Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Improving Cross-Domain Recommendation through Probabilistic Cluster-Level Latent Factor Model

Siting Ren, Sheng Gao, Jianxin Liao and Jun Guo
Beijing University of Posts and Telecommunications
{ rensiting, gaosheng, guojun } @bupt.edu.cn, liaojianxin@ebupt.com

Abstract

Cross-domain recommendation has been proposed to transfer user behavior pattern by pooling together the rating data from multiple domains to alleviate the sparsity problem appearing in single rating domains. However, previous models only assume that multiple domains share a latent common rating pattern based on the user-item co-clustering. To capture diversities among different domains, we propose a novel Probabilistic Cluster-level Latent Factor (PCLF) model to improve the cross-domain recommendation performance. Experiments on several real world datasets demonstrate that our proposed model outperforms the state-of-the-art methods for the cross-domain recommendation task.

Methodology

Traditional recommender systems based on collaborative filtering aim to provide recommendations for users on a set of items belonging to only a single domain (e.g., music or movie) based on the historical user-item preference records. In many cases, in particular when the system is new, the data sparsity problem arises, and prevents the system from generating accurate recommendations. With the increasing of user-generated content, there exists a considerable number of publicly available user-item ratings from different domains. Thus, instead of treating items from each single domain independently, knowledge acquired in a single domain could be transferred and shared in other related domains, which has been referred to as Cross-Domain Recommendation (Gao et al. 2013) problem.

Cross-domain recommendation models have shown that knowledge transfer and sharing among the related domains can be beneficial to alleviate the data sparsity problem in single-domain recommendations. CBT (Li, Yang, and Xue 2009a) is an early transfer learning model which studies knowledge transferability between two distinct domains. A cluster-level rating pattern (a.k.a., codebook) is constructed from the auxiliary data via some user-item co-clustering and transferred to the target data by codebook expansion. A later extension called RMGM (Li, Yang, and Xue 2009b) combines codebook construction and codebook expansion in CBT into one single step with soft membership indicator matrices.

However, related domains do not necessarily share such a common rating pattern. The diversity among related domains might outweigh the advantages of the common rating pattern, which may result in performance degradations. That is, existing models cannot consider the domain-specific knowledge about the rating patterns to improve the mutual strengths in cross-domain recommendation. To learn the shared knowledge and not-shared effect of each domain simultaneously, we propose a novel Probabilistic Cluster-level Latent Factor (PCLF) model. In real-world scenarios, for items, we discover that the domain-specific clusters and the common clusters may exist simultaneously. For example, movies and music can be both classified by regions, but a movie category (e.g., science fiction) may be unable to describe music. Also regions (the common clusters) and the movie category (the domain-specific clusters) affect the rating results of movies in a certain proportion. So including a domain-specific rating pattern may be more accurate than sharing the common knowledge only.

Suppose that we are given Z rating matrices from related domains for rating prediction. And we assume that the hidden cluster-level structures across multiple domains can be extracted to learn the rating-pattern of user clusters on the item clusters. In real world, users may have multiple personalities, causing a user can simultaneously belong to multiple user clusters. Suppose there are K user clusters C_u among all Z domains. The probability of a user u belonging to an exact user cluster k can be described as $P(C_u^{(k)}|u)$, and then the user cluster membership vector for a user u is equation 1. (s.t. $\mathbf{p_u}\mathbf{1} = 1$)

$$\mathbf{p_u} = \left[P(C_u^{(1)}|u), P(C_u^{(2)}|u), \ldots, P(C_u^{(K)}|u)\right] \quad (1)$$

Furthermore, item clusters should be divided into common item clusters and domain-specific item clusters. Common item clusters represent the mutual features of items in all Z domains. Domain-specific item clusters describe the properties of items in each domain. Suppose there are T common item clusters C_{vcom} among Z domains, and L_z domain-specific item clusters C_{vspez} for the z-th domain. Similar to user clusters, common item cluster membership vector and domain-specific item cluster membership vector for an item v in the z-th domain are defined as equation 2,3 respectively.

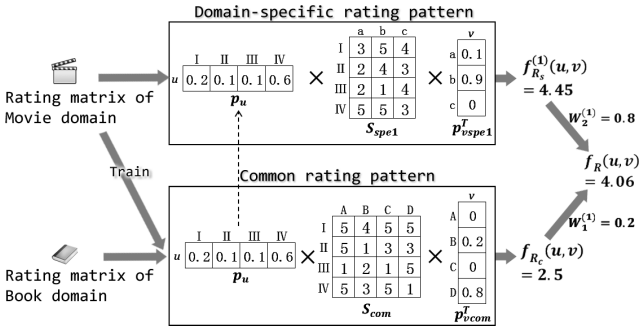

Figure 1: Illustration of PCLF model in the context of two related domains. The rating prediction result $f_R(u,v)$ is the combination of cross-domain rating $f_{R_c}(u,v)$ and single-domain rating $f^{(1)}_{R_s}(u,v)$ with certain proportion.

(s.t. $\mathbf{p_{vcom}}\mathbf{1} = 1$, $\mathbf{p_{vspez}}\mathbf{1} = 1$)

$$\mathbf{p_{vcom}} = \left[P(C^{(1)}_{vcom}|v), \ldots, P(C^{(T)}_{vcom}|v)\right] \quad (2)$$

$$\mathbf{p_{vspez}} = \left[P(C^{(1)}_{vspez}|v), \ldots, P(C^{(L_z)}_{vspez}|v)\right] \quad (3)$$

Then we construct two types of cluster-level rating matrices. One is common cluster-level rating matrix $S_{com} \in \mathbb{R}^{K\times T}$, where each element $S_{com}(k,t) = \sum_r rP(r|C^{(k)}_u, C^{(t)}_{vcom})$ denotes the expectation of rating given by user cluster k to common item cluster t. The other is domain-specific cluster-level rating matrix $S_{spez} \in \mathbb{R}^{K\times L_z}$, where each element $S_{spez}(k,l) = \sum_r rP(r|C^{(k)}_u, C^{(l)}_{vspez})$ denotes the expectation of rating given by user cluster k to domain-specific item cluster l in the z-th domain.

We follow the assumption that random variables u and v are independent (Li, Yang, and Xue 2009b) to get

$$f_{R_c}(u,v) = \mathbf{p_u} S_{com} \mathbf{p^T_{vcom}} \quad (4)$$

$$f^{(z)}_{R_s}(u,v) = \mathbf{p_u} S_{spez} \mathbf{p^T_{vspez}} \quad (5)$$

$f_{R_c}(u,v)$ gives ratings for a user u on an item v from the perspective of common rating pattern. While, $f^{(z)}_{R_s}(u,v)$ gives ratings from the perspective of domain-specific rating pattern. Our model introduces a set of variables $W^{(z)}_1, W^{(z)}_2$ to balance the effect between them. ($W^{(z)}_1 + W^{(z)}_2 = 1$)

Thus, the ultimate predicting rating for the z-th domain is

$$f_R(u,v) = W^{(z)}_1 f_{R_c}(u,v) + W^{(z)}_2 f^{(z)}_{R_s}(u,v) \quad (6)$$

We adopt annealed EM algorithm (AEM) for model training, which is an EM algorithm with regularization to avoid the local maximum problems. The illustration of PCLF model can be found in Figure 1.

Results

We conduct several experiments to examine how our model behaves on real-world datasets and compare it with several single-domain and cross-domain recommendation models.

Dataset	Model	Given 5	Given 10	Given 15
Book-Crossing	NMF	0.6575	0.6375	0.6301
	FMM	0.6451	0.6196	0.6125
	RMGM	0.6378	0.6115	0.6092
	PCLF	**0.6252**	**0.5994**	**0.5969**
Each-Movie	NMF	0.9345	0.8861	0.8799
	FMM	0.9132	0.8831	0.8771
	RMGM	0.9021	0.8743	0.8637
	PCLF	**0.8838**	**0.8677**	**0.8533**

Table 1: MAE performances of the compared models on related domains under different configurations. Best results are in bold.

Some of the MAE (Mean Absolute Error) performances results on *Book-Crossing* [1] vs *EachMovie*[2] datasets are shown in table 1. We normalize the rating scales of each dataset from 1 to 5 for a fair comparison. In each dataset, we randomly choose 500 users (300 for training, 200 for testing) with more than 16 ratings. We keep different sizes of observed ratings as the initialization of test users. Given 5 means 5 ratings of each test user are given for training to avoid cold-start problem and the remaining are used for evaluation.

The experimental results show that our PCLF model can indeed benefit from the combination of two types of cluster-level rating patterns(i.e., common & domain-specific) and outperforms the state-of-the-art methods for recommendation task. There are still several extensions to improve our work. We plan to explore the ability to handle the implicit preferences of users (e.g., visit, click or comment) of our model. Also, most recommendation systems are expected to handle enormous amounts of data ("Big Data") at a reasonable time. So we will evaluate the scalability of our model.

Acknowledgments

This research was supported by the National Natural Science Foundation of China under grant No. 61300080, No. 61273217, the Fundamental Research Funds for the Central Universities of China under grant No. 2013RC0119, the Chinese 111 program of Advanced Intelligence and Network Service under grant No. B08004. The authors are partially supported by the Key Project of Chinese Ministry of Education under grant No. MCM20130310 and Huawei's Innovation Research Program.

References

Gao, S.; Luo, H.; Chen, D.; Li, S.; Gallinari, P.; and Guo, J. 2013. Cross-domain recommendation via cluster-level latent factor model. In *Machine Learning and Knowledge Discovery in Databases*. Springer. 161–176.

Li, B.; Yang, Q.; and Xue, X. 2009a. Can movies and books collaborate? cross-domain collaborative filtering for sparsity reduction. In *IJCAI*, volume 9, 2052–2057.

Li, B.; Yang, Q.; and Xue, X. 2009b. Transfer learning for collaborative filtering via a rating-matrix generative model. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 617–624. ACM.

[1] http://www.informatik.uni-freiburg.de/∼cziegler/BX/

[2] http://www.cs.cmu.edu/∼lebanon/IR-lab.htm