Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Convex Co-Embedding for Matrix Completion with Predictive Side Information

Yuhong Guo
School of Computer Science
Carleton University, Ottawa, Canada
yuhong.guo@carleton.ca

Abstract

Matrix completion as a common problem in many application domains has received increasing attention in the machine learning community. Previous matrix completion methods have mostly focused on exploiting the matrix low-rank property to recover missing entries. Recently, it has been noticed that side information that describes the matrix items can help to improve the matrix completion performance. In this paper, we propose a novel matrix completion approach that exploits side information within a principled co-embedding framework. This framework integrates a low-rank matrix factorization model and a label embedding based prediction model together to derive a convex co-embedding formulation with nuclear norm regularization. We develop a fast proximal gradient descent algorithm to solve this co-embedding problem. The effectiveness of the proposed approach is demonstrated on two types of real world application problems.

Introduction

In many data analysis problems, the relevant information often lies in a low-dimensional subspace of the ambient space and the data matrix exhibits low-rank properties. Matrix completion exploits this observation to recover a low-rank matrix from sparsely observed entries, by implicitly or explicitly identifying low-dimensional vector representations of the row and column objects and decomposing the matrix into the product two low-dimensional matrices (typically in an implicit manner through nuclear norm); e.g., $X = AB$ for $X \in \mathbb{R}^{n\times t}, A \in \mathbb{R}^{n\times m}, B \in \mathbb{R}^{m\times t}$ with $m < \min(n, t)$. The underlying principle is that the distributed low-dimensional representations of the row and column objects can be learned from the sparsely observed entries in a statistical manner, which can then be used to recover the missing entries with their inner products. Matrix completion has broad applications including collaborative filtering (Abernethy et al. 2009; Hu, Koren, and Volinsky 2008), clustering (Yi et al. 2012), and feature learning (Argyriou, Evgeniou, and Pontil 2008). Many algorithms and theoretical results have been developed in the standard setting of matrix completion based on the low-rank property (Cai, Candès, and Shen 2010; Candès and Tao 2010; Candès and Recht 2012; Lin et al. 2009; Keshavan, Montanari, and Oh 2010; Koltchinskii, Lounici, and Tsybakov 2011; Mazumder, Hastie, and Tibshirani 2010; Recht 2011; Richard, Savalle, and Vayatis 2012), without considering side information about the row or column objects.

Recently, it has been observed that in a number of applications of matrix completion, in addition to the observed entries of the target matrix, auxiliary information that describes the row or column objects of the matrix usually exist and can benefit the matrix completion process. For example, in collaborative filtering, besides the user-item recommendation matrix, side information such as product reviews, which comment on the product items, are often available and can be used to improve recommendation performance (Adams, Dahl, and Murray 2010; Fang and Si 2011; Menon et al. 2011; Porteous, Asuncion, and Welling 2010). In (Natarajan and Dhillon 2014), matrix completion with side information has been used to predict gene-disease associations. These works however mostly use non-convex learning techniques without exploiting the structure of the side information. Convex methods for matrix completion have been explored for multi-label learning problems with incomplete labels (Cabral et al. 2011; Goldberg et al. 2010; Xu, Jin, and Zhou 2013), where the partially observed label matrix is the target matrix to be completed and the features of the instances can be used as side information. These works however exploit either the low-rank property of the label matrix or the prediction models, but not both.

In this paper, we propose a novel matrix completion approach to exploit side information within a principled co-embedding framework. The proposed framework not only takes the low-rank property of the target matrix into account, but also simultaneously exploits a label embedding idea to enforce a consistent low-rank structure for the prediction model on the side information. Label embeddings have been exploited in the literature to capture label semantic structures and consequently prediction model structures to improve prediction performance (Akata et al. 2013; Bengio, Weston, and Grangier 2010; Mirzazadeh, Guo, and Schuurmans 2014). The idea of our co-embedding framework is to enforce the consistency of the label embeddings induced from the prediction model and that induced from the low-rank target matrix. We formulate this framework as a convex minimization problem with nuclear norm regulariza-

tion, provide a bound analysis, and develop a fast proximal gradient descent algorithm to solve the problem efficiently. We conduct experiments on two types of applications, transductive incomplete multi-label learning and matrix completion for recommendation systems. Our proposed approach demonstrates very effective empirical results.

Related Work

Matrix Completion The main goal of matrix completion is to exploit the low-rank structure of a data matrix to fill the missing entries based on the observed ones, addressing applications such as collaborative filtering (Rennie and Srebro 2005). Theoretical studies show that one can perfectly recover a low-rank matrix with very high probability from a small number of observed entries under certain assumptions (Candès and Tao 2010; Candès and Recht 2012; Recht 2011). A number of computational algorithms have been developed to solve standard matrix completions (Cai, Candès, and Shen 2010; Lin et al. 2009; Keshavan, Montanari, and Oh 2010; Koltchinskii, Lounici, and Tsybakov 2011; Mazumder, Hastie, and Tibshirani 2010; Richard, Savalle, and Vayatis 2012), which mostly seek efficient optimizations with nuclear norms. Recently, several works have taken side information into account to improve matrix completion, including incorporating side information in probabilistic matrix factorization (Adams, Dahl, and Murray 2010; Porteous, Asuncion, and Welling 2010), conducting inductive matrix factorization (Shin et al. 2015), performing matrix co-factorization (Fang and Si 2011), collective factorization (Bouchard, Guo, and Yin 2013), and weighted factorization (Menon et al. 2011). These methods however are limited to solving non-convex optimization problems.

Convex matrix completion methods have been explored for transductive incomplete multi-label learning in a few recent works (Goldberg et al. 2010; Cabral et al. 2011; Xu, Jin, and Zhou 2013), which use the feature matrix of the instances as side information for label matrix completion. The work in (Goldberg et al. 2010) proposes to perform matrix completion over the concatenation of the input feature matrix and the incomplete label matrix. The work in (Cabral et al. 2011) further adapts the algorithms to address image classifications. The recent work in (Xu, Jin, and Zhou 2013) assumes the label assignments are linear combinations of the feature vectors of the instances, and proposes to speed up convex matrix completion by posing low-rank regularization over the linear combination parameter matrix. The idea of (Xu, Jin, and Zhou 2013) has also been developed in a non-convex form for inductive matrix completion in (Natarajan and Dhillon 2014). A most recent work extends inductive matrix completion by considering noisy side information in a convex formulation with nuclear norm regularization on the parameter matrices (Chiang, Hsieh, and Dhillon 2015). Different from these works, our proposed approach will simultaneously exploit low-rank structures of both the target label matrix and the prediction model to enforce consistent label embeddings, which is expected to enhance the complementary reconstruction of the label matrix from the two information sources.

Label Embedding Label embeddings refer to the distributed vector representations of the label concepts. Here we use *label* in a broad way to refer to *prediction targets*. The learning and employment of label embeddings can facilitate statistical information sharing across different labels and hence improve the learning performance with sparse data. Label embeddings have been exploited in the literature to capture label semantic structures and improve prediction performance in applications such as image tagging (Akata et al. 2013; Weston, Bengio, and Usunier 2010; 2011), multi-class classification (Bengio, Weston, and Grangier 2010), multi-label learning and tag recommendation (Mirzazadeh, Guo, and Schuurmans 2014). These previous methods however do not address learning problems with incomplete labels. Moreover, they mostly rely on local training methods to pursue local optimal solutions. One exception is (Mirzazadeh, Guo, and Schuurmans 2014), which generalizes label embedding into a convex co-embedding framework to handle a set of learning tasks which can be formulated as associations between two types of objects. By contrast, our proposed convex co-embedding framework exploits the low-dimensional associations of three types of objects to enforce consistent label co-embeddings from two information sources for matrix completion.

Approach

In this section, we propose a convex co-embedding solution for general matrix completion problems with side information. In the following presentation, we will refer to the *target matrix* as *label matrix*, which is taken as the prediction target of the auxiliary side feature matrix.

Notation The following notations are used in the presentation: $\mathbf{1}$ is used to denote any column vector with all 1 values, assuming its length can be determined from the context; 1_a is used to denote a $(0,1)$-valued vector with only a single 1 in its a-th entry; I_n denotes an identity matrix of size n; $\mathbf{0}_{n,d}$ denotes a $n \times d$ matrix with all zeros, and $\mathbf{1}_{n,d}$ denotes a $n \times d$ matrix with all 1s. For a vector $\mathbf{x}$, we use $\|\mathbf{x}\|_2$ to denote its Euclidean norm. For a matrix X, we use X_{ij} to denote its (i,j)-th entry, and use X_i to denote its i-th row. Let $[m:n] = \{m, m+1, \cdots, n\}$. We use $X_{[m:n]}$ to denote the submatrix of X formed by the rows with indices in $[m:n]$. We use $\|X\|_F$ to denote the Frobenius norm, use $\|X\|_1 = \sum_{ij} |X_{ij}|$ to denote the entrywise ℓ_1 norm, use $\|X\|_{tr} = \sum_i \sigma_i$ to denote the nuclear norm (trace norm) of X, and use $\|X\|_{sp} = \sigma_{\max}(X) = \sigma_1$ to denote the spectral norm of X, where σ_i denotes the i-th largest singular value of X. We use "$\circ$" to denote the Hadamard product operator such as $[X \circ Y]_{ij} = X_{ij}Y_{ij}$.

Convex Co-Embedding for Matrix Completion

We assume $Y \in \{0,1\}^{N \times L}$ is a target label matrix for N instances over L labels. We only observe a subset of entries of Y and use $\Omega \in \{0,1\}^{N \times L}$ to denote the mask matrix that encodes the observation status for each entry of Y; that is, $\Omega_{ij} = 1$ if Y_{ij} is observed and $\Omega_{ij} = 0$ if Y_{ij} is not observed. Moreover, we also assume the feature vectors for the N instances are given as auxiliary side information, which

forms an input feature matrix $X \in \mathbb{R}^{N\times d}$ with each instance described as a d-dimensional feature vector. We assume d is a reasonable small number such that $d < N$ and use dimensionality reduction to avoid the curse of dimensionality when necessary. We aim to recover the missing entries of Y.

Co-Embedding Perspective of Matrix Completion Due to the existence of statistical correlations between the observations of different labels, we assume the matrix Y is inherently low-rank; i.e., $k = \text{rank}(Y) \ll \min(N, L)$. Standard matrix completion can exploit this low-rank property to recover the target matrix Y by performing a constrained nuclear norm minimization:

$$\min_{\hat{Y}} \quad \|\hat{Y}\|_{tr} \qquad \text{s.t.} \quad \Omega \circ Y = \Omega \circ \hat{Y} \tag{1}$$

This is equivalent to identifying a set of latent factors $Z \in \mathbb{R}^{N\times h}$ for the N instance items and a set of latent factors $B \in \mathbb{R}^{L\times h}$ for the L labels through matrix factorization:

$$\min_{Z,B} \ \frac{1}{2}(\|Z\|_F^2+\|B\|_F^2) \quad \text{s.t.} \ \Omega\circ Y = \Omega\circ(ZB^\top) \tag{2}$$

where the minimization is over Z and B with a dimension value h no less than k. The latent factor matrices Z and B can be viewed as embedding matrices for the instance items and prediction labels respectively, such that each pair of embeddings, e.g., Z_i and B_j, can produce an association score, $Z_i B_j^\top$, to explain the entry $\hat{Y}_{ij}$. This idea is referred to as *co-embedding* since the two sets of objects are embedded into the same space (Mirzazadeh, Guo, and Schuurmans 2014). $\hat{Y}$ hence contains the *co-embedding* scores of the two sets of objects, instance items and prediction labels. The equivalence of (1) and (2) is built based on a well-known identity (Bach, Mairal, and Ponce 2008; Srebro, Rennie, and Jaakkola 2004):

$$\|\hat{Y}\|_{tr} = \min_{Z,B:\, \hat{Y}=ZB^\top} \frac{1}{2}(\|Z\|_F^2 + \|B\|_F^2) \tag{3}$$

Co-Embedding Perspective of Linear Predictions Given the side information matrix X that provides feature descriptions for the instances, we can build a prediction model to predict the entries of the recovered label matrix $\hat{Y}$. The simplest method is to build L predictors independently, one for each label. However, this simple model ignores the correlations/dependencies of the label observations and may degrade the prediction capacity and performance (Zhang and Zhou 2014). Nevertheless, by exploiting label embeddings to build co-embedding based prediction models, we can simultaneously maintain the simplicity of independent prediction models while taking the label dependencies into account. Co-embedding provides a natural way to evaluate the associations between two objects by first embedding them into a common space and then use Euclidean geometry (e.g., inner product) to determine the association score. Given the label embedding matrix $B \in \mathbb{R}^{L\times h}$, we can project the instances from their input feature space into the embedding space with a projection matrix $W \in \mathbb{R}^{d\times h}$. The association score between the i-th instance and the j-th label can then be determined as $s(i, j) = X_i W B_j^\top$, where XW produces the feature-based embeddings of the instances. Note the label embeddings express each label as a vector in terms of h latent common attributes. Hence even the independent computation of the association scores between the instances and each label will naturally capture the label correlations through the latent common attributes. By further considering a bias term for each label, we propose to build the following linear prediction model $f : \mathbb{R}^d \to \mathbb{R}^L$ to predict the entries of the recovered label matrix $\hat{Y}$:

$$f(X_i) = X_i W B^\top + \mathbf{b}^\top \tag{4}$$

where $\mathbf{b} \in \mathbb{R}^L$ is the bias term vector for the L labels. For simplicity, we henceforward will use $f(X) = XWB^\top + \mathbf{1}\mathbf{b}^\top$ to denote the prediction scores from all the N instances. We can train this linear prediction model by minimizing a regularized loss function:

$$\min_{W,\mathbf{b}} \quad \ell(f(X), \hat{Y}) + \frac{\alpha}{2}\|W\|_F^2 \tag{5}$$

Convex Co-Embedding Framework Note the two co-embedding models above share two common components. First, they share the same label embedding matrix B. Second, the prediction target $\hat{Y}$ of $f(X)$ is also the recovery matrix of the matrix completion model such that $\hat{Y} = ZB^\top$. Hence by integrating the co-embedding based matrix completion and linear prediction training together, we can formulate the following joint co-embedding framework:

$$\begin{aligned}\min_{Z,B,W,\mathbf{b}} \ & \ell(f(X), ZB^\top)+\frac{\alpha}{2}\|W\|_F^2+\frac{\gamma}{2}(\|Z\|_F^2+\|B\|_F^2) \\ \text{s.t.} \ & \ \Omega \circ Y = \Omega \circ (ZB^\top)\end{aligned} \tag{6}$$

where $\ell(\cdot, \cdot)$ is a convex loss function in both of its two parameters, α and γ are trade-off parameters. The minimization framework in (6) is in general non-convex due to the existence of the bilinear terms. Below we will reformulate it into a convex learning framework.

Proposition 1 *Let $M \in \mathbb{R}^{(d+N)\times L}$ and let $\tilde{X} = \sqrt{\frac{\gamma}{\alpha}}X$. We define two row selection matrices,*

$$A = [I_d, \mathbf{0}_{d,N}] \quad \textit{and} \quad \overline{A} = [\mathbf{0}_{N,d}, I_N],$$

which extract the first d rows and the last N rows of M respectively, such that $AM = M_{[1:d]}$ and $\overline{A}M = M_{[d+1:d+N]}$. Then the minimization problem (6) over $\{Z, B, W, \mathbf{b}\}$ can be equivalently reformulated into the following convex optimization problem over $\{M, \mathbf{b}\}$ with the same convex loss function $\ell(\cdot, \cdot)$:

$$\begin{aligned}\min_{M,\mathbf{b}} \ & \ell(\tilde{X}AM + \mathbf{1}\mathbf{b}^\top, \overline{A}M) + \gamma\|M\|_{tr} \\ \textit{s.t.} \ & \ \Omega \circ Y = \Omega \circ (\overline{A}M)\end{aligned} \tag{7}$$

Proof: Let $\tilde{W} = \sqrt{\frac{\alpha}{\gamma}}W$. By replacing XW with $\tilde{X}\tilde{W}$, we can rewrite the objective function of (6) equivalently as

$$(6) = \ell(\tilde{X}\tilde{W}B^\top+\mathbf{1}\mathbf{b}^\top, ZB^\top)+\frac{\gamma}{2}\left(\left\|\begin{matrix}\tilde{W}\\ Z\end{matrix}\right\|_F^2+\|B\|_F^2\right) \tag{8}$$

Then we introduce matrix M to replace $\{\tilde{W}, Z, B\}$ with $M = \begin{bmatrix} \tilde{W} \\ Z \end{bmatrix} B^\top$, such that $AM = \tilde{W}B^\top$ and $\overline{A}M = ZB^\top$. Finally based on the equivalence equation (3), we can derive (7). ■

To produce a concrete learning problem, we use a least squares loss function in this work such that

$$\ell(\tilde{X}AM + \mathbf{1}\mathbf{b}^\top, \overline{A}M) = \left\|\tilde{X}AM + \mathbf{1}\mathbf{b}^\top - \overline{A}M\right\|_F^2.$$

Moreover, following a routine of solving equality constrained problems, we relax (7) to an unconstrained optimization problem

$$\begin{aligned}\min_{M,\mathbf{b}} \quad & \|\tilde{X}AM + \mathbf{1}\mathbf{b}^\top - \overline{A}M\|_F^2 \\ & + \rho\|\Omega \circ (Y - \overline{A}M)\|_F^2 + \gamma\|M\|_{tr} \end{aligned} \tag{9}$$

This is equivalent to relaxing the equality constraints into an inequality constraint, $\|\Omega \circ (Y - \overline{A}M)\|_F^2 \leq \delta_\omega$, to handle noise (Candès and Plan 2009).

Theoretical Analysis

Proposition 2 *Let* $\mathbf{h} = [\mathbf{1}_{1,N}, \mathbf{0}_{1,N}]^\top$ *and*

$$\tilde{\Lambda} = \begin{bmatrix} \tilde{X}A - \overline{A} \\ \sqrt{\rho}\overline{A} \end{bmatrix}, \quad \tilde{Y} = \begin{bmatrix} \mathbf{0}_{N,L} \\ \sqrt{\rho}Y \end{bmatrix}, \quad \tilde{\Omega} = \begin{bmatrix} \mathbf{1}_{N,L} \\ \Omega \end{bmatrix}. \tag{10}$$

There exists a proper $\tau > 0$*, such that the minimization problem (9) can be rewritten as below*

$$\min_{M,\mathbf{b}} \ \|\tilde{\Omega} \circ (\tilde{\Lambda}M + \mathbf{h}\mathbf{b}^\top - \tilde{Y})\|_F^2 \quad \textit{s.t.} \ \|M\|_{tr} \leq \tau \tag{11}$$

The re-expression of our proposed model in (11) presents a similar form as the matrix completion with noisy side information model in (Chiang, Hsieh, and Dhillon 2015). But induced from the co-embedding framework our formulation has special structures on both the side information feature matrix $\tilde{\Lambda}$ and the parameter matrix $\mathbf{h}\mathbf{b}^\top$. It is easy to show that the problem (11) has the following closed-form solution for $\mathbf{b}$ by setting the derivative of the objective function regarding $\mathbf{b}$ to zero:

$$\mathbf{b} = \frac{1}{N} M^\top \left(\overline{A} - \tilde{X}A\right)^\top \mathbf{1} \tag{12}$$

Let $f_\theta(i,j) = \tilde{\Lambda}_i M 1_j + \mathbf{h}_i \mathbf{b}_j$ be the prediction function for $\tilde{Y}_{ij}$ parameterized by θ=$\{M, \mathbf{b}\}$, and $\mathcal{F}_\Theta = \{f_\theta | \theta \in \Theta\}$ be the feasible function class. We study the objective loss function of the problem (11) in a relaxed situation where we assume each entry $(i,j) \in \{(i_a, j_a)\}_{a=1}^m$ is sampled i.i.d. under an unknown distribution. Note that the prediction values in the top N rows of $\tilde{Y}$ are automatically obtained constants, hence the assumption above induces no additional requirement but a sampling of the entries in the top N rows. We use $\check{\Omega}$ to denote the relaxed $\tilde{\Omega}$ and then $m = \|\check{\Omega}\|_1$. The objective loss function then can be viewed as an empirical ℓ-risk for function f,

$$\widehat{R}_\ell(f) = \frac{1}{m} \sum\nolimits_{(i,j)\in\check{\Omega}} \ell(f(i,j), \tilde{Y}_{ij})$$

with a squared loss function $\ell(x,y) = |x-y|^2$ and bounded arguments. The corresponding expected ℓ-risk is:

$$R_\ell(f) = \mathbb{E}_{(i,j)}\left[\ell(f(i,j), \tilde{Y}_{ij})\right].$$

Let $\mathcal{L}_\ell$ be a Lipschitz constant for the loss function ℓ with respect to its first argument, and assume it is bounded by $\mathcal{B}_\ell$. Let $\kappa = \max\left(\sqrt{\rho}, \max_i \sqrt{\|\tilde{X}_i\|_2^2 + 1}\right)$, $n_{\max} = \max(2N, L)$ and $d_{\max} = \max(d + N, L)$. We then have the following bound.

Theorem 1 *Consider problem (11) with the closed-form solution for* $\mathbf{b}$ *in (12). Assume the feature matrix* $\tilde{X}$ *is centered around zero mean vector. Then with probability at least* $1-\delta$*, the expected* ℓ*-risk of an optimal solution* f^* *will be bounded by:*

$$\begin{aligned} R_\ell(f^*) \leq \widehat{R}_\ell(f^*) + \mathcal{B}_\ell\sqrt{\frac{\log\frac{2}{\delta}}{2m}} + 4\tau\kappa\mathcal{L}_\ell\sqrt{\frac{\log 2d_{\max}}{m}} + \\ \min\left\{4\mathcal{L}_\ell\tau\sqrt{\frac{\log 2n_{\max}}{m}}, \sqrt{36C\mathcal{L}_\ell\mathcal{B}_\ell\frac{\tau(\sqrt{2N}+\sqrt{L})}{m}}\right\} \end{aligned}$$

The bound above suggests a sample complexity of $O(\tau^2 \log n_{\max})$. A proper chosen τ can significantly reduce the sample complexity. For low-rank matrix Y, this bound can yield a sample complexity of $O(n_{\max} \log n_{\max})$. This shows that by exploiting the co-embedding structure, with predictive auxiliary information from only one side, our model can achieve a lower sample complexity than the $O(n^{3/2})$ reported in (Shamir and Shalev-Shwartz 2014) in a similar distribution-free manner.

Optimization Algorithm

We consider the minimization problem in (9), which is a convex optimization problem with parameters M and $\mathbf{b}$. Given fixed M, by setting the derivative of the objective function regarding $\mathbf{b}$ to zeros, we can derive the same closed-form solution in (12) for $\mathbf{b}$. By plugging (12) back into (9), we obtain the following equivalent minimization problem:

$$\begin{aligned}\min_M \quad F(M) = & \|H(\tilde{X}A - \overline{A})M\|_F^2 + \\ & \rho\|\Omega \circ (Y - \overline{A}M)\|_F^2 + \gamma\|M\|_{tr} \end{aligned} \tag{13}$$

where $H = I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ is a centering matrix. The problem (13) remains to be a convex but non-smooth minimization problem. We develop a fast proximal gradient algorithm (Beck and Teboulle 2009) to solve it with a quadratic convergence rate. The algorithm treats the objective function $F(M)$ of (13) as a combination of the non-smooth nuclear norm regularization term and the remaining convex smooth term $g(M)$ such that $F(M) = g(M) + \gamma\|M\|_{tr}$. In each iteration of the proximal gradient descent algorithm, we need to compute the gradient of $g(M)$ at a given point $Q^{(t)}$ (e.g., in the t-th iteration), $\nabla g(Q^{(t)})$, and then apply a proximal

Algorithm 1 Fast Proximal Gradient Descent Algorithm

Input: $X, Y, \Omega, \rho > 0, \gamma > 0, \eta^* > 0$.
Initialization: $Q^{(1)} = M^{(0)},\ \beta_1 = 1,\ t = 0$.
Repeat
1. Set $t = t + 1$
2. Update: $M^{(t)} = \mathcal{P}_{\eta^*}(Q^{(t)}),\quad \beta_{t+1} = \frac{1+\sqrt{1+4\beta_t^2}}{2}$,
$Q^{(t+1)} = M^{(t)} + \left(\frac{\beta_t - 1}{\beta_{t+1}}\right)(M^{(t)} - M^{(t-1)})$
Until Converge

operator which solves the following intermediate optimization problem for an analytical closed-form solution:

$$\mathcal{P}_\eta(Q^{(t)}) = \arg\min_M \left\{ \frac{\eta}{2}\|M - \widehat{Q}^{(t)}\|_F^2 + \gamma\|M\|_{tr} \right\}$$
$$= U\text{diag}\left((\boldsymbol{\sigma} - \gamma/\eta)_+\right) V^\top \quad (14)$$

where $\widehat{Q}^{(t)} = Q^{(t)} - \frac{1}{\eta}\nabla g(Q^{(t)})$; the U, V and $\boldsymbol{\sigma}$ are the left and right singular vectors and the corresponding singular value vector of $\widehat{Q}^{(t)}$ such as $\widehat{Q}^{(t)} = U\text{diag}(\boldsymbol{\sigma})V^\top$; and $(\cdot)_+ = \max(\cdot, 0)$. With a fast convergence update scheme from (Beck and Teboulle 2009), the overall algorithm is presented in Algorithm 1.

Proposition 3 *Let $\eta^* = 2\sigma_{\max}(\Gamma) + 2\rho$ with $\Gamma = (\tilde{X}A - \overline{A})^\top H(\tilde{X}A - \overline{A})$. Then η^* is the Lipschitz constant of the gradient function ∇g, and the Algorithm 1 has a quadratic rate of convergence $O(1/t^2)$.*

Experiments

We conducted experiments on two types of applications, transductive multi-label learning with incomplete labels and recommendation matrix completion. In this section we present the experimental settings and results.

Transductive Incomplete Multi-label Learning

Experimental setting We conducted experiments for transductive incomplete multi-label learning on ten standard multi-label datasets for web page classification from "yahoo.com" (Ueda and Saito 2002). Each dataset has around 5,000 instances with the number of labels varies from 21 to 40. We treated the label matrix as the target completion matrix and used the web page descriptions as the side feature matrix. We preprocessed the features by reducing the feature vector dimensionality to $d = 500$ with PCA. For each dataset, we randomly sampled 10% instances for testing and used the remaining 90% data for training. The labels for the testing data are completely removed during the training process. Moreover, given a label *observation rate* value $\varepsilon\%$, for each class $\mathcal{L}_j$, we randomly sampled $\varepsilon\%$ positive and negative training instances and kept their label assignment values for class $\mathcal{L}_j$, while ignoring all the other training instances' label assignments for this class. We conducted experiments with a few different $\varepsilon\%$ values in the range of $\{10\%, 30\%, 50\%\}$.

We compared the proposed convex co-embedding method, which we denote as *CoEmbed*, with three state-of-the-art matrix completion methods: *Maxide, IMC* and *DirtyIMC*. *Maxide* is developed for general matrix completion problems with side information, and evaluated on transductive incomplete multi-label learning (Xu, Jin, and Zhou 2013). *IMC* is a non-convex inductive matrix completion method (Natarajan and Dhillon 2014) and *DirtyIMC* is a convex inductive matrix completion method that considers noisy side information (Chiang, Hsieh, and Dhillon 2015). The proposed approach, *CoEmbed*, has three trade-off parameters α, γ and ρ. Note as shown in Proposition 1, any difference between α and γ will simply lead to rescaling the input feature values. Hence we just set $\alpha = \gamma$. Moreover, the ρ parameter controls the degree of the soft approximation for the equality constraints. It should be a reasonably large value. In our experiments, we set $\rho = 100$. We did parameter selection for the regularization parameter γ from the set $2^{\{-9,-8,\cdots,8,9\}}$ by using two-fold cross-validation on the labeled training data. Parameter selection for the three comparison methods are conducted in the same way with the same range of values. For the non-convex *IMC*, we used an inner dimensionality of 200 for the parameter matrices.

Experimental Results For each $\varepsilon\% \in \{10\%, 30\%, 50\%\}$, we randomly partitioned the training and testing data as described above. We performed transductive training using each comparison method on the training and testing data. The label matrix completion results are evaluated on all the unobserved label entries. We used two standard metrics to evaluate the results, average precision (AP) (Zhang and Zhou 2014) and area under the curve (AUC). Both are standard measures used for multi-label learning evaluation. The average results over 10 runs in terms AP and AUC scores are reported in Figure 1 and Figure 2 respectively. We can see *Maxide* outperforms *IMC* and *DirtyIMC* in most cases in terms of AP score, but *DirtyIMC* outperforms *Maxide* in many cases in terms of AUC score. The proposed *CoEmbed* nevertheless consistently produced the best results across all the data sets in terms of both AP and AUC. Moreover, in most cases, *CoEmbed* outperforms the other methods with considerable margins. These results demonstrate the efficacy of the proposed convex co-embedding model.

Recommendation Matrix Completion

We have also conducted recommendation experiments on three real-word Amazon datasets: Beauty, Office and Sports. Each dataset contains the implicit user feedbacks on Amazon products as the target label matrix. These matrices are extremely sparse. After filtering the product items and users that rarely appear, we used the remaining transaction data: Beauty (2,275 product items, 1,965 users, 11,051 transactions), Office (2,107 product items, 1,888 users, 8,307 transactions), and Sports (1,789 product items, 1,324 users, 7,947 transactions). We used the product reviews as side feature information for the items. We preprocessed the reviews into term-frequency feature vectors with the 5,000 most frequent unigram features and then performed dimensionality reduction with PCA to reduce the feature dimension to 500.

We consider performing recommendation matrix completion for *advertisement purpose*. Given observed user-item

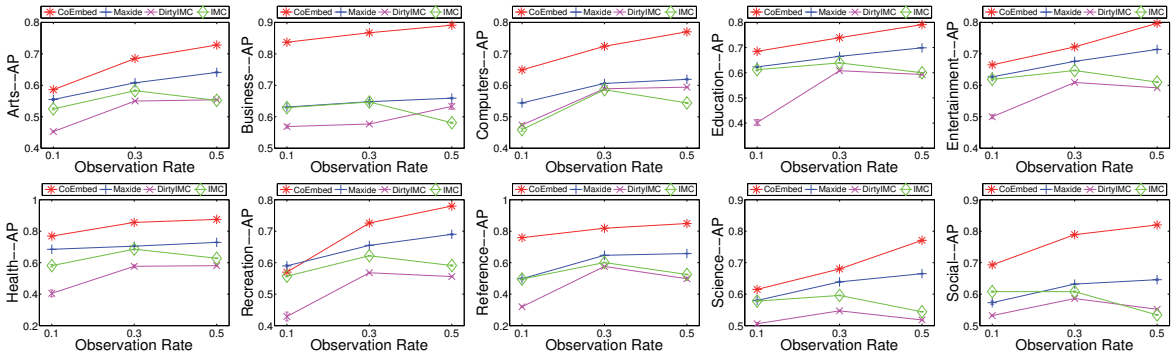

Figure 1: Comparison results in terms of AP on 10 Yahoo datasets with three different observation rates: 10%, 30% and 50%.

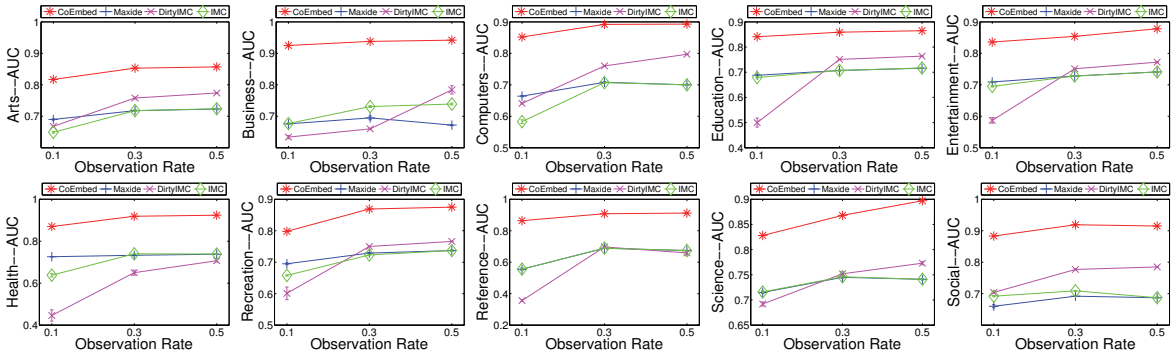

Figure 2: Comparison results in terms of AUC on 10 Yahoo datasets with three different observation rates: 10%, 30% and 50%.

purchase/recommendation history, our goal is to accurately identify the right group of users for each item, to whom we should send advertisement about the particular item. A good recommendation system (matrix completion system) should send the advertisement to users who are most likely to be interested in (or purchase) the item. In this setting, the target label matrix is much larger than the multi-label output matrix. The linear prediction model is built to predict the user tastes from the item features. For each dataset, we randomly dropped 50% entries to perform matrix completion and repeated the experiment five times. We compared our approach, *CoEmbed*, to *Maxide, IMC, DirtyIMC* and a convex collective matrix factorization method (*ConvexCMF*) (Bouchard, Guo, and Yin 2013). We evaluated the matrix completion results using the mean average precision (MAP) at the top-K user recommendations over each item with $K = 5$ and $K = 10$. The average results are reported in Table 1. We can see that *Maxide, IMC* and *DirtyIMC* outperform *ConvexCMF* while our proposed approach outperforms all the four comparison methods on the three datasets. These results again verified the effectiveness of our convex co-embedding method.

Table 1: Matrix completion results in terms of MAP@5 (%) and MAP@10 (%).

Methods		Beauty	Office	Sports
MAP@5	Maxide	19.8± 0.4	20.6± 0.4	21.9± 0.2
	ConvexCMF	12.4± 0.3	9.5± 0.3	15.6± 0.3
	IMC	17.6± 0.4	19.1± 0.4	19.6± 0.2
	DirtyIMC	19.1± 0.1	20.0± 0.1	20.2± 0.0
	CoEmbed	21.1± 0.2	23.0± 0.5	23.5± 0.3
MAP@10	Maxide	20.1± 0.3	20.7± 0.3	22.2± 0.2
	ConvexCMF	12.8± 0.3	10.0± 0.3	16.1± 0.3
	IMC	17.9± 0.4	19.2± 0.4	20.0± 0.2
	DirtyIMC	19.4± 0.1	20.1± 0.1	20.5± 0.0
	CoEmbed	21.4± 0.2	23.1± 0.4	23.8± 0.4

Conclusion

In this paper, we proposed a novel convex co-embedding approach for matrix completion with one side information. It integrates the standard low-rank matrix completion model on the target matrix and the linear prediction model on the auxiliary side information to jointly recover the missing entries of the target matrix within a co-embedding framework. The co-embedding framework can enforce the consistency of the label embeddings induced from the prediction model and from the low-rank target matrix to improve the matrix completion performance. We formulated this framework as a convex minimization problem with nuclear norm regularization, provided a bound analysis, and developed a fast proximal gradient descent algorithm to solve it efficiently. We conducted experiments on two types of applications: trans-

ductive incomplete multi-label learning and recommendation matrix completion. The results show that the proposed approach outperforms a few state-of-the-art methods.

References

Abernethy, J.; Bach, F.; Evgeniou, T.; and Vert, J. 2009. A new approach to collaborative filtering: Operator estimation with spectral regularization. *JMLR* 10:803–826.

Adams, R.; Dahl, G.; and Murray, I. 2010. Incorporating side information in probabilistic matrix factorization with gaussian processes. In *Proc. of UAI*.

Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2013. Label-embedding for attribute-based classification. In *Proc. of CVPR*.

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2008. Convex multi-task feature learning. *MLJ* 73(3):243–272.

Bach, F.; Mairal, J.; and Ponce, J. 2008. Convex sparse matrix factorization. *Technical Report, HAL-00345747*.

Beck, A., and Teboulle, M. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. on Imaging Sciences* 2(1):183–202.

Bengio, S.; Weston, J.; and Grangier, D. 2010. Label embedding trees for large multi-class tasks. In *Proc. of NIPS*.

Bouchard, G.; Guo, S.; and Yin, D. 2013. Convex collective matrix factorization. In *Proc. of AISTATS*.

Cabral, R.; Torre, F.; Costeira, J.; and Bernardino, A. 2011. Matrix completion for multi-label image classification. In *Proc. of NIPS*.

Cai, J.; Candès, E.; and Shen, Z. 2010. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization* 20(4):1956–1982.

Candès, E., and Plan, Y. 2009. Matrix completion with noise. *Proc. of the IEEE* 98(6):925–936.

Candès, E., and Recht, B. 2012. Exact matrix completion via convex optimization. *CACM* 55(6):111–119.

Candès, E., and Tao, T. 2010. The power of convex relaxation: near-optimal matrix completion. *IEEE TIT* 56(5):2053–2080.

Chiang, K.; Hsieh, C.; and Dhillon, I. 2015. Matrix completion with noisy side information. In *Proc. of NIPS*.

Fang, Y., and Si, L. 2011. Matrix co-factorization for recommendation with rich side information and implicit feedback. In *Inter. Workshop on Inform. Heterogeneity and Fusion in Recommender Systems*.

Goldberg, A.; Zhu, X.; Recht, B.; Xu, J.; and Nowak, R. 2010. Transduction with matrix completion: Three birds with one stone. In *Proc. of NIPS*.

Hu, Y.; Koren, Y.; and Volinsky, C. 2008. Collaborative filtering for implicit feedback datasets. In *Proc. of ICDM*.

Keshavan, R.; Montanari, A.; and Oh, S. 2010. Matrix completion from a few entries. *IEEE TIT* 56(6):2980–2998.

Koltchinskii, V.; Lounici, K.; and Tsybakov, A. 2011. Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *The Annals of Statistics* 39(5):2302–2329.

Lin, Z.; Chen, M.; Wu, L.; and Ma, Y. 2009. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. Technical report, UIUC.

Mazumder, R.; Hastie, T.; and Tibshirani, R. 2010. Spectral regularization algorithms for learning large incomplete matrices. *JMLR* 11:2287–2322.

Menon, A.; Chitrapura, K.; Garg, S.; Agarwal, D.; and Kota, N. 2011. Response prediction using collaborative filtering with hierarchies and side-information. In *Proc. of KDD*.

Mirzazadeh, F.; Guo, Y.; and Schuurmans, D. 2014. Convex co-embedding. In *Proc. of AAAI*.

Natarajan, N., and Dhillon, I. S. 2014. Inductive matrix completion for predicting gene-disease associations. *Bioinformatics* 30(12):i60–i68.

Porteous, I.; Asuncion, A.; and Welling, M. 2010. Bayesian matrix factorization with side information and dirichlet process mixtures. In *Proc. of AAAI*.

Recht, B. 2011. A simpler approach to matrix completion. *JMLR* 12:3413–3430.

Rennie, J., and Srebro, N. 2005. Fast maximum margin matrix factorization for collaborative prediction. In *Proc. of ICML*.

Richard, E.; Savalle, P.; and Vayatis, N. 2012. Estimation of simultaneously sparse and low rank matrices. In *Proc. of ICML*.

Shamir, O., and Shalev-Shwartz, S. 2014. Matrix completion with the trace norm: Learning, bounding, and transducing. *JMLR* 15:3401–3423.

Shin, D.; Cetintas, S.; Lee, K.; and Dhillon, I. 2015. Tumblr blog recommendation with boosted inductive matrix completion. In *Proc. of CIKM*.

Srebro, N.; Rennie, J.; and Jaakkola, T. 2004. Large-margin matrix factorization. In *Proc. of NIPS*.

Ueda, N., and Saito, K. 2002. Parametric mixture models for multi-labeled text. In *Proc. of NIPS*.

Weston, J.; Bengio, S.; and Usunier, N. 2010. Large scale image annotation: Learning to rank with joint word-image embeddings. *Machine Learning* 81(1):21–35.

Weston, J.; Bengio, S.; and Usunier, N. 2011. WSABIE: Scaling up to large vocabulary image annotation. In *Proc. of IJCAI*.

Xu, M.; Jin, R.; and Zhou, Z. 2013. Speedup matrix completion with side information: Application to multi-label learning. In *Proc. of NIPS*.

Yi, J.; Yang, T.; Jin, R.; Jain, A.; and Mahdavi, M. 2012. Robust ensemble clustering by matrix completion. In *Proc. of ICDM*.

Zhang, M., and Zhou, Z. 2014. A review on multi-label learning algorithms. *IEEE TKDE* 26(8):1819–1837.