Relational Co-Clustering via Manifold Ensemble Learning

Ping Li Jiajun Bu Chun Chen Zhanying He
Zhejiang Provincial Key Laboratory of Service Robot
College of Computer Science, Zhejiang University
Hangzhou, China, 310027
patriclouis.lee@gmail.com, {bjj, chenc, hezhanying}@zju.edu.cn

ABSTRACT

Co-clustering targets on grouping the samples and features simultaneously. It takes advantage of the duality between the samples and features. In many real-world applications, the data points or features usually reside on a submanifold of the ambient Euclidean space, but it is nontrivial to estimate the intrinsic manifolds in a principled way. In this study, we focus on improving the co-clustering performance via manifold ensemble learning, which aims to maximally approximate the intrinsic manifolds of both the sample and feature spaces. To achieve this, we develop a novel co-clustering algorithm called *Relational Multi-manifold Co-clustering* (RMC) based on symmetric nonnegative matrix tri-factorization, which decomposes the relational data matrix into three matrices. This method considers the inter-type relationship revealed by the relational data matrix and the intra-type information reflected by the affinity matrices. Specifically, we assume the intrinsic manifold of the sample or feature space lies in a convex hull of a group of pre-defined candidate manifolds. We hope to learn an appropriate convex combination of them to approach the desired intrinsic manifold. To optimize the objective, the multiplicative rules are utilized to update the factorized matrices and the *entropic mirror descent algorithm* is exploited to automatically learn the manifold coefficients. Experimental results demonstrate the superiority of the proposed algorithm.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; H.2.8 [**Database Management**]: Database applications—*data mining*

General Terms

Algorithm, Experimentation

Keywords

Co-clustering, manifold ensemble learning, nonnegative matrix tri-factorization, entropic mirror descent algorithm

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

1. INTRODUCTION

A large amount of relational data emerge in a broad range of applications [17], e.g.,various images on the Internet. To handle them, clustering has established itself as a very useful tool and gained an increasing important role in knowledge management and information retrieval, etc. Traditional clustering belongs to unilateral learning paradigm, namely it only emphasizes clustering along the sample or feature dimension individually. Recent works have shown that clustering the samples and features simultaneously, i.e., co-clustering, enables further improving the clustering performance, in the sense that co-clustering fully makes use of the dual interdependence between samples and features to discover hidden structures in data [10].

To this end, many co-clustering algorithms have been proposed, such as graph partitioning based model [10, 19] and matrix factorization based model [12, 22]. In this work, we focus on studying matrix factorization based co-clustering, which models the sample-feature relationship from the data reconstruction perspective. We regard nonnegative matrix tri-factorization [12] as the basis of our co-clustering approach. This typical method imposes the nonnegative constraints on the decomposed matrices, which leads to a parts-based representation [15]. On the other hand, many studies have shown human generated data are usually drawn from a probability distribution that has support on or near to a submanifold [4, 5, 6, 7, 8]. As a result, some researchers strive to consider manifold geometrical structure in co-clustering by dual graph regularization [14, 21]. However, it is nontrivial to seek the intrinsic manifolds. To address this issue, inspired from the work in [13], we propose to approximate the optimal manifold by using a convex combination of some pre-given candidate manifolds, and thus develop a novel co-clustering algorithm called *Relational Multi-manifold Co-clustering*(RMC) to improve the clustering performance via manifold ensemble learning.

We consider the inter-type relation through the relational data matrix and the intra-type information through the affinity matrices encoded on both the sample and feature spaces [21, 22]. In manifold ensemble learning, we assume that the intrinsic manifold of the sample or feature space lies in a convex hull of a group of pre-defined candidate manifolds [13] and hope to approximate the intrinsic manifolds of the sample or feature space by a convex combination of these candidates. To optimize the objective, we alternatively update the factorized matrices and exploit the *entropic mirror descent algorithm* to automatically learn the manifold coefficients. Experiments show its effectiveness.

2. RELATED WORKS

Here, we review some works closely related to our approach. Matrix factorization based co-clustering techniques have been widely studied. The popularized nonnegative matrix factorization (NMF) [15] was proposed to learn a parts-based representation, but it focuses on the unilateral clustering. Motivated by this, the block value decomposition (BVD) was presented for co-clustering dyadic data [18]. It factorizes the data matrix into three components, i.e., the row and column coefficient matrices and the block value matrix. As an extension of NMF, orthogonal nonnegative matrix tri-factorization (ONMTF) was studied in [12], which emphasizes the role of bi-orthogonality in three-factor NMF. Thanks to the successful applications of manifold learning in recent years [2], some researchers consider the local geometrical structure in matrix factorization based co-clustering. For example, a dual regularized co-clustering (DRCC) method based on semi-nonnegative matrix tri-factorization was proposed in [14]. To reduce the computational complexity of DRCC, some fast approaches were proposed [22]. Moreover, a symmetric nonnegative matrix tri-factorization (SNMTF) framework was developed to cluster multi-type relational data [21], which incorporates the intra-type information through manifold regularization.

However, these graph regularization based methods share a shortcoming that the estimated manifold might not be true, and it even deviates far from desired in an adverse situation. To alleviate this problem, inspired by [13], we attempt to approximate the true intrinsic manifold by a convex combination of some candidate manifolds.

3. RELATIONAL MULTI-MANIFOLD CO-CLUSTERING

In this part, we introduce our RMC approach including the optimization framework.

3.1 Problem Formulation

The general problem setting is to co-cluster multi-type relational data. Given a K-type relational data set $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_K\}$, where each $\mathcal{X}_k$ represents the data objects of the k-th type, we define an inter-type relational matrix $\boldsymbol{R}$ with the sub-matrix $\boldsymbol{R}_{ij} \in \mathbb{R}^{n_i \times n_j}$, $i \neq j$, which reflects the inter-type relationship between the i-th type and the j-th type data objects. To model the intra-type structure information, we define an intra-type relational matrix $\boldsymbol{W}$ consisting of a set of affinity matrices $\boldsymbol{W}_k \in \mathbb{R}^{n_k \times n_k}$ encoded on the data, which indicates the intra-type relation of components within the k-th type data.

Typically, we focus on the case $K = 2$, i.e., we employ both the sample and feature type data for co-clustering. The concerned matrices are

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{0}^{n_1 \times n_1} & \boldsymbol{R}_{12}^{n_1 \times n_2} \\ \boldsymbol{R}_{21}^{n_2 \times n_1} & \boldsymbol{0}^{n_2 \times n_2} \end{bmatrix}, \boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_1^{n_1 \times n_1} & \boldsymbol{0}^{n_1 \times n_2} \\ \boldsymbol{0}^{n_2 \times n_1} & \boldsymbol{W}_2^{n_2 \times n_2} \end{bmatrix},$$

where $\boldsymbol{0}$ is a zero matrix. Here, $\boldsymbol{R}_{12}$ and $\boldsymbol{R}_{21}$ represent the feature and sample matrix respectively, $\boldsymbol{R}_{12} = \boldsymbol{R}_{21}^T$. Each column denotes a feature or sample vector.

3.2 Symmetric NMTF

SNMTF [21] decomposes one data matrix into three parts, i.e., ($K = 2$)

$$\boldsymbol{R}_{12} \approx \boldsymbol{G}_1 \boldsymbol{S}_{12} \boldsymbol{G}_2^T, \tag{1}$$

where $\boldsymbol{G}_1 \in \mathbb{R}^{n_1 \times c_1}$ and $\boldsymbol{G}_2 \in \mathbb{R}^{n_2 \times c_2}$ are the cluster indicator matrices for $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively, $c_1 \ll n_1$, $c_2 \ll n_2$. The middle matrix $\boldsymbol{S}_{12} \in \mathbb{R}^{c_1 \times c_2}$ can be treated as a compact representation of $\boldsymbol{R}_{12}$ [18], which absorbs the different scales of other matrices [12]. Note that BVD and ONMTF impose nonnegative constraints on all three matrices $\boldsymbol{G}_1$, $\boldsymbol{G}_2$, $\boldsymbol{S}$, while DRCC and SNMTF both relax the nonnegative constraint on the matrix $\boldsymbol{S}$, thereby allowing negative entries. Based on DRCC, SNMTF employs the symmetric matrix to simultaneously cluster multi-type relational data, and its objective is

$$\min_{\boldsymbol{G},\boldsymbol{S}} \|\boldsymbol{R} - \boldsymbol{G}\boldsymbol{S}\boldsymbol{G}^T\|_F^2 + 2\lambda \mathrm{Tr}(\boldsymbol{G}^T \boldsymbol{L} \boldsymbol{G}), \quad s.t.\ \boldsymbol{G} \succeq 0, \tag{2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ is the graph Laplacian [2], $\boldsymbol{D}$ is the diagonal matrix with $\boldsymbol{D}_{k_{(ii)}} = \sum_j \boldsymbol{W}_{k_{(ij)}}$, and $\boldsymbol{L}_k = \boldsymbol{D}_k - \boldsymbol{W}_k$. The relational matrices $\boldsymbol{G}$ and $\boldsymbol{S}$ are designed below

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{G}_1^{n_1 \times c_1} & \boldsymbol{0}^{n_1 \times c_2} \\ \boldsymbol{0}^{n_2 \times c_1} & \boldsymbol{G}_2^{n_2 \times c_2} \end{bmatrix}, \boldsymbol{S} = \begin{bmatrix} \boldsymbol{0}^{c_1 \times c_1} & \boldsymbol{S}_{12}^{c_1 \times c_2} \\ \boldsymbol{S}_{21}^{c_2 \times c_1} & \boldsymbol{0}^{c_2 \times c_2} \end{bmatrix}.$$

3.3 Manifold Ensemble Learning

It is challenging to discover an appropriate intrinsic manifold for graph based co-clustering methods in reality. In this work, we adopt a novel learning paradigm named *manifold ensemble learning* to maximally approximate the true intrinsic manifold. This idea is inspired from the work in [13], which combines the automatic intrinsic manifold approximation and semi-supervised classification.

We assume a series of initial guesses of graph Laplacian are available and the intrinsic manifold of the sample or feature space lies in the convex hull of these pre-given candidate manifolds. In some sense, this assumption constrains the search space, since the optimal graph Laplacian is an discrete approximation to the intrinsic manifold [13], i.e.,

$$\boldsymbol{L} = \sum_{i=1}^{q} \mu_i \tilde{\boldsymbol{L}}_i, \ s.t.\ \sum_{i=1}^{q} \mu_i = 1, \mu_i \geq 0, \tag{3}$$

where a set of candidate graph Laplacians $C = \{\boldsymbol{L}_1, \ldots, \boldsymbol{L}_q\}$ is defined. Here we use $\tilde{\boldsymbol{L}}_i$ of the i-th candidate manifold to discriminate it from $\boldsymbol{L}_k$ of the k-th type data.

3.4 Objective Function

We take advantage of the manifold ensemble learning into the symmetric nonnegative matrix tri-factorization framework, and thus propose a novel co-clustering approach named *Relational Multi-manifold Co-clustering* (RMC).

Now, it is easy to arrive at our objective, i.e.,

$$\min_{\boldsymbol{G},\boldsymbol{S},\boldsymbol{\mu}} \|\boldsymbol{R} - \boldsymbol{G}\boldsymbol{S}\boldsymbol{G}^T\|_F^2 + \alpha \mathrm{Tr}[\boldsymbol{G}^T (\sum_{i=1}^{q} \mu_i \tilde{\boldsymbol{L}}_i) \boldsymbol{G}] + \beta \|\boldsymbol{\mu}\|_2^2,$$
$$s.t.\ \sum_{i=1}^{q} \mu_i = 1, \boldsymbol{\mu} \succeq 0, \boldsymbol{G} \succeq 0, \tag{4}$$

where $\alpha > 0, \beta > 0$, the tradeoff parameter α is used to govern the contribution of the ensemble manifold regularization to the objective, the l_2-norm of $\boldsymbol{\mu}$ is employed to avoid the coefficient parameter over-fitting to only one manifold and the factor β acts as an over-fitting tolerance parameter for the manifold coefficients. Similar to [14], we do l_2 normalization on columns of $\boldsymbol{G}$ and compensate its norm to $\boldsymbol{S}$. The

components of $\boldsymbol{G}$, i.e., $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ represent the partition matrices of the feature matrix $\boldsymbol{R}_{21}$ and the sample matrix $\boldsymbol{R}_{12}$, respectively. We typically use the partition matrices to derive the co-clustering results.

3.5 Optimization

In this section, we explore how to optimize the objective in Eq. (3.4). It can be readily found that the objective function is non-convex in $\boldsymbol{G}, \boldsymbol{S}, \boldsymbol{\mu}$ jointly, but it is convex in them respectively. So it is unrealistic to find the global minimum since no closed-form solution can be obtained. We present an alternating scheme to optimize the objective involving optimizing the manifold coefficient vector.

3.5.1 Computation of $\boldsymbol{S}$

When fixing $\boldsymbol{G}$ and $\boldsymbol{\mu}$, the objective becomes minimizing $J_S = \|\boldsymbol{R} - \boldsymbol{G}\boldsymbol{S}\boldsymbol{G}^T\|_F^2$. Taking its derivative to $\boldsymbol{S}$ and setting it to 0, then we have the update rule

$$\boldsymbol{S} = (\boldsymbol{G}^T\boldsymbol{G})^{-1}\boldsymbol{G}^T\boldsymbol{R}\boldsymbol{G}(\boldsymbol{G}^T\boldsymbol{G})^{-1}. \quad (5)$$

3.5.2 Computation of $\boldsymbol{G}$

When fixing $\boldsymbol{S}$ and $\boldsymbol{\mu}$, the objective w.r.t. $\boldsymbol{G}$ reduces to minimizing

$$J_G = \|\boldsymbol{R} - \boldsymbol{G}\boldsymbol{S}\boldsymbol{G}^T\|_F^2 + \alpha\mathrm{Tr}(\boldsymbol{G}^T\boldsymbol{L}\boldsymbol{G}),\ s.t.\ \boldsymbol{G} \succeq 0.$$

To solve this problem, we introduce the Lagrangian multiplier matrix $\boldsymbol{\Lambda}$ and its Lagrangian function is formulated as

$$L(\boldsymbol{G}) = \|\boldsymbol{R} - \boldsymbol{G}\boldsymbol{S}\boldsymbol{G}^T\|_F^2 + \alpha\mathrm{Tr}(\boldsymbol{G}^T\boldsymbol{L}\boldsymbol{G}) + \mathrm{Tr}(\boldsymbol{\Lambda}\boldsymbol{G}^T). \quad (6)$$

Requiring its derivative to $\boldsymbol{G}$ be 0, we obtain $\boldsymbol{\Lambda} = 4\alpha\boldsymbol{L}\boldsymbol{G} - 4\boldsymbol{A} + 4\boldsymbol{G}\boldsymbol{B}$, where $\boldsymbol{A} = \boldsymbol{R}\boldsymbol{G}\boldsymbol{S}^T$, $\boldsymbol{B} = \boldsymbol{S}^T\boldsymbol{G}^T\boldsymbol{G}\boldsymbol{S}$.

Since the Karush-Kuhn-Tucker (KKT) condition for the nonnegativity of $\boldsymbol{G}_{ij}$ gives $\boldsymbol{\Lambda}_{ij}\boldsymbol{G}_{ij} = 0$, we have

$$(\alpha\boldsymbol{L}\boldsymbol{G} - \boldsymbol{A} + \boldsymbol{G}\boldsymbol{B})_{ij}\boldsymbol{G}_{ij} = 0. \quad (7)$$

Similar to [11], we define $\boldsymbol{L} = \boldsymbol{L}^+ - \boldsymbol{L}^-$, $\boldsymbol{A} = \boldsymbol{A}^+ - \boldsymbol{A}^-$, $\boldsymbol{B} = \boldsymbol{B}^+ - \boldsymbol{B}^-$, where

$$\boldsymbol{L}_{ij}^+ = (|\boldsymbol{L}_{ij}| + \boldsymbol{L}_{ij})/2,\ \boldsymbol{L}_{ij}^- = (|\boldsymbol{L}_{ij}| - \boldsymbol{L}_{ij})/2.$$

We substitute the decomposed positive and negative parts into Eq. (7), which leads to the update rule

$$\boldsymbol{G}_{ij} \leftarrow \boldsymbol{G}_{ij}\left[\frac{(\alpha\boldsymbol{L}^-\boldsymbol{G} + \boldsymbol{A}^+ + \boldsymbol{G}\boldsymbol{B}^-)_{ij}}{(\alpha\boldsymbol{L}^+\boldsymbol{G} + \boldsymbol{A}^- + \boldsymbol{G}\boldsymbol{B}^+)_{ij}}\right]^{\frac{1}{2}}. \quad (8)$$

3.5.3 Computation of $\boldsymbol{\mu}$

When fixing $\boldsymbol{G}$ and $\boldsymbol{S}$, the objective is simplified to

$$\min_{\boldsymbol{\mu}} f(\boldsymbol{\mu}) = \sum_{i=1}^{q}\mu_i s_i + \beta\|\boldsymbol{\mu}\|_2^2,\ s.t.\ \sum_{i=1}^{q}\mu_i = 1, \boldsymbol{\mu} \succeq 0, \quad (9)$$

where $s_i = \mathrm{Tr}(\boldsymbol{G}^T\tilde{\boldsymbol{L}}_i\boldsymbol{G})$. It is easy to see that if $\beta = 0$, we will get a trivial solution, which is undesirable to learn a composite manifold. If $\beta \to \infty$, all candidate manifolds will receive identical weights, which is also unexpected. Therefore, it is essential to assign a proper parameter β.

It is actually an exactly well-defined problem, i.e., the convex minimization over the unit simplex, which can be solved by the *Entropic Mirror Descent Algorithm* (EMDA) with a global efficiency estimate [1], as shown in Algorithm 1. It has been shown that EMDA owns the natural advantage to solve this convex problem over the unit simplex

Algorithm 1 Entropic Mirror Descent Algorithm

Input: L_f, β, $\boldsymbol{s}$.
Output: $\boldsymbol{\mu}$.
1: **Initialize:** $\mu_i = 1/q$.
2: **for** $i = 1$ to q **do**
3: Compute $t_m = \sqrt{\frac{2\ln q}{mL_f^2}}$, where m is the m-th iteration.
4: Update each μ_i by this rule

$$\mu_i^{m+1} \leftarrow \frac{\mu_i^m \exp[-t_m f'(\mu_i^m)]}{\sum_{i=1}^{q}\mu_i^m \exp[-t_m f'(\mu_i^m)]},$$

where $f'(\mu_i^m) = 2\beta\mu_i^m + s_i$.
5: Repeat Step 3 to 4 until convergence.
6: **end for**

Algorithm 2 Relational Multi-manifold Co-clustering

Input: Relational data matrices $\boldsymbol{R}$ and $\boldsymbol{W}$, the number of sample and feature clusters c_1 and c_2, the tradeoff parameters α and β.
Output: Partition matrix $\boldsymbol{G}$.
1: **Initialize:** Generate $\boldsymbol{G}$ using k-means.
2: Compute $\boldsymbol{S} = (\boldsymbol{G}^T\boldsymbol{G})^{-1}\boldsymbol{G}^T\boldsymbol{R}\boldsymbol{G}(\boldsymbol{G}^T\boldsymbol{G})^{-1}$.
3: Compute the manifold coefficient $\boldsymbol{\mu}$ using Algorithm 1.
4: Update the matrix $\boldsymbol{G}$ according to

$$\boldsymbol{G}_{ij} \leftarrow \boldsymbol{G}_{ij}\left[\frac{(\alpha\boldsymbol{L}^-\boldsymbol{G} + \boldsymbol{A}^+ + \boldsymbol{G}\boldsymbol{B}^-)_{ij}}{(\alpha\boldsymbol{L}^+\boldsymbol{G} + \boldsymbol{A}^- + \boldsymbol{G}\boldsymbol{B}^+)_{ij}}\right]^{\frac{1}{2}}.$$

5: Repeat Step 2 to 4 until convergence.

$\triangle = \{\boldsymbol{\mu} \in \mathbb{R}^q : \sum_{i=1}^{q}\mu_i = 1, \boldsymbol{\mu} \succeq 0\}$. To apply this algorithm, the objective function f should be a convex Lipschitz continuous function with Lipschitz constant L_f with respect to a fixed given norm. We derive this Lipschitz constant from $\|\nabla f(\boldsymbol{\mu})\|_1 \leq 2\beta + \|\boldsymbol{s}\|_1 = L_f$, where $\boldsymbol{s} = \{s_1, \ldots, s_q\}$. Here, we use $\|\cdot\|_1$ norm as suggested in [1].

In summary, we present our *Relational Multi-manifold Co-clustering* (RMC) approach in Algorithm 2. Note that here we omit the convergence proof due to space limit. Please refer to [5, 14].

4. EXPERIMENTS

In this section, we investigate the clustering performance of the proposed method on a broad range of data sets.

4.1 Data Corpora

The important statistics of the data corpora are summarized in Table 1 and the brief descriptions are shown below.

Text corpora. NGroups5 is selected from the newsgroup data collection 20Newsgroups[1]. We use a subset, which contains five different topics, which refer to 4,052 documents. RCV1-5 is a 5-topic subset of a smaller RCV1 [2] collection [9]. We select 1,200 words with the highest contribution to the mutual information between words and documents.

Image databases. The AlphaDigit[3] is a handwritten image data that contains 1,404 images, covering 20×16 dig-

[1]http://people.csail.mit.edu/jrennie/20Newsgroups/
[2]http://www.daviddlewis.com/resources/testcollections/rcv1/
[3]http://cs.nyu.edu/∼roweis/data.html

Table 1: Statistics of the data sets

Data Sets	Domain	Samples	Features	Classes
NGroups5	text	4,052	1,200	5
RCV1-5	text	3,012	1,200	5
AlphaDigit	image	1,404	320	36
UMIST	image	575	644	20
Leukemia2	gene	72	5,551	3
LungCancer	gene	203	2,008	5

its of "0" through "9" and capitals "A" through "Z". UMIST[4] is a face database referring to a range of poses from profile to frontal views. Each image is rescaled to 28×23 pixels.

Gene expression data. Leukemia2 and LungCancer[5] are produced by oligonucleotide-based technology [20]. Similar to [16], we removed the genes which vary little across samples to reduce the computational complexity.

4.2 Performance Comparison

To explore the clustering performance of the proposed RMC algorithm, we compare it with some state-of-the-art approaches: NMF [15], GNMF[5], DRCC [14],ONMTF [12], SNMTF [21] and OSNTF [22]. K-means (KM) is treated as a baseline.

In the experiments, we ran k-means 20 times with different starting points and the best result in terms of the objective function was recorded. The average results over 20 test runs are reported. We adopt two popular criteria to measure the clustering performance [3], i.e., the *accuracy* (AC) and the *normalized mutual information* (NMI).

4.3 Parameter Settings

Except KM, the number of sample or feature clusters is set to the actual number of classes in data sets. Note that there is no parameter selection for KM, NMF and ONMTF, once the number of clusters is given. For GNMF, DRCC, SNMTF, OSNTF and RMC, p is fixed to 5. The regularization parameters are all searched from the grid {0.001, 0.01, 0.1, 1, 10, 100, 500, 1500}. For co-clustering methods, the regularization parameters for the sample and feature graph are set to the same. Except RMC, the other graph-based methods construct the Laplacian matrix using the binary weighting scheme [5].

For RMC, we empirically set $\beta = 0.1\alpha$ [13]. To generate diverse manifolds, we utilized three kinds of weighting schemes to construct the graph, i.e., the binary weighting, the Gaussian kernel, and the cosine similarity. In particular, for the Gaussian kernel, we varied the bandwidth t in a broad range of area, i.e., $t = \{\frac{\tau}{100}, \frac{\tau}{60}, \frac{\tau}{30}, \frac{\tau}{10}, \tau, 10\tau, 30\tau, 60\tau, 100\tau\}$, where $\tau = \left(\frac{1}{n^2}\sum_{i,j=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right)^{-1}$. In total, *eleven* manifolds containing *nine* Gaussian graphs, *one* binary graph and *one* cosine similarity graph were used.

4.4 Results

The average results are tabulated in Table 2 and 3. Several interesting points can be revealed below.

- RMC is consistently better than other algorithms, which verifies that the manifold ensemble learning is advantageous to the graph-based symmetric nonnegative matrix factorization methods for co-clustering.
- The graph regularization based methods, e.g., GNMF, DRCC, SNMTF, OSNTF and RMC, almost perform better than KM, NMF and OSNMTF. This is for the reason that the graph-based approaches consider the local geometrical structure.
- The dual regularization based methods, e.g., DRCC and RMC, generally outperform the one-side method GNMF. This demonstrates that the geometrical structures in both the sample and feature spaces are beneficial to further improve the clustering performance.

4.5 Parameter Selection

Since p is fixed to 5 and $\beta = 0.1\alpha$, we only explore the influences of α here. Figure 1 shows the results ($\alpha = \{0.001, 0.01, 0.1, 1, 10, 100, 500, 1500\}$), which reflect that our approach enjoys satisfactory performances when α takes a higher value, e.g., around 500 or 1000. This indicates that the ensemble manifold regularization term should be imposed larger weights, such that it can make more positive contributions to the objective.

5. CONCLUSIONS

This paper presents a novel co-clustering approach named *Relational Multi-manifold Co-clustering* (RMC), which is based on the symmetric nonnegative matrix tri-factorization. It takes into account the inter-type relation and the intra-type information of both the sample and feature data simultaneously. The basic idea is to make use of manifold ensemble learning to enhance the performance of co-clustering. To achieve this, we attempt to learn a sensible convex combination of candidate manifolds so that it can maximally approximate the true intrinsic manifolds of both the sample and feature spaces. In order to optimize the objective function, we adopt the popular alternating optimization method to update the factorized matrices. However, different from the existing matrix factorization based co-clustering methods, there is a manifold coefficient vector to be optimized in our approach, which poses a challenging task. In this work, we utilize the *entropic mirror descent algorithm* to optimize this coefficient vector. The effectiveness of the proposed approach is demonstrated by a number of interesting experiments on data collections from diverse domains.

6. ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China under Grants 91120302, 60905001 and 61173186, National Basic Research Program of China (2011CB302206), the Fundamental Research Funds for the Central Universities (2012FZA5017) and the Zhejiang Province Key S&T Innovation Group Project (2009R50009).

7. REFERENCES

[1] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[2] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: a geometric framework for learning from examples. *JMLR*, 7:2399–2434, 2006.

[3] D. Cai, X. He, and J. Han. Document clustering using locality preserving indexing. *IEEE TKDE*, 17(12):1624–1637, 2005.

[4]http://images.ee.umist.ac.uk/danny/database.html

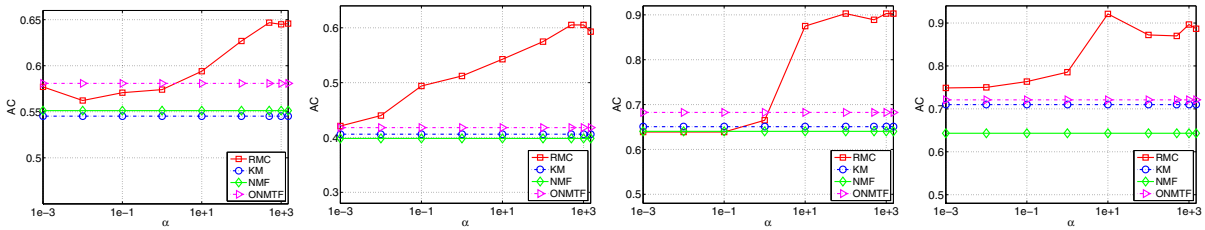
[5]http://www.gems-system.org/datasets/

Table 2: Clustering accuracy of different algorithms (%)

Data Sets	KM	NMF	GNMF	DRCC	ONMTF	SNMTF	OSNTF	RMC
NGroups5	30.70	36.92	44.03	50.37	38.15	47.09	45.56	**51.99**
RCV1-5	54.52	55.12	58.53	61.81	58.08	60.70	60.82	**64.67**
AlphaDigit	43.09	39.67	41.10	45.41	42.18	45.29	42.88	**46.37**
UMIST	40.60	39.83	53.91	55.36	41.80	57.04	54.78	**60.52**
Leukemia2	65.07	64.03	87.50	88.04	68.25	88.40	72.22	**90.28**
LungCancer	70.99	64.31	82.76	83.45	72.10	84.36	75.86	**89.66**

Table 3: Normalized mutual information of different algorithms (%)

Data Sets	KM	NMF	GNMF	DRCC	ONMTF	SNMTF	OSNTF	RMC
NGroups5	13.41	16.62	27.69	36.11	18.66	35.02	28.70	**37.65**
RCV1-5	44.06	45.19	48.78	51.50	46.25	48.94	48.64	**52.02**
AlphaDigit	58.40	55.22	56.01	60.21	57.44	60.69	56.44	**62.28**
UMIST	60.14	58.73	71.03	72.30	59.73	73.72	70.97	**76.61**
Leukemia2	53.65	49.19	65.81	68.25	54.11	69.45	49.43	**71.45**
LungCancer	54.53	53.70	63.46	65.20	55.69	71.07	54.71	**73.97**



(a) RCV1-5 (b) UMIST (c) Leukemia2 (d) LungCancer

Figure 1: Clustering accuracy under varied α of RMC.

[4] D. Cai, X. He, and J. Han. Locally consistent concept factorization for document clustering. *IEEE TKDE*, 23(6):902–913, 2011.

[5] D. Cai, X. He, J. Han, and T. Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE TPAMI*, 33(8):1548–1560, 2011.

[6] D. Cai, X. He, X. Wang, H. Bao, and J. Han. Locality preserving nonnegative matrix factorization. In *IJCAI*, 2009.

[7] D. Cai, X. He, X. Wu, and J. Han. Non-negative matrix factorization on manifold. In *ICDM*, 2008.

[8] D. Cai, X. Wang, and X. He. Probabilistic dyadic data analysis with local and global consistency. In *ICML*, 2009.

[9] W. Chen, Y. Song, H. Bai, C. Lin, and E. Chang. Parallel spectral clustering in distributed systems. *IEEE TPAMI*, 33(3):568–586, 2011.

[10] I. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *SIGKDD*, 2001.

[11] C. Ding, T. Li, and M. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE TPAMI*, 32(1):45–55, 2010.

[12] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *SIGKDD*, 2006.

[13] B. Geng, C. Xu, D. Tao, L. Yang, and X. Hua. Ensemble manifold regularization. In *CVPR*, 2009.

[14] Q. Gu and J. Zhou. Co-clustering on manifolds. In *SIGKDD*, 2009.

[15] D. Lee, H. Seung, et al. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[16] W. Liu, K. Yuan, and D. Ye. On α-divergence based nonnegative matrix factorization for clustering cancer gene expression data. *Artificial Intelligence in Medicine*, 44(1):1–5, 2008.

[17] B. Long, Z. Zhang, X. Wu, and P. Yu. Spectral clustering for multi-type relational data. In *ICML*, 2006.

[18] B. Long, Z. Zhang, and P. Yu. Co-clustering by block value decomposition. In *SIGKDD*, pages 635–640, 2005.

[19] M. Rege, M. Dong, and F. Fotouhi. Co-clustering documents and words using bipartite isoperimetric graph partitioning. In *ICDM*, 2006.

[20] A. Statnikov, C. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631–643, 2005.

[21] H. Wang, H. Huang, and C. Ding. Simultaneous clustering of multi-type relational data via symmetric nonnegative matrix tri-factorization. In *CIKM*, 2011.

[22] H. Wang, F. Nie, H. Huang, and C. Ding. Nonnegative matrix tri-factorization based high-order co-clustering and its fast implementation. In *ICDM*, 2011.