Web Page Classification Based on Uncorrelated Semi-Supervised Intra-View and Inter-View Manifold Discriminant Feature Extraction

Xiao-Yuan Jing[1,2,*], **Qian Liu**[1,2], **Fei Wu**[1,2], **Baowen Xu**[1], **Yangping Zhu**[1,2], **Songcan Chen**[3]
[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] College of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics, China
[*] jingxy_2000@126.com (Corresponding author)

Abstract

Web page classification has attracted increasing research interest. It is intrinsically a multi-view and semi-supervised application, since web pages usually contain two or more types of data, such as text, hyperlinks and images, and unlabeled pages are generally much more than labeled ones. Web page data is commonly high-dimensional. Thus, how to extract useful features from this kind of data in the multi-view semi-supervised scenario is important for web page classification. To our knowledge, only one method is specially presented for this topic. And with respect to a few semi-supervised multi-view feature extraction methods on other applications, there still exists much room for improvement. In this paper, we firstly design a feature extraction schema called semi-supervised intra-view and inter-view manifold discriminant (SI^{2}MD) learning, which sufficiently utilizes the intra-view and inter-view discriminant information of labeled samples and the local neighborhood structures of unlabeled samples. We then design a semi-supervised uncorrelation constraint for the SI^{2}MD schema to remove the multi-view correlation in the semi-supervised scenario. By combining the SI^{2}MD schema with the constraint, we propose an uncorrelated semi-supervised intra-view and inter-view manifold discriminant (USI^{2}MD) learning approach for web page classification. Experiments on public web page databases validate the proposed approach.

1 Introduction

With the rapid development of the Internet, the amount of web pages has been increasing very fast, and people benefit more and more from the information conveyed by these pages. In order to effectively utilize the information of web pages, it is necessary to classify them.

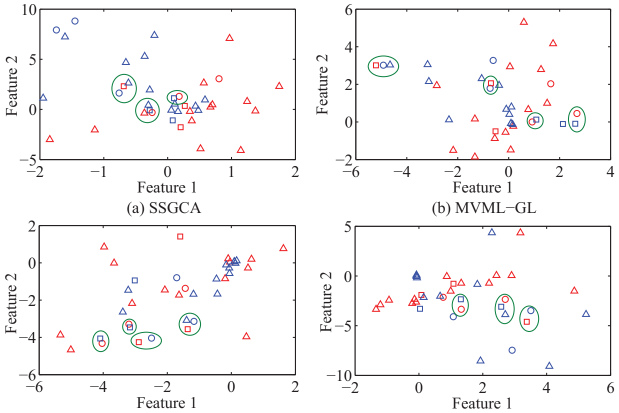
In real-world applications, web page classification has three characteristics: (1) **Web page is a kind of multi-view data** [Zhuang *et al.*, 2012; Li *et al.*, 2012a], since it usually contains two or more types of data, e.g., text, hyperlinks and images, where each type of data can be regarded as a view. (2) **Web page classification is a semi-supervised application** [Li *et al.*, 2012b; Chen *et al.*, 2012]. In the Internet, labeled pages are hard to be collected as compared with unlabeled pages, since the labeling operations are comparatively expensive and time consuming. (3) **Web page data is high-dimensional.** It usually contains much information, and how to extract low-dimensional and useful features from this data is crucial for classification tasks.

1.1 Motivation

To the best of our knowledge, only one web page classification method has taken these three characteristics into consideration, that is semi-paired and semi-supervised generalized correlation analysis (SSGCA) [Chen *et al.*, 2012]. Besides, there exist a few semi-supervised multi-view feature extraction methods designed for other applications, such as multiview 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] and multi-view hypergraph learning (MHL) [Hong *et al.*, 2013].

Existing semi-supervised multi-view feature extraction methods have the following major drawbacks:

(a) **They do not sufficiently utilize the favorable discriminant information within each view and between different views, especially the inter-view discriminant information.** Here, discriminant information refers to the information of class labels and class distributions, which is useful to separate the samples with different class labels. In Figure 1, we separately employ the SSGCA, MVML-GL, VRKHS and MHL methods to extract features from the samples of WebKB database [Chen *et al.*, 2012], and perform the principal component analysis (PCA) [Turk and Pentland, 1991] transform on the extracted features to obtain two major principal components that are used to show the sample distribution. Figure 1 shows that inter-view samples from two classes (in the green circles) tend to cluster together, leading to one of these samples, which actually belongs to one class, may be misclassified into another class, and thus is adverse to classification. This illustrates the necessity of utilizing inter-



(a) SSGCA (b) MVML−GL (c) VRKHS (d) MHL

Figure 1: Sample distribution of two views of 20 web page samples (6 labeled samples from two classes with 3 samples per class and 14 unlabeled samples) in the feature space of SSGCA, MVML-GL, VRKHS and MHL methods on WebKB database, where red and blue colors denote two views, markers ○ and □ denote sample points from two classes, marker △ denotes an unlabeled sample point, and Feature 1 and Feature 2 denote two major principal components of the features extracted by these four methods.

view discriminant information.

(b) **They do not reduce multi-view correlation.** As shown in Figure 1, both labeled and unlabeled web page samples from different views usually mingle together, such that there is much correlation between views. This means much information redundancy exists between views in semi-supervised scenario. The redundancy may interfere the performance of web page classification.

Hence, for the web page classification task, how to sufficiently utilize intra-view and inter-view discriminant information from web pages and effectively reduce the adverse multi-view correlation in semi-supervised scenario is an important research topic.

1.2 Contribution

We provide a novel solution for this topic. The contributions of our work are summarized as follows:

(1) By incorporating the manifold discriminant learning technique into the web page classification task, we propose a feature extraction schema called semi-supervised intra-view and inter-view manifold discriminant (SI^{2}MD) learning. It can fully utilize the intra-view and inter-view discriminant information of labeled samples and the local neighborhood structures of unlabeled samples to extract useful features from web page data. Here, local neighborhood refers to the area near a sample, which contains several nearest neighbor samples of this sample.

(2) We further design a semi-supervised uncorrelation constraint for the SI^{2}MD schema, which can effectively remove adverse multi-view correlation of the extracted features.

(3) By combining the SI^{2}MD schema with the semi-supervised uncorrelation constraint, we propose a novel feature extraction approach for web page classification, which is named uncorrelated semi-supervised intra-view and inter-view manifold discriminant (USI^{2}MD) learning.

The proposed USI^{2}MD approach is verified on the public web page databases, i.e. WebKB [Chen *et al.*, 2012] and Internet advertisements [Kushmerick, 1999] databases. Experimental results demonstrate that the proposed approach performs better than several related multi-view feature extraction and web page classification methods.

2 Related Work

2.1 Web Page Classification

In the last decade, some methods have been presented for web page classification. They can be divided into five categories: (1) Classifier design based semi-supervised multi-view learning method. For example, Muslea *et al.* [2006] introduced a multi-view active learning method called co-testing. Li *et al.* [2012a] presented a two-view transductive SVM (TTSVM) to take advantage of the unlabeled data and their multiple representations. Du *et al.* [2013] designed a multi-view semi-supervised learning method named local co-training (LCT). (2) Feature extraction based semi-supervised multi-view learning method. Chen *et al.* [2012] developed a dimensionality reduction method called semi-paired and semi-supervised generalized correlation analysis (SSGCA). (3) Semi-supervised methods without multi-view learning. For instance, Zhou and Li [2005] designed a semi-supervised co-training algorithm named tri-training. (4) Multi-view methods without semi-supervised learning. For example, Zhuang *et al.* [2012] designed a new generative model for multi-view learning via probabilistic latent semantic analysis. (5) Other methods. For example, Chen and Hsieh [2006] presented a weighted vote based support vector machine to classify web pages with synonymous keywords. Hu *et al.* [2007] presented a framework for recognizing pornographic web pages, where texts, images and content representations of pages are jointly considered. Chen *et al.* [2009] designed a fuzzy ranking analysis paradigm with the discriminating power measure to reduce the dimensionality of input data. Zhang *et al.* [2011] presented a framework using a Bayesian method for content-based phishing web page detection.

Among the above semi-supervised multi-view web classification methods, only SSGCA employs the feature extraction technique. The difference between SSGCA and our USI^{2}MD approach is that SSGCA neither utilizes the inter-view discriminant information nor reduces semi-supervised multi-view correlation, while USI^{2}MD does.

2.2 Unsupervised and Supervised Multi-view Feature Extraction

In recent years, multi-view feature extraction technique has attracted lots of research interest. In this field, multi-view subspace learning is an important research direction. Under this direction, canonical correlation analysis (CCA) [Hardoon *et al.*, 2004] based and discriminant analysis based multi-view subspace learning are two representative techniques. CCA based methods include multi-view CCA (MCCA) [Li *et al.*, 2009], multiset integrated CCA [Yuan *et al.*, 2011], multiple discriminant CCA [Gao *et al.*, 2012], multiple principal angle (MPA) [Su *et al.*, 2012], etc. Discriminant analysis based methods contain multi-view discriminant analysis (MvDA) [Kan *et al.*, 2012], generalized multi-view marginal fisher analysis (GMMFA) [Sharma *et al.*, 2012], multi-view discriminant transfer learning [Yang and Gao, 2013], view-invariant discriminative projection [Hu *et al.*, 2013], etc.

2.3 Semi-supervised Multi-view Feature Extraction

Multi-view learning in semi-supervised scenario has become an important research topic [Li *et al.*, 2012b; Liu and Tao, 2013; Richarz *et al.*, 2014]. To the best of our knowledge, four semi-supervised multi-view feature extraction methods have been addressed. SSGCA has been introduced in Section 2.1. Multiview metric learning with global consistency and local smoothness (MVML-GL) [Zhai *et al.*, 2012] reveals the shared latent feature space of the multi-view observations by embodying global consistency constraints and preserving local geometric structures. Vector-valued reproducing kernel Hilbert spaces (VRKHS) [Minh *et al.*, 2013] learns an unknown functional dependency between a structured input space and a structured output space in the semi-supervised setting. Multi-view hypergraph learning (MHL) [Hong *et al.*, 2013] constructs multi-view hypergraph Laplacian matrix and gets the dimensionality-reduced data by solving the standard eigen-decomposition to obtain the projection matrix.

The difference between these four semi-supervised multi-view feature extraction methods and our USI^{2}MD approach is that these four methods do not consider utilizing the inter-view discriminant information and reducing semi-supervised multi-view correlation, while USI^{2}MD does.

3 Our USI^{2}MD Approach

In this section, we introduce our USI^{2}MD approach. Specifically, we detail the SI^{2}MD schema, semi-supervised uncorrelation constraint and USI^{2}MD approach in the following three subsections. Note that based on the preprocessed high-dimensional web page data, we perform dimension reduction and feature extraction. How to obtain the preprocessed data is not the problem we discuss in this paper, and Refs. [Chen *et al.*, 2012] and [Kushmerick, 1999] address this problem.

3.1 SI^{2}MD Schema

Suppose that $X_1, X_2, \cdots, X_N$ denote the training web page sample sets of N views, where $X_i\,(i = 1, 2, \cdots, N)$ contains c classes (class j comprises n_i^j samples with class label L_j, and $X_1, X_2, \cdots, X_N$ contain the same c classes) and n_i^{c+1} unlabeled samples. All unlabeled samples in each view are regarded as the $(c+1)^{th}$ class, and L_{c+1} is used as the class label of unlabeled samples. In our approach, $X_i\,(i = 1, 2, \cdots, N)$ is mean-normalized (that is, the mean of all samples in X_i is a zero vector), and samples in $X_1, X_2, \cdots, X_N$ are all m-dimensional vectors. Let $n_i = \sum_{j=1}^{c+1} n_i^j$ be the total sample number of X_i, and $n = \sum_{i=1}^{N} n_i$ be the total sample number of all views.

Based on the graph theory that is widely used in the fields of pattern recognition and machine learning, where "graph" usually illustrates the relations among data samples [Wang and Chen, 2009], we use an intrinsic graph G to illustrate the intra-view and inter-view distance relations of web page samples within each class and each local neighborhood. And we use a penalty graph $\bar{G}$ to illustrate the intra-view and inter-view distance relations of web page samples from different classes and different local neighborhoods. The graphs G and $\bar{G}$ are defined as follows.

Definition 1 (Intrinsic graph G): for two training samples $x_s^p \in X_s$ and $x_t^q \in X_t$, whose class labels separately are L_j and L_k, (a) a solid edge is added between x_s^p and x_t^q if $j = k$ and $k \neq c+1$; (b) a dashed edge is added between x_s^p and x_t^q if $j = c+1$ or $k = c+1$, and x_s^p is among the K-nearest neighbors of x_t^q or vice versa. Here $K = N \times \max\left(n_1^1, n_1^2, \cdots, n_N^{c-1}, n_N^c\right)$.

Definition 2 (Penalty graph $\bar{G}$): for two training samples $x_s^p \in X_s$ and $x_t^q \in X_t$, whose class labels separately are L_j and L_k, (a) a solid edge is added between x_s^p and x_t^q if $j \neq k$, $j \neq c+1$ and $k \neq c+1$; (b) a dashed edge is added between x_s^p and x_t^q if $j = c+1$ or $k = c+1$, and x_s^p is outside the K-nearest neighbors of x_t^q or vice versa.

The affinity matrix of G, i.e. $W \in \mathbf{R}^{n\times n}$, whose element (w_{st}^{pq}) refers to the weight of edge between x_s^p and x_t^q, is defined as: if a solid edge exists between x_s^p and x_t^q, then $w_{st}^{pq} = 1$; if a dashed edge exists between x_s^p and x_t^q, then $w_{st}^{pq} = \exp\left(-\|x_s^p - x_t^q\|^2 \big/ r\right)$; otherwise $w_{st}^{pq} = 0$. Here, r is an adjustable parameter of the "heat kernel" [Wang and Chen, 2009]. The affinity matrix of $\bar{G}$, i.e. $\bar{W} \in \mathbf{R}^{n\times n}$, whose element $(\bar{w}_{st}^{pq})$ refers to the weight of edge between x_s^p and x_t^q, is defined in the same way as that of W.

We formulate the objective function of the SI^{2}MD schema as

$$\begin{array}{c}\max\limits_{v_1, v_2, \cdots, v_N} \sum\limits_{s=1}^{N}\sum\limits_{p=1}^{n_s}\sum\limits_{t=1}^{N}\sum\limits_{q=1}^{n_t} \left\|v_s^T x_s^p - v_t^T x_t^q\right\|_2^2 \bar{w}_{st}^{pq} \\ -\beta \sum\limits_{s=1}^{N}\sum\limits_{p=1}^{n_s}\sum\limits_{t=1}^{N}\sum\limits_{q=1}^{n_t} \left\|v_s^T x_s^p - v_t^T x_t^q\right\|_2^2 w_{st}^{pq}\end{array}, \qquad (1)$$

where β is a nonnegative coefficient, and $v_1, v_2, \cdots, v_N$ denote the projective vectors of $X_1, X_2, \cdots, X_N$, respectively. Formula (1) fully utilize the intra-view and inter-view discriminant information of labeled samples and the local neighborhood structures of unlabeled samples. Its first term can separate labeled samples of different classes and unlabeled samples of different neighborhoods, while its second term can congregate labeled samples within each class and unlabeled samples within each neighborhood. The second term also embodies the inter-view consistency, since when $p{=}q$ (the same web page) and $s{\neq}t$ (different views), samples x_s^p and x_t^q are two views of the same web page, a solid edge is added between them and this term minimizes the distance between these two samples in the projected space.

Suppose that $D \in \mathbf{R}^{n\times n}$ and $\bar{D} \in \mathbf{R}^{n\times n}$ are two diagonal matrixes whose elements are separately defined as $d_{ss}^{pp} = \sum_{t=1}^{N}\sum_{q=1}^{n_t} w_{st}^{pq}$ and $\bar{d}_{ss}^{pp} = \sum_{t=1}^{N}\sum_{q=1}^{n_t} \bar{w}_{st}^{pq}$. Let $L_w = D - W$ and $L_b = \bar{D} - \bar{W}$. We can divide L_w into N^2 small matrixes $L_{wst} \in \mathbf{R}^{n_s \times n_t}\,(s, t = 1, 2, \cdots, N)$ and divide L_b into N^2 small matrixes $L_{bst} \in \mathbf{R}^{n_s \times n_t}\,(s, t = 1, 2, \cdots, N)$, so that $L_w = \begin{bmatrix} L_{w11} & L_{w12} & \cdots & L_{w1N} \\ L_{w21} & L_{w22} & \cdots & L_{w2N} \\ \vdots & \vdots & \cdots & \vdots \\ L_{wN1} & L_{wN2} & \cdots & L_{wNN} \end{bmatrix}$ and $L_b = \begin{bmatrix} L_{b11} & L_{b12} & \cdots & L_{b1N} \\ L_{b21} & L_{b22} & \cdots & L_{b2N} \\ \vdots & \vdots & \cdots & \vdots \\ L_{bN1} & L_{bN2} & \cdots & L_{bNN} \end{bmatrix}$. Formula (1) can be rewritten as

$$\max_{v} v^T \left(Q - \beta P\right) v, \tag{2}$$

where $Q = \begin{bmatrix} X_1 L_{b11} X_1^T & X_1 L_{b12} X_2^T & \cdots & X_1 L_{b1N} X_N^T \\ X_2 L_{b21} X_1^T & X_2 L_{b22} X_2^T & \cdots & X_2 L_{b2N} X_N^T \\ \vdots & \vdots & \cdots & \vdots \\ X_N L_{bN1} X_1^T & X_N L_{bN2} X_2^T & \cdots & X_N L_{bNN} X_N^T \end{bmatrix}$, $P = \begin{bmatrix} X_1 L_{w11} X_1^T & X_1 L_{w12} X_2^T & \cdots & X_1 L_{w1N} X_N^T \\ X_2 L_{w21} X_1^T & X_2 L_{w22} X_2^T & \cdots & X_2 L_{w2N} X_N^T \\ \vdots & \vdots & \cdots & \vdots \\ X_N L_{wN1} X_1^T & X_N L_{wN2} X_2^T & \cdots & X_N L_{wNN} X_N^T \end{bmatrix}$ and $v = \left[v_1^T, v_2^T, \cdots, v_N^T\right]^T$.

3.2 Semi-supervised Uncorrelation Constraint

To reduce the adverse multi-view features correlation among web page samples, i.e. multi-view correlation of extracted features from samples not with the same class label, we design the semi-supervised uncorrelation constraint. It is noted that for any two samples, samples not with the same class label include three cases: (a) two samples with different class labels; (b) two unlabeled samples; (c) a labeled sample and an unlabeled sample.

We use an inter-view penalty correlation graph $\hat{G}$ to illustrate the multi-view correlation relations of web page samples not with the same class label. The graph $\hat{G}$ is defined as follows.

Definition 3 (Inter-view penalty correlation graph $\hat{G}$): for two training samples $x_s^p \in X_s$ and $x_t^q \in X_t$ $(s \neq t)$, whose class labels separately are L_j and L_k, (a) a solid edge is added between x_s^p and x_t^q if $j \neq k$, $j \neq c+1$ and $k \neq c+1$; (b) a solid edge is added between x_s^p and x_t^q if $j = c+1$ or $k = c+1$.

The affinity matrix of $\hat{G}$, i.e. $\hat{W} \in \mathbf{R}^{n \times n}$, whose element $(\hat{w}_{st}^{pq})$ refers to the weight of edge between x_s^p and x_t^q, is defined as: if a solid edge exists between x_s^p and x_t^q, then $\hat{w}_{st}^{pq} = 1$; otherwise $\hat{w}_{st}^{pq} = 0$.

We define the semi-supervised multi-view correlation as

$$\text{Corr} = \frac{\sum\limits_{s=1}^{N} \sum\limits_{p=1}^{n_s} \sum\limits_{t=1}^{N} \sum\limits_{q=1}^{n_t} v_s^T x_s^p \hat{w}_{st}^{pq} x_t^{qT} v_t}{\sqrt{\sum\limits_{s=1}^{N} \sum\limits_{p=1}^{n_s} v_s^T x_s^p x_s^{pT} v_s} \sqrt{\sum\limits_{t=1}^{N} \sum\limits_{q=1}^{n_t} v_t^T x_t^q x_t^{qT} v_t}} = \frac{v^T Q' v}{v^T P' v}, \tag{3}$$

where $Q' = \begin{bmatrix} X_1 \hat{W}_{11} X_1^T & X_1 \hat{W}_{12} X_2^T & \cdots & X_1 \hat{W}_{1N} X_N^T \\ X_2 \hat{W}_{21} X_1^T & X_2 \hat{W}_{22} X_2^T & \cdots & X_2 \hat{W}_{2N} X_2^T \\ \vdots & \vdots & \cdots & \vdots \\ X_N \hat{W}_{N1} X_1^T & X_N \hat{W}_{N2} X_2^T & \cdots & X_N \hat{W}_{NN} X_N^T \end{bmatrix}$, $P' = \begin{bmatrix} X_1 X_1^T & 0 & \cdots & 0 \\ 0 & X_2 X_2^T & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & X_N X_N^T \end{bmatrix}$, and $\hat{W}_{st}$ $(s, t = 1, 2, \cdots, N)$ is a $n_s \times n_t$ matrix that satisfies $\hat{W} = \begin{bmatrix} \hat{W}_{11} & \hat{W}_{12} & \cdots & \hat{W}_{1N} \\ \hat{W}_{21} & \hat{W}_{22} & \cdots & \hat{W}_{2N} \\ \vdots & \vdots & \cdots & \vdots \\ \hat{W}_{N1} & \hat{W}_{N2} & \cdots & \hat{W}_{NN} \end{bmatrix}$.

Formula (3) expresses the adverse multi-view correlation among classes and unlabeled samples. To reduce this correlation, we make $\text{Corr} = 0$, which is equivalent to $v^T Q' v = 0$. Thus we use $v^T Q' v = 0$ as the semi-supervised uncorrelation constraint.

3.3 USI^{2}MD Approach Description

By combining the SI^{2}MD schema and semi-supervised uncorrelation constraint, we formulate the objective function of our USI^{2}MD approach as

$$\max_{v} \ v^T \left(Q - \beta P\right) v \quad s.t. \ v^T Q' v = 0. \tag{4}$$

The solution of Formula (4) can be obtained by solving the following eigen-equation:

$$\left(Q - \beta P\right) v = \lambda Q' v. \tag{5}$$

Once the eigenvectors $v^k \left(k = 1, 2, \cdots, d; d < m\right)$ associated with d largest eigenvalues of $Q'^{-1}\left(Q - \beta P\right)$ are obtained, we get $v_1^k, v_2^k, \cdots, v_N^k$ from v^k. Let $V_s = \left[v_s^1, v_s^2, \cdots, v_s^d\right]$, and $y_1, y_2, \cdots, y_N$ denote N views of a query sample, where $s = 1, 2, \cdots, N$. Then we can obtain the projected features of training sample set Z_s^X and query sample Z_s^y separately by $Z_s^X = V_s^T X_s$ and $Z_s^y = V_s^T y_s$, and use the following strategy to fuse these features:

$$Z^X = \left[Z_1^{XT}, Z_2^{XT}, \cdots, Z_N^{XT}\right]^T, \tag{6}$$

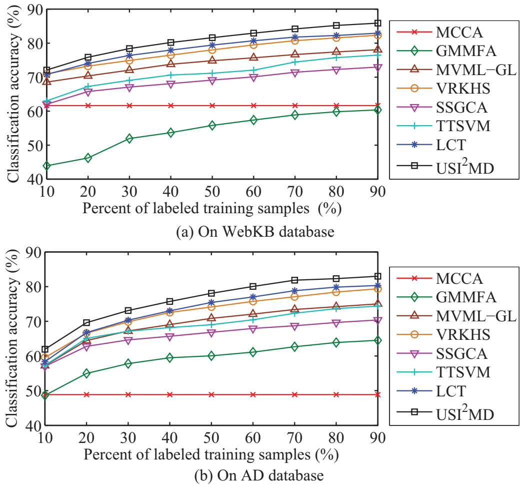
$$Z^y = \left[Z_1^{yT}, Z_2^{yT}, \cdots, Z_N^{yT}\right]^T. \tag{7}$$

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

3.4 Time Complexity Analysis

Here, we analyze the time complexity of our approach and those of three related and representative semi-supervised multi-view feature extraction methods including SSGCA, MVML-GL and VRKHS. The time complexity of our approach is $O\left(NMn_T^2 + M^3\right)$, where $O\left(NMn_T^2\right)$ is the time complexity of calculating the matrixes P, Q and Q', and $O\left(M^3\right)$ is the time complexity of calculating the matrix $Q'^{-1}\left(Q - \beta P\right)$ and conducting the eigen-decomposition of this matrix. The time complexities of SSGCA, MVML-GL and VRKHS are $O\left(Mn_T^2 + M^3\right)$, $O\left(n_L^3 + M^2 n_L\right)$ and $O\left(NMn_T^2 + N^3 n_T^3\right)$, respectively. Here, $M = N \times m$, m denotes the dimensionality of samples from multiple views, N denotes the number of views, n_L and n_T separately denote the numbers of labeled samples and total samples (contains both labeled and unlabeled samples) in each view. It is noted that different views own the same number of samples, i.e. $n_i \left(i = 1, \cdots, N\right) = n_T$, and and different views own the same number of labeled samples, i.e. n_L.

The time complexity of our approach is slightly larger than that of SSGCA. Whether the time complexity of our approach



(a) On WebKB database

(b) On AD database

Figure 2: Average classification accuracies.

Table 1: P-values between USI^{2}MD and other methods.

USI^{2}MD	WebKB database	AD database
MCCA	1.59×10^{-6}	2.54×10^{-6}
GMMFA	1.64×10^{-11}	1.03×10^{-8}
MVML-GL	5.35×10^{-7}	3.60×10^{-7}
VRKHS	5.43×10^{-7}	3.31×10^{-7}
SSGCA	1.50×10^{-9}	5.52×10^{-6}
TTSVM	1.16×10^{-10}	3.57×10^{-6}
LCT	1.24×10^{-6}	5.14×10^{-9}

is lower than that of MVML-GL is mainly determined by the values of M and n_L. And whether the time complexity of our approach is lower than that of VRKHS is determined by the values of m and n_T. In real-world applications, we found that the time cost of our approach is usually close to those of MVML-GL and VRKHS.

4 Experiments

4.1 Introduction of Databases

We use two public web page databases, i.e. the WebKB [Chen *et al.*, 2012] and Internet advertisements [Kushmerick, 1999] databases, as test data.

The WebKB database contains 1051 web pages belonging to two categories: 230 pages in the course category and 821 pages in the non-course category. Each web page has two views: the page view that contains the textual content of this page, and the link view that contains the links from other web pages linking to this page. We use a preprocessed version of this database in our experiment, where 3000-dimensional and 1840-dimensional original features are separately extracted from the page view and link view of a web page.

The Internet advertisements (AD for short) database contains 3279 samples, among which 458 are advertisements and others are not. We use a preprocessed version of this database in the experiment, where each sample is treated as a binary vector with quite large sparsity, which consists of 1558 original features of 9 views. Here, we choose 3 views of them, i.e. 495 base URL features, 472 destination URL features and 457 image URL features, which hold comparatively more features.

4.2 Compared Methods and Experimental Settings

In the experiment, we compare our USI^{2}MD approach with six related and representative multi-view feature extraction and web page classification methods, which can be divided into three categories: **MCCA** [Li *et al.*, 2009] and **GMMFA** [Sharma *et al.*, 2012] are unsupervised or supervised multi-view subspace learning methods; **MVML-GL** [Zhai *et al.*, 2012] and **VRKHS** [Minh *et al.*, 2013] are semi-supervised multi-view feature extraction methods; **SSGCA** [Chen *et al.*, 2012], **TTSVM** [Li *et al.*, 2012a] and **LCT** [Du *et al.*, 2013] are web page classification methods, in which SSGCA is also a semi-supervised multi-view feature extraction method.

For each database, we randomly select half of total samples from each class for training and the rest for testing, and further randomly choose a percentage (e.g., 10%) of the training samples as labeled samples and the rest as unlabeled ones for semi-supervised learning. Note that in the training stage: (1) all labeled and unlabeled training samples are used for unsupervised MCCA method; (2) only labeled training samples are used for supervised GMMFA method. In the AD database, the two views with best classification accuracy are used for SSGCA and TTSVM, since these two methods can deal with only two views. For our USI^{2}MD approach, we set the coefficient $\beta = 0.5$, the parameter of heat kernel $r = 140$ for WebKB database and $r = 3.5$ for AD database, and the feature number $d = 40$ for WebKB database and $d = 50$ for AD database. The parameters of other methods are determined by using the 5-fold cross validation technique.

For all compared methods, we employ the principal component analysis (PCA) [Turk and Pentland, 1991] method to reduce the dimensionality of samples from different views to 1050 for WebKB database and 456 for AD database, because we need to analyze the relations between views for utilizing inter-view discriminant information, and thus demand the same dimensionality of inter-view samples. For all compared methods except TTSVM, we employ the nearest neighbor classifier with the cosine distance to do classification, while TTSVM itself is a classifier and can be directly used for classification.

4.3 Classification Performance Evaluation

We evaluate the classification effects of our USI^{2}MD approach when the percent of labeled training samples varies from 10% to 90%. Figure 2 shows the corresponding average classification accuracies of all methods across 20 runs on two databases. Note that since MCCA is an unsupervised method, its classification accuracies keep unchanged. As seen in Figure 2, USI^{2}MD always outperforms other compared methods, which verifies the effectiveness of our approach.

To statistically analyze the classification accuracies given in Figure 2, we conduct a statistical test, i.e. Mcnemar's test [Draper *et al.*, 2002]. This test can provide statistical significance between USI^{2}MD and other methods. It uses a significance level of 0.05, that is, if the p-value is below 0.05,

Table 2: Average discriminability scores.

Method	WebKB database	AD database
MVML-GL	0.0128	0.0104
VRKHS	0.0199	0.0155
SSGCA	0.0077	0.0068
SI^{2}MD	0.0257	0.0188

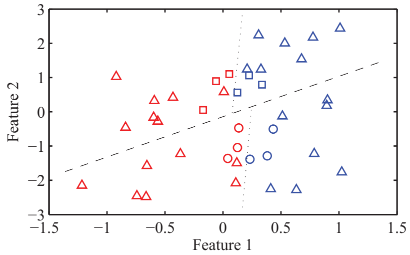


Figure 3: Sample distribution of two views of 20 web page samples (6 labeled samples from two classes with 3 samples per class and 14 unlabeled samples) in the feature space of USI^{2}MD approach on WebKB database, where red and blue colors denote two views, markers ○ and □ denote sample points from two classes, marker △ denotes an unlabeled sample point, and Features 1 and 2 denote two major principal components of the features extracted by USI^{2}MD approach.

the performance difference between two compared methods is statistically significant. Table 1 shows the p-values between USI^{2}MD and other methods on two databases. According to Table 1, the proposed approach makes a statistically significant difference as compared with the related methods.

Corresponding to Figure 1, Figure 3 shows the sample distribution of two views of 20 web page samples in the feature space of our approach on WebKB database. We can find that intra-view and inter-view sample points from different classes are well separated, which demonstrates that our approach can sufficiently utilize intra-view and inter-view discriminant information of labeled samples; and sample points from different views are mostly separated, which indicates that our approach can effectively reduce the adverse multi-view correlation in semi-supervised scenario.

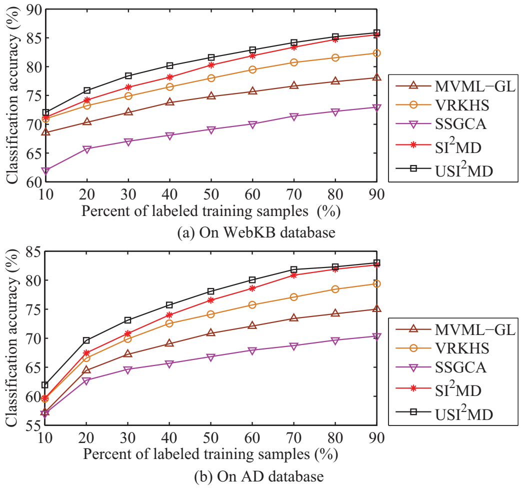
4.4 Effect of SI^{2}MD Schema

To further analyze the discriminant ability of features extracted by the SI^{2}MD schema, we define the following discriminability score:

$$S = tr\left(S_{bL}\right)/tr\left(S_{t(L+U)}\right), \tag{8}$$

where S_{bL} is the between-class scatter matrix constructed by labeled samples in the feature space, $S_{t(L+U)}$ is the total scatter matrix constructed by both labeled and unlabeled samples in the feature space, and $tr(\cdot)$ denotes the trace of a square matrix.

For the SI^{2}MD schema, we use Formula (2) to calculate the projective vectors, and obtain the projected and fused features of training sample set as USI^{2}MD does. Here we use the related semi-supervised multi-view feature extraction methods including MVML-GL, VRKHS and SSGCA as compared methods, and the percent of labeled training samples is set as 20%. Table 2 show the average discriminability scores of these three methods and the SI^{2}MD schema across 20 runs on two databases. In comparison with other methods, the SI^{2}MD schema achieves the highest discriminability scores.



(a) On WebKB database

(b) On AD database

Figure 4: Average classification accuracies of SI^{2}MD, USI^{2}MD, MVML-GL, VRKHS and SSGCA.

4.5 Effect of Semi-supervised Uncorrelation Constraint

To further analyze the effectiveness of our semi-supervised uncorrelation constraint, we compare the classification results of the SI^{2}MD schema (without the constraint), the USI^{2}MD approach (the SI^{2}MD schema and the constraint), MVML-GL, VRKHS and SSGCA. Figure 4 shows the average classification accuracies of these five methods across 20 runs on two databases. We can find that the accuracies of the SI^{2}MD schema are lower than those of the USI^{2}MD approach, but higher than those of other methods.

5 Conclusions

In this paper, we firstly propose a new feature extraction schema called SI^{2}MD, which can sufficiently utilize the intra-view and inter-view discriminant information of labeled samples and the local neighborhood structures of unlabeled samples to extract useful features from web page samples. Then we design a semi-supervised uncorrelation constraint to remove the adverse multi-view features correlation among samples not with the same class label. Finally, by incorporating the constraint into the SI^{2}MD schema, we propose a novel semi-supervised multi-view feature extraction approach named USI^{2}MD for web page classification.

Experimental results on public web page databases demonstrate that USI^{2}MD achieves better classification accuracies than several representative multi-view feature extraction and web page classification methods. The Mcnemar's test experiment shows that the difference between USI^{2}MD and other methods are statistically significant. In addition, experiments validate the effects of the designed SI^{2}MD schema and semi-supervised uncorrelation constraint.

In addition, it should be noted that except web page classification, the proposed approach can be also applied to other semi-supervised applications with high-dimensional multi-view data.

Acknowledgements

The authors want to thank the 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 Projects No. 61272273, No. 61233011, No. 61472178 and No. 61472186, and the Research Project of NJUPT (XJKY14016).

References

[Chen *et al.*, 2012] X.H. Chen, S.C. Chen, H. Xue, and X.D. Zhou. A Unified Dimensionality Reduction Framework for Semi-paired and Semi-supervised Multi-view Data. *Pattern Recognition*, 45(5): 2005-2018, 2012.

[Chen and Hsieh, 2006] R.C. Chen and C.H. Hsieh. Web Page Classification Based on A Support Vector Machine Using A Weighted Vote Schema. *Expert Systems with Applications*, 31(2): 427-435, 2006.

[Chen *et al.*, 2009] C.M. Chen, H.M. Lee, and Y.J. Chang. Two Novel Feature Selection Approaches for Web Page Classification. *Expert Systems with Applications*, 36(1): 260-272, 2009.

[Draper *et al.*, 2002] B.A. Draper, W.S. Yambor, and J.R. Beveridge. Analyzing PCA-based Face Recognition Algorithms: Eigenvector Selection and Distance Measures. *Empirical Evaluation Methods in Computer Vision*, pages 1-15, 2002.

[Du *et al.*, 2013] Y.T. Du, C. Su, Z.M. Cai, and X.H. Guan. Web Page and Image Semi-supervised Classification with Heterogeneous Information Fusion. *Journal of Information Science*, 39(3): 289-306, 2013.

[Gao *et al.*, 2012] L. Gao, L. Qi, E.Q. Chen, and L. Guan. Discriminative Multiple Canonical Correlation Analysis for Multi-feature Information Fusion. *IEEE Int. Symposium on Multimedia*, pages 36-43, 2012.

[Hardoon *et al.*, 2004] D.R. Hardoon, S. Szedmak, and J.S. Taylor. Canonical Correlation Analysis: An Overview with Application to Learning Method. *Neural Computation*, 16(12): 2639-2664, 2004.

[Hong *et al.*, 2013] C.Q. Hong, J. Yu, J. Li, and X.H. Chen. Multi-view Hypergraph Learning by Patch Alignment Framework. *Neurocomputing*, 118: 79-86, 2013.

[Hu *et al.*, 2013] M.D. Hu, Y.H. Wang, Z.X. Zhang, J.J. Little, and D. Huang. View-invariant Discriminative Projection for Multi-view Gait-based Human Identification. *IEEE Trans. Information Forensics and Security*, 8(12): 2034-2045, 2013.

[Hu *et al.*, 2007] W.M. Hu, O. Wu, Z.Y. Chen, Z.Y. Fu, and S. Maybank. Recognition of Pornographic Web Pages by Classifying Texts and Images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(6): 1019-1034, 2007.

[Kan *et al.*, 2012] M. Kan, S.G. Shan, H.H. Zhang, S.H. Lao, and X.L. Chen. Multi-view Discriminant Analysis. *European Conf. Computer Vision*, pages 808-821, 2012.

[Kushmerick, 1999] N. Kushmerick. Learning to Remove Internet Advertisements. *Annual Conf. Autonomous Agents*, pages 175-181, 1999.

[Li *et al.*, 2009] Y.O. Li, T. Adali, W. Wang, and V.D. Calhoun. Joint Blind Source Separation by Multiset Canonical Correlation Analysis. *IEEE Trans. Signal Processing*, 57(10): 3918-3929, 2009.

[Li *et al.*, 2012a] G.X. Li, K.Y. Chang, and S.C.H. Hoi. Multiview Semi-supervised Learning with Consensus. *IEEE Trans. Knowledge and Data Engineering*, 24(11): 2040-2051, 2012.

[Li *et al.*, 2012b] W. Li, L.X. Duan, I.W.H. Tsang, and D. Xu. Co-labeling: A New Multi-view Learning Approach for Ambiguous Problems. *Int. Conf. Data Mining*, pages 419-428, 2012.

[Liu and Tao, 2013] W.F. Liu, and D.C. Tao. Multiview Hessian Regularization for Image Annotation. *IEEE Trans. Image Processing*, 22(7): 2676-2687, 2013.

[Minh *et al.*, 2013] H.Q. Minh, L. Bazzani, and V. Murino. A Unifying Framework for Vector-valued Manifold Regularization and Multi-view Learning. *Int. Conf. Machine Learning*, pages 759-767, 2013.

[Muslea *et al.*, 2006] I. Muslea, S. Minton, and C.A. Knoblock. Active Learning with Multiple Views. *Journal of Artificial Intelligence Research*, 27: 203-233, 2006.

[Richarz *et al.*, 2014] J. Richarz, S. Vajda, R. Grzeszick, and G.A. Fink. Semi-supervised Learning for Character Recognition in Historical Archive Documents. *Pattern Recognition*, 47(3): 1011-1020, 2014.

[Sharma *et al.*, 2012] A. Sharma, A. Kumar, H. Daume, and D.W. Jacobs. Generalized Multiview Analysis: A Discriminative Latent Space. *IEEE Conf. Computer Vision and Pattern Recognition*, pages 2160-2167, 2012.

[Su *et al.*, 2012] Y. Su, Y. Fu, X.B. Gao, and Q. Tian. Discriminant Learning through Multiple Principal Angles for Visual Recognition. *IEEE Trans. Image Processing*, 21(3): 1381-1390, 2012.

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

[Wang and Chen, 2009] R.P. Wang, and X.L. Chen. Manifold Discriminant Analysis. *IEEE Conf. Computer Vision and Pattern Recognition*, pages 429-436, 2009.

[Yang and Gao, 2013] P. Yang, and W. Gao. Multi-View Discriminant Transfer Learning. *Int. Joint Conf. Artificial Intelligence*, pages 1848-1854, 2013.

[Yuan *et al.*, 2011] Y.H. Yuan, Q.S. Sun, Q. Zhou, and D.S. Xia. A Novel Multiset Integrated Canonical Correlation Analysis Framework and Its Application in Feature Fusion. *Pattern Recognition*, 44(5): 1031-1040, 2011.

[Zhai *et al.*, 2012] D.M. Zhai, H. Chang, S.G. Shan, X.L. Chen, and W. Gao. Multiview Metric Learning with Global Consistency and Local Smoothness. *ACM Trans. Intelligent Systems and Technology*, 3(3): article number 53, 2012.

[Zhang *et al.*, 2011] H.J. Zhang, G. Liu, T.W.S. Chow, and W.Y. Liu. Textual and Visual Content-based Anti-phishing: A Bayesian Approach. *IEEE Trans. Neural Networks*, 22(10): 1532-1546, 2011.

[Zhou and Li, 2005] Z.H. Zhou, and M. Li. Tri-training: Exploiting Unlabeled Data Using Three Classifiers. *IEEE Trans. Knowledge and Data Engineering*, 17(11): 1529-1541, 2005.

[Zhuang *et al.*, 2012] F.Z. Zhuang, G. Karypis, X. Ning, Q. He, and Z.Z. Shi. Multi-view Learning via Probabilistic Latent Semantic Analysis. *Information Sciences*, 199: 20-30, 2012.