Robust and graph regularised non-negative matrix factorisation for heterogeneous co-transfer clustering

Yu Ma, Zhikui Chen*, Xiru Qiu and Liang Zhao

School of Software Engineering,
Dalian University of Technology,
Dalian 116620, China
Email: myaue@mail.dlut.edu.cn
Email: zkchen@dlut.edu.cn
Email: brx5qq@mail.dlut.edu.cn
Email: matthew1988zhao@mail.dlut.edu.cn
*Corresponding author

Abstract: Transferring learning is proposed to tackle the problem where target instances are scarce to train an accurate model. Most existing transferring learning algorithms are designed for supervised learning and cannot obtain transferring results on multiple heterogeneous domains simultaneously. Moreover, the performance of transfer learning can be seriously degraded with the appearance of noises and corruptions. In this paper, a robust non-negative collective matrix factorisation model is proposed for heterogeneous co-transfer clustering which introduces the error matrices to capture the sparsely distributed noises. The heterogeneous clustering tasks are handled simultaneously and the graph regularisation is enforced on the collective matrix factorisation model to keep the intrinsic geometric structure of different domains. Experiment results on the real-world dataset show the proposed algorithm outperforms the baselines.

Keywords: transfer learning; non-negative matrix factorisation; NMF; error matrix; graph regularisation; clustering.

Reference to this paper should be made as follows: Ma, Y., Chen, Z., Qiu, X. and Zhao, L. (2019) 'Robust and graph regularised non-negative matrix factorisation for heterogeneous co-transfer clustering', *Int. J. Computational Science and Engineering*, Vol. 18, No. 1, pp.29–38.

Biographical notes: Yu Ma received his BS in Software Engineering from the Dalian University of Technology in 2015. Now, he is a Master student at the Dalian University of Technology. His interests include data mining and transfer learning.

Zhikui Chen received his BS at the Chongqing Normal University, China, and his PhD degree at the Chongqing University, China. He is currently a Professor at the Dalian University of Technology. His current research includes internet of things and big data. He has published more than 20 papers in *IEEE Transactions on Computers, IEEE Transactions on Services Computing*, and so on. He served as a general program chair of the 2015 IEEE Symposium on Smart Data, and 2016 International Conference on Smart X. Moreover, he is an Associate Editor of *International Journal of Communication Systems*.

Xiru Qiu received her BS in Software Engineering from the Dalian University of Technology in 2015. Now, she is a Master student at the Dalian University of Technology. Her interests include data mining, transfer learning and reinforcement learning.

Liang Zhao received his BS and MS at the Dalian University of Technology in 2011. Now, he is a PhD student at the Dalian University of Technology. His interests include data mining and transfer learning. He has published some papers in *IEEE Transactions on Big Data and IEEE Systems Journal*.

1 Introduction

Clustering is a fundamental technique in data mining and machine learning which aims at dividing a set of data into groups according to some similarity or distance strategies. Traditional clustering algorithms such as *k*-means, spectral clustering and non-negative matrix factorisation (NMF) work well when they are provided with a large amount of data in the target domains. However, in real applications, we often encounter the situation that the target instances are insufficient, or the data in some views are scarce (Zhu et al.,

2011). In this case, most existing clustering algorithms fail to learn a good feature representation and will lead to poor performance (Dai et al., 2008b). In order to solve this problem, transfer learning (Pan and Yang, 2010) is proposed to borrow knowledge from auxiliary data and uncover a good feature set for improving the performance of clustering.

In transfer learning, the set of data which needs to be categorised is referred to as target data, and the auxiliary data which used to transmit information is referred to as source data. According to the feature spaces of the source domain and the target domain, transfer learning could be divided into two categories. Early transfer learning algorithms mostly belong to homogeneous transfer where the source and target domains have the same feature space (Pan and Yang, 2010). Homogeneous transfer learning aims at improving generalisation across domains where the source and target data are drawn from different distributions, such as dataset shift (Quiñonero et al., 2009), domain adaptation (Kulis et al., 2011) and multi-task learning (Quadrianto et al., 2010). However, in real applications, it is common that the knowledge is expected to transfer across heterogeneous domains, where the source domain has different feature space from the target domain. For example, in the web image clustering tasks, it would be better to leverage the corresponding text information to help with the image clustering, since the textual features usually contain more semantic information than the visual words. In this case, heterogeneous transfer learning has been proposed to process heterogeneous data and has been successfully applied to text-to-image transferring and cross language classification problems (Yang et al., 2014; Ng et al., 2012).

The key problem of heterogeneous transfer learning is how to bridge different feature spaces. In literature, approaches to solve this problem can be summarised as symmetric transformation (Zhu et al., 2011; Yang et al., 2014) and asymmetric transformation (Kulis et al., 2011). The former separately transforms the source and target domains into a common latent feature space, and the latter tries to transforms the source feature space to the target feature space. However, in real-world applications, the source data collected from web pages or social networks may contain noises and corruptions. Thus, arbitrary knowledge transferring by feature transformation may lead to a degradation of performance. Moreover, most existing heterogeneous transfer learning algorithms only focus on the target domain and cannot handle multiple learning tasks simultaneously. In fact, the performance of some independent tasks could be improved by connecting them together.

In this paper, we propose a co-transfer learning method to deal with the heterogeneous clustering problem. In detail, a collective NMF model is applied to the auxiliary co-occurrence data to learn a common latent subspace. Then the target data from different feature spaces are project to the common subspace using the learned bases. Finally, the projected data are clustered in a unified format to simultaneously get the results of heterogeneous clustering tasks. Furthermore, in order to obtain a clean common subspace, we introduce the error matrices to capture the noises and corruptions of the co-occurrence data during the collective matrix factorisation. Meanwhile, the graph regularisation constrain is enforced on the collective matrix factorisation model to preserve the geometrical structure of the original data spaces. Experiments are conducted on the real-world dataset NUS-WIDE to evaluate the performance of the proposed method. The results demonstrate that the proposed method performs better than *k*-means, symNMF (Kuang et al., 2012) and aPLSA (Yang et al., 2009).

The rest of this paper is organised as follows. In Section 2, we summarise the related work of transfer learning from the aspect of its category. Section 3 gives the problem definition of the heterogeneous co-transfer clustering. The proposed model and its optimisation are presented in Sections 4 and 5. The further co-transfer clustering progress is presented in Section 6. Section 7 gives the experimental evaluation. Finally we conclude our work in Section 8 and give some future work as well.

2 Related work

2.1 Concepts of transfer learning

Transfer learning aims at enhancing the performance of traditional machine learning tasks with the aid of auxiliary data. The basic assumption of transfer learning is that there are some relationships between the auxiliary source data and the target data, which could be used to bridge two domains. According to the transferred components, the approaches of transfer learning can be summarised into four categories (Pan and Yang, 2010).

The first one is instance-based transfer learning, which assumes a part of the source data can be used to boost the target task. The most classical instance-based transfer learning algorithm is the TrAdaBoost (Dai et al., 2007a) which extended the traditional Adaboost algorithm to transfer learning by reweighting the source data at each iteration and filtering out the dissimilar instances. Huang and Smola (2006) proposed a kernel-mean matching (KMM) algorithm to estimate instance weight and increase the distribution similarity by matching the means between the weighted source data and the target data in a reproducing-kernel Hilbert space (RKHS). Sugiyama et al. (2008) further improved Huang's work by estimating the data distribution similarity with the Kullback-Leibler divergence among the source and target domains.

The second kind of popular transferring learning approach is the feature-based transfer learning, which aims at finding the common or implicit shared features of the source and target data. Jiang and Zhai (2007) proposed a two-stage feature selection approach for domain adaption, which gives more importance to the category related features in the training model. Dai et al. (2007a) proposed a co-clustering based approach (CoCC) for out-of-domain documents classification. CoCC identifies the shared word features to transfer knowledge and category information

from the source domain to the target domain. Pan et al. (2008) proposed a dimensionality reduction based transfer learning approach which minimises the maximum mean discrepancy (MMD) of source and target data on the latent semantic space where the feature distributions of different domains are similar.

The latter two kinds of transfer learning approaches are based on parameters and relational-knowledge respectively, which are less investigated than the first two kinds. Parameter-transfer approaches (Daumé and Marcu, 2006; Bonilla et al., 2007) assume that the source tasks and the target tasks share some parameters or prior distributions of the hyper-parameters of the models. In this case, knowledge is transferred across tasks by discovering the shared parameters and priors. Relational-knowledge-transfer approaches (Zheng et al., 2008; Mihalkova et al., 2007) leverage the similar relational scheme among source and target data to deal with the transfer learning problem.

2.2 Heterogeneous and unsupervised transfer learning

Most existing transfer learning algorithms are designed for supervised learning and cannot deal with heterogeneous data directly. Recently, heterogeneous and unsupervised transfer learning has been studied by some researchers. Dai et al. (2008a) first proposed translated learning to enhance the classification with labelled data from different feature space. The core idea of translated learning is to construct a feature-level translator to link different feature spaces. Yang et al. (2009) extend PLSA to help image clustering by annotated web data. The text information is transferred through the annotation relationship for estimating a good latent feature representation. Zhu et al. (2011) also proposed a matrix factorisation based heterogeneous transfer learning methods to enrich the representation of target images with semantic concepts extracted from the auxiliary source data.

In order to further explore the useful non-negative constraints on factors, NMF (Lee and Seung, 1999) is introduced into transfer learning for its simple and interpreted part-based representation. Jing et al. (2012) first adopted a supervised NMF model to solve heterogeneous transferring learning problem. Seichepine et al. (2013) proposed a soft non-negative matrix co-factorisation to improve speaker diarisation results by integrating the audio and video tracks. In order to handle the noise of data in the real world applications, Yang et al. (2015) further proposed a robust and non-negative collective matrix factorisation model for image classification which uses two error matrices to describe the sparsely distributed noise in text and image domain. Yang et al. (2013) also proposed a joint sparse and graph regularised NMF with $\ell_{2,1}$-norm loss function to handle high-dimensional, sparse and noisy data simultaneously.

On the other hand, most existing transfer learning is focused on improving the performance of one target task and cannot obtain learning results on multiple domains at the same time. In this case, some researchers proposed co-transfer learning to improve the performance of multiple tasks from different feature spaces. Ng et al. (2012) present a co-transfer learning (CT-learn) framework which models the knowledge co-transferring problem as a joint transition probability graph. The affinity relationships within domain and the co-occurring relationships cross domains are used to construct the transition probabilities. Yang et al. (2014) also proposed a spectral clustering based co-transfer learning method to address the clustering problem of multi-domain instances with a joint graph. However, CT-learn works under supervised setting and both of them cannot handle the noise in the real-world dataset. In this paper, we consider the co-transfer learning problem with noisy situation and propose a robust non-negative collective matrix factorisation based approach to handle the heterogeneous clustering tasks simultaneously.

3 Problem definition

Given K heterogeneous domains $\mathcal{H} = \{\tilde{X}^k\}_{k=1}^K$, where the k^{th} domain contains n_k instances $\{\tilde{x}_i^k\}_{i=1}^{n_k}$ and the data matrix is denoted as $\tilde{X}^k = [\tilde{x}_1^k, \tilde{x}_2^k, \ldots, \tilde{x}_{n_k}^k] \in \mathbb{R}^{m_k \times n_k}$. $\tilde{x}_i^k \in \mathbb{R}^{m_k}$ is the feature vector with m_k dimensions. The feature spaces of heterogeneous domains are different but the instances in different domains share the same category space. Our goal is to group one or more $\tilde{X}^k$ into c clusters. Traditional clustering methods, such as k-means, handle these problems separately and the performance will be degrade when the data amount in a single domain is insufficient.

In real-world application, the co-occurrence data can be easily collected from web pages and social networks, which could be served as a bridge for heterogeneous domains. The co-occurrence datasets consisted of n_o instances is represented by $\mathcal{O} = \{X^k\}_{k=1}^K$, where each instance appears in K heterogeneous domains. $X^k = [x_1, x_2, \ldots, x_{n_o}] \in \mathbb{R}^{m_k \times n_o}$ is the co-occurrence data matrix in the k^{th} domain, where $x_i^k \in \mathbb{R}^{m_k}$ is the feature vector in the m_k-dimensional feature space as same as $\tilde{X}^k$. Although X^k in the co-occurrence datasets have different feature space, they describe the same set of instances, that is to say $\{X^k\}_{k=1}^K$ have the same latent semantic space. Based on these observation, the first step to transfer knowledge across domains is to map them to a common latent space through the co-occurrence data.

Table 1 Definition of notations

Notations	*Descriptions*
$\mathcal{H}$	Heterogeneous dataset, $\mathcal{H} = \{\tilde{X}^k\}_{k=1}^K$
$\mathcal{O}$	Co-occurrence dataset, $\mathcal{O} = \{X^k\}_{k=1}^K$
$\tilde{X}^k$, X^k	Instances of $\mathcal{H}$ and $\mathcal{O}$ in the k^{th} domain
$\tilde{x}_i^k$, x_i^k	Feature vector in the k^{th} domain

Table 1 Definition of notations (continued)

Notations	*Descriptions*
$\tilde{G}^k, G^k$	The p-nearest neighbour graph in the k^{th} domain
$\tilde{L}^k, L^k$	The Laplacian matrix in the k^{th} domain
$\tilde{S}^k, S^k$	The error matrix in the k^{th} domain
W^k	The basis factor of the k^{th} domain
$\tilde{H}^k, H^k$	The new representation in the k^{th} domain
H	Semantic representation of the instances in the co-occurrence dataset
$\mathcal{U}$	The unified representation of all instances in the heterogeneous dataset
m_k	The dimensionality of the feature vector in the k^{th} domain
n_k	Number of instance in $\tilde{X}^k$
n_o	Number of co-occurrence instances
r	The dimensionality of the common latent subspace

4 The proposed model

In order to identify the common latent space among different domains, we propose a robust and graph regularised non-negative collective matrix factorisation model for heterogeneous co-transfer clustering (RGHCTC). The RGHCTC model is formulated as:

$$\begin{aligned}\min_{\{W^k\}_{k=1}^K, H, \{S^k\}_{k=1}^K} & \sum_{k=1}^{K} \lambda^k \parallel X^k - W^k H - S^k \parallel_F^2 \\ & + \alpha^k \operatorname{Tr}(HL^k H^{\mathrm{T}}) + \beta^k \parallel S^k \parallel_1 \\ & s.t.\ \{W^k\}_{k=1}^K \geq 0, H \geq 0,\end{aligned} \tag{1}$$

where $W^k \in \mathbb{R}^{m_k \times r}$, $H \in \mathbb{R}^{r \times n_k}$, r is the dimensionality of the common latent space. $S^k \in \mathbb{R}^{m_k \times n_k}$ is the error matrix used to capture noises in k^{th} domain. $\|\cdot\|_F$ denotes the Frobenius norm. $L^k = D^k - G^k$ is the Laplacian matrix where $G^k \in \mathbb{R}^{n_o \times n_o}$ is the weight matrix and $D^k = \operatorname{diag}(d^k{}_1, d_2^k \ldots, d_{n_o}^k)$, $d_i^k = \sum_j G_{i,j}^k$. $\operatorname{Tr}(\cdot)$ denotes the trace of a matrix.

The first term of the RGHCTC model is the robust NMF loss function with error matrix S^k, which is introduced to explicitly capture the noises among X^k. As we assume noises are sparely located in the data matrix, the ℓ_1-norm is enforced on S^k to derive sparsity, and $\beta^k \geq 0$ is the regularisation parameter controlling the sparsity of S^k. In the factorisation model, W^k refers to the learned latent feature in the heterogeneous domains, and H refers to the new representation of the co-occurrence instances in the common latent space. Here we use the unified H instead of H^k to denote the new representation of data matrix X^k, since H^k corresponding to different domains describes the same set of objects. Thus, we can set $H^1 = H^2 = \cdots = H^k = H$. $0 \leq \lambda^k \leq 1$ is the weight parameter used to balance the importance of different domains, and $\sum_{k=1}^K \lambda^k = 1$.

On the other hand, in order to keep the original data structure of each domain, the graph regularisation term $\alpha^k \operatorname{Tr}(HL^k H^T)$ is constrained to the factorisation model. The graph regulariser is derived from the natural assumption that if two data points x_i^k and x_j^k are close in the intrinsic geometric structure of the data distribution, then their new representations h_i and h_j, with respect to the learned basis, should also be close to each other. This is usually referred to as *manifold assumption*, which can be realised by constructing a nearest neighbour graph. In our heterogeneous co-transfer clustering problem, we construct a nearest neighbour graph with n_o vertices for each domain, where each vertex corresponds to a data point of the data matrix X^k. If x_i^k is among the p-nearest neighbours of x_j^k, or x_j^k is among the p-nearest neighbours of x_i^k, we put an edge between x_i^k and x_j^k, whose weight is defined as follows:

$$G_{i,j}^k = \begin{cases} \mathcal{K}\left(x_i^k, x_j^k\right), & \text{if } x_i^k \in \mathcal{N}_p\left(x_j^k\right) \text{ or } x_j^k \in \mathcal{N}_p\left(x_i^k\right) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\mathcal{N}_p(x_i^k)$ denotes the p-nearest neighbours of x_i^k, $\mathcal{K}(x_i^k, x_j^k)$ is the kernel function which can be selected depend on the particularity of different applications. $G_{i,j}^k$ describes the closeness of data points x_i^k and x_j^k. According to the *manifold assumption*, the smoothness of the r-dimensional manifold embedded in $\mathbb{R}^{m_k}$ can be measured by:

$$\begin{aligned}\mathcal{R}^k &= \frac{1}{2}\sum_{i,j=1}^{n_o} \parallel h_i - h_j \parallel^2 G_{i,j}^k \\ &= \sum_{i=1}^{N} h_i^{\mathrm{T}} h_i D_{i,i}^k - \sum_{i,j=1}^{N} h_i^{\mathrm{T}} h_j G_{i,j}^k \\ &= \operatorname{Tr}(HD^k H^{\mathrm{T}}) - \operatorname{Tr}(HG^k H^{\mathrm{T}}) \\ &= \operatorname{Tr}(HL^k H^{\mathrm{T}}),\end{aligned} \tag{3}$$

where D^k is a diagonal matrix with $D_{i,i}^k = \sum_j G_{i,j}^k$, and $L^k = D^k - G^k$ is the so called graph Laplacian of the nearest neighbour graph. The smaller the value of $\mathcal{R}^k$ is, the smoother the new representation will be. Enforcing this constraint into the traditional NMF object function leads to the regularisation term $\alpha^k \operatorname{Tr}(HL^k H^T)$, where $\alpha^k \geq 0$ used to control the smoothness of the new representation in the common latent space.

5 Optimisation

The object function (1) is not convex with respect to W^k, H and S^k jointly. There is no realistic algorithm to find a global

minimum. In this section, we introduce an iterative strategy for solving problem (1). The values of W^k, H and S^k are updated individually when fixing other variables. Thus, a local minimum can be achieved by solving a series of sub-optimisation problems.

5.1 Updating S^k

When fixing W^k and H, the objective function (1) degrades into:

$$\min_{S^k} \lambda^k \| A^k - S^k \|_F^2 + \beta^k \| S^k \|_1, \tag{4}$$

where $A^k = X^k - W^k H$. Equation (4) is a ℓ_1-norm regularised convex optimisation problem. We introduce an effective approach for solving this problem via the *soft-thresholding operator* (Hale et al., 2008). The update rule for S^k is formulated as follows:

$$\left(S_{i,j}^k\right)^{(t+1)} \leftarrow \begin{cases} A_{i,j}^k - v^k, & \text{if } A_{i,j}^k > v^k \\ A_{i,j}^k + v^k, & \text{if } A_{i,j}^k < -v^k \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $v^k = \beta^k / 2\lambda^k$ and t denotes the current iteration number. It is obvious that with the increasement of the value of v^k, much more elements in S^k will turn to be zero. When $v^k > \max_{i,j}(A_{i,j}^k)$ all elements in S^k would be zero and our RGHCTC model degenerates to the traditional NMF model (ignoring the graph regularisation term). Thus, it is better to keep $0 < v^k < \max_{i,j}(A_{i,j}^k)$ in the RGHCTC model for handling noises.

5.2 Updating W^k

When fixing S^k and H, the objective function for optimising W^k can be written as:

$$\min_{W^k} \lambda^k \| Z^k - W^k H \|_F^2 \quad s.t.\ W^k \geq 0, \tag{6}$$

where $Z^k = X^k - S^k$. We can prove that with the update rule (5), $X^k - S^k > 0$. In this case, equation (6) becomes a non-negative quadratic programming problem as same as the traditional NMF model, which can be solved by multiplicative updates (Lee and Seung, 1999):

$$\left(W_{i,j}^k\right)^{(t+1)} \leftarrow \left(W_{i,j}^k\right)^{(t)} \left[\frac{(Z^k H^{\mathrm{T}})_{i,j}}{\left((W^k)^{(t)} H H^{\mathrm{T}}\right)_{i,j}} \right]. \tag{7}$$

5.3 Updating H

For the fixed W^k and S^k, the objective function for optimising H is formulated as:

$$\min_{H} \sum_{k=1}^{K} \lambda^k \| Z^k - W^k H \|_F^2 + \alpha^k \operatorname{Tr}(H L^k H^{\mathrm{T}}) \quad s.t.\ H \geq 0, \tag{8}$$

where $Z^k = X^k - S^k$. Due to the presence of the graph regularisation term, it is infeasible to solve equation (8) by multiplicative updates directly. We adopt gradient descent method to derive the update rule for H. Let:

$$\mathcal{F} = \sum_{k=1}^{K} \lambda^k \| Z^k - W^k H \|_F^2 + \alpha^k \operatorname{Tr}(H L^k H^{\mathrm{T}}), \tag{9}$$

gradient descent leads to the following additive update rule:

$$h_{i,j} = h_{i,j} - \delta_{i,j} \frac{\partial \mathcal{F}}{\partial h_{i,j}}, \tag{10}$$

where $\delta_{i,j}$ refers to the step size parameter. The choice of the step size should guarantee the non-negativity of $h_{i,j}$. We use the similar tricks in Cai et al. (2008) to set the step size parameter automatically. Let:

$$\delta_{i,j} = \frac{h_{i,j}}{2\sum_{k=1}^{K} \left(\lambda^k (W^k)^{\mathrm{T}} W^k H + \alpha^k H D^k\right)_{i,j}}, \tag{11}$$

we have:

$$\begin{aligned} & h_{i,j} - \delta_{i,j} \frac{\partial \mathcal{F}}{\partial h_{i,j}} \\ &= h_{i,j} - \frac{h_{i,j}}{2\sum_{k=1}^{K} \left(\lambda^k (W^k)^{\mathrm{T}} W^k H + \alpha^k H D^k\right)_{i,j}} \frac{\partial \mathcal{F}}{\partial h_{i,j}} \\ &= h_{i,j} \frac{\sum_{k=1}^{K} \left(\lambda^k (W^k)^{\mathrm{T}} Z^k + \alpha^k H G^k\right)_{i,j}}{\sum_{k=1}^{K} \left(\lambda^k (W^k)^{\mathrm{T}} W^k H + \alpha^k H D^k\right)_{i,j}}. \end{aligned} \tag{12}$$

Thus, the rule for updating H can be written as:

$$H_{i,j}^{(t+1)} \leftarrow H_{i,j}^{(t)} \left[\frac{\sum_{k=1}^{K} \begin{pmatrix} \lambda^k (W^k)^{\mathrm{T}} Z^k \\ + \alpha^k H^{(t)} G^k \end{pmatrix}_{i,j}}{\sum_{k=1}^{K} \begin{pmatrix} \lambda^k (W^k)^{\mathrm{T}} W^k H^{(t)} \\ + \alpha^k H^{(t)} D^k \end{pmatrix}_{i,j}} \right]. \tag{13}$$

6 Heterogeneous co-transfer clustering

Heterogeneous co-transfer clustering aims at grouping data from different domain simultaneously. By applying the proposed non-negative collective matrix factorisation model on the co-occurrence data, the latent basis factor W^k of different domain is identified, which can be used for mapping data from different feature spaces to a common latent space. Formally, given K heterogeneous datasets $\{\tilde{X}^k\}_{k=1}^K$, which share the same category space as $\{X\}_{k=1}^K$, we firstly use $\{W^k\}_{k=1}^K$, to learn the new representations $\{\tilde{H}^k\}_{k=1}^K$ in the common space. In order to handle the corruptions of the data matrix and preserve the geometric structure of the data manifold, we apply the proposed RGHCTC model to individual domain respectively to

re-represent the datasets. This process can be formulated as the following sequence of optimisation problems:

$$\begin{Bmatrix}\min\limits_{\tilde{H}^k,\tilde{S}^k} \| \tilde{X}^k - W^k\tilde{H}^k - \tilde{S}^k \|_F^2 + \tilde{\alpha}^k \operatorname{Tr}\left(\tilde{H}^k\tilde{L}^k\left(\tilde{H}^k\right)^{\mathrm{T}}\right) \\ + \tilde{\beta}^k \| \tilde{S}^k \|_1 \quad s.t.\ \tilde{H}^k \geq 0\end{Bmatrix}_{k=1}^{K}, \tag{14}$$

where $\tilde{\alpha}^k, \tilde{\beta}^k$ is the regularisation parameters, and $\tilde{H}^k \in \mathbb{R}^{r\times n_k}$ refers to the new representation of $\tilde{X}^k$ in the k^{th} domain. Since $\{\tilde{H}^k\}_{k=1}^K$ have the same feature space, it is feasible to combine them together by:

$$\mathcal{U} = [\tilde{H}^1, \tilde{H}^2, \ldots, \tilde{H}^K] \in \mathbb{R}^{r\times n_a}\left(n_a = \sum\nolimits_{k=1}^{K} n_k\right),$$

which is correspond to all instances in different domains. Thus, the final clustering results can be obtained by grouping the data matrix $\mathcal{U}$. Furthermore, as the basis factors $\{W^k\}_{k=1}^K$ are learned from the co-occurrence datasets, which contain semantic information transferred across heterogeneous domains, the learned new representations $\{\tilde{H}^k\}_{k=1}^K$ gain better ability to explicitly demonstrate the intrinsic category structure. In this case, simple clustering algorithms such as *k*-means, and the traditional NMF could achieve ideal performance on $\mathcal{U}$.

The process of our proposed heterogeneous co-transfer clustering algorithm is summarised in Algorithm 1.

Algorithm 1 RGHCTC co-transfer clustering algorithm

Input:

The heterogeneous datasets $\mathcal{H} = \{\tilde{X}^k\}_{k=1}^K$, co-occurrence datasets $\mathcal{O} = \{X^k\}_{k=1}^K$

Output:

Clustering results for $\mathcal{H}$

Steps:

1 Construct the *p*-nearest neighbour graphs $\{G^k\}_{k=1}^K$ for $\mathcal{O}$ on each domain;

2 Solve the optimisation problem (1) on $\mathcal{O}$ to learn the latent basis factors $\{W^k\}_{k=1}^K$;

3 Construct the *p*-nearest neighbour graphs $\{\tilde{G}^k\}_{k=1}^K$ for $\mathcal{H}$ on each domain;

4 Solve the sequence of optimisation problems (14) on $\mathcal{H}$ with fixed $\{W^k\}_{k=1}^K$, and get the new-representation $\{\tilde{H}^k\}_{k=1}^K$;

5 Combine $\{\tilde{H}^k\}_{k=1}^K$ with $\mathcal{U} = [\tilde{H}^1, \tilde{H}^2, \ldots, \tilde{H}^K] \in \mathbb{R}^{r\times n_a}$

6 Run *k*-means clustering algorithm on $\mathcal{U}$ and assign each instance to the corresponding cluster.

If the target domains are expected to have few noises, we could directly map the target data $\{\tilde{X}^k\}_{k=1}^K$ to the common subspace by $\tilde{H}^k = W^k\tilde{X}^k$. Thus, the complexity of the algorithm could be reduced since the absence of solving additional optimisation problems.

7 Experiments

7.1 Dataset and baseline methods

The dataset used in our experiments is NUS-WIDE (Chua et al., 2009), which is a widely used heterogeneous dataset for multi-view learning and transfer learning. It includes 269,648 images and 5,018 associated tags from Flickr. We follow (Ng et al., 2012) to construct ten binary heterogeneous co-transfer clustering tasks with five selected categories (flowers, rocks, sun, toy, tree). For each task, 600 images, 600 texts, and totally 1,600 co-occurred image-text pairs are sampled with respect to the original data distributions. In order to retain the original trait of images (includes noises), we adopt the 4,096 dimensional feature vectors extracted from the CNN model (Zeiler and Fergus, 2014) to construct the image feature matrix. For text data, 1,000 tags are finally picked for constructing the text feature vectors.

The *k*-means method is one of the most classical and simplest clustering algorithm. It is widely used in the practical applications because of its simplicity. SymNMF (Kuang et al., 2012) is a general framework for graph clustering, which performs factorisation on a similarity matrix of data points with non-negative constraint. SymNMF has better ability to capture the clustering structure embedded in the graph representation and is easily to obtain the cluster assignment compared to spectral clustering. aPLSA (Yang et al., 2009) is an unsupervised heterogeneous transfer learning model which aims at utilising auxiliary annotation information to enhance image clustering performance. aPLSA can be easily extended to transfer knowledge from multiple source domains to one target domain.

7.2 Evaluation metrics

The performance of different clustering methods is evaluated by two metrics. The first one is the accuracy (ACC), which is defined as follows:

$$\text{ACC} = \frac{\sum_{i=1}^{n}\delta\left(map\left(r_i\right), l_i\right)}{n} \tag{15}$$

where r_i denotes the obtained cluster label of the data point $\tilde{x}_i^k$, and l_i denotes the ground truth. $\delta(x, y)$ is the delta function that equals one if $x = y$, otherwise equals zero, while $map(\cdot)$ is the permutation mapping function which maps each cluster label r_i to the equivalent label from the original dataset.

Table 2 Results of image clustering tasks

Task	*Accuracy*				*NMI*			
	k-means	*SymNMF*	*aPLSA*	*RGHCTC*	*k-means*	*SymNMF*	*aPLSA*	*RGHCTC*
1	92.90%	91.33%	92.67%	*93.27%*	0.6455	0.6103	0.6327	*0.6615*
2	92.92%	89.83%	92.83%	*93.73%*	0.6403	0.5485	0.6240	*0.6629*
3	75.13%	74.25%	82.33%	*86.97%*	0.2687	0.1220	0.3239	*0.4317*
4	90.23%	88.17%	90.50%	*91.47%*	0.5784	0.5444	0.5791	*0.6112*
5	89.63%	90.17%	87.50%	*90.58%*	0.5064	0.5303	0.4810	*0.5332*
6	91.52%	*94.33%*	93.33%	93.72%	0.5576	*0.6472*	0.6113	0.6241
7	81.70%	83.33%	79.17%	*86.62%*	0.3280	0.3504	0.2606	*0.4297*
8	94.77%	92.00%	*95.50%*	95.15%	0.7003	0.6190	*0.7336*	0.7170
9	82.13%	*85.00%*	82.17%	83.43%	0.3191	*0.3867*	0.3394	0.3483
10	93.87%	93.00%	94.17%	*94.25%*	0.6521	0.6122	0.6582	*0.6594*
Avg.	88.48%	88.14%	89.02%	*90.92%*	0.5196	0.4976	0.5244	*0.5679*

Table 3 Results of text clustering tasks

Task	*Accuracy*				*NMI*			
	k-means	*SymNMF*	*aPLSA*	*RGHCTC*	*k-means*	*SymNMF*	*aPLSA*	*RGHCTC*
1	92.69%	92.24%	92.50%	*93.51%*	0.6395	0.6004	0.6248	*0.6503*
2	73.56%	94.93%	89.33%	*95.45%*	0.3353	0.6789	0.5031	*0.7033*
3	70.00%	91.91%	90.33%	*93.71%*	0.2137	0.4926	0.4663	*0.5730*
4	77.94%	91.13%	87.17%	*92.13%*	0.3588	0.5562	0.4594	*0.5930*
5	80.02%	75.87%	71.83%	*87.34%*	0.3319	0.2476	0.1935	*0.4328*
6	90.98%	91.29%	90.67%	*94.25%*	0.5328	0.5193	0.5140	*0.6417*
7	66.58%	64.07%	57.50%	*82.51%*	0.1408	0.1307	0.0172	*0.3392*
8	92.00%	90.36%	93.00%	*95.51%*	0.6176	0.5950	0.6293	*0.7319*
9	84.74%	82.70%	80.00%	*85.54%*	0.3956	0.3346	0.2803	*0.4010*
10	84.86%	89.80%	90.33%	*90.82%*	0.4493	0.4906	0.5127	*0.5238*
Avg.	81.34%	86.43%	84.27%	*91.08%*	0.4015	0.4646	0.4201	*0.5590*

The second measure is the normalised mutual information (NMI), which is defined as follows:

$$\mathrm{NMI} = \frac{MI(\mathcal{C}, \mathcal{C}')}{\max(H(\mathcal{C}), H(\mathcal{C}'))} \tag{16}$$

where $\mathcal{C}$ and $\mathcal{C}'$ denote the sets of clusters obtained from the clustering algorithm and the ground truth respectively. $H(\cdot)$ is the entropy of a set of clusters, and $MI(\mathcal{C}, \mathcal{C}')$ is the mutual information which can be computed by:

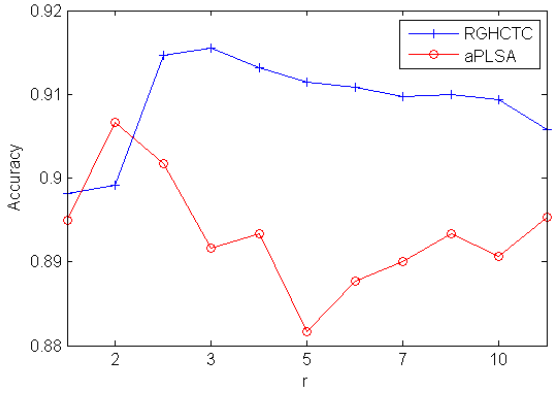
$$MI = \sum_{c_i \in \mathcal{C}, c'_j \in \mathcal{C}'} p(c_i, c'_j) \cdot \log_2 \frac{p(c_i, c'_j)}{p(c_i) \cdot p(c'_j)} \tag{17}$$

where $p(c_i)$, $p(c'_j)$ denote the probabilities that a sample arbitrarily selected from the dataset belongs to the clusters c_i and c'_j respectively, and $p(c_i, c'_j)$ is the joint probability that the arbitrary selected sample belongs to the clusters c_i and c'_j at the same time. NMI ranges from 0 to 1, and the larger value of NMI is, the better clustering performance is.

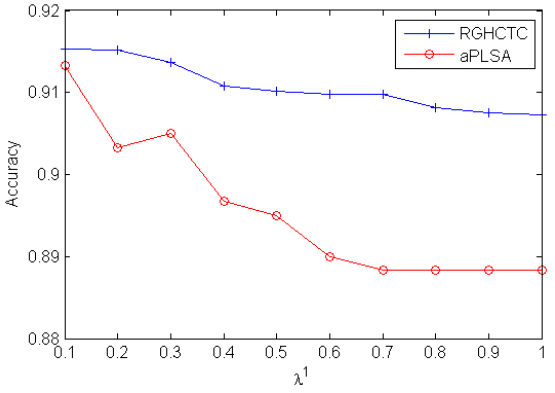
7.3 Results and discussion

NUS-WIDE contains instance from both image and text domains, then K is equal to two. For image data, we use Heat kernel function to yield the nonlinear version of similarity in the intrinsic manifold structure of data, and the threshold parameter σ is set to two in our experiment. For text data, we adopt the cosine similarity. Since *k*-means and SymNMF do not leverage auxiliary information and just perform on a single domain, we run them on target image and text domain respectively to get the final clustering results. As aPLSA only considers to transfer knowledge from source domain to target domain, we regard image and text data as target domain respectively (whereas text and image data is source domain) to get the transferred clustering results on two domains. For each task, we run the algorithms ten times and record the average accuracy and NMI with random initialisation. The results of ten binary clustering tasks on image and text domain are shown in Tables 2 and 3 respectively.

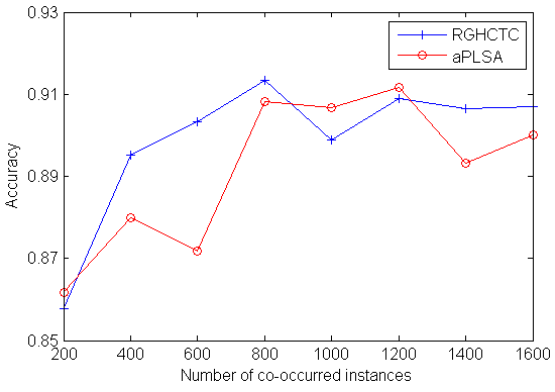
Figure 1 Empirical evaluation of RGHCTC (see online version for colours)



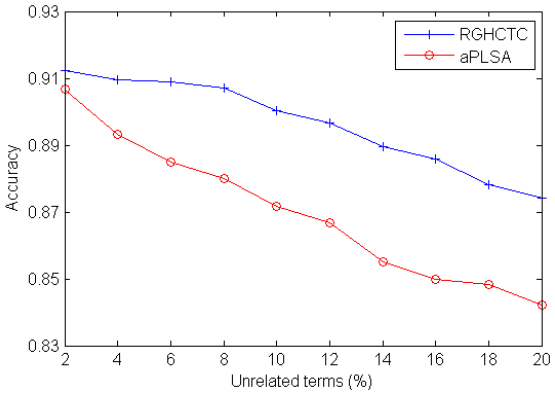
(a)



(b)



(c)



(d)

From Tables 2 and 3, we can see that RGHCTC outperforms the baselines in both image and text clustering tasks on average, and achieves more robust results. This demonstrates the effectiveness of our method that transferring knowledge across heterogeneous domains can improve the performance of data clustering on single domain. It can be noticed that unlike most image-to-text transfer learning algorithms, RGHCTC achieves excellent performance in text domain as well, it indicates that RGHCTC could boost the text clustering performance by leveraging the image knowledge with expressive image features such as which is extracted from the CNN model.

Except the data quality, several key factors affect the performance of RGHCTC such as the dimensionality of the common latent space, the number of co-occurrence instances, and the model parameters. We randomly select a binary image clustering task (with 600 images, 370 for flowers, 230 for tree, and totally 1,600 image-text pair) to illustrate the effects of different parameters used in our experiments.

Figure 1(a) illustrates the image clustering results with varying dimensionality r of the comment latent space. We can see that RGHCTC achieves the best performance around seven, it is much smaller than the number of original image features (4,096). Thus, the computational complexity of target clustering task can be efficiently reduced in the low dimensionality common space while a higher clustering result is realised as well.

In RGHCTC model, λ^k is the weight parameter controlling the balance of different domains. Since NUS-WIDE contains instances from two domains, the weight parameter are λ^1 and λ^2 ($\lambda^1 + \lambda^2 = 1$), where λ^1 refers to the transferred weight of image domain, and λ^2 refers to the transferred weight of text domain. Thus we can only tune λ^1 from 0 to 1 and set $\lambda^2 = 1 - \lambda^1$. Smaller λ^1 indicates more importance of text information. Figure 1(b) shows the results of image clustering where RGHCTC achieves the best performance when λ^1 is in the range of [0.2, 0.3]. This demonstrates that the image clustering benefits from the text information which is transferred in RGHCTC model, and it is reasonable to build a bridge between image domain and text domain.

Another important factor which affects the performance of co-transfer clustering is the size of co-occurrence data n_o. We run the co-transfer clustering algorithm with varying co-occurrence data size from 200 to 1,600 and record the accuracy on image domain. The result is shown on Figure 1(c). As we can see, the image clustering accuracy increases when the size of co-occurrence data n_o increases and becomes relatively steady when n_o up to a point (800). This indicates that more co-occurrence instances make the learned bases more precise and the new representation more helpful for clustering. However, this effect becomes unapparent when the co-occurrence instances are sufficient to find the common latent space and additional instances are not useful anymore.

In order to evaluate the robustness of RGHCTC, we randomly add some irrelevant text terms (from 2% to 20%)

to the co-occurrence data. The results of image clustering tasks are illustrated in Figure 1(d) from which we can see the accuracy of both methods drop owing to the appearance of noisy terms. It is apparent that our method drops more slowly than the compared method. This indicates that the negative information of noisy terms is captured by the error matrix so that the learned subspace is more accurate for image representation.

8 Conclusions

In this paper, we proposed a heterogeneous co-transfer clustering method based on the collective NMF. One property of the proposed method is to introduce the error matrix to capture the noises and corruptions for improving the robustness. In addition, the graph regularisation constraint is enforced on the collective matrix factorisation to preserve the geometric structure of the original data space. Experimental results imply that our proposed method achieves more accurate clustering than other representative clustering methods in terms of ACC and NMI. In the future work, we will investigate the effective initial scheme of the collective NMF to further improve the performance of the proposed method.

Acknowledgements

This work is supported by the project of 'Source Identification and Contamination Characteristics of Heavy Metals in Agricultural Land and Products' (2016YFD0800300), the National Key Research and Development Program of China.

References

Bonilla, E., Chai, K.M. and Williams, C. (2007) 'Multi-task gaussian process prediction', in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pp.153–160.

Cai, D., He, X., Wu, X. and Han, J. (2008) 'Non-negative matrix factorization on manifold', in *Eighth IEEE International Conference on Data Mining*, pp.63–72.

Chua, T-S., Tang, J., Hong. R., Li, H., Luo, Z. and Zheng, Y. (2009) 'NUS-WIDE: a real-world web image database from national university of Singapore', in *Proceeding of the ACM International Conference on Image and Video Retrieval*, pp.8–10.

Dai, W., Chen, Y., Xue, G.R., Yang, Q. and Yu, Y. (2008a) 'Translated learning: transfer learning across different feature spaces', in *Proceedings of the 21st International Conference on Neural Information Processing Systems*, pp.353–360.

Dai, W., Yang, Q., Xue, G-R. and Yu, Y. (2008b) 'Self-taught clustering', in *Proceedings of the 25th International Conference on Machine learning*, pp.200–207.

Dai, W., Xue, G-R., Yang, Q. and Yu, Y. (2007a) 'Co-clustering based classification for out-of-domain documents', *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp.210–219.

Dai, W., Yang, Q., Xue, G-R., and Yu, Y. (2007b) 'Boosting for transfer learning', in *Proceedings of the 24th International Conference on Machine Learning*, pp.193–200.

Daumé, H. and Marcu, D. (2006) 'Domain adaptation for statistical classifiers', *J. Artif. Intell. Res.*, Vol. 26, No. 1, pp.101–126.

Hale, E.T., Yin, W. and Zhang, Y. (2008) 'Fixed-point continuation for l1-minimization: methodology and convergence', *SIAM J. Optim.*, Vol. 19, No. 3, pp.1107–1130.

Huang, J. and Smola, A. (2006) 'Correcting sample selection bias by unlabeled data', in *Proceedings of the 19th International Conference on Neural Information Processing Systems*, pp.601–608.

Jiang, J. and Zhai, C. (2007) 'A two-stage approach to domain adaptation for statistical classifiers', in *Cikm 2007*, pp.401–410.

Jing, L., Zhang, C. and Ng, M.K. (2012) 'SNMFCA: supervised NMF-based image classification and anNotation', *IEEE Trans. Image Process.*, Vol. 21, No. 11, pp.4508–4521.

Kuang, D., Ding, C. and Park, H. (2012) 'Symmetric nonnegative matrix factorization for graph clustering', in *SIAM International Conference on Data Mining 2012*, pp.494–505.

Kulis, B., Saenko, K. and Darrell, T. (2011) 'What you saw is not what you get: domain adaptation using asymmetric kernel transforms', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.1785–1792.

Lee, D. and Seung, H. (2000) 'Algorithms for non-negative matrix factorization', in *Proceedings of the 13th International Conference on Neural Information Processing Systems*, pp.556–562.

Lee, D.D. and Seung, H.S. (1999) 'Learning the parts of objects by non-negative matrix factorization', *Nature*, Vol. 401, No. 6755, pp.788–91.

Mihalkova, L., Huynh, T.N. and Mooney, R.J. (2007) 'Mapping and revising Markov logic networks for transfer learning', in *Proceedings of the 22th AAAI Conference on Artificial Intelligence*, pp.608–614.

Ng, M.K., Wu, Q. and Ye, Y. (2012) 'Co-transfer learning via joint transition probability graph based method', in *Proceedings of the 1st International Workshop on Cross Domain Knowledge Discovery in Web and Social Network Mining*, pp.1–9.

Pan, S.J. and Yang, Q. (2010) 'A survey on transfer learning', *IEEE Trans. Knowl. Data Eng.*, October, Vol. 22, No. 10, pp.1345–1359.

Pan, S.J., Kwok, J.T. and Yang, Q. (2008) 'Transfer learning via dimensionality reduction', in *Proceedings of the 23th AAAI Conference on Artificial Intelligence*, pp.677–682.

Quadrianto, N., Petterson, J., Caetano, T.S., Smola, A.J. et al. (2010) 'Multitask learning without label correspondences', in *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, pp.1957–1965.

Quiñonero-Candela, J., Sugiyama, M. and Schwaighofer, A. et al. (2009) *Dataset Shift in Machine Learning*, pp.27–28, The MIT Press.

Seichepine, N., Essid, S., Fevotte, C. and Cappe, O. (2013) 'Soft nonnegative matrix co-factorization with application to multimodal speaker diarization', in *Processing of the IEEE International Conference on Acoustics, Speech and Signal*, pp.3537–3541.

Sugiyama, M., Suzuki, T., Nakajima, S. and Kashima, H. et al. (2008) 'Direct importance estimation for covariate shift adaptation', *Ann. Inst. Stat. Math.*, Vol. 60, No. 4, pp.699–746.

Yang, L., Jing, L. and Ng, M.K. (2015) 'Robust and non-negative collective matrix factorization for text-to-image transfer learning', *IEEE Trans. Image Process.*, Vol. 24, No. 12, pp.4701–4714.

Yang, L., Jing, L. and Yu, J. (2014) 'Heterogeneous co-transfer spectral clustering', in *9th International Conference on Rough Sets and Knowledge Technology*, pp.352–363.

Yang, Q., Chen, Y., Xue, G-R., Dai, W. and Yu, Y. (2009) 'Heterogeneous transfer learning for image clustering via the social web', in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*, Vol. 1, pp.1–9.

Yang, S., Hou, C., Zhang, C. and Wu, Y. (2013) 'Robust non-negative matrix factorization via joint sparse and graph regularization for transfer learning', *Neural Comput. Appl.*, Vol. 23, No. 2, pp.541–559.

Zhu, Y., Chen, Y., Lu, Z., Pan, S.J., and Xue, G-R. et al. (2011) 'Heterogeneous transfer learning for image classification', in *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, pp.1304–1309.

Zeiler, M.D. and Fergus, R. (2014) 'Visualizing and understanding conVolutional networks', in *European Conference on Computer Vision*, pp.818–833.

Zheng, V.W., Xiang, E.W., Yang, Q. and Shen, D. (2008) 'Transferring localization models over time', in *Proceedings of the 23th AAAI Conference on Artificial Intelligence*, pp.1421–1426.