Hierarchical Latent Factors for Preference Data*

Nicola Barbieri[1,2], Giuseppe Manco[2], and Ettore Ritacco[2]

[1] Department of Electronics, Informatics and Systems - University of Calabria, via Bucci 41c, 87036 Rende (CS) - Italy
nbarbieri@deis.unical.it,

[2] Institute for High Performance Computing and Networks (ICAR)
Italian National Research Council
via Bucci 41c, 87036 Rende (CS) - Italy
{barbieri,manco,ritacco}@icar.cnr.it

Abstract. In this work we propose a probabilistic hierarchical generative approach for users' preference data, which is designed to overcome the limitation of current methodologies in Recommender Systems and thus to meet both prediction and recommendation accuracy. The *Bayesian Hierarchical User Community Model (BH-UCM)* focuses both on modeling the popularity of items and the distribution over item ratings. An extensive evaluation over two popular benchmark datasets shows that the combined modeling of item popularity and rating provides a powerful framework both for rating prediction and for the generation of accurate recommendation lists.

1 Introduction

Recommender systems (RS) play an important role in several domains as they provide users with potentially interesting recommendations within catalogs of available information/products/services [11]. Recent studies [5, 6, 9] have shown that the focus on the prediction accuracy (e.g., in terms of root mean square error, RMSE) does not necessarily help in devising high quality recommendations. It has been shown [1] that probabilistic approaches based on latent-factor models allow the most adequate degree of flexibility, as they: *(i)* allow the specification of complex yet easy to interpret latent structures; *(ii)* achieve the highest recommendation accuracy.

In this paper we propose a new Bayesian Hierarchical latent factor model (*Bayesian Hierarchical User Community Model*, *BH-UCM* in the following) which combines the advantages of both hierarchical modeling and item selection, and comparatively investigate both its recommendation accuracy and prediction error. The underlying generative process takes into account both item selection and rating emission, so that those users who experience the same items and tend to adopt the same rating pattern are gathered into communities. Individual users are modeled as a random mixture of communities, where the individual community is characterized again by a mixture of topics modeling both the popularity of items and the distribution over item ratings.

* Extended Abstract

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,

2 Preliminaries and Context

We introduce in this section the notation used throughout the paper along with some preliminary concepts. Let $\mathcal{U} = \{u_1, \ldots, u_M\}$ be a set of M users and $\mathcal{I} = \{i_1, \ldots, i_N\}$ a set of N items. Users' preferences can be represented as a $M \times N$ matrix $\mathbf{R}$, whose generic entry r_i^u denotes the rating value (i.e., the degree of preference) assigned by user u to item i. For each pair $\langle u, i \rangle$, rating value r_i^u falls within a limited integer range $\mathcal{V} = \{0, \ldots, V\}$, where 0 represents an unknown rating and V is the maximum degree of preference. The number of users M as well as the number of items N are very large and, in practical applications, the rating matrix $\mathbf{R}$ is characterized by an exceptional sparseness (e.g., more than 95%), since the individual users tend to rate a limited number of items.

Given an active user u, the goal of a RS is to provide u with a recommendation list of unexperienced items that are expected to be of interest to u. This clearly involves predicting the interest of u into unrated items.

Based on the underlying mathematical model, probabilistic approaches allow the prediction of the expected interest of a user u into an item i in two different ways [8]:

- *Forced prediction*: the probabilistic model provides an estimate of $P(r|u, i)$;
- *Free prediction*: the item selection process is included in the probabilistic model, which is typically based on the estimate of $P(r, i|u)$.

In general, a recommendation list can be generated by selecting a set of candidate items, sorting them according a ranking function which is provided by the RS and selecting the top k items.

3 Bayesian Hierarchical Model for Preference Data

The key idea of the proposed technique is that there exists a set of user communities, each one describing different tastes of users and their corresponding rating patterns. Each user community is then modeled as a random mixture over latent topics, which can be interpreted as item-categories. Given a user u, we can foresee her preferences on a set of items $\mathcal{I}_u$ by choosing an appropriate user community z (from a set $\{1, \ldots, K\}$) and then choosing an item category w (from a set $\{1, \ldots, L\}$) for each item in the list. The choice of the item category w actually depends on the selected user community z. Finally the preference value is generated by considering the preference of users belonging to the group z on items of the category w. This local modeling of items is the main difference in the generative semantics with respect to state-of-the-art LDA based co-clustering approaches [10].

The generative process for the new *BH-UCM* model, whose corresponding graphical scheme is shown in Fig. 1, is as follows:

- $\forall u \in \mathcal{U}$ sample user community-mixture components $\boldsymbol{\vartheta}_u \sim Dirichlet(\boldsymbol{\alpha})$;

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,

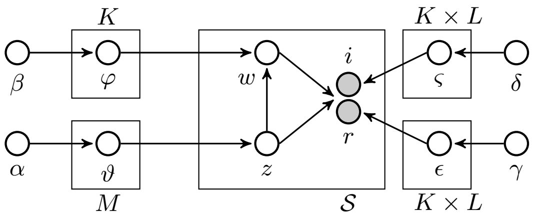


Fig. 1. *BH-UCM* Model

- $\forall z \in \{1, \ldots, K\}$ sample the mixture components $\boldsymbol{\varphi}_z \sim Dirichlet(\boldsymbol{\beta})$
- $\forall w \in \{1, \ldots, L\}, z \in \{1, \cdots, K\}$,
  - Sample item selection components $\boldsymbol{\varsigma}_{z.w} \sim Dirichlet(\boldsymbol{\delta})$
  - Sample rating probabilities $\boldsymbol{\epsilon}_{z,w} \sim Dirichlet(\boldsymbol{\gamma})$
- $\forall u \in \mathcal{U}$
  - Sample the number of items for the user u, $N_u \propto Poisson(\mathcal{K})$
  - For $n = 1$ to N_u
    - Choose a user attitude $z_{u,n} \sim Multinomial(\boldsymbol{\vartheta}_u)$
    - Choose a topic $w_{u,n} \sim Multinomial(\boldsymbol{\varphi}_{z_{u,n}})$
    - Choose an item $i_n \sim Multinomial(\boldsymbol{\varsigma}_{z_{u,n},w_{u,n}})$
    - Generate a rating value for the chosen item according to the distribution $P(r|\boldsymbol{\epsilon}_{z_{u,n},w_{u,n}})$.

Where $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$ and $\mathcal{K}$ are the fixed hyper-parameters of the model.

Unfortunately, the exact inference for this model is however intractable; hence, we propose a Gibbs Sampling parameter estimation, in which at each step inference can be accomplished by exploiting the full conditional $P(z_n = k_n, w_n = l_n|Z_{\neg n}, W_{\neg n}, \mathbf{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$. In the latter, z_n (resp. w_n) is the cell of a matrix Z (resp. W) which corresponds to this observation, and $Z_{\neg n}$ (resp. $W_{\neg n}$) denotes the remaining topic assignments. For the n-th observation we have:

$$P(z_n = k_n, w_n = l_n|Z_{\neg n}, W_{\neg n}, \mathbf{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \propto$$
$$\frac{n_u^{k_n} + \alpha_{k_n} - 1}{\sum_{k'=1}^{K}(n_u^{k'} + \alpha_{k'}) - 1} \cdot \frac{n_{k_n,i}^{l_n} + \beta_{l_n} - 1}{\sum_{l'=1}^{L}(n_{k_n,i}^{l'} + \beta_{l'}) - 1} \cdot$$
$$\frac{n_{r_n}^{k_n,l_n} + \gamma_{r_n} - 1}{\sum_{r=1}^{V}(n_r^{k_n,l_n} + \gamma_r) - 1} \cdot \frac{n_{i_n}^{k_n,l_n} + \delta_{i_n} - 1}{\sum_{i=1}^{N}(n_i^{k_n,l_n} + \delta_i) - 1}$$

The notation used in the Gibbs Sampling derivation is summarized in Tab. 1. Given the state of the Markov chain, denoted my $\mathcal{M} = (\mathbf{R}, Z, W)$, we can obtain the multinomial parameters $\boldsymbol{\varphi}$, $\boldsymbol{\vartheta}$, $\boldsymbol{\epsilon}$ and $\boldsymbol{\varsigma}$ noticing that, by applying Bayes's rule and then by algebraic manipulations and the properties of the Dirichlet distribution. This ultimately yields the following estimations:

$$\vartheta_{u,k} = \frac{n_u^k + \alpha_k}{n_u + \sum_{k=1}^{K} \alpha_k} \qquad \varphi_{k,l} = \frac{n_{k,l} + \beta_l}{n_k + \sum_{l=1}^{L} \beta_l}$$
$$\epsilon_{k,l,r} = \frac{n_r^{k,l} + \gamma_r}{n_{k,l} + \sum_{r'=1}^{V} \gamma_{r'}} \qquad \varsigma_{k,l,i} = \frac{n_i^{k,l} + \delta_i}{n_{k,l} + \sum_{i'=1}^{N} \delta_{i'}}$$

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,

SYMBOL	DESCRIPTION
K	# topics/user communities
L	# item categories
n_u^k	# evaluation of the user u which have been assigned to the user topic k
$n_r^{k,l}$	# times that the rating r has been assigned to each observation when the user topic is k and the item category is l
$n_i^{k,l}$	# times that the item category l has been assigned to observations of the item i when the user topic is k
n_u	# observations for the user u ($\|\mathcal{I}(u)\|$)
n_k	# observations associated with community k
$n_{k,l}$	# times that the category l has been assigned to observations whose user topic is k

Table 1. Summary of notation

Finally, given the pair $\langle u, i\rangle$ we compute the probability of observing the rating value r in a *free prediction* and *forced prediction* context:

$$p(R = r, i|u) = \sum_{k=1}^{K}\sum_{l=1}^{L} \vartheta_{u,k} \cdot \varphi_{k,l} \cdot \epsilon_{k,l,r} \cdot \varsigma_{k,l,i}; \qquad p(R = r|u, i) = \frac{p(R = r, i|u)}{p(i|u)}$$

4 Evaluation

For the experimental evaluation of the proposed model, we use two reference benchmark data sets, namely MovieLens-1M[1] and a sample of Netflix data. Both datasets contain explicit preference data: ratings fall within the range 1 to 5, where the latter denotes the highest preference value.
We compare our model with some state-of-the-art competitors from the Collaborative Filtering literature: (PMF) [12], UCM, HUCM [3], and BUCM [2]; and in particular with a selection of co-clustering approaches: *LDCC* [13], *Bregman-CC* [7] and *Bi-LDA* [10].

All models have been trained by retaining the 1% of the training data as held out to perform *early stopping* and avoid overfitting.

Predictive Accuracy. We start our analysis from the evaluation of the prediction accuracy achieved by the algorithms. Table 2 summarizes the best RMSE obtained on both the considered datasets, together with the associated settings.

BH-UCM outperforms all the other co-clustering approaches on the NetFlix data, and is the runner-up winner after *HUCM* which, however, exhibits a marginal advantage. Minimal differences can also be noticed on MovieLens, where *PMF* (a non co-clustering Probabilistic Matrix Factorization approach) achieves the best RMSE score (as expected).

Recommendation Accuracy. Things change substantially when considering the quality of the recommendation in terms of the precision and recall. Based on the

[1] `http://www.grouplens.org/system/files/million-ml-data.tar\_\_0.gz`

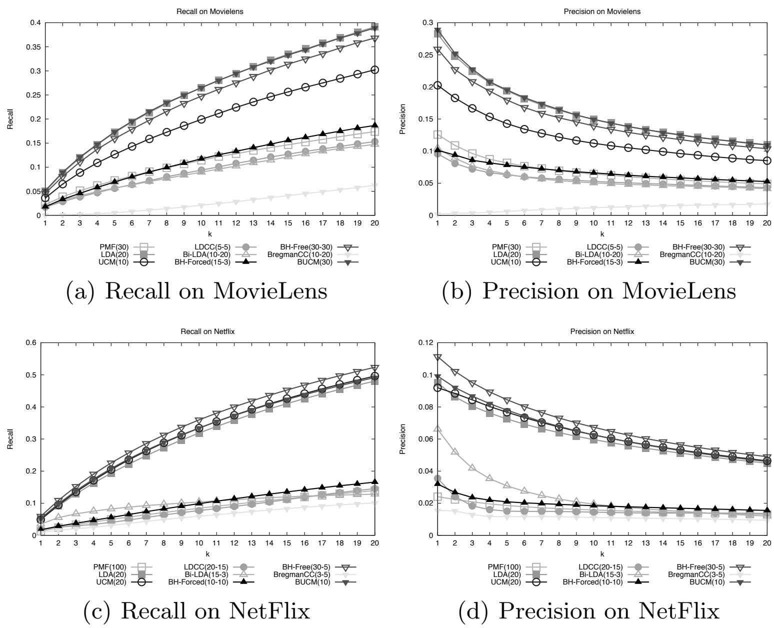
Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,

	MovieLens		Netflix	
Approach	**Best RMSE**	**#Topics**	**Best RMSE**	**#Topics**
PMF	**0.8655**	10	0.9309	100
HUCM	0.9278	2-3	**0.9212**	50-10
Bregman-CC	0.9023	10-20	0.9873	3-5
Bi-LDA	0.9033	30-20	0.9362	30-15
LDCC	0.9074	5-5	0.9419	5-10
BH-UCM	0.9073	30-5	0.9256	30-5
BUCM	0.9292	30	0.9431	10

Table 2. Summary of predictive accuracy over the MovieLens and Netflix datasets

results in [1, 2], we consider here also *LDA* model [4], which has been identified as one of the top-performers from this perspective. Notice that *LDA* was not included in the analysis of predictive accuracy, as it does not explicitly support a way to compute rating prediction.

Figure 2 and shows the results of recommendation accuracy on Movielens and Netflix data, when the size k of the recommendation list varies from 1 to 20. We denote by *BH-Free* and *BH-Forced* two different instantiations of the proposed approach, which focus respectively on free and forced prediction. On Movielens data, *BH-UCM* exhibits a minimal worsening on recall (0.39 vs 0.37) and precision (0.11 vs 0.10), against *LDA* and *BUCM*. Notably, the discrepancy between the recommendation accuracy of Bayesian and non-Bayesian approaches is consistently large. In particular, both *BUCM* and *BH-UCM* outperform *UCM*.



(a) Recall on MovieLens (b) Precision on MovieLens

(c) Recall on NetFlix (d) Precision on NetFlix

Fig. 2. Recommendation Accuracy

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, Copyright (c) 2012 - Edizioni Libreria Progetto and the authors

The outperformance of *BUCM* over *BH-UCM* in Movielens can be explained by the different distribution of these data with respect to Netflix. In this latter case, in fact, the huge volume of data is more likely to exhibit local patterns, which are better modeled by *BH-UCM*. By converse, Movielens exhibits both less users and less ratings, and hence the simpler *BUCM* model can easily fit the data.

5 Conclusion

In this work we proposed a hierarchical Bayesian approach for preference data, which extends state-of-the-art (hierarchical) co-clustering techniques, by modeling dynamic associations and dependencies between user- and item-clusters. An extensive evaluation was performed to assess the skills of the devised model, in terms of both rating prediction and recommendation accuracy, showing that the proposed model is competitive with state-of-the-art approaches w.r.t the studied datasets.

References

1. N. Barbieri and G. Manco, *An Analysis of Probabilistic Methods for Top-N Recommendation in Collaborative Filtering*, in Proc. ECML-PKDD Conf., Athens, Greece, 2011.
2. N. Barbieri, G. Costa, G. Manco and R. Ortale, *Modeling Item Selection and Relevance for Accurate Recommendations: A Bayesian Approach*, in Proc. ACM Conf on Recommendation Systems (RecSys'11), 2011.
3. N. Barbieri, G. Manco and E. Ritacco *A Probabilistic Hierarchical Approach for Pattern Discovery in Collaborative Filtering Data*, in Proc. SDM Conf. , 2011.
4. D. M. Blei, A. Y. Ng and M. I. Jordan, *Latent Dirichlet Allocation*, The Journal of Machine Learning Research, 3 (2003), pp. 993–1022.
5. E. Campochiaro, R. Casatta,P. Cremonesi and R. Turrin *Do Metrics Make Recommender Algorithms?*, in AINA Workshops, 2009.
6. P. Cremonesi, Y. Koren and R. Turrin, *Performance of recommender algorithms on top-n recommendation tasks*, in Proc. ACM Conf on Recommendation Systems (RecSys'10), 2010.
7. T. George and S. Merugu, *A Scalable Collaborative Filtering Framework Based on Co-Clustering*, in ICDM, 2005.
8. T. Hofmann, *Latent semantic models for collaborative filtering*, ACM Transactions on Information Systems (TOIS), 22 (2004), pp. 89–115.
9. S. M. McNee, J. Riedl and J. A. Konstan, *Being Accurate is Not Enough: How Accuracy Metrics have hurt Recommender Systems*, in ACM SIGCHI Conference on Human Factors in Computing Systems, 2006.
10. I. Porteous, E. Bart and M. Welling, *Multi-HDP: a non parametric Bayesian model for tensor factorization*, in AAAI, 2008.
11. F. Ricci, L. Rokach,B. Shapira and P.B. Kantor *Recommender Systems Handbook*, Springer, 2011
12. R. Salakhutdinov and A. Mnih, *Probabilistic Matrix Factorization*, in NIPS, 2008.
13. P. Wang, C. Domeniconi and K. B. Laskey, *Latent Dirichlet Bayesian Co-Clustering*, in ECML-PKDD, 2009.

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,