Convex Subspace Representation Learning from Multi-view Data

Yuhong Guo
Department of Computer and Information Sciences
Temple University
Philadelphia, PA 19122, USA
yuhong@temple.edu

Abstract

Learning from multi-view data is important in many applications. In this paper, we propose a novel convex subspace representation learning method for unsupervised multi-view clustering. We first formulate the subspace learning with multiple views as a joint optimization problem with a common subspace representation matrix and a group sparsity inducing norm. By exploiting the properties of dual norms, we then show a convex min-max dual formulation with a sparsity inducing trace norm can be obtained. We develop a proximal bundle optimization algorithm to globally solve the min-max optimization problem. Our empirical study shows the proposed subspace representation learning method can effectively facilitate multi-view clustering and induce superior clustering results than alternative multi-view clustering methods.

Introduction

In many real world application domains, the data sets are naturally comprised of multiple views. For example, webpages can be represented using both the page-text and the hyperlinks pointing to them, which form their two independent views (Blum and Mitchell 1998). In natural language processing tasks, the same document can have multiple representations in different languages (Amini, Usunier, and Goutte 2009). Although each individual view can be sufficient for characterizing the data object, the multiple views often contain complementary information to each other to alleviate the difficulty of a giving learning problem. Exploiting multiple redundant views to effectively learn from unlabeled data and improve the performance of the target learning task has been a common theme of multi-view learning. Much work on multi-view learning has been focused on classification problems, which share a general principle of maximizing agreement of different views on unlabeled data (Blum and Mitchell 1998; C. Christoudias and Darrell 2008; Collins and Singer 1999; Dasgupta, Littman, and McAllester 2001; Guo and Xiao 2012; Sindhwani and Rosenberg 2008; Sridharan and Kakade 2008).

Recently, exploiting multiple views to improve unsupervised clustering has gained increasing attention from machine learning research community. A number of multi-view clustering methods have been developed in the literature, including the Co-EM method (Bickel and Scheffer 2004), the canonical correlation analysis (CCA) method (Chaudhuri et al. 2009), the generalized multi-view normalized cut method (Zhou and Burges 2007), the two-view spectral clustering over bipartite graphs (de Sa 2005), and the multi-view spectral clustering methods (Kumar and Daumé III 2011; Kumar, Rai, and Daumé III 2011). The CCA method in (Chaudhuri et al. 2009) shows that extracting shared representation across different views can alleviate the difficulty of clustering. The methods in (Kumar, Rai, and Daumé III 2011) conduct co-regularized multi-view spectral clustering, whose nature is to identify consistent low-dimensional representations of the multiple views in terms of eigenvector matrices. They also suggest that learning low-dimensional representations consistent across multiple views can improve the clustering performance. Nevertheless, these existing multi-view clustering methods are limited by either focusing only on two-view learning problems (Chaudhuri et al. 2009; de Sa 2005), or pursuing only alternating optimization procedures to reach arbitrary local optimal solutions for the underlying subspace representations (Bickel and Scheffer 2004; Kumar and Daumé III 2011; Kumar, Rai, and Daumé III 2011).

In this paper, we propose a novel convex subspace representation learning approach for general multi-view clustering. A fundamental assumption of multi-view learning is that the true underlying clustering would assign corresponding points across different views into the same cluster. Our key idea is to identify a common intrinsic subspace representation of the data across multiple views, and then perform standard clustering on this shared representation. Hence the fundamental assumption of multi-view clustering can be automatically captured. We first formulate this common subspace representation learning in the framework of standard matrix factorization, with a group sparsity inducing norm over the shared representation matrix. We then show a convex formulation can be obtained by pursuing its dual relaxation with a matrix trace norm regularizer. We develop a proximal bundle optimization procedure to solve the convex dual optimization problem. Our empirical results demonstrate the efficacy of the proposed approach comparing to a number of alternative methods.

Notations

In this paper, we use capital letters to denote matrices, use boldface lower-case letters to denote vectors, and use lower-case letters to denote scalars. For a matrix X, we use $X_{i:}$ to denote its ith row, use $X_{:j}$ to denote its jth column, and use X_{ij} to denote its entry at the ith row and jth column. We use $\{\theta_v\}^k$ to denote a set of variables $\{\theta_1, \cdots, \theta_k\}$ and use the boldface letter $\boldsymbol{\theta}$ to denote its corresponding column vector. Similarly, we use $\{\mathbf{b}_v\}^k$ to denote a set of vectors $\{\mathbf{b}_1, \cdots, \mathbf{b}_k\}$ where each $\mathbf{b}_v$ is a column vector, and use $\{X^{(v)}\}^k$ to denote a set of matrices $\{X^{(1)}, \cdots, X^{(k)}\}$. We use I_d to denote a $d \times d$ identity matrix, and use $\mathbf{0}_d$ (or $\mathbf{1}_d$) to denote a $d \times 1$ vector with all 0 (or 1) entries. The notation $\circ$ denotes the Hadamard product between two matrices. In terms of norms, $\|\mathbf{b}\|_2$ denotes the Euclidean norm of a vector $\mathbf{b}$, and $\|X\|_F$ denotes the Frobenius norm of a matrix X. The matrix block norm $\|X\|_{2,1}$ is defined as $\|X\|_{2,1} = (\sum_j (\sum_i |X_{ij}|^2)^{\frac{1}{2}})$. The matrix spectral norm is denoted as $\|X\|_{sp} = \max_i \sigma_i(X)$, where $\sigma_i(X)$ denotes the singular value of X. The conjugate of spectral norm is the *trace norm* $\|X\|_* = \sum_i \sigma_i(X)$.

Convex Subspace Representation Learning with Multi-view Data

Given a data set with k views ($k \geq 2$), represented using k matrices $\{X^{(v)} \in I\!R^{t \times d_v}\}^k$, we aim to learn a common subspace representation $\Psi \in I\!R^{t \times m}$ of the data shared across the multiple views. The idea is that such a common subspace representation can capture the intrinsic structure of the data that is consistent across the multiple views, and thus the difficulty of the clustering task can be greatly alleviated. We formulate this multi-view subspace representation learning as a joint optimization problem that minimizes the reconstruction errors over the multiple views of the data while using a $\ell_{2,1}$ norm regularizer over the subspace representation matrix to induce the most intrinsic common representation of the data. Specifically, the optimization problem is

$$\min_{\Psi} \min_{m \in \mathbb{N}} \min_{\{B^{(v)} \in \mathcal{B}_v^m\}^k} \sum_{v=1}^k \frac{\beta_v}{2} \|X^{(v)} - \Psi B^{(v)}\|_F^2 + \gamma \|\Psi\|_{2,1} \tag{1}$$

where $\{\beta_v\}^k$ and γ are tradeoff parameters; each $B^{(v)}$ is a $m \times d_v$ basis matrix for the vth view, which contains row basis vectors such that $\mathcal{B}_v^m = \{\widetilde{B} \in I\!R^{m \times d_v} : \|\widetilde{B}_{i:}\|_2 \leq 1 \ \forall i\}$. The individual basis constraints over each $B^{(v)}$ ensure the multi-view problem (1) differs from a concatenated single view problem. Note m is the size of the basis matrices, which is a model parameter typically being pre-fixed. Instead of selecting such a m parameter beforehand, we would rather determine it automatically within the optimization problem by adding the minimization over $m \in \mathbb{N}$ in (1). Since m can be any natural number, from now on, we will use $\min_{\{B^{(v)} \in \mathcal{B}_v^\infty\}^k}$ as a shorthand for $\min_{m \in \mathbb{N}} \min_{\{B^{(v)} \in \mathcal{B}_v^m\}^k}$.

The optimization problem formulated in (1) is a generalization of the standard single view subspace representation learning by simultaneously conducting subspace representation learning in multiple views with a shared representation matrix Ψ. The problem is convex in $\{B^{(v)}\}^k$ given Ψ and vice versa, but unfortunately is not jointly convex in both and thus does not admit directly global training. Most research for subspace representation learning resorts to alternating minimization even in the single view case, which unavoidable has the drawback of local optimal solutions. Recently, convex reformulations have been developed for single view subspace representation learning (Bach, Mairal, and Ponce 2008; Zhang et al. 2011) and two-view subspace representation learning (White et al. 2012). However, a convex solution for general multi-view subspace representation learning, e.g., the problem we propose to tackle here, is not readily available or extendable from any of these previous work. In this paper, we thus derive a principled convex reformulation of the general multi-view representation learning problem in (1) to facilitate multi-view data analysis.

Proposition 1 *The minimization problem (1) admits the following principled convex dual relaxation*

$$\min_{M} \max_{\boldsymbol{\theta}^\top \mathbf{1}=1, \boldsymbol{\theta}>0} \sum_{v=1}^k \frac{\beta_v}{2} \|X^{(v)} - M^{(v)}\|_F^2 + \gamma \|M E_{\boldsymbol{\theta}}\|_* \tag{2}$$

where

$$M = [M^{(1)}, \cdots, M^{(k)}], \tag{3}$$

$$E_{\boldsymbol{\theta}} = diag([\sqrt{\theta_1}\mathbf{1}_{d_1}; \cdots; \sqrt{\theta_k}\mathbf{1}_{d_k}]). \tag{4}$$

This is the main result of this work. We will prove this proposition by presenting a series of derivation results.

First, by simply setting $M^{(v)} = \Psi B^{(v)}$ for all v, the problem (1) can be equivalently rewritten as

$$\min_{\{M^{(v)}\}^k} \sum_{v=1}^k \frac{\beta_v}{2} \|X^{(v)} - M^{(v)}\|_F^2 + \gamma \min_{\{B^{(v)} \in \mathcal{B}_v^\infty\}^k} \min_{\Psi: \{M^{(v)} = \Psi B^{(v)}\}^k} \|\Psi\|_{2,1} \tag{5}$$

Lemma 1 *For a set of given $\{M^{(v)}\}^k$, assume the inner minimization over $\Psi, \{B^{(v)} \in \mathcal{B}_v^\infty\}^k$ in (5) is within proper bounded closed sets $\{\mathcal{B}_v^\infty\}^k$, one has*

$$\begin{aligned} &\min_{\{B^{(v)} \in \mathcal{B}_v^\infty\}^k} \min_{\Psi: \{M^{(v)} = \Psi B^{(v)}\}^k} \|\Psi\|_{2,1} \\ &= \min_{\{B^{(v)} \in \mathcal{B}_v^\infty\}^k} \max_{\{\Lambda^{(v)}\}^k} \begin{pmatrix} \sum_{v=1}^k tr(\Lambda^{(v)\top} M^{(v)}) - \\ \left\| \sum_{v=1}^k \Lambda^{(v)} B^{(v)\top} \right\|_{2,\infty} \end{pmatrix} \end{aligned} \tag{6}$$

Proof: For any fixed feasible $m \in \mathbb{N}$ and $\{B^{(v)} \in \mathcal{B}_v^\infty\}^k$, we first exploit the following Lagrange dual formulation

$$\begin{aligned} &\min_{\Psi: \{M^{(v)} = \Psi B^{(v)}\}^k} \|\Psi\|_{2,1} \\ &= \min_{\Psi} \max_{\{\Lambda^{(v)}\}^k} \|\Psi\|_{2,1} + \sum_v \text{tr}(\Lambda^{(v)\top}(M^{(v)} - \Psi B^{(v)})) \quad (7) \\ &= \max_{\{\Lambda^{(v)}\}^k} \min_{\Psi} \|\Psi\|_{2,1} + \sum_v \text{tr}(\Lambda^{(v)\top}(M^{(v)} - \Psi B^{(v)})) \quad (8) \end{aligned}$$

The min-max order switching from (7) and (8) is due to the strong Lagrange duality property of the problem (Boyd and Vandenberghe 2004). Since the dual norm of $\|\cdot\|_{2,1}$ is $\|\cdot\|_{2,\infty}$ by norm duality, we then have

$$(8)= \max_{\{\Lambda^{(v)}\}^k} \min_{\Psi} \max_{\Gamma} \begin{pmatrix} \text{tr}(\Gamma^\top \Psi) - \|\Gamma\|_{2,\infty} + \\ \sum_v \text{tr}(\Lambda^{(v)\top}(M^{(v)} - \Psi B^{(v)})) \end{pmatrix} \quad (9)$$

$$= \max_{\{\Lambda^{(v)}\}^k} \max_{\Gamma} \begin{pmatrix} \sum_v \text{tr}(\Lambda^{(v)\top} M^{(v)}) - \|\Gamma\|_{2,\infty} + \\ \min_{\Psi} tr\left(\Psi^\top \left(\Gamma - \sum_v \Lambda^{(v)} B^{(v)^\top}\right)\right) \end{pmatrix} \quad (10)$$

$$= \max_{\substack{\Gamma, \{\Lambda^{(v)}\}^k: \\ \Gamma = \sum_v (\Lambda^{(v)} B^{(v)\top})}} \sum_v \text{tr}(\Lambda^{(v)\top} M^{(v)}) - \|\Gamma\|_{2,\infty} \quad (11)$$

$$= \max_{\{\Lambda^{(v)}\}^k} \sum_v \text{tr}(\Lambda^{(v)\top} M^{(v)}) - \left\| \sum_v \Lambda^{(v)} B^{(v)\top} \right\|_{2,\infty} \quad (12)$$

where (9) follows by the Fenchel conjugate function (Rockafellar 1970), (10) follows by the strong duality (Rockafellar 1970), and (11) follows by eliminating Ψ. Since feasible $m \in \mathbb{N}$ and $\{B^{(v)} \in \mathcal{B}_v^\infty\}^k$ are assumed to exist for the given $\{M^{(v)}\}^k$, thus (6) is proved. □

Proposition 2 *For a set of given $\{\Lambda^{(v)}\}^k$, with proper bounded closed sets $\{\mathcal{B}_v^\infty\}^k$, one has*

$$\max_{\{B^{(v)} \in \mathcal{B}_v^\infty\}^k} \left\| \sum_v \Lambda^{(v)} B^{(v)^\top} \right\|_{2,\infty}^2$$

$$= \max_{\mathbf{b}: \{\|\mathbf{b}_v\|_2^2 \le 1\}^k} \mathbf{b}^\top \Lambda^\top \Lambda \mathbf{b} \quad (13)$$

$$\le \min_{\sum_v \theta_v = 1, \theta_v > 0} \|\Lambda E_{\boldsymbol{\theta}}^{-1}\|_{sp}^2 \quad \textit{(dual relaxation)} \quad (14)$$

where $\Lambda = [\Lambda^{(1)}, \cdots, \Lambda^{(k)}]$; each $\mathbf{b}_v \in \mathbb{R}^{d_v \times 1}$ and $\mathbf{b} \in \mathbb{R}^{d \times 1}$ such that $d = \sum_v d_v$ and $\mathbf{b} = [\mathbf{b}_1; \cdots; \mathbf{b}_k]$; and the matrix $E_{\boldsymbol{\theta}}$ is defined in (4).

Proof: Let

$$B = [B^{(1)}, \cdots, B^{(k)}], \quad \Lambda = [\Lambda^{(1)}, \cdots, \Lambda^{(k)}], \quad (15)$$

we can then rewrite

$$\left\| \sum_v \Lambda^{(v)} B^{(v)\top} \right\|_{2,\infty} = \|\Lambda B^\top\|_{2,\infty}$$

$$= \max_{j \in \{1 \cdots m\}} (\sum_i |\Lambda_{i:} B_{j:}^\top|^2)^{\frac{1}{2}}$$

$$= \max_{j \in \{1 \cdots m\}} (B_{j:} \Lambda^\top \Lambda B_{j:}^\top)^{\frac{1}{2}} \quad (16)$$

This then leads to

$$\max_{\{B^{(v)} \in \mathcal{B}_v^\infty\}^k} \left\| \sum_v \Lambda^{(v)} B^{(v)^\top} \right\|_{2,\infty}^2$$

$$= \max_{\{B^{(v)} \in \mathcal{B}_v^\infty\}^k} \max_{j \in \{1..m\}} B_{j:} \Lambda^\top \Lambda B_{j:}^\top$$

$$= \max_{\mathbf{b}: \{\|\mathbf{b}_v\|_2^2 \le 1\}^k} \mathbf{b}^\top \Lambda^\top \Lambda \mathbf{b} \quad (17)$$

which proves equation (13). We next consider the following Lagrangian for the primal maximization problem in (13)

$$L(\{\mathbf{b}_v\}^k; \{\mu_v\}^k) = \mathbf{b}^\top \Lambda^\top \Lambda \mathbf{b} + \sum_v \mu_v (1 - \|\mathbf{b}_v\|_2^2) \quad (18)$$

where the dual variables satisfy $\mu_v \ge 0$ for all v. In this case, however, the strong duality does not hold in general, which induces a duality gap

$$\max_{\{\mathbf{b}_v\}^k} \min_{\{\mu_v\}^k} L(\{\mathbf{b}_v\}^k; \{\mu_v\}^k)$$

$$\le \min_{\{\mu_v\}^k} \max_{\{\mathbf{b}_v\}^k} L(\{\mathbf{b}_v\}^k; \{\mu_v\}^k) \quad (19)$$

Nevertheless, the optimal solutions of the relaxed problem should satisfy the following KKT conditions

$$\frac{\partial L}{\partial \mathbf{b}_v} = 2\Lambda^{(v)\top} \Lambda \mathbf{b} - 2\mu_v \mathbf{b}_v = \mathbf{0}, \ \forall v; \quad (20)$$

$$\mu_v \ge 0, \quad \|\mathbf{b}_v\|_2^2 \le 1, \ \forall v; \quad (21)$$

$$\mu_v (1 - \|\mathbf{b}_v\|_2^2) = 0, \ \forall v. \quad (22)$$

From (20), it is easy to see that $\mu_v = 0$ implies all entries of $\Lambda^{(v)\top}\Lambda$ are 0s, which is unlikely to happen. It is thus reasonable to assume $\mu_v > 0$ and consider only the interior points $\{\boldsymbol{\mu} : \mu_v > 0, v = 1, \cdots, k\}$ of the feasible dual region in this proof. Then the conditions in (22) lead to $\|\mathbf{b}_v^*\|_2^2 = 1, \forall v$ for optimal solution $\mathbf{b}^*$. The conditions in (20) further induce the following equation system

$$\Lambda^\top \Lambda \mathbf{b}^* = C_\mu \mathbf{b}^* \quad (23)$$

$$\text{for} \quad C_\mu = \text{diag}([\mu_1 \mathbf{1}_{d_1}; \cdots; \mu_k \mathbf{1}_{d_k}]), \quad (24)$$

which suggests a generalized eigenvalue problem. Moreover, based on these conditions, the dual objective function in (19) can be equivalently rewritten as

$$L(\{\mu_v\}^k) = L(\{\mathbf{b}_v^*\}^k; \{\mu_v\}^k) = \sum_{v=1}^k \mu_v \quad (25)$$

Lemma 2 *For any feasible solution $\boldsymbol{\mu}$ of (19), the largest generalized eigenvalue of $(\Lambda^\top \Lambda, C_\mu)$ is upper-bounded by 1; that is, $\lambda_{\max}(\Lambda^\top \Lambda, C_\mu) = \lambda_{\max}(C_\mu^{-\frac{1}{2}} \Lambda^\top \Lambda C_\mu^{-\frac{1}{2}}) = \|\Lambda C_\mu^{-\frac{1}{2}}\|_{sp}^2 \le 1$.*

This Lemma can be proved by simply showing that any $\{\mu_v\}^k$ with $\lambda_{\max}(\Lambda^\top \Lambda, C_\mu) > 1$, will lead to an unbounded objective $L(\{\mathbf{b}_v\}^k; \{\mu_v\}^k)$.

Combing Lemma 2 and the reexpression of dual objective in (25), the dual optimization problem (19) can be equivalently rewritten as

$$\min_{\{\mu_v > 0\}^k} \sum_{v=1}^k \mu_v \quad \text{s.t.} \quad \|\Lambda C_\mu^{-\frac{1}{2}}\|_{sp}^2 \le 1. \quad (26)$$

Now we introduce new variables $\{\alpha_v > 0\}^k$ and $\lambda > 0$. By applying a simple variable replacement, $\mu_v = \lambda \alpha_v, \forall v$, (26) can be equivalently rewritten as

$$\min_{\{\alpha_v > 0\}^k} \sum_{v=1}^k \alpha_v \|\Lambda C_\alpha^{-\frac{1}{2}}\|_{sp}^2 \quad (27)$$

By another variable replacement, $\theta_v = \frac{\alpha_v}{\sum_{v'} \alpha_{v'}}, \forall v$, it is simple to show that (27) can be equivalently reexpressed to

the dual relaxation problem (14). Combing (19), (25), (26) and (27), the dual relaxation from (13) to (14) is proved. □

By combing the results in (6), (13) and (14), we have

$$\begin{aligned}
&\min_{\{B^{(v)}\in\mathcal{B}_v^\infty\}^k}\ \min_{\Psi:\{M^{(v)}=\Psi B^{(v)}\}^k} \|\Psi\|_{2,1} \\
\geq\ & \max_{\Lambda}\ \max_{\sum_v \theta_v=1,\theta_v>0} \mathrm{tr}(\Lambda^\top M) - \|\Lambda E_{\boldsymbol{\theta}}^{-1}\|_{sp} \quad (28)\\
& \text{(dual relaxation)} \\
=\ & \max_{\sum_v \theta_v=1,\theta_v>0}\ \max_{\widetilde{\Lambda}} \mathrm{tr}(\widetilde{\Lambda}^\top M E_{\boldsymbol{\theta}}) - \|\widetilde{\Lambda}\|_{sp} \quad (29)\\
& (\text{set } \widetilde{\Lambda} = \Lambda E_{\boldsymbol{\theta}}^{-1}) \\
=\ & \max_{\sum_v \theta_v=1,\theta_v>0} \|M E_{\boldsymbol{\theta}}\|_* \quad (30)
\end{aligned}$$

where the step from (29) to (30) is based on the simple fact that the trace norm is the dual norm of the spectral norm. Finally, by combining (5) and (30), we can obtain the convex dual relaxation formulation (2) in Proposition 1 from its original primal form (1).

The dual formulation (2) is apparently a convex matrix optimization problem that minimizes the sum of matrix distances while using the trace norm to enforce reduced rank. In next section, we will present a proximal bundle method to solve the convex dual optimization problem (2).

After solving for the optimal M^*, we recover a low-dimensional representation matrix Ψ^* and the concatenated basis matrix B^* by first conducting a singular value decomposition, such that $\Psi^* = U\Sigma, B^* = V^\top$, for $M^* = U\Sigma V^\top$. Starting from these Ψ^* and B^*, we then run an alternating gradient descent procedure to update these matrices by minimizing the squared reconstruction loss in (1).

Optimization Algorithm

Though the optimization problem (2) is a convex optimization problem, it is still difficult to conduct optimization directly due to the non-smooth trace norm. To develop an efficient optimization algorithm, we first derive an equivalent reformulation following a well-known variational formulation of the trace norm (Argyriou, Evgeniou, and Pontil 2006; Grave, Obozinski, and Bach 2011): Let $Z \in \mathbb{R}^{t\times d}$, then the trace norm of Z is equal to

$$\|Z\|_* = \frac{1}{2}\inf_{S\succeq 0} \mathrm{tr}(Z^\top S^{-1} Z) + \mathrm{tr}(S), \quad (31)$$

and the infimum is achieved for $S = (ZZ^\top)^{1/2}$. Based on this result, we can reformulate (2) into the following

$$\begin{aligned}
\min_{\boldsymbol{\theta}:\boldsymbol{\theta}^\top\mathbf{1}=1,\boldsymbol{\theta}\geq 0}\ \max_{M}\ \sup_{S\succeq 0}\ & -\sum_{v=1}^k \frac{\beta_v}{2}\|X^{(v)} - M^{(v)}\|_F^2 \quad (32)\\
& -\frac{\gamma}{2}\mathrm{tr}(E_{\boldsymbol{\theta}}^2 M^\top S^{-1} M) - \mathrm{tr}(S)
\end{aligned}$$

For the convenience of algorithm presentation, here we first switched the order of $\min_M \max_{\boldsymbol{\theta}}$, and then replaced $\min$ with $\max$ and vice versa, while taking a negation of the objective function. The objective function of (32) is a pointwise supremum of linear functions over $\boldsymbol{\theta}$, and thus it remains to be a convex optimization problem. To solve this non-smooth min-max convex optimization problem, we deploy a subgradient based proximal bundle method. In the following, we will first present an efficient coordinate ascent solution for the inner maximization problem over M and S, and then present the overall proximal bundle optimization procedure.

Coordinate Ascent Method

For given $\boldsymbol{\theta}$, the inner maximization problem of (32) is jointly concave in both M and S. We conduct inner maximization using a coordinate ascent procedure which alternately optimizes M and S until convergence is reached. For fixed M, it is known the maximization problem over S has the following closed-form solution

$$S = (ME_{\boldsymbol{\theta}}^2 M^\top)^{1/2} \quad (33)$$

For fixed S, the optimization problem over M can be decomposed into k independently subproblems, one for each view. For the vth view, the subproblem over $M^{(v)}$ is

$$\min_{M^{(v)}}\ \frac{\beta_v}{2}\|X^{(v)} - M^{(v)}\|_F^2 + \frac{\gamma\theta_v}{2}\mathrm{tr}(M^{(v)\top} S^{-1} M^{(v)}) \quad (34)$$

which has a closed-form solution

$$\begin{aligned}
M^{(v)} &= \beta_v(\beta_v I_t + \gamma\theta_v S^{-1})^{-1} X^{(v)} \\
&= \big(I_t - (I_t + \frac{\beta_v}{\gamma\theta_v} S)^{-1}\big) X^{(v)} \quad (35)
\end{aligned}$$

Since the k subproblems can be solved independently from each other, parallel computing can be applied to best use the computer resources.

Proximal Bundle Method

Proximal bundle method is a subgradient based optimization method developed to address non-smooth optimization problems (Kiwiel 1990). We thus deploy a bundle optimization procedure to solve the non-smooth min-max optimization problem (32).

Let $F(\boldsymbol{\theta}, M, S)$ denote the objective function of the optimization problem (32), and let $J(\boldsymbol{\theta})$ denote the objective function for the outer minimization problem over $\boldsymbol{\theta}$, such that

$$\begin{aligned}
F(\boldsymbol{\theta}, M, S) \ =\ & -\frac{\beta_v}{2}\|X^{(v)} - M^{(v)}\|_F^2 \quad (36)\\
& -\frac{\gamma}{2}\mathrm{tr}(E_{\boldsymbol{\theta}}^2 M^\top S^{-1} M) - \mathrm{tr}(S)
\end{aligned}$$

$$J(\boldsymbol{\theta}) = F(\boldsymbol{\theta}, M_{\boldsymbol{\theta}}^*, S_{\boldsymbol{\theta}}^*) = \max_M \sup_{S\succeq 0} F(\boldsymbol{\theta}, M, S) \quad (37)$$

where $M_{\boldsymbol{\theta}}^*$ and $S_{\boldsymbol{\theta}}^*$ are the optimal inner maximization solution for the given $\boldsymbol{\theta}$,

$$\{M_{\boldsymbol{\theta}}^*, S_{\boldsymbol{\theta}}^*\} = \arg\max_{M, S\succeq 0} F(\boldsymbol{\theta}, M, S) \quad (38)$$

Let $I_d^{(v)} = \mathrm{diag}([\mathbf{0}_{d_1}, \cdots, \mathbf{1}_{d_v}, \cdots, \mathbf{0}_{d_k}])$, for $v = 1, \cdots, k$, such that $I_d = \sum_{v=1}^k I_d^{(v)}$. According to Danskin's theorem, the subgradient of $J(\boldsymbol{\theta})$ at point $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_k]^\top$ can be computed as $\mathbf{s} = [s_1, \cdots, s_k]^\top$ for

$$s_v = \frac{\partial J(\boldsymbol{\theta})}{\partial \theta_v} = -\frac{\gamma}{2}\mathrm{tr}(I_d^{(v)} M_{\boldsymbol{\theta}}^{*\top} S_\theta^{*-1} M_{\boldsymbol{\theta}}^*), \quad \forall v. \quad (39)$$

Algorithm 1 Proximal Bundle Method

Input: $\epsilon > 0$, $\xi > 0$, $\rho \in (0, 1)$, $\zeta_0 > 0$, $\boldsymbol{\theta}_0$
Initialize: $r = 0$, $\hat{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0$
Loop:
1. set $r = r + 1$, $\zeta_r = \xi\zeta_{r-1}$
2. compute $J(\boldsymbol{\theta}_{r-1})$, and the subgradient $\mathbf{s}_r$ at $\boldsymbol{\theta}_{r-1}$
3. update model:
$J_r^{CP}(\boldsymbol{\theta}) := \max_{1\leq i\leq r}\{J(\boldsymbol{\theta}_{i-1}) + (\boldsymbol{\theta} - \boldsymbol{\theta}_{i-1})^\top \mathbf{s}_i\}$
4. compute:
$\bar{\boldsymbol{\theta}}_r = \arg\min_{\boldsymbol{\theta}^\top \mathbf{1}=1, \boldsymbol{\theta}\geq 0} J_r^{CP}(\boldsymbol{\theta}) + \frac{\zeta_r}{2}\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{r-1}\|^2$
5. compute:
$\epsilon_r = J(\hat{\boldsymbol{\theta}}_{r-1}) - \left[J_r^{CP}(\bar{\boldsymbol{\theta}}_r) + \frac{\zeta_r}{2}\|\bar{\boldsymbol{\theta}}_r - \hat{\boldsymbol{\theta}}_{r-1}\|^2\right]$
6. if $\epsilon_r < \epsilon$ **then** return $\bar{\boldsymbol{\theta}}_r$ **endif**
7. conduct line search:
$\eta^* = \arg\min_{0<\eta\leq 1} J(\hat{\boldsymbol{\theta}}_{r-1} + \eta(\bar{\boldsymbol{\theta}}_r - \hat{\boldsymbol{\theta}}_{r-1}))$
8. set $\boldsymbol{\theta}_r = \hat{\boldsymbol{\theta}}_{r-1} + \eta^*(\bar{\boldsymbol{\theta}}_r - \hat{\boldsymbol{\theta}}_{r-1})$
9. if $J(\hat{\boldsymbol{\theta}}_{r-1}) - J(\boldsymbol{\theta}_r) \geq \rho\epsilon_r$ **then**
$\hat{\boldsymbol{\theta}}_r = \boldsymbol{\theta}_r$
else
$\hat{\boldsymbol{\theta}}_r = \hat{\boldsymbol{\theta}}_{r-1}$
end if
End Loop

Given subgradients $\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_r$ evaluated at a sequence of feasible points $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{r-1}$, the key idea of the proximal bundle method is based on the following subgradient property:

$$J(\boldsymbol{\theta}) \geq J_r^{CP}(\boldsymbol{\theta}) := \max_{1\leq i\leq r}\{J(\boldsymbol{\theta}_{i-1}) + (\boldsymbol{\theta} - \boldsymbol{\theta}_{i-1})^\top \mathbf{s}_i\} \quad (40)$$

Provided the previous prox-center point $\hat{\boldsymbol{\theta}}_{r-1}$, it seeks the next potential candidate point by minimizing the piecewise linear lower bound augmented with a stabilization term as below

$$\bar{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}^\top \mathbf{1}=1, \boldsymbol{\theta}\geq 0} J_r^{CP}(\boldsymbol{\theta}) + \frac{\zeta_r}{2}\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{r-1}\|^2 \quad (41)$$

The approximation gap at $\bar{\boldsymbol{\theta}}$ can be evaluated as

$$\epsilon_r = J(\hat{\boldsymbol{\theta}}_{r-1}) - \left[J_r^{CP}(\bar{\boldsymbol{\theta}}) + \frac{\zeta_r}{2}\|\bar{\boldsymbol{\theta}}_r - \hat{\boldsymbol{\theta}}_{r-1}\|^2\right] \quad (42)$$

If ϵ_r is less than a pre-defined threshold ϵ, the algorithm exits. Otherwise, a line search is performed along the line between $\hat{\boldsymbol{\theta}}_{r-1}$ and $\bar{\boldsymbol{\theta}}_t$ to produce the new point $\boldsymbol{\theta}_r$. If $\boldsymbol{\theta}_r$ leads to a sufficient decrease of the objective function, it is accepted as the new prox-center point $\hat{\boldsymbol{\theta}}_r$. Otherwise, the new prox-center point is set same as the old prox-center point. The overall algorithm is given in Algorithm 1.

Experiments

In this section, we report our empirical results for multi-view clustering, by comparing the proposed approach to a number of baseline methods over real world multi-view data sets.

Data sets: We constructed a number of multi-view clustering tasks from three real world multi-view data sets. The major information of the seven constructed tasks is summarized in Table 1.

- *3-Sources text data set:* This data set is collected from three online news sources: BBC, Reuters, and the Guardian. In total there are 948 news articles covering 416 distinct news stories. Among them, 169 were reported in all three sources. Each story was manually labeled with one of the six topic labels. We used all 169 news in our experiment, while each source is taken as one independent view of the story.
- *Reuters multilingual data set:* This text collection contains documents originally written in five different languages (English, French, German, Spanish and Italian) and their translations. This multilingual data set covers a common set of six categories (Amini, Usunier, and Goutte 2009). We used the data set downloaded from the Internet [1] , where 1200 documents over 6 labels in five languages are given. From this data set, we constructed two three-view subsets, Reuters1 and Reuters2. Reuters1 is constructed using three languages: English, French and German. Reuters2 is constructed in the same way, but with different languages: English, Spanish and Italian.
- WebKB data set: The WebKB data set has been widely used for multi-view learning. It contains webpages collected from four universities: Cornell, Texas, Washington, and Wisconsin, where each webpage is described in two views: the content view and the link view. We used a version downloaded from the Internet [1] , which contains webpages of the four universities distributed across five classes: course, project, student, faculty and staff.

Approaches: In the experiments, we compared the empirical performance of the following methods.

- *FeatConcate:* Concatenating the features of all views and then applying the standard k-means clustering.
- *ConcatePCA:* Concatenating the features of all views, applying PCA to extract the low dimensional subspace representation, and then applying the standard k-means clustering on the low dimensional representation.
- *PairwiseSC:* The pairwise multi-view spectral clustering method developed in (Kumar, Rai, and Daumé III 2011).
- *CentroidSC:* The centroid multi-view spectral clustering method developed in (Kumar, Rai, and Daumé III 2011), which extracts a low dimensional spectral representation matrix across multiple views.
- *NonConvex:* We used a proximal gradient optimization procedure to solve the original nonconvex multi-view subspace learning problem in (1) directly, which takes alternating gradient descent steps over $\{B^{(v)}\}^k$ and Ψ. Proximal gradient descent is used for minimizing Ψ with a $\ell_{2,1}$-norm regularizer. K-means clustering is applied on the learned common subspace representation matrix Ψ.

[1] http://membres-liglab.imag.fr/grimal/data.html

Table 1: Information of the multi-view tasks.

Info.	3-Sources	Reuters1	Reuters2	Cornell	Texas	Washington	Wisconsin
# of Views	3	3	3	2	2	2	2
# of Clusters	6	6	6	5	5	5	5

Table 2: The clustering results (average$\pm$std) on real world multi-view data sets in terms of normalized mutual information (NMI) measure (%).

Method	3-Sources	Reuters1	Reuters2	Cornell	Texas	Washington	Wisconsin
FeatConcate	36.0$\pm$2.2	11.4$\pm$1.1	8.7$\pm$0.6	9.4$\pm$0.3	14.3$\pm$0.5	15.9$\pm$0.7	9.0$\pm$0.2
ConcatePCA	60.3$\pm$0.5	14.7$\pm$0.3	15.4$\pm$0.3	11.3$\pm$0.2	16.9$\pm$0.2	19.9$\pm$0.2	9.7$\pm$0.2
PairwiseSC	60.1$\pm$0.5	11.6$\pm$0.1	12.1$\pm$0.1	11.2$\pm$0.2	17.9$\pm$0.2	21.2$\pm$0.2	9.8$\pm$0.1
CentroidSC	60.0$\pm$0.6	10.9$\pm$0.0	11.4$\pm$0.1	10.4$\pm$0.2	16.9$\pm$0.2	18.5$\pm$0.2	10.8$\pm$0.2
NonConvex	56.7$\pm$0.4	17.6$\pm$0.2	**19.7**$\pm$0.2	11.5$\pm$0.1	19.8$\pm$0.3	22.5$\pm$0.2	11.8$\pm$0.1
Convex	**61.9**$\pm$0.5	**19.1**$\pm$0.4	18.3$\pm$0.5	**23.3**$\pm$0.1	**24.5**$\pm$0.4	**25.1**$\pm$0.3	**30.3**$\pm$0.3

- *Convex:* This is the proposed approach, which first conducts convex multi-view subspace representation learning, and then applies k-means clustering on the learned common representation matrix.

To maintain a fair comparison, we used the number of clusters as the dimension size of the subspace representations for all the comparison methods except *FeatConcate* which works in the original feature space. For *PairwiseSC* and *CentroidSC*, we tried a set of trade-off parameter values $\lambda = [0.005, 0.01, 0.05, 0.1]$ as suggested in (Kumar, Rai, and Daumé III 2011), and present the best results obtained. For the proposed approach, *Convex*, and its nonconvex version, *NonConvex*, the β_v regularization parameter associated with each view is important, since it controls the degree of importance of each view. The relative informativeness of the multiple views are simple domain knowledge one should exploit. The first three data sets, 3-sources, Reuters1 and Reuters2, have three views, while each view is a description of the topic in different media sources or languages. Thus their multiple views could be equally important. We simply set their β_v values as 1. For the four WebKB data sets with two views, the content view is typically much more informative than the link view. We thus used $\beta = 100$ for the content view and used $\beta = 1$ for the link view in the experiments. We run the *Convex* and *NonConvex* methods using a range of γ values, $\gamma = [0.1, 0.2, 0.3, 0.4] \times \max_v(\beta_v)$, and present the best results obtained. The computational time of the *NonConvex* method is much less than the *Convex* method as the nonconvex optimization procedure can quickly return a local optimal solution. In our experiments, *NonConvex* takes only a few minutes to run, while Convex takes a few minutes on the 3-Sources and WebKB data sets, but takes about up to half or one hour on the two Reuters data sets.

The experimental results are reported in Table 2, where the normalized mutual information (NMI) is used as the clustering quality measure. The results are over 50 runs of k-means with random initializations. We can see that *ConcatePCA* clearly outperforms the simple *FeatConcate* method by simply taking the most informative subspace representations to reduce noise. The two spectral multi-view methods, with local optimal solutions on representation matrix learning, though outperform *ConcatePCA* in some cases, e.g., on Texas, Washington and Wisconsin, demonstrate inferior performance on Reuters1 and Reuters2. The *NonConvex* method however demonstrates comparable or significantly superior performance on most data sets except the 3-Sources, comparing to the previous four methods. This suggests our multi-view subspace representation learning problem (1) is a very reasonable formulation. The proposed *Convex* method on the other hand outperforms all other methods across six out of seven data sets, except on *Reuters2* where *NonConvex* is slightly better. Moreover, the advantage of the *Convex* method is significant in most cases. This clearly demonstrates the efficacy of the derived convex subspace representation learning formulation and the proposed optimization technique. The experimental results suggest the proposed convex subspace representation learning model has great capacity of extracting the intrinsic information of the data shared across multiple views, and facilitating data analysis tasks consequently.

Conclusion

In this paper, we first derived a convex formulation for common subspace representation learning across multiple views. We then developed a proximal bundle method with coordinate ascent subroutines to solve the obtained min-max convex optimization problem. We evaluated the learned subspace representations over seven multi-view clustering tasks, comparing to a number of alternative methods. Our empirical study suggests the proposed convex approach can effectively capture the intrinsic information of the given data and outperform all the other multi-view clustering methods used in the experiments. The proposed convex subspace learning formulation can be directly extended to handle supervised and semi-supervised multi-view learning problems, which we will consider in the future.

References

Amini, M.; Usunier, N.; and Goutte, C. 2009. Learning from multiple partially observed views - an application to multilingual text categorization. In *Advances in Neural Information Processing Systems (NIPS)*.

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2006. Multi-task feature learning. In *Advances in Neural Information Processing Systems (NIPS)*.

Bach, F.; Mairal, J.; and Ponce, J. 2008. Convex sparse matrix factorizations. *arXiv:0812.1869v1*.

Bickel, S., and Scheffer, T. 2004. Multi-view clustering. In *Proceedings of IEEE International Conference on Data Mining (ICDM)*.

Blum, A., and Mitchell, T. 1998. Combing labeled and unlabeled dta with co-training. In *Proceedings of Annual Conference on Learning Theory (COLT)*.

Boyd, S., and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge University Press.

C. Christoudias, R. U., and Darrell, T. 2008. Multi-view learning in the presence of view disagreement. In *Proceedings of Conference on Uncertainty in Artificial Intelligence (UAI)*.

Chaudhuri, K.; Kakade, S.; Livescu, K.; and Sridharan, K. 2009. Multi-view clustering via canonical correlation analysis. In *Proceedings of International Conference on Machine Learning (ICML)*.

Collins, M., and Singer, Y. 1999. Unsupervised models for named entity classification. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Dasgupta, S.; Littman, M.; and McAllester, D. 2001. PAC generalization bounds for co-training. In *Advances in Neural Information Processing Systems (NIPS)*.

de Sa, V. 2005. Spectral clustering with two views. In *Workshop on Learning with Multiple Views, ICML*.

Grave, E.; Obozinski, G.; and Bach, F. 2011. Trace lasso: a trace norm regularization for correlated designs. In *Advances in Neural Information Processing Systems (NIPS)*.

Guo, Y., and Xiao, M. 2012. Cross language text classification via subspace co-regularized multi-view learning. In *Proceedings of International Conference on Machine Learning (ICML)*.

Kiwiel, K. C. 1990. Proximity control in bundle methods for convex nondifferentiable minimization. *Mathematical Programming* 46:105–122.

Kumar, A., and Daumé III, H. 2011. A co-training approach for multi-view spectral clustering. In *Proceedings of International Conference on Machine Learning (ICML)*.

Kumar, A.; Rai, P.; and Daumé III, H. 2011. Co-regularized multi-view spectral clustering. In *Advances in Neural Information Processing Systems (NIPS)*.

Rockafellar, R. 1970. *Convex Analysis*. Princeton University Press.

Sindhwani, V., and Rosenberg, D. 2008. An RKHS for multi-view learning and manifold co-regularization. In *Proceedings of International Conference on Machine Learning (ICML)*.

Sridharan, K., and Kakade, S. 2008. An information theoretic framework for multi-view learning. In *Proceedings of Annual Conference on Learning Theory (COLT)*.

White, M.; Yu, Y.; Zhang, X.; and Schuurmans, D. 2012. Convex multi-view subspace learning. In *Advances in Neural Information Processing Systems (NIPS)*.

Zhang, X.; Yu, Y.; White, M.; Huang, R.; and Schuurmans, D. 2011. Convex sparse coding, subspace learning, and semi-supervised extensions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Zhou, D., and Burges, C. 2007. Spectral clustering and tranductive learning with multiple views. In *Proceedings of International Conference on Machine Learning (ICML)*.