Multi-View Clustering and Feature Learning via Structured Sparsity

Hua Wang HUAWANGCS@GMAIL.COM
Colorado School of Mines, 1610 Illinois Street, Golden, Colorado 80401 USA

Feiping Nie FEIPINGNIE@GMAIL.COM
Heng Huang HENG@UTA.EDU
The University of Texas at Arlington, 500 UTA Boulevard, Arlington, Texas 76019 USA

Abstract

Combining information from various data sources has become an important research topic in machine learning with many scientific applications. Most previous studies employ kernels or graphs to integrate different types of features, which routinely assume one weight for one type of features. However, for many problems, the importance of features in one source to an individual cluster of data can be varied, which makes the previous approaches ineffective. In this paper, we propose a novel multi-view learning model to integrate all features and learn the weight for every feature with respect to each cluster individually via new joint structured sparsity-inducing norms. The proposed multi-view learning framework allows us not only to perform clustering tasks, but also to deal with classification tasks by an extension when the labeling knowledge is available. A new efficient algorithm is derived to solve the formulated objective with rigorous theoretical proof on its convergence. We applied our new data fusion method to five broadly used multi-view data sets for both clustering and classification. In all experimental results, our method clearly outperforms other related state-of-the-art methods.

1. Introduction

Many problems in machine learning involve data sets with multiple views where observations are represented by multiple sources of features. Because different data sources contain different and partly independent information, the multi-view learning is beneficial by reducing the noise, as well as by improving statistical significance and leveraging the interactions and correlations between data sources to obtain more refined and higher-level information, which is also known as data fusion or data integration. Much progress has been made over the last ten years in developing effective multi-view semi-supervised (*e.g.* co-training (Ghani, 2002) and co-EM (Brefeld & Scheffer, 2004)) and unsupervised learning (*e.g.* multi-view clustering (Bickel & Scheffer, 2004)) algorithms. These methods typically utilize multiple redundant views to effectively learn from unlabeled data by mutually training a set of classifiers defined in each view, with the assumption that the multi-view features given the class are conditionally independent. However, in most real-world applications, the independence assumption of the feature sets is not well satisfied, such that these methods may not effectively work (Belkin et al., 2006).

From machine learning point of view, different representations of the same set of objects could give rise to different kernel functions, thus the Multiple Kernel Learning (MKL) approaches (Yu et al., 2010; Suykens et al., 2002; Kloft et al.; Ye et al., 2008a; Lanckriet et al., 2004a; Bach et al., 2004; Sonnenburg et al., 2006) have recently become very popular, because they can easily combine information from multiple views. In general, MKL attempts to form an ensemble of kernels to yield a good fit for a certain application. It has been proven that MKL can offer some needed flexibility and well manipulate the case that involves multiple, heterogeneous data sources.

A core assumption in MKL, as well as many existing graph based multi-view learning methods (Cai et al., 2011; Kumar et al., 2011), is that all features in the same data source are considered as equally important and given the same weight in data fusion, *i.e.*,

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

one weight is learned for one kernel matrix or graph. However, one can expect that the feature-wise importance to different learning tasks can vary significantly. To capture the view-wise relationships among data sources without ignoring the feature-wise importance within each data source, we propose a novel multi-view learning framework via the sparse regularizations to emphasize structured sparsity from both group and multi-task points of views.

In sparsity learning, the sparse representations are typically achieved by imposing non-smooth sparsity-inducing regularization terms. From the sparsity organization perspective of view, we have two types of sparsity: 1) The flat sparsity is often achieved by ℓ_0-norm or ℓ_1-norm regularizer or trace norm in matrix/tensor completion. 2) The structured sparsity is usually obtained via different sparsity-inducing norms such as $\ell_{2,1}$-norm (Obozinski et al., 2010), $\ell_{\infty,1}$-norm (Quattoni et al., 2009), and group ℓ_1-norm (Yuan & Lin, 2006), and many others (Wang et al., 2011; 2012a;c;d).

In this paper, we propose a novel multi-view feature learning and data clustering framework that integrates all features of different views and uses joint structured sparsity-inducing norms to learn a weight for each feature and provide a more flexible method for model selection. The group ℓ_1-norm regularization learns the group-wise features importance of one view on each cluster (task) and the $\ell_{2,1}$-norm regularization explores the feature-wise importance for multiple clusters (tasks). Our new model is designed for multi-view data clustering, which can also be naturally extended to deal with classification tasks when prior labeling knowledge is available. Because our final objective comprises two non-smooth sparsity-inducing norms, the current related optimization methods cannot be efficiently applied. We derive a new efficient algorithm with rigorous theoretical proof on its convergence. We apply our new multi-view learning framework to five broadly used multi-view data sets. Promising results in extensive experiments have validated our new approaches in a number of real-world applications.

2. Multi-View Clustering via Joint Structured Sparsity-Inducing Norms

In this section, we will first systematically propose a novel multi-view learning model for exploring the unsupervised heterogeneous data fusion and clustering, followed by a new efficient iterative algorithm to solve the formulated highly non-smooth objective with a rigorous proof of its convergence. Then we extend the proposed multi-view learning framework to deal with supervised classification tasks.

Notations. In this paper, we write matrices as bold uppercase letters and vectors as bold lowercase letters. Given a matrix $\mathbf{W} = [w_{ij}]$, we denote its i-th row as $\mathbf{w}^i$ and its j-th column as $\mathbf{w}_j$.

2.1. Joint Structured Sparsity-Inducing Norms for Heterogeneous Features Learning

In the setting of clustering, given n data samples $\{\mathbf{x}_i\}_{i=1}^n$, we have data matrix $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \Re^{d\times n}$. $\mathbf{x}_i \in \Re^d$ is the input vector including all features from a total of k views and each view j has d_j features such that $d = \sum_{j=1}^k d_j$. Our goal in multi-view clustering is to partition $\{\mathbf{x}_i\}_{i=1}^n$ into c clusters by exploiting the information stored in all k different views of the input data.

Although the traditional K-means clustering or spectral clustering objectives can be extended for multi-view clustering, similar to MKL, such multi-view clustering objectives still only learn one weight for all features from the same type (due to the objectives' natural limitation). Thus, we need do clustering from another point of view. Previous work (Nie et al., 2009) showed the following regression-like clustering objective, which is equivalent to the Discriminative K-means (Ye et al., 2008b), obtains better results than K-means or spectral clustering methods:

$$\min_{\mathbf{W},\mathbf{F}^T\mathbf{F}=\mathbf{I}} ||\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}||_F^2, \qquad (1)$$

where $\mathbf{b} \in \Re^{c\times 1}$ is the intercept vector, $\mathbf{1}_n$ is $n \times 1$ constant vector of all 1's, $\mathbf{F} = [\mathbf{f}_1, \cdots, \mathbf{f}_n]^T \in \Re^{n\times c}$ is the cluster indicator matrix, and $\mathbf{f}_i \in \Re^c$ is the cluster indicator vector for data point $\mathbf{x}_i$ with f_{ij} indicating how likely $\mathbf{x}_i$ belongs to the j-th cluster. Upon solution, we learn the $d \times c$ parameter matrix $\mathbf{W}$, which includes the weights of each feature for c different clusters. In multi-view clustering, $\mathbf{W} = [\mathbf{w}_1^1, \cdots, \mathbf{w}_c^1;\ \cdots, \cdots, \cdots;\ \mathbf{w}_1^k, \cdots, \mathbf{w}_c^k] \in \Re^{d\times c}$, where, as illustrated in Figure 1, $\mathbf{w}_p^q \in \Re^{d_q}$ indicates the weights of all features in the q-th view with respect to the p-th cluster. Although Eq. (1) learns the weight for each feature to capture the feature-wise importance, more importantly we need design proper regularizers to include the interrelations among multi-view features.

In heterogeneous data fusion, from a multi-view viewpoint, the features of a specific view can be more or less discriminative for different clusters (groups of data objects). For instance, in image clustering, the color features substantially increase the detection of stop signs while they are almost irrelevant for finding cars in images. To address this, we use group ℓ_1-norm (G_1-norm) for regularization, which is defined

as $\|\mathbf{W}\|_{G_1} = \sum_{i=1}^{c}\sum_{j=1}^{k}||\mathbf{w}_i^j||_2$ (Wang et al., 2012b; 2013) and illustrated in Figure 1. Thus, and our objective can be written as:

$$\min_{\mathbf{W},\mathbf{F}^T\mathbf{F}=\mathbf{I}} ||\mathbf{X}^T\mathbf{W}+\mathbf{1}_n\mathbf{b}^T-\mathbf{F}||_F^2+\gamma_1||\mathbf{W}||_{G_1}. \quad (2)$$

Because the group ℓ_1-norm uses ℓ_2-norm within each view and ℓ_1-norm between views, it enforces the sparsity between different views, *i.e.* if one view of features are not discriminative for certain group of objects, the objective in Eq. (2) will assign zeros (in ideal case, usually they are very small values) to them for corresponding clusters; otherwise, their weights are large. Crucially, this group ℓ_1-norm regularizer captures the global relationships between views.

However, in certain cases, even if most features in one view are not discriminative for a group of objects, a small number of features in the same view can still be highly discriminative. From multi-task learning perspective of view, such important features should be shared by all clusters. Thus, we add an additional $\ell_{2,1}$-norm regularizer into Eq. (2) as following:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{F}^T\mathbf{F}=\mathbf{I}} &||\mathbf{X}^T\mathbf{W}+\mathbf{1}_n\mathbf{b}^T-\mathbf{F}||_F^2\\ &+\gamma_1||\mathbf{W}||_{G_1}+\gamma_2||\mathbf{W}||_{2,1}.\end{aligned} \quad (3)$$

The $\ell_{2,1}$-norm has been widely used in multi-task feature learning (Argyriou et al., 2008; Obozinski et al., 2010). Because the $\ell_{2,1}$-norm regularizer imposes the sparsity between all features and non-sparsity between clusters, the features that are discriminative for all clusters will get large weights.

Our regularization items consider the heterogeneous features from both view-wise and individual viewpoints. Figure 1 visualizes the matrix $\mathbf{W}^T$ as a demonstration, in which the elements with deep orange color have large values. The group ℓ_1-norm emphasizes the view-wise weights learning corresponding to each cluster and the $\ell_{2,1}$-norm accentuates the individual weight learning across multiple clusters. Through the joint sparsity-inducing norms, for each task (cluster), many features (not all of them) in the discriminative views and a small number of features (may not be none) in the non-discriminative views will learn large weights as the important and discriminative features.

2.2. Optimization Algorithm

Because the objective in Eq. (3) comprises two non-smooth regularization terms of G_1-norm and $\ell_{2,1}$-norm, it is difficult to solve in general. Thus we derive an alternative iterative algorithm to solve the problem, which employs the iteratively re-weighted method (Gorodnitsky & Rao, 1997) to deal with the non-smooth regularization terms.

First, when $\mathbf{W}$ and $\mathbf{b}$ are fixed, we need to solve the following problem:

$$\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}} ||\mathbf{X}^T\mathbf{W}+\mathbf{1}_n\mathbf{b}^T-\mathbf{F}||_F^2. \quad (4)$$

Due to the orthonormal constraint, Eq. (4) is not easy to solve. We prove the following theorem for generic matrices, which provides the solution to Eq. (4) in a closed form.

Theorem 1 *Given any matrix $\mathbf{A}\in\Re^{n\times c}$ ($c\le n$) and its singular value decomposition (SVD) $\mathbf{A}=\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ ($\mathbf{U}\in\Re^{n\times n}$, $\mathbf{V}\in\Re^{c\times c}$), the solution of the following optimization problem: $(P_1)\ \min_{\mathbf{B}^T\mathbf{B}=\mathbf{I}}\|\mathbf{A}-\mathbf{B}\|_F^2$ is given by $\mathbf{B}=\mathbf{U}[\mathbf{I};\mathbf{0}]\mathbf{V}^T$ ($\mathbf{I}$ is the identity matrix with size c, $\mathbf{0}\in\Re^{(n-c)\times c}$ is a matrix with all zeros).*

Proof: When $\mathbf{A}$ is fixed, it can be verified that the problem (P_1) is equivalent to the following problem: $\max_{\mathbf{B}^T\mathbf{B}=\mathbf{I}}\mathbf{tr}\left(\mathbf{B}^T\mathbf{A}\right)$. We can derive that $\mathbf{tr}\left(\mathbf{B}^T\mathbf{A}\right)=\mathbf{tr}\left(\mathbf{B}^T\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T\right)=\mathbf{tr}\left(\mathbf{\Lambda}\mathbf{V}^T\mathbf{B}^T\mathbf{U}\right)=\mathbf{tr}\left(\mathbf{\Lambda}\mathbf{Z}\right)=\sum_i\lambda_{ii}z_{ii}$ $(1\le i\le c)$, where $\mathbf{Z}=\mathbf{V}^T\mathbf{B}^T\mathbf{U}$, and λ_{ii} and z_{ii} are the (i,i)-th entry of $\mathbf{\Lambda}$ and $\mathbf{Z}$, respectively. Note that $\mathbf{Z}\mathbf{Z}^T=\mathbf{I}$, thus $z_{ii}\le 1$. On the other hand, $\lambda_{ii}\ge 0$ as λ_{ii} is singular value of $\mathbf{A}$. Therefore, $\mathbf{tr}\left(\mathbf{B}^T\mathbf{A}\right)=\sum_i\lambda_{ii}z_{ii}\le\sum_i\lambda_{ii}$, and when $z_{ii}=1\,(1\le i\le c)$, the equality holds. That is to say, $\mathbf{tr}\left(\mathbf{B}^T\mathbf{A}\right)$ reaches its maximum when $\mathbf{Z}=[\mathbf{I},\mathbf{0}^T]$. Recall that $\mathbf{Z}=\mathbf{V}^T\mathbf{B}^T\mathbf{U}$, the solution to $\max_{\mathbf{B}^T\mathbf{B}=\mathbf{I}}\mathbf{tr}\left(\mathbf{B}^T\mathbf{A}\right)$ or the problem (P_1) is $\mathbf{B}=\mathbf{U}\mathbf{Z}^T\mathbf{V}^T=\mathbf{U}[\mathbf{I};\mathbf{0}]\mathbf{V}^T$. Theorem 1 is proved. □

Using Theorem 1, letting SVD of $\mathbf{X}^T\mathbf{W}+\mathbf{1}_n\mathbf{b}^T=\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$, the solution to Eq. (4) is $\mathbf{F}=\mathbf{U}[\mathbf{I};\mathbf{0}]\mathbf{V}^T$.

Next, when $\mathbf{F}$ is fixed, taking the derivative of the objective with respect to $\mathbf{b}$ and $\mathbf{w}_i(1\le i\le c)$, and setting it to zero, we obtain $\mathbf{b}=\mathbf{F}^T\mathbf{1}_n/n$ (because the data are centered) and have[1]:

$$\mathbf{X}\mathbf{X}^T\mathbf{w}_i-\mathbf{X}(\mathbf{f}_i-\mathbf{b}_i)+\gamma_1\mathbf{D}^i\mathbf{w}_i+\gamma_2\tilde{\mathbf{D}}\mathbf{w}_i=\mathbf{0}, \quad (5)$$

where $\mathbf{b}_i$ is an $n\times 1$ vector in which all elements are the i-th element of $\mathbf{b}$, $\mathbf{D}^i(1\le i\le c)$ is a block diagonal

[1]When $\|\mathbf{w}_i^j\|_2=0$, Eq. (3) is not differentiable. Following (Gorodnitsky & Rao, 1997), we can introduce a small perturbation to regularize the j-th diagonal block of $\mathbf{D}^i$ as $\frac{1}{2\sqrt{\|\mathbf{w}_i^j\|_2^2+\zeta}}\mathbf{I}_j$. Similarly, when $\|\mathbf{w}^i\|_2=0$, the i-th diagonal element of $\tilde{\mathbf{D}}$ can be regularized as $\frac{1}{2\sqrt{\|\mathbf{w}^i\|_2^2+\zeta}}$. Then it can be verified that the derived algorithm minimizes the following problem: $\sum_{i=1}^n\|\mathbf{W}^T\mathbf{x}_i+\mathbf{b}-\mathbf{f}^i\|_2^2+\gamma_1\sum_{i=1}^c\sum_{j=1}^k\sqrt{\|\mathbf{w}_i^j\|_2^2+\zeta}+\gamma_2\sum_{i=1}^d\sqrt{\|\mathbf{w}^i\|_2^2+\zeta}$, which is apparently reduced to problem Eq. (3) when $\zeta\to 0$.

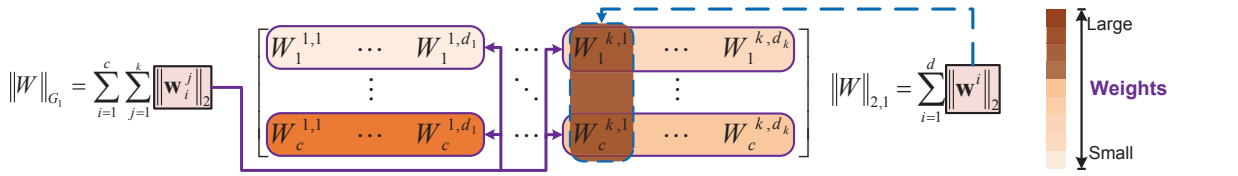

Figure 1. Illustration of the feature weight matrix W^T. The elements in matrix with deep orange color have large values. The group ℓ_1-norm (G_1-norm) emphasizes the learning of the ***group-wise*** weights for a type of features corresponding to each cluster and the $\ell_{2,1}$-norm accentuates the ***individual*** weight learning cross multiple clusters.

matrix with the j-th diagonal block as $\frac{1}{2\|\mathbf{w}_i^j\|_2}\mathbf{I}_j$, $\mathbf{I}_j$ is an identity matrix with size of d_j, $\mathbf{w}_i^j$ is the j-th segment of $\mathbf{w}_i$ and includes the weights of features in j-th view. $\tilde{\mathbf{D}}$ is a diagonal matrix with the i-th diagonal element as $\frac{1}{2\|\mathbf{w}^i\|_2}$. Thus we have:

$$\mathbf{w}_i = (\mathbf{X}\mathbf{X}^T + \gamma_1\mathbf{D}^i + \gamma_2\tilde{\mathbf{D}})^{-1}\mathbf{X}(\mathbf{f}_i - \mathbf{b}_i). \quad (6)$$

Note that $\mathbf{D}^i(1 \le i \le c)$ and $\tilde{\mathbf{D}}$ are dependent on $\mathbf{W}$ and thus are also unknown variables. We propose an iterative algorithm to solve this problem, which is described in Algorithm 1.

Algorithm 1 An efficient iterative algorithm to solve the optimization problem in Eq. (3).

Input: $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in \Re^{d\times n}$.
1. Let $t = 1$. Initialize $\mathbf{F}_t$ by K-means clustering and then initialize $\mathbf{W}_t$ and $\mathbf{b}_t$ by solving $\min_{\mathbf{W},\mathbf{b}} \left\|\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}\right\|_F^2$.
while not converge **do**
2. Calculate $\mathbf{F}_{t+1} = \mathbf{U}[\mathbf{I}; \mathbf{0}]\mathbf{V}^T$ where $\mathbf{U}$ and $\mathbf{V}$ are obtained by SVD on $\mathbf{X}^T\mathbf{W}_t + \mathbf{1}_n\mathbf{b}_t^T$.
3. Calculate $\mathbf{b}_{t+1} = \mathbf{F}_{t+1}^T\mathbf{1}_n/n$. Calculate the block diagonal matrices $\mathbf{D}_{t+1}^i(1 \le i \le c)$, where the j-th diagonal block of $\mathbf{D}_{t+1}^i$ is $\frac{1}{2\|(\mathbf{w}_t)_i^j\|_2}\mathbf{I}_j$. Calculate the diagonal matrix $\tilde{\mathbf{D}}_{t+1}$, where the i-th diagonal element is $\frac{1}{2\|\mathbf{w}_t^i\|_2}$.
4. For each $\mathbf{w}_i(1 \le i \le c)$, $(\mathbf{w}_{t+1})_i = (\mathbf{X}\mathbf{X}^T + \gamma_1\mathbf{D}_{t+1}^i + \gamma_2\tilde{\mathbf{D}}_{t+1})^{-1}\mathbf{X}((\mathbf{f}_{t+1})_i - (\mathbf{b}_{t+1})_i)$.
5. $t = t + 1$.
end while
Output: $\mathbf{W}_t \in \Re^{d\times c}$, $\mathbf{b}_t$, $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n]^T$.

Computational analysis. In Algorithm 1, step 2 solves a SVD problem. Because in typical clustering tasks, the number of clusters d is usually not very large, step 2 can be computed efficiently by many off-the-shelf numerical packages. Step 3 is computationally trivial. In step 4, instead of computing the matrix inverse with cubic complexity, we can solve a system of linear equations with quadratic complexity to obtain $(\mathbf{w}_{t+1})_i$. When sufficient computational resources are available and parallel computing is implemented, both SVD and linear equations, thereby the whole algorithm, can be solved with desired efficiency.

Convergence analysis. The following theorem guarantees the convergence of Algorithm 1.

Theorem 2 *Algorithm 1 decreases the objective value of Eq.* (3) *in each iteration.*

Proof: In each iteration t of Algorithm 1, according to Step 2, we know that

$$\mathbf{F}_{t+1} = \min_{\mathbf{F}} \left\|\mathbf{X}^T\mathbf{W}_t + \mathbf{1}_n\mathbf{b}_t^T - \mathbf{F}\right\|_F^2 + \gamma_1\sum_{i=1}^{c}\mathbf{D}_{t+1}^i\left\|(\mathbf{w}_t)_i\right\|_2^2 + \gamma_2 Tr\mathbf{W}_t^T\tilde{\mathbf{D}}_{t+1}\mathbf{W}_t. \quad (7)$$

According to Steps 3 and 4, we know that

$$\mathbf{W}_{t+1}, \mathbf{b}_{t+1} = \min_{\mathbf{W},\mathbf{b}} \left\|\mathbf{X}^T\mathbf{W} + \mathbf{1}_n\mathbf{b}^T - \mathbf{F}_{t+1}\right\|_F^2 + \gamma_1\sum_{i=1}^{c}\mathbf{D}_{t+1}^i\left\|(\mathbf{w})_i\right\|_2^2 + \gamma_2 Tr\mathbf{W}^T\tilde{\mathbf{D}}_{t+1}\mathbf{W}. \quad (8)$$

Thus, we can derive:

$$\begin{aligned}
&\left\|\mathbf{X}^T\mathbf{W}_{t+1} + \mathbf{1}_n\mathbf{b}_{t+1}^T - \mathbf{F}_{t+1}\right\|_F^2 \quad (9)\\
&+ \gamma_1\sum_{i=1}^{c}\mathbf{D}_{t+1}^i\left\|(\mathbf{w}_{t+1})_i\right\|_2^2 + \gamma_2 Tr\mathbf{W}_{t+1}^T\tilde{\mathbf{D}}_{t+1}\mathbf{W}_{t+1}\\
&\le \left\|\mathbf{X}^T\mathbf{W}_t + \mathbf{1}_n\mathbf{b}_t^T - \mathbf{F}_{t+1}\right\|_F^2\\
&+ \gamma_1\sum_{i=1}^{c}\mathbf{D}_{t+1}^i\left\|(\mathbf{w}_t)_i\right\|_2^2 + \gamma_2 Tr\mathbf{W}_t^T\tilde{\mathbf{D}}_{t+1}\mathbf{W}_t\\
&\le \left\|\mathbf{X}^T\mathbf{W}_t + \mathbf{1}_n\mathbf{b}_t^T - \mathbf{F}_t\right\|_F^2\\
&+ \gamma_1\sum_{i=1}^{c}\mathbf{D}_{t+1}^i\left\|(\mathbf{w}_t)_i\right\|_2^2 + \gamma_2 Tr\mathbf{W}_t^T\tilde{\mathbf{D}}_{t+1}\mathbf{W}_t\,.
\end{aligned}$$

Substituting $\tilde{\mathbf{D}}$ and $\mathbf{D}$ by definitions, we obtain:

$$\mathcal{L}_{t+1} + \gamma_1 \sum_{i=1}^{c}\sum_{j=1}^{k} \frac{\left\|(\mathbf{w}_{t+1})_i^j\right\|_2^2}{2\left\|(\mathbf{w}_t)_i^j\right\|_2} + \gamma_2 \sum_{i=1}^{d} \frac{\left\|\mathbf{w}_{t+1}^i\right\|_2^2}{2\left\|\mathbf{w}_t^i\right\|_2} \leq$$
$$\mathcal{L}_t + \gamma_1 \sum_{i=1}^{c}\sum_{j=1}^{k} \frac{\left\|(\mathbf{w}_t)_i^j\right\|_2^2}{2\left\|(\mathbf{w}_t)_i^j\right\|_2} + \gamma_2 \sum_{i=1}^{d} \frac{\left\|\mathbf{w}_t^i\right\|_2^2}{2\left\|\mathbf{w}_t^i\right\|_2}, \quad (10)$$

where $\mathcal{L}_t = ||\mathbf{X}^T\mathbf{W}_t + \mathbf{1}_n\mathbf{b}_t^T - \mathbf{F}_t||_F^2$. Because it can be easily verified that for function $f(x) = x - \frac{x^2}{2\alpha}$, given any $x \neq \alpha \in \Re$, $f(x) \leq f(\alpha)$ holds, we can derive that

$$\sum_{j=1}^{k}\left\|(\mathbf{w}_{t+1})_i^j\right\|_2 - \sum_{j=1}^{k} \frac{\left\|(\mathbf{w}_{t+1})_i^j\right\|_2^2}{2\left\|(\mathbf{w}_t)_i^j\right\|_2} \leq$$
$$\sum_{j=1}^{k}\left\|(\mathbf{w}_t)_i^j\right\|_2 - \sum_{j=1}^{k} \frac{\left\|(\mathbf{w}_t)_i^j\right\|_2^2}{2\left\|(\mathbf{w}_t)_i^j\right\|_2}, \quad (11)$$

and

$$\sum_{i=1}^{d}\left\|\mathbf{w}_{t+1}^i\right\|_2 - \sum_{i=1}^{d} \frac{\left\|\mathbf{w}_{t+1}^i\right\|_2^2}{2\left\|\mathbf{w}_t^i\right\|_2} \leq$$
$$\sum_{i=1}^{d}\left\|\mathbf{w}_t^i\right\|_2 - \sum_{i=1}^{d} \frac{\left\|\mathbf{w}_t^i\right\|_2^2}{2\left\|\mathbf{w}_t^i\right\|_2}. \quad (12)$$

Adding Eqs. (9-12) on both sides (note Eq. (11) is repeated for $1 \leq i \leq c$), we have

$$\mathcal{L}_{t+1} + \gamma_1 \sum_{i=1}^{c}\sum_{j=1}^{k}\left\|(\mathbf{w}_{t+1})_i^j\right\|_2 + \gamma_2 \sum_{i=1}^{d}\left\|\mathbf{w}_{t+1}^i\right\|_2 \leq$$
$$\mathcal{L}_t + \gamma_1 \sum_{i=1}^{c}\sum_{j=1}^{k}\left\|(\mathbf{w}_t)_i^j\right\|_2 + \gamma_2 \sum_{i=1}^{d}\left\|\mathbf{w}_t^i\right\|_2. \quad (13)$$

Therefore, the algorithm decreases the objective value in each iteration. □

Upon convergence, $\mathbf{W}_t$, $\mathbf{b}_t$, $\mathbf{D}_t^i (1 \leq i \leq c)$ and $\tilde{\mathbf{D}}_t$ will satisfy the Eq. (6), *i.e.*, the K.K.T. conditions is satisfied, which indicates the algorithm converges to a local solution of the problem.

Clustering Rules. Given the input data $\mathbf{X}$, we can compute the projection matrix $\mathbf{W}$ and the cluster indication matrix $\mathbf{F}$ by Algorithm 1. Upon solution, we cluster the data samples $\{\mathbf{x}_i\}_{i=1}^n$ by performing K-means clustering on $\mathbf{F}$.

2.3. Extension to Supervised Classification

The proposed clustering framework can also be extended to a supervised multi-view classification method. Suppose the data samples are labeled into c classes, which are represented by a class indication matrix $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_n]^T$, such that $y_{ij} = 1$ if data point $\mathbf{x}_i$ belongs to the j-th class, $y_{ij} = 0$ otherwise. Then we can perform supervised multi-view classification by simply replacing $\mathbf{F}$ in our objective in Eq. (3) by $\mathbf{Y}$. Because $\mathbf{Y}$ is fixed upon the labeling knowledge, Step 2 in Algorithm 1 is skipped and $\mathbf{b} = \mathbf{Y}^T\mathbf{1}_n/n$. Upon solution, we can classify an unseen data point by $\arg\max_j \left(\mathbf{W}^T\mathbf{x} + \mathbf{b}\right)_j$.

The extended supervised classification method is advantageous for the following reasons. First, similar to the proposed method for clustering, the new classification method explicitly computes the feature weight coefficients $\mathbf{W}$ for both each type of features and each feature within one single type. As a result, compared to the MKL methods, our new classification method has the capability to identify both useful feature types (views) and relevant individual features. Therefore, the features are properly weighted at two levels of granularity upon their relevance to the semantic classes of interest, which could lead to improved classification results. Second, because no SVD is involved in the algorithm for classification task, the computational complexity is approximately linear when sufficient computational resources are available and parallel computing is implemented, *i.e.*, our method scales well to large-scale data and is suitable for practical use to solve real-world problems.

3. Experiments

In this section, we experimentally evaluate the proposed multi-view learning framework in both clustering and classifications tasks on five broadly used multi-view data sets, including three image data sets, one Protein data set and one Multi-Lingual (ML) Text analysis data set. Each data set has a certain number of types of features (views), whose details are described as following and summarized in Table 1.

Image annotation. To minimize the gap between the low-level visual features and high-level semantic concepts, an image can be abstracted by a variety of different descriptors, and each type of these descriptors naturally forms up a view of the images of interest. We evaluate our new multi-view learning framework on the following three broadly used benchmark data sets, including **NUS-WIDE-Object** data set[2], **Animal** data set[3] and **MSRC-v1** data set[4].

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

[3] http://attributes.kyb.tuebingen.mpg.de/

[4] http://research.microsoft.com/en-us/projects/objectclassrecognition/

Table 1. Details of the multi-view data sets used in our experiments (feature type (dimensionality)).

Feature type	NUS-WIDE-Object	Animal	MSRC-v1	Protein	ML Text
1	Color moments (255)	Self-Similarity (2000)	Color moment (48)	Pfam (3375)	English (2000)
2	Color histogram (64)	Color histogram (2688)	LBP (256)	FFT (4910)	German (2000)
3	Color correlogram (144)	PyramidHOG (252)	HOG (100)	Gene expression (441)	French (2000)
4	Wavelet texture (128)	SIFT (2000)	SIFT (1230)	–	Japanese (2000)
5	Edge distribution (73)	colorSIFT (2000)	GIST (512)	–	–
6	Visual words (500)	SURF (2000)	Centrist (1320)	–	–
Data points	30000	30457	240	1500	5000
Classes	31	50	7	2	28

Protein categorization. Protein can be characterized from different aspects, each of which can be seen as a view. The Berkeley genomic data set[5] (Lanckriet et al., 2004b) is used in our studies.

Multi-lingual document analysis. With the advances of machine translation techniques, one can easily get different translations for one document (Prettenhofer & Stein, 2010), and the translation in each language can be considered as a view. We apply our new multi-view learning framework on the **multi-lingual (ML) text** data set[6] for document analysis.

3.1. Improved Multi-View Clustering

In this subsection, we first evaluate the multi-view clustering capability of the proposed method.

Experimental setup. Following (Cai et al., 2011), as a baseline, we apply the spectral clustering (SC) algorithm (Ng et al., 2001) on every data set using each single type of features. Besides, we also apply SC on the concatenation of all the features in different types, which is equivalent to assume that all the feature types are of the same importance and does not distinguish the feature relevance within a feature type. For SC, we need to build a graph from the input data. Following (He et al., 2005), we construct the nearest-neighbor graph, where the neighborhood size for the graph construction is set as optimal by searching the grid of $\{1, 2, \ldots, 10\}$.

We compare our method against two most recent multi-view clustering methods: the Multi-Modal Spectral Clustering (MMSC) method (Cai et al., 2011) and Co-regularized Multi-view Spectral Clustering (CMSC) method (Kumar et al., 2011), which have demonstrated state-of-the-art clustering performance on multi-view data. We implement the compared methods following the original works and set the parameters as optimal by performing cross validations in our preliminary experiments using the ground truth data labels. For our method, we fine tune the parameters γ_1 and γ_2 in Eq. (3) by searching the grid of $\left\{10^{-5}, 10^{-4}, \ldots, 10^{4}, 10^{5}\right\}$ following the same strategy.

We implement four versions of the proposed method to evaluate the effectiveness of its component terms in multi-view learning. First, we implement our method by only using the first term of Eq. (3) and denote it as "loss only", which is equivalent to use linear regression to perform the clustering on the concatenation of all the features from different types. Second, we implement our method by only imposing the group ℓ_1-norm regularization and denote it as "G_1-norm", which, similar to MKL, only takes into account type-wise relevance but not feature-wise relevance. Third, we implement our method by only imposing the $\ell_{2,1}$-norm regularization and denote it as "$\ell_{2,1}$-norm", which thus is reduced a typical multi-task feature selection method and only takes into account feature-wise relevance. Finally, we implement the full version of the proposed method as defined in Eq. (3).

Comparison results. Because the clustering results of compared methods are dependent on the initial values, we repeat each experiment on each setting for 50 times and report the average performance. The comparison results measured by clustering accuracy and normalized mutual information are reported in Table 2 and Table 3 respectively.

A first glance at the experimental results in Table 2 and Table 3 shows that the three multi-view clustering methods, including ours, are generally better than SC on each individual data view, which validate the usefulness of data integration in clustering. In addition, the three multi-view methods also outperform SC on the concatenation of all features, which is reasonable in that multi-view methods learn proper weights for different views (and features by our method) upon their relevance to the data clusters while the simple features concatenation does not has such capability.

Moreover, for the three compared multi-view clustering methods, our method is always better by a large margin. This observation is consistent with our theoretical analysis in that both MMSC and CMSC only learn the weights at feature type level, while ignoring

[5] http://noble.gs.washington.edu/proj/sdp-svm/

[6] http://www.webis.de/research/corpora/

Table 2. Clustering performance comparison measured by clustering accuracy.

Method	NUS-WIDE-object	Animal	MSRC-v1	Protein	ML Text
SC (Type 1)	0.151 ± 0.018	0.508 ± 0.015	0.712 ± 0.023	0.627 ± 0.019	0.505 ± 0.020
SC (Type 2)	0.157 ± 0.020	0.513 ± 0.019	0.716 ± 0.022	0.292 ± 0.023	0.513 ± 0.019
SC (Type 3)	0.153 ± 0.022	0.507 ± 0.018	0.714 ± 0.021	0.651 ± 0.019	0.508 ± 0.024
SC (Type 4)	0.160 ± 0.018	0.518 ± 0.021	0.721 ± 0.025	–	0.519 ± 0.021
SC (Type 5)	0.168 ± 0.023	0.529 ± 0.022	0.734 ± 0.019	–	–
SC (Type 6)	0.171 ± 0.021	0.534 ± 0.021	0.741 ± 0.022	–	–
SC (All by concatenation)	0.201 ± 0.013	0.571 ± 0.017	0.768 ± 0.016	0.656 ± 0.018	0.603 ± 0.021
MMSC	0.214 ± 0.018	0.585 ± 0.016	0.783 ± 0.019	0.662 ± 0.020	0.614 ± 0.015
CMSC	0.217 ± 0.015	0.580 ± 0.018	0.779 ± 0.015	0.667 ± 0.019	0.617 ± 0.021
Our method (loss only)	0.202 ± 0.021	0.563 ± 0.014	0.764 ± 0.015	0.661 ± 0.013	0.601 ± 0.018
Our method (G_1-norm)	0.223 ± 0.015	0.598 ± 0.011	0.797 ± 0.013	0.725 ± 0.015	0.620 ± 0.015
Our method ($\ell_{2,1}$-norm)	0.227 ± 0.013	0.603 ± 0.014	0.804 ± 0.015	0.738 ± 0.015	0.627 ± 0.013
Our method	$\mathbf{0.238 \pm 0.015}$	$\mathbf{0.629 \pm 0.013}$	$\mathbf{0.827 \pm 0.016}$	$\mathbf{0.747 \pm 0.011}$	$\mathbf{0.646 \pm 0.009}$

Table 3. Clustering performance comparison measured by normalized mutual information.

Method	NUS-WIDE-object	Animal	MSRC-v1	Protein	ML Text
SC (Type 1)	0.179 ± 0.019	0.607 ± 0.020	0.851 ± 0.023	0.751 ± 0.021	0.602 ± 0.019
SC (Type 2)	0.182 ± 0.021	0.609 ± 0.022	0.856 ± 0.019	0.757 ± 0.023	0.604 ± 0.025
SC (Type 3)	0.180 ± 0.018	0.613 ± 0.026	0.859 ± 0.021	0.760 ± 0.024	0.611 ± 0.021
SC (Type 4)	0.187 ± 0.021	0.617 ± 0.021	0.864 ± 0.019	–	0.609 ± 0.020
SC (Type 5)	0.184 ± 0.023	0.622 ± 0.024	0.866 ± 0.018	–	–
SC (Type 6)	0.190 ± 0.021	0.628 ± 0.026	0.861 ± 0.022	–	–
SC (All by concatenation)	0.238 ± 0.021	0.681 ± 0.023	0.916 ± 0.019	0.811 ± 0.021	0.718 ± 0.025
MMSC	0.251 ± 0.022	0.698 ± 0.023	0.931 ± 0.023	0.826 ± 0.026	0.730 ± 0.021
CMSC	0.256 ± 0.026	0.695 ± 0.021	0.926 ± 0.024	0.828 ± 0.019	0.733 ± 0.023
Our method (loss only)	0.235 ± 0.016	0.677 ± 0.021	0.911 ± 0.023	0.809 ± 0.024	0.716 ± 0.024
Our method (G_1-norm)	0.262 ± 0.022	0.719 ± 0.016	0.942 ± 0.021	0.841 ± 0.021	0.752 ± 0.023
Our method ($\ell_{2,1}$-norm)	0.271 ± 0.021	0.726 ± 0.026	0.948 ± 0.021	0.850 ± 0.018	0.759 ± 0.016
Our method	$\mathbf{0.282 \pm 0.018}$	$\mathbf{0.751 \pm 0.019}$	$\mathbf{0.987 \pm 0.019}$	$\mathbf{0.865 \pm 0.020}$	$\mathbf{0.769 \pm 0.018}$

the relevance of each individual features, especially for those in low-weight feature types. In contrast, our method is particularly designed to take into account the feature weighting at the two levels of granularity, which is confirmed to be effective in data clustering by all the experimental results reported in Table 2 and 3.

Finally, the full version of our new method outperforms all its three degenerate versions, which demonstrate the correctness of our objective and the usefulness of its two regularization terms that capture both the global and local aspects of feature relevances.

Analysis of learned view relevance and feature relevance. Besides the clustering performance comparison, we examine the feature weight matrix W learned from Eq. (3) with some details, because the most important advantage of our new method over other competing multi-view learning methods lies in its capability for simultaneous view selection and individual feature selection. First, for example, for MSRC-v1 data set we notice that the overall sparsity of the learned coefficient matrix W is 22.1%. In contrast, for the image cluster related to the "outdoor" concept, the sparsity of "color moment" feature type is about 59.7% and the relative weights of the Color/SIFT/LBP/HOG/GIST/CENTRIST are about 1/0.81/0.63/0.35/0.86/0.42, which clearly shows that the color features and GIST features are of the most significant importance when we determine whether an image belongs to the "outdoor" class. This observation perfectly agrees with our common sense and empirically justifies the correctness the proposed method in terms of view selection. Second, although the relative weights of the Color/SIFT/LBP/HOG/GIST/CENTRIST features for the image cluster related to the "car" concept are about 1/1.12/0.60/0.33/0.46/0.68, two HOG features have considerably high relative weights of 0.27% and 0.26%. Such high relative weights, compared to the average relative weights of all non-zero features of 0.12%, indicates the high discriminative power of these two HOG features, although the overall importance of HOG features is the lowest compared to the other 5 types of image features. This result concretely confirms that our new method is able to select the useful individual features from feature groups with very weak influences. In summary, empirical results validate the proposed method for its capability to learn both view-wise and individual feature-wise relevances.

3.2. Improved Multi-View Classification

Now we evaluate the supervised extension of the proposed method in multi-view classification.

We apply SVM on each individual type of features and the concatenation of all types of features of the experimental data sets as baselines. We compare our method against several most recent multiple kernel learning

Table 4. Classification performance comparison measured by classification accuracy.

Method	NUS-WIDE-object	Animal	MSRC-v1	Protein	ML Text
SVM (Type 1)	0.152 ± 0.018	0.542 ± 0.016	0.777 ± 0.019	0.682 ± 0.023	0.543 ± 0.019
SVM (Type 2)	0.149 ± 0.020	0.551 ± 0.019	0.768 ± 0.018	0.318 ± 0.025	0.549 ± 0.021
SVM (Type 3)	0.146 ± 0.016	0.569 ± 0.021	0.781 ± 0.022	0.708 ± 0.021	0.545 ± 0.018
SVM (Type 4)	0.150 ± 0.018	0.541 ± 0.023	0.784 ± 0.026	–	0.548 ± 0.025
SVM (Type 5)	0.141 ± 0.017	0.566 ± 0.021	0.773 ± 0.023	–	–
SVM (Type 6)	0.149 ± 0.018	0.554 ± 0.200	0.789 ± 0.021	–	–
SVM (all by concatenation)	0.138 ± 0.020	0.547 ± 0.019	0.793 ± 0.025	0.714 ± 0.018	0.649 ± 0.020
SVM ℓ_∞ MKL method	0.211 ± 0.023	0.603 ± 0.017	0.820 ± 0.023	0.758 ± 0.021	0.551 ± 0.026
SVM ℓ_1 MKL method	0.207 ± 0.020	0.599 ± 0.019	0.813 ± 0.019	0.744 ± 0.022	0.554 ± 0.029
SVM ℓ_2 MKL method	0.202 ± 0.021	0.593 ± 0.018	0.789 ± 0.022	0.737 ± 0.019	0.556 ± 0.023
LSSVM ℓ_∞ MKL method	0.200 ± 0.018	0.588 ± 0.025	0.778 ± 0.025	0.755 ± 0.024	0.560 ± 0.021
LSSVM ℓ_1 MKL method	0.195 ± 0.022	0.586 ± 0.023	0.808 ± 0.027	0.729 ± 0.025	0.564 ± 0.022
LSSVM ℓ_2 MKL method	0.187 ± 0.021	0.578 ± 0.019	0.796 ± 0.018	0.789 ± 0.018	0.667 ± 0.019
Our method (loss only)	0.141 ± 0.021	0.561 ± 0.022	0.790 ± 0.026	0.716 ± 0.023	0.652 ± 0.016
Our method (G_1-norm)	0.229 ± 0.019	0.641 ± 0.020	0.839 ± 0.023	0.789 ± 0.018	0.667 ± 0.019
Our method ($\ell_{2,1}$-norm)	0.235 ± 0.021	0.648 ± 0.023	0.847 ± 0.021	0.803 ± 0.018	0.682 ± 0.026
Our method	**0.251 ± 0.020**	**0.665 ± 0.018**	**0.869 ± 0.016**	**0.813 ± 0.021**	**0.718 ± 0.018**

(MKL) methods that are able to make use of multiple types of data: (1) SVM ℓ_∞ MKL method (Sonnenburg et al., 2006), (2) SVM ℓ_1 MKL (Lanckriet et al., 2004a), (3) SVM ℓ_2 MKL method (Kloft et al.), (4) least square (LSSVM) ℓ_∞ MKL method (Ye et al., 2008a), (5) LSSVM ℓ_1 MKL method (Suykens et al., 2002) and (6) LSSVM ℓ_2 MKL method (Yu et al., 2010). Same as before, four versions of our method are implemented and evaluated.

We conduct standard 5-fold cross-validation and report the average results. For each of the 5 trials, within the training data, an internal 5-fold cross-validation is performed to fine tune the parameters. The parameters of our method (γ_1 and γ_2 in Eq. (3)) are optimized in the range of $\left\{10^{-5}, 10^{-4}, \ldots, 10^4, 10^5\right\}$. For the SVM method and MKL methods, one Gaussian kernel is constructed for each type of features (*i.e.*, $\mathcal{K}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left(-\gamma \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2\right)$), where the parameters γ are fine tuned in the same range used in our method. We implement the compared MKL methods using the codes published by (Yu et al., 2010). Following (Yu et al., 2010), in LSSVM ℓ_∞ and ℓ_2 methods, the regularization parameter λ is estimated jointly as the kernel coefficient of an identity matrix; in LSSVM ℓ_1 method, λ is set to 1; in all other SVM approaches, the C parameter of the box constraint is fine tuned in the same range used for our method. We use LIBSVM-software package to implement SVM in all our experiments. The classification performances measured by average classification accuracies of all compared methods on the five data sets are reported in Table 4.

Table 4 shows that our method consistently outperforms all other compared methods, which demonstrate the effectiveness of our method in supervised multi-view classification. In addition, the methods that use multiple data sources are generally better than SVM using each one single type of data. This confirms the usefulness of data integration in supervised multi-view learning. Moreover, the results that our method is always better than the MKL methods, though both of them take advantage of data from multiple different sources, are consistent with our theoretical analysis. That is, our method not only assigns proper weight to each type of data, but also considers the relevances of the features inside each individual type of data. In contrast, the MKL methods only address the former while not being able to take into account the latter. These important observations, again, concretely demonstrate the advantages of the proposed multi-view learning framework in classification tasks. Finally, the full version of the proposed method is clearly superior to its degenerate versions, which prove the necessity of the both regularization terms of the proposed method in multi-view learning.

4. Conclusion

In this paper, we proposed a novel multi-view learning model to efficiently learn the weights of individual feature on different clusters when all heterogenous features are integrated. The joint sparsity-inducing norms are utilized to impose the structured sparsity on the learned weight (parameter) matrix from both local and global multi-view viewpoints. Compared to existing state-of-the-art multi-view clustering methods approaches, our new methods capture the importance of local features and achieve better performance in both unsupervised and supervised multi-view learning tasks.

Acknowledgments

Corresponding Author: Heng Huang (heng@uta.edu)
This work was partially supported by NSF CCF-0830780, CCF-0917274, DMS-0915228, IIS-1117965.

References

Argyriou, Andreas, Evgeniou, Theodoros, and Pontil, Massimiliano. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

Bach, F., Lanckriet, G.R.G., and Jordan, M.I. Multiple Kernel Learning, Conic Duality, and the SMO Algorithm. *ICML*, 2004.

Belkin, M, Niyogi, P, and Sindhwani, V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 1:1–48, 2006.

Bickel, Steffen and Scheffer, Tobias. Multi-view clustering. *IEEE International Conference on Data Mining*, 2004.

Brefeld, Ulf and Scheffer, Tobias. Co-em support vector learning. *International Conference on Machine Learning*, 2004.

Cai, X., Nie, F., Huang, H., and Kamangar, F. Heterogeneous image feature integration via multi-modal spectral clustering. In *CVPR*, 2011.

Ghani, R. Combining labeled and unlabeled data for multiclass text categorization. *International Conference on Machine Learning*, 2002.

Gorodnitsky, I.F. and Rao, B.D. Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. *Signal Processing, IEEE Transactions on*, 45(3):600–616, 1997.

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

Kloft, M., Brefeld, U., Laskov, P., and Sonnenburg, S. Non-sparse multiple kernel learning. In *NIPS*.

Kumar, A., Rai, P., and Daumé III, H. Co-regularized multi-view spectral clustering. In *NIPS*, 2011.

Lanckriet, G.R.G., Cristianini, N., Bartlett, P., Ghaoui, L.E., and Jordan, M.I. Learning the kernel matrix with semidefinite programming. *JMLR*, 5:27–72, 2004a. ISSN 1532-4435.

Lanckriet, G.R.G., De Bie, T., Cristianini, N., Jordan, M.I., and Noble, W.S. A statistical framework for genomic data fusion. *Bioinformatics*, 2004b. ISSN 1367-4803.

Ng, Andrew Y., Jordan, Michael I., and Weiss, Yair. On spectral clustering: Analysis and an algorithm. In *NIPS*, pp. 849–856, 2001.

Nie, Feiping, Xu, Dong, Tsang, Ivor W., and Zhang, Changshui. Spectral embedded clustering. In *IJCAI*, pp. 1181–1186, 2009.

Obozinski, Guillaume, Taskar, Ben, and Jordan, Michael I. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20:231–252, 2010.

Prettenhofer, P. and Stein, B. Cross-language text classification using structural correspondence learning. In *ACL*, 2010.

Quattoni, Ariadna, Carreras, Xavier, Collins, Michael, and Darrell, Trevor. An Efficient Projection for $l_{1,\infty}$ Regularization. *ICML*, 2009.

Sonnenburg, S., Rätsch, G., Schäfer, C., and Schölkopf, B. Large scale multiple kernel learning. *JMLR*, 7:1531–1565, 2006. ISSN 1532-4435.

Suykens, J.A.K., Van Gestel, T., and De Brabanter, J. *Least squares support vector machines*. World Scientific Pub Co Inc, 2002. ISBN 9812381511.

Wang, Hua, Nie, Feiping, Huang, Heng, Risacher, Shannon, Saykin, Andrew J, and Shen, Li. Identifying adsensitive and cognition-relevant imaging biomarkers via joint classification and regression. In *MICCAI 2011*, pp. 115–123. Springer, 2011.

Wang, Hua, Nie, Feiping, Huang, Heng, Kim, Sungeun, Nho, Kwangsik, Risacher, Shannon L, Saykin, Andrew J, Shen, Li, et al. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort. *Bioinformatics*, 28(2):229–237, 2012a.

Wang, Hua, Nie, Feiping, Huang, Heng, Risacher, Shannon L, Saykin, Andrew J, Shen, Li, et al. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics*, 28(12):i127–i136, 2012b.

Wang, Hua, Nie, Feiping, Huang, Heng, Yan, Jingwen, Kim, Sungeun, Nho, Kwangsik, Risacher, Shannon L., Saykin, Andrew J., Shen, Li, and for the Alzheimer's Disease Neuroimaging Initiative. From phenotype to genotype: an association study of longitudinal phenotypic markers to alzheimer's disease relevant snps. *Bioinformatics*, 28(18):i619–i625, 2012c.

Wang, Hua, Nie, Feiping, Huang, Heng, Yan, Jingwen, Kim, Sungeun, Risacher, Shannon, Saykin, Andrew, and Shen, Li. High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression prediction. In *NIPS*, 2012d.

Wang, Hua, Nie, Feiping, Huang, Heng, and Ding, Chris. Heterogeneous Visual Features Fusion via Sparse Multimodal Machine. In *CVPR 2013*, 2013.

Ye, J., Ji, S., and Chen, J. Multi-class discriminant kernel learning via convex programming. *JMLR*, 9:719–758, 2008a. ISSN 1532-4435.

Ye, Jieping, Zhao, Zheng, and Wu, Mingrui. Discriminative k-means for clustering. In *NIPS*, pp. 1649–1656, 2008b.

Yu, S., Falck, T., Daemen, A., Tranchevent, L.C., Suykens, J.A.K., De Moor, B., and Moreau, Y. L 2-norm multiple kernel learning and its application to biomedical data fusion. *BMC bioinformatics*, 11(1):309, 2010. ISSN 1471-2105.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, 68(1):49–67, 2006.