Multi-View Visual Classification via a Mixed-Norm Regularizer

Xiaofeng Zhu[1], Zi Huang[1], and Xindong Wu[2,3]

[1] School of Information Technology & Electrical Engineering, The University of Queensland,
Brisbane, QLD4072, Australia
{zhux,huang}@itee.uq.edu.au
[2] School of Computer Science and Information Engineering,
Hefei University of Technology, China
[3] Department of Computer Science, University of Vermont, Burlington, Vermont 05405, USA
xwu@uvm.edu

Abstract. In data mining and machine learning, we often represent instances by multiple views for better descriptions and effective learning. However, such comprehensive representations can introduce redundancy and noise. Learning with these multi-view data without any preprocessing may affect the effectiveness of visual classification. In this paper, we propose a novel mixed-norm joint sparse learning model to effectively eliminate the negative effect of redundant views and noisy attributes (or dimensions) for multi-view multi-label (MVML) classification. In particular, a mixed-norm regularizer, integrating a Frobenius norm and an $\ell_{2,1}$-norm, is embedded into the framework of joint sparse learning to achieve the design goals, which include selecting significant views, preserving the intrinsic view structure and removing noisy attributes from the selected views. Moreover, we devise an iterative algorithm to solve the derived objective function of the proposed mixed-norm joint sparse learning model. We theoretically prove that the objective function converges to its global optimum via the algorithm. Experimental results on challenging real-life datasets show the superiority of the proposed learning model over state-of-the-art methods.

Keywords: Feature selection, Joint sparse learning, Manifold learning.

1 Introduction

In many real-world applications in data mining and machine learning, data are naturally described by multiple views [7]. For example, in document analysis, web pages can be represented by their content or the content of the pages pointing to them; In bioinformatics, genes can be described with the feature space (corresponding to the genetic activity under the biological conditions) as well as the term space (corresponding to the text information related to the genes); Images are represented by different kinds of low-level visual features, such as color histograms, bags of visual words, and so on.

Actually, different views describe different aspects of data. No one among them is absolutely better than others for describing the data [10]. Thus, a good alternative is to simultaneously employ multiple views to learn the data. This is well known as multi-view learning [2]. Multi-view learning has been shown to be more effective than single-view learning, particularly in the scenario where the weaknesses of a single view can be

J. Pei et al. (Eds.): PAKDD 2013, Part I, LNAI 7818, pp. 520–531, 2013.

strengthened by others [14]. For example, in content analysis, color features have been shown to be sensitive to scaling, while SIFT features are robust to scaling. Combining color and SIFT features to perform multi-view learning can boost the performance by complementing each other's robustness on different aspects of the data.

Meanwhile, existing multi-view learning methods have several limitations. First, not all views of data are useful for some specific learning tasks since some of them may be redundant. However, existing methods are often designed to learn from all views of the data, without taking the redundancy issue into account. For example, canonical correlation analysis (CCA) and its kernel edition KCCA (e.g., [4,16]) were designed to learn a common latent feature space by learning from all views of the data. Second, multi-view data often contain noise, which easily affects the effectiveness of learning tasks while learning from all views of the data. Third, the intrinsic group structure of each individual view (i.e., view structure) in the data should also be preserved. Given multi-view data, each view of the data is a natural group to describe an aspect of the data. For example, a color histogram feature naturally forms a group for describing the color characteristics of image data.

Given that data are often represented by multiple views and associated with multiple object categories, this paper focuses on the problem of visual classification with multi-view multi-label (MVML) learning. In this paper, we propose a novel mixed-norm joint sparse learning model, which aims to select representative views and remove noisy attributes for MVML classification. More specifically, we first employ a least square loss function measured via a Frobenius norm (or F-norm in short) in each view to pursue a minimal regression error across all the views. We then introduce a new mixed-norm regularizer (i.e., combining an F-norm with an $\ell_{2,1}$-norm) to avoid redundant views and preserve the intrinsic view structure via the F-norm regularizer, and remove noisy attributes via the $\ell_{2,1}$-norm regularizer. We further devise a novel iterative algorithm to efficiently solve the objective function of the proposed mixed-norm joint sparse learning model, and then theoretically prove that the proposed algorithm enables the objective function to converge to its global optimum. After performing the iterative algorithm, the derived regression coefficient matrix only contains a few non-zero rows in a few selected views due to the mixed-norm regularizer. This makes the test process more efficient. Finally, we conduct an extensive experimental study on real-life datasets to compare the effectiveness of the proposed learning model with state-of-the-art methods for MVML classification.

We summarize the contributions of this paper as follows:

- We identify limitations in traditional multi-view learning, mainly caused by redundant visual features and noisy attributes. In this paper, we devise an effective solution to tackle the limitations via the proposed mixed-norm joint sparse learning model, which can be applied to many real-world applications, such as MVML visual classification.
- The proposed model focuses on embedding a mixed-norm regularizer into the existing joint sparse learning framework. We solve the derived objective function by a simple yet efficient optimization algorithm, which theoretically guarantees that the object ive function converges to its global optimum.

- To perform MVML classification, the proposed model can be regarded as simultaneously performing two types of feature selection, i.e., view-selection and attribute-selection respectively. View-selection aims to discard redundant views and preserve the intrinsic view structure via the F-norm regularizer, while attribute-selection aims to select useful attributes in the selected views of the data via the $\ell_{2,1}$-norm regularizer. These two types of feature selection lead to a new mixed joint sparsity, i.e., the view sparsity and the row sparsity simultaneously. To the best of our knowledge, no research efforts have been proposed on performing two types of feature selection in the joint sparse learning framework. Moreover, extensive experimental results on the public datasets show that the proposed model is more effective than state-of-the-art methods.

2 Approach

In this paper, ℓ_p-norm of a vector $\mathbf{v} \in \mathbb{R}^n$ is defined as $\|\mathbf{v}\|_p = \left(\sum\limits_{i=1}^{n} |v_i|^p\right)^{\frac{1}{p}}$, where v_i is the i^{th} element of $\mathbf{v}$. $\ell_{r,p}$-norm over a matrix $\mathbf{M} \in \mathbb{R}^{n \times m}$ is defined as $\|\mathbf{M}\|_{r,p} = \left(\sum\limits_{i=1}^{n}\left(\sum\limits_{j=1}^{m} \|m_{ij}\|^r\right)^{\frac{p}{r}}\right)^{\frac{1}{p}}$, where m_{ij} is the element of the i^{th} row and j^{th} column. The transpose of $\mathbf{X}$ is denoted as $\mathbf{X}^T$, the inverse of $\mathbf{X}$ is $\mathbf{X}^{-1}$, and the trace operator of a matrix is denoted by the symbol "tr".

2.1 Loss Function

Given the g-th view $\mathbf{X}^g$ of the training data $\mathbf{X}$, we need to obtain its regression coefficients $\mathbf{W}^g$. In MVML learning, we wish to obtain the minimal difference between the training label $\mathbf{Y} = [\mathbf{Y}_1^T, ..., \mathbf{Y}_n^T]^T$ and the summation of all G views on the multiplication between $\mathbf{X}^g$ and $\mathbf{W}^g$, i.e., $\sum\limits_{g=1}^{G} \mathbf{X}^g\mathbf{W}^g$. Therefore, a least square loss function can be defined as follows:

$$\min_{\mathbf{W}} \|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 \tag{1}$$

where $\mathbf{X} = [\mathbf{X}^1, ..., \mathbf{X}^g, ..., \mathbf{X}^G]$. Obviously, Eq.1 meets our goal of minimizing the regression error across all views.

2.2 Mixed-Norm Regularizer

Given a loss function (such as in Eq.1), during the optimization process we also design a mixed-norm regularizer, aiming to meet other goals, such as removing redundant views and noisy attributes. In this paper, we achieve these goals by performing two types of feature selection, i.e., view-selection for removing redundant views and attribute-selection for deleting noisy attributes. To this end, we propose a new mixed-norm regularizer by integrating an F-norm regularizer with an $\ell_{2,1}$-norm regularizer.

More concretely, in the proposed joint sparse learning model, the F-norm regularizer generates the codes of redundant views as zeros and the others as non-zeros; the $\ell_{2,1}$-norm regularizer generates the codes of noisy attributes as zeros and the others as non-zeros. Then with the impact of sparse views and attributes, MVML classification can be effectively and efficiently performed. Moreover, the mixed-norm regularizer enables us to avoid the issue of over-fitting.

In this paper, the $\ell_{2,1}$-norm regularizer is defined as:

$$\|\mathbf{W}\|_{2,1} = \sum_{i=1}^{m} \|(\mathbf{W})^i\|_2 = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{c} m_{ij}^2} \tag{2}$$

where $(\mathbf{W})^j$ is the j-th row of matrix $\mathbf{W}$, and indicates the effect of the j-th attribute to all data points. As mentioned by existing literatures, e.g., [12], the $\ell_{2,1}$-norm regularizer was designed to measure the distance of the attributes via the ℓ_2-norm, while performing summation over all data points via the ℓ_1-norm. Thus the $\ell_{2,1}$-norm regularizer leads to the row sparsity as well as to consider the correlations of all attributes.

The F-norm regularizer is defines as:

$$\|\mathbf{W}\|_F = \sqrt{\sum_{g=1}^{m_g} \|\mathbf{W}^g\|_2^2} = \sqrt{\sum_{g=1}^{m_g} \sum_{j=1}^{c} w_{g,j}^2} \tag{3}$$

where $\mathbf{W}^g$ is the g-th block of matrix $\mathbf{W}$ (or the submatrix formed by all the rows belonging to the g-th view), and indicates the effect of the g-th block (i.e., there are sequential m_g rows in the g-th view) to all data points.

2.3 Objective Function

By considering three equations together, i.e., Eq.1, Eq.2 and Eq.3, we obtain the objective function of the proposed mixed-norm joint sparse learning model as follows:

$$\min_{\mathbf{W}} \tfrac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda_1 \|\mathbf{W}\|_{2,1} + \lambda_2 \sum_{g=1}^{G} \|\mathbf{W}^g\|_F \tag{4}$$

where both λ_1 ($\lambda_1 > 0$) and λ_2 ($\lambda_2 > 0$) are tuning parameters.

Similar to the mixed sparsity using the ℓ_1-norm regularizer and the $\ell_{2,1}$-norm regularizer together in separable sparse learning, the proposed mixed regularizer leads to the mixed joint sparsity. That is, it first discriminates redundant views via the F-norm regularizer, and then detects noisy attributes in the selected views via the $\ell_{2,1}$-norm regularizer.

Actually, some literatures have focused on the mixed sparsity, such as elastic net [19] and sparse group lasso [11] in separable sparse learning, adaptive multi-task lasso [8] in joint sparse learning, and so on. For example, an elastic net combines the ℓ_1-norm regularizer with the ℓ_2-norm regularizer for achieving the element sparsity (via the ℓ_1-norm regularizer) and impact group effect (via the ℓ_2-norm regularizer). Sparse group lasso achieves the mixed sparsity, i.e., the element sparsity via the ℓ_1-norm regularizer as well as the group sparsity via the $\ell_{2,1}$-norm regularizer. As mentioned in Section 2, neither elastic net nor sparse group lasso benefits for feature selection. Recently, adaptive multi-task lasso combines the ℓ_1-norm regularizer with the $\ell_{2,1}$-norm regularizer in multi-task learning to achieve feature selection (via the $\ell_{2,1}$-norm regularizer) and deletes noisy elements (via the ℓ_1-norm regularizer). Obviously, existing literatures mentioned above were not designed to delete redundancy views and to perform feature selection at the same time, as the proposed method in this paper does.

Next we explain why the proposed mixed-norm regularizer leads to the mixed joint sparsity, i.e., simultaneously obtaining two types of sparsity. While the value of λ_2 is larger, the minimization process in Eq.4 drives the value of the F-norm (i.e., the third term in Eq.4) smaller. This tends to force the values of some blocks (e.g., the value of the g-th block is $\|\mathbf{W}^g\|_F$) with small values to be smaller. After several iterations, the values of these blocks in $\mathbf{W}$ are close to zero. Thus we obtain a sparse $\mathbf{W}$ with zero value in some blocks, e.g., the g-th block. This indicates that the corresponding views (e.g., the g-th view) of $\mathbf{X}$ are redundant views since the sparsity appears in those blocks (e.g., the g-th block) of $\mathbf{W}$. The sparse blocks of $\mathbf{W}$ remove the corresponding views of $\mathbf{X}$ from the test process. Meanwhile, we also notice that the larger the value of λ_2, the more the block sparsity. With the same principle, while the value of λ_1 is larger, the minimization process in Eq.4 forces some rows in $\mathbf{W}$ to be zero, i.e., the attributes corresponding to the sparse rows in $\mathbf{W}$ are not involved the test process. Hence, the proposed mixed-norm regularizer leads to the mixed joint sparsity, which achieves the block sparsity as well as the row sparsity.

According to above analysis, Eq.4 can be used to select a few useful attributes from a few representative (or significant) views of the data for the visual classification. This has the following advantages. First, it benefits for improving the efficiency of the test process due to the sparse $\mathbf{W}$. Second, these two kinds of feature selection help to avoid the impact of redundant views and noisy attributes in the test process, thus benefit for effectively performing the MVML classification. Third, it induces the mixed joint sparsity as well as leads to a hierarchical coding model (i.e., non-sparse attributes generated from non-sparse views), which plays an important role in many applications where a feature hierarchy exists. Last but not the least, views-selection via the F-norm regularizer also preserves the individual view structures of the non-sparse views since each view is regarded as a block.

2.4 Classification

By solving Eq.4, we obtain the optimal $\mathbf{W}$. Given a test dataset $\mathbf{X}_{test}$, we obtain the corresponding label set $\mathbf{Y}_{test}$ by $\mathbf{Y}_{test} = \mathbf{X}_{test}\mathbf{W}$ in the test process. Due to inducing by the proposed mixed-norm regularizer, only a few blocks in the derived $\mathbf{W}$ are non-zeros, and also only a few rows in these non-zero blocks are non-zeros. This makes the test process more efficient to be performed.

After ranking $\mathbf{Y}_{test}$ according to the label values, the top-k labels are assigned to the test data as the predicted labels. This rule is the same to existing multi-label methods, e.g., [18].

3 Optimization

Eq.4 is obviously convex since it consists of three norms, which have been shown to be convex [6]. Therefore, Eq.4 has the global optimum. However, its optimization is very challenging because both the $\|\mathbf{W}\|_F$-norm and the $\|\mathbf{W}\|_{2,1}$-norm in Eq.4 are convex but non-smooth. In this section we solve this problem by calculating sub-gradients of the mixed-norm regularizer, i.e., the $\|\mathbf{W}\|_F$-norm and the $\|\mathbf{W}\|_{2,1}$-norm respectively.

3.1 The Proposed Solver

By setting the derivative of Eq.4 with respect to $\mathbf{W}$ as zero, we obtain:

$$(\mathbf{X}^T\mathbf{X} + \lambda_1\mathbf{C} + \lambda_2\mathbf{D})\mathbf{W} = \mathbf{X}^T\mathbf{Y} \tag{5}$$

where $\mathbf{C}$ is a diagonal matrix with the i-th diagonal element:

$$C_{i,i} = \frac{1}{2\|(\mathbf{W})^i\|_2} \tag{6}$$

where $(\mathbf{W})^i$ denotes the i-th row of $\mathbf{W}$, $i = 1, ..., n$. $\mathbf{D} = diag(\mathbf{D}^1, ..., \mathbf{D}^G)$, where the symbol '*diag*' is the diagonal operator and each $\mathbf{D}^g$ $(g = 1, ..., G)$ is also a diagonal matrix with the i-th diagonal element as:

$$D_{j,j} = \frac{1}{2\|\mathbf{W}^g\|_F} \tag{7}$$

where $j = 1, ..., m_g$.

By observing Eq.5, we find that both the matrix $\mathbf{C}$ and the matrix $\mathbf{D}$ depend on the value of matrix $\mathbf{W}$. In this paper we design a novel iterative algorithm to optimize Eq.5 by alternatively computing the $\mathbf{W}$ and the $\mathbf{C}$ (with the $\mathbf{D}$). We first summarize the details in Algorithm 1, and then prove that in each iteration the updated $\mathbf{W}$ and the $\mathbf{C}$ (with the $\mathbf{D}$) make the value of Eq.4 decrease.

Algorithm 1. The proposed method for solving Eq.4

```
Input: $\mathbf{Y} \in \mathbb{R}^{n\times c}, \mathbf{X} \in \mathbb{R}^{n\times D}, \lambda_1$ and $\lambda_2$;
Output: $\mathbf{W} \in \mathbb{R}^{D\times c}$;
Initialize $t = 0$;
Initialize $\mathbf{C}_0$ as a $D \times D$ identity matrix;
Initialize $\mathbf{D}_0$ as a $D \times D$ identity matrix;
repeat
    $\mathbf{W}^{[t+1]} = (\mathbf{X}^T\mathbf{X} + \lambda_1\mathbf{C}^{[t]} + \lambda_2\mathbf{D}^{[t]})^{-1}\mathbf{X}^T\mathbf{Y}$;
    Update $\mathbf{C}^{[t+1]}$ via Eq.6;
    Update $\mathbf{D}^{[t+1]}$ via Eq.7;
    $t = t+1$;
until No change on the objective function value in Eq.4;
```

With Algorithm 1, at each iteration, given the fixed $\mathbf{C}$ and $\mathbf{D}$, the $\mathbf{W}$ is updated by Eq.5. Then the $\mathbf{C}$ and the $\mathbf{D}$ can be updated with the fixed $\mathbf{W}$. The iteration process is repeated until there is no change on the value of Eq.4.

3.2 Convergence

In this subsection we introduce Theorem 1 to guarantee that Eq.4 monotonically decreases in each iteration of Algorithm 1. Following the literature in [9,17], we first give a lemma as follows:

Lemma 1. *For any positive values a_i and b_i, $i = 1, ..., m$, the following holds:*

$$\begin{aligned}&\sum_{i=1}^{m}\frac{b_i^2}{a_i} \le \sum_{i=1}^{m}\frac{a_i^2}{a_i} \Longleftrightarrow \sum_{i=1}^{m}\frac{(b_i+a_i)(b_i-a_i)}{a_i} \le 0\\ &\Longleftrightarrow \sum_{i=1}^{m}(b_i - a_i) \le 0 \Longleftrightarrow \sum_{i=1}^{m} b_i \le \sum_{i=1}^{m} a_i\end{aligned} \tag{8}$$

Theorem 1. *In each iteration, Algorithm 1 monotonically decreases the objective function value in Eq.4.*

Proof. According to the fifth line of Algorithm 1, we denote the $\mathbf{W}^{[t+1]}$ as the results of the $(t + 1)$-th iteration of Algorithm 1, then we have:

$$\begin{aligned}\mathbf{W}^{[t+1]} = \min_{\mathbf{W}}\ &\frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda_1 tr(\mathbf{W}^T\mathbf{C}^{[t]}\mathbf{W})\\ &+\lambda_2\sum_{g=1}^{G} tr((\mathbf{W}^g)^T(\mathbf{D}^g)^{[t]}\mathbf{W}^g)\end{aligned} \tag{9}$$

Then we can get:

$$\begin{aligned}&\frac{1}{2}\|\mathbf{Y} - \mathbf{X}(\mathbf{W}^{[t+1]})^T\|_F^2 + \lambda_1 tr((\mathbf{W}^{[t+1]})^T\mathbf{C}^{[t]}\mathbf{W}^{[t+1]})\\ &\quad + \lambda_2\sum_{g=1}^{G} tr(((\mathbf{W}^g)^{[t+1]})^T(\mathbf{D}^g)^{[t]}(\mathbf{W}^g)^{[t+1]})\\ &\le \frac{1}{2}\|\mathbf{Y} - \mathbf{X}(\mathbf{W}^{[t]})^T\|_F^2 + \lambda_1 tr((\mathbf{W}^{[t]})^T\mathbf{C}^{[t]}\mathbf{W}^{[t]})\\ &\quad + \lambda_2\sum_{g=1}^{G} tr(((\mathbf{W}^g)^{[t]})^T(\mathbf{D}^g)^{[t]}(\mathbf{W}^g)^{[t]})\end{aligned} \tag{10}$$

which indicates that:

$$\begin{aligned}&\frac{1}{2}\|\mathbf{Y} - \mathbf{X}(\mathbf{W}^{[t+1]})^T\|_F^2 + \sum_{i=1}^{n}\frac{\|(\mathbf{W}^{[t+1]})^i\|_2^2}{2\|(\mathbf{W}^{[t]})^i\|_2}) + \sum_{g=1}^{G}\frac{\|(\mathbf{W}^g)^{[t+1]}\|_F^2}{2\|(\mathbf{W}^g)^{[t]}\|_F})\\ &\le \frac{1}{2}\|\mathbf{Y} - \mathbf{X}(\mathbf{W}^{[t]})^T\|_F^2 + \sum_{i=1}^{n}\frac{\|(\mathbf{W}^{[t]})^i\|_2^2}{2\|(\mathbf{W}^{[t]})^i\|_2}) + \sum_{g=1}^{G}\frac{\|(\mathbf{W}^g)^{[t]}\|_F^2}{2\|(\mathbf{W}^g)^{[t]}\|_F})\end{aligned} \tag{11}$$

Substituting b_i and a_i with $\left\|(\mathbf{W}^{[t+1]})^i\right\|_2$ (or $\|(\mathbf{W}^g)^{[t+1]}\|_F$) and $\left\|(\mathbf{W}^{[t]})^i\right\|_2$ (or $\|(\mathbf{W}^g)^{[t]}\|_F$) in Lemma 1, we have:

$$\begin{aligned}&\frac{1}{2}\|\mathbf{Y}-\mathbf{X}(\mathbf{W}^{[t+1]})^T\|_F^2+\lambda_1\sum_{i=1}^{n}\|(\mathbf{W}^{[t+1]})^i\|_2+\lambda_2\sum_{g=1}^{G}\|(\mathbf{W}^g)^{[t+1]}\|_F\\ &\leq\frac{1}{2}\|\mathbf{Y}-\mathbf{X}(\mathbf{W}^{[t]})^T\|_F^2+\lambda_1\sum_{i=1}^{n}\|(\mathbf{W}^{[t]})^i\|_2+\lambda_2\sum_{g=1}^{G}\|(\mathbf{W}^g)^{[t]}\|_F\end{aligned}\qquad(12)$$

This indicates that the objective function value in Eq.4 monotonically decreases in each iteration of Algorithm 1. Therefore, due to the convexity of Eq.4, Algorithm 1 enables Eq.4 to converge to its global optimum.

4 Experimental Analysis

In order to evaluate the performance of the proposed mixed-norm joint sparse learning (denoted as $F2L21F$[1] for short from its objective function), we compare it with several state-of-the-art methods on public datasets (e.g., MIRFLICKR [5] and NUS-WIDE [3]) for MVML classification, by evaluating the average precision and Hamming loss.

4.1 Experiment Setup

We use four datasets, including MIRFlickr, NUS-WIDE, SCENE and OBJECT in our experiments for MVML classification. The comparison methods include the method in [15] (denoted as *F2F* from its objective function, for simplicity) which only considers the block sparsity, the method in [12] (denoted as *F2L21*) which only considers the row sparsity, the *MKCCA* method in [1] which does not consider the feature redundancy and noise, and the single view method *WorstS* (or *BestS*) which has the worst (or best) classification performance from the data represented by a single view via ridge regression (i.e., all single views are tested). We use two popular evaluation metrics (i.e., the average precision (AP) and Hamming loss (HL)) in multi-label learning [13] to evaluate the effectiveness of all the methods in our experiments.

Given the ground true label matrix $Y1 \in \{0,1\}^{n\times c}$ (where n is the number of instances and c is the number of labels) and the predicted one $Y2 \in \{0,1\}^{n\times c}$ obtained by the algorithm for performing MVML learning, average precision (AP) is defined as:

$$AP=\frac{1}{n\times c}\sum_{i=1}^{n}\frac{card(Y1_i\cap Y2_i)}{card(Y1_i\cup Y2_i)}\qquad(13)$$

where the symbol “Card” means the cardinality operation.

HL measuring the recovery error rate is defined as:

$$HL=\frac{1}{n\times c}\sum_{i=1}^{n}\sum_{j=1}^{c}Y1_{i,j}\oplus Y2_{i,j}\qquad(14)$$

where $\oplus$ is an XOR operation, a.k.a. exclusive disjunction.

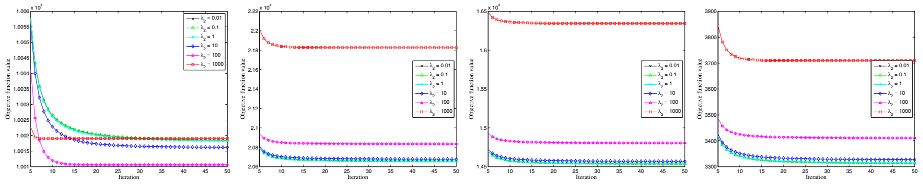
[1] $F2$ means the least square loss function, $L21$ means the $\ell_{2,1}$-norm regularizer, and F means the F-norm regularizer.

According to the literatures, e.g., [13,18], the larger (or smaller) the performance on AP (or HL) is, the better the method.

4.2 Experimental Results

In this subsection, we report the results on MVML classification. First, we evaluate the convergence rate of the proposed *F2L21F* on all four datasets, for evaluating the efficiency of our optimization algorithm, in terms of the objective function value in each iteration. Second, we test the parameters' sensitivity of the proposed model on λ_1 and λ_2, aiming at obtaining the best performance of the proposed *F2L21F*. Finally, we compare *F2L21F* with the comparison algorithms in terms of average precision and Hamming loss.

Convergence Rate. We solve Eq.4 by the proposed Algorithm 1. In this experiment, we want to know the convergence rate of Algorithm 1. Here we report some of the results in Fig.1 and Fig.2 due to the page limit. Fig.1 shows the results on the objective function value while fixing the value of λ_1 (i.e., $\lambda_1 = 1$) and varying λ_2. Fig.2 shows the results on the objective function value while fixing the value of λ_2 (i.e., $\lambda_2 = 1$) and varying λ_1. In both Fig.1 and Fig.2, the x-axis and y-axis denote the number of iterations and the objective function value respectively.



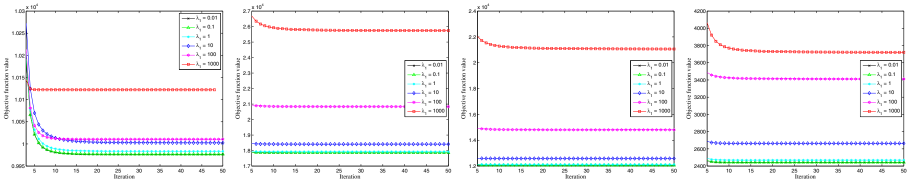
(a) MIRFlickr (b) NUS-WIDE (c) SCENE (d) OBJECT

Fig. 1. An illustration on convergence rate of Algorithm 1 for solving the proposed objective function with fixed λ_1, i.e., $\lambda_1 = 1$

We can observe from both Fig.1 and Fig.2 that: 1) the objective function value rapidly decreases at the first few iterations; and 2) the objective function value becomes stable after about 30 iterations (or even less than 20 in many cases) on all datasets. This confirms a fast convergence rate of Algorithm 1 to solve the proposed optimization problem in Eq.4. Similar results are observed for other λ_1 and λ_2 values.

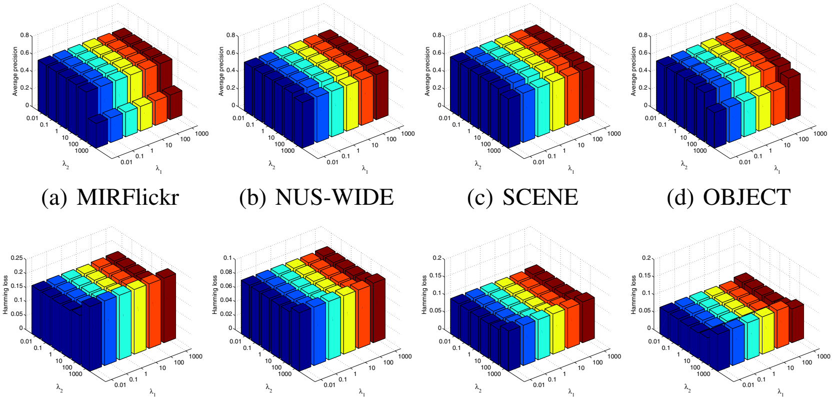
Parameters' Sensitivity. In this experiment, we test different settings on parameters λ_1 and λ_2 in the proposed *F2L21F*, by varying them as $\{0.01, 0.1, 1, 10, 100, 1000\}$. The results on average prediction and Hamming are illustrated in Fig.3.

It is clear that the proposed *F2L21F* is sensitive to the parameters' setting, similar to other sparse learning methods [11,18]. However, we find the worst performance is



(a) MIRFlickr (b) NUS-WIDE (c) SCENE (d) OBJECT

Fig. 2. An illustration on convergence rate of Algorithm 1 for solving the proposed objective function with fixed λ_2, i.e., $\lambda_2 = 1$



(a) MIRFlickr (b) NUS-WIDE (c) SCENE (d) OBJECT

(e) MIRFlickr (f) NUS-WIDE (g) SCENE (h) OBJECT

Fig. 3. The results of average precision (first row) and Hamming loss (second low) on various parameters' settings on different datasets

always obtained when both λ_1 and λ_2 have extremely large values. For example, when the values of parameters pair (λ_1, λ_2) are around (10,10), *F2L21F* achieves the best performance. Actually, in our experiments such a setting simultaneously leads to both the row sparsity (via the λ_1) and the block sparsity (via the λ_2).

Comparison. In this experiment, we compare our proposed method with state-of-the-art methods for MVML classification. We set the values of parameters for the comparison methods by following the instructions in their original papers. For all the methods, we randomly sample 60% of the original data as the training data, and leave the rest as the test data. We randomly generate ten runs, and report the average result and the standard deviation on the average precision and Hamming loss, as shown in Fig.4. Note that we do not use dataset MIRFlickr since it has only two views.

From Fig.4, we have the following observations: 1) The proposed *F2L21F* always achieves the best performance. Among six views in NUS-WIDE and five views in SCENE and OBJECT, the 64-D color histogram is detected as a redundant view in *F2L21F*. *F2L21* and *MKCCA* use all the views to perform MVML classification and obtain worse performance than *F2L21F*. This confirms that some views (e.g., the color histogram in the tested datasets) are not helpful in the learning process and may even

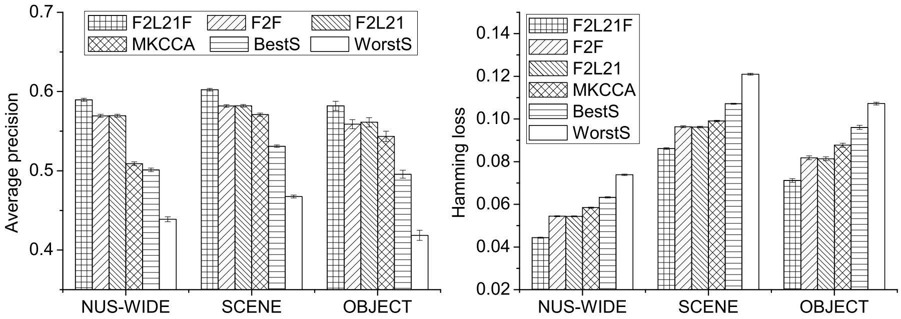


Fig. 4. Comparison on average precision (left) and Hamming loss (right) for all methods on different datasets. Note that the range shown at the top of the bar represents the performance standard deviation.

degrade the performance, especially when many views are available. *F2L21F* is able to identify those redundant views and avoid their negative impact on the classification. Although *F2F* can also discover those redundant views, it is not able to remove noisy attributes from the selected views, leading to worse performance than *F2L21F*. This result proves the effectiveness of *F2L21F* in removing redundant views and noisy attributes by employing the proposed mixed-norm regularizer in MVML classification. 2) The performance of single view learning methods (i.e., *BestS* and *WorstS*) is always worse than those multi-view learning methods. This again confirms the advantages of using multi-views in visual classification. 3) The sparse learning methods (i.e., *F2L21F*, *F2F*, and *F2L21*) consistently outperform *MKCCA*. This shows the superiority of sparse learning which encodes negligible elements as zeros and only selects important elements to perform MVML classification. Moreover, in our implementation the computational cost of *F2L21F* is about tens times faster than that of *MKCCA*, indicating much higher efficiency than *MKCCA*.

In sum, more views can help improve the performance of visual classification since more information can be utilized in the learning process. On the other hand, more views may also potentially introduce higher redundancy and more noise which compromise the performance. The proposed *F2L21F* is able to identify those redundant views and noisy attributes so that MVML classification can be performed for more effective performance.

5 Conclusion

In this paper we proposed a mixed-norm joint sparse learning model for multi-view multi-label (MVML) classification. The proposed method, powered by a mixed-norm regularizer, can effectively avoid the negative impact of redundant views and noisy attributes from the multi-view representation of a large amount of data. Extensive experimental results have shown that the proposed method outperforms state-of-the-art learning methods for MVML classification. In the future, we will extend the proposed method into its kernel edition to project noise more clearly, and involve other learning models, such as semi-supervised MVML classification and transfer MVML classification, to leverage the widely available unlabeled data and heterogenous data.

References

1. Blaschko, M.B., Lampert, C.H., Gretton, A.: Semi-supervised laplacian regularization of kernel canonical correlation analysis. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part I. LNCS (LNAI), vol. 5211, pp. 133–145. Springer, Heidelberg (2008)
2. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Annual Conference on Computational Learning Theory, pp. 92–100 (1998)
3. Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: A real-world web image database from national university of singapore. In: ACM International Conference on Image and Video Retrieval, p. 48 (2009)
4. Dhillon, P.S., Foster, D., Ungar, L.: Multi-view learning of word embeddings via cca. In: Neural Information Processing Systems, pp. 9–16 (2011)
5. Huiskes, M.J., Lew, M.S.: The mir flickr retrieval evaluation. In: ACM International Conference on Multimedia Information Retrieval, pp. 39–43 (2008)
6. Jenatton, R., Audibert, J.-Y., Bach, F.: Structured variable selection with sparsity-inducing norms. Journal of Machine Learning Research 12, 2777 (2011)
7. Kumar, A., DauméIII, H.: A co-training approach for multi-view spectral clustering. In: International Conference on Machine Learning, pp. 393–400 (2011)
8. Lee, S., Zhu, J., Xing, E.P.: Adaptive multi-task lasso: with application to eqtl detection. In: Neural Information Processing Systems, pp. 1306–1314 (2010)
9. Nie, F., Huang, H., Cai, X., Ding, C.: Efficient and robust feature selection via joint l2, 1-norms minimization. In: Neural Information Processing Systems, pp. 1813–1821 (2010)
10. Owens, T., Saenko, K., Chakrabarti, A., Xiong, Y., Zickler, T., Darrell, T.: Learning object color models from multi-view constraints. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 169–176 (2011)
11. Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J.R., Wang, P.: Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. The Annals of Applied Statistics 4(1), 53–77 (2010)
12. Sun, L., Liu, J., Chen, J., Ye, J.: Efficient recovery of jointly sparse vectors. In: Neural Information Processing Systems, pp. 1812–1820 (2009)
13. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining Multi-label Data (2009)
14. Xie, B., Mu, Y., Tao, D., Huang, K.: m-sne: Multiview stochastic neighbor embedding. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 41(4), 1088–1096 (2011)
15. Yuan, X., Yan, S.: Visual classification with multi-task joint sparse representation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3493–3500 (2010)
16. Zhu, X., Huang, Z., Shen, H.T., Cheng, J., Xu, C.: Dimensionality reduction by mixed kernel canonical correlation analysis. Pattern Recognition (2012)
17. Zhu, X., Huang, Z., Yang, Y., Shen, H.T., Xu, C., Luo, J.: Self-taught dimensionality reduction on the high-dimensional small-sized data. Pattern Recognition 46(1), 215–229 (2013)
18. Zhu, X., Shen, H.T., Huang, Z.: Video-to-shot tag allocation by weighted sparse group lasso. In: ACM Multimedia, pp. 1501–1504 (2011)
19. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B 67(2), 301–320 (2005)