Multi-View Embedding Learning for Incompletely Labeled Data

Wei Zhang, Ke Zhang, Pan Gu, Xiangyang Xue
School of Computer Science, Fudan University, China
{weizh,k_zhang,gupan,xyxue}@fudan.edu.cn

Abstract

In many applications, the data may be high dimensional, represented by multiple features, and associated with more than one labels. Embedding learning is an effective strategy for dimensionality reduction and for nearest neighbor search in massive datasets. We propose a novel method to seek compact embedding that allows efficient retrieval with incompletely-labeled multi-view data. Based on multi-graph Laplacian, we achieve the optimal combination of heterogeneous features to effectively describe data, which exploits the feature correlations between different views. We learn the embedding that preserves the neighborhood context in the original spaces, and obtain the complete labels simultaneously. Inter-label correlations are sufficiently leveraged in the proposed framework. Our goal is to find the maps from multiple input spaces to the compact embedding space and to the semantic concept space at the same time. There is semantic gap between the input multi-view feature spaces and the semantic concept space; and the compact embedding space can be looked on as the bridge between the above spaces. Experimental evaluation on three real-world datasets demonstrates the effectiveness of the proposed method.

1 Introduction

Nowadays we are inundated with abundant data such as images, videos, documents, web pages, etc. The main characteristics of these datasets are as follows. i) High-dimension multi-view: For the low-level feature, the feature vector is high-dimensional and often includes multiple kinds of features, i.e., these features are in different spaces and thus heterogeneous; ii) Multi-concept incomplete-label: For the high-level feature, one datum might be associated with more than one semantic concept simultaneously, and semantic richness requires multiple labels to sufficiently describe the datum; however, it is difficult to label the data without missing any concept, and generally the label information of each sample is incomplete. There is semantic gap between the input multi-view feature spaces (low-level) and the semantic concept space (high-level); and it is of significance to learn the bridge between the above spaces.

In real-world applications, there are many datasets with multiple views, that is, one datum point may be represented in several different feature spaces. For example, web images can be described by heterogenous features such as color descriptors, texture descriptors, shape descriptors, and the surrounding texts. Another example is the web categorization task where the web can be described by either the words occurring in web pages or the hyperlinks between web pages [Zhou and Burges, 2007]. We should consider learning from data with multiple views to effectively use multiple representations simultaneously, and such learning issue is usually called multi-view learning. One approach to multi-view learning is co-training [Kumar and Daume, 2011] where multiple learning algorithms are trained for each view and the relationship between a pair of points should be consistent across different views. [Zhou *et al.*, 2007] takes advantage of the correlation between the views using CCA(Canonical Correlation Analysis) [Hotelling, 1936] and performs semi-supervised learning with only one labeled training sample. In [Zhou and Burges, 2007; Fu *et al.*, 2011], multi-view spectral clustering is performed by generalizing Normalized Cut from the single view to the multi-view case, but they ignore the feature correlations between views. Other multi-view learning algorithms are included in [Harel and Mannor, 2011; Quadrianto and Lampert, 2011; Bronstein *et al.*, 2010; Kumar and Udupa, 2011; Kumar *et al.*, 2011; Dhillon *et al.*, 2011a]. However, these methods do not take into account the correlation between concepts which will affect the performance in the multi-label setting.

Multi-label learning deals with the data associated with more than one concepts simultaneously and is often applied to image/video annotation, text categorization, web page classification, and so on. Multi-label learning has received many attentions in the field of machine learning recently, such as TagProp [Guillaumin *et al.*, 2009], MBRM [Feng *et al.*, 2004], ML-GRF [Zha *et al.*, 2009], AFSVM [Chen *et al.*, 2010], RankSVM [Elisseeff and Weston, 2002], RML [Petterson and Caetano, 2010] and references therein. In multi-label classification, the correlations between labels can be captured to improve the performance of classifiers. For example, the concepts $camel$ and $desert$ often co-occur in the same image, while $panda$ and $desert$ may seldom co-occur. Among various studies on multi-label learning, it is still unclear how to learn from multi-view data and how to capture inter-label correlations for embedding learning at the same time. In [Sun *et al.*, 2011], some extensions of CCA(Canonical Correlation Analysis) are proposed for multi-view multi-label learning. One assumption in many multi-label learning algorithms is that for each training sample, all of its associated labels are provided completely, which may not hold in real applications however. In practice, it is hard to get all the proper labels, and usually the label information of samples is incomplete. For example, given one image containing the concepts $bird$, sky, $cloud$ and $tree$, one may only label the image with $bird$ and sky while missing $cloud$ and $tree$. In this case, only an incomplete label set is available, and therefore, for some label which has not been assigned to the sample, the conclusion can not be drawn that this label is not proper for the sample. The existing methods on incompletely labeled data include [Lee and Liu, 2003; Sun *et al.*, 2010; Lu *et al.*, 2012; Liu *et al.*, 2010a; Chen *et al.*, 2010; Wang *et al.*, 2007]; above methods only allows for inter-label correlations but not for inter-feature correlations because they are not developed for multi-view data.

Embedding learning is a popular approach to deal with high-dimensional data, which assumes that all data points reside on an intrinsic manifold and find its embedding in a low-dimensional space. In the past years many algorithms on embedding learning have been proposed [Roweis and Saul, 2000; Belkin and Niyogi, 2001; He *et al.*, 2005; Chen *et al.*, 2005], but they are not designed for multi-view multi-label learning. In [Dhillon *et al.*, 2011b], a learning method is presented to estimate low dimensional context-specific word representations from unlabeled data using a spectral method; [Zhang and Zhou, 2010] performs multi-label dimensionality reduction by maximizing the dependence between the original feature description and the associated class labels; however, it is not clear how to exploit the weak supervised information in the incomplete label problem.

In this paper we propose a novel method to learn compact embedding that captures inter-feature correlations, inter-label correlations, and feature-label associations simultaneously from multi-view incompletely-labeled data. Each data-point is represented by heterogeneous features in different spaces, so multi-view learning is performed to obtain the optimal combined features which exploit the feature correlations across different views. We assume that data are in three kinds of spaces: i) multiple input feature spaces, ii) compact embedding space, and iii) a semantic concept space. By mapping data from multiple input feature spaces to the embedding space and to the concept space, we seek the embedding which preserves the neighborhood context in the original spaces, and complete the labels at the same time. There is semantic gap between the input multi-view feature space and the semantic concept space; and the compact embedding space can be looked on as the bridge between the above spaces. Each data-point is associated with more than one semantic concepts, so inter-label relations are exploited for embedding learning. The proposed method is weakly supervised because the label information available is incomplete.

The rest of this paper is organized as follows: In Section 2 we formulate the proposed model for embedding learning from multi-view incompletely-labeled data. Our experimental results on real datasets are given in Section 3. Finally, we conclude this paper in Section 4.

2 The Proposed Method

Suppose that there are n incompletely labeled samples: $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$, among which each is represented by multiple heterogeneous features; $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n] \in \{0,1\}^{c\times n}$, where $\mathbf{y}_u \in \{0,1\}^c$ is the label vector of the incompletely labeled sample $\mathbf{x}_u, (u = 1, ..., n)$ and c is the number of labels considered. $\mathbf{y}_u^l = 1(l = 1, \dots, c)$ indicates that the l-th label is a proper label for $\mathbf{x}_u$, while $\mathbf{y}_u^l = 0$ tells us nothing (i.e., the absence of a label does not mean that it is not appropriate for the sample). In the multi-label scenario, one datum may be associated with multiple concepts simultaneously, thus there may be more than one positive elements in the c-dimensional binary vector $\mathbf{y}_u$. Let $\mathbf{H} = [\mathbf{h}_1, ..., \mathbf{h}_n] \in \mathbb{R}^{r\times n}$ denote the embedding for all n samples in the r-dimensional space, and the u-th column $\mathbf{h}_u$ corresponds to the embedding code for the sample $\mathbf{x}_u$. We try to seek the compact embedding and to complete the labels of all data-points represented by heterogeneous features in a unified framework. Intuitively, assume that data points lie on a r-dimensional manifold embedded in multiple input feature spaces, our goal is to find the maps from multiple input spaces to the compact embedding space and to the semantic concept space at the same time. It is worthy to note that the multiple input feature spaces are heterogeneous; there is semantic gap between the input multi-view feature space and the semantic concept space; and the compact embedding space can be looked on as the bridge between the above spaces.

2.1 Multi-View Learning with Heterogeneous Features

For multi-view data , it is of significance how to identify the similarity between samples represented by multiple features. Suppose that each datum is represented by J heterogeneous features, and we define J^2 directed graphs $G_{ij} = (V, W_{ij})$, $i, j = 1, ..., J$ over the dataset, where V is the set of nodes and each node represents one sample; $W_{ij} = [w_{ij}(u,v)]_{n\times n}$ is the weight matrix and $w_{ij}(u,v)$ measures the similarity between samples $\mathbf{x}_u$ and $\mathbf{x}_v$ from the i-th and the j-th views, respectively. It should be pointed out that these graphs share the same set of nodes while having different similarity matrices for different pair of views. With respect to the view pair (i,j), the volume of graph G_{ij} is $vol_{ij}V = \sum_{u\in V, v\in V} w_{ij}(u,v)$. The natural random walk on G_{ij} can be defined as follows. That is, given a node u, we try to walk from u to v, and the transition probability on G_{ij} is $p(u \to v|G_{ij}, u) = w_{ij}(u,v)/\sum_{v\in V} w_{ij}(u,v)$ and the probability of u on G_{ij} is $p(u|G_{ij}) = \sum_{v\in V} w_{ij}(u,v)/vol_{ij}V$. Let $p(G_{ij})$ denote the prior probabilities of the random walker choosing the graph $G_{ij}(1 \le i,j \le J)$, and we have $p(G_{ij}) \ge 0$ and $\sum_{ij} p(G_{ij}) = 1$. Then the probability of node u on multiple graphs is computed as $p(u) = \sum_{ij} p(u|G_{ij})p(G_{ij})$.

According to Bayes' theorem, the posterior probability to choose the graph G_{ij} at node u is $p(G_{ij}|u) = \frac{p(u|G_{ij})p(G_{ij})}{\sum_{ij} p(u|G_{ij})p(G_{ij})}$. Thus, for any node u, the transition probability of $u \to v$ on multiple graphs can be computed as $p(u \to v|u) = \sum_{ij} p(u \to v|G_{ij}, u)p(G_{ij}|u)$. Observe that

$$\begin{aligned}
&p(u \to v|u)p(u) \\
=&\sum_{ij} p(u \to v|G_{ij}, u)p(G_{ij}|u)p(u) \\
=&\sum_{ij} p(u \to v|G_{ij}, u)p(u|G_{ij})p(G_{ij}) \\
=&\sum_{ij} \frac{w_{ij}(u, v)}{\sum_{v \in V} w_{ij}(u, v)} \frac{\sum_{v \in V} w_{ij}(u, v)}{vol_{ij}V} p(G_{ij}) \qquad (1) \\
=&\sum_{ij} \frac{w_{ij}(u, v)}{vol_{ij}V} p(G_{ij}) \\
=&\sum_{i} \frac{w_{ii}(u, v)}{vol_{ii}V} p(G_{ii}) + \sum_{i \neq j} \frac{w_{ij}(u, v)}{vol_{ij}V} p(G_{ij})
\end{aligned}$$

where $w_{ii}(u, v) = exp\{-\gamma \|\mathbf{x}_u^{(i)} - \mathbf{x}_v^{(i)}\|^2\}$ measures the similarity between samples represented by the features in the same view, and $w_{ij}(u, v)(i \neq j)$ measures the similarity between samples across different views. We use CCA technique to define $w_{ij}(u, v)(i \neq j)$ which captures the correlations between heterogeneous feature spaces.

Let $\mathbf{X}^{(i)} = [\mathbf{x}_1^{(i)}, ..., \mathbf{x}_n^{(i)}]$ and $\mathbf{X}^{(j)} = [\mathbf{x}_1^{(i)}, ..., \mathbf{x}_n^{(j)}]$ denote the dataset from the i-th and the j-th views, respectively. Assume that they are firstly centered such that $\sum_{v=1}^{n} \mathbf{x}_u^{(i)} = 0$ and $\sum_{v=1}^{n} \mathbf{x}_u^{(j)} = 0$. The covariance matrices are denoted as follows: $C_{ii} = \mathbf{X}^{(i)}\mathbf{X}^{(i)\top}$, $C_{jj} = \mathbf{X}^{(j)}\mathbf{X}^{(j)\top}$, $C_{ij} = \mathbf{X}^{(i)}\mathbf{X}^{(j)\top}$, and $C_{ji} = \mathbf{X}^{(j)}\mathbf{X}^{(i)\top}$. Canonical Correlation Analysis (CCA) is employed to learn the common space to compute the similarity between samples from multi-view spaces. More specifically, the projection directions where samples are most correlated are firstly learned by $\max_{\{\phi^i, \phi^j\}} \frac{\phi^{i\top} C_{ij} \phi^j}{\sqrt{\phi^{i\top} C_{ii} \phi^i \phi^{j\top} C_{jj} \phi^j}}$, which is equivalent to seeking the eigenvectors w.r.t. largest eigenvalues of ${C_{ii}}^{-1}C_{ij}{C_{jj}}^{-1}C_{ji}$ and ${C_{jj}}^{-1}C_{ji}{C_{ii}}^{-1}C_{ij}$, respectively.

Denote $T = {C_{ii}}^{-\frac{1}{2}}C_{ij}{C_{jj}}^{-\frac{1}{2}}$. $TT^\top$ and $T^\top T$ are both symmetric positive semi-definite and share $\tilde{r}$ positive eigenvalues $\lambda_1, ..., \lambda_{\tilde{r}}$, herein $\tilde{r} = rank(TT^\top) = rank(T^\top T)$. Since ${C_{ii}}^{-1}C_{ij}{C_{jj}}^{-1}C_{ji} = {C_{ii}}^{-\frac{1}{2}}TT^\top {C_{ii}}^{\frac{1}{2}}$ and ${C_{jj}}^{-1}C_{ji}{C_{ii}}^{-1}C_{ij} = {C_{jj}}^{-\frac{1}{2}}T^\top T{C_{jj}}^{\frac{1}{2}}$, ${C_{ii}}^{-1}C_{ij}{C_{jj}}^{-1}C_{ji}$ and ${C_{jj}}^{-1}C_{ji}{C_{ii}}^{-1}C_{ij}$ are similar to $TT^\top$ and $T^\top T$ respectively. Thus, ${C_{ii}}^{-1}C_{ij}{C_{jj}}^{-1}C_{ji}$ and ${C_{jj}}^{-1}C_{ji}{C_{ii}}^{-1}C_{ij}$ also share the above r positive eigenvalues $\lambda_1, ..., \lambda_{\tilde{r}}$. Denote $\Lambda = diag(\lambda_1, ..., \lambda_{\tilde{r}})$, $\Phi^i = [\phi_1^i, ..., \phi_{\tilde{r}}^i]$, and $\Phi^j = [\phi_1^j, ..., \phi_{\tilde{r}}^j]$. Because ${C_{ii}}^{-1}C_{ij}{C_{jj}}^{-1}C_{ji}\Phi^i = \Phi^i\Lambda$ and ${C_{jj}}^{-1}C_{ji}{C_{ii}}^{-1}C_{ij}\Phi^j = \Phi^j\Lambda$, so $TT^\top {C_{ii}}^{\frac{1}{2}}\Phi^i = {C_{ii}}^{\frac{1}{2}}\Phi^i\Lambda$ and $T^\top T{C_{jj}}^{\frac{1}{2}}\Phi^j = {C_{jj}}^{\frac{1}{2}}\Phi^j\Lambda$. Suppose that Ψ and $\widetilde{\Psi}$ are the sets of orthonormal eigenvectors with respect to r positive eigenvalues for $TT^\top$ and $T^\top T$ respectively. Then we can easily calculate the transformation matrix $\Phi^i = {C_{ii}}^{-\frac{1}{2}}\Psi$ and $\Phi^j = {C_{jj}}^{-\frac{1}{2}}\widetilde{\Psi}$. The transformed data ${\Phi^i}^\top\mathbf{X}^{(i)}$ and ${\Phi^j}^\top\mathbf{X}^{(j)}$ are in a common $\tilde{r}$-dim feature space, where the correlation matrix is $\Lambda^{\frac{1}{2}}$. Thus the similarity between samples across different views $w_{ij}(u, v)(i \neq j)$ in Eq. (1) can be defined as $w_{ij}(u, v) = exp\{-\sigma\|{\Phi^i}^\top\mathbf{x}_u^{(i)} - {\Phi^j}^\top\mathbf{x}_v^{(j)}\|^2\}$. Note that $w_{ii}(u, v) = w_{ii}(v, u)$, while $w_{ij}(u, v) \neq w_{ij}(v, u), (i \neq j)$. However, $\sum_{i \neq j} \frac{w_{ij}(u,v)}{vol_{ij}V}p(G_{ij}) = \sum_{j \neq i} \frac{w_{ji}(u,v)}{vol_{ji}V}p(G_{ji})$, thus we have $p(u)p(u \to v|u) = p(v)p(v \to u|v) =: w(u, v)$, which can be defined as the similarity measurement for the data points in multiple views.

2.2 Embedding Learning for Incompletely Labeled Data

Based on the above random walk process among multiple graphs, we can seek the compact embedding of multi-view data and predict the labels in a unified framework. Let $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1, ..., \hat{\mathbf{y}}_n] \in \{-1, 1\}^{c \times n}$ denote the label prediction matrix for all the samples where $\hat{\mathbf{y}}_u^l = -1$ means that the l-th label is not a correct label for $\mathbf{x}_u$. $\hat{\mathbf{Y}}$ can be investigated from different perspectives: i)*column view*: Each column of $\hat{\mathbf{Y}}$ corresponds to the estimated label vector for one data, and can be viewed as the high-level feature vector of the u-th sample in the semantic label space, which differs from the low-level feature vectors in the visual feature spaces; ii)*row view*: Each row of $\hat{\mathbf{Y}}$ can be viewed as the voting scores from all the samples for each concept, and can also be viewed as the feature vector of the concept. By leveraging the information from above two perspectives, we can sufficiently capture the semantic context, and the neighborhood consistency between input spaces and embedding space in the proposed framework formulated as follows:

$$\begin{aligned}
\min_{H, \hat{\mathbf{Y}}} &\sum_{u=1}^{n} \mathbf{y}_u^\top(\mathbf{y}_u - \mathbf{col}(\hat{\mathbf{Y}}, u)) + \theta_1 \sum_{u,v \in V} w(u, v)\|h_u - h_v\|^2 \\
&+ \theta_2 \sum_{s,t \in C} \rho(s, t)\|\mathbf{row}(\hat{\mathbf{Y}}, s) - \mathbf{row}(\hat{\mathbf{Y}}, t)\|^2 \\
&s.t. H \in \mathbb{R}^{r \times n}, \hat{\mathbf{Y}} \in \{-1, 1\}^{c \times n}
\end{aligned} \qquad (2)$$

where $\mathbf{col}(\hat{\mathbf{Y}}, u)$ denotes the u-th column of the matrix $\hat{\mathbf{Y}}$, which is the estimated multi-label vector for the sample $\mathbf{x}_u$; $\mathbf{row}(\hat{\mathbf{Y}}, s)$ denotes the s-th row of the matrix $\hat{\mathbf{Y}}$, which indicates for the s-th concept which samples are positive while the others are negative. Solving Eq. (2) is difficult and we can relax the domain of $\hat{\mathbf{Y}}$ from $\{-1, 1\}^{c \times n}$ to $[-1, 1]^{c \times n}$. $\rho(s, t)$ is defined to capture the correlations between concepts s and t. θ_1 and θ_2 are the tradeoff parameters.

Intuitively, the first term in the objective function Eq. (2) forces the label prediction to fit the given labels as much as possible. Since $\mathbf{y}_u \in \{0, 1\}^c$, the penalty occurs only when $\mathbf{y}_u^l = 1$, and $\mathbf{y}_u^\top(\mathbf{y}_u - col(\hat{\mathbf{Y}}, u)) = \sum_{l=1}^{c} \mathbf{y}_u^l(1 - \hat{\mathbf{y}}_u^l) \geq 0$.

The second term constrains that the difference of embedding codes between samples should be as small as possible when they are strongly connected(i.e., with high transition probability). It is reasonable to preserve the neighborhood contexts when mapping from the input feature spaces to the embedding space. The third term achieves that strongly correlated concepts should have similar voting scores from all the samples. The correlation between concepts $\rho(s,t)$ can be initialized as the harmonic mean of the empirical conditional probabilities:$\rho(s,t) = \frac{2p(t|s)p(s|t)}{p(t|s)+p(s|t)}$, where the empirical conditional probability $p(t|s) = \frac{\sum_{u=1}^{l}(\mathbf{y}_u^s)(\mathbf{y}_u^t)}{2\sum_{u=1}^{l}(\mathbf{y}_u^s)}$ is derived from the labeled samples and measure the co-occurrence of concept pair on the given data. ($\mathbf{y}_u^s = 1(s = 1, \dots, c)$ means that the s-th concept is associated with the u-th sample, while $\mathbf{y}_u^s = 0$ tells nothing.)

Suppose that each datum $\mathbf{x}_u$ represented by multiple features is firstly nonlinearly mapped to the r-dimensional embedding space $h_u = \phi(\mathbf{x}_u)$, and then the label vector can be estimated by mapping h_u to the c-dimensional semantic concept space using the discriminant function $\hat{\mathbf{y}}_u = Qh_u$, where $Q \in R^{c\times r}$ is the linear transformation matrix. Denote $\hat{\mathbf{Y}} = QH = Q[h_1, ..., h_n]$, which is the label prediction matrix for all the samples. Above maps from heterogeneous feature spaces to embedding space and to concept space over multiple graphs can be learned in Eq. (2) as

$$\min_{Q,H}\sum_{u=1}^{n}\mathbf{y}_u^\top(\mathbf{y}_u - Qh_u) + \theta_1\sum_{u,v\in V} w(u,v)\big\|h_u - h_v\big\|^2 + \theta_2\sum_{s,t\in C}\rho(s,t)\big\|(\mathbf{row}(Q,s) - \mathbf{row}(Q,t))H\big\|^2 \tag{3}$$

where $\mathbf{row}(Q,s)$ denotes the s-th row of the matrix Q. By introducing the Laplacian matrices we can rewrite Eq. (3) as

$$\begin{aligned}\min_{Q,H} f = &- Tr\big(Y^\top QH\big) + \theta_1 Tr\big(HL_wH^\top\big)\\ &+ \theta_2 Tr\big(H^\top Q^\top L_\rho QH\big)\end{aligned} \tag{4}$$

where the matrices L_ρ and L_w are two different graph Laplacians. More specifically, $L_\rho = D - F$, where $F = [\rho(s,t)]_{c\times c}$ and D is a diagonal matrix with $D(s,s) = \sum_t \rho(s,t)$. Let $G' = (\{s\}_{s=1}^{c}, F)$ denote the semantic context graph with the vertex set corresponding to the concepts $\{s\}_{s=1}^{c}$ and the weight matrix F measuring the correlations between concepts, thus L_ρ is the graph Laplacians over $G' = (\{s\}_{s=1}^{c}, F)$ modeling the semantic context. Likewise, let $\widetilde{G} = (V, W)$ denote the multi-graph with the edge weight matrix $W = [w(u,v)]_{n\times n}$ derived by natural random walk over multiple graphs and let Λ be a diagonal matrix with $\Lambda(u,u) = \sum_v w(u,v)$, thus $L_w = \Lambda - W$ is the multi-graph Laplacian modeling the consistency between neighborhoods in the embedding space and those in multi-view input feature space.

As pointed out before, $\rho(s,t)$ can be initialized by the empirical conditional probabilities using the given labeled dataset, then L_ρ is estimated initially as well. The correlations between concepts derived by the empirical conditional probabilities is simple and effective if the samples are completely labeled. For incompletely labeled dataset, the empirical conditional probabilities might not be estimated correctly, thus $\rho(s,t)$ and L_ρ are not expected to capture the semantic context relationship well. To address this problem, we can learn semantic context, embedding codes, and the label confidence vectors simultaneously by modifying the framework Eq. (4) as $\min_{Q,H,L_\rho} f$, and we can obtain the solutions in an iterative way. We minimize the cost function Eq. (4) by updating H, Q, and L_ρ alternatively. We derive the gradients of the above cost function with respect to H, Q, and L_ρ, respectively:

$$\begin{aligned}\frac{\partial f}{\partial H} &= -Q^\top Y + 2\theta_1 HL_w + 2\theta_2 Q^\top L_\rho QH\\ \frac{\partial f}{\partial Q} &= -YH^\top + 2\theta_2 L_\rho QHH^\top\\ \frac{\partial f}{\partial L_\rho} &= 2\theta_2 QHH^\top Q^\top\end{aligned} \tag{5}$$

Then the optimal H, Q, and L_ρ can be iteratively learned in an alternating updating way as follows:

$$\begin{aligned}H^t &= H^{t-1} - \alpha\frac{\partial f}{\partial H}(H^{t-1}, Q^{t-1}, L_\rho^{t-1})\\ Q^t &= Q^{t-1} - \beta\frac{\partial f}{\partial Q}(H^{t-1}, Q^{t-1}, L_\rho^{t-1})\\ L_\rho^t &= L_\rho^{t-1} - \gamma\frac{\partial f}{\partial L_\rho}(H^{t-1}, Q^{t-1}, L_\rho^{t-1})\end{aligned} \tag{6}$$

where α, β and γ $(0 < \alpha, \beta, \gamma < 1)$ are all the step sizes for gradient search. Based on the available incomplete label matrix $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n] \in \{0,1\}^{c\times n}$, we employ the technique called Non-negative Matrix Factorization(NMF)[Lee and Seung, 1999] to initialize H and Q: $\mathbf{Y} = Q^0H^0$. Once the optimal H and Q are learned, the optimal label prediction matrix for all the samples is computed straightforward $\hat{\mathbf{Y}} = QH$, and we can predict the label vectors easily by choosing a threshold for each row of $\hat{\mathbf{Y}}$.

2.3 Discussion

It is worthy to note that the Laplacian matrix L_w in Eq. (4) is $n \times n$, and the complexity of Eq. (4) is high over large dataset. For scalability, we can also learn an approximate matrix for the Laplacian L_w like[Liu *et al.*, 2010b]. Specially, clustering is firstly performed on n data points to obtain $m(m >> n)$ anchor points. The weight between the non-anchor point u and the neighboring anchor v is defined as $Z_{uv} = \frac{w(u,v)}{\sum_{j\in\langle v\rangle} w(j,v)}$, where $\langle v\rangle$ denotes the indices of nearest anchors of point v. $Z_{uv} = 0 \quad if \quad u \notin \langle v\rangle$, so the weight matrix $Z \in R^{m\times n}$ is highly sparse. We only need to solve the embedding codes associated with anchors which are put in the matrix $A \in R^{r\times m}$ in which each column vector accounts for one anchor. Then the embedding for all data can be computed easily by $H = AZ$. Substitute H with $H = AZ$ in Eq. (4), we can easily learn the optimal A in the iterative way as before. It should be pointed out that the second item in Eq. (4) becomes $Tr\big(AZL_wZ^\top A^\top\big)$. Based on the random walks across data points and anchors, the approximation to the Laplacian matrix is derived as $L_w = I - Z^\top\Lambda^{-1}Z$, where $\Lambda = diag(Z\mathbf{1}) \in R^{m\times m}$. Thus, $ZL_wZ^\top = ZZ^\top - ZZ^\top\Lambda^{-1}ZZ^\top$, which is both memory-wise and computationally tractable, and the second item in Eq. (4) becomes $Tr\big(AZZ^\top A^\top - AZZ^\top\Lambda^{-1}ZZ^\top A^\top\big)$. Furthermore, learning the embedding for anchors enables efficient out-of-sample extension. For any novel datum point, the embedding can be computed as $\phi(\mathbf{x}) = Az(\mathbf{x})$ where $z(\mathbf{x})$ is $m-$dim weigh vector whose $u-$th element is $z(u,\mathbf{x}) = \frac{w(u,\mathbf{x})}{\sum_{j\in\langle\mathbf{x}\rangle} w(j,\mathbf{x})}$.

3 Experiments

We experimentally evaluate the performance of the proposed method, denoted by $'ours'$, and compare it with WELL [Sun *et al.*, 2010] and PU_WLR [Lee and Liu, 2003]. PU_WLR [Lee and Liu, 2003] is a method learning with Positive and Unlabeled data using Weighted Logistic Regression; WELL (WEak Label Learning) [Sun *et al.*, 2010] is the method designed for incompletely labeled dataset. Furthermore, we also evaluate the degenerated version of our method denoted by $'ours-'$ where the multiple heterogeneous features are simply concatenated into a high-dimensional vector without capturing inter-feature correlations. We conduct experimental evaluations on three image datasets: MSRC [Shotton *et al.*, 2006], LabelMe [Russell *et al.*, 2008] and NUS-WIDE [Chua *et al.*, 2009].

3.1 Results on MSRC

MSRC image dataset [Shotton *et al.*, 2006] is widely used in multi-label learning for performance evaluation. It contains 591 photographs with 23 concepts in total. We ignore the concepts $horse$ and $mountain$ since they have few instances. Thus there are totally 21 concepts in our experiments. About 80% images are associated with multiple labels. There are about 3 labels on average per image. Only one label per image is used for models training in the experiments such that MSRC can be employed as the incompletely labeled data. For each image, we extract five kinds of features: 12-dim CLD (Color Layout Descriptor), 64-dim SCD (Scalable Color Descriptor), 256-dim CSD (Color Structure Descriptor), 80-dim EHD (Edge Histogram Descriptor), and 1024-dim SIFT feature.

Fig. 1 shows the experimental results of our method in comparison with other related methods on MSRC image dataset. We use four criterions to evaluate the performance: Macro-F1, Micro-F1, Hamming Loss and Ave-AUROC (Average Area Under ROC). As for the criterions Macro-F1, Micro-F1, and Ave-AUROC, the larger the value is, the better the performance is; meanwhile, as for Hamming Loss, the smaller, the better. It can be seen that our method consistently performs better than other methods in terms of these criterions. Our method sufficiently leverages inter-feature correlation, inter-label correlation, and feature-label association at the same time, which inherits all merits of the state-of-the-art methods. Note that the degenerated version of our method $'ours-'$ still performs better than or comparable with others because $'ours-'$ effectively captures the inter-label correlation from the incompletely labeled data, even though $'ours-'$ simply concatenates into a high-dimensional vector like WELL and PU_WLR.

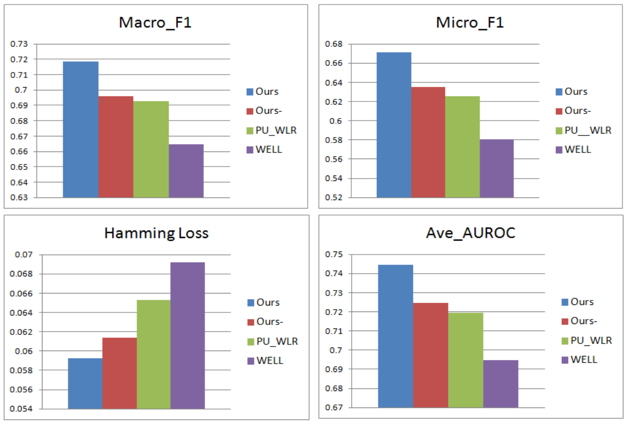

Figure 1: Performance of the proposed method in comparison with the state-of-the-art on MSRC image dataset. The dimension of the learned embedding is set r=64.

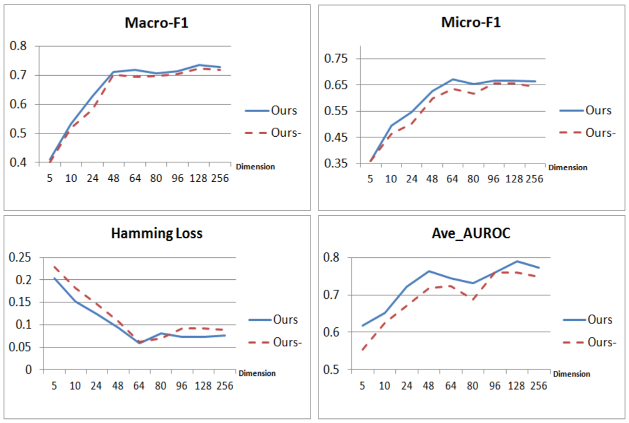

Figure 2: Performance of the proposed method $'ours'$ and the degenerated version $'ours-'$ vs. the dimensions of embedding space on MSRC image dataset.

Fig. 2 illustrates the performance variations on different dimensions of the learned embedding space. As can be seen, the performance increases rapidly with the embedding dimensions are added at the first stage, and it tends to be stable afterward; it even turns to drop when longer codes are used. Thus the longest embedding code can not guarantee the best performance.

3.2 Results on LabelMe

LabelMe image dataset [Russell *et al.*, 2008] is dynamic, free to use, and open to public contribution. We choose a subset of the LabelMe dataset containing the top 33 object categories and totally 1,105 images. In this subset each image has more than 5 labels, and the average number of labels per image in this subset is 6.63. For the purpose of

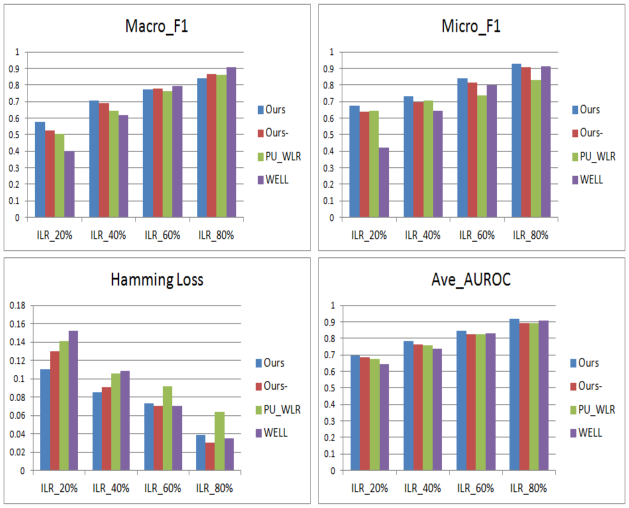

Figure 3: Performance of the proposed method in comparison with the state-of-the-art on LabelMe image dataset. The dimension of the learned embedding is set r=64.

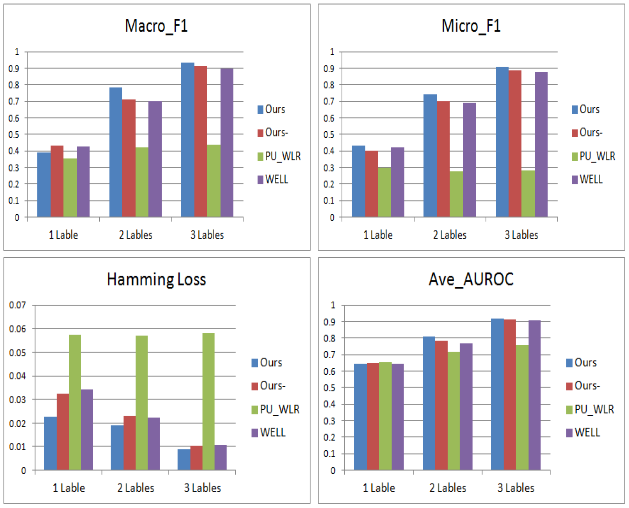

Figure 4: Performance of the proposed method in comparison with the state-of-the-art on NUS-WIDE image dataset. The dimension of the learned embedding is set r=64.

evaluation on incompletely labeled dataset, we present the ground truth of the labels for images in a full label matrix $\dot{\mathbf{Y}} = [\dot{\mathbf{y}}_1, ..., \dot{\mathbf{y}}_n] \in \{-1, +1\}^{c \times n}$ where $\dot{\mathbf{y}}_u^l = +1$ indicates the l-th label is a correct label while $\dot{\mathbf{y}}_u^l = -1$ indicates the l-th label is not a proper label for the u-th image. We vary the Incomplete Label Ratio (ILR, defined as the ratio of the available label number to the ground-truth $\sum_l \mathbf{1}_{\{\mathbf{y}_u^l = 1\}} / \sum_l \mathbf{1}_{\{\dot{\mathbf{y}}_u^l = 1\}}$) of each image from 20% to 80% with 20% as interval to study the performance of our method over various incompletely labeled data. For each image, we also extract five kinds of features: CLD, SCD, CSD, EHD, and SIFT. Fig. 3 shows the experimental results of our method in comparison with other related methods on LabelMe image dataset. Our method outperforms the state-of-the-art methods in most cases.

3.3 Results on NUS-WIDE

NUS-WIDE dataset [Chua *et al.*, 2009] is a challenging collection of 269,648 images and the associated tags from Flickr, with a total number of 5,018 unique tags. Like [Liu *et al.*, 2010a], we select a subset of NUS-WIDE dataset, focusing on images containing at least 5 labels, containing 19,277 images with 81 labels in total. In the experiments, only k (k=1,2,3) label per image are used for models training such that this dataset can be employed as the incompletely labeled data for evaluation. For each image, five types of low-level features are extracted: 144-Dim color correlogram, 73-Dim edge direction histogram, 128-Dim wavelet texture, 225-Dim block-wise color moments, and 500-Dim bag of words based on SIFT descriptions. All these features of NUS-WIDE images are available on web[1] and can be downloaded freely.

Fig. 4 gives the experimental results of our method in comparison with other methods on NUS-WIDE images. In the case that only one label per image is available, the performance of our method is comparable to that of the others on Macro-F1 and Ave-AUROC, and is better on Hamming Loss and Micro-F1. In the case that two or three labels per image are available, the proposed method outperforms the others consistently on these four criterions.

4 Conclusions

In this paper we propose a novel method to learn compact embedding that captures inter-feature correlations, inter-label correlations, and feature-label associations simultaneously from multi-view incompletely-labeled data. Multiple directed graphs are constructed over the dataset to model different similarity matrices across views. We use CCA technique to capture the correlations between heterogeneous feature spaces. By mapping data from multiple feature spaces to the embedding space and further to the concept space, we learn the embedding which preserves the neighborhood context in the original spaces, and complete the labels at the same time. There exists semantic gap between the input multi-view feature spaces and the semantic concept space; and the compact embedding space can be looked on as the bridge over the gap between the above spaces.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. This work was supported in part by the Shanghai Leading Academic Discipline Project (No.B114), the STCSM's Programs (No. 12XD1400900), the NSF of China (No.60903077), and the 973 Program (No.2010CB327906).

References

[Belkin and Niyogi, 2001] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, 2001.

[1] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

[Bronstein *et al.*, 2010] Michael M. Bronstein, Alexander M. Bronstein, Fabrice Michel, and Nikos Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *CVPR*, 2010.

[Chen *et al.*, 2005] H.-T. Chen, H.-W. Chang, and T.-L. Liu. Local discriminant embedding and its variants. In *CVPR*, 2005.

[Chen *et al.*, 2010] Lin Chen, Dong Xu, Ivor W. Tsang, and Jiebo Luo. Tag-based web photo retrieval improved by batch mode re-tagging. In *CVPR*, 2010.

[Chua *et al.*, 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *CIVR*, 2009.

[Dhillon *et al.*, 2011a] Paramveer S. Dhillon, Dean Foster, and Lyle Ungar. Multi-view learning of word embeddings via cca. In *NIPS*, 2011.

[Dhillon *et al.*, 2011b] Paramveer S. Dhillon, Dean Foster, and Lyle Ungar. Multi-view learning of word embeddings via cca. In *NIPS*, 2011.

[Elisseeff and Weston, 2002] Andre Elisseeff and Jason Weston. A kernel method for multi-labelled classification. In *NIPS*, 2002.

[Feng *et al.*, 2004] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR*, 2004.

[Fu *et al.*, 2011] Zhenyong Fu, Horace H.S. Ip, Hongtao Lu, and Zhiwu Lu. Multi-modal constraint propagation for heterogeneous image clustering. In *ACM MM*, 2011.

[Guillaumin *et al.*, 2009] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009.

[Harel and Mannor, 2011] Maayan Harel and Shie Mannor. Learning from multiple outlooks. In *ICML*, 2011.

[He *et al.*, 2005] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang. Face recognition using laplacianfaces. *TPAMI*, 27(3):328–340, 2005.

[Hotelling, 1936] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(4):321–377, 1936.

[Kumar and Daume, 2011] Abhishek Kumar and Hal Daume. A co-training approach for multi-view spectral clustering. In *ICML*, 2011.

[Kumar and Udupa, 2011] Shaishav Kumar and Raghavendra Udupa. Learning hash functions for cross-view similarity search. In *IJCAI*, 2011.

[Kumar *et al.*, 2011] Abhishek Kumar, Piyush Rai, and Hal Daume III. Co-regularized multi-view spectral clustering. In *NIPS*, 2011.

[Lee and Liu, 2003] Wee Sun Lee and Bing Liu. Learning with positive and unlabeled examples using weighted logistic regression. In *ICML*, 2003.

[Lee and Seung, 1999] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[Liu *et al.*, 2010a] Dong Liu, Shuicheng Yan, Yong Rui, and Hong-Jiang Zhang. Unified tag analysis with multi-edge graph. In *ACM MM*, 2010.

[Liu *et al.*, 2010b] Wei Liu, Junfeng He, and Shih-Fu Chang. Large graph construction for scalable semi-supervised learning. In *ICML*, 2010.

[Lu *et al.*, 2012] Yao Lu, Wei Zhang, Ke Zhang, and Xiangyang Xue. Semantic context learning with large-scale weakly-labeled image set. In *CIKM*, 2012.

[Petterson and Caetano, 2010] James Petterson and Tiberio Caetano. Reverse multi-label learning. In *NIPS*, 2010.

[Quadrianto and Lampert, 2011] Novi Quadrianto and Christoph H. Lampert. Learning multi-view neighborhood preserving projections. In *ICML*, 2011.

[Roweis and Saul, 2000] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[Russell *et al.*, 2008] B. C. Russell, A. Torralba, K. Murphy, and W. T. Freeman. Labelme:a database and web-based tool for image annotation. In *InternationalJournal of Computer Vision*, 2008.

[Shotton *et al.*, 2006] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. *ECCV*, 2006.

[Sun *et al.*, 2010] Yu-Yin Sun, Yin Zhang, and Zhi-Hua Zhou. Multi-label learning with weak label. In *AAAI*, 2010.

[Sun *et al.*, 2011] Liang Sun, Shuiwang Ji, and Jieping Ye. Canonical correlation analysis for multilabel classification: A least-squares formulation extensions, and analysis. In *TPAMI*, 2011.

[Wang *et al.*, 2007] Changhu Wang, Feng Jing, Lei Zhang, and Hong-Jiang Zhang. Content-based image annotation refinement. In *CVPR*, 2007.

[Zha *et al.*, 2009] Zheng-Jun Zha, Tao Mei, Jingdong Wang, Zengfu Wang, and Xian-Sheng Hua. Graph-based semi-supervised learning with multiple labels. In *J. Vis. Commun. Image R.*, 2009.

[Zhang and Zhou, 2010] Yin Zhang and Zhi-Hua Zhou. Multi-label dimensionality reduction via dependence maximization. In *TKDD*, 2010.

[Zhou and Burges, 2007] Dengyong Zhou and Christopher J.C. Burges. Spectral clustering and transductive learning with multiple views. In *ICML*, 2007.

[Zhou *et al.*, 2007] Zhi-Hua Zhou, De-Chuan Zhan, and Qiang Yang. Semi-supervised learning with very few labeled training examples. In *AAAI*, 2007.