Balanced Neighborhoods for Fairness-aware Collaborative Recommendation

Robin Burke
School of Computing
DePaul University
Chicago, Illinois
rburke@cs.depaul.edu

Nasim Sonboli
School of Computing
DePaul University
Chicago, Illinois
nsonboli@depaul.edu

Masoud Mansoury
School of Computing
DePaul University
Chicago, Illinois
mmansou4@depaul.edu

Aldo Ordoñez-Gauger
School of Computing
DePaul University
Chicago, Illinois
aordone3@mail.depaul.edu

ABSTRACT

Recent work on fairness in machine learning has begun to be extended to recommender systems. While there is a tension between the goals of fairness and of personalization, there are contexts in which a global evaluations of outcomes is possible and where equity across such outcomes is a desirable goal. In this paper, we introduce the concept of a balanced neighborhood as a mechanism to preserve personalization in recommendation while enhancing the fairness of recommendation outcomes. We show that a modified version of the SLIM algorithm can be used to improve the balance of user neighborhoods, with the result of achieving greater outcome fairness in a real-world dataset with minimal loss in ranking performance.

KEYWORDS

Fairness, Recommender systems, Machine Learning, Neighborhood Models

ACM Reference format:
Robin Burke, Nasim Sonboli, Masoud Mansoury, and Aldo Ordoñez-Gauger. 2017. Balanced Neighborhoods for Fairness-aware Collaborative Recommendation. In *Proceedings of ACM FATRec Workshop, Como, Italy, August 2017 (FATRec'17),* 5 pages.
https://doi.org/10.18122/B2GQ53

1 INTRODUCTION

Bias and fairness in machine learning are topics of considerable recent research interest [1, 3]. A standard approach in this area is to identify a variable or variables representing membership in a protected class, for example, race in an employment context, and to develop algorithms that remove bias relative to this variable. See, for example, [7, 8, 11, 13, 14].

To extend this concept to recommender systems, we must recognize the key role of personalization. Inherent in the idea of recommendation is that the best items for one user may be different than those for another. The dominant recommendation paradigm, collaborative filtering [9], uses user behavior as its input, ignoring user demographics and item attributes. One approach to fairness in recommendation is to examine outcomes only in terms of the level and type of error experienced by different groups [12]. However, there are contexts in which this approach may be insufficient. Consider a recommender system suggesting job opportunities to job seekers. An operator of such a system might wish, for example, to ensure that male and female users with similar qualifications get recommendations of jobs with similar rank and salary. The system would therefore need to defend against biases in recommendation output, even biases that might arise entirely due to behavioral differences: for example, male users might be more likely to click optimistically on high-paying jobs.

Defeating such biases is difficult if we cannot assert a shared global preference ranking over items. Personal preference is the essence of recommendation especially in areas like music, books, and movies where individual taste is paramount. Even in the employment domain, some users might prefer a somewhat lower-paying job if it had other advantages: such as flexible hours, shorter commute time, or better benefits. Thus, to achieve the policy goal of fair recommendation of jobs by salary, a site operator must go beyond personalization as a goal and impose additional constraints on the recommendation algorithm.

In this paper, we investigate fairness-aware recommendation in the context of recommendation. In particular, we develop the idea of segregation in recommendation, its implications for fairness, and show that a regularization-based approach can be used to control the formation of recommendation neighborhoods. We show that this approach can be used to overcome statistical biases in the distribution of recommendations across users in different groups.

2 BALANCED NEIGHBORHOODS IN RECOMMENDATION

In [13], the authors impose a fairness constraint on a classification by creating a *fair representation*, a set of prototypes to which instances are mapped. The prototypes each have an equal representations of users in the protected and unprotected class so that the association between an instance and a prototype carries no information about the protected attribute.

FATRec'17, August 2017, Como, Italy

DOI: 10.18122/B2GQ53

As noted above, the requirement for personalization in recommendation means that we have as many classification tasks as we have users. A direct application of the fair prototype idea would aggregate many users together and produce the same recommendations for all, greatly reducing the level of personalization and the recommendation accuracy. This idea must be adapted to apply to recommendation.

One of the fundamental ideas of collaborative recommendation is that of the *peer user*, a neighbor whose patterns of interest match those of the target user and whose ratings can be extrapolated to make recommendations for the target user. One place where bias may creep into collaborative recommendation may be through the formation of peer neighborhoods.

Consider the situation in Figure 1. The target user here is the solid square, a member of the protected class. The top of the figure shows a neighborhood for this user in which recommendation will be generated only from other square users, that is, other protected individuals. We can think of this as a kind of segregation of the recommendation space. If the peer neighborhoods have this kind of structure relative to the protected class, then this group of users will only get recommendations based on the behavior and experiences of users in their own group. For example, in the job recommendation example above, women would only get recommendations of jobs that have interested other women applicants, potentially leading to very different recommendation experiences across genders.

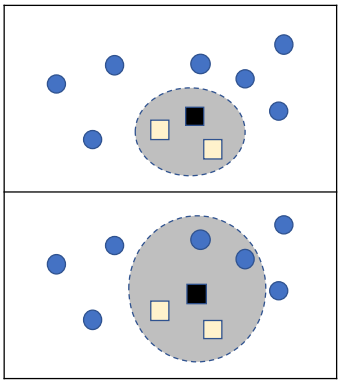

Figure 1: Unbalanced (top) and balanced (bottom) neighborhoods

To counter this type of bias, we introduce the notion of a *balanced neighborhood*. A balanced neighborhood is one in which recommendations for all users are generated from neighborhoods that are balanced with respect to the protected and unprotected classes. This is shown in the bottom half of Figure 1. The target has an equal number of peers inside and outside of the protected class. In the case of job recommendation discussed above, this would mean that female job seekers get recommendations from some female and some male peers.

There are a variety of ways that balanced neighborhoods might be formed. The simplest way would be to create neighborhoods for each user that balance accuracy against group membership. This could be highly computationally inefficient as it would require solving a separate optimization problem for each user. In this research, we explore an extension of the well-known Sparse Linear Method (SLIM), which has been proved very effective in recommendation ranking with implicit data. This extension uses regularization to control the way different neighbors are weighted, with the goal of achieving balance between protected and non-protected neighbors for each user.

3 SLIM

The Sparse Linear Method for recommendation was introduced in [10]. It is a generalization of item-based k-nearest neighbor in which all items are used and weights for these items are learned through optimization to minimize a regularized loss function. Although this is not proposed in the original SLIM paper, it is possible to create a user-based version of SLIM (labeled SLIM-U in [15]), which generalizes the user-based algorithm in the same way.

Assume that there are M users (a set U), N items (a set I), and let us denote the associated 2-dimensional rating matrix by R. SLIM is designed for item ranking and therefore R is typically binary. We will relax that requirement in this work, We use u_i to denote user i and t_j to denote the item j. An entry, r_{ij}, in matrix R represents the rating of u_i on t_j.

SLIM-U predicts the ranking score $\hat{s}$ for a given user, item pair $< u_i, t_j >$ as a weighted sum:

$$\hat{s}_{ij} = \sum_{k \in U} w_{ik} r_{kj}, \tag{1}$$

where $w_{ii} = 0$ and $w_{ik} >= 0$.

Alternatively, this can be expressed as a matrix operation yielding the entire prediction matrix $\hat{S}$:

$$\hat{S} = WR, \tag{2}$$

where W is an MxM matrix of user-user weights. For efficiency, it is very important that this matrix be sparse.

The optimal weights for SLIM-U can be derived by solving the following minimization problem:

$$\min_W \frac{1}{2} \|R - WR\|^2 + \lambda_1 \|W\|^1 + \frac{\lambda_2}{2} \|W\|^2, \tag{3}$$

subject to $W > 0$ and $diag(W) = 0$.

The $\|W\|^2$ term represents the ℓ_2 norm of the W matrix and $\|W\|^1$ represents the ℓ_1 norm. These regularization terms are present to constrain the optimization to prefer sparse sets of weights. Typically, coordinate descent is used for optimization. Refer to [10] for additional details.

3.1 Neighborhood Balance

Recall that our aim in fair recommendation to eliminate segregated recommendation neighborhoods where protected class users only receive recommendations from other users in the same class. Such neighborhoods would tend to magnify any biases present in the system: if users in the protected class only are recommended certain items, then they will be more likely to click on those items and thus increase the likelihood that the collaborative system will make these items the ones that others in the protected group see.

To reduce the probability that such neighborhoods will form, we use the SLIM-U formalization of the recommendation problem, but we add another regularization term to the loss function, which we

call the *neighborhood balance* term. To describe this term, we will enrich our notation further by indicating U^+ to be the subset of U containing users in the protected class with the remaining users in the class U^-. Let W_i^+ be the set of weights for users in U^+ and W_i^- be the corresponding set of weights for the non-protected class. Then the neighborhood balance term b_i for a given user i is the squared difference between the weights assigned to peers in the protected class versus the unprotected class.

$$b_i = (\sum_{w^+ in W_i^+} w^+ - \sum_{w^- in W_i^-} w^-)^2 \tag{4}$$

A low value for the neighborhood balance term means that the user's predictions will be generated by weighting protected and unprotected users on a relatively equal basis.

Note that this is a class-blind optimization that tries to build balanced neighborhoods for both the protected and unprotected users. It is also possible to formulate the objective such that it only impacts the protected class and we will leave this option for future work. If the classes are highly imbalanced, it may be necessary to weight these terms so that the weights are expected to sum in proportion to the size of each group. We will explore this idea in future work.

Another way to express this idea is to create a vector p of dimension M. If u_i is in U^+, then $p_i = 1$; if u_i is in U^-, then $p_i = -1$. Then, the sum expressed above can be rewritten as $b_i = \left\|p^T w_i\right\|^2$. By adding up this term for all users and adding it to the loss function, we can allow the optimization process to derive weights with neighborhood balance in mind. This adapted version of SLIM-U we will call *Balanced Neighborhood SLIM* or BN-SLIM.

As in the case of SLIM, we can apply the method of coordinate descent to optimize the objective. The basic algorithm is to choose one w_{ik} weight and solve the optimization problem for that weight, repeating over all the weights until convergence is reached. The full loss function is as follows:

$$L = \frac{1}{2}\left\|R - WR\right\|^2 + \lambda_1\left\|W\right\|^1 + \frac{\lambda_2}{2}\left\|W\right\|^2 + \frac{\lambda_3}{2}\sum_{i\in U}\left(\sum_{k\in U} p_i w_{ik}\right)^2, \tag{5}$$

where $w_{ii} = 0$ and $w_{ik} >= 0$ and where λ_3 is a parameter controlling the influence of the neighborhood balance calculation on the overall optimization

This loss function retains the property of the original SLIM algorithm in that the rows of the weight matrix are independent, and the weights in each row (those for each user) can be optimized independently. If we take the derivative of L with respect to a single weight w_{ik}, we obtain

$$\frac{\partial L_i}{\partial w_{ik}} = \sum_{j\in I}(r_{ij} - \sum_{l\in U'} w_{il} r_{lj}) + w_{ik}\sum_{j\in I} r_{kj}^2 + \lambda_1 + \lambda_2 w_{ik} + \lambda_3 p_k \sum_{l\in U'} p_l w_{il} \tag{6}$$

where $U' = U - \{u_i, u_k\}$.

We then set this derivative to zero and solve for the value of w_{ik} that produces this minimum. This becomes the coordinate descent update step.

$$w_{ik} \leftarrow \frac{S\left(\sum_{j\in I}(r_{ij} - \sum_{l\in U'} w_{il} r_{lj}) + \lambda_3 p_k \sum_{l\in U'} p_l w_{il}, \lambda_1\right)_+}{\sum_{j\in I} r_{kj}^2 + \lambda_2 + \lambda_3} \tag{7}$$

where $S()_+$ is the soft threshold operator defined in [4].

4 METHODOLOGY

It is very difficult to find datasets that contain the kind of features that would be necessary to evaluate fairness-aware recommendation algorithms based on user demographics. For example, the data from the job search site XING[1] that was made available for the 2017 RecSys Challenge[2] does not have any demographic information about users except their broad geographic region.

For the purposes of demonstration, we are using the MovieLens 1M dataset [6], which contains user gender information. Movie recommendation is, of course, a domain of pure individual taste and therefore not an obvious candidate for fairness-aware recommendation. Following the example of [12], our approach to construct an artificial equity scenario within this data for expository purposes only, with the understanding that real scenarios can be approached with a similar methodology.

Our artificial scenario centers on movie genres. It can be seen in this data that there is a minority of female users (1709 out of the total of 6040). Certain genres display a discrepancy in recommendation delivery to male and female users. For example, in the "Crime" genre, female users rate a very similar number of movies (average of 0.048% of female profiles vs 0.049% of male profiles) and rate them similarly: an average rating of 3.689 for female users vs 3.714 for male users. However, our baseline unmodified SLIM-U algorithm recommends in the top 10 an average of 1.10 "Crime" movies per female user as opposed to 1.18 such movies to male users. We are still exploring the cause of this discrepancy, but it seems likely that there are influential female users with a lower opinion of this genre.

Given that the rating profiles are similar but the recommendation outcomes are different, we can therefore conclude that the female users experience a deprivation (if one wants to call it that) of "Crime" movies compared to their male counter-parts. Similar losses can be observed for other genres. It is, of course, questionable if there is any harm associated with this outcome and we do not claim such. It is sufficient that these differences allow us to validate the properties of the BN-SLIM algorithm.

Our goal, then, is to reduce or eliminate genre discrepancies with minimal accuracy loss by constructing balanced neighborhoods for the MovieLens users. The p vector in Equation 7 therefore will have a 1 for female users and a -1 for male users. In the experiments below, we compare the user-based SLIM algorithm in its unmodified form and the balanced neighborhood version BN-SLIM.

In evaluating fairness of outcome, we measure the number of movies of the chosen genre as the measure of outcome quality.

[1] https://www.xing.com/jobs
[2] http://2017.recsyschallenge.com/

Therefore, we construct a genre-level equity score, $E@k$ for recommendation lists of k items, as the ratio between the outcomes for the different groups. Let $P_i@k = \rho_1, \rho_2, ..., \rho_k$ be the top k recommendation list for user i, and let $c()$ be a function $\rho \rightarrow 0, 1$ that maps to 1 if the recommended movie is in the chosen genre. Then:

$$E@k = \frac{\sum_{i \in U^+} \sum_{\rho \in P_i@k} c(\rho)/|U^+|}{\sum_{i \in U^-} \sum_{\rho \in P_i@k} c(\rho)/|U^-|} \quad (8)$$

$E@k$ will be less than 1 when the protected group is, on average, recommended fewer movies of the desired genre. It may be unrealistic to imagine that this value should approach 1: the metric does not correct for other factors that might influence this score – for example, female users may rate a particular genre significantly lower and an equality of outcome should not be expected. While the absolute value of the metric may be difficult to interpret, it is still useful for comparing algorithms. The one with the higher $E@k$ is providing more movies in the given genre to the protected group.

As in any multi-criteria setting, we must be concerned about any loss of accuracy that results from taking additional criteria into consideration. Therefore, we also evaluate NDCG@10 for our algorithms in the results below.

5 RESULTS

We implemented the SLIM-U and BN-SLIM algorithms using LibRec 2.0 [5]. We used 5-fold cross-validation as implemented within the library. Within the MovieLens 1M dataset, we selected the five genres on which the SLIM-U algorithm produced the lowest equity scores: "Film-Noir", "Mystery", "Horror", "Documentary", and "Crime". The parameters were set as follows: $\lambda_1 = 0.1$, $\lambda_2 = 0.001$, and (for BN-SLIM) $\lambda_3 = 25$[3].

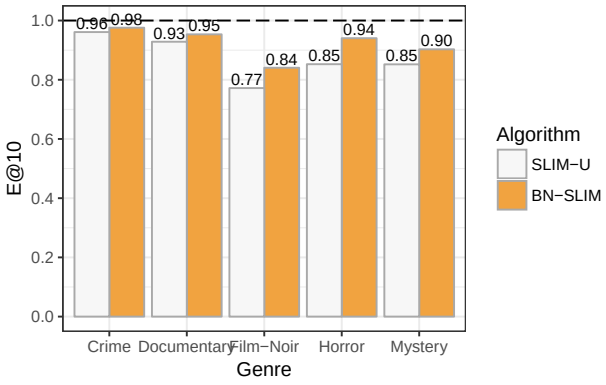

Figure 2: Equity score for SLIM-U and BN-SLIM. Line indicates equal percentage across genders

Figure 2 shows the results of the experiment in terms of the equity scores for each genre. Perfect equity (1.0) is marked with the dashed line. As we can see, in every case, the balanced neighborhood algorithm produced an equity score closer to 1.0 than the unmodified algorithm. The largest jump is seen in the "Horror" genre, about 0.09 in the equity score or around 10%.

[3]Because the balance term measures the difference in weights, it tends to be much smaller than the terms that measure the sums of weights. Therefore, the regularization constant must be much higher for it to have an impact.

Algorithm	NDCG@10
SLIM-U	0.053
BN-SLIM	0.052

Table 1: Ranking accuracy

In terms of accuracy, there was only a small loss of NDCG@10 between the two conditions. See Table 1. The difference amounts to approximately 2% loss in NDCG@10 for the balanced neighborhood version.

Because the balanced neighborhood algorithm is applied across all users, it also has the effect of showing male users movie genres that occur more frequently for female users. To see this effect, we examined the five genres with the highest E@10 values: "Fantasy", "Animation", "War", "Romance", and "Western" using the same parameter values as above. The results appear in Figure 3 and show a similar result. "War" is clearly the anomaly here, both because it is surprising to see it as a one of the more female-recommended genres and because the genre-balance algorithm pushes it to become more skewed rather than less. We are investigating the cause of this phenomenon. Overall, the BN-SLIM algorithm produces a recommendation experience in which the occurrence of gender-specific genres is more closely equalized, with small loss in ranking accuracy.

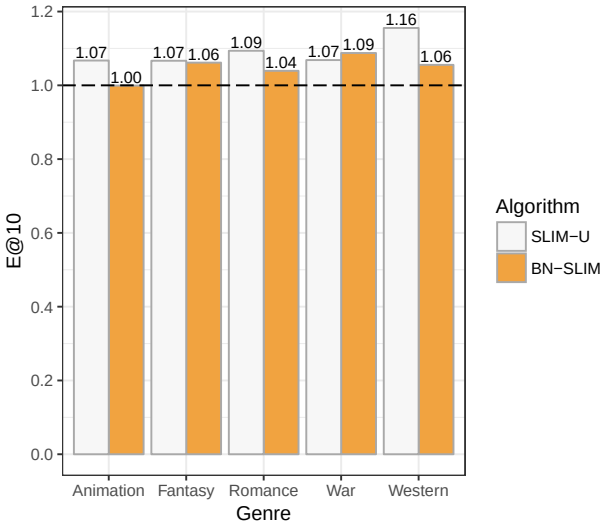

Figure 3: Equity scores for female-preferred genres

6 CONCLUSION

Considerations of fairness and equity are in tension with the focus on personalization that is central to recommender systems research. To ask if a recommendation outcome is fair, by definition, assumes some kind of universal standard for such outcomes, existing outside of individual preference. In some recommendation domains,

such as employment and housing, it is reasonable to expect that recommender systems may be held to such standards.

In this paper, we consider one way in which a fair outcome for a protected group may be sought in the context of personalized recommendation. Drawing on the idea of fair prototypes [13], we propose the construction of balanced neighborhoods as a mechanism for achieving fair outcomes in recommendation and we provide an implementation of the idea using a variant of the Sparse Linear Method.

Although we were not able to demonstrate results in a domain in which fair outcomes are critical, we were able to construct an evaluation using the MovieLens data set and show that our balanced neighborhood implementation overcomes biases inherent in the data with respect to male and female users and the recommendation of different genres with minimal loss in ranking accuracy.

In future work, we hope to acquire appropriate data to evaluate our approach in areas where fairness is of greater societal importance, and to extend the balanced neighborhood approach to other algorithms. Finally, we are also interested in scenarios in which there are fairness considerations for both sides of the recommendation transaction, such as reciprocal recommendation scenarios [2].

7 ACKNOWLEDGMENTS

This work is supported in part by the National Science Foundation under grant IIS-1423368.

REFERENCES

[1] Engin Bozdag. 2013. Bias in algorithmic filtering and personalization. *Ethics and Information Technology* 15, 3 (Sept. 2013), 209–227. https://doi.org/10.1007/s10676-013-9321-6

[2] Burke, Robin. 2017. Multisided Fairness for Recommendation. In *Workshop on Fairness, Accountability and Transparency in Machine Learning (FATML)*. Halifax, Nova Scotia, To appear.

[3] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, 214–226.

[4] Jerome Friedman, Trevor Hastie, Holger HÃűfling, and Robert Tibshirani. 2007. Pathwise coordinate optimization. *The Annals of Applied Statistics* 1, 2 (Dec. 2007), 302–332. https://doi.org/10.1214/07-AOAS131

[5] Guibing Guo, Jie Zhang, Zhu Sun, and Neil Yorke-Smith. 2015. LibRec: A Java Library for Recommender Systems.. In *UMAP Workshops*.

[6] F Maxwell Harper and Joseph A Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 4 (2015), 19.

[7] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 869–874.

[8] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. *Machine Learning and Knowledge Discovery in Databases* (2012), 35–50.

[9] Y. Koren and R. Bell. 2011. Advances in collaborative filtering. *Recommender Systems Handbook* (2011), 145–186.

[10] Xia Ning and George Karypis. 2011. Slim: Sparse linear methods for top-n recommender systems. In *11th IEEE International Conference on Data Mining (ICDM)*. IEEE, 497–506.

[11] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 560–568.

[12] Sirui Yao and Bert Huang. 2017. Beyond Parity: Fairness Objectives for Collaborative Filtering. *CoRR* abs/1705.08804 (2017). http://arxiv.org/abs/1705.08804

[13] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. 325–333.

[14] Lu Zhang and Xintao Wu. 2017. Anti-discrimination learning: a causal modeling-based framework. *International Journal of Data Science and Analytics* (2017), 1–16.

[15] Yong Zheng, Bamshad Mobasher, and Robin Burke. 2014. CSLIM: Contextual SLIM recommendation algorithms. In *Proceedings of the 8th ACM Conference on Recommender Systems*. ACM, 301–304.