Robust multi-label feature selection with shared label enhancement

Juncheng Hu (jchu19@mails.jlu.edu.cn)
Jilin University
Yonghao Li
Jilin University
Wanfu Gao
Jilin University

Research Article

Keywords: Feature selection, Multi-label learning, Label enhancement, Graph regularization

Posted Date: April 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1537380/v1

Robust multi-label feature selection with shared label enhancement

Yonghao Li[1], Juncheng Hu[1*†] and Wanfu Gao[1†]

[1]College of Computer Science and Technology, Jilin University, Changchun, 130012, China.

*Corresponding author(s). E-mail(s): jchu19@mails.jlu.edu.cn;
Contributing authors: yonghao17@mails.jlu.edu.cn;
gaowf@jlu.edu.cn;
†These authors contributed equally to this work.

Abstract

In real-world applications, multi-label feature selection has been widely attract considerable attention due to the importance of multi-label data. However, previous methods do not fully consider the relationship between the feature set and the multi-label set but devote attention to either of them. In addition, the existence of irrelevant and redundant information in the feature set and the multi-label set makes previous methods obtain inaccurate results. Moreover, traditional multi-label learning utilizes logical labels to estimate the relevance between the feature set and the label set so that the importance of labels can not be well-reflected. To deal with these issues, we propose a novel robust multi-label feature selection method named RLEFS in this paper. RLEFS utilizes a shared space by mapping patterns to excavate semantic similarity structure in features and labels. Besides, we reconstruct the label space to obtain numerical labels by a label enhancement regularization term during mining semantic similarity structure process. Furthermore, the local and global structures are considered to ensure effective information can be captured as fully as possible during feature selection process. Finally, we integrate the above terms into one joint learning framework, and then a simple yet effective optimization method with provable convergence is proposed to solve the above problems. Experimental results on multiple data sets show that the superiority of the proposed method.

Keywords: Feature selection, Multi-label learning, Label enhancement, Graph regularization

1 Introduction

In real-world applications, numerous learning methods are confronting great challenges with the increase of high-dimensional data, these data lead to the curse of dimensionality as well. To deal with these issues caused by high-dimensional data, feature selection is designed to reduce the number of irrelevant and redundant features, excavate useful information, improve the classification performance of models simultaneously [1–4]. In light of these, feature selection is used in many domains, such as communications-electronics, biomedical, computational chemistry, etc.

Recent years, a large number of researches regarding feature selection methods are proposed. Generally, feature selection methods are categorized into several main types based on the selection strategy: filter model, wrapper model, and sparse coding based model (a.k.a. embedded model). Filter model is independent of the learning model while wrapper model is dependent the learning model. Therefore, the computation cost of wrapper model is higher than that of filter model. In addition, sparse coding based model utilizes the advantages of the former two models to embed both feature selection and the subsequent learning model into a unified framework. We focus on sparse coding based model for feature selection in this paper.

In the early stage, researchers deal with binary or/and multi-class label data by previous methods. However, with the explosive growth of data, the new emerging multi-label data degenerate the performance of the previous methods. Thus, numerous multi-label learning methods are proposed to handle multi-label data. Most of the existing multi-label learning methods only consider either the feature set or the multi-label set. However, the relationship between them is not considered adequately. We know that the latent structures between feature set and label set is consistent [5]. Therefore, a shared subspace between them is achieved by mapping patterns to capture the useful semantic similarity structure. Besides, there are a lot of noise information in the feature set and label set. Ignoring noise information will construct an inaccurate subspace so that inaccurate label correlations are captured. Furthermore, noise information also degrades the performance of feature selection model. To this end, we introduce a structured sparsity norm—$L_{2,1}$-norm that has been demonstrated to be robust to noise [6]. We impose this norm onto both feature set term and label set term simultaneously. In addition, traditional multi-label feature selection methods utilize logical labels to estimate the relevance between the feature set and label set. However, the importance of each label is different in real-world multi-label data. Hence, the importance of labels can not be well-reflected by logical values. To deal with this problem, we design a label enhancement term to reconstruct label set from logical label set to numeric label set, so that we can enhance the performance of feature selection by numerical labels during mining semantic similarity structure process. From the instance-level perspective, the local and global structure can

provide complementary information to improve multi-label learning according to previous literature [7, 8]. In our method, we exploit local and global structures from the label-level perspective simultaneously.

In light of the above analysis, we propose a novel multi-label feature selection method that integrates the above various terms into one joint learning framework. This joint learning framework is named Robust multi-label Feature Selection with shared Label Enhancement (RLEFS). And then, a simple yet effective optimization method with provable convergence is proposed to solve the above problems.

In summary, the novelties and contributions of this paper are highlighted as follows:

1. Extracting the shared space from feature space and label space by double mapping patterns to capture the useful semantic similarity structure.
2. Reconstructing label set from logical label set to numeric label set by designing a label enhancement term.
3. Imposing structured sparsity norm onto both feature set term and label set term simultaneously, to ensure the model is not disturbed by noise.
4. Combining local and global structure of labels to provide complementary information so as to improve the performance of multi-label feature selection.
5. Designing a joint learning framework that named Robust multi-label Feature Selection with shared Label Enhancement (RLEFS).
6. Developing an optimization method with provable convergence to solve the proposed RLEFS framework.
7. Conducting comprehensively evaluation criteria on multiple benchmark data sets to demonstrate the effectiveness of the proposed framework.

The remainder of the paper is organized as follows. In Section 2, we review some main related works, such as multi-label learning, feature selection methods and learning regularizer, etc. In Sections 3 and 4, we propose a joint learning framework that is named RLEFS and its optimization method with provable convergence respectively. In Section 5, the comprehensively experimental results on multiple multi-label data sets are described. At the same time, we analyze these results to verify the effectiveness and efficiency of the proposed RLEFS. Finally, some concluding remarks and future work are given in Section 6.

2 Related work

2.1 Preliminaries

In this subsection, some definitions of the notations used are introduced. Matrices are denoted by italicized uppercase letters, such as A. For matrix $A \in \mathbb{R}^{n\times m}$, $A_{i\cdot}$ and $A_{\cdot j}$ denote i-th row and j-th column of A. In addition, vectors can be also denoted by rows or columns of the matrix, or bold italicized lowercase letters, such as a. Scalars are denoted by lowercase letters, such as a. Functions can be represented by calligraphic letters. A^T and $\mathrm{Tr}(A)$ denote the

transpose and the trace of A respectively, where A in $\mathrm{Tr}(A)$ is a square matrix. The Frobenius norm of A is defined as $\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}^2}$. The $L_{2,1}$-norm of A is defined as $\|A\|_{2,1} = \sum_{i=1}^{n}\sqrt{\sum_{j=1}^{m} A_{ij}^2}$, where A_{ij} denotes the (i, j)-th entry of matrix A. In this paper, we suppose that the feature matrix $X \in \mathbb{R}^{n\times d}$ has n instances in d-dimensional space. The label matrix $Y \in \mathbb{R}^{n\times c}$ has c column class labels. Generally, the value of Y_{ij} is 1 if the i-th instance is related with the j-th label, and Y_{ij} as 0 otherwise.

2.2 Related work

In this subsection, we review some related works, such as multi-label learning, feature selection methods and several regularizers, etc.

In the past decade, many well-established multi-label learning methods have been widely applied to different fields. Generally, multi-label learning methods are categorized into three categories based on the label correlation strategy: first-order strategy, second-order strategy as well as high-order strategy [9]. In the first strategy, some researchers utilize single-label learning method to deal with multi-label data. This strategy ignores the label correlations, such as Binary Relevance (BR) [10]. In the second strategy, the pairwise label correlations have caused extensive concern. Calibrated Label Ranking (CLR) is a representative second-order method [11]. For the third strategy, the correlations of all labels or a subset of all labels are taken into account, such as LLSF-DL [12]. However, the last strategy consumes too much calculation cost and time cost. Consequently, we choose the second strategy to utilize label correlations. However, most of previous methods that adopt the above strategies assume the importance of labels are equivalent. That is, traditional multi-label learning utilizes logical labels to estimate the relevance between the feature set and label set so that the importance of labels can not be well-reflected. To this end, some researchers study how to transform logical labels into numerical labels for the importance of labels can be well-reflected [13–15].

By investigating numerous literature [16, 17], we find that most multi-label feature selection methods are mainly divided into two categories: problem transformation and algorithm adaption. The former category is to transforms multi-label problem into several single-label problems, such as Pruned Problem Transformation (PPT) [18]. In order to improve the classification performance, Doquire et al propose PPT+MI that is a multi-label feature selection method based on PPT. Besides, PPT+CHI that uses $\mathcal{X}^2$ statistic method is developed to select the important features according to PPT as well. However, these above methods still may be lead to information loss of multi-label data. Consequently, algorithm adaption method is proposed. Next, we introduce some algorithm adaption methods.

We know that it is more difficult to excavate the useful information from multi-label data than from single-label data by traditional feature selection (problem transformation). In order to better excavate the useful information, researchers design many multi-label feature selection methods by different

criteria that have been widely used in multi-label data applications, such as mutual-information-based and sparse-learning-based methods. We briefly review these criteria by several representative multi-label feature selection methods in this subsection.

Jian et al propose a sparse-learning-based method named Multi-label Informed Feature Selection (MIFS). MIFS uses matrix factorization to obtain a low-rank latent label matrix that preserves the local geometrical structure of labels from the instance-level perspective. By the above operations, MIFS eliminates irrelevant and redundant information of label matrix. MIFS is constructed as follows:

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

where $X \in \mathbb{R}^{n\times d}$ and $Y \in \mathbb{R}^{n\times c}$ denote the feature matrix and the label matrix respectively. $W \in \mathbb{R}^{d\times c}$, $V \in \mathbb{R}^{n\times k}$ and $B \in \mathbb{R}^{k\times c}$ are regarded as the weight matrix, the latent label matrix and the basis matrix respectively. $L \in \mathbb{R}^{n\times n}$ denotes Laplacian matrix. α, β and γ are three hyper-parameters of MIFS.

Besides, Cai et al also design a feature selection method based on the sparse theory. This method imposes $L_{2,0}$-norm onto the weight matrix and then uses the Augmented Lagrangian Multiplier method to optimize the following objective function. It is named RALM-FS:

$$\min_{W,V,B} \|Y - XW - 1b^T\|_{2,1} \quad s.t. \quad \|W\|_{2,0} = k \tag{2}$$

where X, Y and W have the same structure as X, Y and W of MIFS. b and 1 denote the bias term and the all-one-element vector respectively. k denotes the number of the selected features.

Alternatively, Lin et al design a multi-label feature selection method named Max-Dependency and Min-Redundancy (MDMR). This method utilizes feature dependency and feature redundancy to conduct feature selection. Besides, Zhang et al propose a novel multi-label feature selection method by using label redundancy (LRFS). LRFS considers a new feature relevance term based on the conditional mutual information. LRFS has the following form:

$$\begin{aligned} J(f_k) &= LR(f_k;\ Y) - \frac{1}{|\ S\ |}\sum_{f_j\in S} I(f_k;\ f_j) \\ &= \sum_{y_i\in Y}\left\{\sum_{y_i\neq y_j, y_j\in Y} I(f_k;\ y_j \mid y_i) - \frac{1}{|\ S\ |}\sum_{f_j\in S} I(f_k;\ f_j)\right\} \end{aligned} \tag{3}$$

where $LR(f_k;\ Y)$ and $I(f_k;\ f_j)$ is regarded as the relevance term and the redundancy term of features respectively. f_k denotes a candidate feature from the full feature set F. y_i and y_j are two labels from the full label set Y. To balance the magnitude between $LR(f_k;\ Y)$ and $I(f_k;\ f_j)$, $I(f_k;\ f_j)$ is divided

over the $\mid S \mid$ of the selected feature subset S. Both MDMR and LRFS belong to multi-label feature selection methods based on mutual-information.

In addition, the various regularization term is used in numerous machine learning algorithms, such as local learning regularizer and sparsity regularizer. In local learning regularizer, an intuitive assumption is adopted, that is, if two data points $X_{i\cdot}$ and $X_{j\cdot}$ in a high-dimensional ambient space are close, then $Y_{i\cdot}$ and $Y_{j\cdot}$ in a low-dimensional space should be similar. Generally, we use graph Laplacian that is a discrete Laplace operator to achieve the above assumption. The following calculation method is obtained:

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}\|Y_{i\cdot} - Y_{j\cdot}\|_2^2 \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}(Y_{i\cdot} - Y_{j\cdot})(Y_{i\cdot} - Y_{j\cdot})^T \\
&= \sum_{i=1}^{n} Y_{i\cdot}Y_{i\cdot}^T A_{ii} - \sum_{i=1}^{n}\sum_{j=1}^{n} Y_{i\cdot}Y_{j\cdot}^T S_{ij} \\
&= \mathrm{Tr}(Y^T(A-S)Y) \\
&= \mathrm{Tr}(Y^T L Y)
\end{aligned} \tag{4}$$

where $L = A - S$ denotes a graph Laplacian matrix of matrix X. S and A denote the symmetric affinity matrix and the degree matrix, respectively. For sparsity regularizer, it can lead to a global structured learning [7, 19] when we take the following form:

$$\min_{W} \|X - XW\|_F^2 + \alpha\|W\|_F^2 \tag{5}$$

where $\|X - XW\|_F^2 = \mathrm{Tr}(X^T(I-W)^T(I-W)X) = \mathrm{Tr}(X^T M X)$, $M = (I-W)^T(I-W)$. This is similar to the above the result of (4), However, function (5) can preserve global structure by adjusting W.

3 The proposed framework

In this section, we propose a novel robust multi-label feature selection method named RLEFS. RLEFS achieves a shared space by mapping patterns to excavate semantic similarity structure between features and labels. Besides, we reconstruct the label space to obtain numerical labels by label enhancement regularization term, local and global regularization term during mining semantic similarity structure process, which have provided excellent guidance for feature selection process.

Generally, the least square regression model is utilized to learn the weight matrix W. However, this model is very sensitive to noise. In order to make the model more robust, we impose $L_{2,1}$-norm onto the learning model, where this

norm has been confirmed to be robust to noise [20]. Therefore, the following learning model is given:

$$\min_{W} \|XW - Y\|_{2,1} \tag{6}$$

where $W \in \mathbb{R}^{d\times c}$ denotes the feature weight matrix which can measure the importance of each feature. That is, if the value of $\|W_{i\cdot}\|_2$ is larger, then the i-th feature of matrix X has greater contribution. In addition, Hu [21] and Shang [22] motivate us to consider the local geometric structure of label set from label-level perspective. It is vital to preserve the local structure of labels due to label correlations in multi-label data. Moreover, the weight matrix W is a mapping matrix from a high-dimension feature space to low-dimension label space. If $W_{ij} = 0 \quad (1 \le i \le d)$, then the i-th feature has no contribution for distinguishing the j-th label $Y_{\cdot j}$. Otherwise, it indicates the i-th feature has contribution to the j-th label. The larger the value is, the greater the contribution is. Thus, W can measure the relevance between labels by the following form:

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{c}\sum_{j=1}^{c} S_{ij}\|W_{\cdot i} - W_{\cdot j}\|_2^2 \\
&= \frac{1}{2}\sum_{i=1}^{c}\sum_{j=1}^{c} S_{ij}(W_{\cdot i} - W_{\cdot j})(W_{\cdot i} - W_{\cdot j})^T \\
&= \sum_{i=1}^{c} W_{\cdot i}W_{\cdot i}^T A_{ii} - \sum_{i=1}^{c}\sum_{j=1}^{c} W_{\cdot i}W_{\cdot j}^T S_{ij} \\
&= \mathrm{Tr}(W(A - S)W^T) \\
&= \mathrm{Tr}(WLW^T)
\end{aligned} \tag{7}$$

We incorporate (7) into (6) to obtain the following function:

$$\min_{W} \|XW - Y\|_{2,1} + \alpha\,\mathrm{Tr}(WLW^T) \tag{8}$$

where $L = A - S$ denotes a graph Laplacian matrix of label set. α is a trade-off parameter between loss function and label graph regularization term. In particular, we need to measure the correlations between X and Y by the weight matrix. However, traditional methods utilize logical labels to estimate the relevance between the feature set and label set so that the importance of labels can not be well-reflected. Therefore, we transform logical labels into numerical labels by a label enhancement regularization term during mining semantic similarity structure process. We obtain the following form:

$$\min_{W,F,B} \|XW - F\|_{2,1} + \alpha\,\mathrm{Tr}(WLW^T) + \beta\|F - YB\|_{2,1} \tag{9}$$

where β denotes a regularization parameter that adjusts the contribution of the third term. $F \in \mathbb{R}^{n\times c}$ denotes a co-embedding space between the feature

space and label space. Moreover, we use F to capture numerical labels of label matrix Y. This reconstruction process is similar to the construction of the shared subspace. However, F is not a low-dimensional embedding subspace for label matrix Y. Besides, we impose $L_{2,1}$-norm onto both loss function and label enhancement regularization term so that the impact of outliers in feature set and label set can be reduced. Similar to the sparse regularizer mentioned in related work, the designed label enhancement regularization term can preserve the global structure as well. Hence, F in the third term is globally consistent with matrix Y, while F preserves local geometric structure by using the first two terms. Next, we impose $L_{2,1}$-norm onto W, which ensures not only the row sparsity of W but also can automatically select features during the multi-label learning. Therefore, we reformulate the following function:

$$\min_{W,F,B} \|XW - F\|_{2,1} + \alpha \operatorname{Tr}(WLW^T) + \beta\|F - YB\|_{2,1} + \gamma\|W\|_{2,1} \tag{10}$$

where γ denotes a regularization parameter that controls the sparsity of the objective function. However, the row-sparse property of W is not always guaranteed by $L_{2,1}$-norm [23]. Consequently, we impose the non-negative constraint on W so that the row-sparse property is further enhanced. We also apply non-negative constraints onto F and B, which can ensure the consistency of F and Y, because Y has only non-negative logical values, 0 and 1. Note that, if the elements of F are zero, the above function always leads to a trivial solution. Consequently, we impose an orthogonality constraint on F. This orthogonal constraint can ensure the minimum redundancy as well. Therefore, the final objective function is reformulated as follows:

$$\begin{gathered}\min_{W,F,B} \|XW - F\|_{2,1} + \alpha \operatorname{Tr}(WLW^T) + \beta\|F - YB\|_{2,1} + \gamma\|W\|_{2,1} \\ s.t.\{W, F, B\} \geq 0, F^T F = I\end{gathered} \tag{11}$$

Next, we develop a simple yet effective optimization method for our objective function, to guarantee the convergence of function (11). This optimization method with provable convergence will be described in detail in the next section.

4 Optimization of RLEFS model

4.1 Optimization Schemes

In this section, we propose an efficient optimization method to solve the proposed objective function in Section 3. we observe that the objective function (11) is joint not-convex. That is, the Hessian matrix that is composed of the second partial derivative of the multivariate function is not a positive semi-definite matrix. In addition, the objective function is non-smooth due to $L_{2,1}$-norm. To solve these problems, we transform the objective function into

several sub-solution processes, that is, a variable is updated and other variables are fixed. At the same time, a relaxed approach is introduced to solve the non-smooth problem. Therefore, the objective function (11) is equivalent to as follows:

$$\begin{aligned}\Theta(W,F,B) =& 2\,\mathrm{Tr}\left[(XW-F)^T D_1(XW-F)\right] + \alpha\,\mathrm{Tr}\left(WLW^T\right) \\ &+ 2\beta\,\mathrm{Tr}\left[(F-YB)^T D_2(F-YB)\right] + 2\gamma\,\mathrm{Tr}\left(W^T D_3 W\right) \\ &\text{s.t. } \{W,F,B\} \geq 0, F^T F = I\end{aligned} \tag{12}$$

where $\Theta(W,F,B)$ denotes the objective function (12) with respect to variables W, F and B. D_1, D_2 and D_3 are defined as follows:

$$\begin{cases} D_{1_{ii}} = \frac{1}{2\|(XW-F)_{i\cdot}\|_2+\epsilon} \\ D_{2_{ii}} = \frac{1}{2\|(F-YB)_{i\cdot}\|_2+\epsilon} \\ D_{3_{ii}} = \frac{1}{2\|W_{i\cdot}\|_2+\epsilon} \end{cases} \tag{13}$$

where $D_{1_{ii}}$, $D_{2_{ii}}$ and $D_{3_{ii}}$ denote the i-th diagonal element of D_1, D_2 and D_3 respectively. ϵ is a non-negative small constant. By integrating non-negative and orthogonal constraints into function (12), we obtain the following Lagrangian function:

$$\begin{aligned}\mathcal{L}(W,F,B) =& 2\,\mathrm{Tr}\left[(XW-F)^T D_1(XW-F)\right] + \alpha\,\mathrm{Tr}\left(WLW^T\right) \\ &+ 2\beta\,\mathrm{Tr}\left[(F-YB)^T D_2(F-YB)\right] + 2\gamma\,\mathrm{Tr}\left(W^T D_3 W\right) \\ &+ \frac{\lambda}{2}\left\|F^T F - I\right\|_F^2 - \mathrm{Tr}\left(\Phi W^T\right) - \mathrm{Tr}\left(\Psi F^T\right) - \mathrm{Tr}\left(\Omega B^T\right)\end{aligned} \tag{14}$$

where $\Phi \in R_+^{d\times c}$, $\Psi \in R_+^{n\times c}$ and $\Omega \in R_+^{c\times c}$ denote the Lagrangian multiplier. λ denotes the regularization parameter of orthogonal constraint. By taking the derivative of function (14) w.r.t W, F and B respectively, we obtain:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial W} = 2X^T D_1 W - 2X^T D_1 F + 2\alpha WL + 2\gamma D_3 W - \Phi \\ \frac{\partial \mathcal{L}}{\partial F} = -2D_1 XW + 2D_1 F + 2\beta D_2 F - 2\beta D_2 YB + 2\lambda FF^T F - 2\lambda F - \Psi \\ \frac{\partial \mathcal{L}}{\partial B} = -2\beta Y^T D_2 F + 2\beta Y^T D_2 YB - \Omega \end{cases} \tag{15}$$

We know that $\Phi_{ij}W_{ij} = 0$, $\Psi_{ij}F_{ij} = 0$ and $\Omega_{ij}B_{ij} = 0$ according to Karush-Kuhn-Tucker conditions. Therefore, we get the following functions:

$$\begin{cases} \left(X^T D_1 W - X^T D_1 F + \alpha WL + \gamma D_3 W\right)_{ij} W_{ij} = 0 \\ \left(-D_1 XW + D_1 F + \beta D_2 F - \beta D_2 YB + \lambda FF^T F - \lambda F\right)_{ij} F_{ij} = 0 \\ \left(-\beta Y^T D_2 F + \beta Y^T D_2 YB\right)_{ij} B_{ij} = 0 \end{cases} \tag{16}$$

According to the above functions, we obtain the following update rules:

$$\begin{cases} W_{ij}^{t+1} \leftarrow W_{ij}^{t} \frac{\left(X^T D_1 + \alpha W S\right)_{ij}}{\left(X^T D_1 X W + \alpha W A + \gamma D_3 W\right)_{ij}} \\ F_{ij}^{t+1} \leftarrow F_{ij}^{t} \frac{\left(D_1 X W + \beta D_2 Y B + \lambda F\right)_{ij}}{\left(D_1 F + \beta D_2 F + \lambda F F^T F\right)_{ij}} \\ B_{ij}^{t+1} \leftarrow B_{ij}^{t} \frac{\left(Y^T D_2 F\right)_{ij}}{\left(Y^T D_2 Y B\right)_{ij}} \end{cases} \tag{17}$$

where t denotes the counter. L is a graph Laplacian matrix with mixed sign, thus we decompose L into two non-negative parts, i.e., $L = A - S$. Besides, some elements in the denominator will become zero during the update process. To solve this issue, we add a very small constant to the denominator. Then, we obtain the top-k features during feature selection process. The pseudo-code of the proposed method is described in Algorithm 1.

Algorithm 1 RLEFS

Input:

The input feature matrix $X \in \mathbb{R}^{n \times d}$ and the output matrix $Y \in \mathbb{R}^{n \times c}$;

Regularization parameters α, β, γ and λ.

Output:

Return the selected top-k selected features index set.

Initialize $W \in \mathbb{R}_{+}^{d \times c}$, $F \in \mathbb{R}^{n \times c}$ and $B \in \mathbb{R}_{+}^{c \times c}$ randomly;

$t = 0$;

Compute the degree matrix A and the affinity matrix S of the label matrix Y;

Repeat

Update the diagonal matrix D_1, D_2 and D_3 by $\begin{cases} D_{1_{ii}} = \frac{1}{2\|(XW-F)_{i\cdot}\|_2 + \epsilon} \\ D_{2_{ii}} = \frac{1}{2\|(F-YB)_{i\cdot}\|_2 + \epsilon} \\ D_{3_{ii}} = \frac{1}{2\|W_{i\cdot}\|_2 + \epsilon} \end{cases}$;

Update $\begin{cases} W_{ij}^{t+1} \leftarrow W_{ij}^{t} \frac{\left(X^T D_1 + \alpha W S\right)_{ij}}{\left(X^T D_1 X W + \alpha W A + \gamma D_3 W\right)_{ij}} \\ F_{ij}^{t+1} \leftarrow F_{ij}^{t} \frac{\left(D_1 X W + \beta D_2 Y B + \lambda F\right)_{ij}}{\left(D_1 F + \beta D_2 F + \lambda F F^T F\right)_{ij}} \\ B_{ij}^{t+1} \leftarrow B_{ij}^{t} \frac{\left(Y^T D_2 F\right)_{ij}}{\left(Y^T D_2 Y B\right)_{ij}} \end{cases}$;

$t = t + 1$;

Until Convergence criterion is satisfied; Sort all features by $\|W_{i\cdot}\|_2$, where i=1,2,3,...,d, and select the top-k features.

4.2 Proof of convergence

In this subsection, we prove the convergence of the proposed optimization method. First, we introduce the conventional gradient descent method to deduce the updating rules in this paper. Then, the proposed optimization

method is proved. Taking variable W as an example, we give the following formula:

$$W_{ij}^{t+1} \leftarrow W_{ij}^{t} - \eta \left(\frac{\partial \Theta}{\partial W_t} \right)_{ij} \tag{18}$$

where the learning rate η is a small positive constant. To ensure non-negative constraints and obtain a data-adaptive learning rate, we set:

$$\eta = \frac{W_{ij}^{t}}{2\left(X^T D_1 X W + \alpha W A + \gamma D_3 W\right)_{ij}} \tag{19}$$

We take (19) into (18), then the following result is obtained:

$$\begin{aligned} W_{ij}^{t+1} &\leftarrow W_{ij}^{t} - \frac{W_{ij}^{t}}{2\left(X^T D_1 X W + \alpha W A + \gamma D_3 W\right)_{ij}} \left(\frac{\partial \Theta}{\partial W_t} \right)_{ij} \\ \Longleftrightarrow W_{ij}^{t+1} &\leftarrow W_{ij}^{t} \frac{\left(X^T D_1 + \alpha W S\right)_{ij}}{\left(X^T D_1 X W + \alpha W A + \gamma D_3 W\right)_{ij}} \end{aligned} \tag{20}$$

Accordingly, we can obtain the above update rule that is a special case of the gradient descent method. The convergence of the optimization method is proved below. First, some related concepts are given [24, 25].

Definition 1. If $\mathcal{G}(\omega, \omega') \geq \mathcal{F}(\omega)$ and $\mathcal{G}(\omega, \omega') = \mathcal{F}(\omega)$ are satisfied, then $\mathcal{G}(\omega, \omega')$ is considered to be an auxiliary function of $\mathcal{F}(\omega)$.

Lemma 1 .If $\mathcal{G}(\omega, \omega')$ is considered to be an auxiliary function of $\mathcal{F}(\omega)$, then $\mathcal{F}(\omega)$ is a non-increasing function according to:

$$\omega^{t+1} = \arg\min_{\omega} \mathcal{G}(\omega.\omega^t) \tag{21}$$

Proof of Lemma 1: According to **Definition 1** and function (21), we can deduce that:

$$\mathcal{F}(\omega^{t+1}) \leq \mathcal{G}(\omega^{t+1}, \omega^t) \leq \mathcal{G}(\omega^t, \omega^t) = \mathcal{F}(\omega^t) \tag{22}$$

where w denotes any elements of W. Next, we prove the convergence of RLEFS w.r.t. W by a proper auxiliary function $\mathcal{G}(\omega, \omega')$. Considering that the update rule happen always on an element-by-element basis, we use W_{ij} to denote the (i, j)-th element of W. And $\mathcal{F}_{ij}$ denotes the part of $\Theta(W)$, which is relevant to W_{ij}. Therefore, the first-order and second-order partial derivatives of $\mathcal{F}(W_{ij})$ are obtained:

$$\mathcal{F}'_{ij} = \left(2X^T D_1 W - 2X^T D_1 F + 2\alpha W L + 2\gamma D_3 W\right)_{ij} \tag{23}$$

$$\mathcal{F}''_{ij} = 2\left(X^T D_1 X\right)_{ii} + 2\alpha(L)_{jj} + 2\gamma\left(D_3\right)_{ii} \tag{24}$$

According to the above process, we obtain the Taylor function of $\mathcal{F}(W_{ij})$:

$$\mathcal{F}_{ij}\left(W_{ij}\right)=\mathcal{F}_{ij}\left(W_{ij}^{t}\right)+\mathcal{F}_{ij}^{\prime}\left(W_{ij}^{t}\right)\left(W_{ij}-W_{ij}^{t}\right)+\frac{1}{2}\mathcal{F}_{ij}^{\prime\prime}\left(W_{ij}^{t}\right)\left(W_{ij}-W_{ij}^{t}\right)^{2} \tag{25}$$

Inspired by NMF methods [26, 27], we set the following auxiliary function about $\mathcal{F}(W_{ij})$:

$$\begin{aligned}\mathcal{G}\left(W_{ij},W_{ij}^{t}\right)=&\mathcal{F}_{ij}\left(W_{ij}^{t}\right)+\mathcal{F}_{ij}^{\prime}\left(W_{ij}^{t}\right)\left(W_{ij}-W_{ij}^{t}\right)\\&+\frac{\left(X^{T}D_{1}XW+\alpha WA+\gamma D_{3}W\right)_{ij}}{W_{ij}^{t}}\left(W_{ij}-W_{ij}^{t}\right)^{2}\end{aligned} \tag{26}$$

Proof: If $W_{ij}=W_{ij}^{t}$ in (26), then $\mathcal{G}(W_{ij},W_{ij}^{t})=\mathcal{F}_{ij}(W_{ij}^{t})$. For another condition $\mathcal{G}(W_{ij},W_{ij}^{t})\geq\mathcal{F}(W_{ij})$ in **Definition 1**, we need to prove the following inequality:

$$\frac{\left(X^{T}D_{1}XW+\alpha WA+\gamma D_{3}W\right)_{ij}}{W_{ij}^{t}}\geq\left(X^{T}D_{1}X\right)_{ii}+\alpha(L)_{jj}+\gamma\left(D_{3}\right)_{ii} \tag{27}$$

It is obvious that (27) is equivalent to the following form:

$$\begin{aligned}(X^{T}D_{1}XW+\gamma D_{3}W)_{ij}&=\sum_{l=1}^{d}(X^{T}D_{1}X+\gamma D_{3})_{il}W_{lj}^{t}\\&\geq(X^{T}D_{1}X+\gamma D_{3})_{ii}W_{ij}^{t}\end{aligned} \tag{28a}$$

$$\begin{aligned}\alpha(WA)_{ij}&=\alpha\sum_{l=1}^{c}W_{il}^{t}(A)_{lj}\geq\alpha W_{ij}^{t}(A)_{jj}\\&\geq\alpha W_{ij}^{t}(A-S)_{jj}=\alpha W_{ij}^{t}(L)_{jj}\end{aligned} \tag{28b}$$

Therefore, $\mathcal{G}(W_{ij},W_{ij}^{t})$ is proved to be an auxiliary function of $\mathcal{F}_{ij}(W_{ij})$. This auxiliary function is brought into (21), then we obtain the update rule of W by $\frac{\partial\mathcal{G}(W_{ij},W_{ij}^{t})}{\partial W_{ij}}=0$:

$$\begin{aligned}W_{ij}^{t+1}&=W_{ij}^{t}-W_{ij}^{t}\frac{\mathcal{F}_{ij}^{\prime}\left(W_{ij}^{t}\right)}{2\left(X^{T}D_{1}XW+\alpha WA+\gamma D_{3}W\right)_{ij}}\\&=W_{ij}^{t}\frac{\left(X^{T}D_{1}+\alpha WS\right)_{ij}}{\left(X^{T}D_{1}XW+\alpha WA+\gamma D_{3}W\right)_{ij}}\end{aligned} \tag{29}$$

Finally, we prove the convergence of the proposed optimization method. The converge proof of other variables (F and B) are similar to the above process.

5 Experimental study

In this section, we use ten multi-label benchmark data sets and six state-of-the-art compared methods to conduct experiments, where all experiments are performed on a 3.4GHz Intel Core (TM) i7-6700 machine with 16 GB main memory.

5.1 Experimental data sets

In our experiment, all data sets used are fetched from Mulan Library [28]. We found that these data sets were adopted in numerous literature about multi-label learning [29, 30]. Besides, these data sets are collected from different fields. For instance, the Enron data set is a subset of the Enron e-mail corpus [31], which comes from text domain. Flags data set is collected from the image field, it has 194 instances and seven labels that contains red, green and blue, etc. Several data sets come from yahoo data sets that belong to multi-label text (web page) categorization. The detailed description of all the data sets is summarized in Table 1.

Table 1 Description of data set

#Data sets	#Instances	#Train	#Test	#Features	#Labels
Arts	5000	2000	3000	462	26
Education	5000	2000	3000	550	33
Enron	1702	1123	579	1001	53
Entertain	5000	2000	3000	640	21
Flags	194	129	65	19	7
Reference	5000	2000	3000	793	33
Scene	2407	1211	1196	294	6
Science	5000	2000	3000	743	40
Social	5000	2000	3000	1047	39
Society	5000	2000	3000	636	27

5.2 Experimental settings

To comprehensively verify the classification performance of RLEFS, the following several classical and state-of-the-art feature selection methods are used as compared methods.

1. PPT+MI [18]: it is a PPT-based multi-label feature selection method that belongs to problem transformation.
2. PPT+CHI [32]: it is a PPT-based multi-label feature selection method that uses χ^2 statistic to select optimal features from feature set.
3. MIFS [33]: it is a sparse-learning-based multi-label feature selection method that decomposes the original label matrix into a low-dimensional label space.
4. MDMR [34]: it is a multi-label feature selection method that utilizes max-dependency and min-redundancy to select the optimal feature subset.

5. LRFS [35]: it is the latest mutual-information-based multi-label feature selection method that proposes a novel feature relevance term based on label redundancy.
6. RALM-FS [36]: it uses the $L_{2,0}$-norm regularization term to conduct feature selection based on sparse-learning.

To facilitate experiments, we set some parameters in advance. The heat-kernel function is adopted during structuring graph Laplacian matrix process, where the parameters p and σ are set as 5 and 1 respectively. To ensure the fairness, the hyper-parameters of all methods are tuned in {0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0}. We utilize the optimal parameters with respect to classification performance during the training process, where 5-fold cross-validation is adopted. Besides, a Binary Relevance model (BR) is used to transform multi-label data into binary classification data so that linear Support Vector Machine (SVM) and K-Nearest-Neighbors (KNN, K=3) can be used in our experiments, where C of SVM is tuned in $\{10^{-4}, 10^{-3},..., 10^{3}, 10^{4}\}$. Next, we adopt $Micro-F_1$ and $Macro-F_1$ (a.k.a. Micro-average and Macro-average) based on $F_1-measure$ as evaluation criteria to evaluate the proposed method and other compared methods. $Micro-F_1$ and $Macro-F_1$ have the following form:

$$\begin{cases} Micro\text{-}F_1 = \frac{\sum_{i=1}^{m} 2TP^i}{\sum_{i=1}^{m}(2TP^i+FP^i+FN^i)} \\ \\ Macro\text{-}F_1 = \frac{1}{m}\sum_{i=1}^{m} \frac{2TP^i}{2TP^i+FP^i+FN^i} \end{cases} \tag{30}$$

where m and i represent the number of class labels and the i-th label respectively. TP, FP and FN are composed by T, F, P and N, where T, F, P and N denote True, False, Positive and Negative respectively. Moreover, the larger both $Micro-F_1$ (or $Macro-F_1$), the better the performance of the method is. In addition, we give our experimental framework in Figure 1.

5.3 Experimental results

To evaluate the classification performance of the RLEFS method, we conduct numerous experiments on ten different multi-label data sets. The experiment results are described in some tables and figures. First, we use the top 20% of the total features in each data set to calculate the average result and standard deviations of different methods (all features of Flags are adopted since it only contains 19 features).

In Tables 2-5, we record the results of $Macro-F_1$ and $Micro-F_1$ of all the methods by using SVM and K-NN (K=3) classifier. The best results are represented by bold fonts in each row of tables. Moreover, the last row "Average" calculates the average values of all data sets under each feature selection method. Observing these tables, we conclude that the proposed RLEFS obtains the best result about the last row "Average", where these best results are 0.177, 0.386, 0.200 and 0.401 in Table 2-5 respectively. Besides, RLEFS obtains the

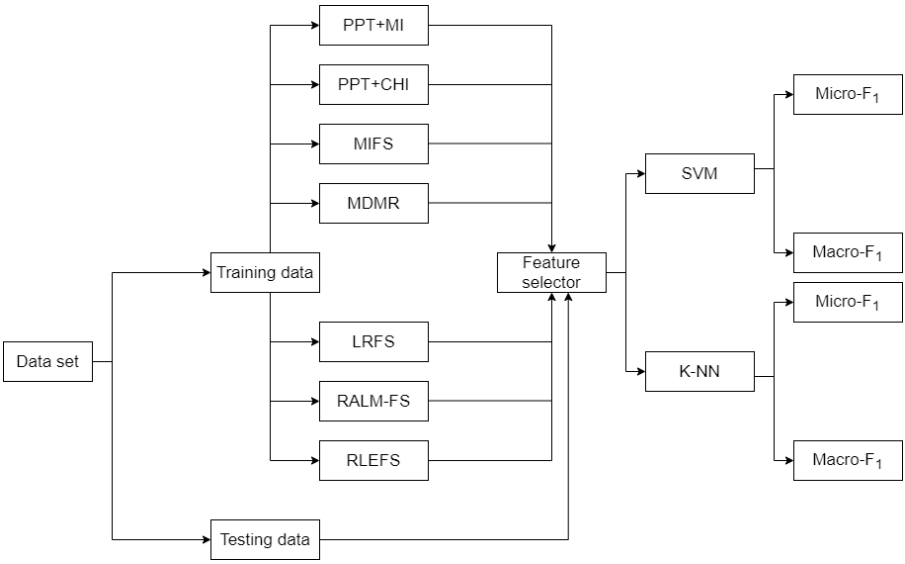

Fig. 1 Experimental framework

best results on most data sets used under all evaluation criteria, and it also obtains the suboptimal result on several data sets according to Tables 2-5.

To better clearly show the classification performance of all compared methods. We use Figs. 2-5 to show the experimental results on six representative data sets, including Arts, Education, Enron, Flags, Reference and Science. In Figs. 2-5, The X-axis and Y-axis are used to indicate the already-selected features and the classification performance of corresponding evaluation criteria, respectively. The number of already-selected features is varied from top-1% to top-20% of all features, where the step size is set to 1%. As shown in Figs. 2-5, we observe that RLEFS achieves the best classification performance. In most cases, the classification performance of RLEFS increases first and then stabilizes based on we have observed. Overall, the proposed RLEFS method outperforms other compared methods in experiments.

Table 2 The results of all methods in terms of $Macro - F_1$ (mean±std).

Data set	PPT+MI	PPT+CHI	MIFS	MDMR	LRFS	RALM-FS	RLEFS
Flags	0.513±0.037	0.517±0.046	0.57±0.1	0.503±0.053	0.511±0.044	0.487±0.044	**0.578±0.091**
Scene	0.163±0.098	0.199±0.114	0.295±0.127	0.406±0.092	0.394±0.104	**0.463±0.165**	0.43±0.163
Enron	0.102±0.035	0.067±0.016	0.074±0.017	0.108±0.032	0.099±0.033	0.074±0.027	**0.141±0.048**
Arts	0.037±0.022	0.04±0.023	0.055±0.034	0.039±0.022	0.043±0.019	0.039±0.026	**0.09±0.033**
Education	0.034±0.025	0.034±0.026	0.019±0.017	0.035±0.025	0.048±0.028	0.052±0.014	**0.068±0.02**
Entertain	0.072±0.033	0.078±0.044	0.097±0.047	0.077±0.032	0.079±0.031	0.081±0.038	**0.13±0.047**
Reference	0.038±0.019	0.037±0.018	0.063±0.024	0.041±0.02	0.051±0.021	0.063±0.028	**0.081±0.021**
Science	0.034±0.022	0.032±0.021	0.034±0.016	0.038±0.028	0.039±0.029	0.036±0.02	**0.069±0.024**
Society	0.048±0.015	0.047±0.014	0.055±0.02	0.049±0.016	0.049±0.015	0.032±0.012	**0.075±0.025**
Social	0.073±0.03	0.081±0.031	0.031±0.016	0.075±0.031	0.083±0.031	0.014±0.012	**0.103±0.039**
Average	0.111	0.113	0.129	0.137	0.14	0.134	**0.177**

Table 3 The results of all methods in terms of $Micro - F_1$ (mean±std).

Data set	PPT+MI	PPT+CHI	MIFS	MDMR	LRFS	RALM-FS	RLEFS
Flags	0.665±0.04	0.663±0.038	**0.722±0.05**	0.649±0.043	0.666±0.04	0.631±0.048	0.719±0.046
Scene	0.168±0.1	0.205±0.117	0.325±0.135	0.429±0.096	0.415±0.108	**0.476±0.164**	0.443±0.162
Enron	0.47±0.043	0.353±0.019	0.372±0.027	0.474±0.049	0.446±0.056	0.389±0.059	**0.525±0.043**
Arts	0.09±0.053	0.098±0.055	0.139±0.078	0.095±0.052	0.104±0.045	0.102±0.061	**0.21±0.073**
Education	0.124±0.083	0.12±0.085	0.073±0.059	0.127±0.081	0.15±0.081	0.193±0.056	**0.228±0.068**
Entertain	0.19±0.088	0.183±0.103	0.228±0.112	0.19±0.087	0.184±0.082	0.214±0.1	**0.285±0.119**
Reference	0.348±0.077	0.355±0.074	0.359±0.105	0.356±0.069	0.377±0.066	0.404±0.132	**0.418±0.108**
Science	0.115±0.059	0.11±0.055	0.129±0.057	0.127±0.074	0.129±0.078	0.097±0.054	**0.194±0.071**
Society	0.316±0.033	0.317±0.032	0.3±0.042	0.319±0.017	0.321±0.016	0.223±0.059	**0.336±0.067**
Social	0.47±0.097	0.445±0.108	0.276±0.136	0.465±0.106	0.465±0.119	0.149±0.112	**0.498±0.142**
Average	0.296	0.285	0.292	0.323	0.326	0.288	**0.386**

Table 4 The results of all methods in terms of $Macro - F_1$ (mean±std).

Data set	PPT+MI	PPT+CHI	MIFS	MDMR	LRFS	RALM-FS	RLEFS
Flags	0.448±0.011	0.449±0.011	0.527±0.078	0.443±0.018	0.453±0.022	0.439±0.011	**0.538±0.084**
Scene	0.348±0.069	0.368±0.072	0.498±0.086	0.555±0.075	0.547±0.082	**0.581±0.108**	0.541±0.117
Enron	0.116±0.019	.074±0.012	0.087±0.014	0.115±0.02	0.109±0.022	0.081±0.026	**0.126±0.018**
Arts	0.088±0.022	0.091±0.022	0.095±0.033	0.095±0.025	0.101±0.025	0.071±0.025	**0.114±0.032**
Education	0.082±0.019	0.083±0.022	0.043±0.018	0.082±0.019	0.087±0.02	0.084±0.023	**0.089±0.021**
Entertain	0.147±0.019	0.151±0.022	0.138±0.042	0.148±0.018	0.145±0.02	0.131±0.04	**0.172±0.043**
Reference	0.079±0.017	0.076±0.016	0.088±0.024	0.081±0.017	0.085±0.017	0.077±0.026	**0.1±0.02**
Science	0.065±0.023	0.066±0.022	0.062±0.015	0.067±0.024	0.067±0.023	0.057±0.021	**0.099±0.024**
Society	0.083±0.015	0.083±0.016	0.089±0.021	0.084±0.014	0.085±0.013	0.053±0.016	**0.095±0.022**
Social	0.099±0.034	0.107±0.03	0.051±0.017	0.107±0.032	0.111±0.027	0.038±0.012	**0.123±0.042**
Average	0.156	0.155	0.168	0.178	0.179	0.161	**0.2**

Table 5 The results of all methods in terms of $Micro - F_1$ (mean±std).

Data set	PPT+MI	PPT+CHI	MIFS	MDMR	LRFS	RALM-FS	RLEFS
Flags	0.615±0.011	0.615±0.011	0.662±0.03	0.611±0.017	0.616±0.018	0.606±0.012	**0.672±0.036**
Scene	0.348±0.067	0.367±0.071	0.493±0.086	0.556±0.075	0.547±0.082	**0.577±0.105**	0.541±0.116
Enron	0.454±0.013	0.345±0.026	0.41±0.024	0.443±0.044	0.419±0.048	0.365±0.073	**0.474±0.041**
Arts	0.181±0.037	0.187±0.033	0.202±0.052	0.189±0.034	0.196±0.027	0.182±0.043	**0.251±0.037**
Education	0.23±0.043	0.227±0.043	0.183±0.055	0.231±0.04	0.243±0.046	0.26±0.05	**0.271±0.054**
Entertain	0.284±0.03	0.285±0.033	0.276±0.065	0.283±0.03	0.281±0.034	0.273±0.065	**0.329±0.074**
Reference	0.373±0.044	0.366±0.04	0.382±0.055	0.378±0.044	0.386±0.046	0.386±0.089	**0.427±0.059**
Science	0.162±0.028	0.162±0.026	0.171±0.037	0.174±0.034	0.177±0.035	0.16±0.048	**0.228±0.04**
Society	0.317±0.023	0.317±0.026	0.305±0.043	0.312±0.032	0.315±0.028	0.255±0.05	**0.328±0.041**
Social	0.451±0.052	0.442±0.055	0.338±0.081	0.451±0.058	0.455±0.054	0.315±0.054	**0.491±0.081**
Average	0.342	0.331	0.342	0.363	0.364	0.338	**0.401**

5.4 Sensitivity analysis of parameters

Like many other methods, we study the sensitivity of parameters of RLEFS, where these parameters contain α, β, γ and λ. In this subsection, we take data set Arts as an experimental subject. First, these parameters are tuned in $\{0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$. If the traditional gird search is performed, it could be time-consuming due to four parameters in RLEFS. To this end, one parameter is tuned while the other three parameters are fixed, where we set the fixed parameters as 0.5 in this paper. We only use the analysis result of SVM classifier for convenience. From Fig. 6 (a)-(d), we observe that the classification performance is sensitive to the values of these parameters. However, RLEFS achieves better results in 0.3-0.7 in most cases. In addition,

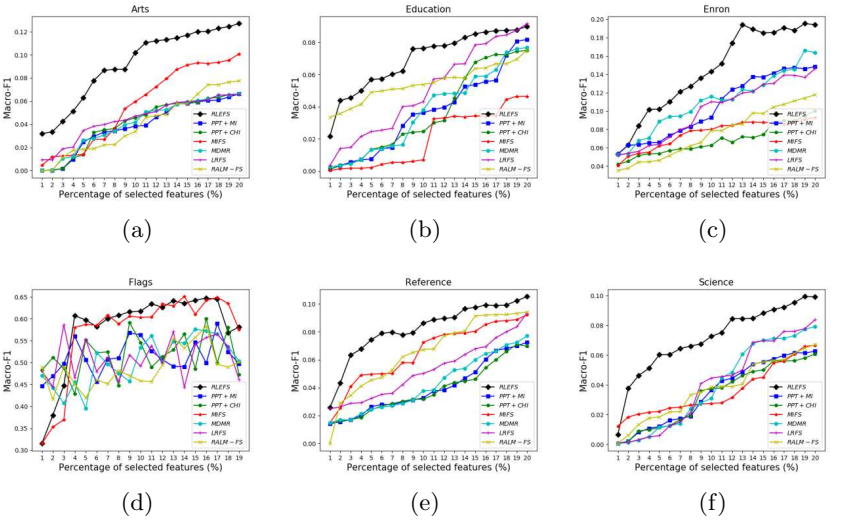

Fig. 2 All compared methods on six data sets using SVM classifier in term of $Macro - F_1$.

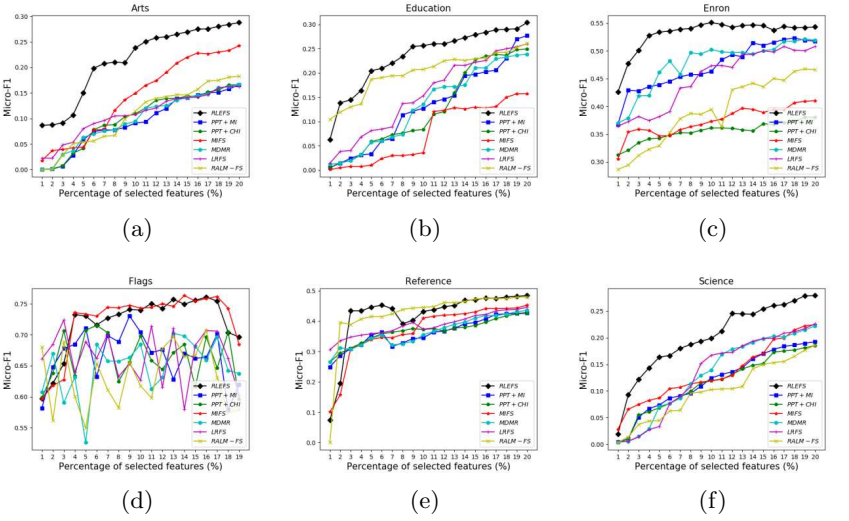

Fig. 3 All compared methods on six data sets using SVM classifier in term of $Micro - F_1$.

a more large candidate gird set can be employed in real-world applications to ensure satisfactory classification performance.

5.5 Convergence and complexity analysis

To verify the convergence of the proposed RLEFS method, we use four benchmark data sets (Arts, Education, Enron and Science) to conduct convergence experiments. The results are shown in Fig. 7. We observe that RLEFS tends to

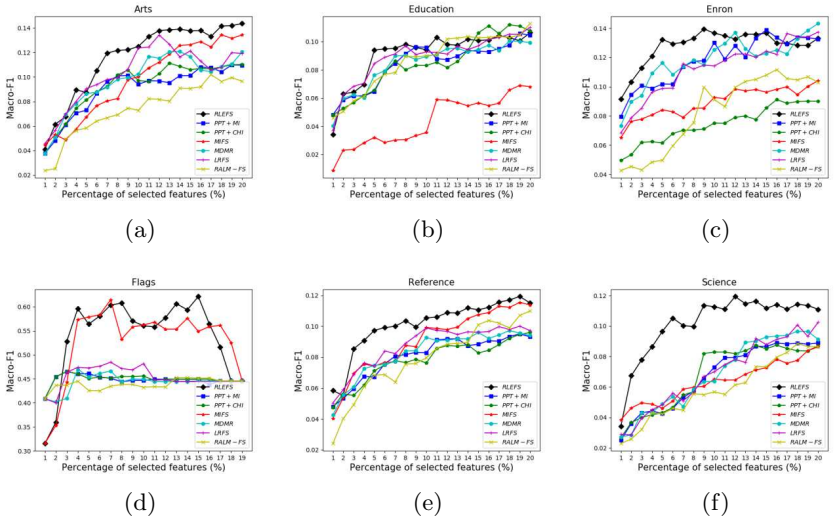

Fig. 4 All compared methods on six data sets using 3NN classifier in term of $Macro - F_1$.

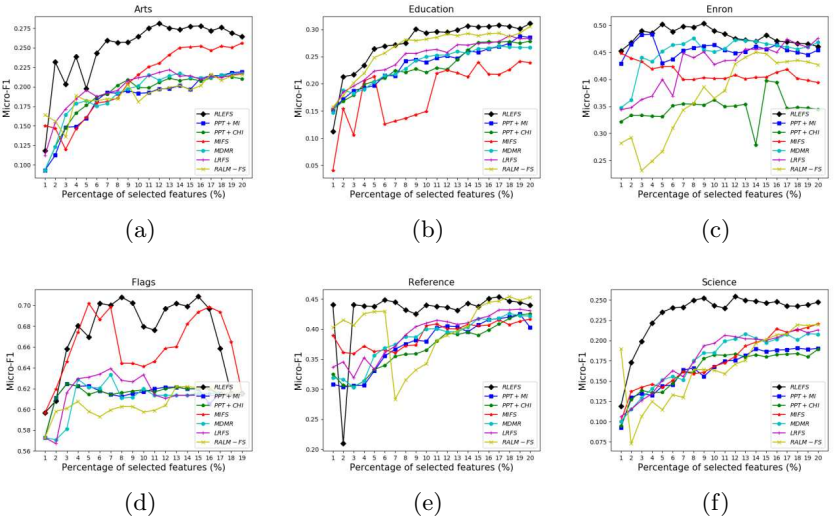

Fig. 5 All compared methods on six data sets using 3NN classifier in term of $Micro - F_1$.

converge in several iterations and then stable. We can also observe similar situations on other data sets. Afterward, the computational complexity of related methods are analyzed. Suppose one data set has n instances with q features, and it contains l labels. The number of already-selected features is recorded as k. The computation complexity of MDMR and LRFS are $O(k(n-k))$ and $O(ql^2 + kq)$ respectively. MIFS requires $O(cnq + n^2)$ in each iteration. The computation complexity of RALM-FS is $O(q^3)$ when an inverse matrix

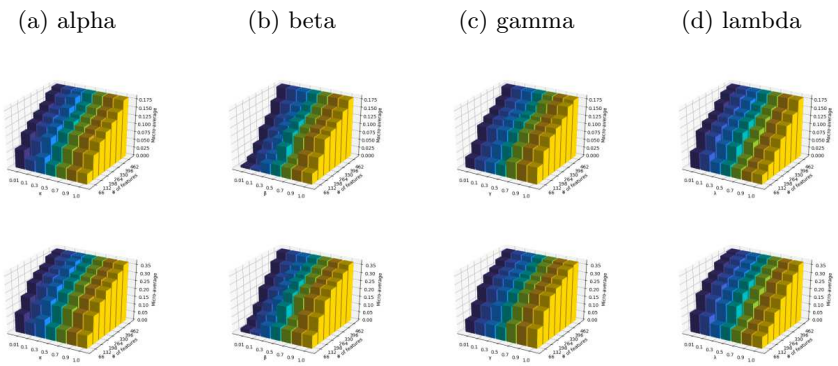

Fig. 6 Micro-average and Macro-average of RLEFS on Arts w.r.t all parameters (SVM classifier).

$(q * q)$ is calculated. The computation complexity of the proposed RLEFS is $O(qn^2 + nq^2)$ in each iteration.

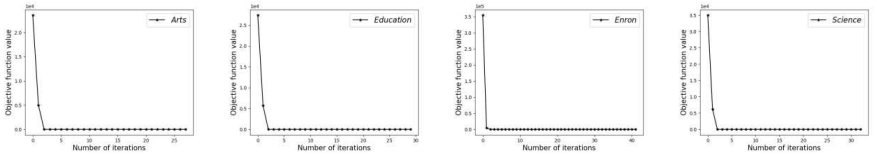

Fig. 7 All compared methods on six data sets using 3NN classifier in term of $Micro - F_1$.

6 Conclusions

In this paper, we propose a joint multi-label feature selection framework RLEFS. RLEFS has the following several appealing characteristics. First, RLEFS utilizes a shared space by mapping patterns to excavate semantic similarity structure in features and labels. It can deal with the problem that previous methods only devote attention to either of feature set or label set. Second, RLEFS reconstructs the label space to obtain numerical labels by a label enhancement regularization term during mining semantic similarity structure process. Therefore, RLEFS can better reflect the importance of labels in real-world applications. Furthermore, the local and global structures are considered to ensure RLEFS can capture effective information as much as possible during feature selection process. To verify the effectiveness of RLFES, we conduct numerous experiments on ten multi-label data sets. Besides, RLEFS is compared to six classical and state-of-the-art feature selection methods (PPT+MI,

PPT+CHI, MIFS, MDMR, LRFS and RALM-FS). By analyzing these experimental results, we conduct that the proposed RLEFS method outperforms other compared methods.

In future work, multi-label feature selection is still our main research direction. However, we will furthermore study personalized multi-label feature selection under non-convex optimization due to its broad prospects.

7 Acknowledgments

This work is funded by: Postdoctoral Innovative Talents Support Program under Grant No. BX20190137, and National Key R&D Plan of China under Grant No. 2017YFA0604500, and by National Sci-Tech Support Plan of China under Grant No. 2014BAH02F00, and by National Natural Science Foundation of China under Grant No. 61701190, and by Youth Science Foundation of Jilin Province of China under Grant No. 20160520011JH & 20180520021JH, and by Youth Sci-Tech Innovation Leader and Team Project of Jilin Province of China under Grant No. 20170519017JH, and by Key Technology Innovation Cooperation Project of Government and University for the whole Industry Demonstration under Grant No. SXGJSF2017-4, and by Key scientific and technological R&D Plan of Jilin Province of China under Grant No. 20180201103GX, Project of Jilin Province Development and Reform Commission No. 2019FGWTZC001.

References

[1] Guo, J., Chang, H., Zhu, W.: Preserving ordinal consensus: Towards feature selection for unlabeled data. Proceedings of the AAAI Conference on Artificial Intelligence **34**(1), 75–82 (2020). https://doi.org/10.1609/aaai.v34i01.5336

[2] Guo, B., Tao, H., Hou, C., Yi, D.: Semi-supervised multi-label feature learning via label enlarged discriminant analysis. Knowledge and Information Systems **62**(6), 2383–2417 (2020). https://doi.org/10.1007/s10115-019-01409-3

[3] Komeili, M., Armanfard, N., Hatzinakos, D.: Multiview feature selection for single-view classification. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(10), 3573–3586 (2021). https://doi.org/10.1109/TPAMI.2020.2987013

[4] Wang, Z., Nie, F., Tian, L., Wang, R., Li, X.: Discriminative feature selection via a structured sparse subspace learning module. Twenty-Ninth International Joint Conference on Artificial Intelligence **3**, 3009–3015 (2020). https://doi.org/10.24963/ijcai.2020/416

[5] Gao, W., Li, Y., Hu, L.: Multilabel feature selection with constrained latent structure shared term. IEEE Transactions on Neural Networks and Learning Systems, 1–10 (2021). https://doi.org/10.1109/TNNLS.2021.3105142

[6] Nie, F., Cai, X., Huang, H., Ding, C.: Efficient and robust feature selection via joint l21-norms minimization. In: Advances in Neural Information Processing Systems, pp. 1813–1821 (2010)

[7] Zhu, X., Zhang, S., Hu, R., Zhu, Y., Song, J.: Local and global structure preservation for robust unsupervised spectral feature selection. IEEE Transactions on Knowledge and Data Engineering **30**(3), 517–529 (2018). https://doi.org/10.1109/TKDE.2017.2763618

[8] Zhu, Y., Kwok, J.T., Zhou, Z.-H.: Multi-label learning with global and local label correlation. IEEE Transactions on Knowledge and Data Engineering **30**(6), 1081–1094 (2018). https://doi.org/10.1109/TKDE.2017.2785795

[9] Siblini, W., Kuntz, P., Meyer, F.: A review on dimensionality reduction for multi-label classification. IEEE Transactions on Knowledge and Data Engineering **33**(3), 839–857 (2021). https://doi.org/10.1109/TKDE.2019.2940014

[10] Luaces, O., Díez, J., Barranquero, J., del Coz, J., Bahamonde, A.: Binary relevance efficacy for multilabel classification. Progress in Artificial Intelligence **1** (2012). https://doi.org/10.1007/s13748-012-0030-x

[11] Fürnkranz, J., Hüllermeier, E., Loza Mencía, E., Brinker, K.: Multilabel classification via calibrated label ranking. Machine Learning **73**(2), 133–153 (2008). https://doi.org/10.1007/s10994-008-5064-8

[12] Huang, J., Li, G., Huang, Q., Wu, X.: Learning label-specific features and class-dependent labels for multi-label classification. IEEE Transactions on Knowledge and Data Engineering **28**(12), 3309–3323 (2016). https://doi.org/10.1109/TKDE.2016.2608339

[13] Xu, N., Liu, Y.-P., Geng, X.: Label enhancement for label distribution learning. IEEE Transactions on Knowledge and Data Engineering **33**(4), 1632–1643 (2021). https://doi.org/10.1109/TKDE.2019.2947040

[14] Xiong, C., Qian, W., Wang, Y., Huang, J.: Feature selection based on label distribution and fuzzy mutual information. Information Sciences **574**, 297–319 (2021). https://doi.org/10.1016/j.ins.2021.06.005

[15] Xu, N., Liu, Y.-P., Zhang, Y., Geng, X.: Progressive enhancement of label distributions for partial multilabel learning. IEEE Transactions on Neural

Networks and Learning Systems, 1–12 (2021). https://doi.org/10.1109/TNNLS.2021.3125366

[16] Zhang, J., Lin, Y., Jiang, M., Li, S., Tang, Y., Tan, K.C.: Multi-label feature selection via global relevance and redundancy optimization. Twenty-Ninth International Joint Conference on Artificial Intelligence, 2512–2518 (2020). https://doi.org/10.24963/ijcai.2020/348

[17] Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A.: A review of feature selection methods on synthetic data. Knowledge and Information Systems **34**(3), 483–519 (2013). https://doi.org/10.1007/s10115-012-0487-8

[18] Read, J.: A pruned problem transformation method for multi-label classification. New Zealand Computer Science Research Student Conference (NZCSRS), 143–150 (2008)

[19] Gui, J., Sun, Z., Ji, S., Tao, D., Tan, T.: Feature selection based on structured sparsity: A comprehensive study. IEEE Transactions on Neural Networks and Learning Systems **28**(7), 1490–1507 (2017). https://doi.org/10.1109/TNNLS.2016.2551724

[20] Nie, F., Tian, L., Huang, H., Ding, C.: Non-greedy l21-norm maximization for principal component analysis. IEEE Transactions on Image Processing **30**, 5277–5286 (2021). https://doi.org/10.1109/TIP.2021.3073282

[21] Hu, J., Li, Y., Gao, W., Zhang, P.: Robust multi-label feature selection with dual-graph regularization. Knowledge-Based Systems **203**, 106126 (2020). https://doi.org/10.1016/j.knosys.2020.106126

[22] Shang, R., Xu, K., Shang, F., Jiao, L.: Sparse and low-redundant subspace learning-based dual-graph regularized robust feature selection. Knowledge-Based Systems **187**, 104830 (2019). https://doi.org/10.1016/j.knosys.2019.07.001

[23] Yan, H., Yang, J., Yang, J.: Robust joint feature weights learning framework. IEEE Transactions on Knowledge and Data Engineering **28**(5), 1327–1339 (2016). https://doi.org/10.1109/TKDE.2016.2515613

[24] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the *EM* algorithm. Journal of the Royal Statistical Society: Series B (Methodological) **39**(1), 1–22 (1977). https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

[25] Cai, D., He, X., Han, J., Huang, T.S.: Graph regularized nonnegative matrix factorization for data representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(8), 1548–1560 (2011). https:

//doi.org/10.1109/TPAMI.2010.231

[26] He, Z., Liu, J., Liu, C., Wang, Y., Yin, A., Huang, Y.: Dropout non-negative matrix factorization. Knowledge and Information Systems **60**(2), 781–806 (2019). https://doi.org/10.1007/s10115-018-1259-x

[27] Lee, D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems, vol. 13, pp. 535–541 (2000)

[28] Tsoumakas, G., Spyromitros-xioufis, E., Vilcek, J., Vlahavas, I.: Mulan: A java library for multi-label learning. Machine Learning **12**(7), 2411–2414

[29] Freitas Rocha, V., Varejão, F.M., Segatto, M.E.V.: Ensemble of classifier chains and decision templates for multi-label classification. Knowledge and Information Systems **64**(3), 643–663 (2022). https://doi.org/10.1007/s10115-021-01647-4

[30] Hou, P., Geng, X., Zhang, M.-L.: Multi-label manifold learning. Proceedings of the AAAI Conference on Artificial Intelligence **30**(1), 1680–1686 (2016)

[31] Klimt, B., Yang, Y.: The enron corpus: A new dataset for email classification research. In: Machine Learning: ECML, pp. 217–226 (2004). https://doi.org/10.1007/978-3-540-30115-8_22

[32] Doquire, G., Verleysen, M.: Feature selection for multi-label classification problems. International Work-Conference on Artificial Neural Networks, 9–16 (2011). https://doi.org/10.1007/978-3-642-21501-8_2

[33] Jian, L., Li, J., Shu, K., Liu, H.: Multi-label informed feature selection. Twenty-Fifth International Joint Conference on Artificial Intelligence, 1627–1633 (2016)

[34] Lin, Y., Hu, Q., Liu, J., Duan, J.: Multi-label feature selection based on max-dependency and min-redundancy. Neurocomputing **168**, 92–103 (2015). https://doi.org/10.1016/j.neucom.2015.06.010

[35] Zhang, P., Liu, G., Gao, W.: Distinguishing two types of labels for multi-label feature selection. Pattern Recognition **95**, 72–82 (2019). https://doi.org/10.1016/j.patcog.2019.06.004

[36] Cai, X., Nie, F., Huang, H.: Exact top-k feature selection via 2,0-norm constraint. Twenty-Third International Joint Conference on Artificial Intelligence, 1240–1246 (2013)