Simultaneously Combining Multi-View Multi-Label Learning with Maximum Margin Classification

Zheng Fang
Dept. of Information Science and Electronic Engineering
Zhejiang University, China
fangzheng354@zju.edu.cn

Zhongfei (Mark) Zhang
Dept. of Information Science and Electronic Engineering
Zhejiang University, China
zhongfei@zju.edu.cn

Abstract—Multiple feature views arise in various important data classification scenarios. However, finding a consensus feature view from multiple feature views for a classifier is still a challenging task. We present a new classification framework using the multi-label correlation information to address the problem of simultaneously combining multiple feature views and maximum margin classification. Under this framework, we propose a novel algorithm that iteratively computes the multiple view feature mapping matrices, the consensus feature view representation, and the coefficients of the classifier. Extensive experimental evaluations demonstrate the effectiveness and promise of this framework as well as the algorithm for discovering a consensus view from multiple feature views.

Keywords—feature mapping; multi-view learning; label dependence maximization; maximum margin classification; consensus representation;

I. INTRODUCTION

In many real-world applications, one object usually has different representations in the form of multiple views and relates to multiple concepts simultaneously. For an example, an image can be related to multiple concepts given as labels, such as bear, grass, and field. At the same time, this image can also be described by different types of features, such as color, texture, shape, and the text surrounding the image.

How to effectively make use of the multi-view features and the label correlation information in the classification has received intensive attention in the literature.

In many data mining applications, though individual feature views might be sufficient on their own for a given learning task, they can often provide complementary information to each other resulting in an improved performance on the learning task [1]–[3]. However, it is still a challenging problem to leverage the consistency between different views in data mining that can be explicitly exploited as the correspondence information hidden in the original multi-view features at the same time taking advantage of the provided complementary information. The basic idea of multi-view learning is to make use of the consistency among different views and complementary information between the views to achieve a better performance [4], [5]. Most of the existing algorithms ignore such correspondence information, and adopt originally provided features for the classifier design leading to a suboptimal solution.

Meanwhile, multi-label classification has received an increasing attention in recent years [6]–[8]. Different from the classic classification approaches where classes are always assumed to be mutually exclusive, in multi-label classification, class concepts given as the labels are normally interdependent upon one another, which violates the fundamental assumption of the mutual exclusiveness in the classic classification literature. Clearly, directly employing the existing multi-view learning methods to multi-label learning would ignore the label correlation information and thus would fail to incorporate the dependence knowledge between the multi-labels and the multi-view features.

Based on these observations, in this paper, we propose a unified framework, Multi-Label Multi-View Learning with Maximum Margin Classification (MLMVL-MM), which learns the latent feature mapping among multi-views as well as conducts a multi-class classification simultaneously with the correlation information explicitly exploited between the class labels.

In particular, we focus on a feature mapping among multiple views to lead to a consensus data representation for the shared common structure across the multiple data descriptions from the different sources. Moreover, a set of binary classifiers constructed for each label respectively are learned with this shared underlying feature structure to discriminate the corresponding label from the rest. Finally, we develop an learning algorithm using an alternating iterative algorithm for discovering the latent consensus data structure, the feature mapping for different views, and the coefficients of the classifiers. Extensive experiments show that the framework achieves a superior performance to the state-of-the-art methods.

The key contributions of this paper are highlighted as follows:

• We develop a general framework MLMVL-MM for learning intrinsic consensus representation for multi-view features in multi-label classification. In this framework, the correlation information among multiple labels is captured by integrating the label dependence maxi-
We show that when the multi-view feature learning constraint and the label dependence maximization criterion are integrated into the classification, the intrinsic consensus data representation, together with the feature mapping and the coefficients of the classifiers can be computed by an effective alternating iterative algorithm.

We report the extensive experiments on two benchmark datasets for the multi-label, multi-view classification to demonstrate the effectiveness and promise of the proposed framework.

II. MULTI-LABEL MULTI-VIEW LEARNING WITH MAXIMUM MARGIN CLASSIFICATION

In this section, we first introduce the problem statement and the data notations. Then, the general framework named MLMVL-MM is described, which integrates the multi-label correlations and multi-view features for maximum margin classification simultaneously. Based on an instantiation of the framework, we develop an optimization method, which achieves the optimum value of the object function iteratively under the MLMVL-MM framework.

A. Data Representations and Notations

For the multi-label, multi-view classification task, we are given \( n \) labeled data instances from \( V \) independent views: \( \{ (x_{i}^{v}, y_{i}^{v}) \} \), \( x_{i}^{v} \in \mathbb{R}^{d_{v} \times 1} \) and \( y_{i}^{v} \in \{ -1,+1 \} \). \( i \in \{ 1,2,\cdots,n \} \), \( v \in \{ 1,2,\cdots,V \} \), \( l \in \{ 1,2,\cdots,L \} \), where \( L \) is the total number of the given class labels and \( d_{v} \) is the dimensionality of the \( v \)th view. \( y_{i}^{v} \) is the \( i \)th class label of \( x_{i}^{v} \). We denote the matrix \( X^{v} = [x_{1}^{v},\cdots,x_{n}^{v}] \), \( Y^{v} \) as \( Y_{ii} = y_{i}^{v} \) and \( Y_{i} = Y - (-1, +1)^{T} \), and the \( i \)th class label vector for the data instances as \( y_{i}^{v} \), where \( y_{i}^{v} = 1 \).

As the one-against-rest scheme commonly applied for multi-class classification, we decompose the multi-label classification into a series of single-class classifications by constructing one binary classifier for each label in which instances relevant to this label form the positive class, and the rest instances form the negative class. Hence, the goal of MLMVL-MM is to construct an accurate classifier for each class label.

Taking into account the label correlations and the redundancy incurred by multiple views, MLMVL-MM makes use of the label correspondence information and combines every feature from different views to obtain a consensus, discriminative, and low dimensional feature representation for a maximum margin classifier. As an appropriate integration of the feature consensus mapping and the maximum margin classification, MLMVL-MM makes an effective use of the label correlation knowledge to deliver an enhanced classification performance with the promise to outperform the existing literature.

B. Perspectives and Problem Statement

1) Consensus Data Description for Multiple Views:
The problem of combining multiple feature views to obtain a consensus data description can be stated as follows. Given a set of data representations from different views, \( C = \{ X^{1},X^{2},\cdots,X^{V} \} \), where \( X^{1} \in \mathbb{R}^{d_{1} \times n} \), \( X^{2} \in \mathbb{R}^{d_{2} \times n} \), \( \cdots \), \( X^{v} \in \mathbb{R}^{d_{v} \times n} \), and a predefined number \( d_{g} \), generate a low dimensional data description as a final consensus \( G \in \mathbb{R}^{d_{g} \times n} \) using a consensus function.

We use Euclidean distance to provide a new consensus function for combining the multiple views. Hence, we define the problem of combining multiple views as an optimization problem below.

Definition 2.1: Given a set of data representations from different views, \( C = \{ X^{1},X^{2},\cdots,X^{V} \} \), and \( d_{g} \in \mathbb{Z}^{+} \), a consensus data description represented by \( G \in \mathbb{R}^{d_{g} \times n} \) and the \( V \) feature mapping matrices \( P^{1} \in \mathbb{R}^{d_{g} \times d_{1}},\cdots, P^{v} \in \mathbb{R}^{d_{g} \times d_{v}} \) are learned by the minimization of the object function:

\[
 f(G,P^{1},\cdots,P^{v}) = \sum_{i=1}^{V} \| G - P^{v} X^{v} \|^{2}
\]

It is clearly noted that there is an obvious trivial zero solution to the object function, on the condition that there are no other constraints imposed to this function.

Hence, the object function must be integrated with other object functions and constraints given in the following section to achieve an optimum solution.

2) Hilbert-Schmidt Independence Criterion: Motivated by the consideration that there should exist relations between the feature description and the labels associated with the same object, we attempt to find a lower dimensional feature space in which the dependence between the features and the labels are maximized. Considering a linear projection \( P \) the instance \( x_{i} \) is projected into the new consensus space \( \tilde{G} \) by \( \phi(x_{i}) = P x_{i} \). Then, we attempt to maximize the dependence between the feature description \( \phi(x_{i}) \in \tilde{G} \) and the class labels \( Y_{i} \). Here we adopt the Hilbert-Schmidt Independence Criterion (HSIC) [8]:

\[
 \text{HSIC}(\tilde{G},Y,P_{xy}) = (n - 1)^{-2} \text{tr}(HKHL),
\]

where \( H = I_{n} - \frac{1}{n}ee^{T} ; e \) is an all-one column vector. \( K = [K_{ij}]_{n \times n} \) and \( L = [L_{ij}]_{n \times n} \) are the matrices of the inner product of instances in \( \tilde{G} \) and \( Y \), respectively. \( K_{ij} = \langle \phi(x_{i}),\phi(x_{j}) \rangle = \langle Px_{i},Px_{j} \rangle \), \( L_{ij} = \langle Y_{i},Y_{j} \rangle \). In our case, the HSIC criterion is rewritten as:

\[
 \text{HSIC}(\tilde{G},Y,P_{xy}) = (n - 1)^{-2} \text{tr}(HX^{T}P^{T}PXHY^{T}Y)
\]

3) MLMVL-MM Framework: In this subsection, we propose a general large margin framework MLMVL-MM, which simultaneously finds a consensus feature description for the \( V \) data representations \( X^{1},\cdots,X^{V} \) from different views and fully exploits the inherent relationship between multi-view features and class labels. For simplicity and without losing the generality, we take the \( i \)th classifier into account for an instance.
First, we employ the consensus function defined in (1) to learn the low dimensional feature representation. In order to capture the intrinsic correspondence information between the different views and the class labels, we incorporate the HISC criterion as a regularization. Maximum margin classifiers are then trained on the learned consensus data description. Finally, the objective functions can be integrated into an unified framework for the optimization. In this paper, we employ the dual kernel version of the traditional maximum margin classifiers and the unified optimization problem is as follows.

\[
\min_{G,(P^v)} \beta \sum_{i=1}^{V} \|G - P^v X^v\|^2 + \lambda \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_j k(g_i, g_j) \\
- \eta \sum_{i=1}^{V} \text{tr}(H(P^v X^v)^T (P^v X^v) H Y^T Y) \\
+ \sum_{i=1}^{n} L(y_i, \sum_{j=1}^{n} k(g_j, g_i) \beta_j)
\]  

(3)

where \( \gamma, \lambda, \eta \geq 0 \) are trade-off regularization parameters.

III. DERIVATION OF THE ALGORITHM

By the reasoning given above, we are essentially to transform the classification task from the domain \( \{X, Y\} \) to domain \( \{G, Y\} \), while at the same time to find the projection matrices \( \{P_v^v\}_{v=1}^{V} \) for each of all the feature views. In the framework, the loss function is chosen to be of the form \( L(y_i, f(x_i)) = \max(0, 1 - y_i f(x_i))^2 \). Although the cost function is not jointly convex for all the unknowns and it is unrealistic to expect an algorithm to find the global minimum, for any one unknown, with all the other unknowns held as constants, the cost function is a convex quadratic function. We derive an EM style algorithm that converges to a local minimum by iteratively updating the feature mapping matrices \( \{P^v\}_{v=1}^{V} \), the consensus data description matrix \( G \), and the parameters \( \beta \) of the classifiers using a set of multiplicative updating rules.

A. Optimizing \( P^v \) with \( G \) and \( \beta \) Fixed

We show that the optimal \( P^v \) that solves the optimization problem in (3) can be expressed in terms of \( G \), as summarized in the following theorem:

**Theorem 3.1**: Denoting \( N = X^v (I - \frac{1}{n-1} M) (X^v)^T \), \( M = H Y^T Y H \), \( U = G(X^v)^T \), the objective function \( F(P) \) in (3) is nonincreasing under the updating rule,

\[
p_k^v = \frac{u_k - P_k^v n_k}{N_{kk}} + P_k^v
\]

(4)

where \( P_k^v = P_k n_k \), \( u_k = U n_k \) are the column vectors of matrices \( P^v \) and \( U \), respectively.

**Proof**: Since the updating rules for all \( P^v \) are the same, for convenience, we simplify the problem to the case with one view only and consider the items containing \( P \).

\[
F(P) = \|G - PX\| - \frac{\eta}{(n-1)^2} \text{tr}(H(PX)^T (PX) HY^T Y)
\]

After a straightforward mathematical derivation and expanding the object function \( F(P) \) with \( P_k \), we finally obtain the item containing \( P_k \):

\[
F(p_k) = p_k^T p_k N_{kk} + 2 \sum_{j \neq k} p_j^T p_j N_{jk} - 2 p_k^T u_k
\]

(5)

Taking the derivative with respect to \( p_k \) and setting \( \frac{\partial F(p_k)}{\partial p_k} = 0 \), we obtain:

\[
p_k = \frac{u_k - P_n k}{N_{kk}} + p_k
\]

(6)

The proof is completed.

B. Optimizing \( \beta \) with \( P^v \) and \( G \) Fixed

We employ the paradigm in [9] to obtain the update rule of \( \beta \). Writing the kernel matrix \( K \), such that \( K_{ij} = k(g_i, g_j) \), and \( k_t \) the \( t \)-th column of \( K \), and taking only the items containing \( \beta \) into account, we obtain the objective function in the following form:

\[
F(\beta) = \lambda \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_j k(g_i, g_j) + \sum_{i=1}^{n} L(y_i, k_i^\top \beta) + \text{const}
\]

\[
- \lambda \beta^T K \beta - 2 \beta^T (K Y^T Y) + \beta^T (K Y^T K) \beta
\]

Here we use Newton Raphson method to update \( \beta \):

\[
\beta^{new} = \beta^{old} - \mu H^{1/2} \nabla \beta^{old}
\]

(7)

We find that the value of \( \mu \) does not affect the convergence, and thus set the default value of \( \mu = 1 \). We obtain the final update formula for \( \beta \) as follows.

\[
\beta = \begin{bmatrix} (\lambda I_{n_{sv}} + K_{sv})^{-1} Y_{sv} \\ 0 \end{bmatrix}
\]

(8)

If the bias term \( \beta_0 \) is incorporated into the function \( f(x) \): \( f(x) = \sum_{i=1}^{n} \beta_i k(x, x) + \beta_0 \), we solve the following system of linear equations to obtain \( \beta \) and \( \beta_0 \):

\[
\begin{bmatrix} (\lambda I_{n_{sv}} + K_{sv}) & 1 \\ 1^T & 0 \end{bmatrix} \begin{bmatrix} \beta \\ \beta_0 \end{bmatrix} = \begin{bmatrix} Y_{sv} \\ 0 \end{bmatrix}
\]

(9)

C. Optimizing \( G \) with \( P^v \) and \( \beta \) Fixed

Fixing \( \{P^v\}_{v=1}^{V}, \beta \), the objective function in (3) is quadratic with respect to \( G \). We derive an iterative updating rule for \( G \) based on the bound optimization procedure [10], [11]. The basic idea is to construct an auxiliary function which is a convex upper bound for the original objective function based on the solution obtained from the previous iteration. Then, a new solution to the current iteration is obtained by minimizing this upper bound. The definition of the auxiliary function and a useful lemma [11] are quoted as follows.

**Definition 3.1**: Given an objective function \( F(g) \), \( Q(g, g') \) is an auxiliary function for \( F(g) \) if the conditions below hold true:

\[
(i) \quad Q(g, g) = F(g); \quad (ii) \quad Q(g, g') \geq F(g)
\]
Lemma 3.2: If \( Q(g, g^t) \) is an auxiliary function for \( F(g) \), then \( F(g) \) is non-increasing under the following update rule:

\[
g^{t+1} = \arg \min_g Q(g, g^t)
\]

where \( g^t \) denotes the current estimation of the model parameter and \( g^{t+1} \) is the new estimation to maximize \( Q(g, g^t) \).

**Proof:** \( F(g^{t+1}) \leq Q(g^{t+1}, g^t) \leq Q(g^t, g^t) \leq F(g^t) \).

Denote the sum of items containing the \( k \) column of \( G \), \( g_k \), in objective function (3) as \( F(g_k) \). We propose an auxiliary function for \( F(g_k) \) in the following theorem.

**Theorem 3.3:** Let \( Q(g_k, g_k^t) \) has the following form:

\[
Q(g_k, g_k^t) = F(g_k^t) + (g_k - g_k^t)^T \nabla_{g_k^t} + \frac{1}{2} (g_k - g_k^t)^T D(g_k^t)(g_k - g_k^t)
\]

where \( D(g_k^t) \) is a diagonal matrix defined as:

\[
D_{i,i}(g_k^t) = \delta_{ik} (2(\gamma V + \lambda \beta_0^2) + 2((q_k - y_k)^2 + \beta_k^2)[k \in n_{svv}])g_{ik}^t + 2(G\beta)(G\beta)^T g_k^t[k \in n_{svv}] + 2\beta_k^2 G_{svv} G_{svv}^T g_{ik}^t)
\]

and \( \nabla_{g_k} \) is the gradient of \( F(g_k^t) \) taking derivative with \( g_k^t \).

\[
\nabla_{g_k^t} = \sum_{v=1}^V (2g_v - 2P^v x_v^t) + \lambda \sum_{j=1}^n 2\beta_k \beta_j g_{j}
\]

\[
+ 2 \left( \left( \sum_{j=1}^n (q_k - y_k) \beta_j g_{j} \right) k \in n_{svv} \right) + \beta_k \sum_{i=1}^n (q_i - y_k) g_{ik}^t
\]

\[
q_i = k^j \beta = \sum_{j=1}^n g_{j}^T g_{j}, \quad G_{svv} G_{svv}^T = \sum_{i=1}^n G_{si} G_{si}^T
\]

Then \( Q(g_k, g_k^t) \) is an auxiliary function for \( F(g_k) \).

Theorem 3.3 can be proven similarly to that in [11] by validating \( Q(g_k, g_k^t) \geq F(g_k) \). Due to the limitation of space, we omit details of the \( F(g_k) \) and the validation here. Based on Theorem 3.3, we minimize \( Q(g_k, g_k^t) \) with respect to \( g_k \) with \( g_k^t \) fixed. Set \( \nabla_{g_k} Q(g_k, g_k^t) = 0 \). We obtain the following updating formula for every \( g_k \).

\[
g_k^{t+1} = g_k^t - D(g_k^t)^{-1} \nabla_{g_k^t}
\]

Note that the update algorithm to the optimization is similar to Newton's method for maximizing \( F(g_k) \) by replacing the Hessian matrix at each iteration with the derived diagonal matrix \( D(g_k^t) \). Thus, the expensive computation of the inverse of Hessian matrix is avoided.

### IV. Experiments

In this section, we evaluate MLMVL-MM on real-world image datasets for multi-label, multi-class classification.

A. Datasets and Evaluation Metric

Two public image datasets NUSWIDE [12] and Corel5K [13] are used in the study. We use the area under the receiver operating characteristic (ROC) curve, called AUC, the classification accuracy (ACC), precision, recall, and microF1 score as the performance measures.

B. Baseline Methods

We compare MLMVL-MM with six peer algorithms: CCA+Ridge, CCA+SVM, MDDM+SVM, CTVF+SVM, MLLS [6], and SVM2K [14]. Of the six methods, the four algorithms, CCA-SVM, MDDM+SVM, CTVF+SVM, and SVM2K adopt the one-vs-rest method and perform binary classification for each label separately. For details of CCA-ridge, CCA-SVM, CTVF+SVM, and MLLS please refer to [6], MDDM+SVM applies MDDM [8] first to reduce the data dimensionality before the linear SVM is applied. The regularization parameters in each algorithms are all tuned using a five-fold cross-validation.

C. Experiments Setup and Performance Evaluation

The parameters learned during the training phase are the feature mapping matrices \( \{P_v\}_{v=1}^V \), the consensus data description matrix \( G \), and the parameters \( \beta, \beta_0 \) of the classifiers. During the testing phase, we perform a feature combination of the test data, with the feature mapping \( \{P_v\}_{v=1}^V \) obtained from the training phase to generate \( \{G_v^t\}_{v=1}^V \) for the different view sources. Once \( \{G_v^t\}_{v=1}^V \) is obtained we have the final consensus data description through combining these new presentations, \( G_{test} = \sum_{v=1}^V G_v^t V \) (this procedure is equivalent to choosing the central point of all the mapped data points as the final representation) and generate the kernel matrix \( K_{test} = G^T G_{test} \). The classification of the test data points can now be obtained by the function \( sign(K_{test}^t \beta + \beta_0) \).

For each dataset, we use the extracted features from two views. For the NUSWIDE dataset, SIFT descriptors and text words surrounding the images are selected as the \( X^1 \) and \( X^2 \) feature views. For the Corel5K dataset, \( X^1 \) and \( X^2 \) correspond to SIFT descriptions and HUE color descriptions.

For MLMVL-MM, we set the reduced dimensionality on the set \{100, 200, 300, 400, 500\} using a five-fold cross validation. The other regularization parameters \( \gamma, \lambda, \eta \) are tuned from the candidate sets, \{0.5, 1, 5, 10, 50, 100\}, \{0.5, 5, 50, 100, 250, 500\}, and \{0.5, 1, 5, 10, 50, 100, 150\}, respectively.

To begin with, we randomly generate a training set for each of the two datasets. Tables I documents the performance results of the six methods in terms of ACC, AUC, and microF1 when we select 60 percent of the positive examples from each class and randomly sample the same number of the negative examples from the rest concept classes to form the training set for the classifiers of the corresponding class. The average performance and standard deviations from the
To further investigate the convergence rate of MLMVL-MM, we show the three performance measure results for NUSWIDE dataset at different iteration numbers in Figure 1(a). From Figure 1(a), all the three measures indicate consistently that MLMVL-MM converges in 10 iterations.

Finally, in order to evaluate the MLMVL-MM learning effectiveness w.r.t. the training data size, MLMVL-MM for each label is trained separately on the training data with different percentages of the dataset, i.e., \{50\%, 60\%, 70\%, 80\%\} for the NUSWIDE dataset. This procedure is repeated for 5-rounds by randomly sampling 5 times for each size of the training data. From Figure 1(b) we observe that MLMVL-MM delivers a better learning performance with more training samples.

From these extensive evaluations, we have the following observations. MLMVL-MM performs the best in all the cases. The reason is that MLMVL-MM models both the consistency between different views and the label dependence information for a discriminatively maximum margin classification simultaneously, whereas the existing methods ignore this useful information (i.e., the underlying structure from different feature views, the correlation information between concept labels and data features, and the redundancy incurred by multiple feature views for the classification).

Compared with CTVF+SVM, it is clear that concatenating the features from different views may not necessarily result in an increase in the classification performance though SVM uses the information from both views. Compared with the existing multi-view learning methods (SVM2K, CCA+Ridge) MLMVL-MM performs better because it explicitly exploits the inherent consensus feature structure underlying the different feature views and discriminatively incorporates the label dependence information in the maximum margin classification. Compared with the existing shared subspace extracting method MLLS and the dimensionality reduction methods based on the label correlation information (CCA+SVM, MDDM+SVM), MLMVL-MM is able to supervise the feature mapping according to the label dependence information. Consequently, it leads to a more discriminative subspace for the subsequent maximum margin classification. Moreover, the updating of G also takes the discriminative maximum margin information into account which in turn helps learn to achieve a more desirable consensus data representation for classification through updating \(\{P\}_{v=1}^{V}\). We also believe that these traditional feature merging methods may not work well in the cases when different kinds of feature views are provided for multi-label classification. Therefore, MLMVL-MM is able to achieve a better performance in all the cases.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>NUSWIDE dataset</th>
<th>Corel5K dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACC</td>
<td>AUC</td>
</tr>
<tr>
<td>CTVF + SVM</td>
<td>0.6667 ± 0.0226</td>
<td>0.7124 ± 0.0381</td>
</tr>
<tr>
<td>CCA + SVM</td>
<td>0.6499 ± 0.0261</td>
<td>0.6937 ± 0.0382</td>
</tr>
<tr>
<td>CCA + Ridge</td>
<td>0.6669 ± 0.0440</td>
<td>0.7134 ± 0.0355</td>
</tr>
<tr>
<td>MDDM + SVM</td>
<td>0.6708 ± 0.0253</td>
<td>0.7169 ± 0.0377</td>
</tr>
<tr>
<td>SVM2K</td>
<td>0.6099 ± 0.0264</td>
<td>0.6921 ± 0.0251</td>
</tr>
<tr>
<td>MLMLS</td>
<td>0.7065 ± 0.0349</td>
<td>0.7669 ± 0.0331</td>
</tr>
<tr>
<td>MLMVL-MM</td>
<td>0.7173 ± 0.0252</td>
<td>0.7916 ± 0.0343</td>
</tr>
</tbody>
</table>

Figure 1. (a) ACC, AUC, and microF1 score of MLMVL-MM at different iteration numbers on the NUSWIDE dataset. (b) ACC, AUC, and microF1 score versus different sizes of the training data for the NUSWIDE dataset by MLMVL-MM.

Table I

Summary of the performance for the 7 comparison methods on the NUSWIDE and Corel5K datasets in terms of ACC, AUC, and micro F1. All the parameters of the 7 methods are tuned by cross-validations, and the averaged performance over 5 random samplings of the training instances is reported. The highest performance value is highlighted in each case.
Recent literature has proven that the multi-view, multi-label learning problem exists everywhere and it is still open and in fact in a high demand to develop an effective solution to this problem. This paper studies this problem and has developed the only and first solution to this problem in the literature that simultaneously combines the multi-view, multi-label learning with maximum margin classification. Theoretical analysis and extensive empirical evaluations have demonstrated that the general framework MLMVL-MM developed in this study is capable of delivering a superior learning performance to several well-known solutions recently developed in the literature to the multi-label, multi-class classification problem.

Acknowledgment

This work is supported in part by National Basic Research Program of China (2012CB316400), ZJU–Alibaba Financial Joint Lab, and Zhejiang Provincial Engineering Center on Media Data Cloud Processing and Analysis. ZZ is also supported in part by US NSF (IIS-0812114, CCF-1017828).

References


