JMLR: Workshop and Conference Proceedings 25:65–79, 2012 Asian Conference on Machine Learning

A Coupled Indian Buffet Process Model for Collaborative Filtering

Sotirios P. Chatzis SOTERI0S@ME.COM
Department of Electrical Engineering, Computer Engineering and Informatics
Cyprus University of Technology

Editor: Steven C.H. Hoi and Wray Buntine

Abstract

The dramatic rates new digital content becomes available has brought collaborative filtering systems in the epicenter of computer science research in the last decade. In this paper, we propose a novel methodology for rating prediction utilizing concepts from the field of Bayesian nonparametrics. The basic concept that underlies our approach is that each user rates a presented item based on the latent genres of the item and the latent interests of the user. Each item may belong to more than one genre, and each user may belong to more than one latent interest class. The number of existing latent genres and interests are not known beforehand, but should be inferred in a data-driven fashion. We devise a novel hierarchical factor analysis model to formulate our approach under these assumptions. We impose suitable priors over the allocation of items into genres, and users into interests; specifically, we utilize a novel scheme which comprises two coupled Indian buffet process priors that allow the number of latent classes (genres/interests) to be automatically inferred. We experiment on a large set of real ratings data, and show that our approach outperforms four common baselines, including two very competitive state-of-the-art approaches.

Keywords: Collaborative filtering; factor analysis; Indian Buffet Process

1. Introduction

Modern online systems are characterized by a continuous explosion in the availability of content. Ratings-based collaborative filtering (CF) systems, which use past ratings to make predictions, provide a more accurate approach to discovering items of interest (Chen et al., 2009). This thriving subfield of machine learning became popular in the late 1990s with the spread of online services that use recommender systems, such as Amazon, Yahoo! Music, and Netflix. In such systems, each user is associated with a profile which comprises votes or ratings for items in the system derived from the user. Such profiles can be obtained either explicitly from user-provided ratings, or be compiled implicitly by tracking user's behavior. Based on the profile of each user, suitable models are trained that allow for the collaborative filtering system to predict the user ratings for any item available in the system, possibly by also exploiting information from other system users.

In this work, we introduce a novel statistical model for collaborative filtering, based on concepts from the field of Bayesian nonparametrics. Specifically, the formulation of our model is based on the notion that there is a set of latent (unobservable) interests that a modeled user may cluster into, and a set of latent genres that each available item may cluster into. Ratings are considered to be obtained by averaging the joint biases for

the interest/genre combinations corresponding to the considered user and item, and the respective user and item biases. One key-element of our approach is our assumption that one given user may simultaneously belong to more than one latent interest classes, and, similar, one available item may simultaneously belong to more than one latent genres. Consider for example the case of movie rating: one user may be interested in both "martial arts" and "science"; similar, a movie may belong to both the "action" and "science fiction" genres. To reflect this assumption in the formulation of our method, we postulate a novel two-level hierarchical factor analysis model to represent the rating function of the users.

A second key-element of our approach that differentiates it from existing approaches is that we impose a novel coupled Indian Buffet Process (IBP) prior over the latent variables of our model that assign items to genres and users to interest classes. The IBP prior is a nonparametric prior for latent feature models in which observations are influenced by a combination of hidden features, exactly as assumed by the proposed model: observed ratings are influenced by a weighted combination of the latent user interests and item genres, where the weights are the biases pertaining to each interest/genre combination. The advantage of the proposed IBP construction consists in that it provides a principled prior in situations where the number of latent features is unknown, as in the case of our model. Therefore, our postulated ratings model is a nonparametric Bayesian hierarchical factor analysis model, obtained by imposition of a novel IBP-based prior over the model latent variables. This construction allows for making no assumptions regarding the number of existing latent genres or interests but, instead, inferring them from the data.

The remainder of this paper is organized as follows: In Section 2, we briefly review the IBP prior. In Section 3, we introduce our proposed model and derive its inference algorithms using a truncated variational Bayesian algorithm. In Section 4, we experimentally evaluate our approach using a large data set of movie ratings. We compare the performance of our approach to four appropriate baselines, and discuss the results of these experiments. As we show, our approach outperforms the considered state-of-the-art alternatives, while also offering exceptional computational efficiency in terms of real-time performance. Finally, in Section 5 we summarize our results, and conclude this paper.

2. The Indian Buffet Process

In many unsupervised learning problems it is necessary to derive a set of latent variables given a set of observations. A characteristic example is collaborative filtering: considering that users are associated with possibly multiple latent interests, and items are associated with possibly multiple latent genres, we may wish to identify these sets of latent classes and determine which users/items belong to which latent classes. Unfortunately, most traditional machine learning approaches require the number of latent classes (features) as an input. In such cases, usually one has to resort to application of a model selection technique to come up with a trade-off between model complexity and model fit.

A solution to this problem is offered in the context of Bayesian nonparametrics. Nonparametric Bayesian approaches treat the number of latent features as a random quantity to be determined as part of the posterior inference procedure. The most common nonparametric prior for latent feature models is the Indian Buffet Process (Griffiths and Ghahramani, 2005). The IBP is a prior on infinite binary matrices that allows us to simultaneously infer

which features influence a set of observations and how many features there are. The form of the prior ensures that only a finite number of features will be present in any finite set of observations, but more features may appear as more observations are received.

Let us consider a set of N objects that may be assigned to a total of K features. Let $\boldsymbol{Z} = [z_{nk}]_{n,k=1}^{N,K}$ be a $N \times K$ matrix of assignment variables, with $z_{nk} = 1$ if the nth object is assigned to the kth feature (multiple z_{nk}'s may be equal to 1 for a given object n), $z_{nk} = 0$ otherwise. The IBP imposes a prior over $[\boldsymbol{Z}]$, a canonical form of $\boldsymbol{Z}$ that is invariant to the ordering of the features (Griffiths and Ghahramani, 2005). The imposed prior takes the form

$$\begin{aligned} p([\boldsymbol{Z}]) = & \frac{\alpha^K}{\prod_{h \in \{0,1\}^N \setminus \{0\}^N} K_h!} \exp\{-\alpha H_N\} \\ & \times \prod_{k=1}^{K} \frac{(N - m_k)!(m_k - 1)!}{N!} \end{aligned} \tag{1}$$

Here, m_k is the number of objects assigned to the kth feature, i.e. with $z_{nk} = 1$, α is the innovation hyperparameter of the IBP prior which regulates the number of effective model features K, H_N is the Nth harmonic number, and K_h is the number of occurrences of the non-zero binary vector h among the columns in $\boldsymbol{Z}$.

Apart from MCMC, mean-field variational Bayesian inference methods, which approximate the true posterior via a simpler distribution, provide a deterministic alternative to sampling-based inference for the IBP (Jordan et al., 1998; Chatzis et al., 2008). Variational Bayesian inference for the IBP is based on an alternative formulation of $p(\boldsymbol{Z})$, namely the *stick-breaking construction* of the IBP (Teh et al., 2007): It has been shown that the prior (1) obtained by the IBP can be equivalently expressed under the following hierarchical Bayesian construction

$$z_{nk} \sim \text{Bernoulli}(\pi_k) \tag{2}$$

$$\pi_k = \prod_{i=1}^{k} v_i \tag{3}$$

$$v_k \sim \text{Beta}(\alpha, 1) \tag{4}$$

In other words, under the stick-breaking construction, an equivalent hierarchical expression for the prior $p(\boldsymbol{Z})$ is obtained by introduction of the Beta-distributed stick variables v_k.

3. Proposed Approach

Before we introduce our model, let us first formally define the problem we are aiming to address: Let us consider a recommender system the registered users of which comprise a set $\mathscr{N} = \{n_u\}_{u=1}^{U}$. Let us also consider that the items registered with the system comprise the set $\mathscr{C} = \{c_m\}_{m=1}^{M}$. Each user is allowed by the system to provide a rating for each item; the rating variable r is assumed to take values in the discrete set $\{1, \ldots, R\}$ of possible ratings. Each user may provide ratings for one or more of the items registered with the system. In essence, at any time point, the considered recommender system has available a training ratings data set $\mathscr{D} = \{(r_d, u_d, m_d)\}_{d=1}^{D}$ comprising D tuples of users with indexes $u_d \in \{1, \ldots, U\}$, items with indexes $m_d \in \{1, \ldots, M\}$, and associated ratings $r_d \in \{1, \ldots, R\}$.

Typically, users provide ratings only for a minuscule fraction of the items registered with a real-life recommender system. Based on this observation, what we aim to achieve is to postulate a suitable statistical model that allows for predicting the rating r each user would assign to each item registered with the system, given the few ratings in the training data set $\mathscr{D}$. To effect this goal, we additionally make the assumption that each user may have one or more interests which comprise latent variables of the sought model of unknown number, while each item may belong to one or more genres which also comprise latent variables of the sought model of unknown number. The ratings users assign to the available items are considered functions of the latent classes (interests/genres) of the users and the available items.

To address the considered problem, we postulate an intricate factor analysis model to represent the rating function of the users of a recommender system. The postulated model assumes that the rating function is the sum of the user and item bias, plus a hierarchical two-level factor analysis term, with the latent factors expressing the assignment of the users into interest classes and of the available items into genres, and the factor weights (loadings) taken as the joint user/item biases of the model, which can be obtained through model training.

Specifically, we postulate the following statistical model for the value of the rating function r given a user $n_u \in \mathscr{N}$ and item $c_m \in \mathscr{C}$:

$$r(u,m) = \mu + \rho_u + \eta_m + \sum_{i=1}^{\infty} z_{ui} \left(\boldsymbol{w}_i^T \boldsymbol{x}_m \right) + \epsilon \tag{5}$$

In the postulated model (5), μ is the mean rating for any user/item pair, which can be initialized as the mean of the ratings in the training data set $\mathscr{D}$, ρ_u is the bias of the uth user, η_m is the bias of the mth item, $\boldsymbol{x}_m = [x_{mg}]_{g=1}^{\infty}$, the x_{mg} are the latent variables of genre assignment, with $x_{mg} = 1$ if the mth item is assigned to the gth genre (multiple genre assignments are possible for each item), $x_{mg} = 0$ otherwise, the z_{ui} are the latent variables of interest assignment, with $z_{ui} = 1$ is the uth user is assigned to the ith interest (multiple interest assignments are possible for each user), $z_{ui} = 0$ otherwise, and the weights $\boldsymbol{w}_i = [w_{ig}]_{g=1}^{\infty}$ are the joint user/item biases that correspond to the combination of the ith latent user interest with the mth latent item genre. The term ϵ in (5) stands for an additive white noise term, with

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{6}$$

Indeed, this selection for the prior of the noise variable ϵ might seem inappropriate, since the observed ranking variables $r(u,m)$ usually take on a set of discrete values. However, the assumption of an additive white noise is not too unrealistic; in addition, it has the major advantage of allowing for a tractable model inference procedure, with simple model update expressions. As such, in the context of our work, we opt for such a simple noise prior selection.

Note that in our model we assume a positive bias associated with "compatible"' interests and genres, but also a possibly negative bias for "conflicting" interests and genres. We also note that, in the above definition (5) of the proposed model, we have considered an infinite number of latent features (classes), i.e. latent interest classes and latent item genres. This assumption reflects our unawareness of the exact number of existing latent features, and

demands imposition of an appropriate prior distribution over the latent variables $\boldsymbol{x}_m$ and $\boldsymbol{z}_u = [z_{ui}]_{i=1}^{\infty}$, so as to conduct Bayesian inference over the number of latent features. Following the discussions of Section 2, to effect this kind of inference over the number of latent features in the model, we need to seek an appropriate prior in the context of Bayesian nonparametrics: we adopt the IBP prior, which, as previously discussed, is the most common nonparametric prior for latent feature models. Specifically, formulation of our model is effected by employing the stick-breaking construction of the IBP, which allows for deriving a computationally efficient and scalable inference algorithm for our model under the variational Bayesian paradigm. We postulate

$$z_{ui} \sim \text{Bernoulli}(\pi_i), \quad i = 1, \ldots, \infty \tag{7}$$

$$\pi_i = \prod_{\lambda=1}^{i} v_\lambda \tag{8}$$

$$v_\lambda \sim \text{Beta}(\alpha, 1) \tag{9}$$

and

$$x_{mg} \sim \text{Bernoulli}(\varpi_g), \quad g = 1, \ldots, \infty \tag{10}$$

$$\varpi_g = \prod_{\lambda=1}^{g} \tilde{v}_\lambda \tag{11}$$

$$\tilde{v}_\lambda \sim \text{Beta}(\gamma, 1) \tag{12}$$

We coin the obtained model the Indian Buffet Process Hierarchical Factor Analysis (IBP-HFA) model for collaborative filtering.

3.1. Variational Bayesian Inference

Our variational Bayesian inference formalism for the IBP-HFA model consists in derivation of a family of variational posterior distributions $q(.)$ which approximate the true posterior distribution over the infinite sets $\boldsymbol{v} = [v_\lambda]_{\lambda=1}^{\infty}$, $\tilde{\boldsymbol{v}} = [\tilde{v}_\lambda]_{\lambda=1}^{\infty}$, and $\boldsymbol{W} = [\boldsymbol{w}_i]_{i=1}^{\infty}$. Apparently, under this infinite dimensional setting, Bayesian inference is not tractable. For this reason, we employ a common strategy in the literature of Bayesian nonparametrics, formulated on the basis of a truncated stick-breaking representation of the IBP (Doshi-Velez et al., 2009). That is, we fix a value I letting the variational posterior over the v_i have the property $q(v_{I+1} = 0) = 1$, and, similar, we fix a value G letting the variational posterior over the $\tilde{v}_g$ have the property $q(\tilde{v}_{G+1} = 0) = 1$. In other words, we set the π_i and ϖ_g equal to zero for $i > I$ and $g > G$, respectively. Note that, under this setting, the treated IBP-HFA model involves two full IBP priors; truncation is not imposed on the model itself, but only on the variational distribution to allow for a tractable inference procedure. Hence, the truncation levels I and G are variational parameters which can be freely set, and not part of the prior model specification.

To conduct variational Bayesian inference for our model, we also impose a prior distribution over the model biases and mean ranking. Specifically, we consider

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_\mu^2) \tag{13}$$

$$p(\rho_u) = \mathcal{N}(\rho_u|0, \varsigma_u^2) \tag{14}$$

$$p(\eta_m) = \mathcal{N}(\eta_m|0, e_m^2) \tag{15}$$

$$p(\boldsymbol{w}_i) = \mathcal{N}(\boldsymbol{w}_i|0, s_i^2\boldsymbol{I}) \tag{16}$$

Let $\Theta = \{\boldsymbol{v}, \tilde{\boldsymbol{v}}, \boldsymbol{W}, \mu, \{\rho_u, \boldsymbol{z}_u\}_{u=1}^U, \{\eta_m, \boldsymbol{x}_m\}_{m=1}^M\}$ be the set of hidden variables and unknown parameters of the IBP-HFA model, and Ξ be the set of the hyperparameters of the imposed priors, $\Xi = \{\alpha, \gamma, \mu_0, \sigma_\mu^2, \{\varsigma_u^2\}_{u=1}^U, \{e_m^2\}_{m=1}^M, \{s_i^2\}_{i=1}^I\}$. Variational Bayesian inference consists in the introduction of an arbitrary distribution $q(\Theta)$ to approximate the actual posterior $p(\Theta|\Xi, \mathscr{D})$, which is computationally intractable (Bishop, 2006). The variational posterior $q(\Theta)$ is obtained by maximization of the variational free energy (Chatzis et al., 2008)

$$\mathcal{L}(q) = \int \mathrm{d}\Theta q(\Theta) \log \frac{p(\mathscr{D}, \Theta|\Xi)}{q(\Theta)} \tag{17}$$

which comprises a lower bound of the log marginal likelihood (log evidence), $\log p(\mathscr{D})$, of the model (Jordan et al., 1998).

Due to the considered conjugate prior configuration of the IBP-HFA model, the variational posterior $q(\Theta)$ is expected to take the same functional form as the prior, $p(\Theta)$ (Chatzis et al., 2008). Derivation of the variational posterior distribution $q(\Theta)$ involves maximization of the variational free energy $\mathcal{L}(q)$ over each one of the factors of $q(\Theta)$ in turn, holding the others fixed, in an iterative manner (Bishop, 2006). By construction, this iterative, consecutive updating of the variational posterior distribution is guaranteed to monotonically and maximally increase the free energy $\mathcal{L}(q)$ (Chatzis et al., 2008). Let us denote as $\langle . \rangle$ the posterior expectation of a quantity. We yield the following posteriors:

1. Regarding the stick-breaking variables of the employed IBPs, the posterior distributions are similar to the one derived in Doshi-Velez et al. (2009). We have

$$q(v_\lambda) = \mathrm{Beta}(v_\lambda|\tau_{\lambda 1}, \tau_{\lambda 2}) \tag{18}$$

where

$$\begin{aligned} \tau_{\lambda 1} =& \alpha + \sum_{k=\lambda}^{I} \sum_{u=1}^{U} q(z_{uk} = 1) \\ &+ \sum_{k=\lambda+1}^{I} \left[U - \sum_{u=1}^{U} q(z_{uk} = 1) \right] \left[\sum_{h=\lambda+1}^{k} q_h \right] \end{aligned} \tag{19}$$

$$\tau_{\lambda 2} = 1 + \sum_{k=\lambda}^{I} \left[U - \sum_{u=1}^{U} q(z_{uk} = 1) \right] q_\lambda \tag{20}$$

and

$$q(\tilde{v}_\lambda) = \mathrm{Beta}(\tilde{v}_\lambda|\tilde{\tau}_{\lambda 1}, \tilde{\tau}_{\lambda 2}) \tag{21}$$

where

$$\begin{aligned}\tilde{\tau}_{\lambda 1} =& \gamma + \sum_{k=\lambda}^{G} \sum_{m=1}^{M} q(x_{mk} = 1) \\ &+ \sum_{k=\lambda+1}^{G} \left[M - \sum_{m=1}^{M} q(x_{mk} = 1) \right] \left[\sum_{h=\lambda+1}^{k} \tilde{q}_h \right]\end{aligned} \tag{22}$$

$$\tilde{\tau}_{\lambda 2} = 1 + \sum_{k=\lambda}^{G} \left[M - \sum_{m=1}^{M} q(x_{mk} = 1) \right] \tilde{q}_\lambda \tag{23}$$

and, following (Teh et al., 2007), the quantities q_λ and $\tilde{q}_\lambda$ are defined as

$$q_\lambda \propto \exp \left[\psi(\tau_{\lambda 2}) + \sum_{h=1}^{\lambda-1} \psi(\tau_{h1}) - \sum_{h=1}^{\lambda} \psi(\tau_{h1} + \tau_{h2}) \right] \tag{24}$$

$$\tilde{q}_\lambda \propto \exp \left[\psi(\tilde{\tau}_{\lambda 2}) + \sum_{h=1}^{\lambda-1} \psi(\tilde{\tau}_{h1}) - \sum_{h=1}^{\lambda} \psi(\tilde{\tau}_{h1} + \tilde{\tau}_{h2}) \right] \tag{25}$$

and are normalized so that the they sum to one.

2. Regarding the posterior distributions over the latent variables of interest assignment, z_{ui}, optimization of $\mathcal{L}(q)$ yields

$$q(z_{ui} = 1) = \frac{1}{1 + \exp(-\nu_{ui})} \tag{26}$$

where

$$\begin{aligned}\nu_{ui} =& \sum_{\lambda=1}^{i} [\psi(\tau_{\lambda 1}) - \psi(\tau_{\lambda 1} + \tau_{\lambda 2})] - \left\langle \log\Big(1 - \prod_{\lambda=1}^{i} v_\lambda\Big) \right\rangle \\ &- \frac{1}{2\sigma^2} \sum_{d \in \mathscr{D}(u)} \Bigg[\left\langle \boldsymbol{w}_i^T \boldsymbol{x}_{m(d)} \boldsymbol{x}_{m(d)}^T \boldsymbol{w}_i \right\rangle \\ &- 2 \left\langle \boldsymbol{w}_i^T \boldsymbol{x}_{m(d)} \right\rangle \left(r_d - \langle \mu \rangle - \langle \rho_u \rangle - \left\langle \eta_{m(d)} \right\rangle \right) \\ &+ 2 \sum_{l \neq i} q(z_{ul} = 1) \left\langle \boldsymbol{w}_l^T \boldsymbol{x}_{m(d)} \boldsymbol{x}_{m(d)}^T \boldsymbol{w}_i \right\rangle \Bigg]\end{aligned} \tag{27}$$

$m(d)$ is the identifier of the item that corresponds to the dth training example, $m(d) \in \{1, \ldots, M\}$, and $\mathscr{D}(u)$ is the subset of the training data set $\mathscr{D}$ that contains examples pertaining to the uth user.

3. Regarding the posterior distributions over the latent variables of genre assignment, x_{mg}, optimization of $\mathcal{L}(q)$ yields

$$q(x_{mg} = 1) = \frac{1}{1 + \exp(-\tilde{\nu}_{mg})} \tag{28}$$

where

$$\begin{aligned}\tilde{\nu}_{mg} = &\sum_{\lambda=1}^{g}\left[\psi(\tilde{\tau}_{\lambda 1})-\psi(\tilde{\tau}_{\lambda 1}+\tilde{\tau}_{\lambda 2})\right]-\left\langle\log\big(1-\prod_{\lambda=1}^{g}\tilde{v}_{\lambda}\big)\right\rangle\\ &-\frac{1}{2\sigma^2}\sum_{d\in\tilde{\mathscr{D}}(m)}\Bigg[\left\langle \boldsymbol{w}_g^{*T}\boldsymbol{z}_{u(d)}\boldsymbol{z}_{u(d)}^{T}\boldsymbol{w}_g^{*}\right\rangle\\ &-2\left\langle \boldsymbol{w}_g^{*T}\boldsymbol{z}_{u(d)}\right\rangle\left(r_d-\langle\mu\rangle-\left\langle\rho_{u(d)}\right\rangle-\langle\eta_m\rangle\right)\\ &+2\sum_{l\neq g}q(x_{ml}=1)\left\langle \boldsymbol{w}_l^{*T}\boldsymbol{z}_{u(d)}\boldsymbol{z}_{u(d)}^{T}\boldsymbol{w}_g^{*}\right\rangle\Bigg]\end{aligned} \tag{29}$$

$u(d)$ is the identifier of the user that corresponds to the dth training example, $u(d)\in\{1,\ldots,U\}$, $\tilde{\mathscr{D}}(m)$ is the subset of the training data set $\mathscr{D}$ that contains examples pertaining to the mth item, and we define $\boldsymbol{w}_g^{*}=[w_{ig}]_{i=1}^{I}$.

4. Regarding the joint biases, assuming a spherical posterior distribution, we have

$$q(\boldsymbol{w}_i)=\prod_{g=1}^{G}q(w_{ig})=\prod_{g=1}^{G}\mathcal{N}(w_{ig}|\hat{\varphi}_{ig},\phi_{ig}^2) \tag{30}$$

where

$$\phi_{ig}^2=\left[\frac{1}{s_i^2}+\frac{1}{\sigma^2}\sum_{d=1}^{D}q(z_{u(d),i}=1)q(x_{m(d),g}=1)\right]^{-1} \tag{31}$$

and

$$\begin{aligned}\hat{\varphi}_{ig}=&\frac{\phi_{ig}^2}{\sigma^2}\sum_{d=1}^{D}q(x_{m(d),g}=1)q(z_{u(d),i}=1)\\ &\times\Bigg[r_d-\langle\mu\rangle-\left\langle\rho_{u(d)}\right\rangle-\left\langle\eta_{m(d)}\right\rangle\\ &-\sum_{l\neq i}\sum_{\xi\neq g}q(z_{u(d),l}=1)q(x_{m(d),\xi}=1)\left\langle w_{l\xi}\right\rangle\Bigg]\end{aligned} \tag{32}$$

5. Regarding the user biases, we have

$$q(\rho_u)=\mathcal{N}(\rho_u|\hat{\rho}_u,\hat{\varsigma}_u^2) \tag{33}$$

where

$$\hat{\varsigma}_u^2=\left[\frac{1}{\varsigma_u^2}+\frac{\#\mathscr{D}(u)}{\sigma^2}\right]^{-1} \tag{34}$$

$\#\mathscr{D}(u)$ is the cardinality of $\mathscr{D}(u)$, and

$$\hat{\rho}_u=\frac{\hat{\varsigma}_u^2}{\sigma^2}\sum_{d\in\mathscr{D}(u)}\left[r_d-\langle\mu\rangle-\left\langle\eta_{m(d)}\right\rangle-\sum_{i=1}^{I}q(z_{ui}=1)\left\langle \boldsymbol{w}_i^{T}\boldsymbol{x}_{m(d)}\right\rangle\right] \tag{35}$$

6. Regarding the item biases, we have

$$q(\eta_m) = \mathcal{N}(\eta_m|\hat{\eta}_m, \hat{e}_m^2) \tag{36}$$

where

$$\hat{e}_m^2 = \left[\frac{1}{e_m^2} + \frac{\#\tilde{\mathscr{D}}(m)}{\sigma^2}\right]^{-1} \tag{37}$$

$\#\tilde{\mathscr{D}}(m)$ is the cardinality of $\tilde{\mathscr{D}}(m)$, and

$$\hat{\eta}_m = \frac{\hat{e}_m^2}{\sigma^2} \sum_{d \in \tilde{\mathscr{D}}(m)} \left[r_d - \langle \mu \rangle - \left\langle \rho_{u(d)} \right\rangle - \sum_{i=1}^{I} q(z_{u(d)i} = 1) \left\langle \boldsymbol{w}_i^T \boldsymbol{x}_m \right\rangle\right] \tag{38}$$

7. The mean ratings μ yield

$$q(\mu) = \mathcal{N}(\mu|\hat{\mu}, \hat{\sigma}_\mu^2) \tag{39}$$

where

$$\hat{\sigma}_\mu^2 = \left[\frac{1}{\sigma_\mu^2} + \frac{D}{\sigma^2}\right]^{-1} \tag{40}$$

and

$$\hat{\mu} = \hat{\sigma}_\mu^2 \left[\frac{1}{\sigma^2} \sum_{d=1}^{D} \left(r_d - \left\langle \rho_{u(d)} \right\rangle - \left\langle \eta_{m(d)} \right\rangle - \sum_{i=1}^{I} q(z_{u(d)i} = 1) \left\langle \boldsymbol{w}_i^T \boldsymbol{x}_{m(d)} \right\rangle\right) + \frac{\mu_0}{\sigma_\mu^2}\right] \tag{41}$$

Finally, given a trained IBP-HFA model, *rating prediction* for any pair of user n_u and item c_m, registered with the system, can be conducted by computing the posterior expectation of the rating function $r(u, m)$. We have

$$\begin{aligned} \hat{r}(u,m) &= \langle \mu \rangle + \langle \rho_u \rangle + \langle \eta_m \rangle + \sum_{i=1}^{\infty} q(z_{ui} = 1) \left\langle \boldsymbol{w}_i^T \right\rangle \langle \boldsymbol{x}_m \rangle \\ &= \hat{\mu} + \hat{\rho}_u + \hat{\eta}_m + \sum_{i=1}^{\infty} q(z_{ui} = 1) \left\langle \boldsymbol{w}_i^T \right\rangle \langle \boldsymbol{x}_m \rangle \end{aligned} \tag{42}$$

where

$$\langle \boldsymbol{x}_m \rangle = [q(x_{mg} = 1)]_{g=1}^{G}$$

3.2. Related Work

Recently, several researchers have considered application of Bayesian nonparametrics to collaborative filtering. One of the earliest works in the field was presented in Meeds et al. (2007). The basic difference between that work and our work is that in the work of Meeds et al. (2007), the postulated prior over the model latent variables is a Beta-Bernoulli model, and a nonparametric nature for this model was induced by taking the limit for $K \to \infty$ features in the context of its inference algorithm. Another major disadvantage of the method of Meeds et al. (2007) is that it considers only one latent model variable, thus it does not

allow for modeling both genres and interests in a joint fashion. Similar is the drawback of the method of Zhou et al. (2010), where IBP priors are used similar to our method, as well as of the work of Ding et al. (2010), where a nonparametric Bayesian matrix factorization algorithm is obtained by using a suitable prior, inspired in part by the IBP prior: in both these works, only one latent variable is considered, thus not allowing for jointly inferring genre and interest latent variables.

4. Experiments

To evaluate the proposed approach, we performed a number of experiments using a large sample of a ranking data set available from the GroupLens website[1]. This data set consists of 10 million ratings for 10,677 movies made by 69,878 users of the MovieLens movie rating web site[2]. The data set used in our experiments consists of almost one million samples and comprises data from randomly selected users that have rated at least 10 movies each. The ratings in the used data set are all given on a scale of 0.5 to 5 stars with increments of 0.5 stars. To be capable of assessing the statistical significance of our findings, in our experimental evaluations we employed 5-fold cross validation, using in each fold a random 20% of the available sample for training and the rest for testing.

We use two metrics to assess the performance of our algorithm, namely the root mean square error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{D}\sum_{d=1}^{D}(r_d - \hat{r}_d)^2} \tag{43}$$

a performance metric that penalizes large errors much more than small errors, and the mean absolute error (MAE), defined as

$$\text{MAE} = \frac{1}{D}\sum_{d=1}^{D}|r_d - \hat{r}_d| \tag{44}$$

which penalizes all errors equally, relative to their size. We compare the performance of our method to four baselines, including two state-of-the-art approaches, namely a mean rating method which returns the average rating as an estimate for all user/item pairs (*mean-rating*), a nearest-neighbor method using the Pearson correlation coefficient as the similarity metric, combined with case amplification and significance weighting (*nearest-neighbor*) (Breese et al., 1998), a regularized gradient descent *SVD* method with user and item biases (Paterek, 2007), and the model-based *BLITR* method (Harvey et al., 2011), which postulates a statistical model inspired by principles employed in LDA. In our evaluations, the settings of the BLITR model were adopted from Harvey et al. (2011); a Rao-Blackwellised Gibbs sampler was used to draw 300 samples from the Markov chain, with a burn-in of 200 samples. For the SVD method, we used a learning rate equal to 0.002, and the 2 regularization constants were set to 0.02 and 0.05, respectively. A neighborhood comprising 100 users was used for the *nearest-neighbor* method. Our source codes were developed in MATLAB R2011b, and were executed on a Macintosh platform running OS X 10.7.3.

1. http://grouplens.org/node/73
2. http://www.movielens.org/

Table 1: Average performance (and statistical significance of differences from IBP-HFA) for the evaluated methods. For latent feature/variable models the (maximum) number of latent features is set to 50.

Model	RMSE	p-value	MAE	p-value
mean-rating	1.0643	1.7×10^{-12}	0.8623	1.7×10^{-12}
nearest-neighbor	0.8449	8.2×10^{-6}	0.6694	8.2×10^{-6}
SVD	0.8387	6.05×10^{-5}	0.6476	6.05×10^{-5}
BLITR	0.8202	2.78×10^{-5}	0.6341	2.78×10^{-5}
IBP-HFA	0.8015		0.5978	

4.1. Average performance results

In Table 1, we provide the average performance figures yielded by our method and its considered alternatives. These results comprise the average model performance obtained over the conducted 5 folds of cross-validation. Additionally, apart from average performance, we also provide the p-values obtained from the Student's-t hypothesis test applied on the pairs of performances of the proposed method and each one of the evaluated alternatives. The Student's-t test allows to assess the statistical significance of the performance difference between two evaluated methods, given a set of performance measurements. Generated p-values of the Student's-t test below 0.05 strongly indicate that the means of the obtained performance statistics of the two methods provide a very good assessment of their actual performance difference. Note that the results depicted in Table 1 were obtained by considering that the number of latent features is equal to 50, wherever applicable (this being the truncation threshold for our method).

As we observe, all methods completely outperform the baseline *mean-rating* method, with the second worst performing method being, quite unsurprisingly, nearest-neighbor ranking. We also observe that the proposed approach outperforms by a large margin the evaluated state-of-the-art SVD method, while a statistically significant advantage for our method is also noticed over the recently proposed, model-based BLITR approach. This latter result is quite significant, as it shows that the assumption of a crisp assignment of users to interests and movies to genres, being quite oversimplistic, does actually have a negative impact on the performance of a real-life movie recommendation system. Additionally, it proves the efficacy and effectiveness of the IBP prior, a method the popularity of which has been relatively confined into the realms of the small Bayesian nonparametrics community, in real-life applications dealing with large-scale data sets.

4.2. Further investigation: Error variance

RMSE and MAE are two standard metrics for quantifying how well a prediction algorithm works. As already discussed, they both provide a quantification of the average error accrued over a number of test cases. However, in real-life recommender systems, apart from the average error, another also significant quality aspect that determines system attractiveness to the average user is error variance. In other words, it is crucial that errors, whenever they (inevitably) happen, are not too large to make system performance seem way too poor.

Table 2: Minimum and maximum absolute error and its standard deviation for the evaluated methods (50 latent features at maximum).

Model	MIN Error	MAX Error	Standard Deviation
SVD	0	3.5	0.73
BLITR	0	2.0	0.59
IBP-HFA	0	2.0	0.51

Table 3: Average RMSE of the evaluated methods for users with less than 20, 50, or 100 available training samples.

Model	≤ 20	≤ 50	≤ 100	All
SVD	0.9366	0.8895	0.8713	0.8387
BLITR	0.8611	0.8468	0.8355	0.8202
IBP-HFA	0.8318	0.8202	0.8041	0.8015

Indeed, even seldom cases of too poor a result being obtained may irrevocably harm user confidence in the system. Table 2 shows the minimum and the maximum absolute prediction error obtained by the SVD, BLITR, and IBP-HFA methods, as well as its standard deviation. As we observe, the model-based approaches, with the IBP-HFA being the better performing of the two, are much less likely to yield completely irrelevant results that could hamper system attractiveness to the average user.

4.3. Effect of the availability of training samples on model performance

We now examine how the performance of our model depends on the availability of training examples. In principle, it is expected that with only few training examples, prediction for a given user or item should become extremely problematic. In real-life systems, this is generally the case for users or items only recently registered with the system. To make these investigations, we proceeded as follows: After carefully analyzing the data set, we found that 11.9% of all users have 20 or fewer ratings and nearly half (48.4%) have 50 or fewer. Presumably, these users should be the most difficult for the system to accurately predict their ratings. To verify this assumption, we recalculate the average error statistics obtained in Section 4.1, taking into account the errors obtained only for users with less than 20, 50, or 100 available training examples; our results are depicted in Table 3. We observe that our approach is more resilient compared to its alternatives, especially the SVD method, which experiences a noticeable performance slump as the number of training samples decreases considerably.

4.4. Effect of the truncation threshold on model performance

Further, we examine how the performance of the IBP-HFA model is related to the employed truncation threshold of the number of existing latent features (interests/genres). For this purpose, we repeat the experiments of Section 4.1 for various truncation threshold values,

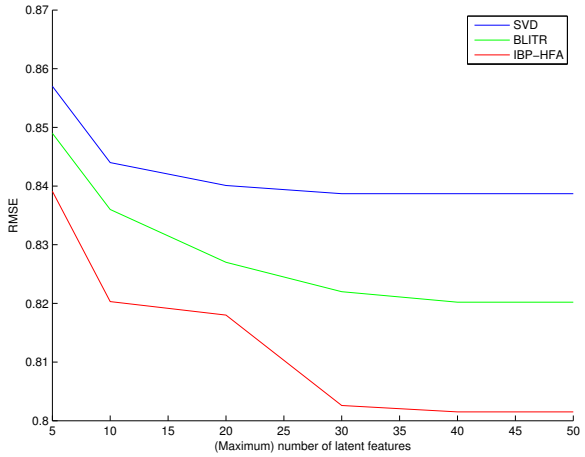

Figure 1: Obtained RMSE as a function of the (maximum) number of latent features (interests/genres).

and we illustrate how model performance changes in Fig. 1. In the same figure, we also illustrate how BLITR and SVD performance changes with the number of latent features varying similar to the truncation thresholds of IBP-HFA. As we observe, the number of latent features has a critical impact on the obtained model performance: allowing only a small number of latent features clearly undermines the performance of all approaches. Increasing the number of latent features yields a rapid increase in model performance, which is more apparent in the case of the proposed approach. However, we also observe that, as the number of latent features increases above a certain level, performance gains seem to stall for all approaches.

4.5. Computational complexity

Finally, let us briefly discuss the computational complexity of our approach, and how it compares to the competition. Regarding prediction generation, the computational costs of our approach are similar to the costs of BLITR for the same number of latent features, since both approaches simply compute a sum of weights (biases). Our method also imposes similar memory requirements compared to BLITR, since the vast majority of the postulated variables are common to both methods. Note also that the computational costs of our method are also similar to the costs of SVD, while orders of magnitude lower than memory-based methods, namely nearest-neighbor ranking, which also impose way higher memory requirements compared to the model-based approaches.

Concerning training costs, despite the fact that our method employs two IBP priors while BLITR adopts a much simpler construction, we have noticed that our model requires almost half the time required by BLITR. This can be expected, as BLITR uses Gibbs sampling, while our approach is based on a much more efficient truncated variational Bayesian expectation-maximization algorithm.

To conclude, both the computational costs and the memory requirements of our method are comparable to existing state-of-the-art approaches, thus making it a viable alternative for real-life recommender systems.

5. Conclusions

In this work, we presented a model-based method for collaborative rating prediction using principles from Bayesian nonparametrics, based on the introduction of a set of latent variables representing user interest and item genre. Even though such an approach is rather common in modern model-based collaborative rating systems, our method is fundamentally different from existing techniques, as it is based on postulation of a novel two-level hierarchical factor analysis model that approximates the ranking function of the system users. Our approach is facilitated by imposing two coupled Indian Buffet Process priors over the variables assigning users to interests and items to genres, which allows for doing inference over the appropriate number of latent features (genres/interests).

We evaluated our approach using a large, freely-available and commonly used data set of real item (movie) ratings. We compared the performance of our approach to four baselines, including two extremely competitive modern methodologies. As we have shown, our model manages to outperform the considered alternatives. Additionally, to demonstrate the robustness of our approach, we investigated how well our model would perform in cases of training with very sparse training data sets (only few item ratings per user). Indeed, the need of training with extremely sparse data sets occurs with very high frequency in real-life systems. As such, the capacity to perform adequately under such a setup is one of the most significant merits of an item rating algorithm. As we showed, our model managed to handle this adverse experimental setup more effectively compared to the alternatives. Finally, we showed that our model possesses the necessary scalability merits real-life systems require.

One issue our model does not address concerns the well-known new user problem i.e., how an established collaborative filtering system based on our algorithm can generate dependable recommendations for users with only few rankings available. In this case, better recommendations may be achieved by considering similarities between users based on additional user features. That is, we may augment our algorithm with some extra observation likelihood, related to features of the system users (e.g., educational level, profession, marital status, and so on). These investigations comprise part of our ongoing research.

References

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proc. of the14th Annual Conference on Uncertainty in Artificial Intelligence*, pages 43–52, 1998.

S. Chatzis, D. Kosmopoulos, and T. Varvarigou. Signal modeling and classification using a robust latent space model based on t distributions. *IEEE Trans. Signal Processing*, 56 (3):949–963, March 2008.

Wen Y. Chen, Jon C. Chu, Junyi Luan, Hongjie Bai, Yi Wang, and Edward Y. Chang. Collaborative filtering for Orkut communities: discovery of user latent behavior. In *Proc. of the 18th international conference on World wide web*, pages 681–690, 2009.

Nan Ding, Yuan (Alan) Qi, Rongjing Xiang, Ian Molloy, and Ninghui Li. Nonparametric bayesian matrix factorization by power-ep. *Journal of Machine Learning Research - Proceedings Track*, pages 169–176, 2010.

Finale Doshi-Velez, Kurt Miller, Jurgen Van Gael, and Yee Whye Teh. Variational inference for the Indian Buffet Process. In *Proc. AISTATS*, 2009.

T. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. Technical Report TR 2005-001, Gatsby Computational Neuroscience Unit, 2005.

Morgan Harvey, Mark J. Carman, Ian Ruthven, and Fabio Crestani. Bayesian latent variable models for collaborative item rating prediction. In *Proc. CIKM '11*, pages 699–708, 2011.

M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. An introduction to variational methods for graphical models. In M.I. Jordan, editor, *Learning in Graphical Models*, pages 105–162. Kluwer, Dordrecht, 1998.

Edward Meeds, Zoubin Ghahramani, Radford M. Neal, and Sam T. Roweis. Modeling dyadic data with binary latent factors. In *Proc. Advances in Neural Information Processing Systems*, 2007.

A. Paterek. Improving regularized singular value decomposition for collaborative filtering. In *In KDDCup '07*, 2007.

Y. W. Teh, D. Gorur, and Z. Ghahramani. Stick-breaking construction for the Indian buffet process. In *Proc. of the 11th Conference on Artificial Intelligence and Statistics*, 2007.

Mingyuan Zhou, Chunping Wang, Minhua Chen, John Paisley, David Dunson, and Lawrence Carin. Nonparametric bayesian matrix completion. In *IEEE Sensor Array and Multichannel Signal Processing Workshop*, 2010.