

Supervised multiview learning based on simultaneous learning of multiview intact and single view classifier

Qingjun Wang¹ · Haiyan Lv² · Jun Yue¹ · Eugene Mitchell³

Received: 20 September 2015 / Accepted: 2 January 2016 / Published online: 21 January 2016
© The Natural Computing Applications Forum 2016

Abstract Multiview learning problem refers to the problem of learning a classifier from multiple view data. In this data set, each data point is presented by multiple different views. In this paper, we propose a novel method for this problem. This method is based on two assumptions. The first assumption is that each data point has an intact feature vector, and each view is obtained by a linear transformation from the intact vector. The second assumption is that the intact vectors are discriminative, and in the intact space, we have a linear classifier to separate the positive class from the negative class. We define an intact vector for each data point, and a view-conditional transformation matrix for each view, and propose to reconstruct the multiple view feature vectors by the product of the corresponding intact vectors and transformation matrices. Moreover, we also propose a linear classifier in the intact space, and learn it jointly with the intact vectors. The learning problem is modeled by a minimization problem, and the objective function is composed of a Cauchy error estimator-based view-conditional reconstruction term over all data points and views, and a classification error term measured by hinge loss over all the intact vectors of all the data points. Some regularization

terms are also imposed to different variables in the objective function. The minimization problem is solved by an iterative algorithm using alternate optimization strategy and gradient descent algorithm. The proposed algorithm shows its advantage in the compression to other multiview learning algorithms on benchmark data sets.

Keywords Multiview learning · Supervised learning · Intact space · Hinge loss

1 Introduction

1.1 Background

Multiview learning has been an important in machine learning community [5, 21, 22, 24, 32, 34, 35, 38, 40, 41]. In traditional machine learning problems, we usually assume that a data point has a feature vector to represent its input information. For example, in image recognition problem, we can extract a visual feature vector from an image, using a texture descriptor [9, 14, 25–28, 36, 39]. In this scene, the texture is a view of the image. However, there could be more than one view of an image. Besides the texture view, we can also extract feature vectors from other views, including shape and color. An other example is the problem of classification of scientific articles, and we may extract a feature vector from the main text of the article [4, 6, 11, 11, 16, 18, 19, 23, 29, 29]. However, the main text is just one view of the article, and we can also have features from other views, such as abstract, reference list. Multiview learning argues that we should learn from more than one views to present the data and construct a predictor. The motive for multiview learning is that single view-based data representation is usually incomplete, and different

✉ Qingjun Wang
qjwang386@hotmail.com

Eugene Mitchell
emitchell328496@outlook.com

¹ School of Information and Electrical Engineering,
Ludong University, Yantai 264025, China

² Naval Aeronautical and Astronautical University,
Yantai 264025, China

³ Department of Computer Science, Ryerson University,
Toronto, ON M5B 2K3, Canada

views can present complementary information for the learning problem. In the problem of multiview learning, the input of a data point is not just one single feature vector of one single view, but multiple feature vectors presenting different views. The target of multiview learning is to learn a predictor to take multiple view feature vectors to predict one single output of a data point. The problem of multiview learning can be classified to two types, supervised multiview learning and unsupervised learning.

- Supervised multiview learning refers to the problem of learning from a data set, where both the multiview input and output are available for each data point [10, 15, 20]. In this problem, the output is usually a class label, or a continuous response. In this case, the learning problem is to build a predictive model from the training data set to predict the output of a input data point, with help the input–output pairs of the training set.
- Unsupervised multiview learning refers to the problem of cluster a set of data points, and the multiview inputs of each data point are given [7, 33, 44]. In this problem, the outputs of the data points are not available.

In this paper, we investigate the problem of supervised multiview learning, and propose a novel algorithm to solve it. The proposed method is based on an assumption that different views of a data point are generated from one single intact feature vector, and the view generation is performed by a linear transformation. We try to recover the intact feature vector for each data point from its multiview feature vectors, with guiding of its corresponding output, i.e., its binary class label.

1.2 Relevant works

There are some existing multiview learning methods. Their state-of-the-arts is as follows.

- Zhang et al. [43] proposed to use local learning (LL) method for the problem of multiview learning problem, and designed a local predictive model for each data point based on the multiview inputs. The local predictive model is learned on the nearest neighbors of a data point.
- Sindhvani et al. [31] proposed to use co-training algorithm for multiview learning problems to improve the classification performance of each view (CT). This method is based on multiview regularization, and the agreement and smoothness over both labeled and unlabeled data points.
- Quadrianto [30] proposed a multiview learning algorithm to solve the problem of view disagreement (VD), i.e., different views of one single data point do not

belong to the same class. This method uses a conditional entropy criterion to find the disagreement among different views, and removes the data points with view disagreement from the training set.

- Zhai [42] proposed multiview metric learning method with global consistency and local smoothness (GL) for the multiview learning problem with partially labeled data set. This method simultaneously considers both the global consistency and local smoothness, by assuming that the different views have a shared latent feature space, and imposing global consistency and local structure to the learning procedure.
- Chen et al. [2] proposed a statistical subspace multiview representation method (SS), by leveraging both multiview dependencies and supervision information. This method is based on a subspace Markov network of multiview latent and assumes that the multiviews and the class labels are conditionally independent. The algorithm is based on the maximization of data likelihood and the minimization of classification error.

1.3 Contributions

In this paper, we propose a novel supervised multiview learning method. This method is based on the assumption of single discriminative intact of different multiview inputs. Under this assumption, although there are different views of one single data point, one single intact feature vector exists for the data point. This intact feature vector is assumed to be discriminative, i.e., it can represent the class information of each data point. Moreover, the feature vector of each view of a data point can be obtained from the intact vector, by performing a linear view-conditional transformation to the intact feature vector. In this way, if we learn the discriminative intact feature vector for each training data point, we can learn a classifier in the intact with the help of the class labels of the training data points. To this end, we proposed a novel method to learn the hidden intact feature vectors, the view-conditional transformation matrices, and the classifier in the intact space simultaneously. We define a intact feature vector for each data point, and a transformation matrix for each view. The feature vector of one view of each data point can be reconstructed as the product of its corresponding transformation matrix and intact feature vector. The reconstruction error for each view of each data point is measured by the Cauchy error estimator [8, 12]. To learn the optimal intact feature vectors and view-conditional transformation matrices, we propose to minimize the Cauchy errors. Moreover, due to the assumption that the intact feature vectors are discriminative, we also argue that we can design a classifier in the intact space, and the classifier can

minimize the classification error. Thus, we also propose to learn a linear classifier in the intact space, and use the hinge loss to measure the classification error the training set in the intact space [1, 3]. To learn the optimal classifier parameter and the intact feature vectors, we also propose to minimize the hinge loss with regard to both the classifier parameter and the intact feature vectors.

To model the problem, we propose a joint optimization problem for learning of intact vectors, view-conditional transformation matrices, and the classifier parameter vector. The objective function of this problem is composed of two error terms, and three regularization terms. The first error term is the view reconstruction term measured by Cauchy estimator over all the data points and views. The second error term is the classification error over all the intact feature vectors of all training data points, measured by hinge losses. The three regularization terms are all squared ℓ_2 norm terms over each intact feature vectors, view-conditional matrices, and the classifier parameter vectors. The purpose of impose the squared ℓ_2 norm to these variables is to reduce the complexity of the learned outputs. To minimize the proposed objective function, we adapt an alternate optimization strategy, i.e., when the objective function is minimized with regard to one variable, other variables are fixed. The optimization with regard to each variable is conducted by using gradient descent algorithm.

The contributions of this paper are of three parts:

1. We propose a novel supervised multiview learning framework by simultaneous learning of intact feature vectors, view-conditional transformation matrices, and classifier parameter vector.
2. We build a novel optimization problem for this learning problem, by considering both the view reconstruction problem, and the classifier learning problem.
3. We develop an iterative algorithm to solve this optimization problem based on alternate optimization strategy and gradient descent algorithm.

1.4 Paper organization

This paper is organized as follows: In Sect. 2, the proposed method for supervised multiview learning is introduced. In this section, we first model this problem as a minimization problem of a objective function, and then solve it with an iterative algorithm. In Sect. 3, the proposed iterative algorithm is evaluated. We first give an analysis of its sensitivity to parameters, and then compare it to some state-of-the-art algorithms, and finally test the running time performance of the proposed algorithm. In Sect. 4, we give the conclusion of this paper.

2 Methods

In this section, we introduce the proposed supervised multiview learning method.

2.1 Problem modeling

We assume we are dealing with supervised binary classification problem with multiview data. A training data set of n data points is given, $X = \{\theta_1, \dots, \theta_n\}$. $\theta_i = (\mathbf{x}_i^1, \dots, \mathbf{x}_i^m, y_i)$ is the i th data point. The information of each data point is composed of feature vectors of m views, and a binary class label y_i . $\mathbf{x}_i^j \in \mathbb{R}^{d_j}$ is the d_j -dimensional feature vector of the j th view of the i th data point, and $y_i \in \{+1, -1\}$ is a the binary class label of the i th data point. The problem of supervised multiview learning is to learn a predictive model from the training set, which can predict a binary class label from the multiview input of a test data point. We assume there is an intact vector $\mathbf{z}_i \in \mathbb{R}^d$ for the i th data point, and its j th view \mathbf{x}_i^j can be reconstructed by a linear transformation,

$$\mathbf{x}_i^j \leftarrow W_j \mathbf{z}_i, \tag{1}$$

where $W_j \in \mathbb{R}^{d_j \times d}$ is the view-conditional linear transformation matrix for the j th view. Please note that the view-conditional transformation matrix for the same view of all the data points is the same. By learning both the W_j and \mathbf{z}_i , we can recover the hidden intact vector for the i th data point, \mathbf{z}_i , and use it for classification problem. To this end, we propose to minimize the reconstruction error. The reconstruction error is measured by Cauchy error estimator, $E(\mathbf{x}_i^j, W_j \mathbf{z}_i)$,

$$E(\mathbf{x}_i^j, W_j \mathbf{z}_i) = \log \left(1 + \frac{\|\mathbf{x}_i^j - W_j \mathbf{z}_i\|_2^2}{c^2} \right). \tag{2}$$

This error estimator has been shown to be robust, and it also provides an offset. We propose to minimize this error estimator over all data points and all views with regard to both $\mathbf{z}_i, i = 1, \dots, n$, and $W_j, j = 1, \dots, m$,

$$\min_{\mathbf{z}_i, W_j} \left\{ \sum_{i=1}^n \sum_{j=1}^m E(\mathbf{x}_i^j, W_j \mathbf{z}_i) = \sum_{i=1}^n \sum_{j=1}^m \log \left(1 + \frac{\|\mathbf{x}_i^j - W_j \mathbf{z}_i\|_2^2}{c^2} \right) \right\} \tag{3}$$

Moreover, we also assume that the intact feature vectors of the data points are discriminative, and presents the class information, thus the intact feature vectors can minimize a classification loss function of the data set. We propose to learn the intact feature vector of the i th data point by jointly learning a liner classifier to predict its class label, y_i . The classifier is designed as linear function,

$$y_i \leftarrow \omega^\top \mathbf{z}_i \quad (4)$$

The classification error can be measured by the hinge loss function,

$$L(y_i, \omega^\top \mathbf{z}_i) = \max(0, 1 - y_i \omega^\top \mathbf{z}_i). \quad (5)$$

The optimization of this loss function can obtain a large margin classifier. To learn the optimal classifier and the discriminative intact feature vectors, we propose to minimize the classifier loss measured by the hinge loss function of the classification result over all the training data points,

$$\min_{\mathbf{z}_i|_{i=1}^n, \omega} \left\{ \sum_{i=1}^n L(y_i, \omega^\top \mathbf{z}_i) = \sum_{i=1}^n \max(0, 1 - y_i \omega^\top \mathbf{z}_i) \right\} \quad (6)$$

Moreover, to prevent the problem of over-fitting of variables, we propose to minimize the squared ℓ_2 norm of the variables to regularize the learning \mathbf{z}_i , W_j , and ω ,

$$\min_{\mathbf{z}_i|_{i=1}^n, W_j|_{j=1}^m, \omega} \left\{ R(\mathbf{z}_i|_{i=1}^n, W_j|_{j=1}^m, \omega) = \sum_{i=1}^n \|\mathbf{z}_i\|_2^2 + \sum_{j=1}^m \|W_j\|_2^2 + \|\omega\|_2^2 \right\}. \quad (7)$$

Our overall learning problem is obtained by considering both the problems of view-conditional reconstruction in (3), and classifier learning in the intact space in (6),

$$\begin{aligned} \min_{\mathbf{z}_i|_{i=1}^n, W_j|_{j=1}^m, \omega} & \left\{ \sum_{i=1}^n \sum_{j=1}^m E(\mathbf{x}_i^j, W_j \mathbf{z}_i) + \alpha L(y_i, \omega^\top \mathbf{z}_i) + \gamma R(\mathbf{z}_i|_{i=1}^n, W_j|_{j=1}^m, \omega) \right. \\ & = \sum_{i=1}^n \sum_{j=1}^m \log \left(1 + \frac{\|\mathbf{x}_i^j - W_j \mathbf{z}_i\|_2^2}{c^2} \right) \\ & + \alpha \sum_{i=1}^n \max(0, 1 - y_i \omega^\top \mathbf{z}_i) \\ & \left. + \gamma \left(\sum_{i=1}^n \|\mathbf{z}_i\|_2^2 + \sum_{j=1}^m \|W_j\|_2^2 + \|\omega\|_2^2 \right) \right\}, \quad (8) \end{aligned}$$

where α is a tradeoff parameter to balance the view-conditional reconstruction terms and the classification error terms, and γ is a tradeoff parameter to balance the view-conditional reconstruction terms and the regularization terms. By optimizing this problem, we can learn intact feature vectors which can present the multiview inputs of the data points, and also is discriminative.

2.2 Optimization

To solve the optimization problem in (21), we propose to use the alternate optimization strategy. The optimization is conducted in an iterative algorithm. When one variable is considered, the others are fixed. After one variable is updated, it will be fixed in the next iteration when other variable is updated. In the following subsections, we will discuss how to update each variable.

2.2.1 Updating \mathbf{z}_i

When we want to update \mathbf{z}_i , we only consider this single variable, while fix all other variables. Thus we have the following optimization problem,

$$\min_{\mathbf{z}_i} \left\{ \sum_{j=1}^m \log \left(1 + \frac{\|\mathbf{x}_i^j - W_j \mathbf{z}_i\|_2^2}{c^2} \right) + \alpha \max(0, 1 - y_i \omega^\top \mathbf{z}_i) + \gamma \|\mathbf{z}_i\|_2^2 \right\}. \quad (9)$$

The second term $\max(0, 1 - y_i \omega^\top \mathbf{z}_i)$ is not a convex function, and it is hard to optimize it directly. Thus we rewrite it as follows,

$$\max(0, 1 - y_i \omega^\top \mathbf{z}_i) = \begin{cases} 1 - y_i \omega^\top \mathbf{z}_i, & \text{if } 1 - y_i \omega^\top \mathbf{z}_i > 0 \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

We define a indicator variable, β_i , to indicate which of the above cases is true,

$$\beta_i = \begin{cases} 1, & \text{if } 1 - y_i \omega^\top \mathbf{z}_i > 0 \\ 0, & \text{otherwise,} \end{cases} \quad (11)$$

and rewrite (10) as follows,

$$\max(0, 1 - y_i \omega^\top \mathbf{z}_i) = \beta_i (1 - y_i \omega^\top \mathbf{z}_i) \quad (12)$$

Please note that β_i is also a function of \mathbf{z}_i ; however, we first update it by using \mathbf{z}_i solved in previous iteration, and then fix it to update \mathbf{z}_i in current iteration. In this way, (9) is rewritten as

$$\min_{\mathbf{z}_i} \left\{ \sum_{j=1}^m \log \left(1 + \frac{\|\mathbf{x}_i^j - W_j \mathbf{z}_i\|_2^2}{c^2} \right) + \alpha \beta_i (1 - y_i \omega^\top \mathbf{z}_i) + \gamma \|\mathbf{z}_i\|_2^2 = g(\mathbf{z}_i) \right\}, \quad (13)$$

where $g(\mathbf{z}_i)$ is the objective of this optimization problem. To seek the minimization of $g(\mathbf{z}_i)$, we use gradient descent algorithm. This algorithm update \mathbf{z}_i by descending it to the direction of gradient of $g(\mathbf{z}_i)$,

$$\mathbf{z}_i \leftarrow \mathbf{z}_i - \mu \nabla g(\mathbf{z}_i), \quad (14)$$

where μ is the descent step, and $\nabla g(\mathbf{z}_i)$ is the gradient function of $g(\mathbf{z}_i)$. We set $\nabla g(\mathbf{z}_i)$ as the partial derivative of $g(\mathbf{z}_i)$ with regard to \mathbf{z}_i ,

$$\begin{aligned} \nabla g(\mathbf{z}_i) &= \frac{\partial g(\mathbf{z}_i)}{\partial \mathbf{z}_i} = \sum_{j=1}^m \frac{2W_j^\top (\mathbf{x}_i^j - W_j \mathbf{z}_i)}{\left(1 + \frac{\|\mathbf{x}_i^j - W_j \mathbf{z}_i\|_2^2}{c^2} \right)} - \alpha \beta_i y_i \omega + \gamma \mathbf{z}_i \\ &= \sum_{j=1}^m \frac{2W_j^\top (\mathbf{x}_i^j - W_j \mathbf{z}_i)}{\left(c^2 + \|\mathbf{x}_i^j - W_j \mathbf{z}_i\|_2^2 \right)} - \alpha \beta_i y_i \omega + \gamma \mathbf{z}_i. \quad (15) \end{aligned}$$

By substituting (15) to (14), we have the final updating rule of \mathbf{z}_i ,

$$\mathbf{z}_i \leftarrow \mathbf{z}_i - \mu \left(\sum_{j=1}^m \frac{2W_j^\top (\mathbf{x}_i^j - W_j \mathbf{z}_i)}{(c^2 + \|\mathbf{x}_i^j - W_j \mathbf{z}_i\|_2^2)} - \alpha \beta_i y_i \boldsymbol{\omega} + \gamma \mathbf{z}_i \right). \tag{16}$$

2.2.2 Updating W_j

When we want to optimize W_j , we fix all other variables and only consider W_j itself. The optimization problem is changed to,

$$\min_{W_j} \left\{ \sum_{i=1}^n \log \left(1 + \frac{\|\mathbf{x}_i^j - W_j \mathbf{z}_i\|_2^2}{c^2} \right) + \gamma \|W_j\|_2^2 = f(W_j) \right\}. \tag{17}$$

where $f(W_j)$ is the objective function of this problem. To solve this problem, we also update W_j by using the gradient descent algorithm,

$$W_j \leftarrow W_j - \mu \nabla f(W_j), \tag{18}$$

where $\nabla f(W_j)$ is the gradient function of $f(W_j)$,

$$\begin{aligned} \nabla f(W_j) &= \frac{\partial f(W_j)}{\partial W_j} = \sum_{i=1}^n \frac{\frac{2(\mathbf{x}_i^j - W_j \mathbf{z}_i) \mathbf{z}_i^\top}{c^2}}{\left(1 + \frac{\|\mathbf{x}_i^j - W_j \mathbf{z}_i\|_2^2}{c^2} \right)} + \gamma W_j \\ &= \sum_{i=1}^n \frac{2(\mathbf{x}_i^j - W_j \mathbf{z}_i) \mathbf{z}_i^\top}{(c^2 + \|\mathbf{x}_i^j - W_j \mathbf{z}_i\|_2^2)} + \gamma W_j. \end{aligned} \tag{19}$$

Substituting (19) to (18), we have the final updating rule of W_j ,

$$W_j \leftarrow W_j - \mu \left(\sum_{i=1}^n \frac{2(\mathbf{x}_i^j - W_j \mathbf{z}_i) \mathbf{z}_i^\top}{(c^2 + \|\mathbf{x}_i^j - W_j \mathbf{z}_i\|_2^2)} + \gamma W_j \right). \tag{20}$$

2.2.3 Updating $\boldsymbol{\omega}$

When we want to update $\boldsymbol{\omega}$ to minimize the objective function of (21), we fix the other variables, and only consider $\boldsymbol{\omega}$. Thus the problem in (21) is transferred to

$$\min_{\boldsymbol{\omega}} \left\{ \alpha \sum_{i=1}^n \beta_i (1 - y_i \boldsymbol{\omega}^\top \mathbf{z}_i) + \gamma \|\boldsymbol{\omega}\|_2^2 = h(\boldsymbol{\omega}) \right\}. \tag{21}$$

Please note that β_i is actually a function of $\boldsymbol{\omega}$. However, similar the strategy to solve \mathbf{z}_i , we also update it according to $\boldsymbol{\omega}$ solved in previous iteration, and fix it to update $\boldsymbol{\omega}$ in current iteration. When $\beta_i, i = 1, \dots, n$ are fixed, we update $\boldsymbol{\omega}$ to minimize $h(\boldsymbol{\omega})$ by using the gradient descent algorithm,

$$\boldsymbol{\omega} \leftarrow \boldsymbol{\omega} - \mu \nabla h(\boldsymbol{\omega}), \tag{22}$$

where $\nabla h(\boldsymbol{\omega})$ is the gradient function of $h(\boldsymbol{\omega})$, and it is defined as follows,

$$\nabla h(\boldsymbol{\omega}) = \frac{\partial h(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}} = -\alpha \sum_{i=1}^n \beta_i y_i \mathbf{z}_i + \gamma \boldsymbol{\omega}. \tag{23}$$

By substituting it to (24), we have the final updating rule for $\boldsymbol{\omega}$,

$$\boldsymbol{\omega} \leftarrow \boldsymbol{\omega} - \mu \left(-\alpha \sum_{i=1}^n \beta_i y_i \mathbf{z}_i + \gamma \boldsymbol{\omega} \right). \tag{24}$$

2.3 Iterative algorithm

After we have the updating rules of all the variables, we can design an iterative algorithm for the learning problem. This iterative algorithm has one outer FOR loop, and two inner FOR loops. The outer FOR loop is corresponding to the main iterations. The two inner FOR loops are corresponding to the updating of n intact feature vectors of n data points, and the updating of m view-conditional transformation matrices. The algorithm is given in Algorithm 1. The iteration number T is determined by cross-validation in our experiments.

- **Algorithm 1.** Iterative algorithm for multiview intact and single-view classifier learning (MISC).
- **Input:** Training data set, $(\mathbf{x}_1^1, \dots, \mathbf{x}_1^m, y_1), \dots, (\mathbf{x}_n^1, \dots, \mathbf{x}_n^m, y_n)$.
- **Input:** Tradeoff parameters, α and γ .
- **Input:** Maximum iteration number, T .
- **Initialization:** $\mathbf{z}_i^0, i = 1, \dots, n, W_j^0, j = 1, \dots, m$ and $\boldsymbol{\omega}^0$.
- **For** $t = 1, \dots, T$

- Update descent step, $\mu^t \leftarrow \frac{1}{t}$
- **For** $i = 1, \dots, n$
Update β_i^t as follows,

$$\beta_i^t = \begin{cases} 1, & \text{if } 1 - y_i \boldsymbol{\omega}^{t-1 \top} \mathbf{z}_i^{t-1} > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

Update \mathbf{z}_i^t by fixing $W_j^{t-1}, j = 1, \dots, m, \beta_i^{t-1}$ and $\boldsymbol{\omega}^{t-1}$,

$$\begin{aligned} \mathbf{z}_i^t &\leftarrow \mathbf{z}_i^{t-1} - \mu^t \\ &\left(\sum_{j=1}^m \frac{2W_j^{t-1 \top} (\mathbf{x}_i^j - W_j^{t-1} \mathbf{z}_i^{t-1})}{(c^2 + \|\mathbf{x}_i^j - W_j^{t-1} \mathbf{z}_i^{t-1}\|_2^2)} - \alpha \beta_i^t y_i \boldsymbol{\omega}^{t-1} + \gamma \mathbf{z}_i^{t-1} \right). \end{aligned} \tag{26}$$

- **End of For**

- **For** $j = 1, \dots, m$
Update W_j^t by fixing $\mathbf{z}_i^t, i = 1, \dots, m$,

$$W_j^t \leftarrow W_j^{t-1} - \mu^t \left(\sum_{i=1}^n \frac{2(\mathbf{x}_i^j - W_j^{t-1} \mathbf{z}_i^t) \mathbf{z}_i^{t\top}}{(c^2 + \|\mathbf{x}_i^j - W_j^{t-1} \mathbf{z}_i^t\|_2^2)} + \gamma W_j^{t-1} \right). \quad (27)$$

- **End of For**
- Update ω^t by fixing $\beta_i^t, i = 1, \dots, n$ and $\mathbf{z}_i^t, i = 1, \dots, n$,

$$\omega^t \leftarrow \omega^{t-1} - \mu^t \left(-\alpha \sum_{i=1}^n \beta_i^t y_i \mathbf{z}_i + \gamma \omega^{t-1} \right). \quad (28)$$

- **End of For**
- **Output:** $W_j^T, j = 1, \dots, m, \mathbf{z}_i^T, i = 1, \dots, n$, and ω^T .

As we can see from the algorithm, in the main FOR loop, descent step variable, μ , is firstly updated, and then the hinge loss indicator variables, $\beta_i, i = 1, \dots, n$ and the intact feature vectors are updated. The view-conditional transformation matrices, $W_j, j = 1, \dots, m$ are updated, and finally, the classifier parameter ω are updated.

3 Experiments

In this section, we will evaluate the proposed algorithm on a few real-world supervised multiview learning problems experimentally.

3.1 Benchmark data sets

3.1.1 PASCAL VOC 07 data set

The first data set used in the experiment is the PASCAL VOC 07 data set [13]. In this data set, there are 9963 images of 20 different object classes. Each image is presented by two different view, which are visual view and tag view. To extract the feature vector from the visual view of an image, we extract local visual features, SIFT, from the image, and represent the local features as a histogram. To extract the feature vector from the tag view from the image, we use the histogram vector of user tags of the image as the feature vector.

3.1.2 CiteSeer data set

The second data set is the CiteSeer data set [37]. In this data set, there are 3312 documents of 6 classes. Each document has three views, which are the text view, inbound reference view, and outbound reference view.

3.1.3 HMDB data set

The third data set is the HMDB data set, which is a video database of human motion recognition problem [17]. In this data set, there are 6849 video clips of 51 action classes. To present each video clip, we extract 3D Harris corners and present them by two different types of local features, which are the histogram of oriented gradient (HOG) and histogram of oriented flow (HOF). We further represent each clip by two feature vectors of two views, which are the histograms of HOG and HOF.

3.2 Experiment protocols

To conduct the experiments, we split each data set into 10 non-overlapping folds, and use the 10-fold cross-validation to perform the training-testing procedure. Each fold is used as a test set in turn, and the remaining 9 folds are used as the training sets. The proposed algorithm is performed on the training set to obtain the view-conditional transformation matrices, and the classifier parameter. Then the learned view-conditional transformation matrices and the classifier parameter are used to represent and classify the data points in the test set. To handle the multiple class problem, we use the one-vs-all strategy.

3.3 Performance measures

To measure the classification performance over the test set, we use the classification accuracy. The classification accuracy is defined as follows,

$$\begin{aligned} &\text{Classification accuracy} \\ &= \frac{\text{Number of correctly classified test data points}}{\text{Number of total test data points}}. \quad (29) \end{aligned}$$

It is obvious that a better algorithm should be able to obtain a higher classification accuracy.

3.4 Experiment results

In this experiment, we first study the sensitivity of the algorithm to the parameters, which are α and γ .

3.4.1 Sensitivity to parameters

To study the performance of the proposed algorithm with different tradeoff parameters, α and β . We perform the algorithm by using the parameters of values 0.1, 1, 10, 100 and 1000, and measured the performance of different parameters. Figure 1 illustrates the performance on the PASCAL VOC 07 data set with respect to different tradeoff parameter α . The proposed algorithm achieves a

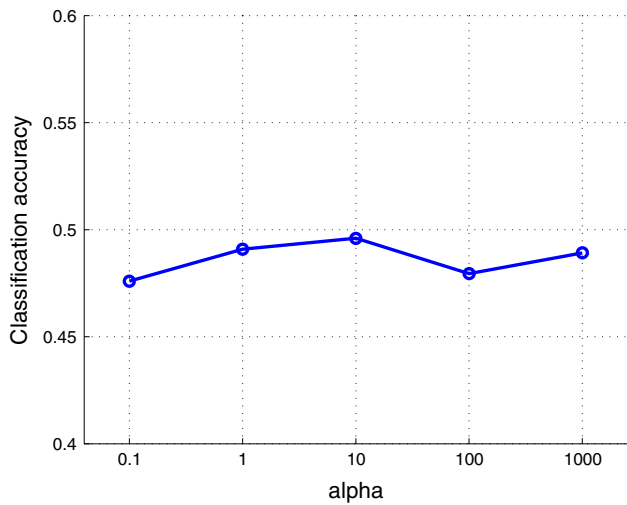


Fig. 1 Sensitivity curve of α over PASCAL VOC 07 data set

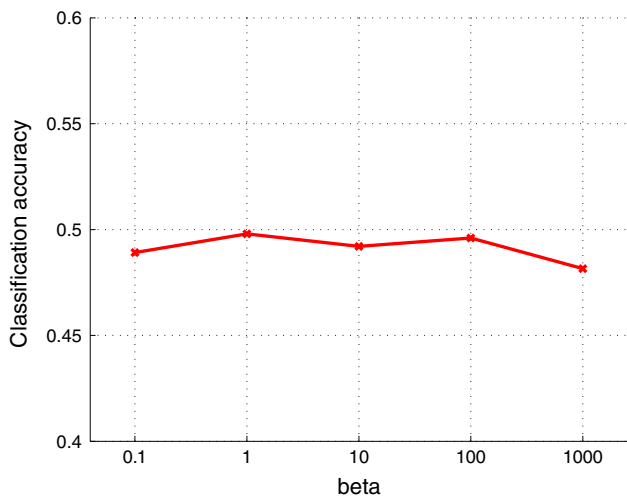


Fig. 2 Sensitivity curve of β over PASCAL VOC 07 data set

stable performance in all the settings of parameter α . In Fig. 2, the performance against different tradeoff parameter β is also shown. From this figure, we can also see that the algorithm is stable to the changes of value of β . This suggests that MISC is not sensitive to the changes of tradeoff parameters.

3.4.2 Comparison to state-of-the-art algorithms

We compare the proposed algorithm to the following methods, multiview learning algorithm using local learning (LL) proposed by Zhang et al. [43], multiview learning algorithm using co-training (CT) proposed by Sindhwani et al. [31], multiview learning algorithm based on view disagreement (VD) proposed by Quadrianto [30], multiview learning algorithm with global consistency and local

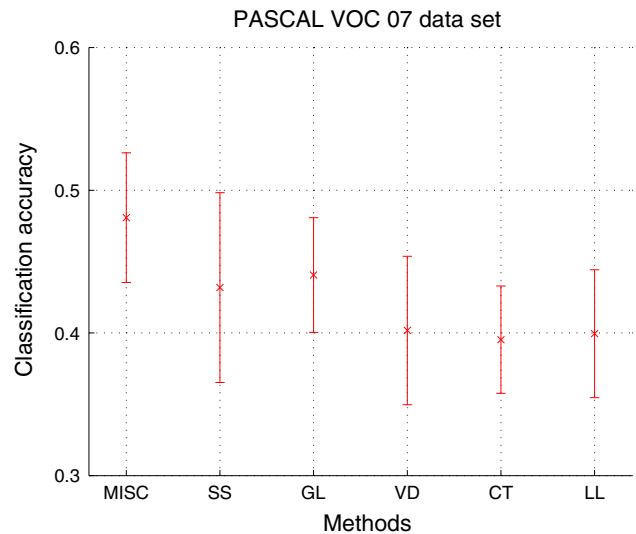


Fig. 3 Results of comparison of different algorithms over PASCAL VOC 07 data set

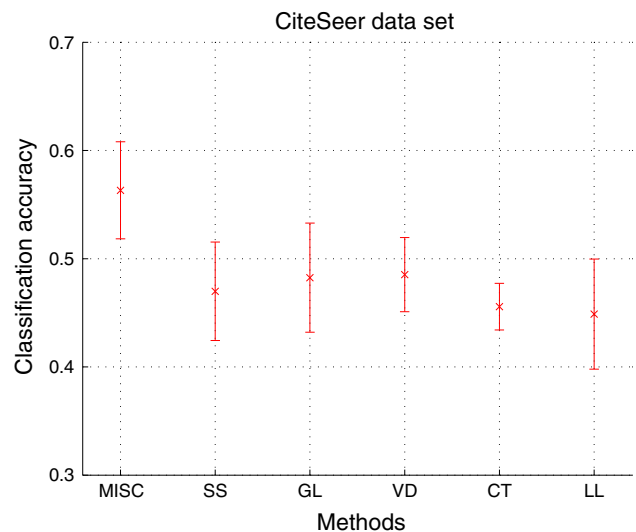


Fig. 4 Results of comparison of different algorithms over CiteSeer data set

smoothness (GL) proposed by Zhai [42], and multiview representation method using statistical subspace learning (SS) proposed by Chen et al. [2]. The error bars of the classification accuracy of the compared methods over three different data sets are given in Figs. 3, 4 and 5. From the figures, we find that the proposed method, MISC, stably outperforms other algorithms at all the data sets. Even on the most difficult data set, HMDB, the proposed method, MISC, also achieves an accuracy as high as about 0.4. The multiple view data are optimally combined by MISC to find the latent intact space and the optimal classifier in the corresponding intact space. The main reason for this is the

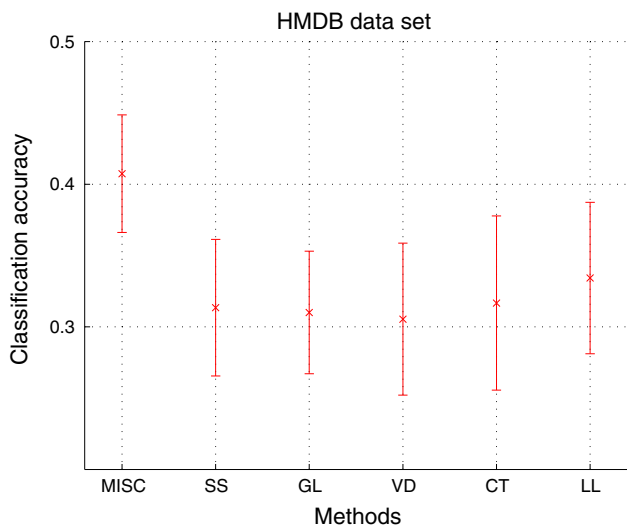


Fig. 5 Results of comparison of different algorithms over HMDB data set

robust property of the proposed algorithm. This algorithm has the ability to appropriately handle the complementary between multiple views, and learn a discriminative hidden intact space with help of classifier learning.

4 Conclusions

We propose a novel multiview learning algorithm by learning intact vectors of the training data points and a classifier in the intact space. The intact vectors are assumed to be a hidden but critical vector for each data point, and we can obtain its multiple view feature vectors by view-conditional transformations. Moreover, we also assume that the intact vectors are discriminative, i.e., can be separated by a linear function according to their classes. We propose a novel optimization problem to model both the learning of intact vectors and classifier. An iterative algorithm is developed to solve this problem. This algorithm outperforms other multiview learning algorithms on benchmark data sets, and it also shows its stability over tradeoff parameters.

Acknowledgments This project was supported by the National Natural Science Foundation of China (Grant No. 61472172), and a research funding of Ludong University (Grant No. 27870301).

References

- Charuvaka A, Rangwala H (2015) Convex multi-task relationship learning using hinge loss. In: IEEE SSCI 2014—2014 IEEE symposium series on computational intelligence—CIDM 2014: 2014 IEEE symposium on computational intelligence and data mining, Proceedings, pp 63–70
- Chen N, Zhu J, Sun F, Xing E (2012) Large-margin predictive latent subspace learning for multiview data analysis. *IEEE Trans Pattern Anal Mach Intell* 34(12):2365–2378
- Chen PT, Chen F, Qian Z (2015) Road traffic congestion monitoring in social media with hinge-loss markov random fields. In: 2014 IEEE international conference on data mining (ICDM), pp 80–89. doi:10.1109/ICDM.2014.139
- Chen YW, Wang JL, Cai YQ, Du JX (2015) A method for chinese text classification based on apparent semantics and latent aspects. *J Ambient Intell Humaniz Comput* 6(4):473–480
- Fakeri-Tabrizi A, Amini MR, Goutte C, Usunier N (2015) Multiview self-learning. *Neurocomputing* 155:117–127
- Feng G, Guo J, Jing BY, Sun T (2015) Feature subset selection using naive bayes for text classification. *Pattern Recogn Lett* 65:109–115. doi:10.1016/j.patrec.2015.07.028
- Feng Y, Xiao J, Zhuang Y, Liu X (2013) Adaptive unsupervised multi-view feature selection for visual concept recognition. In: Lee KM, Matsushita Y, Rehg JM, Hu Z (eds) *Computer Vision – ACCV 2012: 11th Asian conference on computer vision*, Daejeon, Korea, November 5–9, 2012, revised selected papers, Part I. Lecture notes in computer science, vol 7724. Springer, Berlin, Heidelberg, pp 343–357
- Gallagher C, Fisher T, Shen J (2015) A cauchy estimator test for autocorrelation. *J Stat Comput Simul* 85(6):1264–1276
- Gui J, Tao D, Sun Z, Luo Y, You X, Tang Y (2014) Group sparse multiview patch alignment framework with view consistency for image classification. *IEEE Trans Image Process* 23(7):3126–3137
- Hajmohammadi M, Ibrahim R, Selamat A (2014) Cross-lingual sentiment classification using multiple source languages in multiview semi-supervised learning. *Eng Appl Artif Intell* 36:195–203
- Hogenboom A, Frasincar F, De Jong F, Kaymak U (2015) Polarity classification using structure-based vector representations of text. *Decis Support Syst* 74:46–56
- Idan M, Speyer J (2014) Multivariate cauchy estimator with scalar measurement and process noises. *SIAM J Control Optim* 52(2):1108–1141
- Jaszewski M, Parameswaran S, Hallenborg E, Bagnall B (2015) Evaluation of maritime object detection methods for full motion video applications using the PASCAL VOC challenge framework. In: *Proceedings of SPIE—the international society for optical engineering*, vol 9407, p. 94070Y
- Jiang F, Jia L, Sheng X, LeMieux R (2015) Manifold regularization in structured output space for semi-supervised structured output prediction. *Neural Comput Appl* 1–10. doi:10.1007/s00521-015-2029-2
- Jiang Y, Liu J, Li Z, Lu H (2014) Semi-supervised unified latent factor learning with multi-view data. *Mach Vis Appl* 25(7):1635–1645
- Koopman B, Karimi S, Nguyen A, McGuire R, Muscatello D, Kemp M, Truran D, Zhang M, Thackway S (2015) Automatic classification of diseases from free-text death certificates for real-time surveillance. *BMC Med Inform Decis Mak*. doi:10.1186/s12911-015-0174-2
- Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) Hmdb: A large video database for human motion recognition. In: *Proceedings of the IEEE international conference on computer vision*, pp 2556–2563
- Kumar Nagwani N (2015) A comment on “a similarity measure for text classification and clustering”. *IEEE Trans Knowl Data Eng* 27(9):2589–2590
- La L, Wang N, Zhou DP (2015) Improving reading comprehension step by step using online-boost text readability classification system. *Neural Comput Appl* 26(4):929–939
- Li XX, Li RF, Feng FX, Cao J, Wang XJ (2014) Multi-view supervised latent dirichlet allocation. *Acta Electron Sin* 42(10):2040–2044

21. Liu J, Jiang Y, Li Z, Zhou ZH, Lu H (2015) Partially shared latent factor learning with multiview data. *IEEE Trans Neural Netw Learn Syst* 26(6):1233–1246
22. Liu X, Wang J, Yin M, Edwards B, Xu P (2015) Supervised learning of sparse context reconstruction coefficients for data representation and classification. *Neural Comput Appl* 1–9. doi:[10.1007/s00521-015-2042-5](https://doi.org/10.1007/s00521-015-2042-5)
23. Long J, Wang LD, Li ZD, Zhang ZP, Yang L (2015) Wordnet-based lexical semantic classification for text corpus analysis. *J Central South Univ* 22(5):1833–1840
24. Lu H, Hu Z, Gao H (2015) Multiview sample classification algorithm based on L1-graph domain adaptation learning. *Math Prob Eng.* doi:[10.1155/2015/329753](https://doi.org/10.1155/2015/329753)
25. Luo Y, Tao D, Xu C, Xu C, Liu H, Wen Y (2013) Multiview vector-valued manifold regularization for multilabel image classification. *IEEE Trans Neural Netw Learn Syst* 24(5):709–722
26. Mala K, Sadasivam V, Alagappan S (2015) Neural network based texture analysis of CT images for fatty and cirrhosis liver classification. *Appl Soft Comput J* 32:80–86
27. Mohanty A, Senapati M, Beberta S, Lenka S (2013) Texture-based features for classification of mammograms using decision tree. *Neural Comput Appl* 23(3–4):1011–1017
28. Petrov N, Georgieva A, Jordanov I (2013) Self-organizing maps for texture classification. *Neural Comput Appl* 22(7–8):1499–1508
29. Picard D, Gosselin PH, Gaspard MC (2015) Challenges in content-based image indexing of cultural heritage collections: support vector machine active learning with applications to text classification. *IEEE Signal Process Mag* 32(4):95–102
30. Quadrianto N, Lampert C (2011) Learning multi-view neighborhood preserving projections. In: *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp 425–432
31. Sindhwani V, Niyogi P, Belkin M (2005) A co-regularization approach to semi-supervised learning with multiple views. In: *Proceedings of ICML workshop on learning with multiple views*, pp 74–79
32. Sindhwani V, Rosenberg DS (2008) An RKHS for multi-view learning and manifold co-regularization. In: *Proceedings of the 25th international conference on machine learning*, pp 976–983. ACM
33. Sublemontier JH (2013) Unsupervised collaborative boosting of clustering: an unifying framework for multi-view clustering, multiple consensus clusterings and alternative clustering. In: *International joint conference on neural networks (IJCNN 2013)*, pp 1–8. doi:[10.1109/IJCNN.2013.6706911](https://doi.org/10.1109/IJCNN.2013.6706911)
34. Sun S (2013) A survey of multi-view machine learning. *Neural Comput Appl* 23(7–8):2031–2038
35. Wang JY, Wang Y, Jing BY, Gao X (2015) Regularized maximum correntropy machine. *Neurocomputing* 160:85–92
36. Wang Z, Sun X, Sun L, Huang Y (2014) Multiview discriminative geometry preserving projection for image classification. *Sci World J.* doi:[10.1155/2014/924090](https://doi.org/10.1155/2014/924090)
37. Williams K, Wu J, Choudhury S, Khabisa M, Giles C (2014) Scholarly big data information extraction and integration in the CiteSeer digital library. In: *Proceedings—international conference on data engineering*, pp 68–73
38. Wu TX, Lian XC, Lu BL (2012) Multi-view gender classification using symmetry of facial images. *Neural Comput Appl* 21(4):661–669
39. Yadav A, Anand R, Dewal M, Gupta S (2015) Multiresolution local binary pattern variants based texture feature extraction techniques for efficient classification of microscopic images of hardwood species. *Appl Soft Comput J* 32:101–112
40. Yu J, Rui Y, Tang Y, Tao D (2014) High-order distance-based multiview stochastic learning in image classification. *IEEE Trans Cybern* 44(12):2431–2442
41. Zha ZJ, Yang Y, Tang J, Wang M, Chua TS (2015) Robust multiview feature learning for RGB-D image understanding. *ACM Trans Intell Syst Technol.* doi:[10.1145/2735521](https://doi.org/10.1145/2735521)
42. Zhai D, Chang H, Shan S, Chen X, Gao W (2012) Multiview metric learning with global consistency and local smoothness. *ACM Trans Intell Syst Technol* 3(3):1–22
43. Zhang D, Wang F, Zhang C, Li T (2008) Multi-view local learning. In: *Proceedings of the national conference on artificial intelligence*, vol 2, pp 752–757
44. Zhao X, Evans N, Dugelay JL (2013) Unsupervised multi-view dimensionality reduction with application to audio-visual speaker retrieval. In: *Proceedings of the 2013 IEEE international workshop on information forensics and security, WIFS 2013*, pp 7–12