Semi-Supervised Multi-View Correlation Feature Learning with Application to Webpage Classification

Xiao-Yuan Jing,[1,2,*] Fei Wu,[2,*] Xiwei Dong,[2] Shiguang Shan,[3] Songcan Chen[4]

[1] State Key Laboratory of Software Engineering, School of Computer, Wuhan University, China

[2] College of Automation, Nanjing University of Posts and Telecommunications, China

[3] Key Lab of Intelligent Information Process of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, China

[4] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China

[*] Corresponding authors: {jingxy_2000@, wufei_8888@}126.com

Abstract

Webpage classification has attracted a lot of research interest. Webpage data is often multi-view and high-dimensional, and the webpage classification application is usually semi-supervised. Due to these characteristics, using semi-supervised multi-view feature learning (SMFL) technique to deal with the webpage classification problem has recently received much attention. However, there still exists room for improvement for this kind of feature learning technique. How to effectively utilize the correlation information among multi-view of webpage data is an important research topic. Correlation analysis on multi-view data can facilitate extraction of the complementary information. In this paper, we propose a novel SMFL approach, named semi-supervised multi-view correlation feature learning (SMCFL), for webpage classification. SMCFL seeks for a discriminant common space by learning a multi-view shared transformation in a semi-supervised manner. In the discriminant space, the correlation between intra-class samples is maximized, and the correlation between inter-class samples and the global correlation among both labeled and unlabeled samples are minimized simultaneously. We transform the matrix-variable based nonconvex objective function of SMCFL into a convex quadratic programming problem with one real variable, and can achieve a global optimal solution. Experiments on widely used datasets demonstrate the effectiveness and efficiency of the proposed approach.

Introduction

The last several years have witnessed a rapid increase of information available on the World Wide Web, making it difficult to find web pages that contain the information in which one is interested. In order to make an effective and efficient search, we need to classify webpages. Webpage classification refers to the problem of assigning a webpage a class that describes its contents (Qi and Davison 2009; Gollapalli et al. 2013). As stated in (Jing et al. 2015), webpage classification has three characteristics: (1) Webpage is a kind of multi-view data (Wang and Zhou 2008; Zhang et al. 2013; Xu et al. 2013; Kan et al. 2016), since it usually contains two or more types of data, e.g., text, hyperlinks and images, where **each type of data can be regarded as a view**. These multiple views describe the same webpage. (2) Webpage classification is a semi-supervised application (Zhou et al. 2007; Zhu and Goldberg 2009), since labeled pages are harder to collect compared to unlabeled pages in practice. (3) Webpage data is high-dimensional, since webpages usually contain much information. Considering these three characteristics, it is crucial to design effective semi-supervised multi-view feature learning (SMFL) methods for webpage classification. To our knowledge, two webpage classification methods taking these three characteristics into account have been developed, namely semi-paired and semi-supervised generalized correlation analysis (SSGCA) (Chen et al. 2012) and uncorrelated semi-supervised intra-view and inter-view manifold discriminant (USI^{2}MD) (Jing et al. 2015).

For other applications, a few SMFL methods have also been presented, such as multi-view metric learning with global consistency and local smoothness (MVML-GL) (Zhai et al. 2012), vector-valued reproducing kernel Hilbert spaces (VRKHS) (Minh et al. 2013), multi-view hypergraph learning (MHL) (Hong et al. 2013), semi-supervised multi-view canonical correlation analysis based on label propagation (LPbSMCCA) (Shen and Sun 2014), and manifold-regularized semi-supervised kernel canonical correlation analysis (MR-skCCA) (Volpi et al. 2014).

Motivation and Contribution

The correlation information from inter-view and intra-view

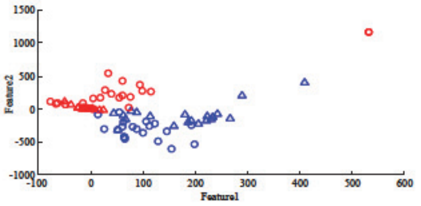


Figure 1. Sample distribution of 40 webpage samples in the WebKB dataset, where red and blue colors denote two views, and markers ○ and △ denote sample points from two classes.

depicts the association relationship among multiple views, which has close connection with classification (Li et al. 2009; Jing et al. 2014). Here, "inter-view" and "intra-view" mean the relationship between samples across different views and within a certain view, respectively. The inter-view correlation contains the within-class and between-class correlation of samples across different views. And the intra-view correlation contains the within-class and between-class correlation of samples within the same view.

Taking the WebKB dataset (Chen et al. 2012) as an example, we randomly select 10 webpage samples of each class in the link and page views, and perform the principal component analysis (PCA) (Turk and Pentland 1991) transformation to obtain two major principal components of each sample for plotting the sample distribution in Fig. 1. From the figure, samples of different views own small correlation, and between-class samples within each view own relatively large correlation. We thus should maximize the correlation of samples from different views while in the same class, and minimize the correlation of samples in the same view while from different classes, to make the distribution more favorable for classification. Therefore, it is necessary to perform correlation analysis on webpage data, which can help us learn features with favorable separability.

Current SMFL methods have following drawbacks:

(1) Most of existing SMFL methods (except SSGCA, LPbSMCCA and MR-skCCA) **do not consider the correlation information**.

(2) For SSGCA, LPbSMCCA and MR-skCCA, there **exists much room to improve their discriminant abilities**. Specifically, SSGCA does not explore the intra-view correlation information. It maximizes the cross-view correlation without differently treating the within-class and between-class correlation. In addition, SSGCA can only be applied to the problems involving two views. LPbSMCCA does not explore the intra-view correlation information either, and it only maximizes the inter-view within-class correlation without considering the inter-view between-class correlation. MR-skCCA simply maximizes the correlation from both intra-view and inter-view, which cannot make full use of the discriminant correlation from the aspect of classification.

Therefore, how to effectively explore the intra-view and inter-view discriminant correlation information to improve the webpage classification performance is an important research topic. We summarize the contributions of our study as following three points:

(1) We propose a novel SMFL approach for webpage classification, which is named semi-supervised multi-view correlation feature learning (SMCFL). The objective function of SMCFL is designed to **maximize the correlation between intra-class samples, and minimize the correlation between inter-class samples and the global correlation among both labeled and unlabeled samples**. SMCFL **can thus effectively explore the intra-view and inter-view discriminant correlation information**.

(2) We transform the matrix-variable based nonconvex objective function of SMCFL into a convex quadratic programming problem with one real variable. **The solution is global optimal and can be derived analytically without iterative calculation**. To our knowledge, **we are the first to present this kind of solution, which can be applied to other correlation-based feature learning problems**.

(3) SMCFL is verified on two widely used webpage datasets. The experimental results show that it can significantly outperform state-of-the-art webpage classification methods.

Related Work

Semi-supervised Multi-view Feature Learning

In recent years, a few SMFL methods have been developed (Xu et al. 2016). MVML-GL (Zhai et al. 2012) jointly considers the global consistency and local smoothness, and formulates metric learning as a convex optimization problem for solution. VRKHS (Minh et al. 2013) gives a general formulation for the problem of learning an unknown functional dependency between a structured input space and a structured output space under the semi-supervised setting. MHL (Hong et al. 2013) uses hypergraph to represent the relationships among samples and applies the constructed hypergraph to semi-supervised dimensionality reduction of multi-view data. LPbSMCCA (Shen and Sun 2014) firstly estimates the labels for the unlabeled samples using sparse representation based label propagation strategy, and then seeks projection directions by considering the inter-view correlation and intra-view compactness of samples from the same class. MR-skCCA (Volpi et al. 2014) exploits both the labels and unlabeled pixels into the computation of cross-correlation, and employs multiset kernel canonical correlation analysis (CCA) algorithm to learn transformation for hyperspectral image classification.

These SMFL methods are not applied to the webpage classification task and cannot fully explore the intra-view and inter-view discriminant correlation information.

Webpage Classification

In the last decade, several webpage classification methods have been presented (Du et al. 2013; Wu et al. 2014; Wang et al. 2015). (Zhang et al. 2008) designs a multi-view local model for each example and presents a multi-view local learning regularization matrix method. (Kim et al. 2009) presents a semi-supervised learning method, which leverages click logs to augment training data by propagating class labels to unlabeled similar documents. (Wang et al. 2011) addresses a nonnegative matrix tri-factorization based dual knowledge transfer algorithm for cross-language webpage classification. Two-view transductive SVM (TTSVM) (Li et al. 2012) uses sufficient unlabeled data and their multiple representations to improve classification performance. (Bing et al. 2014) performs webpage segmentation with structured prediction for webpage classification.

Most of current webpage classification methods do not consider all three characteristics of webpage classification.

SMFL for Webpage Classification

To our knowledge, only two SMFL webpage classification methods have been addressed. SSGCA (Chen et al. 2012) makes as maximal correlation as possible on paired data by performing CCA, with preserving the global structural information of unlabeled data and the local discriminative information of labeled data. USI^{2}MD (Jing et al. 2015) combines the semi-supervised intra-view and inter-view manifold discriminant schema with semi-supervised uncorrelation constraint for webpage classification.

The differences between SSGCA, USI^{2}MD and our approach are: USI^{2}MD does not consider the correlation among multiple views and SSGCA does not explore the intra-view correlation information, while our SMCFL can effectively utilize the intra-view and inter-view discriminant correlation information.

Semi-supervised Multi-view Correlation Feature Learning (SMCFL)

The Objective Function of SMCFL

Suppose that $X^l = \{X_1, X_2, \cdots, X_C\}$ is the labeled training webpage sample set from C classes, where each $X_i\,(i = 1, \cdots, C)$ contains webpage samples of M views and $x_{ip}^s \in \mathbb{R}^{d\times 1}$ denotes the p^{th} webpage sample from the s^{th} view of the i^{th} class. Here, d denotes the dimensionality of samples. Assume that l_i^s denotes the number of samples from the s^{th} view and the i^{th} class, and $l_i = \sum_{s=1}^{M} l_i^s$ denotes the number of samples in the i^{th} class. Let X^u be the unlabeled training sample set, $X = \{X^l, X^u\}$, and N denote the total sample number in X. For simplicity of representation, we regard X^u as the $(C+1)^{th}$ class.

We aim to learn a discriminant projection transformation W that can project samples from M views to one discriminant common space, where the correlation between intra-class samples is maximized, while the correlation between inter-class samples and the global correlation among both labeled and unlabeled samples are minimized simultaneously. We define the within-class correlation S_w, between-class correlation S_b and total correlation S_t as follows:

$$S_w = \frac{1}{C}\sum_{i=1}^{C}\left[\frac{1}{l_i^2}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_i^t}\frac{x_{ip}^{sT}WW^Tx_{iq}^t}{\sqrt{x_{ip}^{sT}WW^Tx_{ip}^s}\sqrt{x_{iq}^{tT}WW^Tx_{iq}^t}}\right], \tag{1}$$

$$S_b = \frac{2}{C(C-1)}\sum_{i=1}^{C-1}\sum_{j=i+1}^{C}\left[\frac{1}{l_il_j}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_j^t}\frac{x_{ip}^{sT}WW^Tx_{jq}^t}{\sqrt{x_{ip}^{sT}WW^Tx_{ip}^s}\sqrt{x_{jq}^{tT}WW^Tx_{jq}^t}}\right], \tag{2}$$

$$S_t = \frac{1}{(C+1)C}\left(\begin{array}{l}\displaystyle\sum_{i=1}^{C}\sum_{j=i+1}^{C+1}\left[\frac{1}{l_il_j}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_j^t}\frac{x_{ip}^{sT}WW^Tx_{jq}^t}{\sqrt{x_{ip}^{sT}WW^Tx_{ip}^s}\sqrt{x_{jq}^{tT}WW^Tx_{jq}^t}}\right]\\ \displaystyle+\frac{C}{2}\sum_{i=1}^{C+1}\left[\frac{1}{l_i^2}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_i^t}\frac{x_{ip}^{sT}WW^Tx_{iq}^t}{\sqrt{x_{ip}^{sT}WW^Tx_{ip}^s}\sqrt{x_{iq}^{tT}WW^Tx_{iq}^t}}\right]\end{array}\right). \tag{3}$$

Then, the objective function of SMCFL can be defined as

$$\max_W \; f(W) = S_w - r_1S_b - r_2S_t, \tag{4}$$

where $r_1 > 0$ and $r_2 > 0$ are weight coefficients. We set $H = WW^T$ and $\|x_{ip}^s\| = 1\ (\forall i, s, p)$, where $\|\cdot\|$ denotes the l_2-norm of a vector. Obviously, H should be symmetric and positive semi-definite, i.e., $H = H^T$ and $H \geq 0$. We relax (4) into the following formulation:

$$\begin{array}{l}\displaystyle\max_H f(H) = S_w' - r_1S_b' - r_2S_t'\\ s\,t.\quad H = H^T,\ H \geq 0\end{array}, \tag{5}$$

where S_w', S_b', and S_t' are separately defined as follows:

$$\begin{aligned}S_w' &= \frac{1}{C}\sum_{i=1}^{C}\left[\frac{1}{l_i^2}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_i^t}\frac{x_{ip}^{sT}WW^Tx_{iq}^t}{\|x_{ip}^s\|\|x_{iq}^t\|\|WW^T\|_F}\right],\\ &= \frac{1}{C}\sum_{i=1}^{C}\left[\frac{1}{l_i^2}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_i^t}\frac{x_{ip}^{sT}Hx_{iq}^t}{\|H\|_F}\right]\end{aligned} \tag{6}$$

$$\begin{aligned}S_b' &= \frac{1}{C(C-1)}\sum_{i=1}^{C}\sum_{j=1,j\neq i}^{C}\left[\frac{1}{l_il_j}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_j^t}\frac{x_{ip}^{sT}WW^Tx_{jq}^t}{\|x_{ip}^s\|\|x_{jq}^t\|\|WW^T\|_F}\right],\\ &= \frac{1}{C(C-1)}\sum_{i=1}^{C}\sum_{j=1,j\neq i}^{C}\left[\frac{1}{l_il_j}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_j^t}\frac{x_{ip}^{sT}Hx_{jq}^t}{\|H\|_F}\right]\end{aligned} \tag{7}$$

$$S_t' = \frac{1}{2(C+1)C}\left(\begin{array}{l}\displaystyle\sum_{i=1}^{C+1}\sum_{j=i,j\neq i}^{C+1}\left[\frac{1}{l_il_j}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_j^t}\frac{x_{ip}^{sT}Hx_{jq}^t}{\|H\|_F}\right]\\ \displaystyle+C\sum_{i=1}^{C+1}\left[\frac{1}{l_i^2}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_i^t}\frac{x_{ip}^{sT}Hx_{iq}^t}{\|H\|_F}\right]\end{array}\right). \tag{8}$$

It is noted that for the transformation from (1) to (6), a similar transformation trick can be found in (Szedmak et al. 2007). As a result, (5) can be further translated into

$$\begin{aligned}&\min_H \ \|H\|_F \\ &s.t.\ \ B \ge 1,\ H = H^T,\ H \ge 0\end{aligned} , \qquad (9)$$

where

$$\begin{aligned}B = {} & \frac{1}{C}\sum_{i=1}^{C}\left[\frac{1}{l_i^2}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_i^t} x_{ip}^{sT} H x_{iq}^{t}\right] \\ & -\frac{r_1}{C(C-1)}\sum_{i=1}^{C}\sum_{j=1,j\ne i}^{C}\left[\frac{1}{l_i l_j}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_j^t} x_{ip}^{sT} H x_{jq}^{t}\right] \\ & -\frac{r_2}{2(C+1)C}\left(\begin{aligned}&\sum_{i=1}^{C+1}\sum_{j=1,j\ne i}^{C+1}\left[\frac{1}{l_i l_j}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_j^t} x_{ip}^{sT} H x_{jq}^{t}\right] \\ &+C\sum_{i=1}^{C+1}\left[\frac{1}{l_i^2}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_i^t} x_{ip}^{sT} H x_{iq}^{t}\right]\end{aligned}\right)\end{aligned} .$$

Solution of SMCFL

To achieve the solution of Formula (9), we design the following optimization scheme, which can obtain an analytical and global optimal solution.

We provisionally leave the constraints $H = H^T$ and $H \ge 0$ off (Note that we will show that H is symmetric in the following derivation, namely in Eqs. (15) and (17); and we make H positive semi-definite in Eq. (22) or (23)). Thus, we can simplify (9) as

$$\min_H \ \|H\|_F \ \ s.t.\ \ B \ge 1 , \qquad (10)$$

which can be expressed as the following **convex quadratic programming problem**

$$\min_H \ \frac{1}{2}\|H\|_F^2 \ \ s.t.\ \ B \ge 1 . \qquad (11)$$

To make the solution of (11) robust, we introduce the slack variable $\varepsilon \ge 0$ to relax the corresponding constraint. With such a relaxation, (11) is reformulated as

$$\begin{aligned}&\min_H \ \frac{1}{2}\|H\|_F^2 + \eta\varepsilon \\ &s.t.\ \ B \ge 1-\varepsilon, \varepsilon \ge 0\end{aligned} , \qquad (12)$$

where η is a regularization parameter.

By applying the Lagrangian technique to constrained optimization problem, we define the Lagrange function as

$$L(H,\varepsilon,\alpha,\mu) = \frac{1}{2}\|H\|_F^2 + \eta\varepsilon - \alpha(B-1+\varepsilon) - \mu\varepsilon , \qquad (13)$$

where α and μ are Lagrangian multipliers. By making the derivatives with respect to H and ε equal to zeros, we can obtain

$$\frac{\partial L}{\partial H} = H - \alpha R = 0 \ \text{ and } \ \frac{\partial L}{\partial \varepsilon} = \eta - \alpha - \mu = 0 , \qquad (14)$$

where

$$\begin{aligned}R = {} & \frac{1}{C}\sum_{i=1}^{C}\left[\frac{1}{l_i^2}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_i^t} x_{ip}^{s} x_{iq}^{tT}\right] \\ & -\frac{r_1}{C(C-1)}\sum_{i=1}^{C}\sum_{j=1,j\ne i}^{C}\left[\frac{1}{l_i l_j}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_j^t} x_{ip}^{s} x_{jq}^{tT}\right] \\ & -\frac{r_2}{2(C+1)C}\left(\begin{aligned}&\sum_{i=1}^{C+1}\sum_{j=1,j\ne i}^{C+1}\left[\frac{1}{l_i l_j}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_j^t} x_{ip}^{s} x_{jq}^{tT}\right] \\ &+C\sum_{i=1}^{C+1}\left[\frac{1}{l_i^2}\sum_{s=1}^{M}\sum_{t=1}^{M}\sum_{p=1}^{l_i^s}\sum_{q=1}^{l_i^t} x_{ip}^{s} x_{iq}^{tT}\right]\end{aligned}\right)\end{aligned} . \qquad (15)$$

The corresponding Karush-Kuhn-Tucher (KKT) conditions (Chen et al. 2011) are

$$\alpha\left(B-1+\varepsilon\right) = 0 ,\ \ \mu \ge 0 \ \text{ and } \ \alpha \ge 0 . \qquad (16)$$

According to (14), we obtain

$$H = \alpha R . \qquad (17)$$

It can be easily proved that R is a symmetric matrix, and thus $H = \alpha R$ is symmetric. Then, we can get

$$B = tr\left(HR^T\right) , \qquad (18)$$

where $tr(\cdot)$ denotes the trace of a square matrix.

Substituting Eqs. (14), (17) and (18) into (13), we obtain its Wolfe dual objective as follows

$$DL(H,\varepsilon,\alpha,\mu) = -\frac{\alpha^2}{2}tr\left(RR^T\right) + \alpha . \qquad (19)$$

Hence, to get the solution of (12) is equivalent to solving the following optimization problem:

$$\max_\alpha \ \alpha - \frac{A}{2}\alpha^2 , \qquad (20)$$

where $A = tr\left(RR^T\right)$ is positive. (20) is a convex quadratic programming problem. If $\eta \ge 1/A$, the solution of (20) is $\alpha^* = 1/A$; otherwise, the solution is $\alpha^* = \eta$.

Finally, substituting α into (17), we can get H . To obtain the projective transformation matrix W , H is eigen-decomposed as

$$H = U\Lambda U^T , \qquad (21)$$

where Λ is a diagonal eigenvalue matrix of H , and U is an orthogonal matrix whose columns correspond to the eigenvectors of H .

If H is positive semi-definite, we can obtain W by

$$W = U\sqrt{\Lambda} . \qquad (22)$$

When H is not positive semi-definite, namely some eigenvalues of H are negative, like the solution trick in (Ma et al. 2007), we select the positive eigenvalues and corresponding eigenvectors to construct a new diagonal matrix Λ_+ ($\Lambda_+ = \sqrt{\Lambda_+}\left(\sqrt{\Lambda_+}\right)^T$) and a new orthogonal matrix U_+, respectively. Then we can obtain W by

$$W = U_+\sqrt{\Lambda_+} . \qquad (23)$$

Let $y_1, y_2, \cdots, y_M$ be M views of a given query sample and $\hat{X}_1,\hat{X}_2,\cdots,\hat{X}_M$ be M views of labeled training samples, where each $\hat{X}_s\left(s=1,\cdots,M\right)$ contains C classes. With the obtained transformation matrix W , we achieve the projected features of training sample set and query sample separately by $Z_s^X=W^T\hat{X}_s$ and $Z_s^y=W^Ty_s$ for each view. Then, we use the following strategy to fuse these features:

$$Z^X=\left[Z_1^{XT},Z_2^{XT},\cdots,Z_M^{XT}\right]^T \text{ and } Z^y=\left[Z_1^{yT},Z_2^{yT},\cdots,Z_M^{yT}\right]^T . \quad (24)$$

Finally, we use the nearest neighbor classifier with the cosine distance to classify Z^y.

Algorithm 1 summarizes the proposed SMCFL approach.

```
Algorithm 1. SMCFL
Input:   Training sample sets $X^l$ and $X^u$, test sample $y$.
Output:  Class label of $y$.
Step 1.  Calculate $\alpha$ according to (20).
Step 2.  Calculate $H$ according to (17).
Step 3.  Calculate $W$ according to (22) or (23).
Step 4.  Obtain the projected test sample $Z^y$ and the projected labeled training sample set $Z^X$.
Step 5.  Use the nearest neighbor classifier with the cosine distance to classify $Z^y$ according to $Z^X$.
```

Time Complexity Analysis

The main computational burden of SMCFL consists of the matrix calculation for R and the eigen-decomposition problem in (21). Thus, the time complexity of SMCFL is $O\left(d^3+N^2d\right)$. As reported in (Jing et al. 2015; Volpi et al. 2014), the time complexities of representative SMFL methods including MVML-GL, VRKHS, MR-skCCA, SSGCA and USI^{2}MD are $O\left(N_l^3+\left(Md\right)^2N_l\right)$, $O\left(\left(MN\right)^3+M^2N^2d\right)$, $O\left(N^3+d^2N\right)$, $O\left(\left(Md\right)^3+MN^2d\right)$ and $O\left(\left(Md\right)^3+M^2N^2d\right)$, respectively. Here, N_l represents the number of labeled training samples.

It is obvious that the time complexity of our SMCFL approach is smaller than those of SSGCA and USI^{2}MD. Whether the time complexity of our approach is lower than those of MVML-GL, VRKHS or MR-skCCA is mainly determined by the values of N_l , N and d .

Experiments

Data Set

In this paper, we evaluate our approach on two widely used datasets, namely WebKB (Chen et al. 2012) and Internet advertisements (Kushmerick 1999).

The WebKB dataset contains 1051 webpages from two classes (230 pages in the course class and 821 pages in the non-course class). Each webpage is characterized by the page view and the link view. We use a preprocessed version of this dataset, where 3000-dimensional and 1840-dimensional original features are extracted from the page view and link view of a webpage, respectively.

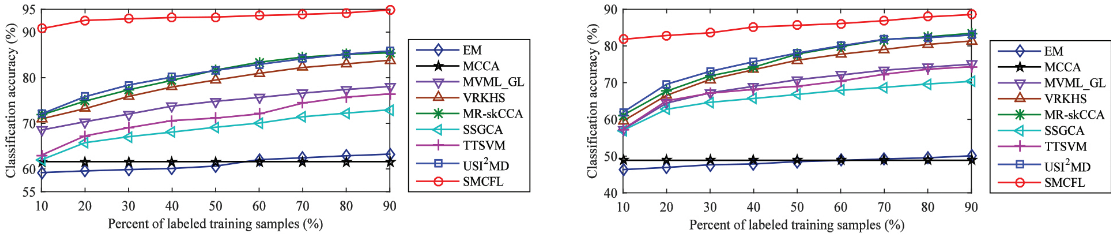
The Internet advertisements (AD for short) dataset includes 3279 samples containing 458 advertisements and 2821 non-advertisement samples. Like in (Jing et al. 2015), we utilize a preprocessed version of this dataset, where each sample is regarded as a binary vector with quite large sparsity and consists of 1558 original features. Three views that have similar numbers of features are used for experiment, including 495 base URL features, 472 destination URL features and 457 image URL features.

Compared Methods and Experimental Settings

In experiment, we compare our SMCFL with six state-of-the-art related methods, including three SMFL methods, i.e., **MVML-GL** (Zhai et al. 2012), **VRKHS** (Minh et al. 2013) and **MR-skCCA** (Volpi et al. 2014); and three webpage classification methods, i.e., **SSGCA** (Chen et al. 2012), **TTSVM** (Li et al. 2012) and **USI^{2}MD** (Jing et al. 2015). We also compare SMCFL with a representative unsupervised multi-view feature learning method, i.e., multi-view canonical correlation analysis (**MCCA**) (Li et al. 2009). Expectation Maximization (EM) based algorithm is an important kind of the semi-supervised learning method (Nigam et al. 2000; McLachlan and Krishnan 2007; Saluja et al. 2012; Zhao et al. 2016). Here, we also compare our SMCFL with **EM** (Nigam et al. 2006). And the view with the best classification accuracy (Jing et al. 2015) is used for EM.

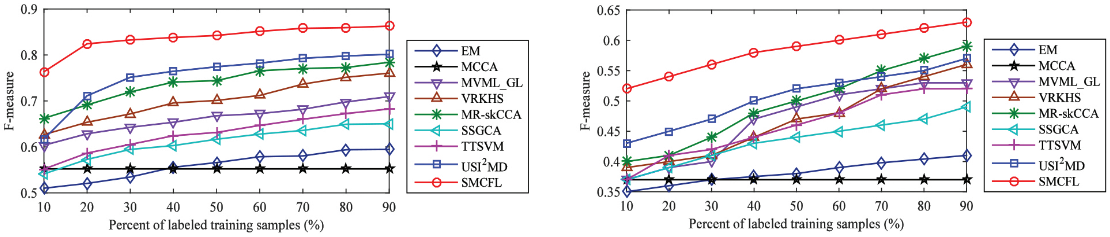
For each dataset, we randomly select 50% samples per class to construct the training set and use the remaining samples for testing. We further randomly select a certain percentage of the training samples as the labeled samples and regard the remaining training samples as unlabeled ones for semi-supervised learning. For the unsupervised method MCCA, we use all training samples (both labeled and unlabeled samples) for its training. For SSGCA and TTSVM, we utilize two views with the best classification accuracy for them in the AD dataset, since these two methods only apply to two views based learning problem.

Like in (Jing et al. 2015), we employ the PCA method (Turk and Pentland 1991) to reduce the samples' dimensionalities of different views to 1050 for WebKB and 456 for AD. For our SMCFL, the parameters r_1 and r_2 in (5) and η in (12) are determined by using 5-fold cross validation technique on training data. Concretely, we set $r_1=r_2=0.01$ and $\eta=1$ for WebKB, and $r_1=r_2=0.1$ and $\eta=1$ for AD. The evaluation measures including classification accuracy (CA) (Jing et al. 2015) and F-measure (Banerjee and Pedersen 2003) are used to evaluate the



(a) On WebKB dataset (b) On AD dataset

Figure 2. Average classification accuracies of compared methods on two datasets.



(a) On WebKB dataset (b) On AD dataset

Figure 3. Average F-measure values of compared methods on two datasets.

Table 1. Average performances of all compared methods on two datasets.

Dataset	Measure	EM	MCCA	MVML-GL	VRKHS	MR-skCCA	SSGCA	TTSVM	USI^{2}MD	SMCFL
WebKB	CA	61.11	61.62	74.14	78.63	80.41	68.75	71.09	80.80	**93.27**
	F-measure	0.56	0.55	0.66	0.70	0.74	0.61	0.63	0.76	**0.84**
AD	CA	48.33	48.88	69.28	73.93	75.58	65.96	68.62	76.25	**85.40**
	F-measure	0.38	0.37	0.47	0.47	0.50	0.43	0.46	0.51	**0.58**

Table 2. P-values between SMCFL and other methods on measures of classification accuracy and F-measure.

Dataset	Measure	EM	MCCA	MVML-GL	VRKHS	MR-skCCA	SSGCA	TTSVM	USI^{2}MD
WebKB	CA	2.74×10^{-13}	5.86×10^{-13}	4.81×10^{-9}	1.27×10^{-6}	7.38×10^{-6}	1.13×10^{-9}	2.64×10^{-8}	4.55×10^{-6}
	F-measure	4.55×10^{-10}	2.83×10^{-9}	3.45×10^{-10}	9.05×10^{-8}	1.13×10^{-7}	2.08×10^{-11}	5.98×10^{-10}	2.71×10^{-5}
AD	CA	7.29×10^{-12}	4.07×10^{-11}	8.59×10^{-7}	1.30×10^{-4}	5.79×10^{-4}	3.26×10^{-9}	2.40×10^{-7}	4.46×10^{-4}
	F-measure	3.10×10^{-10}	1.29×10^{-7}	2.33×10^{-6}	2.08×10^{-6}	4.17×10^{-5}	1.08×10^{-14}	3.20×10^{-8}	3.26×10^{-8}

Table 3. Average training time (seconds) of all compared methods on two datasets.

Dataset	EM	MCCA	MVML-GL	VRKHS	MR-skCCA	SSGCA	TTSVM	USI^{2}MD	SMCFL
WebKB	0.26	0.12	0.38	5.97	1.72	5.18	-	4.89	0.45
AD	4.50	4.12	4.87	154.72	21.45	13.83	-	29.86	5.58

classification performances.

Evaluation of Classification Results

We evaluate the classification results of our SMCFL approach with the percentage of labeled training samples increasing from 10% to 90%. Figs. 2 and 3 separately illustrate the average classification accuracies and F-measure values of all compared methods across 20 random running on two datasets. Table 1 reports the average results (across 9 percentages) corresponding to Figs. 2 and 3. From Fig. 2, SMCFL significantly outperforms all the other competing methods with regard to classification accuracy in all cases. On average, SMCFL improves the classification accuracy at least by **12.47%** on WebKB and **9.15%** on AD. According to Fig. 3, SMCFL always obtains better F-measure results than other methods. On average, SMCFL improves the F-measure value at least by **0.08** on WebKB and **0.07** on AD. This illustrates that SMCFL can well balance the classification effects of positive (minority) class and negative (majority) class. **The main reason for the obvious improvement lies in that SMCFL fully and effectively utilizes the intra-view and inter-view discriminant correlation information**.

To statistically analyze the classification results shown in Figs. 2 and 3, we conduct a statistical test, i.e., Mcnemar's test (Yambor et al. 2002), at the confidence level of 95%. If the p-value is below 0.05, the performance difference between two compared methods is statistically significant. Table 2 shows the p-values between SMCFL and other methods on two measures. According to the table, the proposed approach makes a statistically significant difference as compared with other methods. Besides the WebKB and AD datasets, we also evaluated our proposed approach on the practical and large-scale Open Directory Project (ODP) (http://www.dmoz.org) dataset. The experimental results indicated that our SMCFL outperforms the competing methods. Since using the experimental data downloaded in different time from dmoz.org may generate different classification results, in this paper, we only report the detailed experimental results on WebKB and AD.

Evaluation of Running Time

In Table 3, we take the semi-supervised learning case that 20 percent of labeled training samples are used as an example, and report the average training time of compared methods across 20 runs. It is noted that TTSVM itself is a classifier requiring no training. Since the testing time of all compared methods is close to each other, we do not report the testing time. Our hardware configuration comprises a 2.93-GHz CPU and a 8GB RAM. We can see that SMCFL costs less training time than semi-supervised multi-view feature learning methods VRKHS and MR-skCCA, and webpage classification methods SSGCA and USI^{2}MD.

Parameter Analysis

Our approach involves three parameters including r_1, r_2 and η. We found that changing η would not affect the results that much and we set it as 1. Here, we show that the performance of SMCFL is not sensitive when r_1 and r_2 are sampled in some value ranges. For these two parameters, we search the following parameter space $\{10^{-3},10^{-2},10^{-1},10^{0},10^{1},10^{2},10^{3}\}$. Fig. 4 illustrates the classification accuracies of our approach versus different values of r_1 and r_2 on WebKB with 20 percent of labeled training samples. We can see that the performances of our approach are stable with respect to the parameters in the range of $[10^{-3},10^{-1}]$. For simplicity, we set r_1 and r_2 as 0.01 on WebKB. A similar phenomenon also exists on AD dataset.

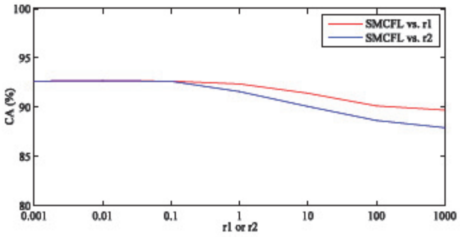


Figure 4: Classification accuracy versus r_1 or r_2 on WebKB.

Conclusion

In this paper, we propose a novel semi-supervised multi-view feature learning approach named SMCFL for webpage classification. It can effectively explore the intra-view and inter-view discriminant correlation information. The objective function is converted into a convex quadratic programming problem, and we can obtain a global optimal solution analytically.

Experimental results on two public webpage datasets demonstrate that SMCFL significantly outperforms several state-of-the-art SMFL and webpage classification methods with respect to classification accuracy and F-measure. The experimental results also indicate the efficiency of the training procedure of our approach.

Acknowledgments

The authors want to thank the editors and anonymous reviewers for their constructive comments and suggestions. The work described in this paper was supported by the National Nature Science Foundation of China under Project Nos. 61272273, 61671182, 61672281 and 61233011, and the Research Project of NJUPT (XJKY14016).

References

Banerjee, S.; and Pedersen, T. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, 805-810.

Bing, L.; Guo, R.; Lam, W.; Niu, Z. Y.; and Wang, H. 2014. Web page segmentation with structured prediction and its application in web page classification. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 767-776.

Chen, X. H.; Chen, S. C.; and Xue, H. 2011. Large correlation analysis. *Applied Mathematics and Computation* 217(22): 9041-9052.

Chen, X. H.; Chen, S. C.; Xue, H.; and Zhou, X. D. 2012. A unified dimensionality reduction framework for semi-paired and semi-supervised multi-view data. *Pattern Recognition* 45(5): 2005-2018.

Du, Y.; Su, C.; Cai, Z.; and Guan, X. 2013. Web page and image semi-supervised classification with heterogeneous information fusion. *Journal of Information Science* 39(3): 289-306.

Gollapalli, S. D.; Caragea, C.; Mitra, P.; and Giles, C. L. 2013. Researcher homepage classification using unlabeled data. In *Proceedings of the 22nd International Conference on World Wide Web*, 471-482.

Hong, C.; Yu, J.; Li, J.; and Chen, X. 2013. Multi-view hypergraph learning by patch alignment framework. *Neurocomputing* 118(11): 79-86.

Jing, X. Y.; Hu, R.; Zhu, Y. P.; Wu, S. S.; Liang, C.; and Yang, J. Y. 2014. Intra-view and inter-view supervised correlation analysis for multi-view feature learning. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 1882-1889.

Jing, X. Y.; Liu, Q.; Wu, F.; Xu, B.; Zhu, Y.; and Chen, S. 2015. Web page classification based on uncorrelated semi-supervised intra-view and inter-view manifold discriminant feature extraction. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 2255-2261.

Kan, M.; Shan, S.; Zhang, H.; Lao, S.; and Chen, X. 2016. Multi-view discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(1): 188-194.

Kim, S.; Pantel, P.; Duan, L.; and Gaffney, S. 2009. Improving web page classification by label-propagation over click graphs. In *Proceedings of the 18th International Conference on Information and Knowledge Management*, 1077-1086.

Kushmerick, N. 1999. Learning to remove internet advertisements. In *Proceedings of the 3rd Annual Conference on Autonomous Agents*, 175-181.

Li, G.; Chang, K.; and Hoi, S. C. 2012. Multiview semi-supervised learning with consensus. *IEEE Transactions on Knowledge and Data Engineering* 24(11): 2040-2051.

Li, Y. O.; Adali, T.; Wang, W.; and Calhoun, V. D. 2009. Joint blind source separation by multiset canonical correlation analysis. *IEEE Transactions on Signal Processing* 57(10): 3918-3929.

Nigam, K.; McCallum, A.; and Mitchell, T. 2006. Semi-supervised text classification using EM. *Semi-Supervised Learning*, 33-56.

Nigam, K.; McCallum, A. K.; Thrun, S.; and Mitchell, T. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning* 39(2-3): 103-134.

Ma, Y.; Lao, S.; Takikawa, E.; and Kawade, M. 2007. Discriminant analysis in correlation similarity measure space. In *Proceedings of the 24th International Conference on Machine learning*, 577-584.

McLachlan, G.; and Krishnan, T. 2007. The EM algorithm and extensions: second edition. *Wiley Blackwell*, 1-369.

Minh, H. Q.; Bazzani, L.; and Murino, V. 2013. A unifying framework for vector-valued manifold regularization and multi-view learning. In *Proceedings of the 30th International Conference on Machine Learning*, 100-108.

Qi, X.; and Davison, B. D. 2009. Web page classification: features and algorithms. *ACM Computing Surveys* 41(2): 75-79.

Saluja, A. S.; Sundararajan, P. K.; and Mengshoel, O. J. 2012. Age-layered expectation maximization for parameter learning in Bayesian networks. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, 984-992.

Shen X. B.; and Sun, Q. S. 2014. A novel semi-supervised canonical correlation analysis and extensions for multi-view dimensionality reduction. *Journal of Visual Communication and Image Representation* 25(8): 1894-1904.

Szedmak, S.; De Bie, T.; and Hardoon, D. R. 2007. A metamorphosis of canonical correlation analysis into multivariate maximum margin learning. In *Proceedings of the 11th European Symposium Artificial Neural Networks*, 211-216.

Turk, M.; and Pentland, A. 1991. Eigenfaces for recognition. *Journal of Cognitive Neuroscience* 3(1): 71-86.

Volpi, M.; Matasci, G.; Kanevski, M.; and Tuia, D. 2014. Semi-supervised multiview embedding for hyperspectral data classification. *Neurocomputing* 145(18): 427-437.

Wang, H.; Huang, H.; Nie, F.; and Ding, C. 2011. Cross-language web page classification via dual knowledge transfer using nonnegative matrix tri-factorization. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 933-942.

Wang, H.; Nie, F.; and Huang, H. 2015. Large-scale cross-language web page classification via dual knowledge transfer using fast nonnegative matrix tri-factorization. *ACM Transactions on Knowledge Discovery from Data* 10(1): 933-942.

Wang, W.; and Zhou, Z. H. 2008. On multi-view active learning and the combination with semi-supervised learning. In *Proceedings of the 25th International Conference on Machine learning*, 1152-1159.

Wu, O.; Hu, R.; Mao, X.; and Hu, W. 2014. Quality-based learning for web data classification. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 194-200.

Xu, X.; Li, W.; Xu, D.; and Tsang, I. W. 2016. Co-labeling for multi-view weakly labeled learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(6): 1113-1125.

Xu, C.; Tao, D.; and Xu, C. 2013. A survey on multi-view learning. CoRR abs/1304.5634.

Yambor, W.; Draper, B.; and Beveridge, R. 2002. Analyzing PCA-based face recognition algorithms: eigenvector selection and distance measures. In *Proceedings of the 2nd Workshop on Empirical Evaluation Methods in Computer Vision*, 1-15.

Zhai, D. M.; Chang, H.; Shan, S. G.; Chen, X. L.; and Gao, W. 2012. Multiview metric learning with global consistency and local smoothness. *ACM Transactions on Intelligent Systems and Technology* 3(3): 451-458.

Zhang, D.; Wang, F.; Zhang, C.; and Li, T. 2008. Multi-view local learning. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, 752-757.

Zhang, W.; Zhang, K.; Gu, P.; and Xue, X. 2013. Multi-view embedding learning for incompletely labeled data. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, 1910-1916.

Zhao, L.; Huang, M.; Yao, Z.; Su, R.; Jiang, Y.; and Zhu, X. 2016. Semi-supervised multinomial naive Bayes for text classification by leveraging word-level statistical constraint. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2877-2883.

Zhou, Z. H.; Zhan, D. C.; and Yang, Q. 2007. Semi-supervised learning with very few labeled training examples. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, 675-680.

Zhu, X.; and Goldberg, A. B. 2009. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 3(1): 1-130.