Fast Inference in Infinite Hidden Relational Models

Zhao Xu ZHAO.XU@WEB.DE
Institute for Computer Science, University of Munich, Germany

Volker Tresp VOLKER.TRESP@SIEMENS.COM
Corporate Technology, Siemens AG, Germany

Shipeng Yu SHIPENG.YU@SIEMENS.COM
Siemens Medical Solutions, USA

Kai Yu KYU@SV.NEC-LABS.COM
NEC Laboratories America, USA

Hans-Peter Kriegel KRIEGEL@DBS.IFI.LMU.DE
Institute for Computer Science, University of Munich, Germany

1. Introduction

Relational learning (Dzeroski & Lavrac, 2001; Friedman et al., 1999; Raedt & Kersting, 2003) is an area of growing interest in machine learning. Xu et al. (2006) introduced the infinite hidden relational model (IHRM) which views relational learning in context of the entity-relationship database model with entities, attributes and relations (compare also (Kemp et al., 2006)). In the IHRM, for each entity an auxiliary latent variable is introduced. The latent variable is the only parent of attributes of the entity and is a parent of attributes of relations the entity participates. The number of hidden states is entity class specific. Therefore it is sensible to work with Dirichlet process (DP) mixture models in which each entity class can optimize its own representational complexity in a self-organized way. For our discussion it is sufficient to say that we integrate a DP mixture model into the IHRM by simply letting the number of hidden states for each entity class approach infinity. Thus, a natural outcome of the IHRM is clustering effect providing interesting insight into the structure of the domain.

Figure 1 left shows an IHRM on a movie recommendation system. In the system, there are entity classes User, Movie and relationship class Like. In addition there are User Attributes, Movie Attributes and Relationship Attributes R with various parameters and hyperparameters. In the IHRM, for each entity an infinite-dimensional latent variable is introduced (Z^u and Z^m). They can be thought of as unknown attributes of users and movies, and are the parents of user attributes, movie attributes and relationship attributes. The underlying assumption is that if the latent variable was known, these attributes can be well predicted. The most important result of introducing the latent variables is that information can *propagate* through the ground network (Figure 1 right) via inter-connected latent variables. Let us consider the prediction of relationship attribute R for user i and movie j. If both user i and movie j have strong known attributes A_i^u and A_j^m, these will determine the state of latent variables Z_i^u and Z_j^m, and prediction for R is mostly based on A_i^u and A_j^m. In terms of a recommender system we would obtain a content-based recommendation system. Conversely, if the known attributes A_i^u are weak, the states of Z_i^u for user i might be determined by its relations with other movies and the states of those movies' latent variables. This also applies for the movie j. Again in terms of a recommender system we would obtain a collaborative-filtering system. So with the help of the latent variables, information can distribute globally in the ground network defined by the relationship structure. This reduces the need for extensive structural learning, which is particularly difficult in relational models due to the huge number of potential parents.

As in other approaches to relational learning, inference is executed in a large interconnected ground network. Thus being able to perform efficient inference is critical for the success of the IHRM. The main contribution in this paper is the analysis of four inference methods: blocked Gibbs sampler with truncated stick-breaking (TSB), blocked GS with the Dirichlet-multinomial allocation (DMA) and the corresponding mean field solutions. These methods are evaluated in two domains: movie recommendation and gene function prediction.

2. Inference based on Gibbs Sampling

Inference in (Xu et al., 2006) is based on the Chinese restaurant process (CRP), which is a collapsed version of Pólya urn sampling. The sampler updates latent variables one at a time which potentially slows down the method. Blocked sampling (Ishwaran & James, 2001) typically shows better mixing. Thus we

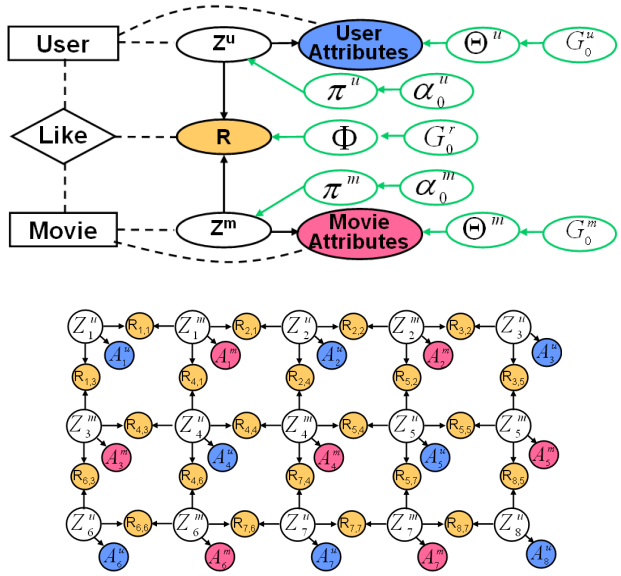

Figure 1. Left: An IHRM for movie recommendation with the DAPER representation. Right: The ground network.

extend the efficient blocked sampler with truncated stick-breaking (TSB) or Dirichlet-Multinomial allocation (DMA) to the IHRM.

We now introduce some notations. Let the number of entity classes be C, and let G_0^c and α_0^c be the base distribution and concentration parameter of entity class c. In an entity class c, there are N^c entities e_i^c indexed by i, and K^c mixture components θ_k^c indexed by k. θ_k^c are the parameters of distributions of the entity attributes A^c. The number of relationship classes is B and G_0^b is the base distribution for a relationship attribute R^b. For a relationship class b between two entity classes c_i and c_j, there are $K^{c_i} \times K^{c_j}$ correlation mixture components $\phi_{k,\ell}^b$ indexed by hidden states k for c_i and ℓ for c_j. $\phi_{k,\ell}^b$ are the parameters of distributions of relationship attributes. Here we assume that the entity and relationship attributes are drawn from exponential family distributions. G_0^c and G_0^b are the conjugate priors with the hyperparameters β^c and β^b.

Gibbs Sampling with TSB In the method, the posteriors of parameters θ_k^c and $\phi_{k,\ell}^b$ are explicitly sampled in the form of truncated stick breaking representation (TSB). The advantage is that given the posterior, we can independently sample the latent variables in a block, which highly accelerates the computation. The Markov chain is thus defined not only on the latent variables Z_i^c, but also the parameters: π^c, θ^c and ϕ^b. Note, that there are additional parameters K^c in block GS, which specify the positions to truncate the DPs. In practice, we set K^c as the number of entities of class c, K^c will be automatically reduced to a suitable value based on the complexity of the data in the sampling process. Taking some initial values for Z, π^c, θ^c and ϕ^b, the following steps are repeated until convergence:

1. For each entity class c,
   (a) Update hidden variable Z_i^c for each entity i independently:
   $$P(Z_i^c = k | D_i^c, Z_{-i}, \pi^c, \theta^c, \{\phi^{b'}\}_{b'=1}^{B'}) \propto \pi_k^c P(A_i^c|\theta_k^c) \prod_{b'} \prod_{j'} P(R_{i,j'}^{b'}|Z_i^c = k, Z_{j'}^{c_{j'}}, \phi^{b'}). \quad (1)$$
   Where D_i^c denotes all information of the entity i, i.e. its attributes A_i^c and relations $R_{i,j'}^{b'}$.
   (b) Update π^c as follows:
   i. Sample v_k^c from $\text{Beta}(\lambda_{k,1}^c, \lambda_{k,2}^c)$ for $k = \{1, \ldots, K^c - 1\}$ with
   $$\lambda_{k,1}^c = 1 + \sum_{i=1}^{N^c} \delta_k(Z_i^c), \quad (2)$$
   $$\lambda_{k,2}^c = \alpha_0^c + \sum_{k'=k+1}^{K^c} \sum_{i=1}^{N^c} \delta_{k'}(Z_i^c), \quad (3)$$
   and set $v_{K^c}^c = 1$. Where $\delta_k(Z_i^c)$ equals to 1 if $Z_i^c = k$ and 0 otherwise.
   ii. Compute $\pi_1^c = v_1^c$, $\pi_k^c = v_k^c \prod_{k'=1}^{k-1}(1 - v_{k'}^c), k > 1$.
2. Update the parameters from their posteriors given the sampled Z: $\theta_k^c \sim P(\cdot|A^c, Z^c, G_0^c)$ and $\phi_{k,\ell}^b \sim P(\cdot|R^b, Z, G_0^b)$.

Gibbs Sampling with DMA The *Dirichlet-Multinomial allocation* (DMA) approximation to DP (Green & Richardson, 2000; Yu et al., 2005) has a similar truncation form as TSB, but differs in that the prior $P(\pi^c|\alpha_0^c)$ now takes an exchangeable K^c-dimensional Dirichlet distribution $\text{Dir}(\alpha_0^c/K^c, \ldots, \alpha_0^c/K^c)$, not a stick-breaking prior as in TSB. Therefore, the blocked sampling with DMA is the same as that with TSB except in step 1.*b*, where we directly sample the mixing weight π^c from the posterior $\text{Dir}\left(\frac{\alpha_0^c}{K^c} + \sum_{i=1}^{N^c} \delta_1(Z_i^c), \ldots, \frac{\alpha_0^c}{K^c} + \sum_{i=1}^{N^c} \delta_K(Z_i^c)\right)$.

3. Mean Field Approximations

Since the proposed IHRM model has multiple DPs which interact through the relations, blocked sampling is still slow due to the slow exchange of information between DPs. Thus we explore two variational inference methods, which both assume a specific form for the posterior of all the unobservable variables, and maximize the lower bound of data log likelihood via coordinate ascent algorithm.

Mean-Field with TSB Blei and Jordan (2005) introduce a mean-field method to approximate the posterior of unobserved variables using a factorized variational distribution q. We now extend it to IHRM and define $q(\{Z^c, V^c, \theta^c\}_{c=1}^C, \{\phi^b\}_{b=1}^B)$ as
$$\left[\prod_c^C \prod_i^{N^c} q(Z_i^c|\eta_i^c) \prod_k^{K^c} q(V_k^c|\lambda_k^c) q(\theta_k^c|\tau_k^c)\right] \left[\prod_b^B \prod_k^{K^{c_i}} \prod_\ell^{K^{c_j}} q(\phi_{k,\ell}^b|\rho_{k,\ell}^b)\right].$$

Where c_i and c_j denote the entity classes involved in the relationship class b. k and ℓ denote the hidden states for c_i and c_j. $\{\eta_i^c, \lambda_k^c, \tau_k^c, \rho_{k,\ell}^b\}$ are variational parameters. $q(Z_i^c|\eta_i^c)$ is a multinomial distribution. $q(V_k^c|\lambda_k^c)$ is a Beta distribution. $q(\theta_k^c|\tau_k^c)$ and $q(\phi_{k,\ell}^b|\rho_{k,\ell}^b)$ are distributions with the same forms as G_0^c and G_0^b, respectively.

Based on Jensen's inequality, we can obtain a lower bound of the log likelihood of the data given the variational distribution q. Then we use a coordinate ascent algorithm to optimize the lower bound and yields the following updates for the variational parameters:

$$\lambda_{k,1}^c = 1 + \sum_{i=1}^{N^c} \eta_{i,k}^c, \lambda_{k,2}^c = \alpha_0^c + \sum_{i=1}^{N^c} \sum_{k'=k+1}^{K^c} \eta_{i,k'}^c, \tag{4}$$

$$\tau_{k,1}^c = \beta_1^c + \sum_{i=1}^{N^c} \eta_{i,k}^c \mathrm{T}(A_i^c), \tau_{k,2}^c = \beta_2^c + \sum_{i=1}^{N^c} \eta_{i,k}^c, \tag{5}$$

$$\rho_{k,\ell,1}^b = \beta_1^b + \sum_{i,j} \eta_{i,k}^{c_i} \eta_{j,\ell}^{c_j} \mathrm{T}(R_{i,j}^b), \rho_{k,\ell,2}^b = \beta_2^b + \sum_{i,j} \eta_{i,k}^{c_i} \eta_{j,\ell}^{c_j}, \tag{6}$$

$$\eta_{i,k}^c \propto \exp\Bigg(E_q[\log V_k^c] + \sum_{k'=1}^{k-1} E_q[\log(1 - V_{k'}^c)] + E_q[\log P(A_i^c|\theta_k^c)] + \sum_b \sum_j \sum_\ell \eta_{j,\ell}^{c_j} E_q[\log P(R_{i,j}^b|\phi_{k,\ell}^b)] \Bigg). \tag{7}$$

Where τ_k^c are parameters of exponential family distributions $q(\theta_k^c|\tau_k^c)$, we decompose τ_k^c such that $\tau_{k,1}^c$ contains the first $dim(\theta_k^c)$ components and $\tau_{k,2}^c$ is a scalar. $\rho_{k,\ell,1}^b$ and $\rho_{k,\ell,2}^b$ are defined equivalently. $\mathrm{T}(A_i^c)$ denotes the *sufficient statistic* of the exponential family distribution $P(A_i^c|\theta_k^c)$. It is clear that Equation 4 and Equation 5 updates the variational parameters for entity class c, and follow equations in (Blei & Jordan, 2005). Equation 6 updates the variational parameters for relationship attributes, which is computed on the involved entities. The most interesting updates are Equation 7, where the posteriors of entity-component assignment are *coupled together*. These essentially connect the DPs.

Mean-Field with DMA The other variational algorithm extends to the DMA approximation of DP (Green & Richardson, 2000; Yu et al., 2005). The basic difference from the above approximation is that we directly assume a variational distribution $\mathrm{Dir}(\pi^c|\lambda^c)$ to the mixing weights π^c, instead of K^c Beta distributions $q(V_k^c|\lambda_k^c)$. A similar coordinate ascent algorithm is derived as the one based on TSB, except the updates for λ^c and η^c:

$$\lambda_k^c = \frac{\alpha_0^c}{K^c} + \sum_{i=1}^{N^c} \eta_{i,k}^c; \tag{8}$$

$$\eta_{i,k}^c \propto \exp\Big(E_q[\log \pi_k^c] + E_q[\log P(A_i^c|\theta_k^c)] + \sum_b \sum_j \sum_\ell \eta_{j,\ell}^{c_j} E_q[\log P(R_{i,j}^b|\phi_{k,\ell}^b)] \Big). \tag{9}$$

The coupling of entity assignments $\eta_{i,k}^c$ remains the same as the one based on TSB.

4. Experimental Analysis

We demonstrate the proposed inference algorithms in two domains, including movie recommendation system and prediction of gene functions. For space limitation, we only list some results on the MovieLens data (Sarwar et al., 2000). There are in total 943 users and 1680 movies in the data set, and we obtain 702 users and 603 movies after removing low-frequent entities. The average number of ratings of each user is 112. We used data from 546 users for training and 156 users for testing. The performances of all approaches are analyzed from 3 points: prediction accuracy for ratings, convergence time and clustering effect.

We compare the following methods: Chinese restaurant process Gibbs sampling (CRPGS), truncated SB Gibbs sampling (TSBGS), Dirichlet-multinomial allocation Gibbs sampling (DMAGS), and the two corresponding mean field methods TSBMF and DMAMF, as well as Pearson-coefficient collaborative filtering. For TSBMF and DMAMF we consider $\alpha_0 = \{5, 10, 100, 1000\}$, and obtain the best prediction when $\alpha_0 = 100$. For CRPGS, TSBGS and DMAGS α_0 is set to be 100. For the variational methods, the change of variational parameters between two iterations is monitored to determine the convergence. For the Gibbs samplers, the convergence was analyzed by three measures: Geweke statistic on likelihood, Geweke statistic on the number of components for each entity class, and autocorrelation. Table 1 shows that the two blocked Gibbs samples converge approximately by a factor 5 faster than CRPGS. The mean field methods are again by a factor around 10 faster than the blocked Gibbs samplers and thus almost two orders of magnitude faster than CRPGS. CRPGS is much slower than the other two Gibbs samplers mainly due to the large time cost per iteration shown as Table 1. The reason is that CRPGS samples the hidden variables one by one, which causes two additional time costs. First, the expectations of attribute parameters and relational parameters have to be updated when sampling each user/movie. Second, the posterior of hidden variables have to be computed one by one, thus we can not use fast matrix multiplication technology to accelerate the computation. The prediction results are measured with prediction accuracy (shown in Table 1). For each test user, we respectively select 5, 10, 15 and 20 ratings as the known ratings, and predict the remaining ones. The results are denoted as *Given5*, *Given10*, *Given15* and *Given20* in Table 1. All methods achieved comparably good; the best results are achieved by the Gibbs samplers. The IHRM outperforms the traditional collaborative filtering method, especially when there are a few known ratings for the test users.

IHRM also provides cluster assignments for all entities involved. The columns $\#Comp.^u$ and $\#Comp.^m$

Table 1. Performances of the proposed inference methods on MovieLens data.

	Prediction Accuracy							
	Given5	Given10	Given15	Given20	Time (s)	Time(s/iter.)	$\#Comp.^u$	$\#Comp.^m$
CRPGS	65.13	65.71	66.73	68.53	164993	109	47	77
TSBGS	65.51	66.35	67.82	68.27	33770	17	59	44
DMAGS	65.64	65.96	67.69	68.33	25295	17	52	34
TSBMF	65.26	65.83	66.54	67.63	2892	19	9	6
DMAMF	64.23	65.00	66.54	66.86	2893	19	8	12
Pearson	57.81	60.04	61.25	62.41	-	-	-	-

Table 2. Clustering result of CRP-based Gibbs sampler on MovieLens data.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
My Best Friend's Wedding (1997) G.I. Jane (1997) The Truth About Cats and Dogs (1996) Phenomenon (1996) Up Close and Personal (1996) Tin Cup (1996) Bed of Roses (1996) Sabrina (1995) Clueless (1995).....	Big Night (1996) Antonia's Line (1995) Three Colors: Red (1994) Three Colors: White (1994) Cinema Paradiso(1989) Henry V (1989) Jean de Florette (1986) A Clockwork Orange (1971) Citizen Kane (1941) Mr. Smith Goes to Washington (1939).....	Swingers (1996) Get Shorty (1995) Mighty Aphrodite (1995) Welcome to the Dollhouse (1995) Clerks (1994) Ed Wood (1994) The Hudsucker Proxy (1994) What's Eating Gilbert Grape (1993) Groundhog Day (1993).....	Event Horizon (1997) Batman and Robin (1997) Escape from L.A. (1996) Batman Forever (1995) Batman Returns (1992) 101 Dalmatians (1996) The First Wives Club (1996) Nine Months (1995) Casper (1995).....

in Table 1 denote the number of clusters for User class and Movie class, respectively. The mean field solutions have a tendency to converge to a smaller number of clusters than Gibbs samplers. Further analysis shows that the clustering results of the methods are actually similar. First, the sizes of most clusters generated by the Gibbs samplers are very small, e.g., there are 72% (72.55%, 75.47%) user clusters with less than 5 members in CRPGS (DMAGS, TSBGS). Intuitively, the Gibbs samplers tend to assign the outliers to new clusters. Second, we compute the rand index (0-1) of the clustering results of the methods, e.g. the values are 0.8071 between CRPGS and TSBMF, 0.8221 between TSBGS and TSBMF, which also demonstrates the similarity of the clustering results. Table 2 illustrates the movies with highest posterior probability in the 4 largest clusters generated from CRPGS. It is quite surprising that the clustering result is highly interpretable.

5. Conclusions

The IHRM and the related IRM (Kemp et al., 2006) are novel and principled approaches to relational learning but the full potential can only be developed in combination with fast inference approaches. The blocked samplers proposed in this paper are more than a factor of five faster than the originally proposed Chinese restaurant Gibbs sampler. Another factor of 10 in speed up can be achieved by using variational methods. Thus the presented work makes the IHRM applicable to considerably larger domains. In addition we analyzed the clustering results in the experiments and found that IHRM provides quite interpretable clusters in the movie recommendation and gene domains.

References

Blei, D., & Jordan, M. (2005). Variational inference for dp mixtures. *Bayesian Analysis*, *1*, 121–144.

Dzeroski, S., & Lavrac, N. (Eds.). (2001). *Relational data mining*. Berlin: Springer.

Friedman, N., Getoor, L., Koller, D., & Pfeffer, A. (1999). Learning probabilistic relational models. *Proc. 16th IJCAI* (pp. 1300–1309). Morgan Kaufmann.

Green, P. J., & Richardson, S. (2000). Modelling heterogeneity with and without the dirichlet process.

Ishwaran, J., & James, L. (2001). Gibbs sampling methods for stick breaking priors. *JASA*, *96*, 161–174.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. *Proc. 21st AAAI*.

Raedt, L. D., & Kersting, K. (2003). Probabilistic logic learning. *SIGKDD Explor. Newsl.*, *5*, 31–48.

Sarwar, B. M., Karypis, G., Konstan, J. A., & Riedl, J. (2000). Analysis of recommender algorithms for e-commerce. *Proc. ACM E-Commerce Conference* (pp. 158–167). ACM.

Xu, Z., Tresp, V., Yu, K., & Kriegel, H.-P. (2006). Infinite hidden relational models. *Proc. 22nd UAI*.

Yu, K., Yu, S., & Tresp, V. (2005). Dirichlet enhanced latent semantic analysis. *AISTATS 05* (pp. 437–444).