# Manifold regularized discriminative feature selection for multi-label learning

Jia Zhang [a], Zhiming Luo [b], Candong Li [c], Changen Zhou [c], Shaozi Li [a],*

[a] Department of Cognitive Science, Xiamen University, Xiamen 361005, PR China
[b] Postdoc Center of Information and Communication Engineering, Xiamen University, Xiamen 361005, PR China
[c] College of Traditional Chinese Medicine, Fujian University of Traditional Chinese Medicine, Fuzhou 350122, PR China

## ARTICLE INFO

## ABSTRACT

In multi-label learning, objects are essentially related to multiple semantic meanings, and the type of data is confronted with the impact of high feature dimensionality simultaneously, such as the bioinformatics and text mining applications. To tackle the learning problem, the key technology, i.e., feature selection, is developed to reduce dimensionality, whereas most of the previous methods for multi-label feature selection are either directly transformed from traditional single-label feature selection methods or half-baked in the label information exploitation, and thus causing the redundant or irrelevant features involved in the selected feature subset. Aimed to seek discriminative features across multiple class labels, we propose an embedded multi-label feature selection method with manifold regularization. To be specific, a low-dimensional embedding is constructed based on the original feature space to fit the label distribution for capturing the label correlations locally, which is also constrained using the label information in consideration of the co-occurrence relationships of label pairs. Following this principle, we design an optimization objective function involving $l_{2,1}$-norm regularization to achieve multi-label feature selection, and the convergence is guaranteed. Empirical studies on various multi-label data sets reveal that the proposed method can obtain highly competitive performance against some state-of-the-art multi-label feature selection methods.

## 1. Introduction

For pattern recognition, traditional supervised learning mainly focuses on labeling an unseen instance with an associated class label. This formulation of the learning problem entails the restriction of that each instance is relevant to only one label. In fact, instances might be associated with multiple semantic meanings simultaneously [28,35,37], which have been widely existed in real-world applications. For example, a news document could be related to several topics in text categorization [26], an image could be annotated with multiple scenes [2], and a disease is regularly represented by several patterns of syndromes in TCM diagnosis [6]. For the classification problem of these complex scenarios, multi-label learning technologies are regarded as feasible schemes to deal with multi-label data.

However, high-dimensional data is often notorious to tackle for multi-label learning, which not only leads to the increasing of

the computational cost and memory storage requirement, but also overfits many machine learning models due to the curse of dimensionality [17,18]. Feature selection is one of popular and effective methods in reducing dimensionality. It helps to prevent the curse of dimensionality and select an optimal feature subset from the original feature space with a certain criterion [20,33]. According to the difference of recognition tasks, feature selection methods can be roughly classified into two categories: single-label learning and multi-label learning. For the single-label task, the purpose of feature selection is to keep relevant features for classification and eliminate redundant and irrelevant features to improve the generalization performance [21]. Based on this, many well-established feature selection methods have been proposed in past years, such as mRMR [25] and RFS [23]. For the multi-label task, feature selection is performed for the same purpose with the single-label task, in the meanwhile, mining the correlations among labels is deemed to be crucial in improving the performance [16,34].

As we know, class labels in multi-label data are not independent but inherently correlated. According to the order of correlations, the existing strategies for multi-label learning can be

* Corresponding author.
*E-mail addresses:* zhiming.luo@xmu.edu.cn (Z. Luo), szlig@xmu.edu.cn (S. Li).

categorized into three families [37], as follow: *first-order* strategy deals with multiple class labels with a label-by-label style, and representative methods include ML-KNN [38] and BR [2]. *Second-order* strategy explores pairwise relationships between labels. For modeling the *second-order* correlations, a common strategy is to exploit the interaction of label pairs, such as LLSF [12] and CLR [9], and another one is to optimize the multi-label model where *ranking loss* is involved in the objective function, such as MLCM-r [32] and Rank-SVM [8]. *High-order* strategy tackles multi-label problem by mining relations among all labels or subsets of labels, whereas this strategy could be less effective due to the time complexity and scalability [37]. Furthermore, some researchers devote to multi-label learning by capturing the label correlations locally, and the local correlations shared by instances can be discovered via clustering [14], or locally linear embedding [11].

To address the above challenges for multi-label learning, in this paper, we propose an embedded feature selection method via manifold regularization, namely MDFS, to select discriminative features that are shared by multiple class labels. In detail, we project the original feature information into a low-dimensional space for exploiting the local label correlations, and employ the reduced space to design the label information based manifold regularizer to take into account the global correlations between labels. Based on this, an optimization objective function is constructed via joint $l_{2,1}$-norm minimization. With alternating optimization, the sparse representation can be learned for multi-label feature selection. Finally, experimental results on various multi-label data sets, including a collected *TCM* data set, show that the proposed method is superior to some state-of-the-art multi-label feature selection methods. The major contributions of this paper are summarized as follow:

- An optimization framework is built to exploit the label correlations via manifold regularization. The label space is recovered to exploit the local correlations, which is further utilized to enhance the pairwise label relationships.
- An embedded multi-label feature selection method is proposed. By combining $l_{2,1}$-norm regularization with the constraint for capturing the label correlations, the feature importance can be assessed across multiple class labels.
- An iterative optimization algorithm is developed to solve the optimization problem of the proposed method with convexity.
- Extensive experiments on various benchmark multi-label data sets demonstrate the feasibility and effectiveness of the proposed method.

The rest of this paper is organized as follows. We give a brief review of related work in Section 2. In Section 3, we describe the proposed method in detail, and analyze the experimental results in Section 4. Finally in Section 5, the discussion and conclusion is given.

## 2. Related work

With the rapid development of the information technology, the expansion size in data keeps growing, and the curse of dimensionality is becoming a tricky problem in practical applications [16,17]. Feature selection is a high-efficiency tool to reduce dimensionality, which can be classified into three categories, i.e., filter, wrapper, and embedded, from the selection strategy perspective [18]. Filter methods are applied to generating the feature ranking before classification, which is based on certain characteristics of data, such as feature correlations [25]. Wrapper methods include the interactions with an off-the-shelf classifier, and the classifier performance is viewed as the indicator to evaluate the selected feature subset [36]. Embedded methods directly incorporate the feature selection process as a part of the classifier training [23].

Based on the above selection strategies, many multi-label feature selection methods have been proposed, in which information theoretical based methods are widely used for the reduction of multi-label data [18]. In consideration of that the designed information-theoretic functions can be applied to measuring the uncertainty of random variables or features, which are conducive to capturing the feature-feature correlations and feature-label correlations, plenty of information-theoretic feature selection methods have been put forward to select relevant features. For example, inspired by the strategy with integrated relevance and redundancy criterion [25], some feature selection methods have been proposed to search for discriminative features [20,27]. Moreover, Lin et al. [19] proposed a MDMR method based on max-dependency and min-redundancy. To be specific, the mutual information-based measure was employed to maximize the dependency between the candidate feature and all class labels, and minimize the conditional redundancy between the candidate feature and the selected feature subset simultaneously. Li et al. [16] achieved multi-label feature selection from a granular computing viewpoint. By abstracting the relevant labels into the same information granule, the criteria of maximal dependency on information granules and minimal intra-redundancy were adopted to exploit the correlations among class labels.

In addition, some sparse learning based methods have been designed for multi-label feature selection [12,15]. Among these methods using sparse representation, both $l_{2,1}$-norm regularization and $l_1$-norm regularization based methods have attracted increasing interests. Huang et al. [12] presented a multi-label feature selection method LLSF based on $l_1$-norm regularization, which designed an optimization framework to learn the low-dimensional data representation of each label, and considered the shared features to exploit the pairwise relations between labels simultaneously. Nie et al. [23] proposed an embedded feature selection method RFS for multi-class classification, which is expandable in multi-label setting. By combining the $l_{2,1}$-norm regularization, the proposed optimization objective was proved to be robust to the outlines. Jian et al. [15] first mapped the original label information into a low-dimensional latent semantics matrix for the label correlations exploitation, and then combined the low-dimensional space with $l_{2,1}$-norm regularization to achieve multi-label feature selection. Noticeably, the feature representation using $l_2$-norm regularization is not beneficial to obtain a strictly sparse solution, however, it has been shown that it possesses the capability of discriminability [39].

Furthermore, there are some other well-established multi-label feature selection methods via manifold learning. For example, Huang et al. [13] developed a new method called MCLS. By exploiting the local topological structures from feature space, the method first transformed the original logical labels into the numerical ones, and then obtained the laplacian score with the regularization based on the similarity between the numerical labels. Cai and Zhu [5] proposed a new multi-label feature selection method with sparsity, namely MSSL, which reconstructed the feature manifold that has taken the similarity between features into account, and then embedded the feature graph regularization into the feature selection process. Zhu et al. [40] designed a multi-label classification model GLOCAL. This method used the low-rank of label space as the central role to generate the classifier output. Based on this, the authors designed global and local manifold regularizers to the label correlations exploitation, and thus minimizing the reconstruction error. In this paper, we also use manifold learning to achieve the algorithm design. Different to the aforementioned methods, we consider to construct a low-dimensional embedding, which is induced from the feature manifold and also restricted by the global label correlations information, for learning a feature mapping with sparsity under strong supervision.

## 3. The MDFS approach

In this section, a novel feature selection method is proposed for multi-label learning, and the framework of the proposed method is shown in Fig. 1. First, we map the original feature space into a low-dimensional embedding based on manifold assumption [1]. In light of that two instances are more similar in the low-dimensional embedding while they are closer to each other in the original feature space, the local label correlations are captured to guide the feature selection process. Next, the low-dimensional embedding is further utilized to construct the label information based manifold regularizer, thus facilitating the global label correlations exploitation. Finally, with the aid of the local and global label correlations, we conduct manifold regularized discriminative feature selection for multi-label learning.

### 3.1. Problem formulation

Formally, let $\mathcal{D} = \{(\mathbf{x}_i, Y_i)|1 \leq i \leq n\}$ be a multi-label data set. $X = \{\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_n\}$ is the feature matrix, and $\mathbf{x}_i$ is a instance represented by a $d$-dimensional input vector, which is associated with a finite set of $q$ possible labels, denoted by $L = \{l_1, l_2, \cdots, l_q\}$. $Y = \{Y_1; Y_2; \ldots; Y_n\}$ is the label matrix, and $Y_i = \{y_{i1}, y_{i2}, \ldots, y_{iq}\}$ is the ground-truth label vector of $\mathbf{x}_i$. For $y_{ij} \in Y_i$ ($1 \leq j \leq q$), $y_{ij} = 1$ in case that $l_j$ is relevant to $\mathbf{x}_i$, otherwise $y_{ij} = -1$.

Following the preset, we search for a set of relevant features from multi-label data, such that some loss functions or specific evaluation is satisfied, thus achieving the purpose. As we know, each instance is associated with multiple labels, and these labels may be correlated with each other. Thus, the appropriate usage of the label correlations information is a crucial step to improve the performance. Aimed at designing such an effective feature selection method, we define an optimization framework for revealing the features that are discriminative for all the labels. Suppose $F$ is the low-dimensional embedding induced from the feature matrix $X$, $W$ is a mapping matrix, and both $\lambda$ and $\gamma$ are tradeoff parameters, we define the optimization framework as follow:

$$\min_{W,F} \quad V(X, F, W) + \lambda C(F, Y) + \gamma \Omega(W), \tag{1}$$

In Eq. (1), the first term $V$ represents the mapping function while the local label correlations are taken into consideration, the second term $C$ is used to enforce the multi-label learning using the

global label correlations information, and the third term $\Omega$ is the regularizer to control the complexity of the model. For simplicity, we implement $\Omega(W)$ with $l_{2,1}$-norm regularization.

$$\Omega(W) = \|W\|_{2,1}, \tag{2}$$

The term imposed on $W$ makes the optimal solution sparse [23,39]. Thus, discriminative features for multi-label learning have large weights, and the weights of features that are nonsignificant are close to zero. Besides, the definitions of terms $V$ and $C$ will be described in the following section.

### 3.2. Exploiting label correlations

Inspired by the literature [24], one important property of the low-dimensional embedding, which is induced from the feature space, is that most of the structures in the feature manifold can be recovered. Based on this, the low-dimensional embedding $F$ (induced from the feature matrix $X$) has similar local structures with $X$. In other words, the similarity among instances can be reconstructed during the mapping procedure. Moveover, similar instances are more likely to share a label [11], hence we utilize $F$ for effective exploitation of the local label correlations. However, the mapping $F$ doesn't help to project test data. To achieve this goal, we define the linear mapping function $\mathbf{h}(X, W) = XW + 1_n b^T$, in which $W \in \Re^{d \times q}$ is utilized as a medium to relate $F$ with the feature matrix $X$. Thus, the term $V$ can be formulated as follow:

$$V(X, F, W) = \mathtt{Tr}(F^T L F) + \|XW + 1_n b^T - F\|_F^2, \tag{3}$$

where $\|.\|_F$ denotes the Frobenius norm, $b \in \Re^q$ is the bias term, and $1_n \in \Re^n$ denotes a column vector, whose all the elements are 1. $L = D - S$ is the graph Laplacian, in which $D$ is a diagonal matrix, and $D_{ii} = \sum_{j=1}^{n} S_{ij}$. Besides, $S$ is the weight matrix of instances, and element $S_{ij} \in S$ measures the similarity between instances $\mathbf{x}_i$ and $\mathbf{x}_j$, and can be calculated by a heat kernel.

$$S_{ij} = \begin{cases} \exp\left(-\dfrac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}\right) & if \quad \mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}_j) \quad or \quad \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i) \\ 0 & otherwise, \end{cases} \tag{4}$$

where $\sigma$ denotes the parameter, which is usually set to $+\infty$ for graph construction. $\mathcal{N}_k(\mathbf{x}_i)$ denotes the set of the Top-$k$ nearest neighbors of $\mathbf{x}_i$, and the Euclidean distance is employed to search for the neighbors.
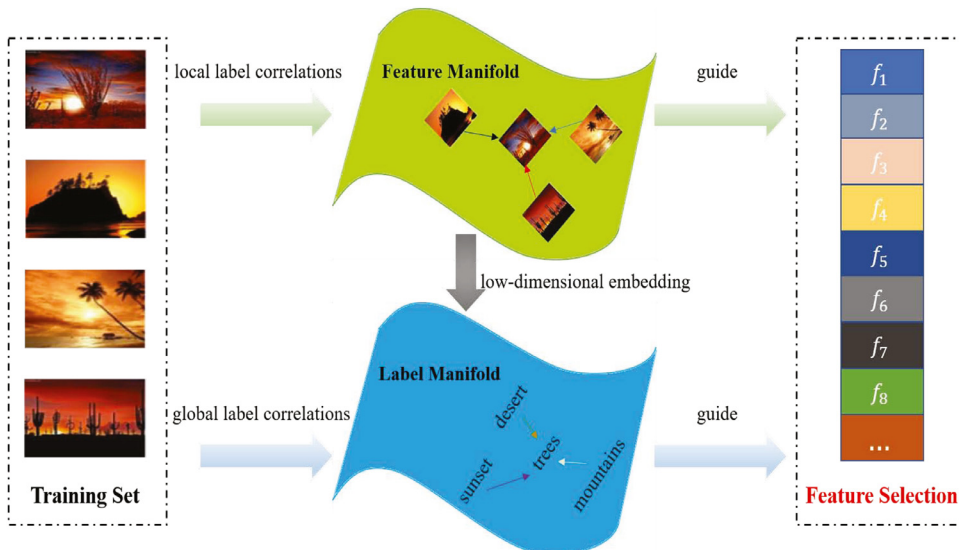


Fig. 1. The framework of the proposed method.

Furthermore, we define the label manifold regularization to design the term $C$. Specially, if the positive correlation between labels is stronger, the predictions on the two labels should be more similar, and vise versa. Then we introduce a feasible *second-order* strategy to consider the global label correlations, and the term $C$ can be defined as follow:

$$C(F, Y) = \frac{1}{2} \sum_{i=1}^{q} \sum_{j=1}^{q} [S_0]_{ij} \|[F^T]_i - [F^T]_j\|^2, \qquad (5)$$

where $S_0 \in \Re^{q \times q}$ is the similarity matrix of labels, and each element in $S_0$ represents the similarity between two labels calculated by the heat kernel. $[F^T]_i \in \Re^n$ is the predicted result on the $i$th label. With some algebraic steps, Eq. (5) can be equivalently written as $\text{Tr}(FL_0F^T)$, in which $L_0 = D_0 - S_0$ is the label Laplacian matrix, $D_0$ is a diagonal matrix, whose diagonal element is defined as $[D_0]_{ii} = \sum_{j=1}^{q} [S_0]_{ij}$. Based on this, if $[S_0]_{ij}$ is more larger, $[F^T]_i$ and $[F^T]_j$ are more similar, and vice versa. Thus, the expected $F$ can be generated as the pairwise label information is involved.

### 3.3. The objective function of MDFS

As the previous discussion, the objective function of Eq. (1) can be clearly defined for manifold regularized discriminative feature selection. Considering the values of ground-truth labels are known, we utilize a reliable approach by incorporating the label information to guide the optimization process. In a supervised manner, we enhance the consistency between $F$ and the real label distribution. Thus, the objective function of MDFS becomes the following one:

$$\min_{F,W,b} \quad \text{Tr}(F^T LF) + \|XW + 1_n b^T - F\|_F^2 + \alpha \|F - Y\|_F^2 \\ + \beta \text{Tr}(FL_0F^T) + \gamma \|W\|_{2,1}, \qquad (6)$$

where $\alpha$, $\beta$, and $\gamma$ are the parameters to trade off the three terms. Eq. (6) involves three variables, including $W$, $F$ and $b$. It is natural to obtain the optimal solution with iterative optimization, which will be introduced in the next section. Note that the objective function in Eq. (6) can change by substituting $XW + 1_n b^T$ for $F$. Nevertheless, the solution for $W$ is also obtained based on iterative optimization. Therefore, from the perspective of the efficiency, it is inessential to omit $F$ for discriminative feature analysis.

### 3.4. Solution

The objective function, as shown in Eq. (6), contains multiple variables to optimize, which is also non-smooth in light of the $l_{2,1}$-norm regularization imposed on $W$. To solve the learning problem, an alternating minimization strategy is introduced to generate the optimal solution. First, we set the derivative of the objective function in Eq. (6) w.r.t. $b$ to zero, and we can derive:

$$b^T = \frac{1}{n}(1_n^T F - 1_n^T XW), \qquad (7)$$

Plugging Eq. (7) into Eq. (6), we can rewrite Eq. (6) as follow:

$$\min_{F,W} \quad \text{Tr}(F^T LF) + \|XW + \frac{1}{n}(1_n 1_n^T F - 1_n 1_n^T XW) - F\|_F^2 \\ + \alpha \|F - Y\|_F^2 + \beta \text{Tr}(FL_0F^T) + \gamma \|W\|_{2,1}, \\ \Longrightarrow \min_{F,W} \quad \text{Tr}(F^T LF) + \|HXW - HF\|_F^2 + \alpha \|F - Y\|_F^2 \\ + \beta \text{Tr}(FL_0F^T) + \gamma \|W\|_{2,1}, \qquad (8)$$

where $H = I - \frac{1}{n} 1_n 1_n^T$, which is used to center data by subtracting the mean of the data, and satisfies the property $H = H^T = H^2$. Then we solve for $F$ and $W$ in Eq. (8) with an EM style iterative process. By using the matrix properties:

$$\frac{\partial \text{Tr}(F^T LF)}{\partial F} = (L + L^T)F, \quad \frac{\partial \text{Tr}(FL_0F^T)}{\partial F} = F(L_0 + L_0^T),$$

The solution of $F$ can be obtained by setting the derivative of the objective function in Eq. (8) w.r.t. $F$ to zero. We have:

$$(L + L^T)F - 2H^T(HXW - HF) + 2\alpha(F - Y) + \beta F(L_0 + L_0^T) = 0, \qquad (9)$$

Considering that both $L$ and $L_0$ are symmetric matrices, we can transform Eq. (9) into the following one:

$$(L + H + \alpha I)F + \beta FL_0 = HXW + \alpha Y, \qquad (10)$$

Eq. (10) is a matrix equation with the form of $AF + FB = C$, where $A = L + H + \alpha I$, $B = \beta L_0$, and $C = HXW + \alpha Y$. To address the mathematic issue, several practical methods [22,30] can be employed to obtain the numerical solution w.r.t. $F$, such as the existed software library LAPACK and the *Lyapunov* function in Matlab. Similarly, we derive $W$ by fixing $F$ by minimizing the reduced formula of Eq. (8), as follow:

$$\min_{W} \quad \delta(W) = \|HXW - HF\|_F^2 + \gamma \|W\|_{2,1}, \qquad (11)$$

To solve the optimization problem, $Q$ is defined as a diagonal matrix, whose diagonal element $Q_{ii} = \frac{1}{2\|W_i\|_2}$ $(1 \le i \le d)$, we can solve for $W$ by requiring $\partial \delta(W)/\partial W$ to zero, as follow:

$$W = (X^T HX + \gamma Q)^{-1} X^T HF, \qquad (12)$$

In brief, the solution is designed by executing the iterative process for unknown variables, and the pseudo code is shown in Algorithm 1. After obtaining $W$, we assess the importance of each feature based on the value of $\|W_i\|_2$ $(1 \le i \le d)$, and thus the subset of top ranked features can be obtained.

### 3.5. Convergence analysis

In this section, we prove that the iterative optimization converges due to the fact that the optimization problem is convex in terms of both $W$ and $F$.

**Theorem 1.** *Algorithm 1 gives the optimal solution with convexity.*

---

**Algorithm 1** The MDFS approach.

**Input**: Training set $\mathcal{D} = \{(\mathbf{x}_i, Y_i) | 1 \le i \le n\}$, parameters $\alpha$, $\beta$, $\gamma$, and $k$.
**Output**: Projection matrix $W$.

1: Construct the graph Laplacian matrices $L$ and $L_0$;
2: $H = I - \frac{1}{n} 1_n 1_n^T$;
3: Initialize $W$ randomly;
4: **repeat**
5:     Update the optimal $F$ by solving Eq. (10);
6:     Compute the $d$-dimensional diagonal matrix $Q$ as:

$$Q = \begin{bmatrix} \frac{1}{2\|W_1\|_2} & & \\ & \ddots & \\ & & \frac{1}{2\|W_d\|_2} \end{bmatrix};$$

7:     Update the optimal $W = (X^T HX + \gamma Q)^{-1} X^T HF$;
8: **until** *Convergence*;
9: **return** $W$ for feature selection.

---

**Proof.** Suppose $W^t$ and $F^t$ are prepared in the $t$th iteration, we fix $W$ as $W^t$ for obtaining $F^{t+1}$. Considering $\varphi(F)$ is a standard quadratic form:

$$\varphi(F) = \text{Tr}(F^T LF) + \|HXW - HF\|_F^2 + \alpha \|F - Y\|_F^2 + \beta \text{Tr}(FL_0F^T), \qquad (13)$$

It is obvious that $L$ and $L_0$ are positive semidefinite, and $F^T L F \geq 0$ and $F L_0 F^T \geq 0$ for any nonzero F hold. Thus, both $\text{Tr}(F^T L F)$ and $\text{Tr}(F L_0 F^T)$ are convex. Besides, the other terms are also convex w.r.t. F, hence the sum of the four functions is a convex program. Then, we can easily conclude that:

$$\varphi(F^{t+1}) \leq \varphi(F^t), \tag{14}$$

In the same manner, we fix $F$ as $F^t$, and solve for $W^{t+1}$. As we know, the standard $l_{2,1}$-norm regularization is convex, and the convergence has been well studied and proved in the literature [23]. Therefore, we have:

$$\|HXW^{t+1} - HF^t\|_F^2 + \gamma \|W^{t+1}\|_{2,1} \leq \|HXW^t - HF^t\|_F^2 + \gamma \|W^t\|_{2,1}, \tag{15}$$

Integrating Eqs. (14) and (15), the following inequation arrives at:

$$\begin{aligned}
&\text{Tr}((F^{t+1})^T L F^{t+1}) + \|HXW^{t+1} - HF^{t+1}\|_F^2 + \alpha \|F^{t+1} - Y\|_F^2 \\
&\quad + \beta \text{Tr}(F^{t+1} L_0 (F^{t+1})^T) + \gamma \|W^{t+1}\|_{2,1} \\
&\leq \text{Tr}((F^t)^T L F^t) + \|HXW^t - HF^t\|_F^2 + \alpha \|F^t - Y\|_F^2 \\
&\quad + \beta \text{Tr}(F^t L_0 (F^t)^T) + \gamma \|W^t\|_{2,1}, 
\end{aligned} \tag{16}$$

Eq. (16) shows the objective value decreases after each iteration. It is proved that the solution found by Algorithm 1 is convex. For a convex problem, any local minimum is also a global minimum [3]. Thus, the propose method converges to the global minimum. □

## 4. Experiments

### 4.1. Experimental data sets

We conduct experiments on eleven multi-label data sets, and details of the data sets are listed in Table 1. Among these data sets, *Birds* is obtained based on bird sounds collected in field conditions, which consists of 645 ten-second audio recordings associated with 19 species of birds in uncompressed WAV format. *Emotions* contains 593 instances, each of which represents a song labeled with 6 possible emotions. *Scene* contains 2407 images where each image is related to a subset of six semantic scenes. On text, seven *Yahoo* data sets, including *Business, Computers, Education, Entertainment, Health, Reference*, and *Science*, are used in the experiments. All the above data sets are publicly available and widely used in multi-label learning, and can be downloaded from Mulan Library [29]. Moveover, *TCM* is collected from the Second People's Hospital Health Management of Fujian Province in China, which is for the purpose of the multi-label learning task regarding the identification of health-state in traditional Chinese medicine. Specially, this data set is created with 461 symptoms from 1146 patients and has 43 class labels indicating disease natures and locations.

**Table 1**
Characteristics of the benchmark multi-label data sets.

| Data set | #Instances | #Training | #Test | #Features | #Labels |
|---|---|---|---|---|---|
| Birds | 645 | 322 | 323 | 260 | 19 |
| Business | 5000 | 2000 | 3000 | 438 | 30 |
| Computers | 5000 | 2000 | 3000 | 681 | 33 |
| Education | 5000 | 2000 | 3000 | 550 | 33 |
| Emotions | 593 | 391 | 202 | 72 | 6 |
| Entertainment | 5000 | 2000 | 3000 | 640 | 21 |
| Health | 5000 | 2000 | 3000 | 612 | 32 |
| Reference | 5000 | 2000 | 3000 | 793 | 33 |
| Scene | 2407 | 1211 | 1196 | 294 | 6 |
| Science | 5000 | 2000 | 3000 | 743 | 40 |
| TCM | 1146 | 606 | 540 | 461 | 43 |

### 4.2. Experimental evaluation metrics

To evaluate the performance of multi-label feature selection, we use *hamming loss, ranking loss, coverage, average precision, macro-F1*, and *micro-F1* as evaluation metrics [31]. Let $\mathcal{U} = \{(\mathbf{x}_i, Y_i) | 1 \leq i \leq t\}$ be a test set, and $\mathbf{f}(\mathbf{x}_i)$ be a multi-label classifier's predicted label set for unseen instance $\mathbf{x}_i$.

(1) *Hamming loss*: This metric evaluates the average error rate over all the binary labels. Assume that $\oplus$ denotes the symmetric difference between two sets (XOR operation), and $|.|_1$ denotes the $l_1$-norm.

$$HL(\mathbf{f}, \mathcal{U}) = \frac{1}{t} \sum_{i=1}^t \frac{1}{q} |\mathbf{f}(\mathbf{x}_i) \oplus Y_i|_1, \tag{17}$$

(2) *Ranking loss*: This metric evaluates the fraction of reversely ordered label pairs. Assume that $\mathbf{f}_j(\mathbf{x}_i)$ denotes the $j$th entry of $\mathbf{f}(\mathbf{x}_i)$, and $\bar{Y}_i$ denotes the complementary set of $Y_i$ in $L$.

$$RL(\mathbf{f}, \mathcal{U}) = \frac{1}{t} \sum_{i=1}^t \frac{|\{(l_j, l_k) | \mathbf{f}_j(\mathbf{x}_i) \leq \mathbf{f}_k(\mathbf{x}_i), (l_j, l_k) \in Y_i \times \bar{Y}_i\}|}{|Y_i||\bar{Y}_i|}, \tag{18}$$

(3) *Coverage*: This metric evaluates how many steps are needed, on average, to go down the label ranking list so as to cover all the ground-truth labels of the instance. Assume that $rank(\mathbf{x}_i, l_k) = \sum_{j=1}^q [\![\mathbf{f}_j(\mathbf{x}_i) \geq \mathbf{f}_k(\mathbf{x}_i)]\!]$ returns the rank of $l_k$ when all labels in $L$ are sorted in descending order based on $\mathbf{f}$.

$$CV(\mathbf{f}, \mathcal{U}) = \frac{1}{q} \left( \frac{1}{t} \sum_{i=1}^t \max_{l_k \in Y_i} rank(\mathbf{x}_i, l_k) - 1 \right), \tag{19}$$

(4) *Average precision*: This metric evaluates the average fraction of relevant labels ranked higher than a particular label $l_k \in Y_i$.

$$AP(\mathbf{f}, \mathcal{U}) = \frac{1}{t} \sum_{i=1}^t \frac{1}{|Y_i|} \sum_{l_j, l_k \in Y_i} \frac{|L_i = \{l_j | rank(\mathbf{x}_i, l_j) \leq rank(\mathbf{x}_i, l_k)\}|}{rank(\mathbf{x}_i, l_k)}, \tag{20}$$

(5) *Macro-F1*: This metric evaluates a classifier's label set prediction performance, which considers F-measure averaging on each label.

$$MaF(\mathbf{f}, \mathcal{U}) = \frac{1}{q} \sum_{j=1}^q \frac{2 \sum_{i=1}^t y_{ij} \mathbf{f}_j(\mathbf{x}_i)}{\sum_{i=1}^t y_{ij} + \sum_{i=1}^t \mathbf{f}_j(\mathbf{x}_i)}, \tag{21}$$

(6) *Micro-F1*: This metric evaluates a classifier's label set prediction performance, which considers F-measure averaging on the prediction matrix.

$$MiF(\mathbf{f}, \mathcal{U}) = \frac{2 \sum_{i=1}^t |\mathbf{f}(\mathbf{x}_i) \cap Y_i|_1}{\sum_{i=1}^t |Y_i|_1 + \sum_{i=1}^t |\mathbf{f}(\mathbf{x}_i)|_1}, \tag{22}$$

For *hamming loss, ranking loss*, and *coverage*, the smaller value brings to better performance, whereas for *macro-F1, micro-F1*, and *average precision*, the larger value brings to better performance.

### 4.3. Comparing methods

We compare the proposed method against some state-of-the-art multi-label feature selection methods, including MCLS [13], MSSL [5], GLOCAL [40], LLSF [12], GMBA [33], MIFS [15], MDMR [19], and RFS [23], and ML-KNN ($K = 10$) [38] is employed as the classifier to evaluate the performance of the learned feature subset. For MSSL, the parameters $\alpha$ and $\gamma$ are searched in

(a) *Hamming loss*　　　　(b) *Ranking loss*　　　　(c) *Coverage*

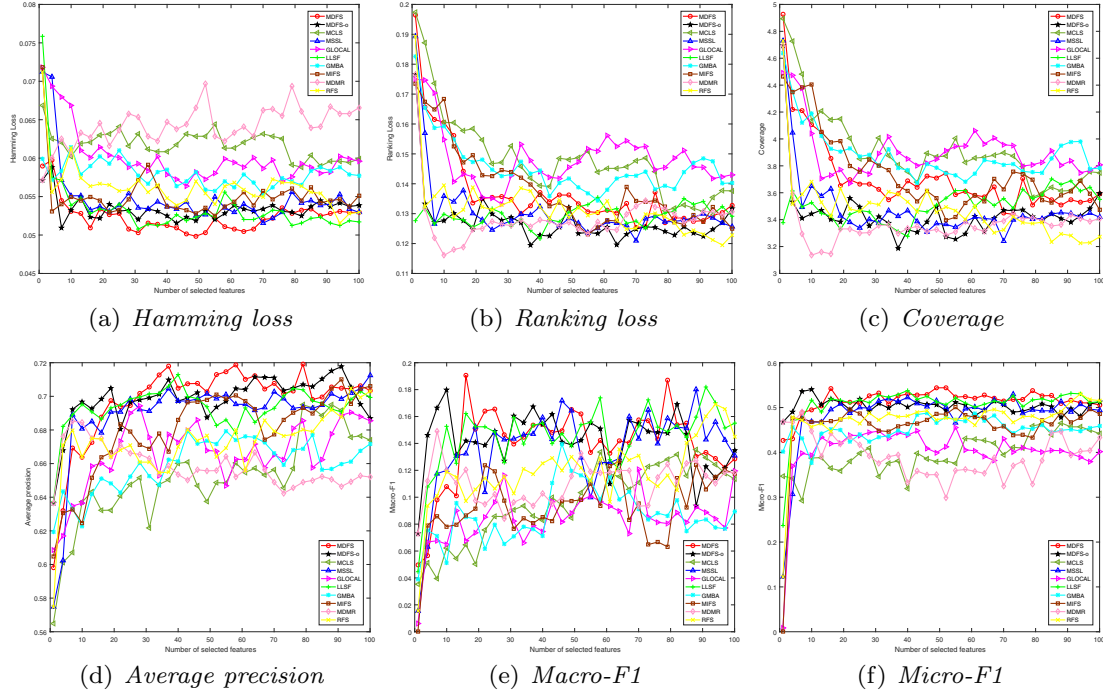(d) *Average precision*　　　(e) *Macro-F1*　　　(f) *Micro-F1*

**Fig. 2.** Comparison results of multi-label feature selection on *Birds*.

$\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. For GLOCAL, the label manifold regularizers $\lambda_3$ and $\lambda_4$ are tuned in $\{10^{-4}, 10^{-3}, \ldots, 1\}$, the latent representation dimensionality $k$ is tuned in $\{5, 10, 15, 20, 30\}$, and the number of clusters $g$ is varied from $\{4, 8, 16, 32, 64\}$. Besides, we define $\lambda = 1$ and $\lambda_2 = 10^{-3}$ in light of GLOCAL is not sensitive to these parameters. For LLSF, $\alpha$ and $\beta$ are searched in $\{2^{-10}, 2^{-9}, \ldots, 2^{10}\}$, and $\rho$ is searched in $\{0.1, 1, 10\}$. As the literature [4] suggested, the parameters $\alpha$, $\beta$, and $\gamma$ of MIFS are tuned in $\{0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 1\}$, and the approximate rank $c$ is tuned in $\{2, \lfloor 0.25q \rfloor, \lfloor 0.5q \rfloor, \lfloor 0.75q \rfloor, q\}$, where $q$ is the number of class labels. For RFS, the parameter $\gamma$ is searched in $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$. MDMR is a parameter-free method, and the parameters of MCLS and GMBA are defined as the default setting. For MDFS, the parameter $\alpha$ is simply set to 1, $\beta$ and $\gamma$ are searched in $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$, and $k$ is fixed to $q-1$ to compute the graph Laplacian. Furthermore, as the literature [10] suggested, PCA is employed as a preprocessing step to preserve 95% energy of multi-label data in all the experiments. The code of MDFS is made available on https://github.com/JiaZhang19/MDFS. For revealing the advantage, we also run a variant of MDFS that uses the original features to achieve feature selection, and call it as MDFS-o.
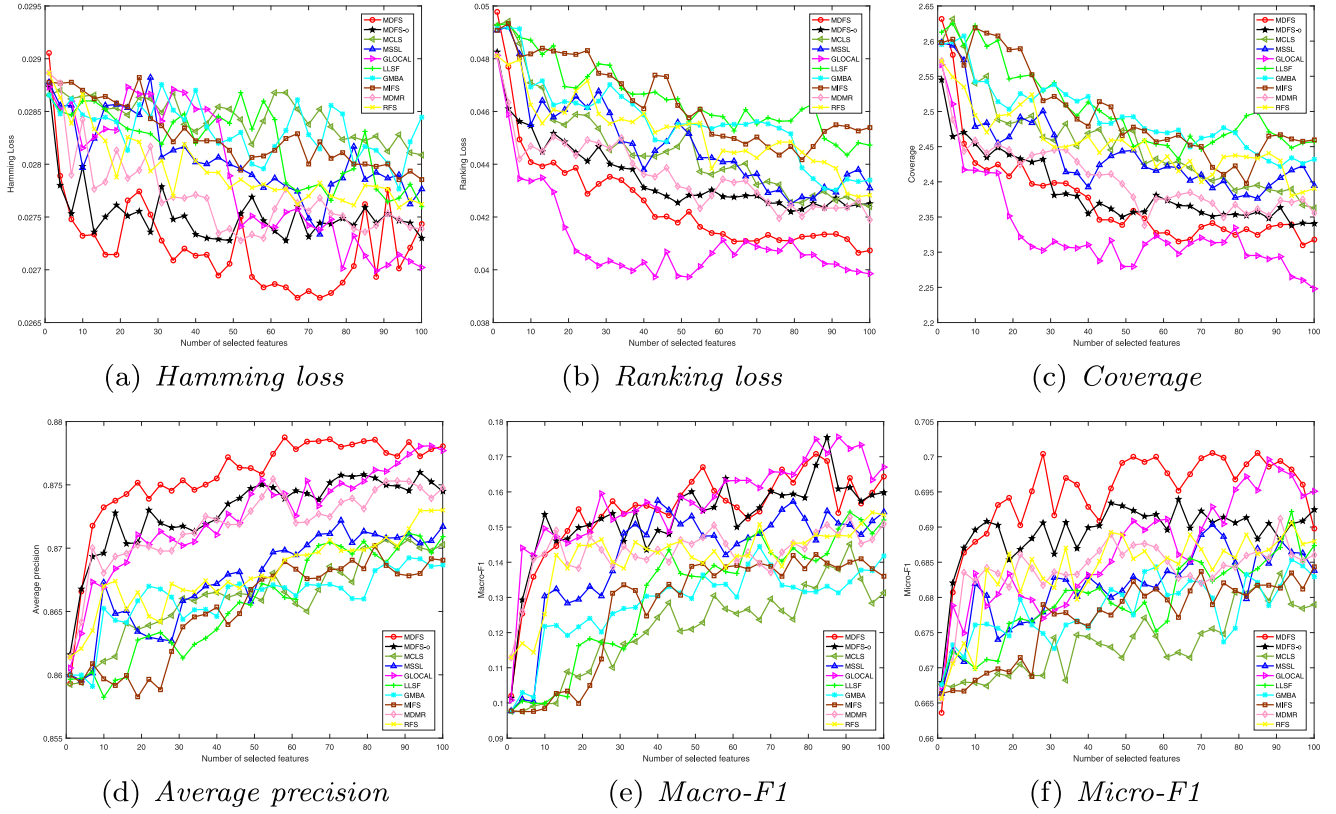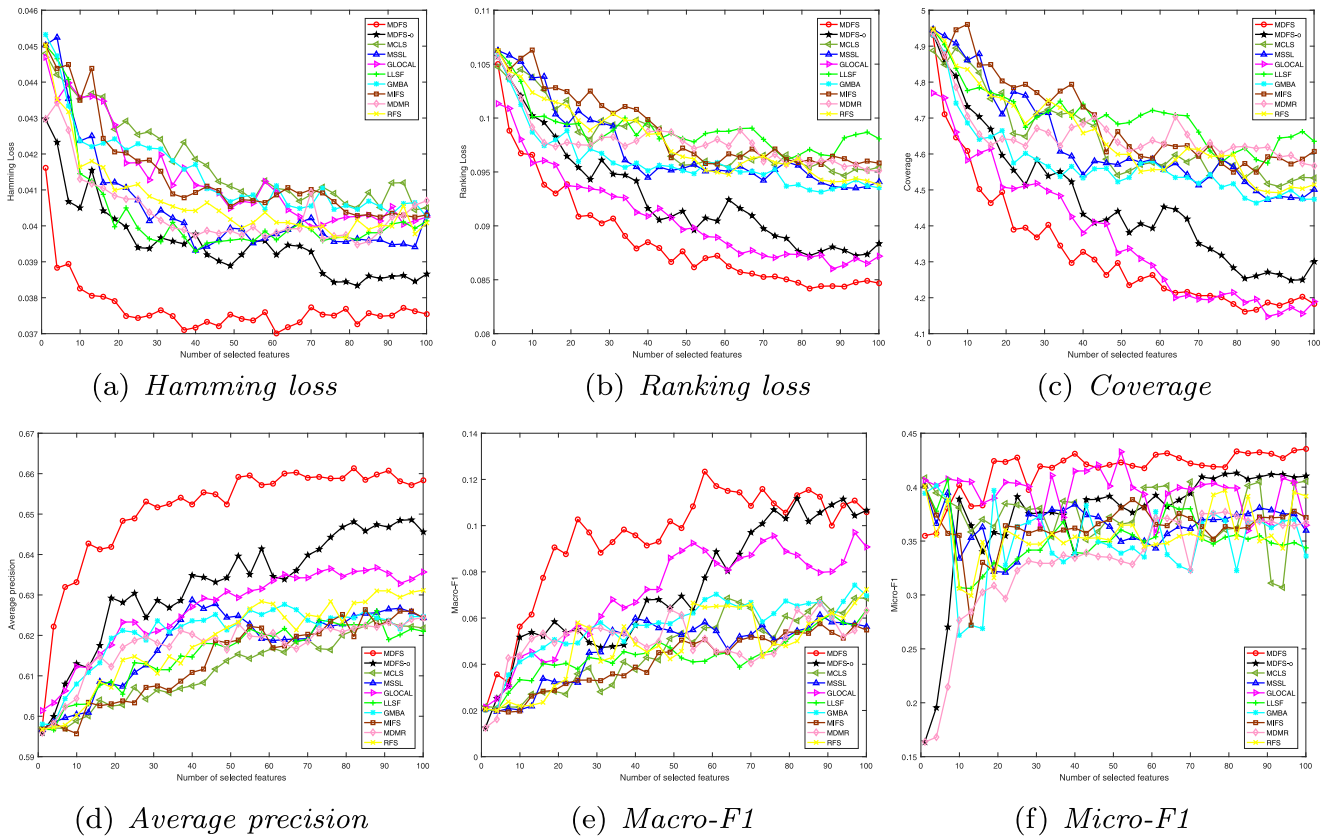
In addition, Top-30 features are sequentially selected to the parameter-tunning for MSSL, GLOCAL, LLSF, MIFS, RFS, MDFS, and MDFS-o, and the best result of each method is reported. Specially, the number of selected features sequentially increases from 1 to 30, and the average *hamming loss* generated by ML-KNN is recorded and used to seek the optimum parameters, i.e., the parameters of each method are selected which makes the average *hamming loss* smallest.

### 4.4. Comparison with other multi-label feature selection methods

Figs. 2–12 demonstrate the performance of different multi-label feature selection methods in terms of *hamming loss, ranking loss, coverage, average precision, macro-F1* and *micro-F1* on the eleven data sets, respectively. In these figures, with Top-100 features (*Emotions* only has 72 features), the learning performance

varies *w.r.t.* the number of selected features. Based on these experimental results from Figs. 2–12, we have a couple of observations. With the increasing of the number of selected features, the performance of MDFS first has a remarkable improvement, and then keeps stable or even degrades. This observation reveals that it is meaningful to conduct feature selection for multi-label learning. In addition, MDFS can achieve good learning performance on most of the data sets even when a few number of features are selected. This observation reveals that the proposed method is effective in practice.

Furthermore, we compare MDFS with the comparing methods on each data set. As shown in Figs. 2–12, MDFS achieves better performance against MCLS, MSSL, GLOCAL, LLSF, GMBA, MIFS, MDMR, and RFS in terms of each evaluation metric in most cases. To be specific, MDFS significantly outperforms these comparing methods on *Computers* (Fig. 4), *Education* (Fig. 5), *Scene* (Fig. 10), and *Science* (Fig. 11), and (can obtain better feature selection results on *Emotions* (Fig. 6) and *TCM* (Fig. 12). As shown in Fig. 2, MDFS is inferior to some multi-label feature selection methods on *Birds* in terms of *ranking loss* and *coverage*, such as MDMR and MSSL, but it achieves statistically superior or at least comparable performance against these comparing methods on the other four metrics. Fig. 3 shows the feature selection results on *Business*, from which we can see that MDFS has the best results on *hamming loss, average precision*, and *micro-F1*. In addition, MDFS achieves a comparable performance with GLOCAL in consideration of *macro-F1*, while on *ranking loss* and *coverage*, the performance of MDFS is worse than GLOCAL. Fig. 7 shows the comparison results of the multi-label feature selection methods on *Entertainment*, and we can observe from Fig. 7 that MDFS significantly outperforms the comparing methods in terms of *ranking loss, coverage*, and *average precision*, and obtains better performance on the other metrics when the number of selected features is less than 30. However, when the number of selected features is more than 30, MDFS performs worse than LLSF on *hamming loss* and *micro-F1*. Figs. 8 and 9 show the feature selection results on *Health* and *Reference* respectively, and we can observe that MDFS significantly

(a) *Hamming loss*              (b) *Ranking loss*              (c) *Coverage*

(d) *Average precision*         (e) *Macro-F1*                 (f) *Micro-F1*

**Fig. 3.** Comparison results of multi-label feature selection on *Business*.



(a) *Hamming loss*              (b) *Ranking loss*              (c) *Coverage*

(d) *Average precision*         (e) *Macro-F1*                 (f) *Micro-F1*

**Fig. 4.** Comparison results of multi-label feature selection on *Computers*.

(a) *Hamming loss*     (b) *Ranking loss*     (c) *Coverage*

(d) *Average precision*     (e) *Macro-F1*     (f) *Micro-F1*
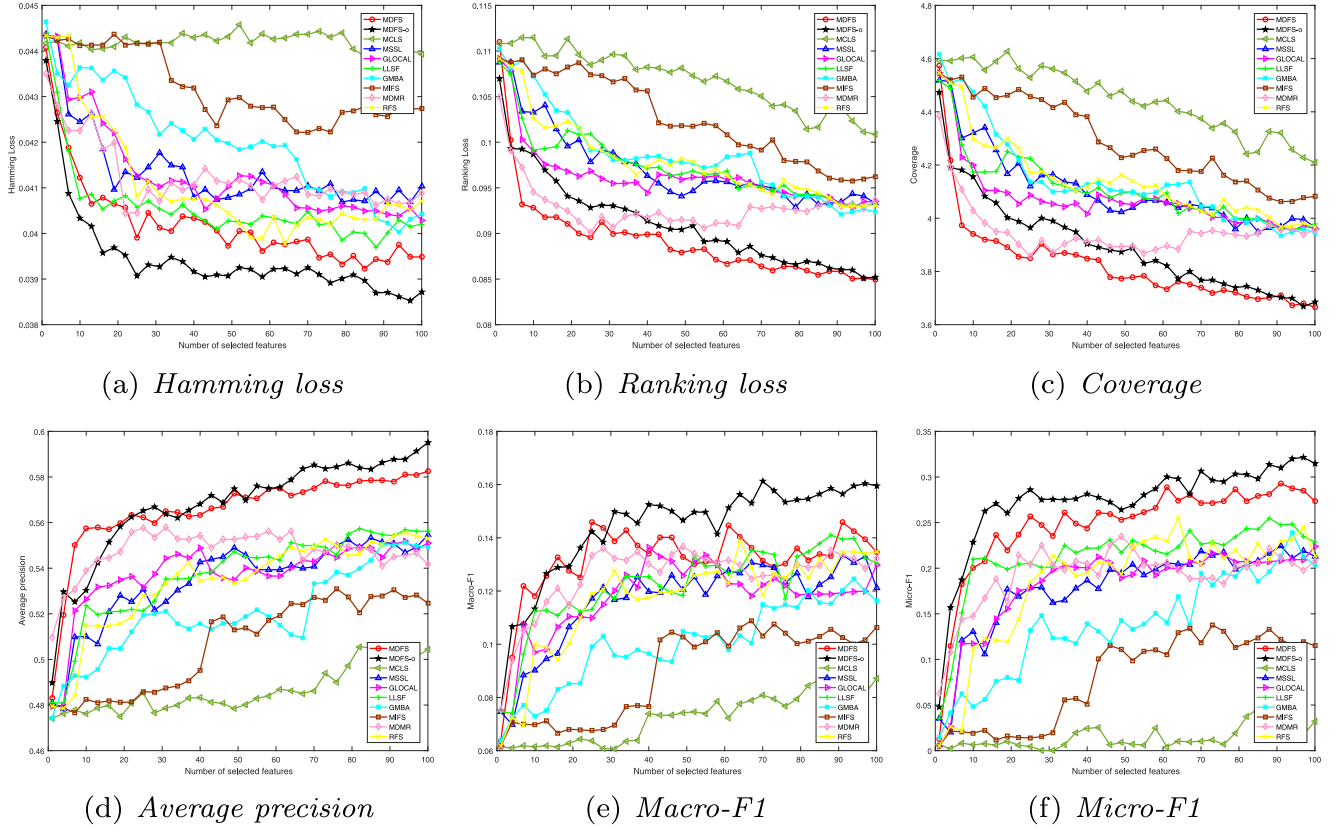
**Fig. 5.** Comparison results of multi-label feature selection on *Education*.



(a) *Hamming loss*     (b) *Ranking loss*     (c) *Coverage*

(d) *Average precision*     (e) *Macro-F1*     (f) *Micro-F1*
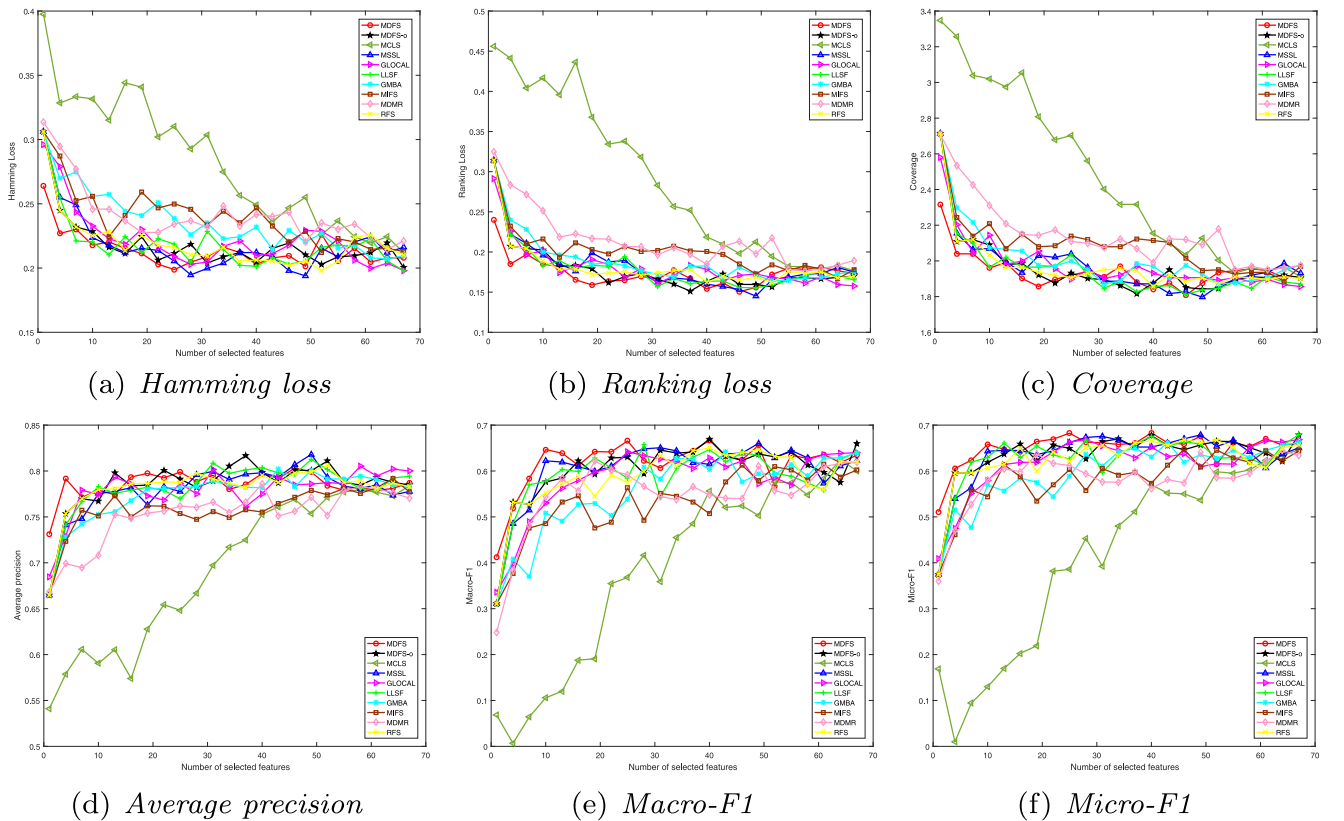
**Fig. 6.** Comparison results of multi-label feature selection on *Emotions*.

(a) *Hamming loss*  (b) *Ranking loss*  (c) *Coverage*

(d) *Average precision*  (e) *Macro-F1*  (f) *Micro-F1*
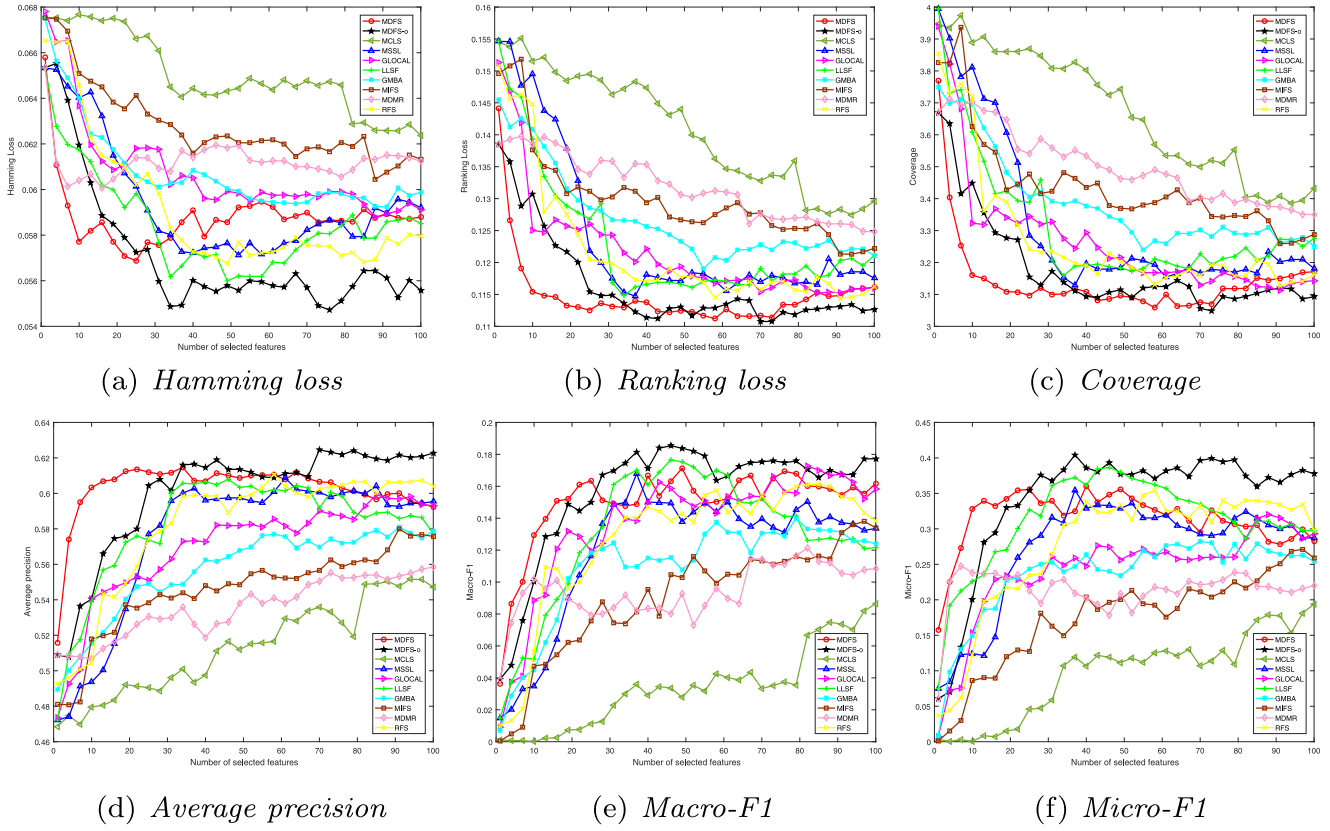
**Fig. 7.** Comparison results of multi-label feature selection on *Entertainment*.



(a) *Hamming loss*  (b) *Ranking loss*  (c) *Coverage*

(d) *Average precision*  (e) *Macro-F1*  (f) *Micro-F1*
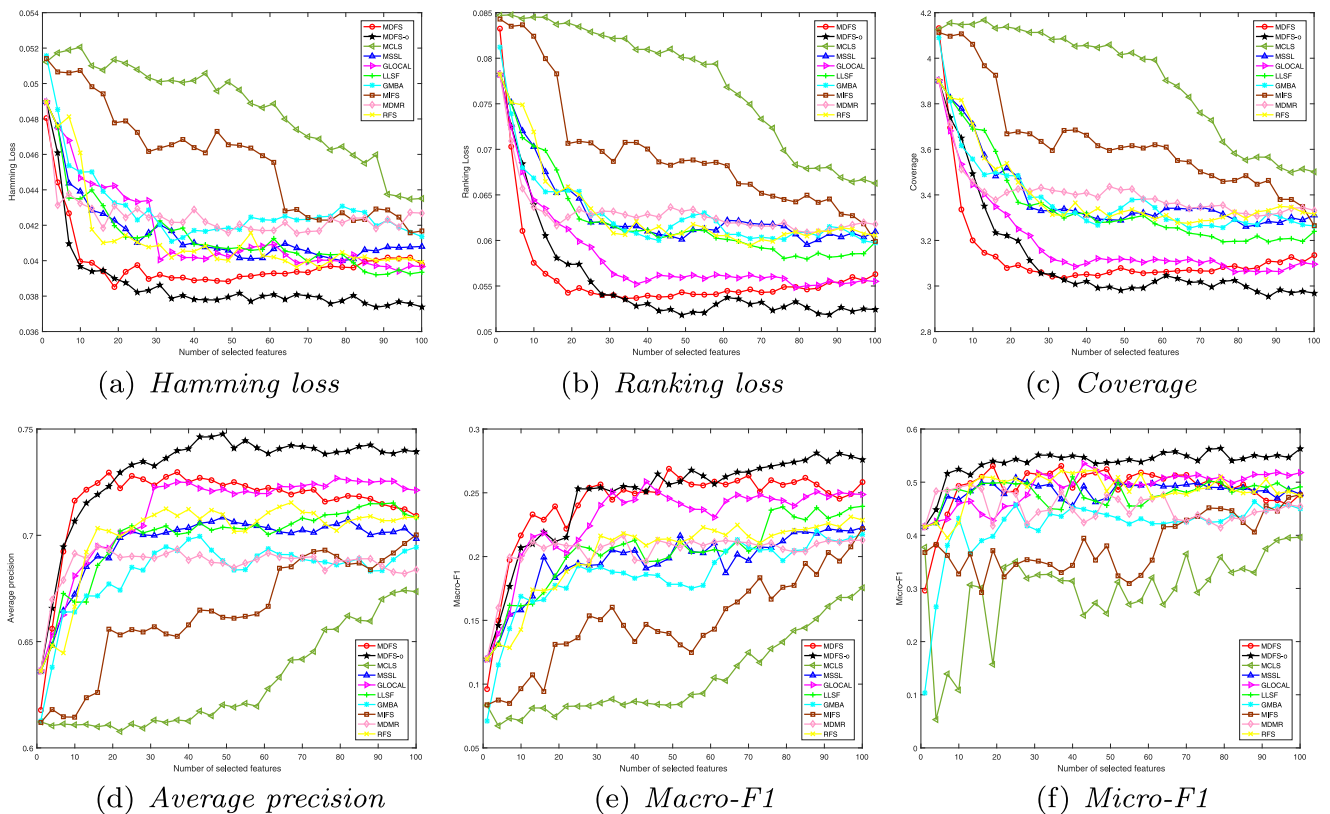
**Fig. 8.** Comparison results of multi-label feature selection on *Health*.
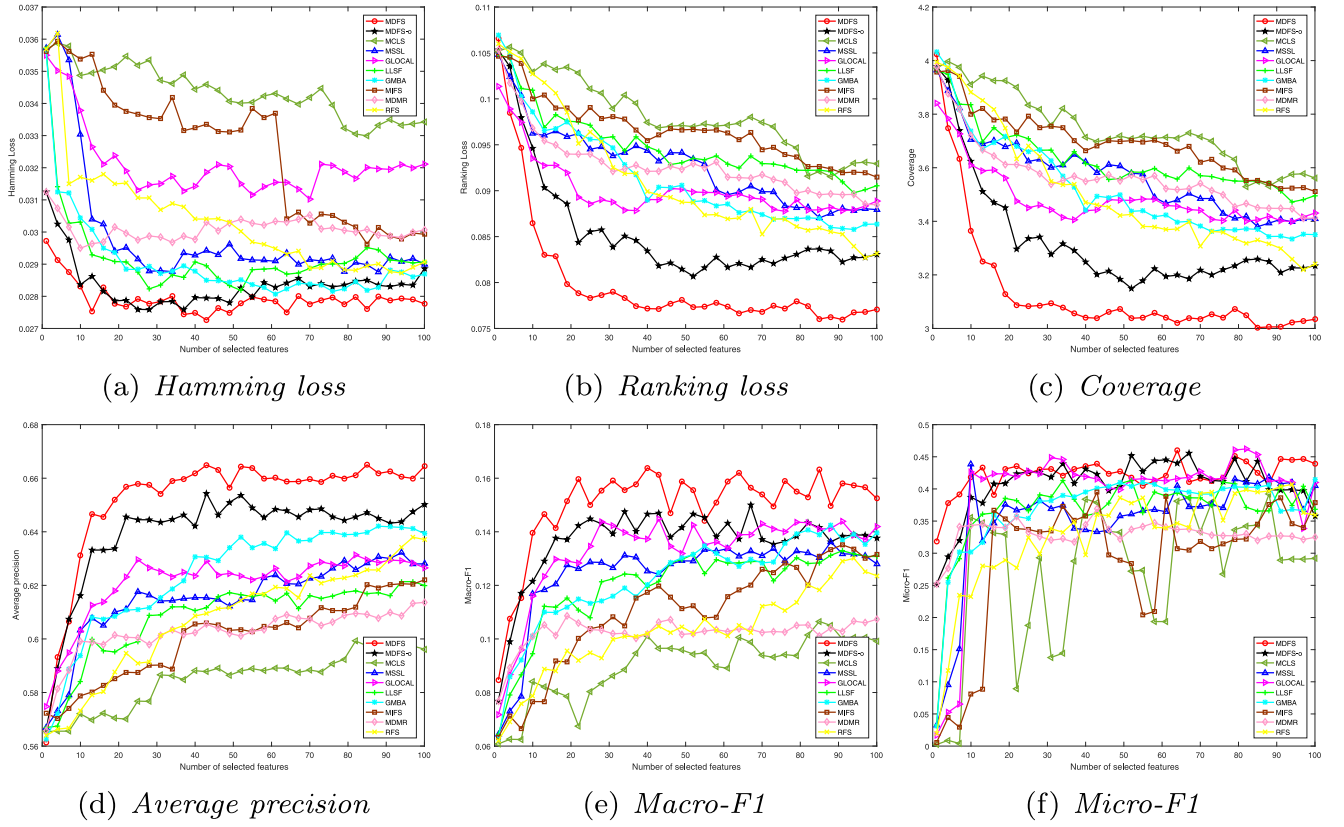
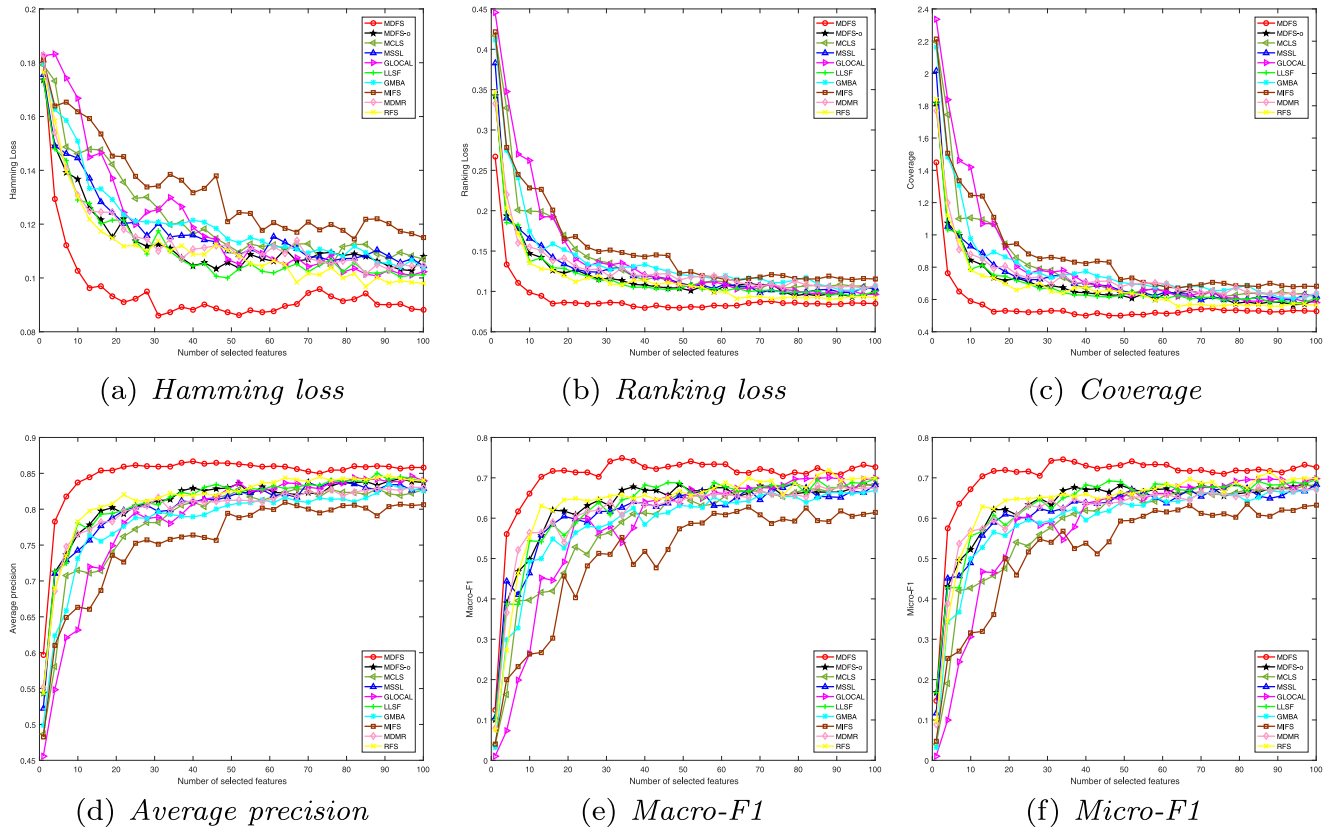**Fig. 9.** Comparison results of multi-label feature selection on *Reference*.



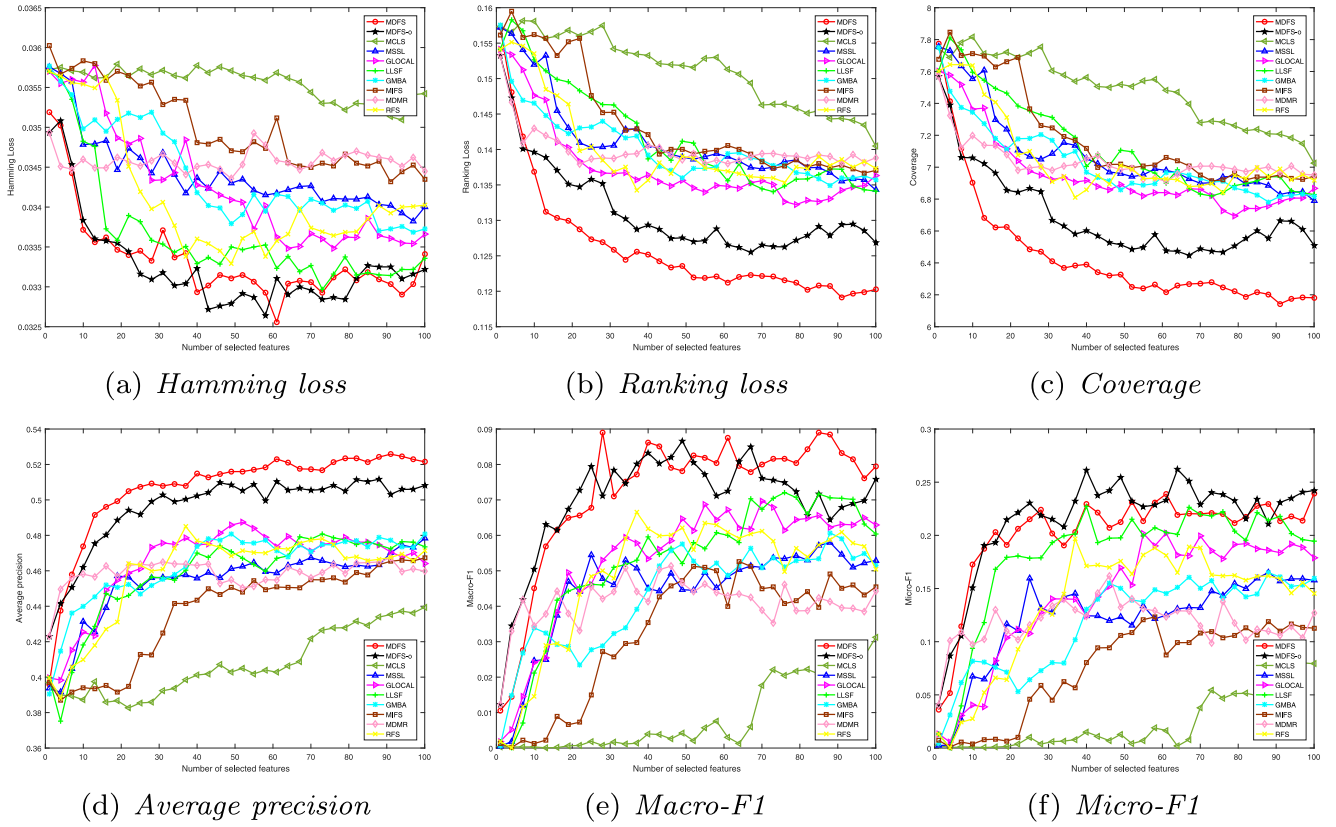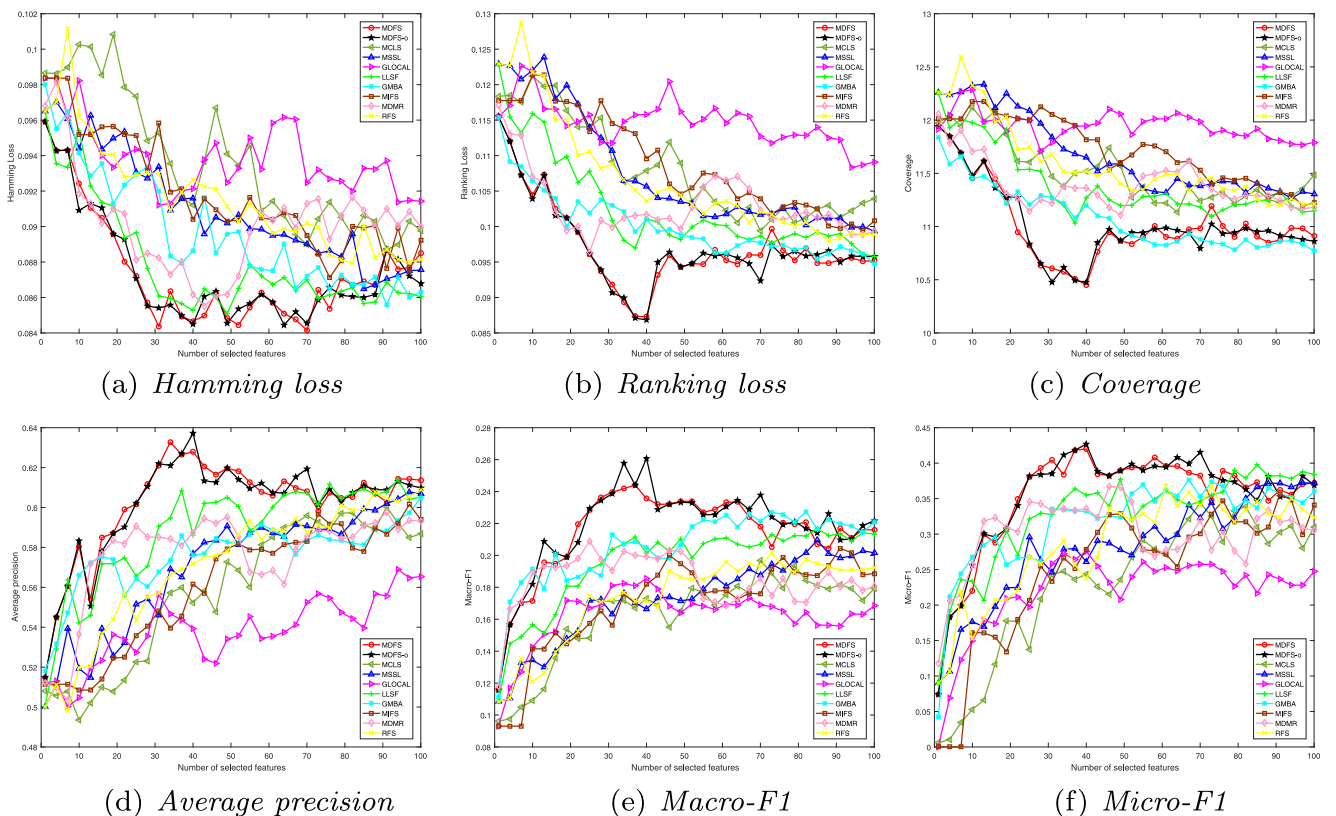**Fig. 10.** Comparison results of multi-label feature selection on *Scene*.

(a) *Hamming loss*     (b) *Ranking loss*     (c) *Coverage*

(d) *Average precision*     (e) *Macro-F1*     (f) *Micro-F1*

**Fig. 11.** Comparison results of multi-label feature selection on *Science*.



(a) *Hamming loss*     (b) *Ranking loss*     (c) *Coverage*

(d) *Average precision*     (e) *Macro-F1*     (f) *Micro-F1*

**Fig. 12.** Comparison results of multi-label feature selection on *TCM*.

**Table 2**
Summary of the Friedman statistics $F_F$ ($k = 10, N = 11$) and the critical value on evaluation metrics($k$: comparing methods; $N$: data sets).

| Evaluation metric | $F_F$ | Critical value ($\alpha = 0.05$) |
|---|---|---|
| Hamming loss | 32.5875 | 2.0 |
| Ranking loss | 14.4909 | |
| Coverage | 11.8675 | |
| Average precision | 19.7985 | |
| Macro-F1 | 21.3915 | |
| Micro-F1 | 22.3687 | |

outperforms MCLS, MSSL, GLOCAL, LLSF, GMBA, MIFS, MDMR, and RFS except on *micro-F1*. Moreover, we can see from Figs. 2–12 that MDFS-o can obtain good performance, which compares favorably with MDFS on some text data sets, such as *Education* and *Health*, but MDFS performs better in most cases, and is capable to handle various tasks from different domains effectively, such as audio concept detection (Fig. 6) and image annotation (Fig. 10).

Based on the observations, we can conclude that the proposed method benefits to the performance with feature selection, and has the advantages compared with some other well-established multi-label feature selection methods.

To further analyze the performance among all the methods, Top-30 features of each method are selected as the feature subset for multi-label learning, and Friedman test [7] is used as the favorable statistical significance test for the method comparison on the eleven data sets. Table 2 illustrates the Friedman statistic $F_F$ and the corresponding critical value on each metric, and we can see from Table 2 that the null hypothesis, which follows the principle that all the methods have equal performance, is clearly rejected in terms of each metric at significance level $\alpha = 0.05$. Thus, we proceed with certain post-hoc test [7] to complete the performance analysis, and the Nemenyi test [7] is utilized to serve this purpose where MDFS or MDFS-o is regarded as the control method respectively. Further, the difference between two methods is dis-

tinguished with the critical difference (CD), as follow:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}, \tag{23}$$

where $q_\alpha = 3.164$ at significance level $\alpha = 0.05$, and then we can calculate $CD = 4.0847$ ($k = 10, N = 11$).

Fig. 13 shows the CD diagrams [7] *w.r.t.* each metric. To be specific, any comparing method whose average rank is within one CD to that of MDFS or MDFS-o is connected. Otherwise, the method, which is not connected with MDFS or MDFS-o, is considered to have the significant different performance with the control method. From Fig. 13, we can see that MDFS is superior to MCLS, MIFS, and GMBA, and obtains better performance than GLOCAL on *hamming loss*, and outperforms LLSF on *ranking loss* and *average precision*. Compared with MSSL, MDFS has better performance on *ranking loss, average precision*, and *macro-F1*. Compared with MDMR, MDFS has the advantages on *hamming loss, average precision, macro-F1*, and *micro-F1*, while on *hamming loss, average precision*, and *macro-F1*, MDFS performs better compared with RFS. Note that MDFS and its variant MDFS-o have no significant difference on all the metrics, but MDFS has the higher ranking than MDFS-o on each metric. Thus, we conclude that MDFS can achieve highly competitive performance against MCLS, MSSL, GLOCAL, LLSF, GMBA, MIFS, MDMR, and RFS.

### 4.5. Sensitivity to parameters

In this section, the influence of parameters $\beta$ and $\gamma$ is studied for the proposed method. Specifically, with the increasing of the number of selected features from 1 to 30, we calculate the average *hamming loss*, as follow:

$$ave - HL(\beta, \gamma) = \frac{1}{30} \sum_{k=1}^{30} HL_k(\mathbf{f}, \mathcal{U}), \tag{24}$$

where $HL_i(\mathbf{f}, \mathcal{U})$ denotes the result of *hamming loss* when Top-$k$ features are selected as input, and $\mathbf{f}$ denotes the multi-label
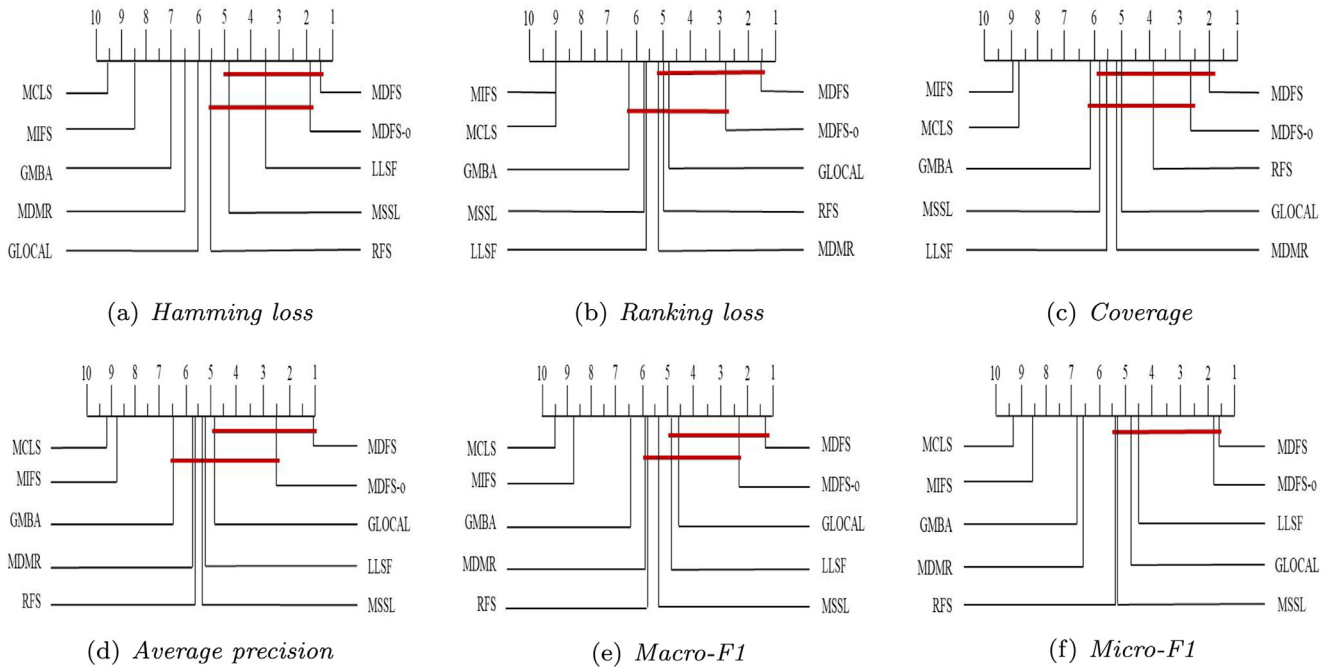


(a) *Hamming loss*     (b) *Ranking loss*     (c) *Coverage*

(d) *Average precision*     (e) *Macro-F1*     (f) *Micro-F1*
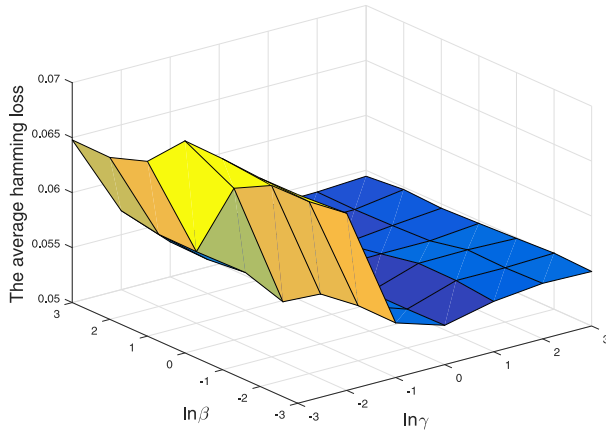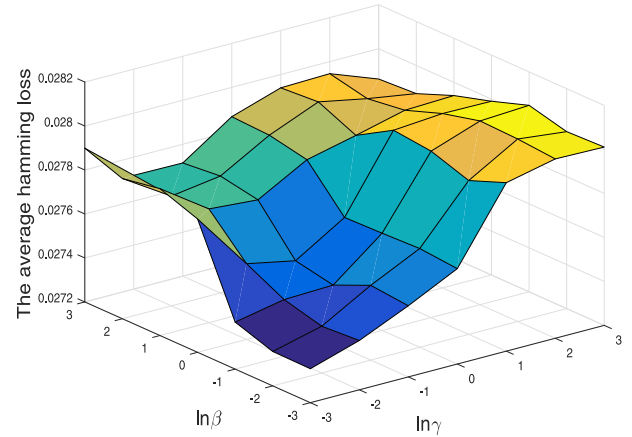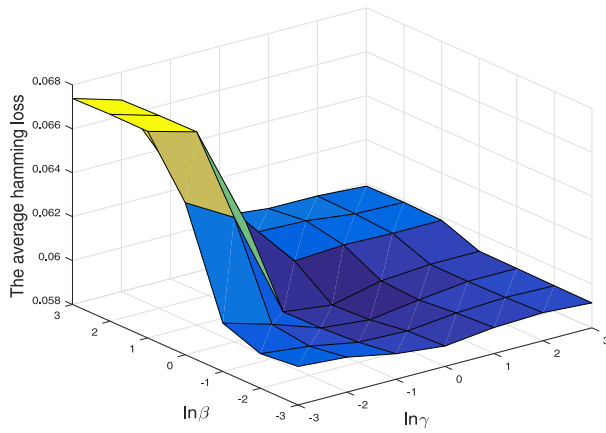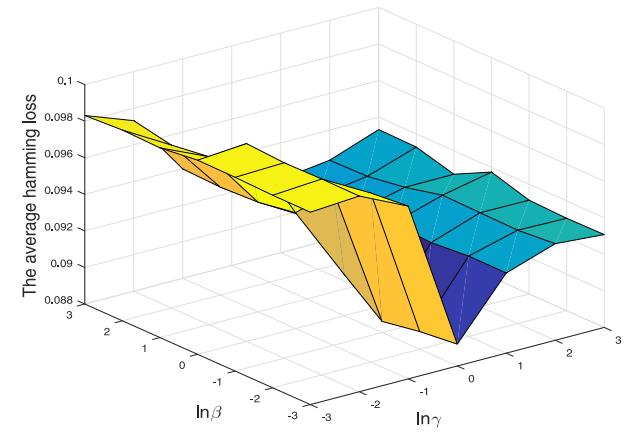
**Fig. 13.** Comparison of MDFS or MDFS-o (control method) against other comparing methods with the Nemenyi test.

(a) *Birds*



(b) *Business*



(c) *Entertainment*



(d) *TCM*

**Fig. 14.** The average *hamming loss* when different pairs of $\beta$ and $\gamma$ are employed.

**Table 3**
Running time (sec) of the multi-label feature selection methods.

| Data set | MCLS | MSSL | GLOCAL | LLSF | GMBA | MIFS | MDMR | RFS | MDFS |
|---|---|---|---|---|---|---|---|---|---|
| Birds | 0.70 | 0.11 | 6.23 | 0.12 | 9.97 | 1.57 | 183.23 | 0.23 | 0.17 |
| Business | 9.32 | 0.45 | 37.18 | 0.16 | 610.47 | 22.91 | 2533.24 | 2.13 | 12.52 |
| Scene | 1.88 | 0.25 | 10.92 | 0.15 | 143.63 | 5.79 | 165.59 | 0.74 | 5.12 |
| TCM | 2.76 | 0.32 | 14.48 | 0.14 | 51.78 | 3.07 | 1515.12 | 1.04 | 0.64 |

classifier, i.e., MLKNN. The average *hamming loss* changes when different pairs of $\beta$ and $\gamma$ are employed. Fig. 14(a)–(d) show the influence of $\beta$ and $\gamma$ learned by the proposed method on *Birds*, *Business*, *Entertainment*, and *TCM* respectively, from which it can be inferred that the performance is generally better when $\beta$ and $\gamma$ are comparable on *Business*, *Entertainment*, and *TCM*. However, $\beta$ is suggested to be larger than $\gamma$ for enhancing the influence of the label correlations when the class labels in multi-label data, such as *Birds*, have strong co-occurrence relationships with each other, and vice versa.

### 4.6. Convergence analysis

As stated in Section 3.3, the convergence of the proposed method is guaranteed, and we further display the convergence rate by using different scale data sets. Fig. 15(a) and (b) show the change of the objective value *w.r.t.* each iteration on *Birds* and *Entertainment* respectively, and we can observe that the objective decreases fast in a few iterations and then reaches a stable stage, showing that the proposed method converges quickly in practice.

### 4.7. Running time comparison

In Table 3, we show the timing results (in second) on multi-label feature selection. MSSL, LLSF, MIFS, RFS, and MDFS train a linear classifier with sparse representation, and have their own strength to achieve the purpose. Compared with MSSL and LLSF, MDFS requires additional steps to obtain local structures of feature (or label) space, hence it is slower than the two methods. For the similar reason, MDFS is less efficient than RFS on large-scale data analysis. MIFS takes the advantage by using a low-rank label correlation matrix, but it has low convergence rate with iterative optimization. Thus, MDFS is more computationally efficient than MIFS. Furthermore, we can see from Table 3 that MDFS outperforms GLOCAL (Manifold based method with global and local label correlations), GMBA (graph-margin based method for the exploitation of *high-order* label correlations), and MDMR (greedy mutual information based method) in running time comparison. Besides, compared with MCLS, MDFS is more efficient while the data size is small.
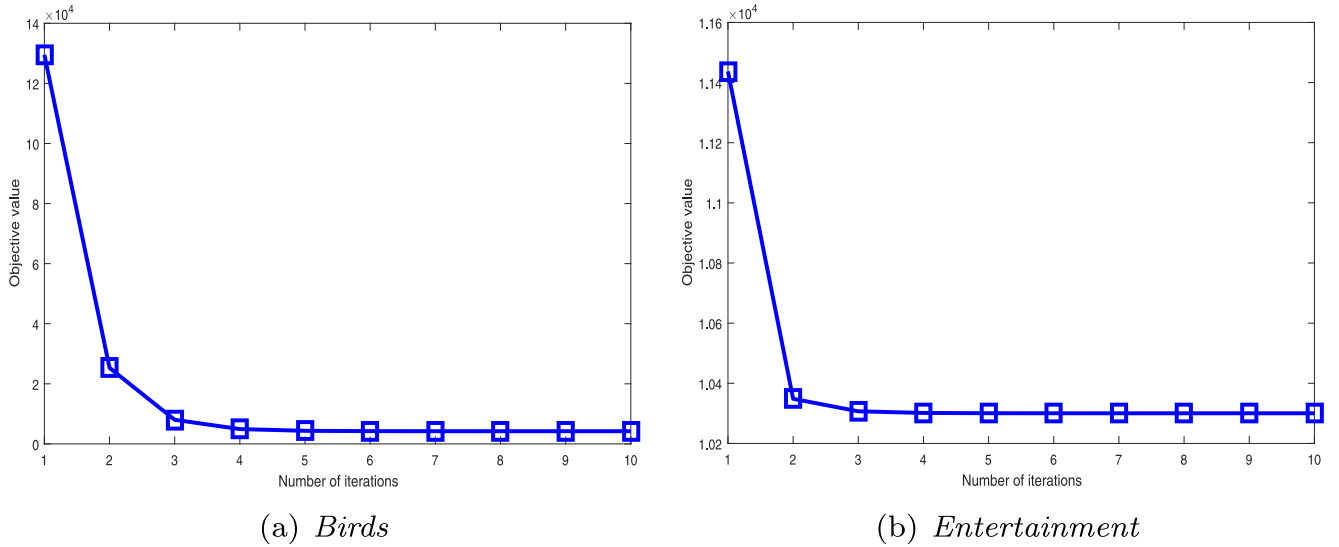
(a) *Birds*　　　　　　　　　　　(b) *Entertainment*

**Fig. 15.** Convergence analysis. The objective value in Eq. (6) using Algorithm 1 *w.r.t.* the number of iterations.

## 5. Discussion and conclusion

In this paper, we introduced a novel manifold regularized optimization framework for multi-label feature selection. The proposed optimization framework has two appealing properties. First, it made use of manifold regularization to generate the low-dimensional embedding from the original feature space for the local and global label correlations exploitation. Second, by involving $l_{2,1}$-norm regularization into the learning framework, the feature selection procedure was executed to search for discriminative features for multi-label learning. Aimed to achieve the above purposes, an efficient alternating optimization algorithm was developed to solve the optimization problem with convexity. Empirical studies on various real-world multi-label data sets demonstrated the efficiency and efficacy of the proposed method. Compared with some other state-of-the-art multi-label feature selection methods, the proposed method had the advantages in view of the performance. By the parameter-tunning, the performance of the proposed method had a general improvement by resolving label correlations.

In the future, it is interesting to research the strategy of *high-order* correlations, and we will also explore how to perform feature selection incorporating the label correlations information for weakly supervised multi-label learning.

## References

[1] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from labeled and unlabeled examples„ J. Mach. Learn. Res. 7 (2006) 2399–2434.

[2] M.R. Boutell, J. Luo, X. Shen, C.M. Brown, Learning multi-label scene classification, Pattern Recognit. 37 (9) (2004) 1757–1771.

[3] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

[4] A. Braytee, W. Liu, D.R. Catchpoole, P.J. Kennedy, Multi-label feature selection using correlation information, in: Proceedings of the ACM Conference on Information and Knowledge Management, 2017, pp. 1649–1656.

[5] Z. Cai, W. Zhu, Multi-label feature selection via feature manifold learning and sparsity regularization, Int. J. Mach. Learn. Cybern. 9 (8) (2018) 1321–1334.

[6] L. Dai, J. Zhang, C. Li, C. Zhou, S. Li, Multi-label feature selection with application to TCM state identification, Concurr. Comput. Pract. Exp. doi:10.1002/cpe.4634.

[7] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.

[8] A. Elisseeff, J. Weston, A kernel method for multi-labelled classification, in: Advances in Neural Information Processing Systems 14, 2001, pp. 681–687.

[9] J. Fürnkranz, E. Hüllermeier, E.L. Mencía, K. Brinker, Multilabel classification via calibrated label ranking, Mach. Learn. 73 (2) (2008) 133–153.

[10] X. He, S. Yan, Y. Hu, P. Niyogi, H. Zhang, Face recognition using laplacianfaces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 328–340.

[11] P. Hou, X. Geng, M. Zhang, Multi-label manifold learning, in: Proceedings of the 30th AAAI Conference on Artificial Intelligence, 2016, pp. 1680–1686.

[12] J. Huang, G. Li, Q. Huang, X. Wu, Learning label-specific features and class-dependent labels for multi-label classification, IEEE Trans. Knowl. Data Eng. 28 (12) (2016) 3309–3323.

[13] R. Huang, W. Jiang, G. Sun, Manifold-based constraint laplacian score for multi-label feature selection, Pattern Recognit. Lett. 112 (2018) 346–352.

[14] S. Huang, Z. Zhou, Multi-label learning by exploiting label correlations locally, in: Proceedings of the 26th AAAI Conference on Artificial Intelligence, 2012, pp. 945–955.

[15] L. Jian, J. Li, K. Shu, H. Liu, Multi-label informed feature selection, in: Proceedings of the 25th International Joint Conference on Artificial Intelligence, 2016, pp. 1627–1633.

[16] F. Li, D. Miao, W. Pedrycz, Granular multi-label feature selection based on mutual information, Pattern Recognit. 67 (2017) 410–423.

[17] J. Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang, H. Liu, Feature selection: a data perspective, ACM Comput. Surv. 50 (6) (2017). 94:1–94:45

[18] Y. Li, T. Li, H. Liu, Recent advances in feature selection and its applications, Knowl. Inf. Syst. 53 (3) (2017) 551–577.

[19] Y. Lin, Q. Hu, J. Liu, J. Duan, Multi-label feature selection based on max-dependency and min-redundancy, Neurocomputing 168 (2015) 92–103.

[20] Y. Lin, Q. Hu, J. Zhang, X. Wu, Multi-label feature selection with streaming labels, Inf. Sci. 372 (2016) 256–275.

[21] J. Liu, Y. Lin, M. Lin, S. Wu, J. Zhang, Feature selection based on quality of information, Neurocomputing 225 (2017) 11–22.

[22] Y. Liu, K. Wen, Q. Gao, X. Gao, F. Nie, SVM based multi-label learning with missing labels for image annotation, Pattern Recognit. 78 (2018) 307–317.

[23] F. Nie, H. Huang, X. Cai, C.H.Q. Ding, Efficient and robust feature selection via joint l2, 1-norms minimization, in: Advances in Neural Information Processing Systems, 23, 2010, pp. 1813–1821.

[24] F. Nie, D. Xu, I.W. Tsang, C. Zhang, Flexible manifold embedding: a framework for semi-supervised and unsupervised dimension reduction, IEEE Trans. Image Process. 19 (7) (2010) 1921–1932.

[25] H. Peng, F. Long, C.H.Q. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.

[26] R.E. Schapire, Y. Singer, Boostexter: a boosting-based system for text categorization, Mach. Learn. 39 (2/3) (2000) 135–168.

[27] Z. Sun, J. Zhang, L. Dai, C. Li, C. Zhou, J. Xin, S. Li, Mutual information based multi-label feature selection via constrained convex optimization, Neurocomputing 329 (2019) 447–456.

[28] G. Tsoumakas, I. Katakis, I.P. Vlahavas, Mining multi-label data, in: Data Mining and Knowledge Discovery Handbook, Second ed., 2010, pp. 667–685.

[29] G. Tsoumakas, E.S. Xioufis, J. Vilcek, I.P. Vlahavas, MULAN: a java library for multi-label learning, J. Mach. Learn. Res. 12 (2011) 2411–2414.

[30] B. Wu, Z. Liu, S. Wang, B. Hu, Q. Ji, Multi-label learning with missing labels, in: Proceedings of the 22nd International Conference on Pattern Recognition, 2014, pp. 1964–1968.

[31] X. Wu, Z. Zhou, A unified view of multi-label performance measures, in: Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 3780–3788.

[32] S. Xie, X. Kong, J. Gao, W. Fan, P.S. Yu, Multilabel consensus classification, in: Proceedings of the 13th IEEE International Conference on Data Mining,, 2013, pp. 1241–1246.

[33] P. Yan, Y. Li, Graph-margin based multi-label feature selection, in: Proceedings of the European Conference on Machine Learning, 2016, pp. 540–555.

[34] J. Zhang, C. Li, D. Cao, Y. Lin, S. Su, L. Dai, S. Li, Multi-label learning with label-specific features by resolving label correlations, Knowl.-Based Syst. 159 (2018) 148–157.

[35] J. Zhang, C. Li, Z. Sun, Z. Luo, C. Zhou, S. Li, Towards a unified multi–source-based optimization framework for multi-label learning, Appl. Soft Comput. 76 (2019) 425–435.

[36] M. Zhang, J.M.P. Sánchez, V. Robles, Feature selection for multi-label naive Bayes classification, Inf. Sci. 179 (19) (2009) 3218–3229.

[37] M. Zhang, Z. Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (8) (2014) 1819–1837.

[38] M. Zhang, Z. Zhou, ML-KNN: a lazy learning approach to multi-label learning, Pattern Recognit. 40 (7) (2007) 2038–2048.

[39] Z. Zhang, Y. Xu, J. Yang, X. Li, D. Zhang, A survey of sparse representation: algorithms and applications, IEEE Access 3 (2015) 490–530.

[40] Y. Zhu, J.T. Kwok, Z. Zhou, Multi-label learning with global and local label correlation, IEEE Trans. Knowl. Data Eng. 30 (6) (2018) 1081–1094.

**Jia Zhang** received the M.S. degree from the School of Computer Science, Minnan Normal University, Zhangzhou, China, in 2016. He is currently working towards the Ph.D. degree from the Cognitive Science Department, Xiamen University, Xiamen, China. His research interests include weakly multi-label learning, information fusion, feature selection, and artificial intelligence in medicine.

**Zhiming Luo** received the B.S. degree from the Cognitive Science Department, Xiamen University, Xiamen, China, in 2011; the Ph.D. degree in computer science with Xiamen University and University of Sherbrooke, Sherbrooke, QC, Canada, in 2017. His research interests include traffic surveillance video analytics, computer vision, and machine learning.

**Candong Li** is the president of Fujian University of Traditional Chinese Medicine (FUTCM) as well as the chairman of the Research Base of TCM Syndrome in FUTCM. Prof Li graduated from FUTCM in 1984 and became a M.D in TCM. He got his M.S. degree in 1991 and Ph.D. degree in 2002, and became a full Prof from 2003. Prof Li is the leading researcher in the field of diagnostics of TCM (DTCM). He is the director of the committee of DTCM in the Chinese Society of TCM along with the chairman of the health management branch of World Federation of TCM. He is a member of the WHO ICD-11 expert group. He is a supervisor and the principle investigator (PI) of 30 research programs including a National Program on Key Basic Research Pro-ject of China (973 Program). Till 2018 April, he achieved 270 publications in journals and conferences. Nowadays, his research focuses on the fields of healthy management based on theories of TCM and DTCM.

**Changen Zhou** received the M.S. degree from the School of Computer Science, Fuzhou University, Fuzhou, China, in 2007. His research interests include machine learning and data mining.

**Shaozi Li** received the B.S. degree from Hunan University, Changsha, China; the M.S. degree from Xi'an Jiaotong University, Xi'an, China; and the Ph.D. degree from the National University of Defense Technology, Changsha. He is the Chair and Professor of the Cognitive Science Department, Xiamen University, Xiamen, China, the Vice Director of the Technical Committee on Collaborative Computing of CCF, and the Vice Director of the Fujian Association of Artificial Intelligence. He has directed and completed more than 20 research projects, including several national 863 Programs, National Nature Science Foundation of China, and the Ph.D. Programs Foundation of the Ministry of Education of China. Furthermore, he has authored nearly 300 papers in journals and international conferences. His research interests include artificial intelligence and its applications, moving objects detection and recognition, machine learning, computer vision, and multimedia information retrieval. Dr. Li is a Senior Member of ACM and the China Computer Federation.