Relational Gaussian Processes for Learning Preference Relations

Kristian Kersting KRISTIAN.KERSTING@IAIS.FRAUNHOFER.DE
Zhao Xu ZHAO.XU@IAIS.FRAUNHOFER.DE
Fraunhofer IAIS, Schloss Birlinghoven, 53754 Sankt Augustin, Germany

1. Introduction

Preference learning has received increasing attention in both machine learning and information retrieval. The goal of preference learning is to automatically learn a model to rank entities (e.g., documents, webpages, products, music, etc.) according to their degrees of relevance. The particularity of preference learning might be that the training data is a set of pairwise preferences between entities, instead of explicit entity-wise values. For example, we may only know that a user prefers an item to another one $e_i \succ e_j$, but we do not know the exact preference degrees of items.

Gaussian processes have successfully been used to learn preferences among entities (Chu & Ghahramani, 2005), as they provide nonparametric Bayesian approaches for model selection and probabilistic inference. Basically, GP based preference learning models introduce for each entity a latent variable, which is a function value $f(x_i)$ (shortened as f_i in the rest of the paper) of entity attributes x_i. We can intuitively view the latent function values as preference degrees of entities. Then entities are ranked according to the latent values. Namely if an entity is preferred to another one $e_i \succ e_j$, then latent function value of the entity is larger than that of another one $f_i > f_j$.

However, existing GP preference models only exploit the information about entity attributes. The information about relations among entities is not taken into account in these models, whereas such information is very important and informative in many applications. To incorporate relations into nonparametric Bayesian ranking models, we present a relational Gaussian process (XPGP) approach. The key insight of XPGP is that some hidden common causes lurk in relational graphs, and the hidden common causes are important factors to influence the preference degrees of entities. The overall preference degree of an entity is combined result of entity attributes and the hidden common causes. Technically, under the GP framework, we introduce to each entity an additional latent function value $g(r_i)$ (shortened as g_i) of relations r_i of the entity, which encodes the preference causes hidden in relations. Then, we model entity preference degree ξ_i as linear combination of related function values (i.e. f_i and g_i). Each preference $(e_i \succ e_j)$ is modeled as a random variable conditioned on an indicator that is a function of preference degrees (ξ_i and ξ_j) of involved entities. The method of integrating relations into probabilistic kernels by graph Laplacian was first introduced by (Silva et al., 2007), and was proven to be successful in classification. Here, we extend it to learning preferences. The XPGP model can be straightforwardly extended to scenarios with multiple relations, directed and bipartite relations. The experimental analysis on real-world dataset LETOR demonstrates that incorporating relations can improve the quality of preference learning.

2. Model Description

In this section, we will describe the XPGP model for learning preferences. Assume that there are a set of n entities $E = \{e_1, \ldots, e_n\}$, with attributes $X = \{x_i : x_i \in \mathbb{R}^D, i = 1, \ldots, n\}$, relations between them $R = \{r_{i,j} : i, j \in 1, \ldots, n\}$, and a set of m observed pairwise ordinal relations (preferences/ranks) among entities, $O = \{e_{i_s} \succ e_{j_s} : s = 1, \ldots, m; i_s, j_s \in 1, \ldots, n\}$ (i_s and j_s are entities involved in s-th observed preference). All relations of entity e_i are denoted as r_i.

In XPGP, we introduce to each entity two latent function values $f(x_i)$ and $g(r_i)$ (shortened as f_i, g_i). $f(\cdot)$ and $g(\cdot)$ are functions of attributes and relations, respectively. Their values preserve the attribute factor and relational factor of the entity on preferences. The linear combination of latent function values $\xi_i = \omega_1 f_i + \omega_2 g_i$ is an indicator, which specifies preference degree of the entity, taking two factors (attribute- and relation-wise) into account. A preference $e_i \succ e_j$ is conditioned on indicators of e_i and e_j with a likelihood distribution $P(e_i \succ e_j | \xi_i, \xi_j)$. The key idea of the XPGP model is illustrated in Fig. 1.

Prior Distributions: We assume that an infinite

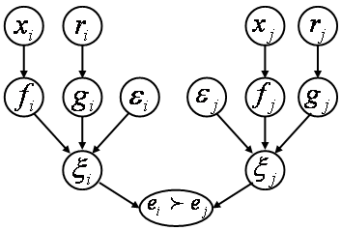

Figure 1. Graphical representation of the XPGP model. f_i is a latent function value of entity attributes. $\{f_1, f_2, \ldots\}$ follow Gaussian process prior. g_i is a latent function value of relations, $\{g_1, g_2, \ldots\}$ follow another GP prior. ϵ_i denotes Gaussian noise. ξ_i is the overall preference degree of the entity, which is weighted sum of f_i, g_i and ϵ_i. f_j, g_j and ϵ_j are defined equivalently.

number of latent function values follow a GP prior with zero mean and covariance function $k_a(x_i, x_j)$ (the subscript a emphasizes they are attribute-wise). In terms of definition of GP, any finite set of function values $f = \{f_i : i = 1, \ldots, n\}$ have a multivariate Gaussian distribution, which mean and covariance matrix K_a are defined by mean and covariance functions of the GP (Rasmussen & Williams, 2006).

Equivalently, we define $\{g_1, g_2, \ldots\}$ as another zero-mean GP, so that $g = \{g_i : i = 1, \ldots, n\}$ follows a multivariate Gaussian distribution with covariance matrix K_r. However, we note that the covariance function $k_r(r_i, r_j)$ should represent correlation of i and j on relations. There are generally two strategies to define the kernel function. The simplest way is to represent the relations of entity i as a vector. The kernel function $k_r(r_i, r_j)$ can then be any Mercer kernel functions, and the computation is similar with that of attributes. Alternatively, we can employ graph-based kernels to obtain the covariances (Smola & Kondor, 2003).

Now we can write the prior $P(f, g|A, R)$ as:

$$(2\pi)^{-n}|K_a|^{-\frac{1}{2}}|K_r|^{-\frac{1}{2}}\exp(-\frac{f^T K_a^{-1} f + g^T K_r^{-1} g}{2}). \quad (1)$$

Likelihood of Preference: Essentially, we extend the likelihood distribution introduced by (Chu & Ghahramani, 2005) to relational domains. In relational data, the preference degree of an entity consists of two components: attribute-wise factor and relation-wise factor, respectively represented as latent function values f_i and g_i. The overall preference degree of the entity is technically represented as weighted sum of the two latent functions: $\xi_i = \omega_1 f_i + \omega_2 g_i$. The likelihood function is

$$P(e_i \succ e_j|\xi_i - \xi_j) = \int_{-\infty}^{\frac{\xi_i - \xi_j}{\sqrt{2}\sigma}} N(t|0,1)dt \equiv \Phi(\frac{\xi_i - \xi_j}{\sqrt{2}\sigma}),$$

which means: the larger the preference degree of e_i is than that of e_j, the more likely e_i is preferred to e_j.

Directed, Bipartite and Multiple Relations: In multi-relational scenarios, we use distinct latent function values to represent preference factors driven by different types of relations, i.e., introduce to each entity multiple relational function values, one for each type of relations: $\{g_i^{r1}, g_i^{r2}, \ldots\}$. The latent function values of the same type of relations, e.g. $\{g_1^{r1}, \ldots, g_n^{r1}\}$, share a GP prior. The overall preference degree is weighted sum of all latent function values associated with the entity: $\xi_i = \omega_1 f_i + \omega_2 g_i^{r1} + \omega_3 g_i^{r2} + \ldots + \epsilon_i$.

In a directed relation, the two involved entities play different roles. Consider e.g. links of webpages. The entities typically serve as the linking and linked webpages. It is reasonable to introduce to a webpage two latent function values to respectively represent preference factors from linking and linked "roles" of the entity. The preference degree of a webpage becomes: $\xi_i = \omega_1 f_i + \omega_2 g_i^{linking} + \omega_3 g_i^{linked} + \epsilon_i$.

If relations are bipartite, i.e., there are different types of entities involved in relations, then graph kernels for univariate relations are not applicable. We address the problem by projecting bipartite relations to univariate ones. Specifically, we add a relation between entities i and j iff. both entities link to the same entity. Then we can compute the kernels on the projected graphs.

3. EMEP-based Inference and Learning

This section will present the inference and learning methods for the XPGP model based on the EM-EP method (Kim & Ghahramani, 2006). The key inference problem is computing the posterior distribution over the latent function values given attributes X, relations R, and preference relations O:

$$P(f, g|X, R, O) \propto P(f, g|X, R)\prod_s P(e_{i_s} \succ e_{j_s}|\xi_{i_s}, \xi_{j_s}),$$

where prior and likelihood distributions are defined as Sec. 2. Unfortunately, computing this posterior distribution is intractable. We thus stick to the expectation propagation (EP) algorithm to approximate the posterior distribution, i.e., we use unnormalized Gaussian distributions $t_s(\xi_{i_s}, \xi_{j_s}|\tilde{\mu}_s, \tilde{\sigma}_s^2, \tilde{Z}_s)$ to approximate the real likelihood distribution $\Phi(\frac{\xi_{i_s} - \xi_{j_s}}{\sqrt{2}\sigma})$. In the inference process, we update the approximations for each observed preference pair sequentially until convergence.

We learn the hyperparameters under the empirical Bayesian framework with a Expectation Maximum approach. In the E step, the EP parameters ($\tilde{\mu}_{i,k}$, $\tilde{\sigma}_{i,k}^2$, $\tilde{Z}_{i,k}$) are optimized to approximate the posterior of latent variables with the current values of hyperparameters. In the M step, the hyperparameters are optimized to maximize the lower bound of log marginal

likelihood with scaled conjugate gradient method.

4. Experiments

We apply XPGP to relevance feedback, and evaluate it with the OHSUMED dataset in LETOR (Liu et al., 2007). In the experiments, we use XPGP model to predict preferences on articles (resp. webpages) based on some known preferences. This corresponds to transductive preference learning. We compare the XPGP model with standard GP (Chu & Ghahramani, 2005) and SVM models (Joachims, 2002). For GP-based approaches, we use Gaussian kernels to compute the covariance matrixes on entity attributes, and use two different graph kernels to obtain correlations on relations: one is regularized Laplacian and the other is the kernel from (Silva et al., 2007). The RBF kernel is employed in the SVM method.

In the OHSUMED dataset, there are 106 queries, each of which is associated with some relevant documents evaluated by humans. The relevance degree has three levels: definitely relevant, partially relevant and not relevant. (Liu et al., 2007) sampled some "possible" relevant documents from the document collection and got about 152 documents on average for a query. They extracted 25 dimension vector for each query-document pair. The relations between documents are based on similarity matrix, i.e. there is a weighted complete graph between documents and the weight of each edge is cosine similarity between the two documents. Although the relations here are "pseudo" ones, the model can deal with real relations, e.g. friendship.

In the dataset, each document is associated with a relevance degree. We transfer the entity-wise degrees to the pair-wise preferences. This is not only due to the need of modeling with XPGP, most importantly, preferences are more realistic in real-world applications. In the experiments, we randomly select 100, 150, 200 preference pairs for each query as evidence, and predict the remaining ones. For each setting, the selection was repeated 20 times. Fig. 2 shows the experimental results averaged over randomly selected 10 queries. We compute the prediction error rate and the area under ROC curve (AUC). In all settings, the XPGP model outperforms the other two models, especially when the number of known preference pairs is small. The two different graph kernels obtain similar results.

5. Summary

We propose a relational GP model XPGP for preference learning, which integrates relations with enhanced priors using graph kernels. The empirical analysis on the LETOR dateset demonstrates that relational information can improve the performance of a probabilistic kernel method on learning preference.

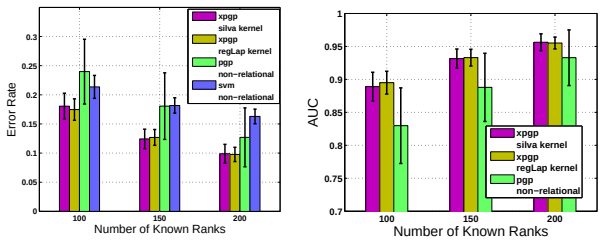

Figure 2. Experimental results on predicting preference pairs given different number of known ones. Left: Prediction error rate. The less the better. Right: Area under ROC curve. The larger the better.

Acknowledgments

The authors would like to thank the anonymous reviewers for their comments. This research was supported by the Fraunhofer ATTRACT fellowship STREAM.

References

Chu, W., & Ghahramani, Z. (2005). Preference learning with gaussian processes. *ICML 2005*.

Joachims, T. (2002). Optimizing search engines using clickthrough data. *SIGKDD 2002*.

Kim, H.-C., & Ghahramani, Z. (2006). Bayesian gaussian process classification with the em-ep algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (pp. 1948–1959).

Liu, T.-Y., Qin, T., Xu, J., Xiong, W.-Y., & Li, H. (2007). Letor: benchmark dataset for research on learning to rank for information retrieval. *SIGIR 2007 Workshop on LR4IR*.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. The MIT Press.

Silva, R., Chu, W., & Ghahramani, Z. (2007). Hidden common cause relations in relational learning. *Neural Information Processing Systems*.

Smola, A. J., & Kondor, I. (2003). Kernels and regularization on graphs. *Annual Conference on Computational Learning Theory*.