brazil), nos concedeu acesso aos recursos do computador de alto desempenho chamado SDumont.

References

[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The Impact of Popularity Bias on Fairness and Calibration in Recommendation. *ArXiv* (10 2019).

[2] 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*. 726–731.

[3] Sung-Hyuk Cha. 2007. Comprehensive Survey on Distance/Similarity Measures Between Probability Density Functions. *International Journal of Mathematical Models and Methods in Applied Sciences* 1, 4 (04 2007), 300–307.

[4] Diego Corrêa da Silva and Frederico Araújo Durão. 2022. Introducing a Framework and a Decision Protocol to Calibrate Recommender Systems.

[5] Diego Corrêa da Silva, Marcelo Garcia Manzato, and Frederico Araújo Durão. 2021. Exploiting Personalized Calibration and Metrics for Fairness Recommendation. *Expert Systems with Applications* (2021), 115112.

[6] Nicolas Hug. 2017. Surprise, a Python library for recommender systems. (2017).

[7] Toshihiro Kamishima, Shotaro Akaho, and Hideki Asoh. 2012. Enhancement of the neutrality in recommendation. In *Proc. of the 2nd Workshop on Human Decision Making in Recommender Systems*. 8–14.

[8] Mesut Kaya and Derek Bridge. 2019. A Comparison of Calibrated and Intent-Aware Recommendations. 151–159.

[9] Kun Lin, Nasim Sonboli, Bamshad Mobasher, and Robin Burke. 2020. Calibration in Collaborative Filtering Recommender Systems: A User-Centered Analysis. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*. 197–206.

[10] Harald Steck. 2018. Calibrated Recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 154–162.