Robust multi-label feature selection with shared label enhancement

Yonghao Li¹,² · Juncheng Hu¹,² · Wanfu Gao¹,²

Received: 8 April 2022 / Revised: 30 June 2022 / Accepted: 13 August 2022 / Published online: 1 September 2022
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

Abstract
Feature selection has attracted considerable attention due to the wide application of multi-label learning. However, previous methods do not fully consider the relationship between feature sets and label sets but devote attention to either of them. Furthermore, conventional multi-label learning utilizes logical labels to estimate relevance between feature sets and label sets so that the importance of corresponding labels cannot be well reflected. Additionally, numerous irrelevant and redundant labels degrade the classification performance of models. To this end, we propose a multi-label feature selection method named Robust multi-label Feature Selection with shared Label Enhanced (RLEFS). First, we obtain a robust label enhancement term by reconstructing labels from logical labels to numerical labels and imposing $l_{2,1}$-norm onto the label enhancement term. Second, RLEFS utilizes the robust label enhancement term to share the similar latent semantic structure between feature matrix and label matrix. Third, local structure is considered to ensure the consistency of label information during the feature selection process. Finally, we integrate the above terms into one joint learning framework, and then, a simple but effective optimization method with provable convergence is proposed to solve RLEFS. Experimental results demonstrate the classification superiority of RLEFS in comparison with seven state-of-the-art methods.

Keywords Feature selection · Multi-label learning · Sparse learning · Label enhancement

Juncheng Hu and Wanfu Gao have contributed equally to this work.

✉ Juncheng Hu
jchu19@mails.jlu.edu.cn

Yonghao Li
yonghao17@mails.jlu.edu.cn

Wanfu Gao
gaowf@jlu.edu.cn

¹ College of Computer Science and Technology, Jilin University, Changchun 130012, China
² Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China
1 Introduction

In real-world applications, data mining is confronting a great challenge with the increase in high-dimensional data, because the high dimensionality of data brings a lot of noise information and corrodes the classification performance of the learning models [1]. To address the issue, numerous feature selection methods are designed to reduce the number of irrelevant and redundant features, to excavate useful information, and improve the classification performance [2–5]. Feature selection technique is widely used in extensive domains, such as communications-electronics, biomedical, computational chemistry and virus detection, etc.

In recent years, a large number of researches regarding feature selection methods are proposed. Generally, feature selection methods are categorized into three types based on the selection strategy [6]: filter models [7], wrapper models [8] and embedded models [4, 9]. Filter models are independent of the subsequent learning approach in the implementation process, while wrapper models are dependent on the specific learning approach. Embedded models embed feature selection and the learning approach into one unified framework, which becomes a hot topic in the areas of feature selection [10–12]. Embedded-based feature selection is focused in this paper.

In the early stage, researchers use previous methods to deal with binary or multi-class label data. With the explosive growth of multi-label data, numerous multi-label learning methods have been proposed to handle multi-label data. Most of the existing multi-label learning methods do not fully consider the relationship between the feature set and the label set simultaneously but devote attention to either of them. It means that these methods are good at dealing with either the relationship between features or the relationship between labels. For instance, Jian et al. propose a multi-label method named MIFS [13]. MIFS method focuses on obtaining the low-dimensional form of label matrix by matrix decomposition. Besides, Cai and Zhu consider the feature manifold of features to conduct multi-label feature selection but ignoring the label correlation which is important for multi-label learning [14]. Moreover, there exists close relationship between the feature set and the label set. It is indicated that the similar latent semantic structure between the feature set and the label set is consistent in the literature [15]. Therefore, we propose a new term to share the latent semantic similar structure between the feature set and the corresponding label set, combining local and global structures of labels to provide complementary information so as to improve the performance of multi-label feature selection, i.e., we simultaneously exploit local and global structures from the label-level perspective. The literature works [16–18] are several highly related studies. In [16, 17], researchers demonstrate that two kinds of geometry structure (both local and global geometry structure) among instances can provide complimentary information to each other so that the performance of unsupervised feature selection methods is strengthened. However, unsupervised methods ignore the relationship between labels and the significance of label enhancement. In addition, the local and global structures may occur in any real-world applications, i.e., some structures provide global information and some information are shared in local structure. Therefore, Zhu et al. [18] propose a global and local label correlation method to deal with multi-label learning. However, this method not only ignores multi-label enhancement but is also unable to conduct feature selection. Therefore, these approaches cannot be applied to joint feature selection and multi-label learning with label enhancement directly. But the proposed method solves this intractable problem. Additionally, traditional multi-label feature selection methods utilize logical labels to estimate the relevance between the feature set and label set [15, 19, 20]. For example, there is a picture that includes sunshine, flying bird, sky and white cloud. Usually, we use 1 and 0 (or -1) to denote the
relevant or irrelevant object of this picture. In this case, we find that the importance of each label is the same for each instance. However, the importance of labels cannot be well reflected by logical values [21, 22]. To this end, we design a robust label enhancement term, which reconstructs labels from logical labels to numerical labels. Furthermore, there exist a lot of irrelevant and redundant information in labels, as mentioned above. As a result, low-quality labels cannot offer the accurate guidance for feature selection. We impose $l_{2,1}$-norm onto the label enhancement term so that we can obtain robust labels to guide feature selection process. Finally, the consistency of label information is taken into account in the design of the proposed method.

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, which is named Robust multi-label Feature Selection with shared Label Enhancement (RLEFS). And then, an effective optimization method with provable convergence is proposed to solve RLEFS. In summary, the novelties and contributions of this paper are highlighted as follows:

- We design a robust label enhancement term by reconstructing labels from logical labels to numerical labels and imposing $l_{2,1}$-norm onto the label enhancement term.
- The proposed method shares the new robust labels that we design between the feature set and the corresponding label set to capture the latent semantic similar structure. At the same time, the consistency of label information is taken into account in the design of the proposed method.
- Designing a joint learning framework that named Robust multi-label Feature Selection with shared Label Enhancement (RLEFS), and developing an optimization method with provable convergence to solve the proposed RLEFS framework.
- Conducting comprehensive evaluation criteria on multiple benchmark data sets to demonstrate the effectiveness of the proposed framework.

The remainder of the paper has the following organizational form. In Sect. 2, we review the related work including some related notations and representative studies. In Sects. 3 and 4, we propose a joint learning framework named RLEFS and one optimization method with provable convergence, respectively. In Sect. 5, the comprehensive experimental results on multiple multi-label data sets are presented and discussed. Finally, some concluding remarks are given in Sect. 6.

2 Related work

2.1 Preliminaries

In this subsection, some notations are defined, which are used through this paper. Matrices are denoted by italicized uppercase letters, such as $A$. For matrix $A \in \mathbb{R}^{n \times m}$, $A_i$ and $A_{j\cdot}$ denote $i$th row vector and $j$th column vector of $A$, respectively. Scalars are denoted by lowercase letters, such as $a$. Functions can be represented by calligraphic letters, such as $\mathcal{L}$. $A^T$ and $Tr(A)$ denote the transpose and the trace of $A$, respectively, where $A$ in $Tr(A)$ 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 corresponding label matrix $Y \in \mathbb{R}^{n \times c}$ has $c$ column class labels. Generally, $Y_{ij} = 1$ if the $i$th instance is related with the $j$th label, and $Y_{ij} = 0$ otherwise.
2.2 Related work

In the past decade, a large number of established multi-label learning methods have been widely applied to different fields. Generally, these methods are categorized into three groups from the perspective of label correlations, i.e., first-order strategy, second-order strategy and high-order strategy [23]. The first-order strategy, such as Binary Relevance (BR) [24], utilizes single-label learning method to deal with multi-label data, which is not only time-consuming but also ignores the label correlations. In the second-order strategy, the pairwise label correlations attract extensive concern. For instance, Global Relevance and Redundancy Optimization (GRRO) is a state-of-the-art second-order-based method [25]. For the high-order strategy, the correlations of multiple labels or all labels are taken into account. For example, Huang et al. propose a label-specific-features-based multi-label learning method named LLSF-DL [3]. LLSF-DL utilizes a sparse stacking way to conduct high-order label correlations learning. However, previous methods based on the three strategies assume that the importance of labels is equivalent. That is, the traditional multi-label learning utilizes logical labels to estimate the relevance between the feature set and label set so that the importance of labels cannot be well reflected. To this end, some researchers transform logical labels into numerical labels. For instance, Hou et al. propose manifold-based multi-label learning method to translate logical labels into numerical labels [21]. However, this method focuses on the manifold learning of multi-label set. Huang et al. design a feature selection method that considers Laplacian score based on manifold learning named MCLS [26]. Inspired by [21], MCLS method utilizes the manifold learning to transform the binary label to the numerical label. However, these methods only focus on handling label sets.

Based on whether or not to deal directly with multi-label data, multi-label learning is mainly divided into two types: problem transformation methods and algorithm adaption methods [23]. The former class is to transform the multi-label classification into several single-label classification problems, such as Pruned Problem Transformation (PPT) [27]. Doquire et al. propose a multi-label feature selection method based on PPT, which employs mutual information to evaluate features named PPT+MI [28]. Similarly, PPT+CHI employs $\chi^2$ statistic method to select the important features after the preprocessing of PPT [27]. However, there exist some problems in these problem transformation methods. First, the geometry structure of the labels was not considered. Second, these methods use logical labels in design of methods. The latter one is to transform algorithms to fit multi-label data, which handles multi-label data directly. Generally, algorithm adaption multi-label feature selection methods adopt different criteria in design of methods, such as mutual information-based and sparse learning-based multi-label feature selection methods [1]. Several representative methods are reviewed.

MIFS [13] uses matrix factorization to obtain a low-rank latent label subspace that preserves the local geometrical structure of labels based on the instance-level regularizer. The obtained low-rank latent label subspace eliminates irrelevant and redundant information of the original label matrix. However, the multi-label matrix of MIFS is decomposed into two factor matrices. The latent label matrix may lose effective information when the value of the rank $k$ of the latent label matrix is inaccurate. MIFS is constructed as follows:

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

(1)

where $W \in \mathbb{R}^{d \times c}$, $V \in \mathbb{R}^{n \times k}$ 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. $\alpha$, $\beta$ and $\gamma$ are three regularization parameters of MIFS method.
Besides, Cai et al. also design a feature selection method based on the sparse-learning theory, which imposes $l_{2,0}$-norm onto the weight matrix and then uses the Augmented Lagrangian Multiplier method to optimize the following objective function \cite{29}. It is named RALM-FS:

$$\min_{W, V, B} \| Y - X^T W - 1b^T \|_{2,1} \quad \text{s.t.} \quad \| W \|_{2,0} = k$$  \hspace{1cm} (2)

where $b$ and $I$ denote the bias term and the all-one-element vector, respectively. $k$ denotes the number of the already-selected features. However, this method ignores the relationship between labels.

Alternatively, Lin et al. design a multi-label feature selection method, which maximizes the feature dependency while minimizing feature redundancy named MDMR \cite{30}. Besides, Zhang et al. propose a new feature relevance term based on label redundancy, and they combine the new feature relevance term with feature redundancy term to design a new method named Label Redundancy-based multi-label Feature Selection (LRFS) \cite{2}. LRFS has the following form:

$$J(f_k) = LR(f_k; Y) - \frac{1}{|S|} \sum_{f_j \in S} I(f_k; f_j)$$

$$= \sum_{y_i \in Y} \sum_{y_j \neq y_i, y_j \in Y} I(f_k; y_j | y_i) - \frac{1}{|S|} \sum_{f_j \in S} I(f_k; f_j)$$  \hspace{1cm} (3)

where $LR(f_k; Y)$ and $I(f_k; f_j)$ are regarded as the feature relevance term and the feature redundancy term, respectively. $f_k$ denotes a candidate feature that is from the feature set $F$. $y_i$ and $y_j$ are two labels that are from the label set $Y$. To balance the magnitude between feature relevance term and feature redundancy term, feature redundancy term is divided over the number of the already-selected feature subset $S$. Both MDMR and LRFS methods belong to multi-label feature selection methods based on information theory. In addition, we analyze the advantages and disadvantages of these existing state-of-the-art methods. For instance, Jian et al. \cite{13} realize the existence of invalid labels that could restrain the prediction performance of feature selection. Therefore, multi-label informed feature selection method (MIFS) is proposed. The advantages are that it utilizes non-negative matrix factorization technology to reduce invalid original label information and maintains the structural consistency by constructing a conventional Laplacian matrix. In addition, the disadvantages are that the original labels can be decomposed into two factors with mixed values \cite{31}. Moreover, MIFS utilizes the input instance space to generate local manifold graph structure. However, the local and global structure have been demonstrated to provide complementary information to each other \cite{16, 17}. Furthermore, MIFS ignores the importance of label enhancement for feature selection. As a result, MIFS results in serious classification error. To this end, we tend to exploit label enhancement with label-level manifold regularization term to guide feature selection. RALM-FS holds that $l_{2,0}$-norm has better sparsity than $l_{2,1}$-norm. Therefore, 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. However, RALM-FS only selects the fixed number of features and the weight of other features is forcibly adjusted to zero. At the same time, the importance of label is ignored due to the nature of $l_{2,0}$-norm. Moreover, RALM-FS is unable to achieve label enhancement. Gao et al. \cite{32} design a feature selection method based on shared patterns named SSFS. The advantage is that it utilizes a constrained latent structure to achieve a shared regularization term so that multi-label instances with high-dimensional features and labels can be addressed. This shared space encapsulates important information between the features and the labels simultaneously. The disadvantage is that it
not only constructs a fixed local instance-level manifold graph regularizer, but also ignores the label enhancement.

In addition, local learning regularizer is adopted in this paper. In local learning regularizer, an intuitive assumption is adopted, that is, if \( X_i \) and \( X_j \) in a high-dimensional space are close, then \( Y_i \) and \( Y_j \) in a low-dimensional space should be similar [33]. The graph Laplacian is used to achieve the above assumption. Therefore, the following calculation method is obtained:

\[
\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} S_{ij}^X \| Y_i - Y_j \|_2^2
\]

\[
= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} S_{ij}^X (Y_i - Y_j)(Y_i - Y_j)^T
\]

\[
= \sum_{i=1}^{n} Y_i Y_i^T A_{ii}^X - \sum_{i=1}^{n} \sum_{j=1}^{n} Y_i Y_j^T S_{ij}^X
\]

\[
= Tr(Y^T (A^X - S^X) Y)
\]

\[
= Tr(Y^T L^X Y)
\]

where \( L^X = A^X - S^X \) denotes a graph Laplacian matrix of matrix \( X \). \( S^X \) and \( A^X \) denote the symmetric affinity matrix and the degree matrix, respectively.

Finally, we review the recent advanced work. Hashemi et al. [34] discover the ParetoCluster method lacks discrimination in high-dimensional cases, where ParetoCluster uses Pareto dominance and cluster analysis to deal with multi-label feature selection. To deal with this problem, they propose a Pareto-based multi-label feature selection method named PMFS. PMFS transforms the multi-label issue into a bi-objective optimization issue by considering feature relevancy and feature redundancy simultaneously. Paniri et al. [35] propose an Ant Colony Optimization-based feature selection method to deal with the multi-label learning problem. This method uses a heuristic learning scheme with Temporal Difference (TD) reinforcement learning instead of the static heuristic function when dealing with ACO-based feature selection problems. Hashemi et al. [36] impose joint feature selection and multi-label learning method into a bipartite graph (feature graph and label graph) matching process. This method achieves superior classification performance. Moreover, Hashemi et al. [37] also propose a multi-label feature selection method during the multi-criteria decision-making process for the first time. These method belong to the state-of-the-art multi-label feature selection methods. In addition, Fan et al. [38] propose a state-of-the-art structured subspace multi-label feature selection method. This method utilizes global and local label correlations by manifold learning to uncover a compact latent subspace. However, this method ignores the importance of label enhancement. In [39], the adaptive spectral graph with information entropy is designed due to its nature, i.e., simple and effective. However, the information entropy theory-based strategy leads to be more time-consuming. [39] consists of two parts, i.e., the local cliques are integrated into the global discriminant model and exploring clustering results by high-order label correlations. Similar to the aforementioned literature, this method ignores the importance of label enhancement. Liu et al. [40] propose an Online Multi-label Group Feature Selection (OMGFS) method by considering online intra-group and online inter-group sections. OMGFS uses an information-theoretic-based strategy so that it has more time-consuming, while label enhancement is ignored. To this end, we propose the following method.
3 The proposed framework

In this section, we propose a novel robust multi-label feature selection method that is named Robust multi-label Feature Selection with shared Label Enhancement (RLEFS). The core contribution of the proposed method can be concluded as follow: RLEFS utilizes the robust label enhancement term to share the similar latent semantic structure between the feature matrix and the label matrix. Additionally, the local structure is considered in the design of RLEFS to ensure the consistency of label information as much as possible during the feature selection process. The details of the proposed method are presented in the followings.

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 $l_{2,1}$-norm has been confirmed to be robust to noise information \[20\]. Therefore, the following learning model is given:

$$\min_W \|XW - Y\|_{2,1}$$

where $W \in \mathbb{R}^{d \times c}$ denotes the feature weight matrix that measures the importance of each feature. The greater the value of $\|W_i\|_2$ is, the larger contribution the $i$th feature of matrix $X$ is. Furthermore, motivated by literature \[18, 41\], it is vital to preserve the local structure of labels, which means if $W_i$ and $W_j$ are close, then $Y_i$ and $Y_j$ should be close. Thus, the corresponding graph Laplacian is presented by the following form:

$$\frac{1}{2} \sum_{i=1}^c \sum_{j=1}^c S_{ij} \|W_i - W_j\|^2_2$$

$$= \frac{1}{2} \sum_{i=1}^c \sum_{j=1}^c S_{ij}(W_i - W_j)(W_i - W_j)^T$$

$$= \sum_{i=1}^c W_i W_i^T A_{ii} - \sum_{i=1}^c \sum_{j=1}^c W_i W_j^T S_{ij}$$

$$= Tr(W(A - S)W^T)$$

$$= Tr(WLW^T)$$

where Formula (6) utilizes the weight matrix to preserve the relationship between labels from the perspective of label level. We incorporate Formula (6) into Formula (5) to obtain Formula (7):

$$\min_W \|XW - Y\|_{2,1} + \alpha Tr(WLW^T)$$

where $L = A - S$ denotes a graph Laplacian matrix of label set. $\alpha$ is a trade-off parameter between loss function and label graph regularization term. However, traditional multi-label feature selection methods utilize logical labels to estimate the relevance between the feature set and label set so that the importance of labels cannot be well reflected. Therefore, we design a robust label enhancement term by reconstructing labels from logical labels to numerical labels and imposing $l_{2,1}$-norm onto the label enhancement term. We obtain the following form:

$$\min_{W, F, B} \|XW - F\|_{2,1} + \alpha Tr(WLW^T) + \beta \|F - YB\|_{2,1}$$
where $\beta$ denotes a regularization parameter that adjusts the contribution of the robust label enhancement term. $F \in \mathbb{R}^{n \times c}$ denotes an enhanced label space, which shares the similar latent semantic structure between the feature space and label space. It is worthy to pointed that $F$ is not a low-dimensional embedding subspace for label matrix $Y$, but a homomorphic space. This homomorphic label space will not lead to information loss of the original logical label space compared with some methods that directly reduce the dimension of high-dimensional space to low-dimensional space. Besides, we impose $l_2,1$-norm onto the label enhancement regularization term to enhance the robustness of the enhanced label space $F$, which can prevent noise interference. $B \in \mathbb{R}^{c \times c}$ denotes a self-representation mapping term, which can adjust the correlation between the enhanced label space $F$ and the original logical label space $Y$. Finally, we impose $l_2,1$-norm onto the feature weight matrix $W$, which ensures not only the row sparsity of $W$ but also can automatically select the most discriminating and effective features during the multi-label learning. Afterward, we reformulate the following function:

$$
\min_{W, F, B} \| X W - F \|_{2,1} + \alpha Tr(W L W^T) + \beta \| F - Y B \|_{2,1} + \gamma \| W \|_{2,1}$$

where $\gamma$ 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 [42]. 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 in the used data sets. Note that, if the elements of $F$ are all zero, the above function always leads to a trivial solution. To this end, we impose an orthogonality constraint onto $F$. The final objective function is reformulated as follows:

$$
\min_{W, F, B} \| X W - F \|_{2,1} + \alpha Tr(W L W^T) + \beta \| F - Y B \|_{2,1} + \gamma \| W \|_{2,1}
$$

s.t. $\{ W, F, B \} \geq 0$, $F^T F = I$ (10)

Next, we develop a simple yet effective optimization method to guarantee the convergence of function (10). The 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 a simple but efficient optimization method to solve the proposed objective function (10). We observe that the objective function (10) is joint non-convex because the Hessian matrix 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, one variable is adjusted, while other variables are fixed. At the same time, a relaxed approach is introduced to solve the non-smooth problem. The objective function (10) is equivalent to objective function (11) as follows:

$$
\Theta(W, F, B) = 2 Tr[(XW - F)^T D_1(XW - F)] + \alpha Tr(W L W^T)
+ 2 \beta Tr[(F - Y B)^T D_2(F - Y B)] + 2 \gamma Tr(W^T D_3 W)
$$

s.t. $\{ W, F, B \} \geq 0$, $F^T F = I$ (11)

Springer
where $\Theta(W, F, B)$ denotes the objective function (11) with respect to variables $W, F$ and $B$. $D_1, D_2$ and $D_3$ are defined as follows, respectively:

$$
\begin{align*}
D_{1i} &= \frac{1}{\|(XW-F)_i\|_2^2 + \epsilon} \\
D_{2i} &= \frac{1}{\|(F-YB)_i\|_2^2 + \epsilon} \\
D_{3i} &= \frac{1}{\|W_i\|_2^2 + \epsilon}
\end{align*}
$$

(12)

where $\| \cdot \|_2$ denotes the $l_2$-norm of a vector. $D_{1i}$, $D_{2i}$ and $D_{3i}$ denote the $i$th diagonal element of $D_1$, $D_2$ and $D_3$, respectively. $\epsilon$ is a non-negative small constant. By integrating non-negative and orthogonal constraints into function (11), we obtain the following Lagrangian function:

$$
\mathcal{L}(W, F, B) = 2Tr[(XW - F)^TD_1(XW - F)] + \alpha Tr(WLW^T) + 2\beta Tr[(F - YB)^TD_2(F - YB)] + 2\gamma Tr(W^TD_3W) + \frac{\lambda}{2}F^TF - I ||_F^2 - Tr(\Phi W^T) - Tr(\Psi F^T) - Tr(\Omega B^T)
$$

(13)

where $\Phi \in \mathbb{R}^{d \times c}$, $\Psi \in \mathbb{R}^{n \times c}$ and $\Omega \in \mathbb{R}^{d \times n}$ represent the Lagrangian multiplier, respectively. $\lambda$ denotes the regularization parameter of orthogonal constraint. By taking the derivative of function (13) w.r.t. $W$, $F$ and $B$, respectively, we obtain

$$
\begin{align*}
\frac{\partial \mathcal{L}}{\partial W} &= 2XTD_1XW - 2XTD_1F + 2\alpha WL + 2\gamma D_3W - \Phi \\
\frac{\partial \mathcal{L}}{\partial F} &= -2D_1XW + 2D_1F + 2\beta D_2F - 2\beta D_2YB + 2\lambda FF^TF - 2\lambda F - \Psi \\
\frac{\partial \mathcal{L}}{\partial B} &= -2\beta YT D_2F + 2\beta YT D_2YB - \Omega
\end{align*}
$$

(14)

Due to $\Phi_{ij}W_{ij} = 0$, $\Psi_{ij}F_{ij} = 0$ and $\Omega_{ij}B_{ij} = 0$ according to Karush-Kuhn-Tucker conditions, we get the following Formulas:

$$
\begin{align*}
(X^TD_1XW - XT^TD_1F + \alpha WL + \gamma D_3W)_{ij}W_{ij} &= 0 \\
(-D_1XW + D_1F + \beta D_2F - \beta D_2YB + \lambda FF^TF - \lambda F)_{ij}F_{ij} &= 0 \\
(-\beta YT D_2F + \beta YT D_2YB)_{ij}B_{ij} &= 0
\end{align*}
$$

(15)

According to Formula (15), we obtain the following update rules:

$$
\begin{align*}
W_{ij}^{t+1} &\leftarrow \frac{(X^TD_1F + \alpha WL)_{ij}}{(X^TD_1XW + \alpha WL + \gamma D_3W)_{ij}} \\
F_{ij}^{t+1} &\leftarrow \frac{(D_1XW + \beta D_2YB + \lambda F)_{ij}}{(D_1F + \beta D_2YB + \lambda FF^TF)_{ij}} \\
B_{ij}^{t+1} &\leftarrow \frac{(\gamma YT D_2F)_{ij}}{(\gamma YT D_2YB)_{ij}}
\end{align*}
$$

(16)

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., $A$ and $S$. Besides, some elements in the denominator could be zero during the updated process. To solve this issue, we add a very small constant to the denominator. Then, we obtain the top-$k$ features during the feature selection process. The pseudo-code of the proposed method is described in Algorithm 1. In Algorithm 1, RLEFS has three phases. Phase 1 (lines 1, 3, 4 and 5) provides the input terms. Line 1 gives the input data set and regularization parameters, such as the input feature matrix $X \in \mathbb{R}^{d \times d}$, the output matrix $X \in \mathbb{R}^{n \times c}$ as well as regularization parameters $\alpha$, $\beta$, $\gamma$ and $\lambda$. Lines 3–5 provide initialized value of all variables. For instance, we randomly initialize the variables $W$, $F$ and $B$ in line 3. Phase 2 (lines 6–10) shows the update process by using the aforementioned update
rules. Phase 3 (lines 11 and 12) sorts all features by $\|W_i\|_2$, where $i = 1, 2, 3, \ldots, d$, and select the top-$k$ features. Then, we use top-$k$ selected features to train subsequent classifiers.

**Algorithm 1 RLEFS**

**Input:**
1: The input feature matrix $X \in \mathbb{R}^{n \times d}$ and the output label matrix $Y \in \mathbb{R}^{n \times c}$; The regularization parameters $\alpha, \beta, \gamma$ and $\lambda$.

**Output:**
2: Return the top-$k$ selected features index set.

3: Initialize $W \in \mathbb{R}^{d \times c}$, $F \in \mathbb{R}^{n \times c}$ and $B \in \mathbb{R}^{c \times c}$ randomly;
4: $t = 0$;
5: Compute the degree matrix $A$ and the affinity matrix $S$ of the label matrix $Y$;
6: **Repeat:**
7: Update the diagonal matrix $D_1, D_2$ and $D_3$ by formula (12);
8: Update $W$, $F$ and $B$ by formula (16);
9: $t = t + 1$;
10: **Until** convergence criterion is satisfied;
11: Return $W$;
12: Sort all features by $\|W_i\|_2$, where $i = 1, 2, 3, \ldots, d$, and select the top-$k$ features.

**4.2 Proof of convergence**

In this subsection, we offer the proof of the convergence of the proposed optimization method. Taking the variable $W$ as one example, we obtain the following formula according to the gradient descent method:

$$W_{ij}^{t+1} \leftarrow W_{ij}^t - \eta \left( \frac{\partial \Theta}{\partial W_i} \right)_{ij}$$  \hspace{1cm} (17)

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

$$\eta = \frac{W_{ij}^t}{2(X^TD_1XW + \alpha WA + \gamma D_3W)_{ij}}$$  \hspace{1cm} (18)

We take Formula (18) into Formula (17), then the following result is presented:

$$W_{ij}^{t+1} \leftarrow W_{ij}^t - \frac{W_{ij}^t}{2(X^TD_1XW + \alpha WA + \gamma D_3W)_{ij}} \left( \frac{\partial \Theta}{\partial W_i} \right)_{ij}$$

$$\Leftrightarrow W_{ij}^{t+1} \leftarrow W_{ij}^t \frac{(X^TD_1F + \alpha WS)_{ij}}{(X^TD_1XW + \alpha WA + \gamma D_3W)_{ij}}$$  \hspace{1cm} (19)

Accordingly, the above update rule is a special case of the gradient descent method. The convergence of the optimization method is proved based on [43, 44].

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

**Lemma 1** If $G(w, w')$ is considered to be an auxiliary function of $F(w)$, then $F(w)$ is a non-increasing function according to

$$w_{t+1} = \arg \min_w G(w, w')$$  \hspace{1cm} (20)
Proof of Lemma 1  According to Definition 1 and Formula (20), we can deduce that
\[
\mathcal{F}(w^{t+1}) \leq \mathcal{G}(w^{t+1}, w^t) \leq \mathcal{G}(w^t, w^t) = \mathcal{F}(w^t)
\]  
(21)

where \( w \) denotes any element of \( W \). Next, we give the proof of the convergence of RLEFS w.r.t. \( W \) by a proper auxiliary function \( \mathcal{G}(w, w^t) \). Considering that the update rules are always based 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:
\[
\mathcal{F}_{ij}' = 2(X^TD_1XW - 2X^TD_1F + 2\alpha WL + 2\gamma D_3 W)_{ij}
\]
(22)
\[
\mathcal{F}_{ij}'' = 2(X^TD_1X)_{ii} + 2\alpha(L)_{jj} + 2\gamma(D_3)_{ii}
\]
(23)

Afterward, we obtain the Taylor function of \( \mathcal{F}_{ij}(W_{ij}) \):
\[
\mathcal{F}_{ij}(W_{ij}) = \mathcal{F}_{ij}(W_{ij}^t) + \mathcal{F}_{ij}'(W_{ij}^t)(W_{ij} - W_{ij}^t) + \frac{1}{2}\mathcal{F}_{ij}''(W_{ij}^t)(W_{ij} - W_{ij}^t)^2
\]
(24)

Inspired by [44], we set the following auxiliary function regarding \( \mathcal{F}_{ij}(W_{ij}) \):
\[
\mathcal{G}(W_{ij}, W_{ij}^t) = \mathcal{F}_{ij}(W_{ij}^t) + \mathcal{F}_{ij}'(W_{ij}^t)(W_{ij} - W_{ij}^t) + \frac{(X^TD_1XW + \alpha WA + \gamma D_3 W)_{ij}}{W_{ij}^t}(W_{ij} - W_{ij}^t)^2
\]
(25)

Proof  If \( W_{ij} = W_{ij}^t \) in the Formula (25), then \( \mathcal{G}(W_{ij}, W_{ij}^t) = \mathcal{F}_{ij}(W_{ij}^t) \). As for the other condition \( \mathcal{G}(W_{ij}, W_{ij}^t) \geq \mathcal{F}(W_{ij}) \) in Definition 1, we need to prove the following inequality:
\[
\frac{(X^TD_1XW + \alpha WA + \gamma D_3 W)_{ij}}{W_{ij}^t} \geq (X^TD_1X)_{ii} + \alpha(L)_{jj} + \gamma(D_3)_{ii}
\]
(26)

It is obvious that Formula (26) is equivalent to the following form:
\[
\begin{align*}
(X^TD_1XW + \gamma D_3 W)_{ij} &= \sum_{l=1}^{d} (X^TD_1X + \gamma D_3)_{il}W_{ij}^l \\
\alpha(WA)_{ij} &= \alpha \sum_{l=1}^{d} W_{ij}^l(A)_{lj} \geq \alpha W_{ij}^l(A)_{lj} \geq \alpha W_{ij}^l(A - S)_{lj} = \alpha W_{ij}^l(L)_{lj}
\end{align*}
\]
(27)

Therefore, \( \mathcal{G}(W_{ij}, W_{ij}^t) \) is proved to be an auxiliary function of \( \mathcal{F}_{ij}(W_{ij}) \). We brought the auxiliary function into Formula (20), then we obtain the update rule of \( W \) by \( \frac{\partial \mathcal{G}(W_{ij}, W_{ij}^t)}{\partial W_{ij}} = 0 \):
\[
W_{ij}^{t+1} \leftarrow W_{ij}^t - W_{ij}^t \frac{\mathcal{F}_{ij}'(W_{ij}^t)}{2(X^TD_1XW + \alpha WA + \gamma D_3 W)_{ij}}
\]
\[
\leftarrow W_{ij}^t \frac{(X^TD_1F + \alpha WS)_{ij}}{(X^TD_1XW + \alpha WA + \gamma D_3 W)_{ij}}
\]
(28)

As a consequence, the objective function (10) is non-increasing under (28) according to the above proof. The convergent proof of the other two variables \( (F \) and \( B) \) is similar to the variable \( W \).
Table 1 Description of data sets

| Data sets | #Instances | #Train | #Test | #Features | #Labels | Domains |
|-----------|------------|--------|-------|-----------|---------|---------|
| Flags     | 194        | 129    | 65    | 19        | 7       | Images  |
| Enron     | 1702       | 1123   | 579   | 1001      | 53      | Text    |
| Arts      | 5000       | 2000   | 3000  | 462       | 26      | Text(web) |
| Education | 5000       | 2000   | 3000  | 550       | 33      | Text(web) |
| Entertain | 5000       | 2000   | 3000  | 640       | 21      | Text(web) |
| Reference | 5000       | 2000   | 3000  | 793       | 33      | Text(web) |
| Scene     | 2407       | 1211   | 1196  | 294       | 6       | Images  |
| Science   | 5000       | 2000   | 3000  | 743       | 40      | Text(web) |
| Society   | 5000       | 2000   | 3000  | 636       | 27      | Text(web) |
| Social    | 5000       | 2000   | 3000  | 1047      | 39      | Text(web) |
| Corel5K   | 5000       | 3350   | 1650  | 499       | 374     | Image   |

5 Experimental settings and results

In this section, we conduct experiments on eleven multi-label benchmark data sets in comparison with the seven state-of-the-art multi-label feature selection methods, where all the experiments are performed on the 3.4GHz i7-6700 machine with 16 GB RAM.

5.1 Experimental data sets

In our experiments, all data sets are fetched from Mulan Library [45] in which these data sets are also adopted in numerous literature regarding multi-label learning [4, 12, 46–48]. Besides, these data sets are collected from different fields. For example, the Enron data set is a subset of the Enron e-mail corpus [49], which comes from text domain. Flags data set is collected from the image field, which has 194 instances and 7 labels that contains the color of red, green and blue, etc. Several data sets come from the yahoo data set that belongs to multi-label text (web page) categorization. The detailed description of all the used data sets is summarized in Table 1.

5.2 Experimental settings

To demonstrate the effectiveness of the proposed method RLEFS, the following classic and state-of-the-art methods are used as compared methods, i.e., PPT+MI [28], PPT+CHI [27], MIIFS [13], MDMR [30], LRFS [2], RALM-FS [29] and SSFS [32]. The detailed analysis of these compared methods has shown in the related work. Then, we introduce some experimental parameters in advance. The heat-kernel function is adopted for graph Laplacian matrix:

\[
S_{ij} = \begin{cases} 
    e^{-\frac{|Y_i - Y_j|^2}{\sigma}}, & \text{if } Y_i \in N_p(Y_j) \text{ or } Y_j \in N_p(Y_i) \\
    0, & \text{otherwise}
\end{cases}
\]  

(29)

where the parameters \( p \) and \( \sigma \) are set as 5 and 1, respectively. To ensure the fairness, the hyperparameters of all the methods are tuned in the same grid, i.e., \{0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}. 

 Springer
We utilize the optimal parameters with respect to classification performance during the training process, where fivefold cross-validation is adopted. Besides, BR [24] 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}, \ldots, 10^{3}, 10^{4}\}. At the same time, ML-kNN (k=10) is adopted as well, where ML-kNN is a common multi-label learning classifier [50]. Next, we adopt Micro-$F_1$ and Macro-$F_1$ (a.k.a. Micro-average and Macro-average) based on $F_1$-measure and Hamming Loss as evaluation criteria to evaluate the proposed method and other compared methods [51]. Micro-$F_1$ and Macro-$F_1$ are two label-based evaluation metrics. They have the following forms:

\[
\begin{align*}
\text{Micro-} F_1 & = \frac{\sum_{i=1}^{m} 2TP_i}{\sum_{i=1}^{m} (2TP_i + FP_i + FN_i)} \\
\text{Macro-} F_1 & = \frac{1}{m} \sum_{i=1}^{m} \frac{2TP_i}{2TP_i + FP_i + FN_i}
\end{align*}
\]

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 value of Micro-$F_1$ (or Macro-$F_1$) indicates the better classification performance of the method. Besides, Hamming Loss ($HL$) belongs to the example-based evaluation metric. It has the following form:

\[
\text{Hamming Loss} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m} |Y'_{ij} \oplus Y_{ij}|
\]

where $n$ and $m$ denote the number of instances and labels, respectively. $Y'_{ij}$ and $Y_{ij}$ denote the $(i, j)\text{th}$ predicted label and the $(i, j)\text{th}$ label of the original label set. In addition, we suppose $D = \{(X_i, Y_i) | 1 \leq i \leq n\}$ denotes a test data set. $\hat{Y}$ denotes a predicted label set by the learned multi-label classifier. $\oplus$ denotes the XOR operator. The Hamming Loss measures the number of misclassified instances and label pairs. Moreover, the smaller value of Hamming Loss indicates the better classification performance of the method. These different types of evaluation metrics can measure multi-label methods from multiple aspects.

5.3 Experimental results

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

Tables 2, 3, 4 and 5 record the results of Macro-$F_1$ and Micro-$F_1$ of all the methods on the classifiers SVM and KNN ($K = 3$). Table 6 records the results of Hamming Loss of all the methods on the ML-kNN classifiers. In these tables, all methods are ranked on each data set, such as the optimal method ranks first, the suboptimal method ranks second. The average rank is shown if several methods have the same rank result. The average result of all data sets by each method is shown in the final row in each table. Observing these tables, we conclude that the proposed method RLEFS obtains the best average results, where the best average results of the RLEFS are 0.161, 0.355, 0.183, 0.375 and 0.0722 in Tables 2, 3, 4, 5 and 6, respectively. The results of all methods using SVM classifier in terms of Macro-$F_1$ are described in Table 2. Through observing Table 2, we find that RLEFS outperforms PPT+MI,
| Data sets  | PPT+MI        | PPT+CHI        | MIFS          | MDMR          | LRFS          | RALM-FS       | SSFS          | RLEFS         |
|------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| Flags      | 0.513±0.037(4)| 0.517±0.046(3)| 0.57±0.1(2)   | 0.503±0.053(6)| 0.511±0.044(5)| 0.487±0.044(8)| 0.499±0.048(7)| 0.578±0.091(1)|
| Scene      | 0.163±0.098(8)| 0.199±0.114(6)| 0.295±0.127(5)| 0.406±0.092(3)| 0.394±0.104(4)| 0.463±0.165(1)| 0.186±0.144(7)| 0.43±0.163(2) |
| Enron      | 0.102±0.035(4)| 0.067±0.016(8)| 0.074±0.017(6.5)| 0.108±0.032(3)| 0.099±0.033(5)| 0.074±0.027(6.5)| 0.122±0.045(2)| 0.141±0.048(1) |
| Arts       | 0.037±0.022(8)| 0.04±0.023(5) | 0.055±0.034(3) | 0.039±0.022(6.5)| 0.043±0.019(4)| 0.039±0.026(6.5)| 0.079±0.038(2)| 0.09±0.033(1)  |
| Education  | 0.034±0.025(6.5)| 0.034±0.026(6.5)| 0.019±0.017(8) | 0.035±0.025(5) | 0.048±0.028(4) | 0.052±0.014(3) | 0.065±0.02(2) | 0.068±0.02(1) |
| Entertain  | 0.072±0.033(8)| 0.078±0.044(6) | 0.097±0.047(3)| 0.077±0.032(7) | 0.079±0.031(5) | 0.081±0.038(4) | 0.116±0.047(2)| 0.13±0.047(1) |
| Reference  | 0.038±0.019(7)| 0.037±0.018(8)| 0.063±0.024(3.5)| 0.041±0.02(6) | 0.051±0.021(5)| 0.063±0.028(3.5)| 0.068±0.026(2)| 0.081±0.021(1) |
| Science    | 0.034±0.022(6.5)| 0.032±0.021(8)| 0.034±0.016(6.5)| 0.038±0.028(4)| 0.039±0.029(3)| 0.036±0.02(5) | 0.07±0.028(1) | 0.069±0.024(2) |
| Society    | 0.048±0.015(6)| 0.047±0.014(7)| 0.055±0.02(2)| 0.049±0.016(4.5)| 0.049±0.015(4)| 0.032±0.012(8)| 0.055±0.019(2.5)| 0.075±0.025(1) |
| Social     | 0.073±0.03(6)| 0.081±0.031(4)| 0.031±0.016(7)| 0.075±0.031(5)| 0.083±0.031(3)| 0.014±0.012(8)| 0.099±0.032(2)| 0.103±0.039(1) |
| Corel5K    | 0.0062±0.0029(3.5)| 0.0075±0.0031(2)| 0.0062±0.0029(3.5)| 0.0039±0.0024(7)| 0.0041±0.0025(6)| 0.0061±0.0033(5)| 0.008±0.0008(8)| 0.0086±0.0028(1) |
| Average(rank)| 0.102(6.136) | 0.104(5.773) | 0.118(4.591) | 0.125(5.182) | 0.127(4.409) | 0.122(5.318) | 0.124(3.409) | 0.161(1.182)  |
Table 3: The results of all methods using SVM classifier in terms of Micro-$F_1$ (mean±std.)

| Data sets   | PPT+MI     | PPT+CHI    | MIFS       | MDMR       | LRFS       | RALM-FS    | SSFS       | RLEFS       |
|-------------|------------|------------|------------|------------|------------|------------|------------|-------------|
| Flags       | 0.665±0.04(4) | 0.663±0.038(5) | 0.722±0.05(1) | 0.649±0.043(7) | 0.666±0.04(3) | 0.631±0.048(8) | 0.652±0.034(6) | 0.719±0.046(2) |
| Scene       | 0.168±0.1(8)  | 0.205±0.117(7)  | 0.325±0.135(5) | 0.429±0.096(3) | 0.415±0.108(4) | 0.476±0.164(1) | 0.214±0.162(6) | 0.443±0.162(2)  |
| Enron       | 0.47±0.043(4) | 0.353±0.019(8)  | 0.372±0.027(7) | 0.474±0.049(3) | 0.446±0.056(5) | 0.389±0.059(6) | 0.505±0.042(2) | 0.525±0.043(1)  |
| Arts        | 0.09±0.053(8) | 0.098±0.055(6)  | 0.139±0.078(3) | 0.095±0.052(7) | 0.104±0.045(4) | 0.102±0.061(5) | 0.188±0.092(2) | 0.21±0.073(1)   |
| Education   | 0.124±0.083(6) | 0.12±0.085(7)   | 0.073±0.059(8) | 0.127±0.081(5) | 0.15±0.081(4)  | 0.193±0.056(3) | 0.25±0.084(1)  | 0.228±0.068(2)  |
| Entertain   | 0.19±0.088(5.5) | 0.183±0.103(8)  | 0.228±0.112(3) | 0.19±0.087(5.5) | 0.184±0.082(7) | 0.214±0.1(4)    | 0.281±0.106(2) | 0.285±0.119(1)  |
| Reference   | 0.348±0.077(8) | 0.355±0.074(7)  | 0.359±0.105(5) | 0.356±0.069(6) | 0.377±0.066(4) | 0.404±0.132(3)  | 0.423±0.068(1) | 0.418±0.108(2)  |
| Science     | 0.115±0.059(6) | 0.11±0.055(7)   | 0.129±0.057(3.5) | 0.127±0.074(5) | 0.129±0.078(3.5) | 0.097±0.054(8) | 0.199±0.069(1) | 0.194±0.071(2)  |
| Society     | 0.316±0.033(5) | 0.317±0.032(4)  | 0.3±0.042(6)   | 0.319±0.017(3) | 0.321±0.016(2) | 0.223±0.059(8) | 0.297±0.067(6) | 0.336±0.067(1)  |
| Social      | 0.47±0.097(3)  | 0.445±0.108(6)  | 0.276±0.136(7) | 0.465±0.106(4,5) | 0.465±0.119(4,5) | 0.149±0.112(8) | 0.538±0.075(1) | 0.498±0.142(2)  |
| Corel5K     | 0.032±0.016(4) | 0.036±0.017(2)  | 0.033±0.016(3) | 0.016±0.011(7) | 0.017±0.011(6) | 0.03±0.016(5) | 0.009±0.01(8)  | 0.052±0.016(1)  |
| Average(rank)| 0.272(5.591)  | 0.262(6.091)   | 0.269(4.682)  | 0.295(5.091) | 0.298(4.273) | 0.264(5.364) | 0.323(3.364) | 0.355(1.545)    |
Table 4 The results of all methods using 3NN classifier in terms of Macro-$F_1$ (mean±std.)

| Data sets | PPT+MI   | PPT+CHI   | MIIF  | MDMR   | LRFS    | RALM-FS  | SSFS    | RLEFS   |
|-----------|----------|-----------|-------|--------|---------|----------|---------|---------|
| Flags     | 0.448±0.011(6) | 0.449±0.011(5) | 0.527±0.078(2) | 0.443±0.018(7) | 0.453±0.022(3) | 0.439±0.011(8) | 0.45±0.02(4) | 0.538±0.084(1) |
| Scene     | 0.348±0.069(8) | 0.368±0.072(7) | 0.498±0.086(5) | 0.555±0.075(2) | 0.547±0.082(3) | 0.581±0.108(1) | 0.382±0.113(6) | 0.541±0.117(4) |
| Enron     | 0.116±0.019(3) | 0.074±0.012(8) | 0.087±0.014(6) | 0.115±0.02(4)  | 0.109±0.022(5) | 0.081±0.026(7) | 0.122±0.02(2)  | 0.126±0.018(1) |
| Arts      | 0.088±0.022(7) | 0.091±0.022(6) | 0.095±0.033(4.5)| 0.095±0.025(4.5)| 0.101±0.025(3) | 0.071±0.025(8) | 0.119±0.035(1) | 0.114±0.032(2) |
| Education | 0.082±0.019(6.5)| 0.083±0.022(5) | 0.043±0.018(8) | 0.082±0.019(6.5)| 0.087±0.02(3)  | 0.084±0.023(4) | 0.096±0.02(1)  | 0.089±0.021(2) |
| Entertain | 0.147±0.019(5) | 0.151±0.022(3) | 0.138±0.042(7) | 0.148±0.018(4) | 0.145±0.02(6)  | 0.131±0.048(1) | 0.163±0.035(2) | 0.172±0.043(1) |
| Reference | 0.079±0.017(6) | 0.076±0.016(8) | 0.088±0.024(3) | 0.081±0.017(5) | 0.085±0.017(4) | 0.077±0.026(7) | 0.096±0.022(2) | 0.1±0.02(1)    |
| Science   | 0.065±0.023(6) | 0.066±0.022(5) | 0.062±0.015(7) | 0.067±0.024(3.5)| 0.067±0.023(3.5)| 0.057±0.021(8) | 0.093±0.028(2) | 0.099±0.024(1) |
| Society   | 0.083±0.015(6.5)| 0.083±0.016(6.5)| 0.089±0.021(2) | 0.084±0.014(5) | 0.085±0.013(3.5)| 0.053±0.016(8) | 0.085±0.018(3.5)| 0.095±0.022(1) |
| Social    | 0.099±0.034(6) | 0.107±0.03(4.5) | 0.051±0.017(7) | 0.107±0.032(4.5)| 0.111±0.027(3) | 0.038±0.012(8) | 0.12±0.035(2)  | 0.123±0.042(1) |
| Corel5K   | 0.008±0.003(7) | 0.01±0.002(4.5) | 0.01±0.003(4.5) | 0.011±0.004(2.5)| 0.011±0.004(2.5)| 0.009±0.003(6) | 0.007±0.003(8) | 0.012±0.004(1) |
| Average(rank) | 0.142(6.091) | 0.142(5.682) | 0.153(5.091) | 0.163(4.409) | 0.164(3.591) | 0.147(6.636) | 0.158(3.045) | 0.183(1.455) |
Table 5  The results of all methods using 3NN classifier in terms of Micro-$F_1$ (mean±std.)

| Data sets | PPT+MI     | PPT+CHI     | MIFS        | MDMR        | LRFS        | RALM-FS     | SSFS        | RLEFS       |
|-----------|------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Flags     | 0.615±0.011(4.5) | 0.615±0.011(4.5) | 0.662±0.03(2) | 0.611±0.017(7) | 0.616±0.018(3) | 0.606±0.012(8) | 0.613±0.014(6) | 0.672±0.036(1) |
| Scene     | 0.348±0.067(8) | 0.367±0.071(7) | 0.493±0.086(5) | 0.556±0.075(2) | 0.547±0.082(3) | 0.577±0.105(1) | 0.389±0.117(6) | 0.541±0.116(4) |
| Enron     | 0.454±0.013(3) | 0.345±0.026(8) | 0.41±0.024(6) | 0.443±0.044(4) | 0.419±0.048(5) | 0.365±0.073(7) | 0.464±0.015(2) | 0.474±0.041(1) |
| Arts      | 0.181±0.037(8) | 0.187±0.033(6) | 0.202±0.052(3) | 0.189±0.034(5) | 0.196±0.027(4) | 0.182±0.043(7) | 0.23±0.049(2) | 0.251±0.037(1) |
| Education | 0.23±0.043(6) | 0.227±0.043(7) | 0.183±0.055(8) | 0.231±0.049(5) | 0.243±0.046(4) | 0.26±0.05(3) | 0.286±0.043(1) | 0.271±0.054(2) |
| Entertain | 0.284±0.034(3) | 0.285±0.033(3) | 0.276±0.065(7) | 0.283±0.03(5) | 0.281±0.034(6) | 0.273±0.065(8) | 0.317±0.056(2) | 0.329±0.074(1) |
| Reference | 0.373±0.044(7) | 0.366±0.048(4) | 0.382±0.055(5) | 0.378±0.044(6) | 0.386±0.046(3.5) | 0.386±0.089(3.5) | 0.404±0.034(2) | 0.427±0.059(1) |
| Science   | 0.162±0.028(6.5) | 0.162±0.026(6.5) | 0.171±0.037(5) | 0.174±0.034(4) | 0.177±0.035(3) | 0.16±0.048(8) | 0.224±0.042(2) | 0.228±0.04(1) |
| Society   | 0.317±0.023(6.5) | 0.317±0.026(2.5) | 0.305±0.043(7) | 0.312±0.032(5) | 0.315±0.028(4) | 0.255±0.05(8) | 0.311±0.032(6) | 0.328±0.041(1) |
| Social    | 0.451±0.052(4.5) | 0.442±0.055(6)  | 0.338±0.081(7) | 0.451±0.058(4.5) | 0.455±0.054(3) | 0.315±0.054(8) | 0.495±0.028(1) | 0.491±0.081(2) |
| Corel5K   | 0.073±0.02(6)  | 0.07±0.026(7)  | 0.099±0.005(2) | 0.085±0.024(4.5) | 0.085±0.023(4.5) | 0.086±0.028(3) | 0.063±0.018(8) | 0.112±0.015(1) |
| Average(rank) | 0.317(5.455) | 0.308(5.955) | 0.32(5.182) | 0.338(4.727) | 0.338(3.909) | 0.315(5.864) | 0.345(3.455) | 0.375(1.455) |
Table 6  The results of all methods using ML-kNN classifier in terms of Hamming Loss (mean±std.)

| Data sets  | PPT+MI       | PPT+CHI      | MIFS         | MDMR         | LRFS         | RALM-FS       | SSFS         | RLEFS         |
|------------|--------------|--------------|--------------|--------------|--------------|---------------|--------------|---------------|
| Flags      | 0.3171±0.0082(5) | 0.3184±0.0083(6) | 0.292±0.0205(2) | 0.3146±0.0111(3) | 0.3169±0.0087(4) | 0.3239±0.0087(8) | 0.3191±0.0065(7) | 0.2888±0.0202(1) |
| Scene      | 0.178±0.0105(8)  | 0.1708±0.0126(7)  | 0.1507±0.0173(5) | 0.1354±0.0144(3) | 0.1362±0.0164(4) | 0.1267±0.0225(1) | 0.1658±0.0192(6) | 0.1328±0.023(2)  |
| Enron      | 0.0528±0.0015(3) | 0.0589±0.00058(8) | 0.0974±0.0012(7) | 0.0531±0.0027(4) | 0.055±0.003(5) | 0.057±0.0023(6) | 0.0515±0.0015(2) | 0.0509±0.0019(1) |
| Arts       | 0.0647±0.0006(7) | 0.0646±0.0009(5)  | 0.0632±0.0012(3) | 0.0647±0.0007(7) | 0.0647±0.0009(7) | 0.0633±0.0008(4) | 0.0629±0.0015(2) | 0.0602±0.0011(1) |
| Education  | 0.0448±0.0013(5) | 0.0446±0.0014(3)  | 0.0449±0.0007(6.5)| 0.0449±0.0011(4) | 0.0447±0.0011(4) | 0.0433±0.0005(2) | 0.0632±0.0019(8) | 0.0422±0.0009(1) |
| Entertain  | 0.0654±0.0014(5) | 0.0659±0.0013(8)  | 0.0653±0.0018(4) | 0.0655±0.0016(6) | 0.0657±0.0015(7) | 0.0638±0.0021(3) | 0.0631±0.0018(2) | 0.0622±0.0028(1) |
| Reference  | 0.0325±0.0015(5) | 0.0329±0.0015(8)  | 0.0313±0.0015(3) | 0.0325±0.0014(5) | 0.0326±0.0013(7) | 0.0325±0.0013(5) | 0.0301±0.0008(1) | 0.0304±0.0017(2) |
| Science    | 0.0362±0.0004(5.5) | 0.0364±0.0003(8)  | 0.036±0.0006(3)  | 0.0363±0.0005(7) | 0.0362±0.0005(5.5) | 0.0361±0.0004(4) | 0.0354±0.0004(2) | 0.0349±0.0006(1) |
| Society    | 0.0589±0.0009(7)  | 0.0588±0.0009(6)  | 0.0582±0.0012(3) | 0.0585±0.0011(5) | 0.0584±0.0014(4) | 0.0593±0.0006(8) | 0.0577±0.0012(2) | 0.0575±0.0016(1) |
| Social     | 0.0273±0.0015(3)  | 0.0283±0.0018(6)  | 0.0311±0.0016(7) | 0.0275±0.0018(4.5)| 0.0275±0.0024(4.5)| 0.0324±0.0018(1) | 0.0249±0.0012(1) | 0.025±0.0024(2)  |
| Corel5K    | 0.00958±0.00007(5)| 0.00956±0.00006(4)| 0.0095±0.00003(2)| 0.0096±0.00009(7)| 0.0096±0.00009(7)| 0.00949±0.00003(1)| 0.0096±0.00009(7)| 0.00954±0.00006(3) |
| Average    | 0.0807(5.3182)   | 0.0808(6.2727)   | 0.0763(4.1364)   | 0.0766(5.2727)   | 0.077(5.3636)    | 0.0771(4.5455)    | 0.0803(3.6364)   | 0.0722(1.4545)   |
Table 7 The Friedman statistics $F_F$ and critical values in terms of each evaluation criterion

| Metric                      | $F_F$    | Critical value ($\alpha = 0.05$) |
|-----------------------------|----------|----------------------------------|
| Macro-$F_1$ (SVM)           | 7.2476   | 2.143                            |
| Micro-$F_1$ (SVM)           | 5.5105   |                                   |
| Macro-$F_1$ (3NN)           | 10.0632  |                                   |
| Micro-$F_1$ (3NN)           | 6.2286   |                                   |
| HL (ML-kNN)                 | 5.7362   |                                   |

Fig. 1 Comparison of the proposed method against other compared methods with the Bonferroni-Dunn test (Significance level $\alpha = 0.05$)

PPT+CHI, MIFS, MDMR, LRFS, RALM-FS and SSFS on nine data sets except the data sets Scene and Science. Moreover, RLEFS improves the classification performance by 57.84%, 54.81%, 36.44%, 28.8%, 26.77%, 31.97% and 29.84% in comparison with other compared methods. In other tables, we can observe similar situations. Besides, RLEFS also obtains the sub-optimal results on several data sets (such as Scene and Science) according to Tables 2, 3, 4, 5 and 6. To sum up, these results show that RLEFS method achieves the best classification performance compared to other methods on most data sets under all the evaluation criteria.

To further verify the effectiveness of RLEFS, the statistical tests, i.e., the Friedman test and its post hoc test are added to analyze the proposed and all compared methods. Friedman statistic $F_F$ and critical values of all evaluation criteria are shown in Table 7. As can be seen, the null hypothesis will be rejected, when we set the significance level $\alpha$ as 0.05, which implies all used methods have the same performance if the null hypothesis holds. As a result, the Bonferroni–Dunn test is used as the post hoc test due to observed results. The significant difference between RLEFS and other compared methods is estimated by computing whether the average rank of the two methods is at the critical distance (CD), where $CD = q_\alpha \sqrt{k(k+1)/6N}$, $q_\alpha = 2.69$, $\alpha = 0.05$, $k = 8$ ($k$ denotes the number of algorithms), $N = 11$ ($N$ denotes the number of data sets), and thus, $CD = 2.8096$. Figure 1 shows that CD diagrams on all used evaluation metrics. In these CD diagrams, we use a red line to denote the distance between RLEFS method and other compared methods within
As we can see in these CD diagrams, RLEFS and other methods have significant differences. At the same time, RLEFS has a competitive advantage over other methods. Moreover, the proposed method ranks first under all evaluation metrics.

Besides, we present the experimental results on six representative data sets, including Arts, Education, Enron, Flags, Reference and Science in Figs. 2, 3, 4, 5 and 6. 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 the features, where the step size is set to 1%. As shown in Figs. 2, 3, 4, 5 and 6, we observe that RLEFS achieves the best classification performance.
In most cases, the classification performance of RLEFS increases at first and then stabilizes according to Figs. 2, 3, 4 and 5. The classification performance of RLEFS decreases at first and then stabilizes according to Fig. 6. Overall, the proposed RLEFS method outperforms other seven compared methods in the experiments.

5.4 Sensitivity analysis of parameters

We present and discuss the sensitivity of parameters regarding the proposed method RLEFS, where these parameters contain $\alpha$, $\beta$, $\gamma$ and $\lambda$. In this section, we take data set Arts as an
example. First, these parameters are tuned in \{0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}. One parameter is adjusted while the other three parameters are fixed, where we set the fixed parameters as 0.5 in this paper. We show the analysis results on SVM classifier. In Fig. 7a–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. Therefore, a larger and finer range should be used in practical applications.
Next, we display the execution time of RLEFS and other compared methods on all data sets. The execution time is recorded in Table 8. By analyzing the execution time, RLEFS costs the least execution time on the most data sets. Afterward, RALM-FS, PPT+MI and SSFS obtain similar execution time. And then, the running time of PPT+CHI approaches to that of MIFS. MDMR and LRFS cost more running time than the other embedded methods due to interactive correlations among features and each label. Overall, RLEFS has acceptable execution time and classification performance compared to other methods. Furthermore, we
make an analysis of time complexity regarding the proposed method and other algorithm adaption-based multi-label feature selection methods. Suppose the data set has \( n \) instances with \( q \) features, and it contains \( l \) labels. The number of already-selected features is recorded as \( k \). The time complexity of MDMR and LRFS methods is \( O(k(n - k)) \) and \( O(ql^2 + kq) \), respectively. MIFS requires \( O(cnq + n^2) \) in each iteration. The time complexity of RALM-FS is \( O(q^3) \) when an inverse matrix \( (\mathbb{R}^{q \times q}) \) is calculated. The time complexity of the proposed method RLEFS is \( O(qn^2 + nq^2) \) in each iteration.

**Fig. 6** Eight methods on six data sets using ML-kNN classifier in terms of Hamming Loss

![Images showing Hamming Loss for different datasets](image-url)
### Table 8 The execution time (secs) of all different methods

| Data sets  | PPT+MI | PPT+CHI | MIFS  | MDMR  | LRFS  | RALM-FS | SSFS  | RLEFS |
|------------|--------|---------|-------|-------|-------|---------|-------|-------|
| Flags      | 0.0020 | 0.2075  | 37.7202 | 0.5027 | 0.5316 | 0.0878  | 0.1855 | 0.0070 |
| Scene      | 0.7799 | 659.087 | 21.0767 | 524.0462 | 346.0897 | 0.9016  | 1.9660 | 0.4009 |
| Enron      | 0.7879 | 13.1508 | 155.9359 | 21074.2350 | 18092.3621 | 1.3050  | 6.5934 | 40.6992 |
| Arts       | 1.0381 | 91.4707 | 32.4866 | 5234.9059 | 4180.9622 | 1.8629  | 3.5235 | 0.2354 |
| Education  | 1.2427 | 93.0931 | 35.6676 | 9313.5424 | 7306.1785 | 1.8311  | 6.5799 | 0.3341 |
| Entertain  | 1.5170 | 113.5574 | 72.6079 | 7986.8245 | 5440.0593 | 1.8870  | 11.1263 | 0.3920 |
| Reference  | 1.9258 | 100.1712 | 41.6147 | 18737.2117 | 13960.6576 | 1.9428  | 6.5246 | 0.5859 |
| Science    | 1.9797 | 196.2386 | 20.2100 | 20579.7952 | 16294.1888 | 1.9518  | 5.9112 | 0.7670 |
| Society    | 1.3683 | 140.5333 | 55.1017 | 10269.9873 | 7874.2184 | 1.9897  | 5.9670 | 0.4528 |
| Social     | 2.7656 | 151.7017 | 48.0066 | 38600.2489 | 27130.5610 | 2.1822  | 7.1180 | 1.4790 |
| Corel5K    | 3.6242 | 196.9421 | 75.5486 | 45861.4514 | 38546.5373 | 2.4514  | 6.1546 | 1.8936 |
Fig. 7 Sensitivity analysis of parameters on the data set Arts on SVM classifier in terms of Micro-average and Macro-average

6 Conclusions

The most existing feature selection methods adopt the following scheme: extracting logical label correlations and maintaining the consistency between the features and labels. However, the label information is described by logical value that cannot reflect the importance of label. In this paper, we propose a joint multi-label feature selection framework RLEFS. RLEFS has the following several appealing characteristics. First, RLEFS transforms the logical labels to numerical labels and then imposes $l_{2,1}$-norm onto the new numerical label matrix, i.e., we extract numerical label information by label correlations between the feature space and the reconstructed label space with three $l_{2,1}$-norm constraints to guide feature selection. Afterward, the proposed method shares the new numerical label matrix between the feature set and the label set. Furthermore, RLEFS utilizes the graph Laplacian matrix to ensure the consistency of label information. To verify the effectiveness of the proposed method, we conduct numerous experiments on multiple multi-label data sets in comparison with the multiple state-of-the-art multi-label feature selection methods. By analyzing these experimental results, we conclude that the proposed method RLEFS outperforms other compared methods in terms of different evaluation criteria, i.e., Micro-$F_1$, Macro-$F_1$ and Hamming Loss. In other word, the proposed method RLEFS can make full use of effective label information to select the most discriminative features. In addition, RLEFS has some limitations. For instance, RLEFS involves the fixed Laplace graph, which is not suitable for joint multi-label learning and feature selection process. How to effectively combine the advantages of joint Laplace graph learning and label enhancement learning is an open problem.

In future work, multi-label feature selection is still our research focus. However, we will furthermore study feature selection under non-convex optimization under causal mechanisms due to the existence of interpretability problems. The causal sequence relationship can explain the internal mechanism of function from a deeper level, which helps people understand the black box problem.

Acknowledgements 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.

**Author Contributions**  YL mainly collected and processed data. He proposed algorithm framework. WG and JH conceived and designed the analysis, and then design contrast experiment. YL prepared the figures and experimental data. YL and JH wrote the main manuscript text. WG completed the audit and proofreading. JH is responsible for the delivery of final manuscript.

**Declaration**

**Competing interests**  The authors declare no competing interests.

**References**

1. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H (2018) Feature selection: A data perspective. ACM Comput Surv 50(6):94. https://doi.org/10.1145/3136625
2. Zhang P, Liu G, Gao W (2019) Distinguishing two types of labels for multi-label feature selection. Pattern Recogn 95:72–82. https://doi.org/10.1016/j.patcog.2019.06.004
3. Huang J, Li G, Huang Q, Wu X (2016) Learning label-specific features and class-dependent labels for multi-label classification. IEEE Trans Knowl Data Eng 28(12):3309–3323. https://doi.org/10.1109/TKDE.2016.2608339
4. Zhang J, Luo Z, Li C, Zhou C, Li S (2019) Manifold regularized discriminative feature selection for multi-label learning. Pattern Recogn 95:136–150. https://doi.org/10.1016/j.patcog.2019.06.003
5. Zhu R, Dornaika F, Ruichek Y (2019) Learning a discriminant graph-based embedding with feature selection for image categorization. Neural Netw 111:35–46. https://doi.org/10.1016/j.neunet.2018.12.008
6. Bolón-Canedo V, Sánchez-Marono N, Alonso-Betanzos A (2016) Feature selection for high-dimensional data. Progr Artif Intell 5(2):65–75. https://doi.org/10.1007/s10287-008-0070-7
7. Ditzler G, Polikar R, Rosen G (2017) A sequential learning approach for scaling up filter-based feature subset selection. IEEE Trans Neural Netw Learn Syst 29(6):2530–2544. https://doi.org/10.1109/TNNLS.2017.2697407
8. Wang A, An N, Chen G, Li L, Alterovitz G (2015) Accelerating wrapper-based feature selection with k-nearest-neighbor. Knowl-Based Syst 83:81–91. https://doi.org/10.1016/j.knosys.2015.03.009
9. Hu L, Li Y, Gao W, Zhang P, Hu J (2020) Multi-label feature selection with shared common mode. Pattern Recogn 104:107344. https://doi.org/10.1016/j.patcog.2020.107344
10. Yuan H, Li J, Lai LL, Tang YY (2019) Joint sparse matrix regression and nonnegative spectral analysis for two-dimensional unsupervised feature selection. Pattern Recogn 89:119–133. https://doi.org/10.1016/j.patcog.2019.01.014
11. Zhang R, Nie F, Li X (2018) Self-weighted supervised discriminative feature selection. IEEE Trans Neural Netw Learn Syst 29(8):3913–3918. https://doi.org/10.1109/TNNLS.2017.2740341
12. Huang J, Li G, Huang Q, Wu X (2018) Joint feature selection and classification for multilabel learning. IEEE Trans Cybern 48(3):876–889. https://doi.org/10.1109/TCYB.2017.2663838
13. Jian L, Li J, Shu K, Liu H (2016) Multi-label informed feature selection. In: IJCAI, pp. 1627–1633. https://doi.org/10.5555/3060832.3060848
14. Cai Z, Zhu W (2018) Multi-label feature selection via feature manifold learning and sparsity regularization. Int J Mach Learn Cybern 9(8):1321–1334. https://doi.org/10.1007/s13042-017-0647-y
15. Xu Y, Wang J, An S, Wei J, Ruan J (2018) Semi-supervised multi-label feature selection by preserving feature-label space consistency. In: Proceedings of the 27th ACM international conference on information and knowledge management, pp 783–792. https://doi.org/10.1145/3269206.3271760
16. Saul LK, Roweis ST (2003) Think globally, fit locally: unsupervised learning of low dimensional manifolds. J Mach Learn Res 4(Jun):119–155. https://doi.org/10.1162/153244303322972667
17. Zhu X, Li X, Zhang S, Xu Z, Yu L, Wang C (2017) Graph pca hashing for similarity search. IEEE Trans Multimedia 19(9):2033–2044. https://doi.org/10.1109/TMM.2017.2703636
18. Zhu Y, Kwok JT, Zhou Z-H (2017) Multi-label learning with global and local label correlation. IEEE Trans Knowl Data Eng 30(6):1081–1094. https://doi.org/10.1109/TKDE.2017.2785795
19. Chen G, Song Y, Wang F, Zhang C (2008) Semi-supervised multi-label learning by solving a sylyvester equation. In: Proceedings of the 2008 SIAM international conference on data mining, pp 410–419. https://doi.org/10.1137/1.9781611972788.37
20. Nie F, Huang H, Cai X, Ding CH (2010) Efficient and robust feature selection via joint l2, l-norms minimization. In: Advances in neural information processing systems, pp 1813–1821. https://doi.org/10.5555/2997046.2997098
21. Hou P, Geng X, Zhang M-L (2016) Multi-label manifold learning. In: Thirtieth AAAI conference on artificial intelligence, pp 1680–1686. https://doi.org/10.5555/3016100.3016134
22. Shao R, Xu N, Geng X (2018) Multi-label learning with label enhancement. In: 2018 IEEE international conference on data mining (ICDM), pp 437–446. https://doi.org/10.1109/ICDM.2018.00059
23. Zhang M-L, Zhou Z-H (2014) A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng 26(8):1819–1837. https://doi.org/10.1109/TKDE.2013.39
24. Boutilier MR, Luo J, Shen X, Brown CM (2004) Learning multi-label scene classification. Pattern Recogn 37(9):1757–1771. https://doi.org/10.1016/j.patcog.2004.03.009
25. Zhang J, Lin Y, Jiang M, Li S, Tang Y, Tan KC (2020) Multi-label feature selection via global relevance and redundancy optimization. In: Proceedings of 29th international joint conference artificial intelligence, pp 2512–2518. https://doi.org/10.24963/ijcai.2020/348
26. Huang R, Jiang W, Sun G (2018) Manifold-based constraint laplacian score for multi-label feature selection. Pattern Recogn Lett 112:346–352. https://doi.org/10.1016/j.patrec.2018.08.021
27. Read J (2008) A pruned problem transformation method for multi-label classification. In: Proceedings of the 2008 New Zealand computer science research student conference (NZCSRS 2008), vol. 143150, p 41
28. Doquire G, Verleysen M (2011) Feature selection for multi-label classification problems. In: International work-conference on artificial neural networks, pp 9–16. https://doi.org/10.1007/978-3-642-21501-8_2
29. Cai X, Nie F, Huang H (2013) Exact top-k feature selection via l2, l-norms constraint. In: Twenty-third international joint conference on artificial intelligence, pp 1240–1246. https://doi.org/10.5555/2540128.2540307
30. Lin Y, Hu Q, Liu J, Duan J (2015) Multi-label feature selection based on max-dependency and min-redundancy. Neurocomputing 168:92–103. https://doi.org/10.1016/j.neucom.2015.06.010
31. Braytee A, Liu W, Catchpoole DR, Kennedy PJ (2017) Multi-label feature selection using correlation information. In: Proceedings of the 2017 ACM on conference on information and knowledge management, pp 1649–1656. https://doi.org/10.1145/3132847.3132858
32. Gao W, Li Y, Hu L (2021) Multilabel feature selection with constrained latent structure shared term. IEEE Trans Neural Netw Learn Syst 1:1–10. https://doi.org/10.1109/TNNLS.2021.3105142
33. Von Luxburg U (2007) A tutorial on spectral clustering. Stat Comput 17(4):395–416. https://doi.org/10.1007/s11222-007-9033-z
34. Hashemi A, Dowlatshahi MB, Nezamabadi-pour H (2021) An efficient pareto-based feature selection algorithm for multi-label classification. Inf Sci 581:428–447. https://doi.org/10.1016/j.ins.2021.09.052
35. Paniri M, Dowlatshahi MB, Nezamabadi-pour H (2021) Ant-td: Ant colony optimization plus temporal difference reinforcement learning for multi-label feature selection. Swarm Evol Comput 64:100892. https://doi.org/10.1016/j.swevo.2021.100892
36. Hashemi A, Dowlatshahi MB, Nezamabadi-pour H (2021) A bipartite matching-based feature selection for multi-label learning. Int J Mach Learn Cybern 12(2):459–475. https://doi.org/10.1007/j.10667-022-109085
37. Hashemi A, Dowlatshahi MB, Nezamabadi-pour H (2021) A bipartite matching-based feature selection for multi-label learning. Int J Mach Learn Cybern 12(2):459–475. https://doi.org/10.1007/j.10667-022-109085
43. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc: Ser B (Methodol) 39(1):1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x
44. Cai D, He X, Han J, Huang TS (2010) Graph regularized nonnegative matrix factorization for data representation. IEEE Trans Pattern Anal Mach Intell 33(8):1548–1560. https://doi.org/10.1109/TPAMI.2010.231
45. Tsoumakas G, Spyromitros-Xioufis E, Vilcek J, Vlahavas I (2011) Mulan: A java library for multi-label learning. J Mach Learn Res 12(7):2411–2414
46. Huang K-H, Lin H-T (2017) Cost-sensitive label embedding for multi-label classification. Mach Learn 106(9–10):1725–1746. https://doi.org/10.1007/s10994-017-5659-z
47. Zhang Y, Miao D, Pedrycz W, Zhao T, Xu J, Yu Y (2020) Granular structure-based incremental updating for multi-label classification. Knowl-Based Syst 189:105066. https://doi.org/10.1016/j.knosys.2019.105066
48. Jeong J-Y, Kang J-S, Jun C-H (2020) Regularization-based model tree for multi-output regression. Inf Sci 507:240–255. https://doi.org/10.1016/j.ins.2019.08.034
49. Klimt B, Yang Y (2004) The enron corpus: A new dataset for email classification research. In: European conference on machine learning, pp 217–226. https://doi.org/10.1007/978-3-540-30115-8_22
50. Zhang M-L, Zhou Z-H (2007) Ml-knn: A lazy learning approach to multi-label learning. Pattern Recogn 40(7):2038–2048. https://doi.org/10.1016/j.patcog.2006.12.019
51. Yu K, Yu S, Tresp V (2005) Multi-label informed latent semantic indexing. In: Proceedings of the 28th Annual International ACM SIGIR conference on research and development in information retrieval, pp 258–265. https://doi.org/10.1145/1076034.1076080

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Yonghao Li received his B.E degree from Henan Agricultural University in 2017. Now, he is working toward Ph.D. degree in College of Computer Science, Jilin University, China. His research focuses on feature selection.
Juncheng Hu received the B.S. and Ph.D. degrees from the Jilin University in 2017 and 2022. He is currently a lecturer at the College of Computer Science and Technology, Jilin University, China. His research interests include data mining, machine learning and parallel computing.

Wanfu Gao received his M.S. and Ph.D. degrees in the College of Computer Science from Jilin University in 2016 and 2019. He is doing post-doctoral research in the College of Chemistry in Jilin University. His research interests include feature selection and information theory.