PERSPeCT: Collaborative Filtering for Tailored Health Communications

Roy J. Adams
School of Computer Science
UMass Amherst
rjadams@cs.umass.edu

Rajani S. Sadasivam
Department of Quantitative Health Sciences
UMass Medical School

Kavitha Balakrishnan
Department of Quantitative Health Sciences
UMass Medical School

Rebecca L. Kinney
Department of Quantitative Health Sciences
UMass Medical School

Thomas K. Houston
eHealth QUERI
Bedford VA Medical Center
UMass Medical School

Benjamin M. Marlin
School of Computer Science
UMass Amherst

ABSTRACT

The goal of computer tailored health communications (CTHC) is to elicit healthy behavior changes by sending motivational messages personalized to individual patients. One prominent weakness of many existing CTHC systems is that they are based on expert-written rules and thus have no ability to learn from their users over time. One solution to this problem is to develop CTHC systems based on the principles of collaborative filtering, but this approach has not been widely studied. In this paper, we present a case study evaluating nine rating prediction methods for use in the *Patient Experience Recommender System for Persuasive Communication Tailoring*, a system developed for use in a clinical trial of CTHC-based smoking cessation support interventions.

Categories and Subject Descriptors

H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval

Keywords

Recommender systems; tailored health communications

1. INTRODUCTION

Computer tailored health communications (CTHC) systems aim to elicit healthy behavior changes by sending motivational messages personalized to individual patients. Current CTHC systems start by creating a pool of messages and a fixed set of theory-driven, expert-written rules that relate the characteristics of messages to the characteristics of the patients [4]. When a new patient is enrolled in the system, a baseline patient profile is collected that includes demographic information and assessments of the relevant tailoring characteristics. Messages are then selected for the patient on the basis of the patient's characteristics and the expert-derived message selection rules.

CTHC systems have been deployed to address a range of high-profile health problems including supporting smoking cessation [12]. One prominent weakness of many existing CTHC systems is that they do not solicit feedback from their users and thus have no ability to learn over time. The fixed rules these systems are based on may fail to account for socio-cultural concepts that have intrinsic importance to the targeted population, thus limiting their success [1].

One solution to this problem is to develop CTHC systems based on the principles of collaborative filtering. In electronic commerce settings, collaborative filtering recommender systems have been successfully used to derive personalized recommendations for items like books or movies based on implicit or explicit feedback about the items collected from an online community of users. The promise of collaborative filtering-based recommendation in the CTHC setting derives from its ability to learn associations between types of patients and types of messages directly from data, allowing the CTHC system to adapt to individual users over time [9].

Collaborative filtering in the CTHC setting presents a number of challenges. Our prior work has addressed the problem of defining semantics for explicit feedback ratings in the CTHC setting as a precursor to developing the *Patient Experience Recommender System for Persuasive Communication Tailoring* (PERSPeCT) [5]. The PERSPeCT system will use collaborative filtering-based rating prediction to drive the personalized selection of smoking cessation support messages that are currently in use in a deployed expert system operated through our online smoking cessation support portal, decide2quit.org.

The primary challenge faced by PERSPeCT is the need to construct a system with maximal accuracy in the small data regime. In this case study, we present an empirical evaluation of nine classical and state of the art rating prediction methods applied to the novel problem of predicting smoking cessation support message ratings. We begin by reviewing the problem of rating semantics in the CTHC setting. We then present the details of a new data set collected

RecSys'14, October 6–10, 2014, Foster City, Silicon Valley, CA, USA.

ACM 978-1-4503-2668-1/14/10
http://dx.doi.org/10.1145/2645710.2645768.

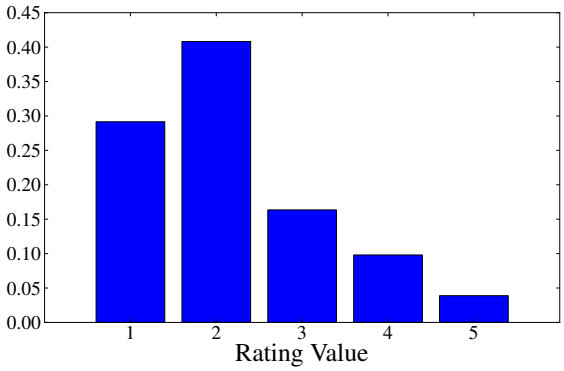

Figure 1: The marginal distribution over ratings.

to bootstrap collaborative filtering model learning and selection for PERSPeCT and provide a brief overview of the models evaluated in our study. Finally, we describe our results, which underscore the importance of Bayesian methods in small data regimes and surprisingly show that the inclusion of patient and message features has no marginal benefit in addition to the available rating information.

2. STUDY DESIGNS & DATA COLLECTION

To date, two studies have been performed to guide the design of PERSPeCT. The first study was designed to inform the choice of rating semantics, which was a significant open question resulting from the differing goals of e-commerce recommender systems (recommending items that users will like) and CTHC systems (selecting messages to promote healthy behavior changes for individual patients).

To address this issue, four candidate factors including how much the message influenced the user not to smoke, how much the message effected the subject emotionally, how relevant the message was to the subject, and how much the user liked the message were identified. To assess the differences between these factors, we collected a novel multi-factor rating data set consisting of ratings for all four message factors collected from 100 study participants. Each participant provided ratings on a 5-point scale for each of five unique messages drawn randomly from a 50 message subset of all messages available on decide2quit.org. The results showed that the ratings for each of the four factors were highly correlated, rendering the collection of all four ratings for each message unnecessary. These results are presented in detail in our prior work [5], which also includes ratings analysis and a small-scale rating prediction feasibility study.

Following the completion of the rating semantics study, a second study was performed to collect a larger rating data set to bootstrap the learning and evaluation of collaborative filtering models for PERSPeCT. We collected ratings of both influence and relevance from a total of 846 users with respect to all 261 actively used messages from decide2quit.org. Each study participant was asked to rate 20 messages. This resulted in a total of 16920 ratings, yielding a data density of 7.7%. Subjects also supplied demographic data and values for a variety of tailoring variables related to smoking behavior including number of cigarettes smoked per day and readiness to quit. Study subjects were recruited through a mix of local and online sources and received a $50 incentive.

The marginal distribution of influence ratings for this data set is shown in Figure 1. In the remainder of this paper, we present the first empirical analysis of influence rating prediction using the data collected in this second study.

3. METHODS AND MODELS

Baselines: Our evaluation includes a number of common baselines including predicting the global mean rating (Global Mean) and the item mean rating (Item Mean). The item mean model includes smoothing hyperparameters that pull the means toward the global mean. We also include a simple additive model that learns an optimal additive combination of ℓ_2 regularized user and item bias terms.

KNN: User and item-based K-Nearest Neighbors are classical models in collaborative filtering. A user-based KNN model works by predicting user u's rating for item i as a weighted sum of other users' ratings of i, where the weights are a similarity measure. Following [2], we used Pearson Correlation between ratings on co-rated items, adjusted to account for the number of co-rated items. KNN models have two hyperparameters: the neighborhood size, K, and the weight for pairs with few shared ratings.

PMF: Probabilistic Matrix Factorization represents each user and item as a K-dimensional vector with a multivariate normal prior. User u's rating for item i is then a normal random variable with mean equal to the inner product of u's and i's factors. The user and item factors can be learned using standard numerical optimization techniques [8]. Additionally, global, user, and item biases can be added to the model with little additional complexity. The hyperparameters in PMF include the ℓ_2 regularization parameters for the user/item biases and factor matrices, as well as the dimensionality of the model K.

BPMF: Bayesian Probabilistic Matrix Factorization extends PMF by adding Normal-Wishart hyper-priors to the mean and covariance matrices of the user and item factors and replaces learning the primary parameters with optimization by full Bayesian inference using Gibbs sampling [7].The hyperparameters in BPMF include the parameters for the Normal-Wishart hyper-priors, the rating noise variance, and the dimensionality of the factors.

PMF-F/BPMF-F: A simple way to incorporate user and item side-information into both of the previous models is to add linear regression terms to the models. In PMF, the feature weights can be ℓ_2 regularized and estimated along with all of the other parameters. In BPMF, the individual feature weights are given normal priors with zero mean and tunable variance and added to the Gibbs sampling routine.

CMF: Collective Matrix Factorization jointly factorizes the rating matrix and a related feature matrix into a set of user, item, and feature factors [10]. In our case, the item factors are shared between the two factorizations and allow information to flow between the feature and ratings matrices. A hyperparameter, α, balances the importance of the reconstruction error on the ratings against the reconstruction error on the features. Parameters in this model are learned via numerical optimization as in PMF. The hyperparameters in CMF are the same as those in PMF plus α and an ℓ_2 regularization parameter for the feature factors.

BCMF: Bayesian Collective Matrix Factorization extends CMF in the same way that BPMF extends PMF, with Normal-Wishart hyper-priors on the user, item, and feature hyperparameters [11]. The same distributions as in BPMF are in-

volved, so we used a similar Gibbs sampler to run inference in this model. We used the same Normal-Wishart parameters as in BPMF and the the only additional hyperparameter is a noise variance parameter for the features.

HFT: In addition to the user demographics and theory based message codings, the text of the messages themselves can be used as features. Similar to CMF, the recently introduced Hidden Factors as Topics model incorporates text by jointly fitting a PMF model to the ratings data and a latent Dirichlet allocation (LDA) model to the message text where a softmax transformation of the item factor is used as the distribution over topics [6]. The model is trained by alternating numerical optimization of the user/item factors and topic specific distributions over words and sampling the topic assignments for individual words. The only additional hyperparameter is similar to α in CMF and balances the importance of the LDA portion of the model.

4. PROTOCOLS AND METRICS

The fielded PERSPeCT system will select messages for users based on their predicted influence ratings. One previously unseen message will be selected and sent per day for each user. The system will initially be evaluated on its ability to select messages that are rated higher on average than messages selected by an existing expert system. These design factors are reflected in our evaluation methodology as described below. All methods were implemented by the authors using Python with the exception of HFT, for which we use the originally published code [6].

Empirical Protocol: We use a strong-generalization protocol that evaluates the ability of the system to generalize to non-training users, which better matches our application. This involves completely separating test users from train users, learning a model using all of the train users' ratings, freezing all non-user-specific parameters, and finally training the user-specific parameters (e.g. user factors and biases in PMF) on a subset of each test user's observed ratings. To implement this protocol, we first divided the users randomly into five folds and then generated three random train and validation sets for each test fold. We further divided each test user's ratings into five folds. To evaluate each method's performance given varying levels of information about a test user, we evaluated all methods with 5, 10 and 16 of each test user's ratings available for inference and learning of user-specific parameters. Each test user has a constant set of 4 test ratings per test fold. The validation sets were used to set the hyper-parameters of each method (e.g. K in KNN). Exhaustive grid search was used and the hyper-parameter ranges were iteratively extended to ensure that no selected hyperparameter values occurred at the endpoints of the search intervals.

Performance Metrics: In evaluating rating prediction methods, we considered a range of standard performance metrics including root mean squared error (RMSE), Kendal's Tau-b (KTAU-b), and normalized discounted cumulative gain (NDCG) to analyze rating prediction and ranking performance. We also evaluate some less common metrics including the fraction of times a test item with the maximal rating is ranked first (Hit Rate), the absolute difference between the maximum test item rating and the rating of the top ranked test item (Top Loss), and the average rating of the top-ranked test items (Avg. Top). We report averages for these metrics over all test folds.

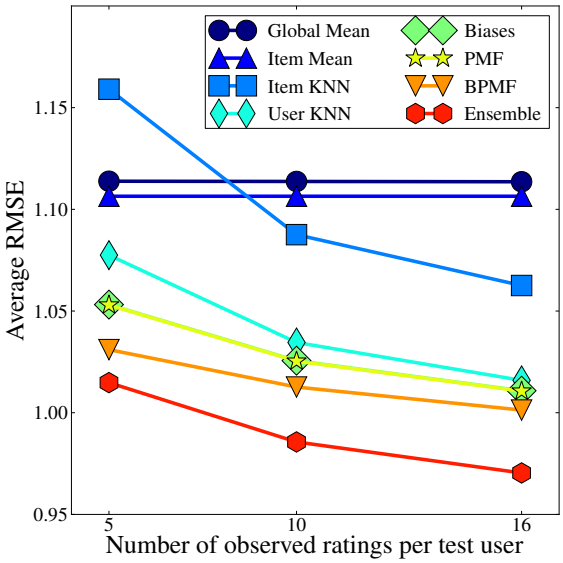

Figure 2: Average RMSEs for all models without features using 5, 10, and 16 training examples.

5. EXPERIMENTS AND RESULTS

In this section, we present an empirical evaluation of the methods described in the previous section applied to the prediction of the influence ratings collected in our second PERSPeCT study.

Pure Methods: The average test performance for all collaborative filtering models based on 10 test user observations is reported in Table 1 along with standard errors. In terms of RMSE, the best performing pure (Type P) collaborative filtering method is BPMF, with an RMSE of 1.001. The absolute performance gap between BPMF and the other base methods is small, ranging between 1 and 10 percent improvement. However, the RMSE gap between BPMF and all the other pure methods is statistically significant at the $p = 0.01$ level as determined by a paired t-test with Bonferroni correction. Figure 2 shows the performance of each of the pure collaborative filtering methods across 5, 10 and 16 observed ratings per test user. We can see that the performance gap is larger with fewer observed ratings, which highlights the superiority of Bayesian models in small data settings. Further investigation showed that BPMF was able to effectively use more factors than PMF. Across all runs, BPMF selected between 4 and 7 factors, while PMF selected only 1 or 2 factors. This highlights the ability of Bayesian methods to learn richer models while avoiding over fitting problems that effect optimization-based approaches in small data settings. A surprising finding is that many methods, including PMF, perform worse or no better than the Biases baseline approach in terms of RMSE. This indicates the difficulty of extracting useful structure in the small data setting. In general, the performance gap is much smaller with respect to the ranking metrics; however, BPMF is always either the best model or statistically indistinguishable from the best model.

Ensemble Methods: The Netflix competition demonstrated the success of ensemble methods which aggregate the predictions of a number of distinct models [3]. To evaluate ensemble methods in our context, we considered combing the results of the best neighborhood-based method (user

Model	Type	RMSE	KTAU-b	NDCG	Hit Rate	Avg. Top	Top Loss
Global Mean	P	1.114 ± 0.019	0.000 ± 0.000	0.876 ± 0.002	0.000 ± 0.000	2.302 ± 0.032	0.851 ± 0.009
Item Mean	P	1.106 ± 0.019	0.105 ± 0.013	0.894 ± 0.004	0.552 ± 0.012	2.158 ± 0.042	0.706 ± 0.031
Biases	P	1.026 ± 0.014	0.113 ± 0.013	0.895 ± 0.004	0.561 ± 0.014	2.142 ± 0.050	0.690 ± 0.031
Item KNN	P	1.088 ± 0.015	0.078 ± 0.008	0.889 ± 0.003	0.525 ± 0.011	2.201 ± 0.027	0.749 ± 0.020
User KNN	P	1.035 ± 0.012	0.000 ± 0.000	0.876 ± 0.002	0.000 ± 0.000	2.302 ± 0.032	0.851 ± 0.009
PMF	P	1.025 ± 0.014	0.112 ± 0.013	0.895 ± 0.004	0.561 ± 0.014	2.141 ± 0.048	0.689 ± 0.031
BPMF	P	1.013 ± 0.013	0.112 ± 0.014	0.895 ± 0.004	0.560 ± 0.014	2.144 ± 0.047	0.692 ± 0.033
PMF-F	H	1.022 ± 0.013	0.112 ± 0.013	0.895 ± 0.004	0.561 ± 0.013	2.144 ± 0.050	0.692 ± 0.032
BPMF-F	H	1.013 ± 0.013	0.113 ± 0.013	0.896 ± 0.004	0.564 ± 0.013	2.134 ± 0.043	0.682 ± 0.027
CMF	H	1.035 ± 0.012	0.105 ± 0.013	0.894 ± 0.004	0.557 ± 0.012	2.146 ± 0.047	0.694 ± 0.028
BCMF	H	1.012 ± 0.013	0.108 ± 0.011	0.894 ± 0.003	0.559 ± 0.009	2.152 ± 0.039	0.701 ± 0.026
HFT	H	1.031 ± 0.012	0.110 ± 0.014	0.895 ± 0.004	0.559 ± 0.014	2.140 ± 0.051	0.688 ± 0.032

Table 1: Predictive performance for pure (P) and hybrid (H) models with 10 observed ratings per test user.

KNN) and the best matrix factorization method (BPMF) using an unweighted average of their predictions. The average RMSE at each training data size is shown in Figure 2. With 10 test user observations, the ensemble method has an average RMSE of 0.986, approximately 3% better than the best pure method, BPMF. This confirms the utility of ensemble methods in the small data regime.

Hybrid Methods: The models indicated as hybrid (Type H) in Table 1 include features in various ways. The best performing of these models in terms of RMSE is BCMF, however, its performance is statistically indistinguishable from BPMF. In fact, PMF with linear features, CMF, BPMF with linear features, and BCMF are all very close in performance to their counterparts that do not include features. Finally, the addition of text features through HFT also results in performance that is slightly worse than PMF, the corresponding base model. Interestingly, these results indicate that the demographic and behavioral theory-based features, which are currently the basis for assigning messages to users in an existing expert system, do not add any information over ratings alone within the state-of-the-art modeling frameworks we have investigated. To further investigate this finding, we evaluated the mutual information between the rating values and each of the patient and message features. The results showed a maximum absolute mutual information value of 0.027 for patient features and 0.002 for message features. These values are remarkably low and explain why the features do not result in a significant boost in performance.

6. CONCLUSIONS

In this paper we have described PERSPeCT, a novel collaborative filtering-based computer tailored health communications (CTHC) system targeted at the smoking cessation support domain. Our primary contribution is an evaluation of a wide range of classic and state-of-the art collaborative filtering methods including pure methods, ensemble methods, and hybrid methods in a new and important small-data domain. Due to the success of Bayesian Probabilistic Matrix Factorization (BPMF) as the best single model identified in our evaluation, a pure BPMF-based instance of PERSPeCT is currently being deployed and evaluated relative to our existing expert system through a randomized control trial to determine its effectiveness in an online message selection context. The online evaluation of ensemble methods is reserved for future work and future versions of PERSPeCT will be evaluated based on patient outcomes.

Acknowledgments

This work is supported by the Patient Centered Outcomes Research Institute (1IP2P1000582) and the University of Massachusetts Center for Clinical and Translational Sciences (UL1TR000161). Dr Sadasivam is also funded by a National Cancer Institute Career Development Award (K07CA172677).

7. REFERENCES

[1] K. Ashing-Giwa. Health behavior change models and their socio-cultural relevance for breast cancer screening in african american women. *Women Health*, 28(4):53–71, 1999.

[2] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering, 1999.

[3] M. Jahrer, A. Töscher, and R. Legenstein. Combining predictions for accurate recommender systems. In *Proceedings of the 16th ACM SIGKDD*, pages 693–702. ACM, 2010.

[4] M. Kreuter. *Tailoring health messages : customizing communication with computer technology*. LEA's communication series. L. Erlbaum, Mahwah, N.J., 2000.

[5] B. Marlin, R. Adams, R. Sadasivam, and T. Houston. Towards collaborative filtering recommender systems for tailored health communications. In *Proceedings of the AMIA 2013 Annual Symposium*, pages 1600–1607, 2013.

[6] J. McAuley and J. Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172. ACM, 2013.

[7] R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. pages 880–887. ACM, 2008.

[8] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. *Advances in neural information processing systems*, 20:1257–1264, 2008.

[9] T. Segaran. *Programming collective intelligence : building smart web 2.0 applications*. O'Reilly, Beijing ; Sebastapol CA, 1st edition, 2007.

[10] A. P. Singh and G. J. Gordon. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD*, pages 650–658. ACM, 2008.

[11] A. P. Singh and G. J. Gordon. A bayesian matrix factorization model for relational data. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, 2010.

[12] V. J. Strecher, J. B. McClure, G. L. Alexander, B. Chakraborty, V. N. Nair, J. M. Konkel, S. M. Greene, L. M. Collins, C. C. Carlier, C. J. Wiese, R. J. Little, C. S. Pomerleau, and O. F. Pomerleau. Web-based smoking-cessation programs: results of a randomized trial. *Am J Prev Med*, 34(5):373–81, 2008.