Cross-Domain Recommendation via Coupled Factorization Machines

Lile Li, Quan Do, Wei Liu
Advanced Analytics Institute, School of Software, Faculty of Engineering and Information Technology
Univeristy of Technology Sydney, Sydney, Australia
Lile.Li@student.uts.edu.au, Quan.Do@student.uts.edu.au, Wei.Liu@uts.edu.au

Abstract

Data across many business domains can be represented by two or more coupled data sets. Correlations among these coupled datasets have been studied in the literature for making more accurate cross-domain recommender systems. However, in existing methods, cross-domain recommendations mostly assume the coupled mode of data sets share identical latent factors, which limits the discovery of potentially useful domain-specific properties of the original data. In this paper, we proposed a novel cross-domain recommendation method called Coupled Factorization Machine (CoFM) that addresses this limitation. Compared to existing models, our research is the first model that uses factorization machines to capture both common characteristics of coupled domains while simultaneously preserving the differences among them. Our experiments with real-world datasets confirm the advantages of our method in making across-domain recommendations.

Introduction

Recent innovations in the Web have enabled many coupled datasets to co-describe cross-domain businesses. For example, MovieLens and Netflix websites each published their data of their users' ratings on movies. Although users on MovieLens and Netflix are not identical, they may rate the same movies. In other words, these datasets are coupled in their movie field. Joint analysis of the coupled datasets can provide a deeper understanding of their true nature, improving the recommendation performance (Gao et al. 2013).

Coupled datasets have strong correlations on their coupled field. For instance, the same action movies in MovieLens and Netflix websites are highly rated by action movie fans in both sites. However, MovieLens allows ratings from 0.5 to 5 with 0.5 increments whereas Netflix only allows 1 to 5 ratings with 1 increases. Thus, there are scenarios where action movie fans in MovieLens rate action movies with 3.5 or 4.5 stars while those in Netflix rate them with 4 or 5 stars. Due to this scale difference across sites, existing models that assume coupled datasets to share the same coupled factor or the same parameters on their coupled field are unlikely to capture the differences.

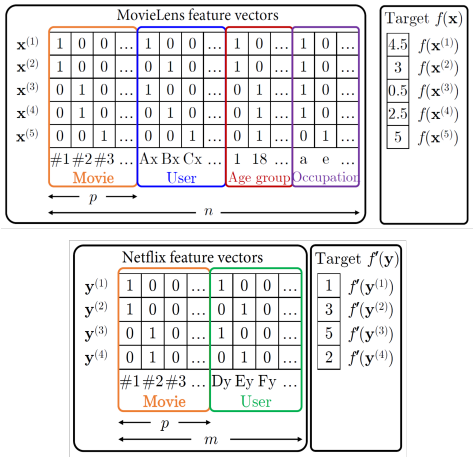

Figure 1: Ratings extracted as feature vectors from MovieLens and Netflix. Note that the target variables f and f' are not of the same incremental scale.

Developing from this observation, we hypothesize that the coupled movie field across MovieLens and Netflix have two essential properties needed to be considered in recommendation models: they share similar movie genres' characteristics, and at the same time they differ in rating scales. From this perspective, in this research our main contribution is the design of Coupled Factorization Machines (CoFM), which allows the latent vectors of the coupled field to have two critical parts. The first part captures the shared characteristics crossed datasets to exploit their correlations and the second one preserves their domain-specific uniqueness to maintain the scale differences. We test the effectiveness of CoFM on real-world across-domains datasets, and our experiments demonstrate clear superiority of CoFM over other existing methods.

Our approach

Suppose two cross-domain datasets are coupled as illustrated in Fig. 1: one with feature vectors $\mathbf{x}$ and targets $f(\mathbf{x})$ and the other with feature vectors $\mathbf{y}$ and targets $f'(\mathbf{y})$. Both $\mathbf{x}$ and $\mathbf{y}$ have the same movie field of size p but differ in all other fields (e.g., different users). Moreover, $f(\mathbf{x})$ and $f'(\mathbf{y})$

are of different scales. We aim to exploit their correlations for recommending their unseen data.

As $\mathbf{x}$ and $\mathbf{y}$ are coupled in one field of size p, the first p latent variables of their feature interactions ($\mathbf{v}_i$ and $\mathbf{v}'_i$, $\forall i \in [1, p]$) are expected to have strong coupled relationship. Furthermore, that $f(\mathbf{x})$ and $f'(\mathbf{y})$ differ in their scales suggests the two datasets possess unique characteristics of their domains. Our model enables the latent vectors $\mathbf{v}$ and $\mathbf{v}'$ of the coupled field to have an non-shared part to preserve their domain-specific differences. So our model's objective function is:

$$\hat{f}(\mathbf{x}, \mathbf{y}) = w_0 + \sum_{i=1}^{n} w_i x_i + w'_0 + \sum_{i=1}^{m} w'_i y_i$$
$$+ \sum_{i=1}^{p} \sum_{j=i+1}^{n} < \mathbf{v}_i, \mathbf{v}_j >_1^c x_i x_j + \sum_{i=1}^{p} \sum_{j=i+1}^{m} < \mathbf{v}_i, \mathbf{v}'_j >_1^c y_i y_j$$
$$+ \sum_{i=1}^{p} \sum_{j=i+1}^{n} < \mathbf{v}_i, \mathbf{v}_j >_{c+1}^k x_i x_j + \sum_{i=1}^{p} \sum_{j=i+1}^{m} < \mathbf{v}'_i, \mathbf{v}'_j >_{c+1}^k y_i y_j$$
$$+ \sum_{i=p+1}^{n} \sum_{j=i+1}^{n} < \mathbf{v}_i, \mathbf{v}_j >_1^k x_i x_j + \sum_{i=p+1}^{m} \sum_{j=i+1}^{m} < \mathbf{v}'_i, \mathbf{v}'_j >_1^k y_i y_j$$

where c is the size of the shared part in the coupled field. The second line of the equation above present *the shared part* of the coupled field (i.e., $\mathbf{v}_i, \forall i \in [1, p]$). The terms in the third line capture *the domain-specific differences* in the coupled field by different $\mathbf{v}_i$ and $\mathbf{v}'_i$ ($\forall i \in [1, p]$). The last line terms are latent vectors of the non-coupled fields.

Performance Evaluation

We compare CoFM with CDCF (Loni et al. 2014), CMF (Singh and Gordon 2008), CLFM (Gao et al. 2013), FM (Rendle 2012), and a deep learning based recommendation system DSSM (Elkahky, Song, and He 2015). In our experiments, we use two public datasets, MovieLens and Netflix. Experiments on more datasets can be found in our supplementary material. The MovieLens (Harper and Konstan 2015) dataset includes ratings of 247,753 users for 34,208 movies whereas the Netflix[1] one contains ratings of 480,189 users for 17,770 movies. Among the two sites' movies, 5,911 of them are identical. We randomly extract $10^4 \times 5,911$ dense rating matrices for each dataset. Then, we randomly take two sub-matrices of 5,000×5,000, each has 500,000 ratings of no common users. Ratings are 0.5 to 5 with 0.5 increments in MovieLens and 1 to 5 with 1 increments in Netflix.

Each algorithm is run five times and we report the means and standard deviations of RMSEs. Table 1 shows the our experiment results. One can observe that CoFM outperforms existing models statistically significantly.

We also analyze how CoFM works with different values of shared latent space c by plotting the RMSEs of different algorithms, as illustrated in Fig. 2. As other models do not have the parameter c, their results are plotted as reference lines. The figure shows that CoFM is constantly better than other alternative models.

[1]Netflix's movie ratings dataset: http://www.netflixprize.com/

Table 1: Mean and standard deviation of tested RMSE.

k	Dataset	FM	CDCF	CMF	Our CoFM
5	MovieLens	0.8585±0.0013	0.8844±0.0011	0.8885±0.0038	**0.8527**±0.0019
	Netflix	0.8962±0.0032	0.9090±0.0032	**0.8890**±0.0036	0.8965±0.0045
10	MovieLens	0.8653±0.0026	0.8901±0.0015	0.8761±0.0020	**0.8552**±0.0013
	Netflix	0.9001±0.0009	0.9154±0.0019	0.8979±0.0020	**0.8952**±0.0013
15	MovieLens	0.8643±0.0011	0.8978±0.0018	0.8786±0.0012	**0.8573**±0.0010
	Netflix	0.9018±0.0001	0.9176±0.0017	0.9089±0.0006	**0.8991**±0.0008
20	MovieLens	0.8661±0.0007	0.8939±0.0003	0.8825±0.0004	**0.8594**±0.0012
	Netflix	0.9006±0.0006	0.9153±0.0006	0.9166±0.0019	**0.8997**±0.0023
25	MovieLens	0.8680±0.0013	0.8943±0.0008	0.8865±0.0015	**0.8576**±0.0018
	Netflix	0.9013±0.0002	0.9164±0.0015	0.9192±0.0028	**0.8998**±0.0014
	t-tests	4.41×10^{-3}	6.00×10^{-6}	1.09×10^{-4}	–

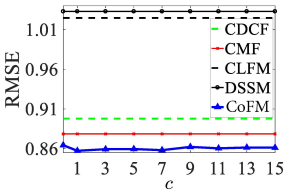

Figure 2: Performance evaluation on shared latent space c.

Conclusions

We have introduced Coupled Factorization Machines (CoFM) for joint analysis and recommendation on across-domain datasets. As cross-domain coupled datasets not only have strong correlations but also contain domain-specific differences, CoFM models their latent variables of the coupled field to capture both the similarities and differences. Our experiments with real-world datasets demonstrate the power of CoFM in utilizing correlations cross domains. In future, we plan to explore the shared latent variables more deeply by analyzing the relationship of cross-domain datasets and extending the CoFM model to higher-order interactions.

References

Elkahky, A. M.; Song, Y.; and He, X. 2015. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *International World Wide Web Conference*.

Gao, S.; Luo, H.; Chen, D.; Li, S.; Gallinari, P.; and Guo, J. 2013. Cross-domain recommendation via cluster-level latent factor model. In *ECML PKDD 2013*.

Harper, F. M., and Konstan, J. A. 2015. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*

Loni, B.; Shi, Y.; Larson, M.; and Hanjalic, A. 2014. Cross-domain collaborative filtering with factorization machines. In *the 36th European Conference on IR Research on Advances in Information Retrieval*.

Rendle, S. 2012. Factorization machines with libFM. *ACM Trans. Intell. Syst. Technol.*

Singh, A. P., and Gordon, G. J. 2008. Relational learning via collective matrix factorization. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.