Cross-Dataset Propensity Estimation for Debiasing Recommender Systems

Anonymous Author(s)
Affiliation
Address
email

Abstract

Datasets for training recommender systems are often subject to severe distribution shift induced by users' selection biases. In this paper, we study the impact of selection bias on datasets with different quantization. We then leverage two differently quantized datasets from different source distributions to improve propensity estimation for the inverse probability scoring method from causal inference. Empirically, our approach gains significant performance improvement over single-dataset methods and alternative ways of combining two datasets.

1 Introduction

Selection bias is one of the most prevalent sources of biases for recommender systems [3]. Selection bias happens when there is a pattern in the users' ratings that is unique to the training set. For example, in a recommender system for movies, users might mainly rate movies that are recommended to them, which is a small section of movies already tailored to the users' tastes [9]. However, the environment the recommender system is deployed on contains all candidates movies, which is not biased toward personal tastes. This discrepancy produces a misalignment between training and deployed settings, which is known as a distribution shift. Tackling the selection bias in a recommender dataset has been a constant challenge in designing recommender algorithms [11].

When recommender systems are deployed in real-world platforms, clients are usually able to collect preference-associated data from different source distributions. These data, or feedbacks, are either implicit or explicit [1], highly-quantized (e.g. binary) or less quantized (e.g. one to five stars.) However, state-of-the-art recommender systems tend to only train on one dataset, which is often less quantized. We hypothesize that combing more than one datasets with different degrees of bias in the training procedure would make the model more robust against selection bias.

In this paper, we attempt to take advantage of datasets from differently biased sources with different quantization. More specifically, we propose a way to feed both a highly quantized and a less quantized dataset to a gradient-based recommender algorithm, given the assumption that the highly quantized dataset contains less bias. Practically, this assumption stands true under many settings. Being less biased toward personal taste has been an important reason that implicit dataset, which is often more quantized, is adopted for modeling [2] [5].

To ensure both datasets do not significantly lose their values in the presence of selection bias, we first examine the susceptibility to selection bias of differently quantized datasets from a source distribution. We design experiments under a simulated environment and shows that susceptibility to selection bias is not correlated with the way a dataset is quantized.

Then, since a less-quantized dataset inherently contains more information than a more-quantized dataset [14] and thus is more suitable for training the model, we decide to use it as the training

Submitted to Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

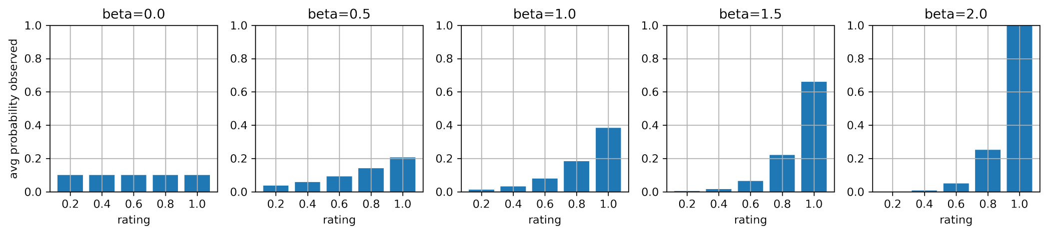

Figure 1: Visualizing the effect of controlling bias

data of matrix factorization and use the more quantized dataset for propensity scoring and deriving the inverse-probability-scoring (IPS) estimator [13] [4], a causal inference approach applicable to matrix completion-based recommender algorithms. In this way, our cross-dataset learning framework empowers existing recommender algorithms by using previously neglected, highly quantized dataset. We carried out experiment and found our method outperforming baselines by a significant margin.

2 Related Work

Prior works on overcoming selection bias-induced distribution shift via a propensity-based approach begins with the seminal paper [11], which introduces the IPS method into recommender systems. Follow-up works aim at providing a learning-based or behavioral model of user feedbacks for propensity estimation [6] [15], which remains the central concern of this approach. While many works rely on a propensity matrix that is assumed to be known, e.g. [10], our work provides an idea of using a more quantized, implicit dataset for accurate propensity estimation.

The IPS method in our context is, essentially, a method of weighting training examples to correct the bias in the training data. Equivalent approaches such as importance weighting are widely used for domain adaptation in fields other than recommender systems [12] [16]. Discussions on the IPS-based domain adaptation for countering selection bias, which is most relevant to recommender systems, remain limited.

3 Susceptibility to Selection Bias

We first examine the susceptibility to selection bias of differently quantized data by manually introducing biased distributions of various degrees to the differently quantized training sets. It is crucial for us that the highly quantized dataset does not exhibit particular weakness when facing selection bias so that they can be properly adopted for propensity estimation.

3.1 Simulated Environment for Controlling Bias

Since selection bias is uncontrollable in a dataset completely drawn from real-world, we have to adopt a simulated environment [8] with both semi-synthetic and synthetic datasets, which shall be explained in section 5. In our environment, we propose the **softmax observation model** and introduce a hyperparameter β to control the degree of bias. For a rating matrix R, the corresponding probability matrix of each rating being observed is $\Pr(R_{u,i} \text{ is observed}) = k\text{softmax}(\beta R_{u,i})$, where k is pre-set so that the expected proportion of observed ratings is controlled. The effect of β on probability of being observed for different ratings is visualized in figure 1. In our experiment we assume constantly 10% of ratings are observed among 1000 users and 1000 items.

3.2 Results

Figure 2 shows our results on a synthetic dataset (see section 5) and 3 classic algorithms: user-KNN, item-KNN, and SVD matrix factorization. The three differently quantized variants come from the same source distribution (the dataset). Gaussian noise is introduced for simulating real-world user noise. Shaded regions denote the regions between quartiles.

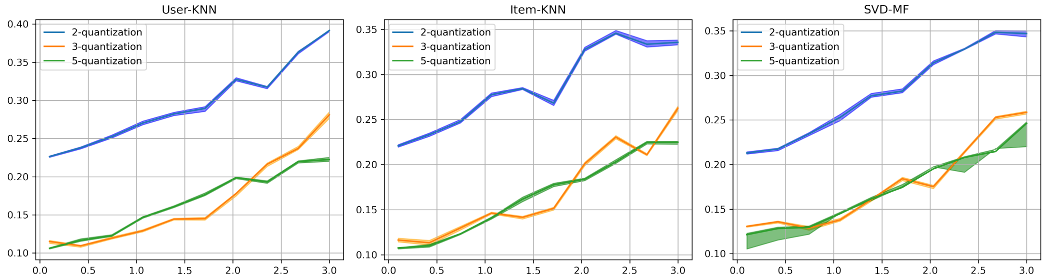

Figure 2: The effect of selection bias is not affected by quantization

Although the RMSE grows consistently as β increases for algorithms, differently quantized datasets do not exhibit significantly different growth rates. The discrepancy between the RMSE of differently quantized datasets with the same β can be ascribed to the inherent information loss from high quantization [14]. We thus conclude it is a viable approach to use the more quantized, less biased dataset for propensity estimation. Furthermore, figure 2. visualizes the damaging effect of selection bias. The less quantized datasets breaks even with the most coarse binary dataset at $\beta = 2.5$ versus $\beta = 0$, suggesting the practical importance of debiasing recommender systems. The above experiments are repeated under a semi-synthetic environment and the conclusion holds.

4 Cross-Dataset Propensity Estimation

With the above safety guarantee, we now turn to our cross-dataset matrix factorization model. Matrix factorization is a simple recommender model that decomposes the rating matrix based on known entries and then predicts the unknown entries [7]. We integrate propensities as [11] did and formulated the recommendation problem as the empirical risk minimization framework below.

$$\arg\min_{V,W,A} \frac{1}{N} \left(\frac{(Y_{u,i} - (V_u^T W_i + A))^2}{P_{u,i}} \right) + c\|A\|^2 \tag{1}$$

where $A = \{b_u, b_i, \mu\}$ represents the standard bias parameters (offset), V and W are the decomposed vectors, $\hat{Y} = V_u^T W_i + A$ is the predicted rating, N is the number of ratings, and $c\|A\|^2$ is the regularizer. The inverse propensity scores ${}^1/_{P_{u,i}}$ are multiplied to each rating during learning, which is analogous to re-weighting ratings based on their biases.

Denote the biased training set as D. we propose the **naive-bayes propensity estimator** (NBPE-MF) from a more quantized (binary) dataset D'. Essentially,

$$P_{u,i} = \Pr(Y_{u,i} \text{ is observed} \mid Y_{u,i} = r_{u,i}) = \frac{\Pr(Y_{u,i} = r_{u,i} \mid Y_{u,i} \text{ is observed})}{\Pr(Y_{u,i} = r_{u,i})}, \tag{2}$$

where $r_{u,i}$ denotes the value of a rating. The numerator can be easily approximated from the dataset, but the denominator requires additional data from less biased sources. In this paper, we estimates the denominator from the less biased dataset using a categorical naive-bayes model. We use D' as a mask to hide biased ratings in the training set from propensity estimation. In other words, $Y_{u,i}$ is considered observed only when $D'_{u,i}$ is observed:

$$\Pr(Y_{u,i} = r_{u,i}) = \frac{\sum_1^N \mathbf{1}_{(Y_{u,i}^D = r_{u,i})} \cdot \mathbf{1}_{(Y_{u,i}^{D'} \text{ is observed})}}{N}. \tag{3}$$

If the corresponding rating is not captured in D', we set its propensity to the average propensity.

5 Experiment

We designed experiments to verify the performance of our cross-dataset model. We trained on two datasets with $\beta = 1$ in training set and used three baseline algorithms for comparison. In this paper we assumed the highly quantized dataset is unbiased with $\beta = 0$, which is probable when the implicit

Table 1: Test set RMSE and MAE for NBPE-MF and baselines

	ML100K		Latent Factors	
	RMSE	MAE	RMSE	MAE
MF	0.1046	0.0833	0.1331	0.1045
NPE-MF	0.1048	0.0834	0.132	0.1049
MD-MF	0.1044	0.083	0.1299	0.1027
NBPE-MF	**0.076**	**0.062**	**0.1065**	**0.0841**

dataset reflects users' spontaneous choices. Theoretically, our framework will work as long as the highly quantized dataset is less biased. We selected the root mean square error (RMSE) and the mean absolute error (MAE) as performance metrics.

5.1 Datasets

Imputed ML100K Dataset. The ML100K dataset provides 100 thousand MNAR ratings across 1683 movies rated by 944 users and is the standard large-scale dataset used for recommender systems. Since we need the ground truth ratings for controlling bias, we impute the missing ratings using standard matrix factorization. Ratings are normalized between 0 and 1 for consistency with the other dataset.

Latent Factors Simulated Dataset. The latent factors dataset is a synthetic dataset that models real-world user behavior. Users and item both have random latent vectors to simulate preferences; both also have biases. The environment is provided by [8].

5.2 Baselines

Matrix Factorization (MF.) As a simple baseline, we adopt the standard matrix factorization that does not adopt propensity estimation or importance weighting.

Naive Propensity Estimator (NPE-MF.) The naive propensity estimator naively estimates the propensity scores from the already biased training data and plugs in the results to equation 1.

Mixing Datasets Matrix Factorization (MD-MF.) To analyze whether our cross-dataset model more efficiently exploits all existing data, we create a larger dataset by mixing the two differently quantized datasets and train the MF algorithm on it.

5.3 Results

Table 1 shows the the experiment's results, each entry as the average of five independent trials. Averaged across datasets and metrics, our approach gains 23.1% reduction in error compared to matrix factorization, 23.1% compared to naive propensity estimator, and 22.1% compared to mixing datasets matrix factorization.

Additionally, we observe that NPE-MF and MF have very similar performance across all settings. This suggests that a IPS-based model will not gain any performance boost if propensities are inaccurately estimated. Furthermore, we argue that our model is data-efficient, i.e., it beats the competitor model (MD-MF), demonstrating that utilizing less biased dataset for propensity estimation is a more viable approach than mixing it with other data, though the latter is sometimes a common engineering practice. We nudge the way two datasets are mixed in MD-MF and discover similar results. On our benchmarks, mixing datasets MF consistently achieves less than 2% advantage compared to using only one dataset.

We conclude that our cross-dataset model significantly outperforms baselines and provides a efficient way to make use of more quantized (implicit) data.

6 Conclusion

We propose an effective and data-efficient method of combing differently sourced datasets for training recommender systems. Our approach provides yet another way of accurate propensity estimation and explores a new potential of cross-dataset joint learning. We believe real-world recommender systems will benefit from data-centric frameworks like ours.

Acknowledgements

This work is supported in part by a Cornell CSURP Grant, *is there some source of funding you'd like to acknowledge?*

References

[1] Charu C Aggarwal et al. *Recommender systems*, volume 1. Springer, 2016.

[2] Xavier Amatriain, Josep M Pujol, and Nuria Oliver. I like it... i like it not: Evaluating user ratings noise in recommender systems. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 247–258. Springer, 2009.

[3] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *arXiv preprint arXiv:2010.03240*, 2020.

[4] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[5] Gawesh Jawaheer, Martin Szomszor, and Patty Kostkova. Comparison of implicit and explicit feedback from an online music recommendation service. In *proceedings of the 1st international workshop on information heterogeneity and fusion in recommender systems*, pages 47–51, 2010.

[6] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the tenth ACM international conference on web search and data mining*, pages 781–789, 2017.

[7] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[8] Karl Krauth, Sarah Dean, Alex Zhao, Wenshuo Guo, Mihaela Curmei, Benjamin Recht, and Michael I Jordan. Do offline metrics predict online performance in recommender systems? *arXiv preprint arXiv:2011.07931*, 2020.

[9] Bruno Pradel, Nicolas Usunier, and Patrick Gallinari. Ranking with non-random missing ratings: influence of popularity and positivity on evaluation metrics. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 147–154, 2012.

[10] Masahiro Sato, Sho Takemori, Janmajay Singh, and Tomoko Ohkuma. Unbiased learning for the causal effect of recommendation. In *Fourteenth ACM Conference on Recommender Systems*, RecSys '20, page 378–387, New York, NY, USA, 2020. Association for Computing Machinery.

[11] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*, pages 1670–1679. PMLR, 2016.

[12] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

[13] Steven K Thompson. *Sampling*, volume 755. John Wiley & Sons, 2012.

[14] Bernard Widrow, Istvan Kollar, and Ming-Chang Liu. Statistical theory of quantization. *IEEE Transactions on instrumentation and measurement*, 45(2):353–361, 1996.

[15] Longqi Yang, Yin Cui, Yuan Xuan, Chenyang Wang, Serge Belongie, and Deborah Estrin. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *Proceedings of the 12th ACM conference on recommender systems*, pages 279–287, 2018.

[16] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. Importance weighted adversarial nets for partial domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8156–8164, 2018.