# Multi-Label Informed Feature Selection

**Ling Jian**[1,2*], **Jundong Li**[1*], **Kai Shu**[1], **Huan Liu**[1]
1. Computer Science and Engineering, Arizona State University, Tempe, 85281, USA
2. College of Science, China University of Petroleum, Qingdao, 266555, China
{ling.jian,jundong.li,kai.shu,huan.liu}@asu.edu

## Abstract

Multi-label learning has been extensively studied in the area of bioinformatics, information retrieval, multimedia annotation, etc. In multi-label learning, each instance is associated with multiple interdependent class labels, the label information can be noisy and incomplete. In addition, multi-labeled data often has noisy, irrelevant and redundant features of high dimensionality. As an effective data preprocessing step, feature selection has shown its effectiveness to prepare high-dimensional data for numerous data mining and machine learning tasks. Most of existing multi-label feature selection algorithms either boil down to solving multiple single-labeled feature selection problems or directly make use of imperfect labels. Therefore, they may not be able to find discriminative features that are shared by multiple labels. In this paper, we propose a novel multi-label informed feature selection framework MIFS, which exploits label correlations to select discriminative features across multiple labels. Specifically, to reduce the negative effects of imperfect label information in finding label correlations, we decompose the multi-label information into a low-dimensional space and then employ the reduced space to steer the feature selection process. Empirical studies on real-world datasets demonstrate the effectiveness and efficiency of the proposed framework.

## 1 Introduction

Recent years has witnessed an increasing number of applications involving multi-labeled data in which each instance is associated with multiple labels simultaneously [Zhang and Zhou, 2006; Hua and Qi, 2008; Katakis *et al.*, 2008; Song *et al.*, 2008; Tang *et al.*, 2009; Gopal and Yang, 2010]. For example, in bioinformatics, a gene may be related to multiple functions [Elisseeff and Weston, 2001]; in information retrieval, each document may cover several topics [Huang *et al.*, 2012]; in image processing, an image may be annotated with different scenes [Boutell *et al.*, 2004].

Normally, multi-labeled data in the aforementioned applications such as gene sequences, texts and images are represented by feature vectors with very high dimensionality. The high dimensionality of multi-labeled data not only significantly increases the memory storage requirements and computational costs for many learning algorithms, but also limits the usage of these learning algorithms in real applications due to the curse of dimensionality [Duda *et al.*, 2012]. Previous studies on feature selection have shown that the most discriminative information are usually carried by only a subset of relevant features [Liu and Motoda, 2007]. In other words, many noisy, redundant, and irrelevant features which negatively affect the learning performance can be eliminated.

However, it is not easy to directly perform feature selection on multi-labeled data due to its unique characteristics. First, different from traditional single-labeled feature selection problems where class labels are mutually exclusive, different classes in multi-labeled data are typically not independent but inherently correlated. For example, in text categorization, "sports" are more closely related to the category of "athletics" than to the category of "soap stars". Therefore, it is crucial to find some common features for the classes of "sports" and "athletics". Second, multiple labels of instances are often annotated by human beings. It is natural for us to make some incorrect or incomplete annotations especially when we are provided with hundreds or even thousands of labeling options. Therefore, it is important to seek a reasonable way of exploiting the label correlation for multi-labeled feature selection. Existing methods for multi-labeled feature selection either transform the problem to multiple single-labeled sub-problems or directly make use of the flawed labels. Not surprisingly, they do not perform well in finding relevant features that are shared by multiple labels.

To tackle above challenges in multi-labeled feature selection problem, in this paper, we propose a multi-label informed feature selection framework, named MIFS. In particular, we first map the label information into a low-dimensional reduced space that captures the correlations among multiple labels. Then we employ the reduced space instead of the original noisy and incomplete label information to guide the feature selection phase. In this way, it alleviates the negative influences of imperfect label information to find relevant features across multiple labels. The major contributions of this paper are summarized as follows:

[*] Indicates equal contribution.

- Introducing a principled way of exploiting label correlations for feature selection in the presence of noisy and incomplete label information;
- Proposing a novel multi-label informed feature selection framework MIFS which is able to select discriminative features across multiple class labels;
- Developing an efficient algorithm to address the optimization issue of MIFS and;
- Conducting experiments on multiple benchmark datasets to demonstrate the effectiveness and efficiency of the proposed MIFS framework.

The remainder of this paper is organized as follows. In Section 2, we introduce the proposed multi-label informed feature selection framework MIFS with an efficient alternating optimization algorithm. In Section 3, empirical evaluations on benchmark datasets are given to show the superiority of the proposed framework. In Section 4, we briefly review related work on multi-label feature selection. The conclusion and future work are presented in Section 5.

## 2 Multi-Label Informed Feature Selection Framework - MIFS

In this section, we first summarize some symbols used throughout this paper and then introduce the formulation of the proposed MIFS framework.

### 2.1 Preliminaries

We use bold uppercase character to denote matrix (e.g., $\mathbf{A}$), bold lowercase character to denote vector (e.g., $\mathbf{a}$), the $i$-th entry of $\mathbf{a}$ as $\mathbf{a}_i$, $(i,j)$-th entry of $\mathbf{A}$ as $\mathbf{A}_{ij}$, $i$-th row of $\mathbf{A}$ as $\mathbf{A}_{i:}$, transpose of $\mathbf{A}$ as $\mathbf{A}^T$, trace of $\mathbf{A}$ as Tr($\mathbf{A}$) if $\mathbf{A}$ is square matrix. For any matrix $\mathbf{A}\in\mathbb{R}^{n\times d}$, its Frobenius norm is defined as $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{d}\mathbf{A}_{ij}^2}$, its $\ell_{2,1}$-norm is $\|\mathbf{A}\|_{2,1} = \sum_{i=1}^{n}\sqrt{\sum_{j=1}^{d}\mathbf{A}_{ij}^2}$. Suppose that in the multi-labeled dataset, we have $n$ instances $\mathbf{x}_1,\cdots,\mathbf{x}_n\in\mathbb{R}^d$ and $k$ different labels $\mathcal{Y}=\{c_1,\cdots,c_k\}$. Each instance $\mathbf{x}_i$ is associated with a subset of labels in $\mathcal{Y}$, we represent this subset of labels by a binary vector $\mathbf{y}_i=[y_i^1,\cdots,y_i^k]\in\{0,1\}^k$ where $y_i^j=1$ $(j=1,\cdots,k)$ if and only if $\mathbf{x}_i$ is associated with label $c_j$. Following the expression of MATLAB, we denote the data matrix as $\mathbf{X}=[\mathbf{x}_1;\mathbf{x}_2;\cdots;\mathbf{x}_n]\in\mathbb{R}^{n\times d}$ and label matrix as $\mathbf{Y}=[\mathbf{y}_1;\mathbf{y}_2;\cdots;\mathbf{y}_n]\in\{0,1\}^{n\times k}$.

### 2.2 Formulation

In multi-label learning problems, each instance is attributed with multiple class labels and these labels may be correlated with each other. Therefore, during the feature selection process, it is beneficial to explicitly take the label correlations into consideration. In this way, we are able to find common features for strongly correlated labels and different features for weakly correlated labels. However, it is problematic to directly extract label correlations from multiple labels. In reality, multi-labeled data often consists of hundreds or even thousands of human annotated labels. It is inevitable to make some incorrect and incomplete labeling during the arduous annotation work. It is inappropriate to directly apply multiple labels for feature selection in the presence of flawed labels.

Motivated by Latent Semantic Indexing (LSI) [Dumais, 2004], we propose an effective way to decompose the multi-labeled output space to a low-dimensional space, and employ this low-dimensional space to guide the feature selection process. One encouraging property of this low-dimensional space is that most of the structures in the original output label space can be explained and recovered. Meanwhile, noisy information in the output space is greatly reduced [Yu *et al.*, 2005]. Mathematically, it decomposes the multi-labeled output space $\mathbf{Y}$ to a product of two low-dimensional matrices $\mathbf{V}\in\mathbb{R}^{n\times c}$ and $\mathbf{B}\in\mathbb{R}^{c\times k}$ by minimizing the following reconstruction error:

$$\min_{\mathbf{V},\mathbf{B}}\|\mathbf{Y}-\mathbf{V}\mathbf{B}\|_F^2, \tag{1}$$

where $\mathbf{V}$ denotes the latent semantics of the multi-label information. It can be interpreted that we cluster the original $k$ labels into $c$ different clusters, each cluster has a specific semantic meaning. For example, in text categorization, labels of "sports" and "athletics" are more likely to encode similar latent semantic meanings. $\mathbf{B}$ is a coefficient matrix, and each column in $\mathbf{B}$ shows the coefficient of each label in these $c$ latent semantic variables.

Since the low-dimensional latent semantics matrix $\mathbf{V}$ encodes label correlations and greatly reduces the noise in the original multi-label output space, we take advantage of it to perform feature selection via a regression model and the features that are most related to the latent semantics $\mathbf{V}$ are selected. In particular, we propose to perform label decomposition and feature selection simultaneously via the following optimization problem:

$$\min_{\mathbf{W},\mathbf{V},\mathbf{B}}\|\mathbf{X}\mathbf{W}-\mathbf{V}\|_F^2+\alpha\|\mathbf{Y}-\mathbf{V}\mathbf{B}\|_F^2+\gamma\|\mathbf{W}\|_{2,1}, \tag{2}$$

where $\mathbf{W}$ is a feature coefficient matrix and each row of $\mathbf{W}$ measures the importance of $i$-th feature in approximating the latent semantics $\mathbf{V}$. The $\ell_{2,1}$-norm regularization term is imposed on $\mathbf{W}$ to ensure that $\mathbf{W}$ is sparse in rows, i.e., it achieves joint feature sparsity for all $c$ latent semantic variables. The parameter $\alpha$ balances the contribution of feature learning and label decomposition. The other parameter $\gamma$ controls the sparsity of the model.

In addition, since we employ low-dimensional latent semantics $\mathbf{V}$ to guide the feature selection process, we need to ensure that local geometry structures are consistent between the input space $\mathbf{X}$ and the reduced low-dimensional semantics $\mathbf{V}$. In other words, if two instances are close to each other in the input space $\mathbf{X}$ then they should also be close to each other in the latent semantic space $\mathbf{V}$ by minimizing the following:

$$\begin{aligned}&\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{S}_{ij}(\mathbf{V}_{i:}-\mathbf{V}_{j:})^2\\ =&\mathrm{Tr}(\mathbf{V}^T(\mathbf{A}-\mathbf{S})\mathbf{V})\\ =&\mathrm{Tr}(\mathbf{V}^T\mathbf{L}\mathbf{V})\end{aligned} \tag{3}$$

where $\mathbf{L}=\mathbf{A}-\mathbf{S}$ is the graph laplacian matrix and $\mathbf{A}$ is a diagonal matrix with $\mathbf{A}_{ii}=\sum_{j=1}^{n}\mathbf{S}_{ij}$. $\mathbf{V}_{i:}$ denotes the

latent semantics of $\mathbf{y}_i$ and $\mathbf{S}_{ij}$ is some similarity measure of instances $\mathbf{x}_i$ and $\mathbf{x}_j$. In this paper we follow [Cai *et al.*, 2010] to build a nearest neighbor graph to effectively model local geometry structure in the input space $\mathbf{X}$ and the affinity graph is defined as:

$$\mathbf{S}_{ij} = \begin{cases} \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}) & \text{if } \mathbf{x}_i \in \mathcal{N}_p(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathcal{N}_p(\mathbf{x}_i) \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathcal{N}_p(\mathbf{x}_i)$ denotes the $p$-nearest neighbors of instance $\mathbf{x}_i$.

Integrating the local geometry structure of the data, the final objective function for the multi-label informed feature selection (MIFS) can be formulated as follows:

$$\min_{\mathbf{W},\mathbf{V},\mathbf{B}} \quad \|\mathbf{XW} - \mathbf{V}\|_F^2 + \alpha\|\mathbf{Y} - \mathbf{VB}\|_F^2 + \beta\mathrm{Tr}(\mathbf{V}^T\mathbf{LV}) + \gamma\|\mathbf{W}\|_{2,1}, \tag{4}$$

where $\beta$ is a parameter that measures how the local geometry structure of the data is preserved in the latent semantic space. It can be observed from the objective function of MIFS in Eq. (4) that the latent semantics $\mathbf{V}$ involves in three terms, it captures label correlations, preserves local geometry structure and guides feature selection process simultaneously. The feature coefficient matrix $\mathbf{W}$ involves in two terms, it makes $\mathbf{X}$ approximate $\mathbf{V}$ via a regression model and achieves feature selection by a $\ell_{2,1}$-norm regularization. There are three variables in the above objective function, we will introduce an efficient way to obtain these model parameters in the next subsection.

### 2.3 Optimization Algorithm for MIFS

It can be observed that the objective function of MIFS in Eq. (4) is not convex w.r.t. all these variables $\mathbf{V}$, $\mathbf{B}$ and $\mathbf{W}$ jointly. In addition, it is also not smooth due to the $\ell_{2,1}$-norm regularization term on $\mathbf{W}$. Therefore, following [Nie *et al.*, 2010], we relax the term $\|\mathbf{W}\|_{2,1}$ by $2\mathrm{Tr}(\mathbf{W}^T\mathbf{DW})$, where $\mathbf{D}$ is a diagonal matrix with its diagonal element $\mathbf{D}_{ii} = \frac{1}{2\sqrt{\mathbf{W}_{i:}^T\mathbf{W}_{i:}+\epsilon}}$ and $\epsilon$ is a small positive constant.

To optimize the objective function $\Theta(\mathbf{W},\mathbf{V},\mathbf{B}) = \|\mathbf{XW} - \mathbf{V}\|_F^2 + \alpha\|\mathbf{Y} - \mathbf{VB}\|_F^2 + \beta\mathrm{Tr}(\mathbf{V}^T\mathbf{LV}) + 2\gamma\mathrm{Tr}(\mathbf{W}^T\mathbf{DW})$, we propose to use an efficient alternating optimization algorithm. Specifically, in each iteration, we update one variable while fix the other two variables since the objective function is convex when any two variables are fixed. In addition, the objective function is differentiable, thus we can apply gradient descent method in an alternating way to update the model parameters. By taking the derivative of objective function w.r.t. variables $\mathbf{W}$, $\mathbf{V}$ and $\mathbf{B}$ respectively, we have the following formulations:

$$\begin{cases} \frac{\partial\Theta}{\partial\mathbf{W}} = 2[(\mathbf{X}^T(\mathbf{XW} - \mathbf{V}) + \gamma\mathbf{DW})] \\ \frac{\partial\Theta}{\partial\mathbf{V}} = 2[(\mathbf{V} - \mathbf{XW}) + \alpha(\mathbf{VB} - \mathbf{Y})\mathbf{B}^T + \beta\mathbf{LV}] \\ \frac{\partial\Theta}{\partial\mathbf{B}} = 2\alpha\mathbf{V}^T(\mathbf{VB} - \mathbf{Y}). \end{cases} \tag{5}$$

With these, the update rule of the alternating algorithm for MIFS is summarized as follows:

$$\begin{cases} \mathbf{W} := \mathbf{W} - \lambda_W\frac{\partial\Theta}{\partial\mathbf{W}} \\ \mathbf{V} := \mathbf{V} - \lambda_V\frac{\partial\Theta}{\partial\mathbf{V}} \\ \mathbf{B} := \mathbf{B} - \lambda_B\frac{\partial\Theta}{\partial\mathbf{B}}, \end{cases} \tag{6}$$

where $\lambda_W$, $\lambda_V$ and $\lambda_B$ are stepsizes for these three gradient descent update rules. In each iteration, it needs $\mathcal{O}(ndc + n^2)$ operations to update all these three variables. Hence, the chosen of suitable stepsizes is of crucial importance to accelerate the convergence rate and to reduce the total running time of MIFS, especially for large-scale (large $n$) and high-dimensional (large $d$) problems. In the current work, we employ Armijo rule [Bertsekas, 1999] to adaptively determine the stepsizes $\lambda_W$, $\lambda_V$, and $\lambda_B$ in each iteration.

After obtaining the model parameters, we rank each feature according to the value of $\|\mathbf{W}_{i,:}\|_2$ ($i = 1, .., d$) in a descending order and return the top ranked features. The pseudocode of the multi-label informed feature selection framework MIFS is illustrated in Algorithm 1.

**Algorithm 1** Multi-label Informed Feature Selection (MIFS)

**Input:** Initialize $\mathbf{W}$, $\mathbf{V}$, $\mathbf{B}$. Parameters $\alpha, \beta, \gamma$
**Output:** Top ranked features

**Repeat**
$\frac{\partial\Theta}{\partial\mathbf{W}} = 2[(\mathbf{X}^T(\mathbf{XW} - \mathbf{V}) + \gamma\mathbf{DW})]$;
$\frac{\partial\Theta}{\partial\mathbf{V}} = 2[(\mathbf{V} - \mathbf{XW}) + \alpha(\mathbf{VB} - \mathbf{Y})\mathbf{B}^T + \beta\mathbf{LV}]$;
$\frac{\partial\Theta}{\partial\mathbf{B}} = 2\alpha\mathbf{V}^T(\mathbf{VB} - \mathbf{Y})$;
determine stepsizes $\lambda_V$, $\lambda_B$ and $\lambda_W$ with Armijo rule;
$\mathbf{V} := \mathbf{V} - \lambda_V\frac{\partial\Theta}{\partial\mathbf{V}}$;
$\mathbf{B} := \mathbf{B} - \lambda_B\frac{\partial\Theta}{\partial\mathbf{B}}$;
$\mathbf{W} := \mathbf{W} - \lambda_W\frac{\partial\Theta}{\partial\mathbf{W}}$;
Update $\mathbf{D}$;
**Until** Convergence
Return $\mathbf{W}^*$;
Rank features according to $\|\mathbf{W}^*_{i,:}\|_2$ in a descending order and return the top ranked features.

## 3 Experimental Study

In this section, we conduct experiments on real-world multi-labeled datasets to assess the performance of the proposed multi-label informed feature selection framework MIFS.

### 3.1 Data Sets

Experiments are conducted on four publicly available benchmark datasets[1], including one image dataset (i.e. Scene [Boutell *et al.*, 2004]) and three text datasets from RCV1 [Lewis *et al.*, 2004]. The Scene dataset consists of 400 images from Corel stock photo library and some personal images. Each image is associated with a subset of six semantic scenes (bench, sunset, fall foliage, field, urban, and mountain). RCV1, i.e., Reuters Corpus Volume 1, is an archive of over 80,000 newswire stories. Each document is represented with TF-IDF format and has been cosine normalized. From the RCV1 repository, we choose three representative multi-labeled text datasets Topics, Regions, and Industries. Details of these datasets are listed in Table 1.

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

Table 1: Details of four benchmark datasets.

| Dataset | # Training | # Test | # Features | # Labels |
|---|---|---|---|---|
| Scene | 1211 | 1196 | 294 | 6 |
| Topics | 3000 | 3000 | 14171 | 101 |
| Regions | 11575 | 13953 | 7236 | 228 |
| Industries | 6967 | 9967 | 18894 | 354 |

### 3.2 Experimental Settings

Following the standard way to validate supervised feature selection, we evaluate the proposed MIFS framework on classification task. To have a fair comparison with existing methods, we decompose the multi-labeled classification problem into multiple binary classification problems, and then employ SVM to learn these binary classifiers with a five-fold cross validation. The SVM implementation in the Liblinear toolbox [Fan *et al.*, 2008] is used in the experiments.

Two widely adopted evaluation criteria based on F-measure, i.e., macro-average and micro-average are used to measure the performance of multi-label classification algorithms. F-measure is one of the most popular metric for evaluation of binary classification [Yu *et al.*, 2005] and it is defined as the harmonic mean of precision and recall:

$$\text{F-measure} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \tag{7}$$

where TP denotes the number of true positives, FP denotes the number of false positives and FN denotes the number of false negatives.

Micro-average can be considered as a weighted average of F-measure over all $k$ labels:

$$\text{Micro-average} = \frac{\sum_{i=1}^{k} 2\text{TP}^i}{\sum_{i=1}^{k} (2\text{TP}^i + \text{FP}^i + \text{FN}^i)}. \tag{8}$$

Macro-average is an arithmetic average of F-measure of all output labels:

$$\text{Macro-average} = \frac{1}{k} \sum_{i=1}^{k} \frac{2\text{TP}^i}{(2\text{TP}^i + \text{FP}^i + \text{FN}^i)}. \tag{9}$$

$\text{TP}^i$, $\text{FP}^i$ and $\text{FN}^i$ denotes the number of true positives, false positives and false negatives in the $i$-th class label, respectively. The higher the micro-average and macro-average values are, the better the classification performance is.

MIFS is measured against the following state-of-the-art feature selection methods for multi-label classification problems. The number of selected features are varied as $\{2\%, 4\%, ..., 20\%\}$ of total number of features.

1. F-Score: Fisher Score [Duda *et al.*, 2012] selects features by assigning similar feature values to the instances within the same class and different feature values to instances from different classes. The features with the highest discriminative power are selected.
2. RFS: Robust Feature Selection [Nie *et al.*, 2010] applies $\ell_{2,1}$-norm regularization for both the loss function and the regularization term. It is robust to outliers in the input space and is able to select features across all instances with a joint sparsity.

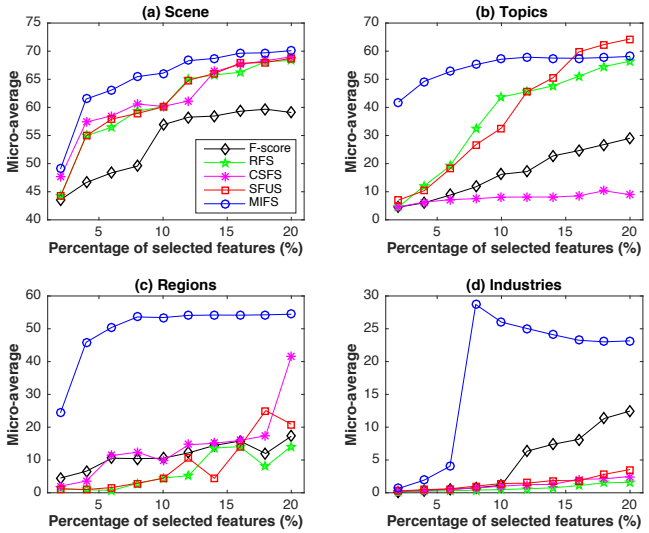


Figure 1: Micro-average comparisons of five feature selection algorithms on four datasets.

3. CSFS: Convex Semi-supervised multi-label Feature Selection [Chang *et al.*, 2014] is a convex algorithm designed for large-scale multi-label feature selection problems. In this experiment, we adopt its supervised version for a fair comparison.
4. SFUS: Sub-Feature Uncovering with Sparsity [Ma *et al.*, 2012] incorporates joint sparse feature selection with multi-label learning to uncover shared feature subspace.

### 3.3 Experimental Results

In MIFS, there are some parameters need to be set in advance. First, to model the local geometry structure in the input space $\mathbf{X}$, we set the parameters $\sigma^2$ and $p$ as 1 and 5, respectively. There are three important regularization parameters $\alpha$, $\beta$ and $\gamma$ in MIFS. Similarly, RFS, CSFS and SFUS also have different regularization parameters. For a fair comparison between these feature selection methods, we tune these regularization parameters for all methods with a grid-search strategy by varying its value in the range of $\{10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.2, 0.4, 0.6, 0.8, 1, 10\}$. For all these methods, we report the best results of the optimal parameters in terms of classification performance. The experiments are repeated 5 times and averaged.

Figure 1 and Figure 2 show the classification performance of different feature selection algorithms in terms of Micro-average and Macro-average on four datasets. To have a more comprehensive comparison, we also use the Statlog' ordering method [Michie *et al.*, 1994] to list the average ranks of these four feature selection algorithms. The ranks are presented in Table 2. The following observations can be found from Figure 1, Figure 2 and Table 2.

1. With the increase of the number of selected features, the classification performance first tends to increase and then keeps stable or even degrades.
2. MIFS and the other two multi-labeled learning algorithms CSFS and SFUS obtains better Micro-average

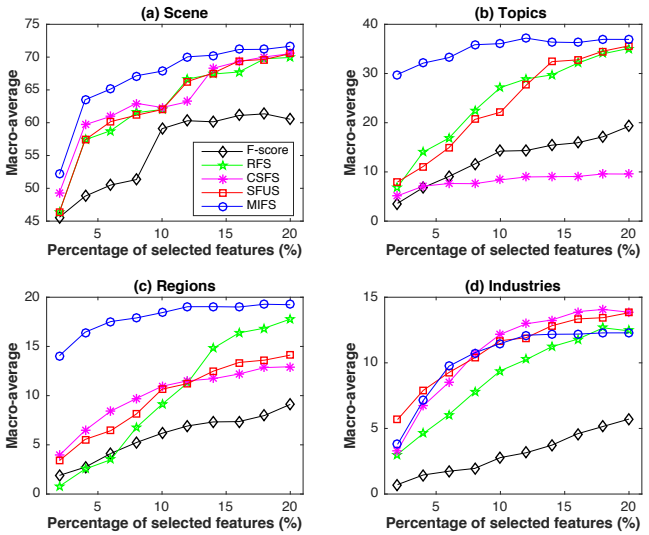


Figure 2: Macro-average comparisons of five feature selection algorithms on four datasets.

Table 2: The comprehensive rank of different methods.

| Dataset | F-score | RFS | CSFS | SFUS | MIFS |
|---|---|---|---|---|---|
| Scene | 4.00 | 3.10 | 2.25 | 3.00 | 1.00 |
| Topics | 4.05 | 2.55 | 4.75 | 2.30 | 1.15 |
| Regions | 3.80 | 3.30 | 2.45 | 3.50 | 1.00 |
| Industries | 3.50 | 3.95 | 2.30 | 2.25 | 1.75 |

and Macro-average than single-labeled feature selection algorithms F-score and RFS on these four datasets. It shows that when performing feature selection, it is beneficial to explicitly incorporate the multi-label information into the model.

3. MIFS outperforms the other two multi-labeled learning algorithms CSFS and SFUS in most cases. It demonstrates that by decomposing the label information to a low-dimensional semantic space, we can still capture the label correlations and alleviate the negative effects of flawed labels to find relevant features.

4. On datasets Scene, Topics, and Regions, the proposed MIFS outperforms other four methods significantly when the percentage of select features are less than 14%. This observation shows that MIFS can achieve good classification performance even when a few number of features are selected.

### 3.4 Convergence Analysis

As mentioned before, the proposed alternating optimization algorithm monotonically decreases the objective function value in Eq. (4) iteratively until convergence. In addition, the Armijo update rule can accelerate the convergence process significantly. In this subsection, we conduct an empirical experiment to show the efficiency of the alternating optimization algorithm with Armijo update rule. In the experiment, the parameters $\alpha$, $\beta$, and $\gamma$ are all fixed as 0.1. The following stopping criterion is used:

$$\frac{|\Theta^t - \Theta^{t-1}|}{\Theta^{t-1}} < 10^{-5}, \tag{10}$$

where $\Theta^t$ indicates the objective function value in the $t$-th iteration.

Table 3: The running time (sec) of different methods.

| Dataset | F-score | RFS | CSFS | SFUS | MIFS |
|---|---|---|---|---|---|
| Scene | 0.16 | 58.96 | 10.71 | 10.82 | 4.47 |
| Topics | 91.59 | 962.92 | 3089.09 | 9859.06 | 1460.29 |
| Regions | 198.87 | 1434.64 | 998.44 | 7855.08 | 1002.76 |
| Industries | 686.92 | 6019.93 | 13688.60 | 25808.25 | 5532.39 |

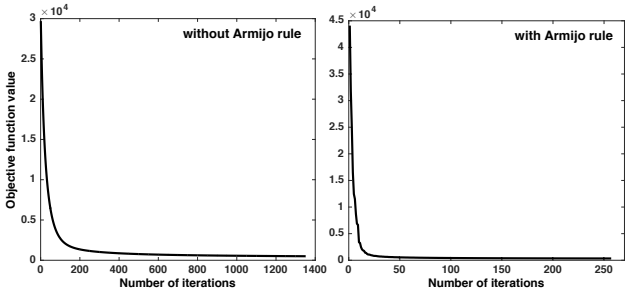


Figure 3: Convergence comparison of MIFS with and without Armijo update rule on the Scene dataset.

Figure 3 shows the convergence of the proposed alternating optimization algorithm with/without Armijo update rule. It can be seen clearly that the alternating optimization algorithm with Armijo update rule converges more quickly. It converges in 100 iterations, while the optimization algorithm without Armijo update rule converges needs 1000 iterations.

### 3.5 Running Time Comparison

In this subsection, we show the efficiency of the proposed MIFS framework by comparing its running time with other baseline methods. In Table 3, we show the running time that each method needs to converge to the optimal solution. SFUS and CSFS perform an eigen-decomposition and matrix inverse computation each iteration which require $\mathcal{O}(d^3)$ operations, respectively. Hence, they are not suitable to handle high-dimensional data. RFS also needs to compute the matrix inverse each iteration, which requires $\mathcal{O}(n^3 + n^2 d)$. In contrast, MIFS only requires simple matrix multiplication operations, its computational complexity is $\mathcal{O}(ndc + n^2)$ each iteration. In a nutshell, compared with the state-of-the-art methods RFS, SFUS and CSFS, the proposed MIFS method is more computationally efficient.

### 3.6 Parameter Sensitivity Study

MIFS has three important parameters: $\alpha$, $\beta$, and $\gamma$. The parameter $\alpha$ measures the contribution of the multi-label decomposition process. The parameter $\beta$ controls how strongly the low-dimensional latent semantics preserves the local geometry structure in the original input space. The third parameter $\gamma$ controls the sparseness of the proposed model.

To study how these parameters affect the feature selection results and the consequent multi-label classification problems, we conduct an experiment to study the effectiveness of these parameters and report the performance variances in Figure 4. Due to space limit, we only present the results of dataset Scene. We turn the parameters $\alpha$, $\beta$, $\gamma$ from $\{10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.2, 0.4, 0.6, 0.8, 1, 10\}$. Figures of $\alpha$ are shown with the other parameters $\beta$ and $\gamma$ fixed as 0.1. The same setting is used for the figures of $\beta$ and figures of $\gamma$.

It can be shown that the classification performance is not very sensitive to the changes of parameters $\alpha, \beta$, and $\gamma$. Therefore, in practice, we can safely set these parameters in a wide range. The classification performance is the best when $\alpha = 0.4$, $\beta = 1$ and $\gamma = 0.8$. Another observation is that compared with these three regularization parameters, the classification performance is more sensitive to the number of selected features. Specifically, when we increase the number of selected features, the classification performance first increase, keeps stable and then decreases. How to determine the suitable number of selected features is still an open issue in feature selection research.

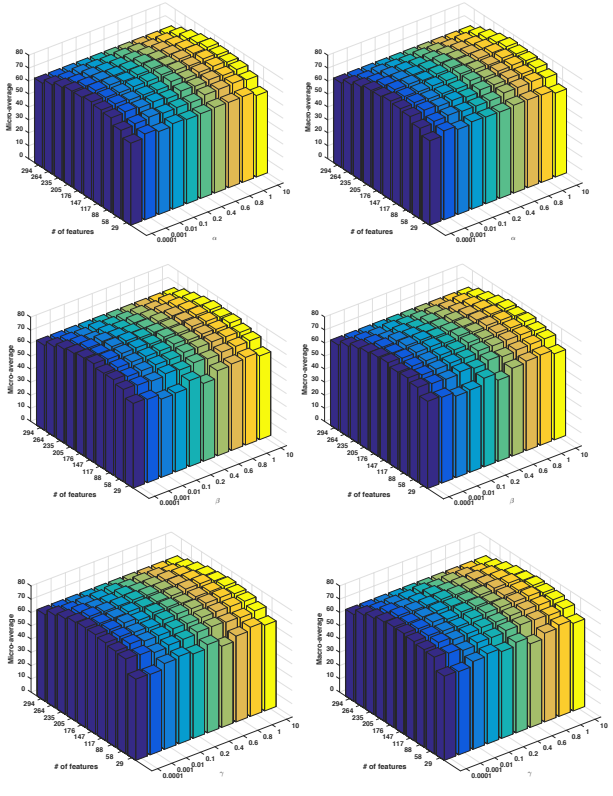


Figure 4: Micro-average and Macro-average of MIFS on Scene dataset with respect to different $\alpha$, $\beta$, $\gamma$ and number of selected features.

## 4 Related Work

Our work is most related to sparse learning based feature selection and multi-label learning. Therefore, in this section, we briefly review some related work on these two aspects.

Over the past two decades, numerous feature selection methods have been proposed [Tibshirani, 1996; Peng *et al.*, 2005; Li *et al.*, 2015; 2016a]. Recently, sparse learning based methods have received increasing attention due to their good performance and interpretability. Typically, these methods embed the feature selection process into the classification model such these two phases compliment with each other. Among these approaches, $\ell_{2,1}$-norm regularization based methods are extremely popular [Liu *et al.*, 2009; Nie *et al.*, 2010; Yang *et al.*, 2011; Ma *et al.*, 2012; Chang *et al.*, 2014; Li *et al.*, 2016b] due to its ability to handle multi-class problems. By imposing the $\ell_{2,1}$-norm sparse regularization, the feature coefficients are guaranteed to be sparse across multiple targets. However, these methods cannot be directly applied for multi-label feature selection as they do not explicitly consider the label correlations in the feature selection process.

With the prevalence of multi-labeled data in many real-world applications, multi-label learning emerges to be another hot research topic. Similar to many data mining and machine learning tasks, multi-label learning also suffers from the curse of dimensionality. An effective strategy is to exploit the label correlations to reduce the feature dimensionality. To tackle this issue, MLLS extracts a common subspace shared among multiple labels [Ji *et al.*, 2008]. MDDM projects the original data into a low dimensional space by maximizing the dependence between the original feature description and the associated class labels [Zhang and Zhou, 2010]. SFUS joint selects features via a sparse regularization and uncovers the shared feature subspace of original features [Ma *et al.*, 2012]. These work are different from our proposed MIFS framework - (1) Most of exiting multi-label learning methods focus on transforming the original feature space to a new space, while our method performs feature selection directly which preserves the physical meanings of the original data; (2) To reduce the negative effects of imperfect label information, we decompose it to a low-dimensional space and take advantage of it to perform feature selection with sparse regularization.

## 5 Conclusion and Future Work

In this study, we propose a novel multi-label informed feature selection framework MIFS. The proposed method has two appealing properties. First, it makes use of latent semantics of the multi-labels to guide the feature selection phase. Therefore, it alleviates the negative affects of noisy and incomplete labels in finding relevant features. Second, it exploits the label correlations in the output space to find features that are shared across multiple labels. An efficient alternating optimization algorithm is developed to solve the optimization problem of MIFS. Empirical studies on real-world datasets demonstrate the efficiency and efficacy of the proposed framework. Future research can be focused on investigating how to perform online multi-labeled feature selection in which data samples arrive in a streaming fashion.

## Acknowledgements

This material is, in part, supported by National Natural Science Foundation of China (NSFC No. 61403419, 61503412) and National Science Foundation (NSF No. IIS-1217466).

## References


[Bertsekas, 1999] Dimitri P Bertsekas. *Nonlinear Programming*. Athena scientific, 1999.

[Boutell *et al.*, 2004] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

[Cai *et al.*, 2010] Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In *KDD*, pages 333–342, 2010.

[Chang *et al.*, 2014] Xiaojun Chang, Feiping Nie, Yi Yang, and Heng Huang. A convex formulation for semi-supervised multi-label feature selection. In *AAAI*, pages 1171–1177, 2014.

[Duda *et al.*, 2012] Richard O Duda, Peter E Hart, and David G Stork. *Pattern Classification*. John Wiley & Sons, 2012.

[Dumais, 2004] Susan T Dumais. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1):188–230, 2004.

[Elisseeff and Weston, 2001] André Elisseeff and Jason Weston. A kernel method for multi-labelled classification. In *NIPS*, pages 681–687, 2001.

[Fan *et al.*, 2008] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[Gopal and Yang, 2010] Siddharth Gopal and Yiming Yang. Multilabel classification with meta-level features. In *SIGIR*, pages 315–322, 2010.

[Hua and Qi, 2008] Xian-Sheng Hua and Guo-Jun Qi. Online multi-label active annotation: towards large-scale content-based video search. In *ACM MM*, pages 141–150, 2008.

[Huang *et al.*, 2012] Sheng-Jun Huang, Zhi-Hua Zhou, and ZH Zhou. Multi-label learning by exploiting label correlations locally. In *AAAI*, pages 949–955, 2012.

[Ji *et al.*, 2008] Shuiwang Ji, Lei Tang, Shipeng Yu, and Jieping Ye. Extracting shared subspace for multi-label classification. In *KDD*, pages 381–389, 2008.

[Katakis *et al.*, 2008] Ioannis Katakis, Grigorios Tsoumakas, and Ioannis Vlahavas. Multilabel text classification for automated tag suggestion. *ECML PKDD discovery challenge*, 75, 2008.

[Lewis *et al.*, 2004] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397, 2004.

[Li *et al.*, 2015] Jundong Li, Xia Hu, Jiliang Tang, and Huan Liu. Unsupervised streaming feature selection in social media. In *CIKM*, pages 1041–1050, 2015.

[Li *et al.*, 2016a] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *arXiv preprint arXiv:1601.07996*, 2016.

[Li *et al.*, 2016b] Jundong Li, Xia Hu, Liang Wu, and Huan Liu. Robust unsupervised feature selection on networked data. In *SDM*, 2016.

[Liu and Motoda, 2007] Huan Liu and Hiroshi Motoda. *Computational Methods of Feature Selection*. CRC Press, 2007.

[Liu *et al.*, 2009] Jun Liu, Shuiwang Ji, and Jieping Ye. Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization. In *UAI*, pages 339–348, 2009.

[Ma *et al.*, 2012] Zhigang Ma, Feiping Nie, Yi Yang, Jasper RR Uijlings, and Nicu Sebe. Web image annotation via subspace-sparsity collaborated feature selection. *IEEE Transactions on Multimedia*, 14(4):1021–1030, 2012.

[Michie *et al.*, 1994] Donald Michie, David J Spiegelhalter, and Charles C Taylor. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, 1994.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS*, pages 1813–1821, 2010.

[Peng *et al.*, 2005] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[Song *et al.*, 2008] Yang Song, Lu Zhang, and C Lee Giles. A sparse gaussian processes classification framework for fast tag suggestions. In *CIKM*, pages 93–102, 2008.

[Tang *et al.*, 2009] Lei Tang, Suju Rajan, and Vijay K Narayanan. Large scale multi-label classification via metalabeler. In *WWW*, pages 211–220, 2009.

[Tibshirani, 1996] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[Yang *et al.*, 2011] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. $\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI*, pages 1589–1594, 2011.

[Yu *et al.*, 2005] Kai Yu, Shipeng Yu, and Volker Tresp. Multi-label informed latent semantic indexing. In *SIGIR*, pages 258–265, 2005.

[Zhang and Zhou, 2006] Min-Ling Zhang and Zhi-Hua Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1338–1351, 2006.

[Zhang and Zhou, 2010] Yin Zhang and Zhi-Hua Zhou. Multilabel dimensionality reduction via dependence maximization. *ACM Transactions on Knowledge Discovery from Data*, 4(3):14, 2010.