Exploiting Variational Domain-Invariant User Embedding for Partially Overlapped Cross Domain Recommendation

Weiming Liu
College of Computer Science,
Zhejiang University
China
21831010@zju.edu.cn

Xiaolin Zheng
College of Computer Science,
Zhejiang University
China
xlzheng@zju.edu.cn

Jiajie Su
College of Computer Science,
Zhejiang University
China
jiajesu@zju.edu.cn

Mengling Hu
College of Computer Science,
Zhejiang University
China
humengling@zju.edu.cn

Yanchao Tan
College of Computer Science,
Zhejiang University
China
yctan@zju.edu.cn

Chaochao Chen
College of Computer Science,
Zhejiang University
China
zjuccc@zju.edu.cn

ABSTRACT

Cross-Domain Recommendation (CDR) has been popularly studied to utilize different domain knowledge to solve the cold-start problem in recommender systems. Most of the existing CDR models assume that both the source and target domains share the same overlapped user set for knowledge transfer. However, only few proportion of users simultaneously activate on both the source and target domains in practical CDR tasks. In this paper, we focus on the Partially Overlapped Cross-Domain Recommendation (POCDR) problem, that is, how to leverage the information of both the overlapped and non-overlapped users to improve recommendation performance. Existing approaches cannot fully utilize the useful knowledge behind the non-overlapped users across domains, which limits the model performance when the majority of users turn out to be non-overlapped. To address this issue, we propose an end-to-end dual-autoencoder with Variational Domain-Invariant Embedding Alignment (VDEA) model, a cross-domain recommendation framework for the POCDR problem, which utilizes dual variational autoencoders with both local and global embedding alignment for exploiting domain-invariant user embedding. VDEA first adopts variational inference to capture collaborative user preferences, and then utilizes Gromov-Wasserstein distribution co-clustering optimal transport to cluster the users with similar rating interaction behaviors. Our empirical studies on Douban and Amazon datasets demonstrate that VDEA significantly outperforms the state-of-the-art models, especially under the POCDR setting.

CCS CONCEPTS

• Information systems → Recommender systems.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR ’22, July 11–15, 2022, Madrid, Spain
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-8732-3/22/07...
https://doi.org/10.1145/3477495.3531975

1 INTRODUCTION

Recommender systems become more and more indispensable in practice since it is a sharp tool for users to easily find what they desire. Nowadays, users always participate in multiple domains (platforms) for different purposes, e.g., buying books on Amazon and watching movies on Youtube. Therefore, Cross Domain Recommendation (CDR) has emerged to utilize both the source and target information for establishing high accuracy recommendation system and overcoming the long-standing data sparsity problem. Most existing CDR models assume the existence of fully overlapped users or items, who have similar tastes or attributions [7], across domains. For example, a user who likes romantic movies and dislikes horror movies would always like romantic books and dislike horror books as well. Although the movie and book domains are rather different in the specific form of expression, the latent user embedding should properly represent the domain-invariant user personal characteristics such as preferring romantic items and disliking horror items, as shown in Fig. 1. Therefore, existing CDR models can leverage the information from a (source) domain to improve the recommendation accuracy of other (target) domains [34, 48]. However, in practice, not all users or items are strictly overlapped in CDR settings, which limits the applications of existing CDR approaches.

In this paper, we focus on a general problem in CDR based on our observation. That is, two or multiple domains only share a certain degree of overlapped users or items. Specifically, we term this problem as Partially Overlapped Cross Domain Recommendation (POCDR). On the one hand, POCDR has some overlapped users who have rating interactions in both the source and target domains, similar
as the conventional CDR problem. On the other hand, POCDR also has several non-overlapped users who only have rating interactions in at most one domain, making it different from the conventional CDR problem. Fig. 2 shows an example of the POCDR problem, where Jimmy and Mary are overlapped users, while John, Tom, Lily, and Rosa are non-overlapped users. The prime task of POCDR is to solve the data sparsity problem by leveraging the user-item interactions, with only partially overlapped users across the source and target domains. Solving POCDR is challenging, mainly because it is hard to fully exploit and transfer useful knowledge across domains, especially when extremely few users are overlapped.

Existing researches on CDR cannot solve the POCDR problem well. Firstly, conventional CDR approaches can only adjust to the situations where the source and target domains share partially overlapped items. For example, Mike and Anne who both like romantic items and dislike horror items should have similar domain-invariant embeddings. In contrast, Mike and Alice have different preferences, and thus their embeddings should be pushed away.

In a nutshell, current CDR models cannot achieve satisfying performance because of the lacking of sufficient techniques to fully leverage the data in both domains under the POCDR setting.

To address the aforementioned issue, in this paper, we propose VDEA, an end-to-end dual-autoencoder with Variational Domain-invariant Embedding Alignment (VDEA), for solving the POCDR problem. The core insight of VDEA is exploiting domain-invariant embedding across different domains, for both the overlapped and non-overlapped users. Again, in the example of Fig. 1, Mike and Anne who share similar tastes, i.e., preferring romantic items rather than horror ones, should have similar domain-invariant embeddings. In contrast, users with different tastes, e.g., Mike and Alice, should be pushed away in the latent embedding space. In order to better represent user embedding with high-quality rating prediction and knowledge transfer ability, we utilize two modules in VDEA, i.e., variational rating reconstruction module and variational embedding alignment module, as will be shown in Fig. 3. On the one hand, the variational rating reconstruction module aims to capture user collaborative preferences in the source and target domains with cross-entropy loss. On the other hand, the variational local embedding alignment module aims to match the users with similar tastes or characteristics on Mixture-Of-Gaussian distribution simultaneously. The variational embedding alignment module further includes two main components, e.g., variational local embedding alignment and variational global embedding alignment. Variational local embedding alignment tends to match the overlapped users across domains to obtain more expressive user embeddings. Variational global embedding alignment can co-cluster the similar user groups via distribution optimal transport, transferring useful knowledge among both overlapped and non-overlapped users. By doing this, we can not only obtain discriminative domain-invariant user embedding in a single domain, but also cluster users with similar preferences between the source and target domains.

We summarize our main contributions as follows: (1) We propose a novel framework, i.e., VDEA, for the POCDR problem, which can utilize the rating interactions of both overlapped and non-overlapped users to improve the recommendation performance. (2) In order to better exploit users with similar tastes or characteristics, we equip VDEA with variational inference and distribution co-clustering optimal transport to obtain more discriminative and domain-invariant latent user embedding. (3) Extensive empirical studies on Douban and Amazon datasets demonstrate that VDEA significantly improves the state-of-the-art models, especially under the POCDR setting.

2 RELATED WORK

Cross Domain Recommendation. Cross Domain Recommendation (CDR) emerges as a technique to alleviate the long-standing data sparsity problem in recommendation by assuming the same user set across domains [18, 23, 33, 41]. According to [48], existing CDR models have three main types, i.e., transfer-based methods, clustered-based methods, and multitask-based methods. Transfer-based methods [26, 46] learn a linear or nonlinear mapping function across domains. Some recent method [4] even adopts equivalent transformation to obtain more reliable knowledge across domains.
with shared users. Clustered-based methods [40] adopt the co-clustering approach to learn cross-domain comprehensive embeddings by collectively leveraging single-domain and cross-domain information within a unified framework. Multi-task-based methods [12, 47, 49] enable dual knowledge transfer across domains by introducing shared connection modules in neural networks. However, the majority of current CDR frameworks have the underlying assumption that the source and target domains share the same set of users or items [48]. Conventional CDR approaches cannot perfectly solve the POCDR problem where a large number of users are non-overlapped. As a comparison, in this paper, we focus on the POCDR problem, and propose VDEA which adopt variational local and global embedding alignment for better exploiting domain-invariant embeddings for all the users across domains.

Domain Adaptation. Domain adaptation aims to transfer useful knowledge from the source samples with labels to target samples without labels for enhancing the target performance. Existing work on this is mainly of three types, i.e., discrepancy-based methods [3, 28], adversarial-based methods [9, 24], and reconstruction-based methods [32, 50]. Discrepancy-based methods, e.g., Maximum Mean Discrepancy (MMD) [3], Correlation Alignment (CORAL). Recently, ESAM [5] extends CORAL with attribution alignment for solving the long-tailed item recommendation problem. Adversarial-based methods integrate a domain discriminator for adversarial training, including Domain Adversarial Neural Network (DANN) [9] and Adversarial Discriminative Domain Adaptation (ADDA) [36]. DARec [44] even transferred knowledge across domains with shared users via deep adversarial adaptation technique. Reconstruction-based methods utilize autoencoder to exploit the domain-invariant latent representations, e.g., Unsupervised Image-to-Image Translation Networks (UNIT) [13, 22], which indicates that exploiting shared latent space across different domains can enhance the model performance and generalization. However, most existing reconstruction-based models fail to adopt a dual discriminator network to reduce the distribution discrepancy across different domains. In this paper, we propose to combine reconstruction-based model with discrepancy-based method for better exploiting domain-invariant embeddings.

3 MODELING FOR VDEA

3.1 Framework of VDEA

First, we describe notations. We assume there are two domains, i.e., a source domain $\mathcal{S}$ and a target domain $\mathcal{T}$. We assume $\mathcal{S}$ has $|\mathcal{U}^S|$ users and $|\mathcal{V}^S|$ items, and $\mathcal{T}$ has $|\mathcal{U}^T|$ users and $|\mathcal{V}^T|$ items. Unlike conventional cross domain recommendation, in POCDR, the source and target user sets are not completely the same. We use the overlapped user ratio $\mathcal{K}_u \in (0, 1)$ to measure how many users are concurrence. Bigger $\mathcal{K}_u$ means more overlapped users across domains, while smaller $\mathcal{K}_u$ means the opposite situation.

Then, we introduce the overview of our proposed VDEA framework. The aim of VDEA is providing better single domain recommendation performance by leveraging useful cross domain knowledge for both overlapped and non-overlapped users. VDEA model mainly has two modules, i.e., variational rating reconstruction module and variational embedding alignment module, as shown in Fig. 3. The variational rating reconstruction module establishes user embedding according to the user-item interactions on the source and target domains. Meanwhile, the variational rating reconstruction module also clusters the users with similar characteristics or preferences based on their corresponding latent embeddings to enhance the model generalization. The basic intuition is that users with similar behaviors should have close embedding distances through this module with deep unsupervised clustering approach [25, 39, 40]. The variational embedding alignment module has two parts, i.e., variational local embedding alignment and variational global embedding alignment, sufficiently aligning the local and global user embeddings across domains. The local embedding alignment can utilize the user-item ratings of the overlapped users to obtain more expressive user embeddings [16]. Intuitively, expressive overlapped user embedding can enhance model training of the non-overlapped users, since the non-overlapped users share the similar preference (e.g., romantic or realistic) as the overlapped ones [45]. Moreover, the global embedding alignment can also directly co-cluster both the overlapped and non-overlapped users with similar tastes, which can boost the procedure of knowledge transfer. To this end, the variational embedding alignment module can exploit domain-invariant latent embeddings for both overlapped and non-overlapped users. The framework of VDEA is demonstrated in Fig. 3 and we will introduce its two modules in details later.
Variational evidence lower bound. To begin with, we first introduce the variational rating reconstruction module. We define \( X^S \in \mathbb{R}^{N \times |V^S|} \) and \( X^T \in \mathbb{R}^{N \times |V^T|} \) as the user-item interactions on the source and target domains respectively, and \( N \) is batch size. We adopt dual variational autoencoder as the main architecture for accurately representing and clustering user embedding according to the rating interactions. We utilize two variational autoencoders in VDEA, i.e., a source variational autoencoder and a target variational autoencoder, for source and target domains, respectively. The source variational autoencoder includes the encoder \( E_S \) and decoder \( D_S \). Similarly, the target variational autoencoder includes the encoder \( E_T \) and decoder \( D_T \). The encoders \( E_S \) and \( E_T \) can predict the mean \( (\mu^S, \sigma^S) \) and log variance \( (\log(\sigma^S)^2) \) and \( (\log(\sigma^T)^2) \) for the input data as \( \mu^S, \sigma^S \) and \( \mu^T, \sigma^T \). In order to guarantee gradient descent in back propagation, we also apply the reparameterization mechanism [21, 30] as \( Z^S = \mu^S + \epsilon^S \sigma^S \) and \( Z^T = \mu^T + \epsilon^T \sigma^T \), where \( \epsilon^S, \epsilon^T \sim N(0,1) \) are randomly sampled from normal distribution, and \( \mu^S, \mu^T, \sigma^S, \) and \( \sigma^T \in \mathbb{R}^{N \times D} \) have the same dimension \( D \). Inspired by variational deep embedding [14], we expect the latent user embedding conform to Mixture-Of-Gaussian distribution for grouping similar users. The dual variational autoencoder also reconstructs the user-item interaction as \( \hat{X}^S = D_S(Z^S) \) and \( \hat{X}^T = D_T(Z^T) \). In this section, we will first present the variational evidence lower bound according to the Mixture-Of-Gaussian distributions, and then illustrate the rating reconstruction and distribution matching for the total loss. By proposing the variational rating reconstruction module, the users with similar characteristics will be clustered in each domain, as can be seen in Fig. 4(a).

Variational evidence lower bound. Since we tend to cluster the users with similar preferences, we choose Mixture-Of-Gaussian as the user latent distribution. To do this, we let user embedding obey the Mixture-Of-Gaussian distribution and properly predict the user-item rating interactions through variational evidence lower bound [14]. The variational evidence lower bound for the variational rating reconstruction module is to maximize the likelihood of the given user embeddings in Mixture-Of-Gaussian distribution. Specifically, we define the prior source and target latent distributions as \( \mathbb{P}^S = \text{GMM}(\pi^S, \mu^S, (\sigma^S)^2) \) and \( \mathbb{P}^T = \text{GMM}(\pi^T, \mu^T, (\sigma^T)^2) \) respectively, where \( \text{GMM}(\cdot, \cdot, \cdot) \) denotes the distribution of Mixture-Of-Gaussian. \( \pi^S \) and \( \pi^T \) denote the prior probability of each cluster components in Mixture-Of-Gaussian. \( (\mu^S, (\sigma^S)^2) \) and \( (\mu^T, (\sigma^T)^2) \) denote the mean and variance of the Gaussian distribution in the source domain and target domain, respectively. For the \( i \)-th user, \( \sum^K c_{ik} = 1 \) for \( K \) denoting the total number of clusters. Taking the joint probability in the source domain for example, it can be factorized as \( p_S(X^S, Z^S, C^S) \) in the perspective of the generative process:

\[
p_S(X^S, Z^S, C^S) = p_S(X^S|Z^S)p_S(Z^S|C^S)p_S(C^S) \\
p_S(C^S) = \text{Cat}(C^S|\pi^S) \\
p_S(Z^S|C^S) = N(\tilde{\mu}^S_i, (\tilde{\sigma}^S_i)^2) \\
p_S(X^S|Z^S) = \text{Ber}(X^S|\mu^S_i) 
\]

where \( X^S \) is the observed sample and \( Z^S \) is the latent user embedding. The corresponding cluster is \( C^S \sim \text{Cat}(\pi^S) \) with \( \text{Cat}(\cdot) \) denoting the prior categorical distribution. \( \tilde{\mu}^S_i \) and \( (\tilde{\sigma}^S_i)^2 \) denote the mean and variance of the Gaussian distribution to the \( c \)-th cluster. \( \text{Ber}(X^S|\mu^S_i) \) denotes multivariate Bernoulli distribution for the output of the decoder.

We then adopt Jensen’s inequality with the mean-field assumption to estimate the Evidence Lower Bound (ELBO) by maximizing \( \log p_S(X^S) \). Following [14], ELBO can be rewritten in a simplified form as:

\[
\mathcal{L}_{\text{ELBO}}(X^S) = \mathbb{E}_{q_S(Z^S,C^S|X^S)} \left[ \log \frac{p_S(Z^S,C^S|X^S)}{q_S(Z^S,C^S|X^S)} \right] = \mathbb{E}_{q_S(Z^S,C^S|X^S)} \left[ \log p_S(X^S|Z^S) + \log \frac{p_S(Z^S,C^S)}{q_S(Z^S,C^S|X^S)} \right] ,
\]

where the first term represents the rating reconstruction loss that encourages the model to fit the user-item interactions well. The second term calculates the Kullback-Leibler divergence between Mixture-of-Gaussians prior \( q_S(Z^S,C^S|X^S) \) and the variational posterior \( p_S(Z^S,C^S) \). Furthermore, in order to balance the reconstruction capacity (the first term) efficiently and disentangle the latent factor (the second term) automatically, we change the KL divergence into the following optimization equation with constraints:

\[
\max \mathbb{E}_{q_S(Z^S,C^S|X^S)} \log p_S(X^S|Z^S) \\
t s.t. \text{KL}(q_S(Z^S,C^S|X^S)||p_S(Z^S,C^S)) < \delta \nonumber ,
\]

where \( \delta \) specifies the strength of the KL constraint. Rewriting the optimization problem in Equation (3) under the KKT conditions and neglecting the small number of \( \epsilon \), we have:

\[
\mathcal{L}_{\text{ELBO}}(X^S, \beta) = \mathbb{E}_{q_S(Z^S,C^S|X^S)} \log p_S(X^S|Z^S) - \beta \text{KL}(q_S(Z^S,C^S|X^S)||\mathbb{E}_{q_S(Z^S,C^S|X^S)}[p_S(Z^S,C^S)]) ,
\]

where \( \beta \) denotes the hyper-parameter between rating modelling and latent distribution matching [11].

Loss computing. The loss of the variational rating reconstruction module can be deduced from ELBO, which has two main parts, i.e., rating reconstruction and distribution matching. For the rating reconstruction part, the reconstruction loss is given as below:

\[
\mathbb{E}_{q_S(Z^S,C^S)} \log p_S(X^S|Z^S) = -\frac{1}{N} \sum_{i=1}^{N} \mathcal{F}(X^S_i, X^S_i) ,
\]

where \( \mathcal{F}(\cdot, \cdot) \) is the cross-entropy loss. For the distribution matching part, it can be further substituted through Kullback-Leibler
By the detailed process on the source domain through maximization of ELBO w.r.t. the parameters of \( \{\pi^S, \mu^S, \sigma^S, E_S, D_S\} \), the latent user embeddings will be clustered and become much more discriminative. Likewise, we can also optimize ELBO on the target domain parameters \( \{\pi^T, \mu^T, \sigma^T, E_T, D_T\} \). Therefore, the loss of the whole variational rating reconstruction module is given as:

\[
\min L_{VR} = \min \left[-L_{ELBO}(X^S, \beta) - L_{ELBO}(X^T, \beta)\right]
\]

\[
= \frac{1}{N} \sum_{i=1}^{N} F(x_i^S, x_i^T) + \beta \text{KL}(q_S(Z^S, C^S|X^S)||p_S(Z^S, C^S))
+ \frac{1}{N} \sum_{i=1}^{N} F(x_i^T, x_i^T) + \beta \text{KL}(q_T(Z^T, C^T|X^T)||p_T(Z^T, C^T)).
\]

By the detailed process on the source domain through maximizing ELBO w.r.t. the parameters of \( \{\pi^S, \mu^S, \sigma^S, E_S, D_S\} \), the latent user embeddings will be clustered and become much more discriminative. Likewise, we can also optimize ELBO on the target domain parameters \( \{\pi^T, \mu^T, \sigma^T, E_T, D_T\} \). Therefore, the loss of the whole variational rating reconstruction module is given as:

\[\min L_{VR} = \min \left[-L_{ELBO}(X^S, \beta) - L_{ELBO}(X^T, \beta)\right] = \frac{1}{N} \sum_{i=1}^{N} F(x_i^S, x_i^T) + \beta \text{KL}(q_S(Z^S, C^S|X^S)||p_S(Z^S, C^S)) + \frac{1}{N} \sum_{i=1}^{N} F(x_i^T, x_i^T) + \beta \text{KL}(q_T(Z^T, C^T|X^T)||p_T(Z^T, C^T)).\]

### 3.3 Variational Embedding Alignment Module

As we have motivated before, the embeddings of similar users should be clustered through variational deep embedding to make them more discriminative, as is shown in Fig. 4. However, how to align these Mixture-Of-Gaussian distribution across domains remains a tough challenge. In order to tackle this issue, we propose **variations embedding alignment** from both local and global perspectives.

**Variational local embedding alignment.** Variational local embedding alignment in VDEA aims to align the corresponding overlapped user embedding distribution across domains. Since variational encoder can measure the mean and variance of the input data, we adopt the 2-Wasserstein distance of different Gaussians distributions to calculate the variational local user embedding alignment loss:

\[
L_{VA} = \sum_{i,j=1}^{N} \delta(U_i^S, U_j^T) d_W(N(\mu_i^S, (\sigma_i^S)^2), N(\mu_j^T, (\sigma_j^T)^2))
= \sum_{i,j=1}^{N} \delta(U_i^S, U_j^T) \left(||\mu_i^S - \mu_j^T||_2^2 + ||\sigma_i^S - \sigma_j^T||_2^2\right),
\]

where \( d_W(\cdot, \cdot) \) denotes the 2-Wasserstein distance of different Gaussians distributions that measures the discrepancy across domains, \( \delta(\cdot) \) indicates whether a user occurs simultaneously in both domains under the POCDR setting. \( \delta(U_i^S, U_j^T) = 1 \) means the \( i \)-th user in the source domain has the same identification with the \( j \)-th user in the target domain, otherwise \( \delta(U_i^S, U_j^T) = 0 \). The distribution estimation can further help the model to better capture the uncertainties of the learned embeddings [35].

**Variational global embedding alignment.** VDEA should also align the global distributions of both the overlapped and non-overlapped users across domains. Similar intra-domain users can be clustered through variational local embedding alignment, however, there is still a distribution discrepancy between the source and target domains. Thus, how to match the whole user distributions across domains without explicit label information remains the key issue.

In order to obtain the alignment results effectively, we propose the *Gromov-Wasserstein Distribution Co-Clustering Optimal Transport (GDOT)* approach. Unlike conventional optimal transport [29, 37] that only aligns user embeddings, our proposed GDOT for domain adaptation performs user embedding distribution alignment in the source and target domains. The optimization of optimal transport is based on the Kantorovich problem [1], which seeks a general coupling \( \psi \in \mathcal{X}(S, T) \) between \( S \) and \( T \):

\[
\psi^* = \arg \min_{\psi \in \mathcal{X}(S, T)} \int_{S \times T} M(z_i^S, z_j^T) d\psi(z_i^S, z_j^T) d\psi(z_i^S, z_j^T),
\]

where \( \mathcal{X}(S, T) \) denotes the latent embedding probability distribution between \( S \) and \( T \). The cost function matrix \( M(z_i^S, z_j^T) \) measures the difference between each pair of distributions. To facilitate calculation, we formulate the discrete optimal transport formulation as \( \psi^* = \arg \min_{\phi \in \mathcal{P}(S \times T)} \frac{1}{2} \int_{S \times T} H(\psi^*) d\psi(\mathcal{P}(z_i^S, z_j^T)) \), where \( \mathcal{P}(S \times T) \) is the ideal coupling matrix between the source and target domains, representing the joint probability of the source probability \( \mathcal{P}(S \times T) \) and the target probability \( \mathcal{P}(T) \). The matrix \( \mathcal{M} \in \mathbb{R}^{K \times K \times K \times K} \) denotes a pairwise Gaussian distribution distance among the source probability \( \mathcal{P}(S \times T) \) and target probability \( \mathcal{P}(T) \) as:

\[
M(i,j) = d_W(z_i^S, z_j^T) \leq \mathcal{M}(z_i^S, z_j^T) \leq d_W(z_i^S, z_j^T).
\]

Meanwhile \( \otimes \) denotes the tensor matrix multiplication, which means \( \mathcal{M} \otimes \mathcal{P} = (\sum_{i,j} M_{i,j} \mathcal{P}_{i,j})_{i,j} \). The second term \( H(\psi^*) = \sum_{i,j=1}^{K} \psi_{ij} (\log(\psi_{ij}) - 1) \) is a regularization term and \( \epsilon \) is a balance hyper parameter. The matching matrix follows the constraints that \( \psi_{1K} = \psi_{K1} = \psi_{1T} = \psi_{T1} = 0 \) with \( K \) denoting the number of clusters. We apply the Sinkhorn algorithm [6, 42] for solving the optimization problem by adopting the Lagrangian multiplier to minimize the objective function \( f \) as below:

\[
f = \langle \mathcal{M} \otimes \psi, \psi \rangle - \epsilon H(\psi) - f(\psi_{1K} - \psi_{K1}) - g(\psi_{1T} - \psi_{T1}).
\]

where \( f \) and \( g \) are both Lagrangian multipliers. We first randomly generate a matrix \( \psi \) that meets the boundary constraints and then define the variable \( \tilde{G} = \mathcal{M} \otimes \psi \). Taking the differentiation of Equation (12) on variable \( \psi \), we can obtain:

\[
\psi_{ij} = \exp \left( \frac{\tilde{G}_{ij}}{\epsilon} \right) \exp \left( -\frac{\tilde{G}_{ij}}{\epsilon} \right) \exp \left( \frac{\psi_{ij}}{\epsilon} \right).
\]

Taking Equation (13) back to the original constraints \( \psi_{1K} = \psi_{K1} = \psi_{1T} = \psi_{T1} = 0 \), we can obtain that:

\[
u \otimes (\mathcal{H} \psi) = \psi \otimes \mathcal{H} T = \psi T.
\]
Since we define $u = \text{diag}(\frac{\mu_s}{c}), H = \exp\left(-\frac{uH}{c}\right)$, and $v = \text{diag}(\frac{\mu_t}{c})$. Through iteratively solving Equation (14) on the variables $u$ and $v$, we can finally obtain the coupling matrix $\pi^*$ on Equation (13). We show the iteration optimization scheme of GDOT in Algorithm 1. In summary, variational embedding alignment can obtain the optimal transport solution as $q^*$ for each corresponding sub-distribution, as shown with dot lines in Fig. 4(b).

Algorithm 1 The optimization scheme of GDOT.

1: Input: $K$: Number of clusters; $N(\tilde{\mu}_c^S, (\tilde{\sigma}_c^S)^2)$: The source user latent distribution for the $c$-th cluster group; $N(\tilde{\mu}_c^T, (\tilde{\sigma}_c^T)^2)$: The target user latent distribution for the $c$-th cluster group; $\varepsilon$: Hyper parameter to balance the matching loss and the entropy maximization constraint.
2: Initialize the coupling matrix $\pi$.
3: for $i = 1$ to $T$ do
4: Calculate the distance between different Normal distribution in source and target domains as matrix $M \in \mathbb{R}^{K \times K \times K}$ through Equation (11).
5: Calculate matrix $G = M \otimes \pi$ and $H = \exp\left(-\frac{G}{\varepsilon}\right)$.
6: Randomly initialize vectors $u \in \mathbb{R}^K$ and $v \in \mathbb{R}^K$.
7: for $j = 1$ to $t$ do
8: Update $u \leftarrow \pi^S \otimes (Hu)$.
9: Update $v \leftarrow \pi^T \otimes (H^T u)$.
10: end for
11: $\pi^* = \text{diag}(u)H\text{diag}(v)$.
12: end for
13: Return: $\pi^* = \pi$.

After that, the variational global embedding alignment loss is given as:

$$L_{VG} = \sum_{i=1}^{K} \sum_{j=1}^{K} \psi_{ij}^*d_W \left(\left|N(\tilde{\mu}_i^S, (\tilde{\sigma}_i^S)^2), N(\tilde{\mu}_j^T, (\tilde{\sigma}_j^T)^2)\right|\right).$$  \hspace{1cm} (15)

The bigger $\psi_{ij}^*$ for the matching constraint, the closer between the distributions $N(\tilde{\mu}_i^S, (\tilde{\sigma}_i^S)^2)$ and $N(\tilde{\mu}_j^T, (\tilde{\sigma}_j^T)^2)$.

3.4 Putting Together

The total loss of VDEA could be obtained by combining the losses of the variational rating reconstruction module and the variational embedding alignment module. That is, the losses of VDEA is given as:

$$L_{\text{VDEA}} = L_{VR} + \lambda_{VL}L_{VA} + \lambda_{VG}L_{VG}.$$  \hspace{1cm} (16)

where $\lambda_{VL}$ and $\lambda_{VG}$ are hyper-parameters to balance different type of losses. By doing this, users with similar rating behaviors will have close embedding distance in single domain through the variational inference, and the user embeddings across domains will become domain-invariant through GDOT which has been shown in Fig. 4(c).

4 EMPIRICAL STUDY

In this section, we conduct experiments on several real-world datasets to answer the following questions: (1) RQ1: How does our approach perform compared with the state-of-the-art single-domain or cross-domain recommendation methods? (2) RQ2: How do the local and global embedding alignments contribute to performance improvement? Can proposed DCOT better align the global distribution? (3) RQ3: How does the performance of VDEA vary with different values of the hyper-parameters?

4.1 Datasets and Tasks

We conduct extensive experiments on two popularly used real-world datasets, i.e., Douban and Amazon, for evaluating our models on POCDIR tasks [27, 47, 49]. First, the Douban dataset [47, 49] has two domains, i.e., Book and Movie, with user-item ratings. Second, the Amazon dataset [27, 46] has four domains, i.e., Movies and TV (Movie), Books (Book), CDs and Vinyl (Music), and Clothes (Clothes). The detailed statistics of these datasets after pre-process will be shown in Table 1. For both datasets, we binarize the ratings to 0 and 1. Specifically, we take the ratings higher or equal to 4 as positive and others as 0. We also filter the users and items with less than 5 interactions, following existing research [44, 49]. We provide four main tasks to evaluate our model, i.e., Amazon Movie & Amazon Music, Amazon Movie & Amazon Book, Amazon Movie & Amazon Clothes, and Douban Movie & Douban Book. It is noticeable that some tasks are rather simple because the source and target domains are different, e.g., Amazon Movie & Amazon Book, while some tasks are difficult since these two domains are different, e.g., Amazon Movie & Amazon Clothes [43].

4.2 Experimental Settings

We randomly divide the observed source and target data into training, validation, and test sets with a ratio of 8:1:1. Meanwhile, we vary the overlapped user ratio $\mathcal{K}_u$ in [30%, 60%, 90%]. Different user overlapped ratio represents different situations, e.g., $\mathcal{K}_u = 30\%$ represents only few users are overlapped while $\mathcal{K}_u = 90\%$ means most of users are overlapped [45]. For all the experiments, we perform five random experiments and report the average results. We choose Adam [15] as optimizer, and adopt Hit Rate@k (HR@k) and NDCG@k [38] as the ranking evaluation metrics with $k = 5$.

4.3 Baseline

We compare our proposed VDEA with the following state-of-the-art cross domain recommendation models. (1) KerKT [45] Kernel-induced knowledge transfer for overlapping entities (KerKT) which is the first attempt on POCDIR problem. It uses a shallow model with multiple steps to model and align user representations. (2) NeuMF [10] Neural Collaborative Filtering for Personalized Ranking (NeuMF) replaces the traditional inner product with a neural network to learn an arbitrary function on the single domain accordingly. (3) MultiVAE [21] Variational Autoencoders for Collaborative Filtering (MultiVAE) extends variational autoencoders for
We set batch size $N = 256$ for both the source and target domains. The latent embedding dimension is set to $D = 128$. We assume both source and target users share the same embedding dimension according to previous research [4, 22]. For the variational rating reconstruction module, we adopt the heuristic strategy for selecting the disentangle factor $\beta$ by starting from $\beta = 0$ and gradually increasing to $\beta = 0.2$ through KL annealing [21]. Empirically, we set the cluster number as $K = 30$. Meanwhile, we set hyper-parameters $\epsilon = 0.1$ for Gromov-Wasserstein Distribution Co-Clustering Optimal Transport. For VDEA model, we set the balance hyper-parameters as $\lambda_L = 0.7$ and $\lambda_G = 1.0$.

### 4.5 Recommendation Performance (for RQ1)

The comparison results on Douban and Amazon datasets are shown in Table 1. From them, we can find that: (1) With the decrease of the overlapped user ratio $K_u$, the performance of all the models will also decrease. This phenomenon is reasonable since fewer overlapped users may raise the difficulty of knowledge transfer across different domains. (2) The conventional shallow model KerKT with multiple separated steps cannot model the complex and nonlinear user-item relationship under the POCDR setting. (3) Most of the

---

**Table 2: Experimental results on Amazon and Douban datasets.**

| (Amazon) Movie & Music | (Amazon) Movie & Clothes | (Amazon) Movie & Book | (Douban) Movie & Book |
|------------------------|-------------------------|-----------------------|-----------------------|
| **$K_u = 30\%$**       | **$K_u = 60\%$**        | **$K_u = 90\%$**      |                       |
| Movie                  | Book                    | Movie                 | Book                  |
| KerKT                  | 0.3507                  | 0.3534                | 0.3903                |
| NDCG                   | 0.2705                  | 0.2559                | 0.3013                |
| NeuMF                  | 0.3722                  | 0.3774                | 0.4155                |
| NDCG                   | 0.2638                  | 0.2672                | 0.3021                |
| MultVAE                | 0.3878                  | 0.3940                | 0.4239                |
| NDCG                   | 0.2763                  | 0.2728                | 0.3145                |
| BiVAE                  | 0.3983                  | 0.3992                | 0.4246                |
| NDCG                   | 0.2852                  | 0.2837                | 0.3289                |
| DARec                  | 0.3947                  | 0.3918                | 0.4364                |
| NDCG                   | 0.2816                  | 0.2783                | 0.3230                |
| ETL                    | 0.4022                  | 0.4075                | 0.4269                |
| NDCG                   | 0.2959                  | 0.2938                | 0.3297                |
| DML                    | 0.4070                  | 0.4013                | 0.4252                |
| NDCG                   | 0.3073                  | 0.2934                | 0.3389                |
| VDEA                   | 0.4481                  | 0.4342                | 0.4813                |
| NDCG                   | 0.3338                  | 0.3206                | 0.3581                |

**Table 3: Experimental results on Amazon and Douban datasets.**

| (Amazon) Movie & Music | (Amazon) Movie & Clothes | (Amazon) Movie & Book | (Douban) Movie & Book |
|------------------------|-------------------------|-----------------------|-----------------------|
| **$K_u = 30\%$**       | **$K_u = 60\%$**        | **$K_u = 90\%$**      |                       |
| Movie                  | Book                    | Movie                 | Book                  |
| KerKT                  | 0.2540                  | 0.2559                | 0.2947                |
| NDCG                   | 0.2594                  | 0.2572                | 0.2858                |
| NeuMF                  | 0.2662                  | 0.2686                | 0.2944                |
| NDCG                   | 0.2602                  | 0.2686                | 0.2944                |
| MultVAE                | 0.2807                  | 0.2861                | 0.3019                |
| NDCG                   | 0.2654                  | 0.2758                | 0.2995                |
| BiVAE                  | 0.2880                  | 0.2956                | 0.3104                |
| NDCG                   | 0.2754                  | 0.2836                | 0.3109                |
| DARec                  | 0.2831                  | 0.2898                | 0.4055                |
| NDCG                   | 0.2699                  | 0.2740                | 0.3038                |
| ETL                    | 0.2893                  | 0.4002                | 0.4160                |
| NDCG                   | 0.2751                  | 0.3145                | 0.3415                |
| DML                    | 0.3941                  | 0.4064                | 0.4182                |
| NDCG                   | 0.2763                  | 0.2934                | 0.3156                |
| VDEA                   | 0.4404                  | 0.4457                | 0.4602                |
| NDCG                   | 0.3021                  | 0.3207                | 0.3322                |

---
Table 3: Ablation results on Amazon datasets.

| Algorithm | HR (Movie) | Music | Book | Clothes |
|-----------|------------|-------|-------|---------|
| VDEA-Base | 0.4404     | 0.4456| 0.4956| 0.5147  |
|           | 0.3021     | 0.3207| 0.3659| 0.3966  |
| VDEA-Local| 0.4217     | 0.4303| 0.4764| 0.5034  |
|           | 0.2889     | 0.3076| 0.3561| 0.3850  |
| VDEA-Global| 0.4330    | 0.4315| 0.4913| 0.5082  |
|           | 0.2936     | 0.3134| 0.3614| 0.3901  |
| VDEA      | 0.4330     | 0.4315| 0.4913| 0.5082  |
|           | 0.2936     | 0.3134| 0.3614| 0.3901  |

Figure 5: The t-SNE visualization of user latent embeddings on Amazon Movie & Book (first row) and Douban Movie & Book (second row). The user latent embeddings in the source domain are shown with red dots and those in the target domain are shown with blue dots.

4.6 Analysis (for RQ2 and RQ3)

Ablation. To study how does each module of VDEA contribute to the final performance, we compare VDEA with its several variants, including VDEA-Base, VDEA-Local, and VDEA-Global. (1) VDEA-Base only includes the variational rating reconstruction loss $L_{VR}$, that is $\lambda_{VL} = \lambda_{VG} = 0$. (2) VDEA-Local excludes global embedding alignment for user embedding, that is $\lambda_{VG} = 0$. (3) VDEA-Global adopts local embedding alignment while utilizing the adversarial domain alignment method in DARec for the global embedding alignment. The comparison results are shown in Table 3. From it, we can observe that: (1) VDEA-Base can already achieve good results compared with some other models (e.g., MultVAE), indicating that clustering the users with similar preference inside a single domain can boost the performance. However, due to the lacking of knowledge transfer, it cannot exceed the latest cross-domain recommendation model (e.g., DML). (2) VDEA-Local and VDEA-Global achieve better results than VDEA-Base, showing that adopting the variational embedding alignment module has the positive effect on solving the POCDR problem. Aligning the overlapped users can enhance the representation ability for the user embeddings which is also conforms to our intuitions. However, cross domain recommendation models (e.g., DML) perform better than single domain recommendation model (e.g., NeuMF), indicating that cross domain recommendation can integrate more useful information to enhance the model performance, which is always conform to our common sense. (4) Although ETL and DML can get reasonable results on conventional cross domain recommendation problem where two domains are highly overlapped, it cannot achieve satisfied solutions on the general POCDR problem. When the majority of users are non-overlapped, ETL and DML cannot utilize the knowledge transfer from the source domain to the target domain and thus lead to model degradation. The main reason is that their loss functions only act on the overlapped users without considering the non-overlapped users. The same phenomenon happens on other baseline models, e.g., DARec. (5) VDEA with deep latent embedding clustering can group users with similar behaviors and can further enhance the performance under the POCDR settings. We also observe that our proposed VDEA can also have great prediction improvement even when the overlapped user ratio $K_u$ is much smaller (e.g., $K_u = 30\%$). Meanwhile our proposed VDEA still obtains better results when the source and target domains are different (e.g., Amazon Movie $\rightarrow$ Amazon Clothes) which indicates the high efficacy of VDEA.
VDEA-Local can only transfer the knowledge on the overlapped users and thus its performance is limited, especially when the overlapped ratio is relatively small. VDEA-Global with the adversarial alignment based on DARec still outperforms than VDEA-Local indicates that global distribution alignment is important and essential. Nevertheless, generative adversarial network with discriminator is unstable and hard to train in practice which hinders the model performance[31]. (3) By complementing both local and global alignment, VDEA can further promote the performance. Meanwhile, we can conclude that our proposed distribution co-clustering optimal transport method is also much better than traditional distribution alignment (e.g., VDEA-Global with DARec). Overall, the above ablation study demonstrates that our proposed embedding alignment module is effective in solving the POCDR problem.

Visualization and discrepancy measurement. To better show the user embeddings across domains, we visualize the t-SNE embeddings [17] for VDEA-Base, VDEA-Local, and VDEA. The results of Amazon Movie & Book are shown in the first row of Fig. 5(a)-(c) and the results of Douban Movie & Book are shown in the second row of Fig. 5(a)-(c). From it, we can see that: (1) VDEA-Base cannot bridge the user gap across different domains, leading to insufficient knowledge transfer, as shown in Fig. 5(a). (2) VDEA-Local can align some overlapped users across domains while it cannot fully minimize the global user preference discrepancy, as shown in Fig. 5(b). (3) VDEA with local and global embedding distribution alignment can better match and cluster the source and target users with similar characteristics, as shown in Fig. 5(c). The visualization result is consistent with Fig. 4(a)-(c), which illustrates the validity of our model. Moreover, we adopt $d_{KL}(S, T) = 2(1 − 2e(h))$ to analyze the distance between two domains, where $e(h)$ is the generalization error of a linear classifier $h$ which discriminates the source domain $S$ and the target domain $T$ [2]. Table 4 demonstrates the domain discrepancy on Amazon Movie & Amazon Book and Douban Movie & Douban Book tasks using VDEA and several baseline methods. From it, we can conclude that adopting both local and global alignment in VDEA can better reduce the domain discrepancy between the source and target domains.

Effect of hyper-parameters. We finally study the effects of hyper-parameters on model performance, including $\lambda_{VL}$, $\lambda_{VG}$, $K$, and $D$. We first conduct experiments to study the effects of $\lambda_{VL}$ and $\lambda_{VG}$ by varying them in $\{0.1, 0.3, 0.5, 0.7, 1, 3, 5, 10\}$ and report the results in Fig. 6(a)-(b). When $\lambda_{VL}, \lambda_{VG} \to 0$, the alignment loss cannot produce the positive effect. When $\lambda_{VL}$ and $\lambda_{VG}$ become too large, the alignment loss will suppress the rating reconstruction loss, which also decreases the recommendation results. Empirically, we choose $\lambda_{VL} = 0.7$ and $\lambda_{VG} = 1.0$. We then analyse the effect of the cluster number by varying $K = \{10, 20, 30, 40, 50\}$ on Amazon Music & Movie ($\mathcal{K}_a = 30\%$ , $\mathcal{K}_d = 90\%$), and report the result in Fig. 6(c). We can conclude that more clusters will enhance the model performance but it will naturally consume more time and space as well. In order to balance the model performance and operation efficiency, we recommend setting $K = 30$ in practice. In order to investigate how the latent user embedding dimension affect the model performance, we finally perform experiment by varying dimension $D$ on VDEA. The result on Amazon Music & Movie ($\mathcal{K}_a = 30\%$ , $\mathcal{K}_d = 90\%$) is shown in Fig. 6(d), where we range $D$ in $\{8, 16, 32, 64, 128\}$. From it, we can see that, the recommendation accuracy of VDEA increases with $D$, which indicates that a larger embedding dimension can provide more accurate latent embeddings for both users and items.

5 CONCLUSION

In this paper, we first propose Variational Dual autoencoder with Domain-invariant Embedding Alignment (VDEA), which includes the variational rating reconstruction module and the variational embedding alignment module, for solving the Partially Overlapped Cross-Domain Recommendation (POCDR) problem. We innovatively propose local and global embedding alignment for matching the overlapped and non-overlapped users across domains. We further present Mixture-Of-Gaussian distribution with Gromov-Wasserstein Distribution Co-clustering Optimal Transport (GDOT) to group the users with similar behaviors or preferences. We also conduct extensive experiments to demonstrate the superior performance of our proposed VDEA models. In the future, we plan to extend VDEA to item-based POCDR tasks and conduct more comprehensive experiments on new datasets. We are also interested in taking side-information into consideration to better solve the POCDR problem.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (No.72192823 and No.62172362) and Leading
Expert of "Ten Thousands Talent Program" of Zhejiang Province (No.2021R52001).

REFERENCES

[1] Sigurd Angenent, Steven Haker, and Allen Tannenbaum. 2003. Minimizing flows for the Monge–Kantorovich problem. SIAM journal on mathematical analysis 35, 1 (2003), 61–97.

[2] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2007. Anal-

ysis of representations for domain adaptation. In Advances in neural information processing systems. 137–144.

[3] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bern-

hard Scholkopf, and Alex J Smola. 2006. Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics 22, 14 (2006), e49–e57.

[4] Xu Chen, Ya Zhang, Ivor Tsang, Yuangang Pan, and Jingchao Su. 2020. Towards Equiva-

lent Transformation of User Preferences in Cross Domain Recommendation. arXiv preprint arXiv:2009.06884 (2020).

[5] Zhihong Chen, Rong Xiao, Chenliang Li, Gangfeng Ye, Haochuan Sun, and Hongbo Deng. 2020. Exam. Discriminative domain adaptation with non-displayed items to improve long-tail performance. In SIGIR 579–588.

[6] Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems 26 (2013), 2292–2300.

[7] Maurizio Ferrari Dacrema, Iván Cantador, Ignacio Fernández-Tobías, Shlomo Berko-

vsky, and Paolo Cremonesi. [n. d.]. Design and Evaluation of Cross-domain Recommender Systems. (n. d.).

[8] Aleksandr Farseev, Ivan Samborskii, Andrey Filchenkov, and Tat-Seng Chua. 2017. Unsupervised image-to-image translation. In Proceedings of the European conference on computer vision (ECCV). 172–189.

[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. The journal of machine learning research 17, 1 (2016), 2096–2030.

[10] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In Proceedings of the 26th international conference on machine learning on computer vision. 9, 2605 (2008), 2579–2605.

[11] D. Kingma and J. Ba. 2014. Adam: A Method for Stochastic Optimization. In Proceedings of the 31st International Conference on Machine Learning (ICML-14). PMLR, 10576–10586.

[12] Cédric Villani. 2008. Transport咙guns. Berlin: Springer Science & Business Media.

[13] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discrimination domain adaptation. In CVPR. 7167–7176.

[14] Zhihong Chen, Rong Xiao, Chenliang Li, Gangfeng Ye, Haochuan Sun, and Hongbo Deng. 2020. Exam. Discriminative domain adaptation with non-displayed items to improve long-tail performance. In SIGIR 579–588.

[15] Van Der Maaten Laurens and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. Journal of Machine Learning Research 9, 2605 (2008), 2579–2605.

[16] Optimal transport: old and new. Vol. 137. Springer Science & Business Media.

[17] Qin Zhang, Jie Huang, Wei Lu, Tai-Sung Chua. 2017. Domain-adversarial training of neural networks. The journal of machine learning research 17, 1 (2016), 2096–2030.

[18] Hongbo Deng. 2020. Esam: Discriminative domain adaptation with non-displayed items to improve long-tail performance. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 950–958.

[19] Cédric Villani. 2008. Transport咙guns. Berlin: Springer Science & Business Media.

[20] Hongfeng Xu, Dixin Luo, Ricardo Henao, Svatl Hall, and Lawrence Carin. 2020. Learning autoencoders with relational regularization. In International Conference on Machine Learning. PMLR, 10576–10586.

[21] Fengyuan Yuan, Xin Rong, and Xiaolin Zheng. 2019. Domain adaptation via transfer component analysis. The journal of machine learning research 188–197. https://doi.org/10.18653/v1/D19-1018

[22] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2007. Anal-

ysis of representations for domain adaptation. In Advances in neural information processing systems. 137–144.

[23] Tomohiro Kato, Shinji Matsubara, and Yu Ishii. 2019. A study of feature alignment in cross-domain recommendation. SIGIR '22, July 11–15, 2022, Madrid, Spain

[24] Shengliang Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. 2019. Learning disentangled representations for recommendation. arXiv preprint arXiv:1910.14238 (2019).

[25] Hongbo Deng. 2020. Esam: Discriminative domain adaptation with non-displayed items to improve long-tail performance. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 950–958.