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

Multi-View Correlated Feature Learning by Uncovering Shared Component

Xiaowei Xue,[1] Feiping Nie,[2*] Sen Wang,[3] Xiaojun Chang,[4] Bela Stantic,[3] Min Yao[1]
[1]College of Computer Science, Zhejiang University, P.R. China
[2]School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China
[3]School of Information and Communication Technology, Griffith University, Australia
[4]Centre for Quantum Computation and Intelligent Systems (QCIS), University of Technology Sydney, Australia
{xwxue, myao}@zju.edu.cn, {feipingnie, cxj273}@gmail.com, {sen.wang, b.stantic}@griffith.edu.au

Abstract

Learning multiple heterogeneous features from different data sources is challenging. One research topic is how to exploit and utilize the correlations among various features across multiple views with the aim of improving the performance of learning tasks, such as classification. In this paper, we propose a new multi-view feature learning algorithm that simultaneously analyzes features from different views. Compared to most of the existing subspace learning methods that only focus on exploiting a shared latent subspace, our algorithm not only learns individual information in each view but also captures feature correlations among multiple views by learning a shared component. By assuming that such a component is shared by all views, we simultaneously exploit the shared component and individual information of each view in a batch mode. Since the objective function is non-smooth and difficult to solve, we propose an efficient iterative algorithm for optimization with guaranteed convergence. Extensive experiments are conducted on several benchmark datasets. The results demonstrate that our proposed algorithm performs better than all the compared multi-view learning algorithms.

Introduction

In recent years, due to the fact that the data representation is becoming more diverse than before, the heterogeneous features fusion has attract much research attention in various applications, such as computer vision (Peng et al. 2016; Wang et al. 2013; Zhu, Li, and Zhang 2016; Chang et al. 2016b), social media analysis (Meng, Tan, and Xu 2014; Yang et al. 2013b), biomedical research (Liu et al. 2015b; Nie et al. 2015; Zhao et al. 2016; Wang et al. 2016), etc. For example, an image can have many representations with respect to different types of visual features, e.g. texture and color features. Content-based understanding of the image can significantly benefit from properly fusing visual features in different aspects. Multi-view learning has been well studied as the solution to feature fusion over the past years. A number of previous works (Zheng et al. 2015; Liu et al. 2015a; Nie, Li, and Li 2016; Chang et al. 2016a; Gong et al. 2014; Wang et al. 2014) on multi-view learning has demonstrated that sophisticated learning algorithms can perform remarkably better performance than single-view learning that only uses one type of feature or simply uses all types of features as one feature.

In literatures (Xu, Tao, and Xu 2013; Sun 2013), multi-view learning can be roughly categorized into three groups: 1) co-training, 2) multiple kernel learning, and 3) subspace learning. Being one of representative works on semi-supervised learning, the co-training was firstly introduced in (Blum and Mitchell 1998). Assuming that each data point is described by independent features in two views, co-training trains a classifier using labeled data in each view. Predictions on new unlabeled data in one view are mutually used to enlarge the training set of the other view. Other learning techniques have also been combined to achieve better learning results in different applications. Expectation-maximization has been combined with co-training in (Nigam and Ghani 2000; Parker and Khan 2015) for lower errors. In (Brefeld and Scheffer 2004), SVM was used to develop an extended version of co-EM for multi-view learning. In (Yu et al. 2011), a Bayesian undirected graphical model was used for co-training. Muslea et al. (Muslea, Minton, and Knoblock 2002) claimed that active learning is beneficial to the co-training regarding robustness in the multi-view learning problem. In Multiple Kernel Learning (MKL), each kernel can be regarded as a view. A typical MKL algorithm aims to learn an ensemble of multiple kernels for better performance of a certain application. Over the past few years, a number of MKL algorithms (Kloft et al. 2008; Gönen and Alpaydin 2008; Xu et al. 2009; Cortes, Mohri, and Rostamizadeh 2009; Varma and Babu 2009; Yu et al. 2010) have been proposed and demonstrated that either linear or non-linear combinations of multiple kernels can lead to better performance eventually. However, most of MKL algorithms fail to consider correlations between views that might broadly exist in the real-world scenarios, such as multimedia analysis.

In the practical applications, such as multimedia domain, each view is usually represented in a high-dimensional feature space which always leads to the *curse of dimension-*

ality problem. Rather than adopting the traditional feature analysis technologies to reduce dimensionality for single-view learning problems (Zhu et al. 2016; Chang et al. 2015; Chang and Yang 2016; Yang et al. 2013a), many works on multi-view learning assume that there is a low-dimensional subspace shared by different views. Exploiting such a shared subspace not only solves the *curse of dimensionality* problem when facing multiple high-dimensional features in different views but obtains improved performance from discovering latent variables in the shared latent subspace. Canonical Correlation Analysis (CCA) mutually maximizes the correlations between two views and learns a common subspace in an unsupervised manner. The kernel extension of CCA, namely KCCA, applies a kernel function mapping data into a high-dimensional space in which the original non-linearly separable data becomes linearly separable. Diethe et.al (Diethe, Hardoon, and Shawe-Taylor 2008) explored the latent subspace spanned by the multiple views by generalizing the Fisher's discriminant analysis. Similarly, some other methods, such as Gaussian process (Sigal, Memisevic, and Fleet 2009) and Markov network (Chen, Zhu, and Xing 2010) are adopted to find the latent space. However, since this kind of methods assumes that a latent subspace is shared by all the views, they may lose the specific information of different views.

To address the aforementioned issues, in this paper, we propose a novel multi-view feature learning algorithm to exploit correlations between different views. Compared to the MKL-based algorithms that learn combination of multiple kernels in different views and ignore the potential correlations between the views, the proposed method learns all the features from multiple views and simultaneously considers correlated information across views by exploiting a low-dimensional subspace shared by different views. Moreover, different from most of the existing subspace learning algorithms that only focus on exploiting the shared latent subspace, we assume that each view should possess its unique information. To achieve this goal, for each view, the proposed algorithm is enforced to learn a subspace that consists of two kinds of information: shared information across all views and individual information within that single view. Simultaneously, we train a linear classifier based on all the individual information in each view and the shared information. To step further, we propose an iterative algorithm in a joint framework to squeeze two different types of knowledge in multi-view data until global optima is reached. By taking both shared component and individual information in each view into consideration, we can further improve the performance for multi-view learning tasks. We name our proposed algorithm Multi-View Correlated feature learning with Shared Component(MVCS). The main contributions of this work are summarized as follows:

- We propose a novel multi-view feature learning algorithm that can simultaneously learn features in all views and exploit both a common component and individual information in each view.
- We propose an iterative algorithm with guaranteed convergence to efficiently optimize the objective function. Experimental results show that the proposed algorithm converges within 10 iterations on all benchmark datasets.
- Extensive experiments are conducted on several benchmark datasets to evaluate the effectiveness of the proposed algorithm. The results demonstrate that our algorithm performs better than state-of-the-art multi-view feature learning algorithms across all the datasets.

The rest of this paper is organized as follows. We detail the proposed feature learning framework in Section 2, followed by optimization algorithm to this problem in Section 3 and convergence analysis in Section 4. The experimental results are shown in Section 5. Section 6 draws the conclusion.

Multi-View Correlated Feature Learning with Shared component Framework

In this section, we first systematically describe the novel multi-view feature learning framework by mining correlations between different views to improve subsequent classification performances, followed by an efficient algorithm with guaranteed convergence to solve the objective function. In this paper, we write the matrices as bold uppercase letters and vector as bold lowercase letters. For arbitrary matrix $\mathbf{A}$, $\|\mathbf{A}\|_F$ denotes the Frobenius norm of matrix $\mathbf{A}$.

Given a set of n data samples $\mathbf{x}_i|_{i=1}^n$, we have data matrix $\mathbf{X}_i = [\mathbf{x}_1^i, \cdots, \mathbf{x}_n^i] \in \mathbb{R}^{d_i \times n} (i = 1, \cdots, k)$, where k is the number of views and d_i denotes the feature dimension of the i-th view. The label matrix $\mathbf{Y} \in \mathbb{R}^{n \times c}$, where c is the number of classes. Our goal in multi-view classification is to classify each sample into c classes by exploiting the correlations among all k different views of the training samples. We propose to learn features by minimizing the following objective function:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{P}_i,\mathbf{b}_i,\mathbf{b},\mathbf{Z}_i} & \sum_{i=1}^{k} \|\mathbf{X}_i^T\mathbf{P}_i + \mathbf{1}\mathbf{b}_i^T - \mathbf{Z}_i\|_F^2 \\ & +\alpha\|[\mathbf{Z}_1 \cdots \mathbf{Z}_k]\mathbf{W} + \mathbf{1}\mathbf{b}^T - \mathbf{Y}\|_F^2\end{aligned} \tag{1}$$

where $\mathbf{Z}_i \in \mathbb{R}^{n \times d}$ is the learned features of the i-th view, d is dimension of the learned features, $\alpha > 0$ is a trade-off parameter and $\mathbf{1} \in \mathbb{R}^n$ is a vector full of 1. With the first least square loss function, the transformation matrix $\mathbf{P}_i \in \mathbb{R}^{d_i \times d}$ is used to map the i-th view into a subspace $\mathbf{Z}_i$ and the $\mathbf{b}_i \in \mathbb{R}^d$ is the bias term. By concatenating all the projected subspaces, the learned features $[\mathbf{Z}_1 \cdots \mathbf{Z}_k] \in \mathbb{R}^{n \times kd}$ are obtained. The second least square loss function is used to measure the loss incurred by $\mathbf{W} \in \mathbb{R}^{kd \times c}$ on the learned features and the $\mathbf{b} \in \mathbb{R}^c$ is the bias term for the learned features. We choose least square loss function for its good performance and simplicity.

To step further, we assume that different views share a subspace feature component. By denoting the shared component as $\mathbf{Z} \in \mathbb{R}^{n \times d_s}$ where d_s is dimension of the shared component, the learned feature for the i-th view becomes $[\mathbf{Z}\ \mathbf{Z}_i] \in \mathbb{R}^{n \times (d+d_s)}$ and the learned features arrive at

$[\mathbf{Z}\ \mathbf{Z}_1 \cdots \mathbf{Z}_k] \in \mathbb{R}^{n\times(d_s+kd)}$. In this case, the dimensions of projection matrixes and bias term are changed, namely $\mathbf{P}_i \in \mathbb{R}^{d_i\times(d+d_s)}$, $\mathbf{W} \in \mathbb{R}^{(kd+d_s)\times c}$ and $\mathbf{b}_i \in \mathbb{R}^{(d+d_s)}$. Hence, our objective function becomes:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{P}_i,\mathbf{b}_i,\mathbf{b},\mathbf{Z}_i} &\sum_{i=1}^{k} \|\mathbf{X}_i^T\mathbf{P}_i + \mathbf{1}\mathbf{b}_i^T - [\mathbf{Z}\ \mathbf{Z}_i]\|_F^2 \\ &+\alpha\|[\mathbf{Z}\ \mathbf{Z}_1 \cdots \mathbf{Z}_k]\mathbf{W} + \mathbf{1}\mathbf{b}^T - \mathbf{Y}\|_F^2\end{aligned} \tag{2}$$

Note that if there is no shared component among different modalities of features, we set shared component $\mathbf{Z}$ to an empty matrix. The first term of Eq. (2) is a loss function for projecting the original feature space into a subspace while the second of term consider the loss between the prediction results with the ground-truth results.

Optimization Algorithm

The difficulty of solving the objective function in Eq.(2) lies in the concatenation of learned intermediate representations. By setting the derivative of Eq. (2) w.r.t $\mathbf{b}, \mathbf{b}_i$ to zero, we have

$$\mathbf{b}_i = \frac{1}{n}[\mathbf{Z}\ \mathbf{Z}_i]^T\mathbf{1} - \frac{1}{n}\mathbf{P}_i^T\mathbf{X}_i\mathbf{1} \tag{3}$$

$$\mathbf{b} = \frac{1}{n}\mathbf{Y}^T\mathbf{1} - \frac{1}{n}\mathbf{W}^T[\mathbf{Z}\ \mathbf{Z}_1 \cdots \mathbf{Z}_k]^T\mathbf{1} \tag{4}$$

Substituting $\mathbf{b}_i$ and $\mathbf{b}$ in Eq. (3) and (4), the original problem in Eq. (2) becomes:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{P}_i,\mathbf{Z}_i,\mathbf{Z}} &\sum_{i=1}^{k} \|(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{X}_i^T\mathbf{P}_i - (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)[\mathbf{Z}\ \mathbf{Z}_i]\|_F^2 \\ &+ \alpha\|(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)[\mathbf{Z}\ \mathbf{Z}_1 \cdots \mathbf{Z}_k]\mathbf{W} - (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{Y}\|_F^2\end{aligned} \tag{5}$$

where $\mathbf{I} \in \mathbb{R}^{n\times n}$ is an identity matrix. Denoting the $\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ with $\mathbf{H} \in \mathbb{R}^{n\times n}$, the problem in Eq. (5) arrives:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{P}_i,\mathbf{Z}_i,\mathbf{Z}} &\sum_{i=1}^{k} \|\mathbf{H}\mathbf{X}_i^T\mathbf{P}_i - \mathbf{H}[\mathbf{Z}\ \mathbf{Z}_i]\|_F^2 \\ &+ \alpha\|\mathbf{H}[\mathbf{Z}\ \mathbf{Z}_1 \cdots \mathbf{Z}_k]\mathbf{W} - \mathbf{H}\mathbf{Y}\|_F^2\end{aligned} \tag{6}$$

We replace the variables $\mathbf{HZ}$, $\mathbf{HZ}_i$ in Eq. (6) with $\mathbf{Z}$, $\mathbf{Z}_i$ respectively, then the problem becomes

$$\begin{aligned}\min_{\mathbf{W},\mathbf{P}_i,\mathbf{Z}_i,\mathbf{Z}} &\sum_{i=1}^{k} \|\mathbf{H}\mathbf{X}_i^T\mathbf{P}_i - [\mathbf{Z}\ \mathbf{Z}_i]\|_F^2 \\ &+ \alpha\|[\mathbf{Z}\ \mathbf{Z}_1 \cdots \mathbf{Z}_k]\mathbf{W} - \mathbf{H}\mathbf{Y}\|_F^2\end{aligned} \tag{7}$$

Taking derivative of Eq. (7) w.r.t. $\mathbf{P}_i$ and $\mathbf{W}$ respectively, we have

$$\mathbf{P}_i = (\mathbf{X}_i\mathbf{H}\mathbf{X}_i^T)^{-1}\mathbf{X}_i\mathbf{H}[\mathbf{Z}\ \mathbf{Z}_i] \tag{8}$$

and

$$\mathbf{W} = ([\mathbf{Z}\ \mathbf{Z}_1 \cdots \mathbf{Z}_k]^T\mathbf{H}[\mathbf{Z}\ \mathbf{Z}_1 \cdots \mathbf{Z}_k])^{-1}[\mathbf{Z}\ \mathbf{Z}_1 \cdots \mathbf{Z}_k]^T\mathbf{H}\mathbf{Y} \tag{9}$$

After substituting $\mathbf{P}_i$ and $\mathbf{W}$, we set the derivative w.r.t $[\mathbf{Z}\ \mathbf{Z}_1 \cdots \mathbf{Z}_k]$, we obtain:

$$\begin{aligned}&[\mathbf{Z}\ \mathbf{Z}_1 \cdots \mathbf{Z}_k]\begin{bmatrix} k\mathbf{I} & & & \\ & \mathbf{I} & & \\ & & \cdots & \\ & & & \mathbf{I}\end{bmatrix} \\ &- \begin{bmatrix}\sum_{i=1}^{k}\mathbf{H}\mathbf{X}_i^T\mathbf{P}_{i1} & \mathbf{H}\mathbf{X}_i^T\mathbf{P}_{12} \cdots \mathbf{H}\mathbf{X}_i^T\mathbf{P}_{k2}\end{bmatrix} \\ &+\alpha\,[\mathbf{Z}\ \mathbf{Z}_1 \cdots \mathbf{Z}_k]\,\mathbf{W}\mathbf{W}^T - \alpha\mathbf{H}\mathbf{Y}\mathbf{W}^T = 0\end{aligned} \tag{10}$$

$\Rightarrow$

$$\begin{aligned}&[\mathbf{Z}\ \mathbf{Z}_1 \cdots \mathbf{Z}_k] = \\ &\left(\begin{bmatrix}\sum_{i=1}^{k}\mathbf{H}\mathbf{X}_i^T\mathbf{P}_{i1} & \mathbf{H}\mathbf{X}_i^T\mathbf{P}_{12} \cdots \mathbf{H}\mathbf{X}_i^T\mathbf{P}_{k2}\end{bmatrix} + \alpha\mathbf{H}\mathbf{Y}\mathbf{W}^T\right) \\ &\left(\begin{bmatrix} k\mathbf{I} & & & \\ & \mathbf{I} & & \\ & & \cdots & \\ & & & \mathbf{I}\end{bmatrix} + \alpha\mathbf{W}\mathbf{W}^T\right)^{-1}\end{aligned} \tag{11}$$

where $\mathbf{P}_i = [\mathbf{P}_{i1}\ \mathbf{P}_{i2}]$, $\mathbf{P}_{i1} \in \mathbb{R}^{d_i\times d_s}$ is the projection matrix of the i-th view for shared component $\mathbf{Z}$ and $\mathbf{P}_{i2} \in \mathbb{R}^{d_i\times d}$ is the projection matrix of the i-th view for $\mathbf{Z}_i$

Based on the above mathematical deduction, an iterative algorithm is proposed to optimize the objective function in Eq. (2), which is summarized in Algorithm 1.

Algorithm 1 Multi-view Correlated feature Learning

Input: data $\mathbf{X}_i|_{i=1}^k \in \mathbb{R}^{d_i\times n}$, label $\mathbf{Y} \in \mathbb{R}^{n\times c}$, penalty parameter scalar α.

1: Initialize $\mathbf{Z}_i$ with PCA on $\mathbf{X}_i\mathbf{X}_i^T$;
2: Initialize $\mathbf{Z}$ with PCA on $\sum_{i=1}^{k}\mathbf{X}_i\mathbf{X}_i^T$;
3: Initialize $\mathbf{W} \in \mathbb{R}^{d\times c}$;
4: **Repeat:**
5: Update $\mathbf{P}_i$ and $\mathbf{W}$ according to Eq. (8) and Eq. (9);
6: Update $[\mathbf{Z}\ \mathbf{Z}_1 \cdots \mathbf{Z}_k]$ using Eq. (11);
7: **until Convergence**
8: Update $\mathbf{b}_i$ by $\mathbf{b}_i = \frac{1}{n}[\mathbf{Z}\ \mathbf{Z}_i]^T\mathbf{1} - \frac{1}{n}\mathbf{P}_i^T\mathbf{X}_i\mathbf{1}$
9: Update $\mathbf{b}$ by $\mathbf{b} = \frac{1}{n}\mathbf{Y}^T\mathbf{1} - \frac{1}{n}\mathbf{W}^T[\mathbf{Z}\ \mathbf{Z}_1 \cdots \mathbf{Z}_k]^T\mathbf{1}$
10: Return $\mathbf{W}$, $\mathbf{P}_i$,$\mathbf{b}$ and $\mathbf{b}_i$ for $1 \le i \le k$.

Convergence Analysis

In this section, we prove that Algorithm 1 converges by the following theorem. It begins with the following lemma.

Lemma 1. By fixing $\mathbf{P}_i|_{i=1}^k$ and $\mathbf{W}$, the global solutions for $\mathbf{Z}_i|_{i=1}^k$. Similarly, we can get the global solutions for $\mathbf{P}_i|_{i=1}^k$ and $\mathbf{W}$ with fixed $\mathbf{Z}_i|_{i=1}^k$.

Proof: By fixing $\mathbf{Z}_i|_{i=1}^k$, the objective function can be converted to a convex optimization problem wrt $\mathbf{P}_i|_{i=1}^k$ and $\mathbf{W}$. Hence, the global solutions for $\mathbf{P}_i|_{i=1}^k$ and $\mathbf{W}$ can be obtained by setting the derivative of Eq. (7) to zero respectively. In the same manner, we can also prove that by fixing $\mathbf{P}_i|_{i=1}^k$ and $\mathbf{W}$, we can get the global solutions for $\mathbf{Z}_i|_{i=1}^k$.

Theorem 1. The objective function value shown in Eq. (2) monotonically decreases until converged by applying the proposed algorithm.

Proof: Suppose after the r-th iteration, we get $\mathbf{P}_i^r|_{i=1}^k$, $\mathbf{b}_i^r|_{i=1}^k$, $\mathbf{W}^r$, $\mathbf{b}^r$, $\mathbf{Z}^r$ and $\mathbf{Z}_i^r$. In the next iteration, we fix $\mathbf{Z}$ as $\mathbf{Z}^r$, $\mathbf{Z}_i$ as $\mathbf{Z}_i^r$ and solve for $\mathbf{P}_i$ and $\mathbf{W}$. We can get the following inequality according Lemma 1:

$$\begin{aligned}
&\sum_{i=1}^{k} \|\mathbf{X}_i^T\mathbf{P}_i^{r+1} + \mathbf{1}(\mathbf{b}_i^{r+1})^T - [\mathbf{Z}^r\ \mathbf{Z}_i^r]\|_F^2 \\
&+ \alpha\|[\mathbf{Z}^r\ \mathbf{Z}_1^r \cdots \mathbf{Z}_k^r]\mathbf{W}^{r+1} + \mathbf{1}(\mathbf{b}^{r+1})^T - \mathbf{Y}\|_F^2 \\
\leq &\sum_{i=1}^{k} \|\mathbf{X}_i^T\mathbf{P}_i^{r} + \mathbf{1}(\mathbf{b}_i^{r})^T - [\mathbf{Z}^r\ \mathbf{Z}_i^r]\|_F^2 \\
&+ \alpha\|[\mathbf{Z}^r\ \mathbf{Z}_1^r \cdots \mathbf{Z}_k^r]\mathbf{W}^{r} + \mathbf{1}(\mathbf{b}^{r})^T - \mathbf{Y}\|_F^2
\end{aligned} \tag{12}$$

In the same manner, when fixing $\mathbf{W}$ and $\mathbf{P}_i$, the following inequality holds:

$$\begin{aligned}
&\sum_{i=1}^{k} \|\mathbf{X}_i^T\mathbf{P}_i^{r} + \mathbf{1}(\mathbf{b}_i^{r})^T - [\mathbf{Z}^{r+1}\ \mathbf{Z}_i^{r+1}]\|_F^2 \\
&+ \alpha\|[\mathbf{Z}^{r+1}\ \mathbf{Z}_1^{r+1} \cdots \mathbf{Z}_k^{r+1}]\mathbf{W}^{r} + \mathbf{1}(\mathbf{b}^{r})^T - \mathbf{Y}\|_F^2 \\
\leq &\sum_{i=1}^{k} \|\mathbf{X}_i^T\mathbf{P}_i^{r} + \mathbf{1}(\mathbf{b}_i^{r})^T - [\mathbf{Z}^r\ \mathbf{Z}_i^r]\|_F^2 \\
&+ \alpha\|[\mathbf{Z}^r\ \mathbf{Z}_1^r \cdots \mathbf{Z}_k^r]\mathbf{W}^{r} + \mathbf{1}(\mathbf{b}^{r})^T - \mathbf{Y}\|_F^2
\end{aligned} \tag{13}$$

By integrating Eq. (12) and Eq. (13), we can obtain:

$$\begin{aligned}
&\sum_{i=1}^{k} \|\mathbf{X}_i^T\mathbf{P}_i^{r+1} + \mathbf{1}(\mathbf{b}_i^{r+1})^T - [\mathbf{Z}^{r+1}\ \mathbf{Z}_i^{r+1}]\|_F^2 \\
&+ \alpha\|[\mathbf{Z}^{r+1}\ \mathbf{Z}_1^{r+1} \cdots \mathbf{Z}_k^{r+1}]\mathbf{W}^{r+1} + \mathbf{1}(\mathbf{b}^{r+1})^T - \mathbf{Y}\|_F^2 \\
\leq &\sum_{i=1}^{k} \|\mathbf{X}_i^T\mathbf{P}_i^{r} + \mathbf{1}(\mathbf{b}_i^{r})^T - [\mathbf{Z}^r\ \mathbf{Z}_i^r]\|_F^2 \\
&+ \alpha\|[\mathbf{Z}^r\ \mathbf{Z}_1^r \cdots \mathbf{Z}_k^r]\mathbf{W}^{r} + \mathbf{1}(\mathbf{b}^{r})^T - \mathbf{Y}\|_F^2
\end{aligned} \tag{14}$$

From Eq. (14), we can see that the objective function value decreases after each iteration. Thus, Theorem 1 has been proved.

Experiment

In this section, systematical experiments have been conducted to evaluate the performance of the proposed MVCS. We first compare our algorithm with other related methods, followed by the study on shared component evaluation. Additional experiments are conducted on the convergence of Algorithm 1.

Dataset Description

Our experiments carried out on the following four datasets that are broadly used in multi-view studies.

- NUS-WIDE-OBJECT: NUS-WIDE-OBJECT dataset (Chua et al. 2009) is used to compare different multi-view algorithms in terms of object categorization. This dataset consists of 30,000 real-world object images, falling into 30 object categories. In this experiment, we use the official split: 17,927 training images and 12,073 testing images.
- OUTDOOR SCENE: The outdoor scene dataset (Monadjemi, Thomas, and Mirmehdi 2002) contains 2,688 color images that belong to 8 outdoor scene categories: coast, mountain, forest, open country, street, inside city, tall buildings and highways.
- MSRC-V1: This dataset is a scene recognition data set consisting of 240 images and 8 classes in total. Following the setting in (Grauman and Darrell 2006), we select 7 classes and each class has 30 images. All the classes include tree, building, airplane, cow, face, car and bicycle.
- Handwritten Digit: Handwritten Digit dataset contains 0 to 9 ten digit classes and 2,000 data points in total. Five public available features are used in our experiment.

The feature descriptors of the datasets used in the experiments are described in Table1.

Experiment Setup

We compare the multi-view classification performance of the proposed algorithm with their corresponding single-view counterpart and the concatenation of all types of features. We apply SVM on each individual type of features and the concatenation of all types of features of the experimental datasets as baseline. In addition, we also compare the results of our proposed algorithm with several well-known multiple kernel learning (MKL) methods that are able to make use of multiple types of data, including: (1) SVM l_∞ MKL (Kloft et al. 2011), (2) SVM l_1 MKL method (Kloft et al. 2011), (3) SVM l_2 MKL method (Kloft et al. 2008), (4) least square (LSSVM) l_∞ MKL method (Ye, Ji, and Chen 2008), (5) LSSVM l_1 MKL method (Suykens, Van Gestel, and De Brabanter 2002), (6) LSSVM l_2 MKL method (Yu et al. 2010). Besides, we compare with another two multi-view classification methods, including LPboost-β (Gehler and Nowozin 2009) and LPboost-B (Gehler and Nowozin 2009) that have demonstrated state-of-the-art classification performance. Furthermore, other multi-view correlated algorithms that are compared in our experiments, including Multi-view CCA (Rupnik and Shawe-Taylor 2010) and Multirelational classification (Guo and Viktor 2008), take correlations among different views into consideration.

In all the experiments, we apply standard 5-fold cross-validation and report the average results with standard deviation. For the last three datasets, they are randomly split into equally sized training and test sets. The parameter of our method (α in Eq.(2)) is optimized in the range of $\{10^{-6}, 10^{-4}, ..., 10^4, 10^6\}$. For SVM method and MKL methods, one Gaussian kernel is constructed for each for each type of features (i.e., $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = exp\big(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2\big)$), where the parameter γ is the fine tuned in the same range used as our method. We implement the compared MKL methods using the codes published by (Yu et al. 2010; Kloft et al. 2011). Following the setting in

Table 1: Details of Multi-view datasets in the experiment

Feature Type ID	NUS-WIDE-OBJECT	Outdoor Scene	MSRC-V1	Handwritten Digit
1	Color Histogram(64-D)	GIST(512-D)	Color Moment(48-D)	FOU(76-D)
2	Color Correlogram(144-D)	Color Moment(432-D)	LBP(256-D)	FAC(216-D)
3	Ege Direction Histogram(73-D)	HOG(256-D)	HOG(100-D)	KAR(64-D)
4	Wavelet Texture(128-D)	LBP(48-D)	SIFT(1230-D)	PIX(240-D)
5	Block-Wise Color Moments(225-D)	-	GIST(512-D)	ZER(47-D)
6	BoW SIFT(500-D)	-	CENTRIST(1320-D)	-
Classes	31	8	7	10
Dataset Size	30000	2688	210	2000

Table 2: Classification results of the compared methods in terms of mAP (mean and std)

Methods	NUS-WIDE-OBJECT	Scene	MSRC-v1	Handwritten Digit
SVM (Type 1)	0.161±0.016	0.830±0.018	0.786±0.026	0.964±0.023
SVM (Type 2)	0.152±0.018	0.743±0.015	0.774±0.022	0.764±0.021
SVM (Type 3)	0.144±0.020	0.665±0.017	0.794±0.021	0.923±0.018
SVM (Type 4)	0.153±0.019	0.581±0.019	0.798±0.019	0.958±0.023
SVM (Type 5)	0.142±0.021	-	0.781±0.018	0.798±0.026
SVM (Type 6)	0.147±0.017	-	0.799±0.025	-
SVM (All)	0.187±0.021	0.846±0.014	0.802±0.018	0.969±0.022
SVM l_∞ MKL	0.223±0.019	0.852±0.021	0.829±0.021	0.975±0.018
SVM l_1 MKL	0.215±0.026	0.848±0.017	0.824±0.018	0.968±0.023
SVM l_2 MKL	0.212±0.024	0.847±0.018	0.801±0.022	0.966±0.022
LSSVM l_∞ MKL	0.211±0.020	0.835±0.021	0.795±0.024	0.969±0.020
LSSVM l_1 MKL	0.198±0.021	0.837±0.019	0.812±0.026	0.971±0.019
LSSVM l_2 MKL	0.192±0.022	0.840±0.014	0.819±0.019	0.967±0.022
GP method	0.190±0.019	0.835±0.018	0.826±0.017	0.969±0.025
LPboost-β	0.229±0.017	0.859±0.021	0.818±0.022	0.972±0.017
LPboost-B	0.227±0.014	0.861±0.023	0.810±0.022	0.970±0.014
Multi-view CCA	0.236±0.025	0.874±0.028	0.832±0.021	0.975±0.023
Multirelational Classification	0.268±0.022	0.894±0.024	0.865±0.013	0.987±0.012
MVCS (no shared)	0.297±0.011	0. 911±0.017	0.918±0.12	0.983±0.003
MVCS (shared)	**0.309±0.008**	**0.929±0.013**	**0.928±0.013**	**0.991±0.002**

(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 as α. For LPBboost-β and LPboost-B methods, we use the code published by the author [1]. LIBSVM [2] software package is used to implement SVM in all our experiments. As for our method, PCA is used to remove the null space of the dataset. And the sum of the dimensionalities of the $\mathbf{Z}$ and $\mathbf{Z}_i$, namely $d + d_s$, is equal to the number of the class c and we optimize the d_s in the range of $\{0, 1, 2, 3\}$. The performance is evaluated by mean Average Precision (mAP).

Results and Analysis

The performance of the compared methods on the four datasets are reported in Table 2. Both of mean and standard deviation of mAP are presented. The results show that MVCS outperforms all other compared methods, which demonstrates the effectiveness of our method for supervised classification problems.

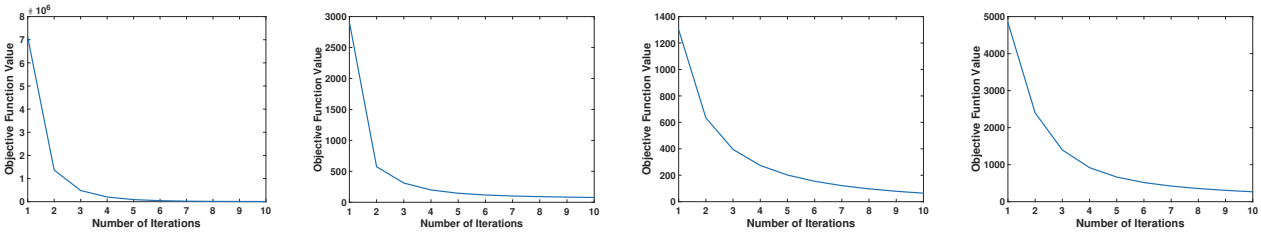
Firstly, from the results in Table 2, we can see that the methods using multiple data sources are always superior to SVM using one single type of features. For example, compared with SVM (Type 2), MVCS achieves an improvement of 15.7% for NUS dataset and 18.6% for Scene dataset. This confirms the usefulness of data integration from different views that contributes to the performance improvement.

In addition, compared with the MKL methods and boosting-enhanced MKL methods, our method achieves significant improvements in terms of mAP. In particular, we obtain 11.7% improvement in comparison with LSSVM l_2 MKL. Although the MKL methods take advantage of the information from different views, they fail to consider the correlations among different views as MVCS.

In Multi-view CCA and Multirelational Classification,

[1] http://files.is.tue.mpg.de/pgehler/projects/iccv09/

[2] https://www.csie.ntu.edu.tw/ cjlin/libsvm/



(a) NUS-OBJECT-WIDE (b) Scene (c) MSRCV1 (d) Digit

Figure 1: Convergence curves of the objective function value in Eq.(2) using Algorithm 1. From this figure, we can observe that the proposed algorithm monotonically decrease the objective function value until convergence.

they both exploit the correlations among different views. The results show these two methods outperform MKL methods, which demonstrates that considering correlations further facilitates classification performance. However, our method always performs better than these two methods since we not only consider the correlations but also utilize the individual information of each view. In contrast, Multi-view CCA and Multirelational Classification only address the former while not being able to take into account the latter.

To evaluate the benefit of mining the shared component among different views, we present the results of MVCS (no shared) without shared component by setting $\mathbf{Z}$ to an empty matrix. The MVCS with shared component always performs better than MVCS without shared component. This observation indicates that mining the shared component among different views is beneficial. Another interesting finding is that the result of Multirelational Classification method in Handwritten Digit dataset performs a bit better than MVCS without shared component, which indicates that the shared component for the Digit dataset makes limited contributions for improving the final results.

Finally, we show the behavior of the objective values by increasing the iteration number in Fig 1. From the figure, we can see that only a few iteration steps are needed to reach the convergence, which is very efficient.

Conclusion

In this paper, we have proposed a new multi-view learning algorithm called MVCS that efficiently and effectively finds the correlation between views and an intermediate representation of each view for the subsequent classification tasks. MVCS assumes that a latent subspace exists and is shared by all different views to some extent. Exploitation of the shared information and individual information of each view are conducted simultaneously. At the same time, a classifier is trained using the exploited information in a batch mode. In this way, a common component that is shared by all different views can be captured. As the objective function is non-smooth and difficult to solve, we propose an efficient iterative algorithm with guaranteed convergence. Intensive experiments on four benchmark datasets show that MVCS performs better than traditional single-view algorithms as well as some well-known multi-view learning counterparts for classification tasks.

Acknowledgments

The work is supported by National Key Research and Development Plan under Grant No. 2016YFB1001203.

References

Blum, A., and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *COLT*, 92–100. ACM.

Brefeld, U., and Scheffer, T. 2004. Co-em support vector learning. In *ICML*, 16. ACM.

Chang, X., and Yang, Y. 2016. Semi-supervised feature analysis by mining correlations among multiple tasks. *IEEE TNNLS* PP(99):1–12.

Chang, X.; Nie, F.; Wang, S.; Yang, Y.; Zhou, X.; and Zhang, C. 2015. Compound rank-k projections for bilinear analysis. *IEEE TNNLS* 27(7):1502–1513.

Chang, X.; Ma, Z.; Yang, Y.; Zeng, Z.; and Hauptmann, A. G. 2016a. Bi-level semantic representation analysis for multimedia event detection. *IEEE Transactions on Cybernetics* PP(99):1–18.

Chang, X.; Yu, Y.-L.; Yang, Y.; and Xing, E. P. 2016b. Semantic pooling for complex event analysis in untrimmed videos. *IEEE TPAMI* PP(99).

Chen, N.; Zhu, J.; and Xing, E. P. 2010. Predictive subspace learning for multi-view data: a large margin approach. In *NIPS*, 361–369.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. Nus-wide: a real-world web image database from national university of singapore. In *ACM CIVR*, 48. ACM.

Cortes, C.; Mohri, M.; and Rostamizadeh, A. 2009. Learning non-linear combinations of kernels. In *NIPS*, 396–404.

Diethe, T.; Hardoon, D. R.; and Shawe-Taylor, J. 2008. Multiview fisher discriminant analysis. In *NIPS workshop on learning from multiple sources*.

Gehler, P., and Nowozin, S. 2009. On feature combination for multiclass object classification. In *IEEE CVPR*, 221–228. IEEE.

Gönen, M., and Alpaydin, E. 2008. Localized multiple kernel learning. In *ICML*, 352–359. ACM.

Gong, Y.; Ke, Q.; Isard, M.; and Lazebnik, S. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV* 106(2):210–233.

Grauman, K., and Darrell, T. 2006. Unsupervised learning of categories from sets of partially matching image features. In *IEEE CVPR*, volume 1, 19–25. IEEE.

Guo, H., and Viktor, H. L. 2008. Multirelational classification: a multiple view approach. *Knowl. Inf. Syst.* 17(3):287–312.

Kloft, M.; Brefeld, U.; Laskov, P.; and Sonnenburg, S. 2008. Non-sparse multiple kernel learning. In *NIPS Workshop on Kernel Learning: Automatic Selection of Optimal Kernels*, volume 4.

Kloft, M.; Brefeld, U.; Sonnenburg, S.; and Zien, A. 2011. Lp-norm multiple kernel learning. *JMLR* 12(Mar):953–997.

Liu, M.; Luo, Y.; Tao, D.; Xu, C.; and Wen, Y. 2015a. Low-rank multi-view learning in matrix completion for multi-label image classification. In *AAAI*, 2778–2784.

Liu, S.; Liu, S.; Cai, W.; Che, H.; Pujol, S.; Kikinis, R.; Feng, D.; Fulham, M. J.; et al. 2015b. Multimodal neuroimaging feature learning for multiclass diagnosis of alzheimer's disease. *IEEE TBME* 62(4):1132–1140.

Meng, L.; Tan, A.-H.; and Xu, D. 2014. Semi-supervised heterogeneous fusion for multimedia data co-clustering. *IEEE TKDE* 26(9):2293–2306.

Monadjemi, A.; Thomas, B.; and Mirmehdi, M. 2002. Experiments on high resolution images towards outdoor scene classification. Technical report, Tech. Rep., University of Bristol, Department of Computer Science.

Muslea, I.; Minton, S.; and Knoblock, C. A. 2002. Active+ semi-supervised learning= robust multi-view learning. In *ICML*, volume 2, 435–442.

Nie, L.; Zhang, L.; Yang, Y.; Wang, M.; Hong, R.; and Chua, T.-S. 2015. Beyond doctors: Future health prediction from multimedia and multimodal observations. In *ACM MM*, 591–600. ACM.

Nie, F.; Li, J.; and Li, X. 2016. Parameter-free auto-weighted multiple graph learning: A framework for multi-view clustering and semi-supervised classification.

Nigam, K., and Ghani, R. 2000. Analyzing the effectiveness and applicability of co-training. In *CIKM*, 86–93. ACM.

Parker, B. S., and Khan, L. 2015. Detecting and tracking concept class drift and emergence in non-stationary fast data streams. In *AAAI*, 2908–2913.

Peng, X.; Wang, L.; Wang, X.; and Qiao, Y. 2016. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *CVIU* 150:109–125.

Rupnik, J., and Shawe-Taylor, J. 2010. Multi-view canonical correlation analysis. In *Conference on Data Mining and Data Warehouses (SiKDD 2010)*, 1–4.

Sigal, L.; Memisevic, R.; and Fleet, D. J. 2009. Shared kernel information embedding for discriminative inference. In *IEEE CVPR*, 2852–2859. IEEE.

Sun, S. 2013. A survey of multi-view machine learning. *Neural Comput. Appl.* 23(7-8):2031–2038.

Suykens, J. A.; Van Gestel, T.; and De Brabanter, J. 2002. *Least Squares Support Vector Machines*. World Scientific.

Varma, M., and Babu, B. R. 2009. More generality in efficient multiple kernel learning. In *ICML*, 1065–1072. ACM.

Wang, H.; Nie, F.; Huang, H.; and Ding, C. 2013. Heterogeneous visual features fusion via sparse multimodal machine. In *IEEE CVPR*, 3097–3102.

Wang, S.; Ma, Z.; Yang, Y.; Li, X.; Pang, C.; and Hauptmann, A. G. 2014. Semi-supervised multiple feature analysis for action recognition. *IEEE TMM* 16(2):289–298.

Wang, S.; Chang, X.; Li, X.; Long, G.; Yao, L.; and Sheng, Q. 2016. Diagnosis code assignment using sparsity-based disease correlation embedding. *IEEE TKDE* 28(12):3191–3202.

Xu, Z.; Jin, R.; King, I.; and Lyu, M. 2009. An extended level method for efficient multiple kernel learning. In *NIPS*, 1825–1832.

Xu, C.; Tao, D.; and Xu, C. 2013. A survey on multi-view learning. *Neural Comput. Appl.* 23(7–8):2031–2038.

Yang, Y.; Ma, Z.; Hauptmann, A. G.; and Sebe, N. 2013a. Feature selection for multimedia analysis by sharing information among multiple tasks. *IEEE TMM* 15(3):661–669.

Yang, Y.; Song, J.; Huang, Z.; Ma, Z.; Sebe, N.; and Hauptmann, A. G. 2013b. Multi-feature fusion via hierarchical regression for multimedia analysis. *IEEE TMM* 15(3):572–581.

Ye, J.; Ji, S.; and Chen, J. 2008. Multi-class discriminant kernel learning via convex programming. *JMLR* 9(Apr):719–758.

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

Yu, S.; Krishnapuram, B.; Rosales, R.; and Rao, R. B. 2011. Bayesian co-training. *JMLR* 12(Sep):2649–2680.

Zhao, F.; Qiao, L.; Shi, F.; Yap, P.-T.; and Shen, D. 2016. Feature fusion via hierarchical supervised local cca for diagnosis of autism spectrum disorder. *Brain Imaging and Behavior* 1–11.

Zheng, S.; Cai, X.; Ding, C. H.; Nie, F.; and Huang, H. 2015. A closed form solution to multi-view low-rank regression. In *AAAI*, 1973–1979.

Zhu, X.; Li, X.; Zhang, S.; Ju, C.; and Wu, X. 2016. Robust joint graph sparse coding for unsupervised spectral feature selection. *IEEE TNNLS* pp(99):1–13.

Zhu, X.; Li, X.; and Zhang, S. 2016. Block-row sparse multiview multilabel learning for image classification. *IEEE transactions on cybernetics* 46(2):450–461.