Complex & Intelligent Systems
https://doi.org/10.1007/s40747-024-01439-7

ORIGINAL ARTICLE

Sparse semi-supervised multi-label feature selection based on latent representation

Xue Zhao[1] · Qiaoyan Li[1] · Zhiwei Xing[1] · Xiaofei Yang[1] · Xuezhen Dai[2]

Received: 4 August 2023 / Accepted: 23 March 2024

Abstract

With the rapid development of the Internet, there are a large number of high-dimensional multi-label data to be processed in real life. To save resources and time, semi-supervised multi-label feature selection, as a dimension reduction method, has been widely used in many machine learning and data mining. In this paper, we design a new semi-supervised multi-label feature selection algorithm. First, we construct an initial similarity matrix with supervised information by considering the similarity between labels, so as to learn a more ideal similarity matrix, which can better guide feature selection. By combining latent representation with semi-supervised information, a more ideal pseudo-label matrix is learned. Second, the local manifold structure of the original data space is preserved by the manifold regularization term based on the graph. Finally, an effective alternating iterative updating algorithm is applied to optimize the proposed model, and the experimental results on several datasets prove the effectiveness of the approach.

Keywords Feature selection · Multi-label learning · Manifold regularization · Latent representation · Similarity matrix

Introduction

With the advent of the information explosion era, a large amount of multi-label data are widely used in many machine learning tasks. For example, text classification [1, 2], bioinformatics [3], gene expression, speech recognition, image recognition [4], etc. Depending on the data, each instance can be assigned to several different categories simultaneously. For example, in text classification, each document may belong to multiple themes, such as society, economy, culture, and even politics; in image recognition, an image can always be labeled with multiple labels at the same time, such as blue sky, white clouds, big trees, rivers, etc; a gene is associated with several functional categories, such as "metabolism", "energy", "cellular biogenesis", etc. To deal with these multi-labeled data, multi-label learning [5–9] has emerged and received increasing attention from scholars.

Due to the rapid development of the Internet, most of the real-life data are high-dimensional, and processing such high-dimensional data directly is both time-consuming and computationally unreliable. Therefore, researchers extract a small amount of features by reducing dimensionality to remove irrelevant, redundant, and noisy information, to alleviate the computational burden brought by high-dimensional information. There are two traditional dimensionality reduction methods: feature extraction [10, 11] and feature selection [12–14]. Feature extraction is to project features into a new low dimensional space, while feature selection is to select a subset of features to minimize redundancy and maximize the correlation with the target. Since feature selection can maintain the original features, but only remove some features that are considered redundant, it has better readability and interpretability, so multi-label feature selection has attracted extensive attention of scholars.

✉ Qiaoyan Li
liqiaoyan@xpu.edu.cn

Xue Zhao
872421556@qq.com

Zhiwei Xing
zwxing@xpu.edu.cn

Xiaofei Yang
yangxiaofei2002@163.com

Xuezhen Dai
420073948@qq.com

[1] School of Science, Xi'an Polytechnic University, 19 Jinhua South Road, Xi'an 710048, China

[2] The Public Sector, Xi'an Traffic Engineering Institute, No. 1, Meibei West Road, Huyi District, Xi'an 710300, China

Published online: 17 April 2024

However, in practical applications, obtaining data labels is expensive and time-consuming [15]. For example, in text classification, we must traverse the entire list of multiple labels to find a set of all possible labels in an article. Manually labeling each article with all labels requires a significant amount of time, effort, and resources. Therefore, more and more scholars want to use a large amount of unlabeled data and limited labeled data to improve the learning performance. Semi-supervised learning methods have been widely used in single-label learning environments to effectively use unlabeled data to improve generalization performance [16–19] with good results. Therefore, semi-supervised learning will also be an appropriate strategy to address the lack of labeled data in multi-label feature selection. In practical applications, some scholars have borrowed the advantages of this method and applied it to practice, achieving good results. [20, 21].

The existing semi-supervised learning mainly learns in two ways: one is to use label propagation to assist the supervision method [22, 23]. Another approach is to use label information as a constraint in unsupervised methods [24, 25]. The semi-supervised multi-label feature selection algorithm considers the relationship between samples and labels and between labels on the basis of semi-supervised learning. Guo et al. [26] proposed a semi-supervised multi-label feature learning method based on label extension discriminant analysis by extending single-label propagation to multi-label, taking into account label correlation between multiple labels. Xu et al. [27] proposed a semi-supervised multi-label feature selection that maintains the consistency of feature label space by combining feature selection with semi-supervised multi-label learning. This method captures reliable and discriminative local information in the projected feature space by constructing an improved similarity matrix, and uses this information to optimize the correlation in the predicted label space, thereby ensuring the consistency of the feature label space. Zhang et al. [28] proposed a semi-supervised multi-label feature selection (SMLFS) that preserves local logistic information. This method combines the logistic regression model with graph regularization and sparse regularization to form a joint framework for semi-supervised multi-label feature selection; Lv et al. [29] integrated manifold learning and adaptive global structure learning into the semi-supervised feature selection framework and proposed a semi-supervised multi-label feature selection (SFAM) based on adaptive structure learning and manifold learning. For label similarity and data point similarity, semi-supervised learning and label space representation learning using two different graphs at the same time have received some attention [30]. Inspired by this, Kraus et al. [31] proposed a semi-supervised multi-label regression based on Laplace operator by combining the least squares term, semi-supervised regularization term and multi-label regularization extension, and explored the similarity of multi-label data on the basis of semi-supervised framework through graph Laplace matrix.

Although most semi-supervised multi-label feature selection uses graphs to preserve the local structure of data, affinity graphs in previous methods are only used to preserve the local geometric structure, and the potential information implied in graphs has not been fully mined and utilized. To solve this problem, this paper proposes a **s**parse **s**emi-supervised multi-label feature selection based on **l**atent **r**epresentation (SSLR).

First, supervisory information is added in building the initial similarity matrix to make the built initial similarity matrix more desirable. Then, the similarity matrix S is learned by the established initial similarity matrix $\tilde{A}$. Since there is a distance between the learned similarity matrix S and the given initial similarity matrix $\tilde{A}$, learn the optimal similarity matrix through the Frobenius norm, and mine the potential information implied in the optimal similarity graph, decompose it into the product of the pseudo-label matrix F and its transposed F^T. And it can complete the update of the dynamic manifold diagram S. And to ensure the accuracy of known labels, the constraint $F_l = Y_l$ is added in the text. Second, considering that S and W should have similar manifold structures, dynamic manifold diagrams are used to constrain W's manifold structure. A new semi-supervised multi-label feature selection method is proposed by combining the sparse regular term, latent representation, and dynamic graph to constrain the learning of W. Finally, an alternating iterative optimization algorithm was used to solve the target problem, and the effectiveness of the proposed algorithm was verified on multiple datasets. Figure 1 is the framework diagram of this article. The main contributions of this article are as follows:

(1) The use of latent correlation information as supervisory information has not yet been used in multi-label feature selection. Existing methods only learn latent information directly from samples or label matrices, which is insufficient. This cannot excavate the deep latent correlation of samples. The latent information mined from similar matrices in the article includes both the grassroots latent information of samples and the deep latent correlation information between samples, more conducive to guiding model learning.

(2) Embedding latent representation learning into a semi-supervised feature selection task exploits the correlation between data samples to decompose the similarity matrix into a product of the pseudo-label matrix and its transpose.

(3) Use dynamic manifold diagrams to constrain the manifold structure of W.

(4) Make full use of the supervisory information. First, the supervisory information is added to the creation of the initial similarity matrix. Second, to ensure the accuracy of the known labels, the known part of the pseudo label is the same as the marked part of the real label.

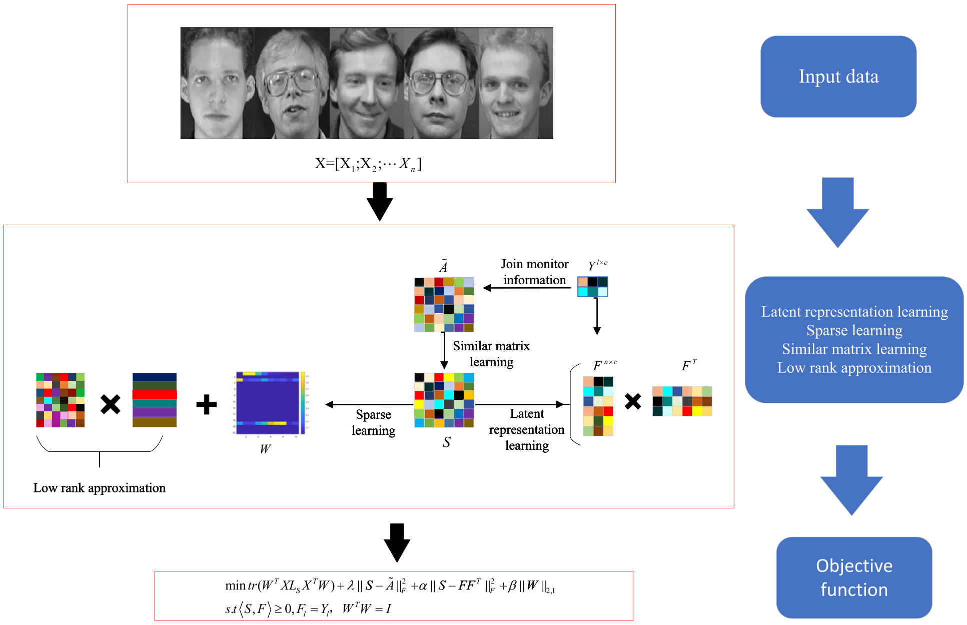


Fig. 1 Framework of this article

(5) A convergent alternating algorithm was designed to optimize the proposed model, and experiments were conducted on various datasets to verify the superiority of the proposed method compared to other state-of-the-art methods.

The rest of the paper is organized as follows. The first section describes the related work, the second section describes the model building process, the third section describes the model solving methods and algorithms, and the fourth section analyzes the experimental results and compares them with seven other algorithms in order to demonstrate the advantages of the algorithm in this paper.

Related work

Notations

In this paper, the sample data matrix is denoted as $X = [X_l, X_u] \in R^{d\times n}$, where X_l denotes the labeled data matrix, X_u denotes the unlabeled data matrix. d and n denote the feature dimension and the number of samples, respectively. The true label matrix is denoted as $Y = \begin{bmatrix} Y_l \\ Y_u \end{bmatrix} \in R^{n\times c}$, where $Y_l \in R^{l\times c}$ denotes the label matrix with the labeled data part, where l and c denote the number of labeled samples and the number of classes, respectively. If the i-th sample belongs to the j-th class, then $Y_{ij} = 1$, otherwise $Y_{ij} = 0$. It is worth noting that in multi-label data, a sample may belong to more than one class. $Y_u \in R^{u\times c}$ denotes the label matrix of the unlabeled data part. The pseudo-label matrix is denoted as $F = \begin{bmatrix} F_l \\ F_u \end{bmatrix} \in R^{n\times c}$.

Notations: For any matrix $M \in R^{n\times d}$, where $m_{i,.}$ represent the vector of the i-th row of matrix M, $m_{.,j}$ represents the vector of the j-th column of matrix M, and $m_{i,j}$ represents the element of the i-th row, the j-th column of the matrix M; The Forbenius norm of matrix M is $||M||_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{d} m_{ij}^2}$, and the $L_{2,1}$-norm of matrix M is $||M||_{2,1} = \sum_{i=1}^{n}\sqrt{\sum_{j=1}^{d} m_{ij}^2} = \sum_{i=1}^{n} ||m_{i,.}||_2$; The $L_{2,0}$-norm of matrix M is expressed as $||M||_{2,0} = \sum_{i=1}^{n}(\sum_{j=1}^{d} m_{ij}^2)^0 = \sum_{i=1}^{n} ||m_{i,.}||_0$. $Tr(M)$ represents the trace of matrix M. M^T represents the transposition of matrix M.

Latent representation

Latent representations can benefit many data mining and machine learning tasks and have recently received increasing attention, especially for network data. Latent representations can be obtained from the similarity matrix of samples, since the more similar two samples are, the more likely they are to influence each other. Typically, the information represented is generated through nonnegative factorization [32]. The similarity matrix A is decomposed into the product of the nonnegative matrix Q and its transposed Q^T. The specific form of this model is shown in Eq. (1).

$$\min_{Q\geq 0} \| A - QQ^T \|_F^2, \tag{1}$$

where $Q \in R^{n\times c}$ is the latent representation of n data instances, which is the latent representation matrix. c is the number of categories, so Q is also the clustering structure of data, which can guide feature selection. The similarity matrix is constructed by Gaussian function and used to represent the interconnection information between data. The similarity matrix A is defined as Eq. (2).

$$A = \exp\left(\frac{- \parallel x_{.,i} - x_{.,j} \parallel \cdot_{2}^{2}}{\sigma^2}\right). \tag{2}$$

Expand OGSSL [33]

OGSSL adaptively learns S from data. If any two samples $x_{.,i}$, and $x_{.,j}$ are similar (i.e., their Euclidean distance is small), they are more likely to have the same labels, so the weight s_{ij} should be large. Otherwise, s_{ij} should be very small. If $s_{ij} = 0$, it indicates that there is no correlation between $x_{.,i}$ and $x_{.,j}$, therefore the following model is established.

$$\min_{S \geq 0} \sum_{i,j=1}^{n} (\parallel x_{.,i} - x_{.,j} \parallel_2^2 s_{ij} + \gamma s_{ij}^2). \tag{3}$$

Intuitively speaking, we can project the above data into a subspace where we can more accurately model data similarity. Assuming that the subspace is identified by the projection matrix $W \in R^{d \times m}$, where m is the dimension of the subspace. We force W to be row sparse to select more discriminative features. Mathematically, we replace $x_{.,i}$ in (3) with a linear combination $W^T x_{.,i}$ (similarly, we replace $x_{.,j}$ with $W^T x_{.,j}$ and merge $||W||_{2,1}$ into regularization; thus, we can obtain the following objective function:

$$\min_{S,W} \sum_{i,j=1}^{n} (\parallel W^T x_{.,i} - W^T x_{.,j} \parallel_2^2 s_{ij} + \alpha s_{ij}^2) + \beta \parallel W \|_{2,1}. \tag{4}$$

Model establishment

In semi-supervised multi-label feature selection, the acquisition of label information is time-consuming and expensive. Therefore, we hope to make full use of label information to guide feature selection, so that more accurate and efficient features can be selected. First, we add the supervision information to the construction of the initial similarity matrix to construct a more accurate initial similarity matrix. When building the initial similarity matrix, we need to measure the pairwise similarity between the real labels in the marking part through the Jaccard index (expressed in P_{ij}, the larger P_{ij}, the more similar instances are). And the larger the pairwise constraint P_{ij} between labels, the larger the similarity matrix $\tilde{A}_{ij}$ constructed. Therefore, the specific structure of the similarity matrix is as follows:

$$P_{i,j} = \frac{|y_i \cap y_j|}{|y_i \cup y_j|},$$

$$\tilde{A}_{ij} = \begin{cases} \frac{A_{ij}}{1.1 - P_{ij}} & P_{ij} \geq 0.5, \\ A_{ij} P_{ij} & P_{ij} < 0.5, \\ A_{ij} & i > l \ or \ j > l. \end{cases} \tag{5}$$

However, if the constructed similarity matrix is not good enough, it will affect its guidance of feature selection. Therefore, this paper uses the constructed initial matrix similarity $\tilde{A}$ to learn similarity matrix S, so as to guide feature selection. In the following, we designed ablation experiments to verify that the method we designed to learn the similarity matrix is better than using the similarity matrix directly to guide feature selection. Secondly, to ensure the correctness of known labels, we set $F_l = Y_l$ and obtain the objective function as follows:

$$\min_{S, F_l = Y_l} \parallel S - \tilde{A} \parallel_F^2 . \tag{6}$$

Based on the significance of latent representations, we hope that the greater the similarity between two samples, the more shared labels they have. Therefore, the matrix similarity is decomposed into the product of the pseudo-label matrix F and its transposed F^T. And by combining latent representations with semi-supervised information, a more ideal pseudo-label matrix is learned, resulting in the following objective function:

$$\begin{aligned} &\min \parallel S - \tilde{A} \parallel_F^2 + \alpha \parallel S - F F^T \parallel_F^2 \\ &s.t \ \langle S, F \rangle \geq 0, F_l = Y_l. \end{aligned} \tag{7}$$

Due to the latent representation of information generated through nonnegative factorization, $F \geq 0$. Where α is a parameter that balances latent representation learning and feature selection in the latent space.

With the development of spectral analysis and manifold learning, many feature selection methods try to preserve the local manifold structure, which is better than the global structure. Therefore, the similarity matrix is critical to the final performance of spectral methods. However, most methods simply construct a similarity matrix from the original features containing many redundant and noise samples or features. This will inevitably damage the learned structure, and the similarity matrix is definitely unreliable and inaccurate. Therefore, this paper will apply the adaptive process to determine the similarity matrix with the probability neighbor through this algorithm [34]. In other words, we perform both feature selection and local structure learning.

This article represents XW as the linear combination and finds the best linear combination of the original features, which can approximate a low dimensional manifold. Where $W \in R^{d \times m}$ is the projection matrix, d and m are the original dimension and projection dimension, respectively. And considering that S and W should have similar manifold structures, we can add a graph Laplace regularization term. Use dynamic manifold diagrams to constrain the manifold structure of W. Therefore, the following objective function is obtained:

$$\begin{aligned}&\min tr(W^T X L_S X^T W) + \lambda \parallel S - \tilde{A} \parallel_F^2 + \alpha \parallel S - FF^T \parallel_F^2 \\ &s.t \ \langle S, F \rangle \geq 0, F_l = Y_l, W^T W = I. \end{aligned} \tag{8}$$

Among them, we add the constraint $W^T W = I$ to ensure that the original data are still statistically irrelevant after being mapped to a low-dimensional space.

In recent years, sparse learning technology has been widely used in feature selection models to obtain row sparse weight matrix. Therefore, the sparse regularization term of W is added in this paper to obtain a more sparse projection matrix W for feature selection. In this article, we use the $l_{2,1-}$ norm regularization term to obtain the following objective function:

$$\begin{aligned}&\min tr(W^T X L_S X^T W) + \lambda \parallel S - \tilde{A} \parallel_F^2 \\ &\quad + \alpha \parallel S - FF^T \parallel_F^2 + \beta \parallel W \parallel_{2,1} \\ &s.t \ \langle S, F \rangle \geq 0, F_l = Y_l, W^T W = I. \end{aligned} \tag{9}$$

In the objective function, the $l_{2,1}$-norm regularization term is used to constrain the sparsity of row W, while the $l_{2,0}$-norm regularization term is not used because the projection matrix W learned using the $l_{2,1}$-norm regularization term in this model is more sparse. This article sets the number of selected features to 50 and the projection dimension to 10 on the dataset Emotions. By comparing the sparsity of W constrained by the $l_{2,1}$-norm regularization term and the $l_{2,0}$-norm regularization term, the superiority of the $l_{2,1}$-norm regularization term in our model is verified. Due to the large size of the dimension, it is difficult to observe, so we take the first 20 rows of the projection matrix W. As shown in Fig. 2, the darker the color, the sparser the image. Figure (a) represents the initial visual image of W, Figure (b) represents the visual image of W learned using the $l_{2,1}$-norm, and Figure (c) represents the visual image of W learned using the $l_{2,0}$-norm. Although both methods of learning W are sparse, it is evident that using the $l_{2,1}$-norm regularization term results in a more sparse learning of W.

Model solution

In this article, we adopt an alternating iterative algorithm to solve this problem.
1. Fix W and S, solve F

$$J(F) = \min_{F_l = Y_l, F \geq 0} \alpha \parallel S - FF^T \parallel_F^2 . \tag{10}$$

Due to $F = \begin{bmatrix} F_l \\ F_u \end{bmatrix}$, the above equation can be written as equation (11)

$$J(F) = \min_{F_l = Y_l, F \geq 0} \alpha \parallel \begin{bmatrix} S_{ll} & S_{lu} \\ S_{ul} & S_{uu} \end{bmatrix} - \begin{bmatrix} F_l \\ F_u \end{bmatrix} \begin{bmatrix} F_l^T & F_u^T \end{bmatrix}^T \parallel_F^2 . \tag{11}$$

According to the Lagrange method, equation (11) can be written as equation (12)

$$\begin{aligned} L(F_u) = \min \alpha (tr(& F_l F_u^T F_u F_l^T - 2S_{ul}^T F_u F_l^T + F_u F_l^T F_l F_u^T \\ & - 2F_u F_l^T S_{lu} + F_u F_u^T F_u F_u^T - 2F_u F_u^T S_{uu})) \\ & + tr(\theta {F_u}^T). \end{aligned} \tag{12}$$

According to the Kuhn–Tucker condition, for any i, j, there is $\theta_{ij} F_{ij} = 0$. Therefore, we obtain the following equation by setting $\frac{\partial L(F)}{\partial F_u} = 0$:

$$\begin{aligned} &(F_u F_l^T F_l - S_{ul}^T F_l + F_u F_l^T F_l - S_{lu}^T F_l \\ &+ 2F_u F_u^T F_u - 2S_{uu} F_u)_{ij} (F_u)_{ij} = 0. \end{aligned} \tag{13}$$

Through the optimization framework of nonnegative quadratic problems, we can obtain the update formula for each element $(Fu)_{ij}$ of F_u:

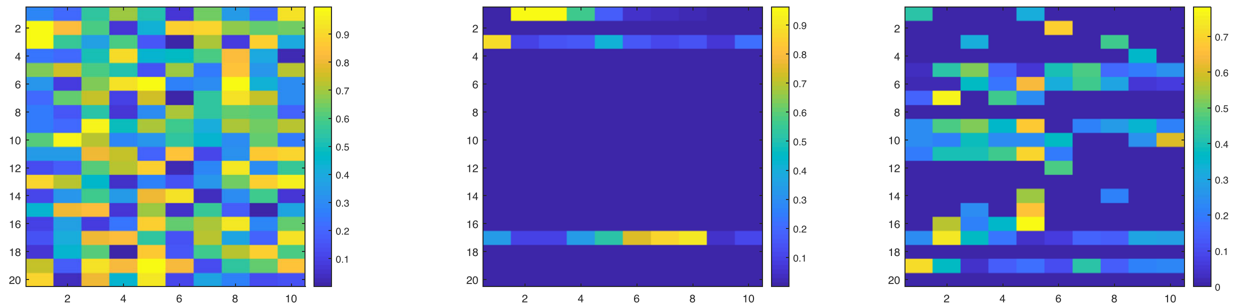
$$(F_u)_{ij} \leftarrow (F_u)_{ij} \frac{(S_{ul}^T F_u + S_{lu}^T F_l + 2S_{uu} F_u)_{ij}}{(F_u F_l^T F_l + F_u F_l^T F_l + 2F_u F_u^T F_u)_{ij}}, \tag{14}$$

$$F_l = Y_l.$$

Therefore, we obtain the update formula for F as follows:

$$F = \begin{bmatrix} F_l \\ F_u \end{bmatrix}. \tag{15}$$

2. Fix F and S, solve W



(a) the initial visual image of W (b) SSLR($\| W \|_{2,1}$) (c) SSLR($\| W \|_{2,0}= k$)

Fig. 2 Visual view of W obtained by different methods on the Emotions dataset

$$J(W) = \min_{W^TW=I} tr(W^T X L_S X^T W) + \beta \parallel W\|_{2,1}. \tag{16}$$

Equation (16) can be equivalent to equation (17) by organizing it

$$J(W) = \min_{W^TW=I} tr(W^T X L_S X^T + \beta D W), \tag{17}$$

$$J(W) = \min_{W^TW=I} tr(W^T R W), \tag{18}$$

where D is a diagonal matrix, $D_{ii} = \frac{1}{2\|w_i\|_2}$, $R = X L_S X^T + \beta D$. The solution to such problem (18) is the eigenvector corresponding to the minimum k eigenvalues of matrix R.

3. Fix F and W, solve S

$$\begin{aligned} J(S) = \min_{S\geq 0} & tr(W^T X L_S X^T W) \\ & + \lambda \parallel S - \tilde{A} \parallel_F^2 + \alpha \parallel S - FF^T \parallel_F^2, \end{aligned} \tag{19}$$

(19) we can be equivalent to equation (20)

$$J(S) = \min_{S\geq 0} \parallel S - \frac{\lambda A + \alpha F F^T - \frac{1}{4} W^T X}{\alpha + \lambda} \parallel_F^2 . \tag{20}$$

Let $T = \frac{\lambda A + \alpha F F^T - \frac{1}{4} W^T X}{\alpha + \lambda}$, and the optimal solution to the above equation is $S_{t+1} = \max(T, 0)$.

```
Algorithm 1: SSLR

Input: data matrix $X \in R^{d\times n}$, label matrix $Y$, projection dimension $m$, select the number of features $k$, parameter $\alpha, \beta, \lambda$.
Initialization: random initialization $W \in R^{d\times m}$, initialize $F$ to all 0 matrices.

1. Update process:
    (a) Update $F_u$ through the equation (14).
    (b) $F_l = Y_l$.
    (c) Then, $F$ is updated through updates to $F_l$ and $F_u$.
    (d) Update $W$, which is formed by $k$ eigenvectors of $R = X L_s X^T + \beta D$, corresponding to $k$ minimum eigenvalues.
    (e) Update $S_{t+1} \leftarrow max(T, 0)$.
2. Perform steps (a)–(e) iteratively until the algorithm converges.

Output: A feature subset of $k$ selected features.
```

Generally, $m < n, m < d, k < n$, and $k < d$. According to the algorithm, during each iteration, the complexity of updating$W$ is $O(n^2 d)$, the complexity of updating S is $O(n^2 c)$, the complexity of updating F is $O(u^2 c)$. Therefore, the total time complexity is $O(t(n^2 d + n^2 c + u^2 c))$, where t is the number of iterations, due to the small t-value, time depends on the feature dimension d of the data and the number of samples n.

Experiments and results

In this section, we introduced the dataset used, the comparison algorithm, and the evaluation indicators used to validate the effectiveness of our algorithm. Use charts to illustrate the results of different experiments.

Experimental data and comparative algorithms

To verify the effectiveness of the proposed algorithm, we compared it with the following seven algorithms: S2MFSHMRMR [35], S-CLS [36], 3-3FS [37], PMU [38], SCLS [39], SSFS [40], SLMDS [41].

S2MFSHMRMR: Hessian energy semi-supervised multi-label feature selection based on maximum correlation and minimum redundancy.

S-CLS: A unified framework for semi-supervised multi-label feature selection based on Laplacian fraction.

3-3FS: Semi-supervised multi-label feature selection based on three-way data resampling integration method.

PMU: A multi-label feature selection algorithm based on mutual information. Perform multi-label feature selection by selecting the correlation between the selected feature and the label.

SCLS: A multi-label feature selection method based on extensible standards.

SSFS: A multi-label feature selection with constraint potential structure shared items.

SLMDS: A multi-label feature selection that preserves global label correlation and dynamic local label correlation by preserving the graph structure.

Six common datasets were used in the experiment. Among them, the Image and Scene datasets belong to the Image dataset; Emotion belongs to the music dataset; Enron and Computer belong to the text dataset; Yeast belongs to the biological dataset. These data are all from (http://mulan.sourceforge.net/datasets.html). Table 1 shows the specific parameters of the dataset:

Experimental design

(1) Because some of the comparison algorithms are supervised multi-label feature selection algorithms, the proportion of labeled data in the training set of supervised multi-label feature selection comparison algorithm is set to 1. In the semi-supervised multi-label feature selection comparison algorithm and the SSLR algorithm, the proportion of labeled data in the training set is set to 0.2 and 0.4 respectively. Therefore, compared to algorithms in label sets, it has significant advantages.

(2) The nearest neighbor parameter of all feature selection algorithms is set to $k = 5$, and the maximum iteration is set to 50.

(3) In the ML-KNN algorithm, set the smoothing parameter $S = 1$ and the adjacent parameter $k = 10$.

(4) Adjust regularization parameters α and λ through the "grid search" strategy. The search scope is set to $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}, 10^{6}\}$, and the number of selected functions is set to $\{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$. For all algorithms, when the parameters are optimal, the best result is obtained (each algorithm solves ten times to find the average).

Evaluating indicator

Let $D \in R^{n\times d}$ be the transpose of training set sample data and $Y \in R^{n\times c}$ be the corresponding label set data. $h(D_{i,.})$ is a binary label vector, and $rank_{i,.(q)}$ represents the rank predicted by label $y_{q,.}$.

(1) Hamming loss: The proportion of labels that are misclassified.

$$HL\left(D\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\left\|h\left(D_{i,.}\right)\Delta y_{i,.}\right\|_{1},$$

where Δ is the sign of symmetric difference.

(2) Ranking loss: It is the proportion of labels in reverse order.

$$RL\left(D\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\mathbf{1}_{m}^{T}y_{i,.}\mathbf{1}_{m}^{T}\tilde{y}_{i,.}}\sum_{q:y_{i,.}^{q}=1}\;\sum_{q':y_{i,.}^{q'}=0}\left(P\right),$$

where $P = \delta\left(rank_{i,.}\left(q\right) \geq rank_{i,.}\left(q'\right)\right)$, $\delta\left(z\right)$ are indicator function, and $\tilde{y}_{i,.}$ is the complement of $y_{i,.}$ on Y.

(3) One error: There is no sample proportion of "predicted most relevant labels" in the "real labels".

$$OE\left(D\right) = \frac{1}{n}\sum_{i=1}^{n}\delta\left(y_{i,.}^{l} = 0\right),$$

where $l = \arg\min_{q\in[1,m]} rank_{i,.}\left(q\right)$.

(4) Coverage: How many steps does the "sorted labels" need to be moved on average to cover the real label correlation set.

$$CV\left(D\right) = \frac{1}{n}\sum_{i=1}^{n}\arg\max_{q:y_{i,.}^{q}=1} rank_{i,.}\left(q\right) - 1.$$

(5) Average precision: The proportion of tags with higher correlation than specific tags in the ranking.

$$AP\left(D\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\mathbf{1}_{m}^{T}y_{i,.}}\sum_{q:y_{i,.}^{q}=1}\frac{\sum_{q':y_{i,.}^{q'}=1} P}{rank_{i,.}(q)}.$$

Among these five evaluation indicators, only a higher value of the metric of average accuracy indicates better algorithm performance, while smaller values of other metrics indicate better algorithm performance.

Table 1 Description of multi-label datasets

DID	Datasets	Domain	Instances	Features	Labels	Cardinality	Density
1	Image	Image	600	294	5	1.236	0.247
2	Emotions	Music	593	72	6	1.869	0.311
3	Enron	Text	1702	1001	43	3.378	0.064
4	Yeast	Biology	2417	103	14	1.237	0.303
5	Scene	Image	2407	294	6	1.074	0.179
6	Computers	Text	5000	681	33	1.507	0.046

Table 2 Comparison of average accuracy (%) of different algorithms under each dataset (scale of labeled data is 20%)

Datesets	S2MFSHMRMR	S-CLS	3-3FS	PMU	SCLS	SSFS	SLMDS	SSLR
Image	71.44	69.83	68.80	70.02	67.64	72.08	70.88	**72.52**
Emotions	76.68	74.04	74.90	71.43	75.1	74.91	75.48	**77.98**
Enron	59.01	63.77	59.73	64.66	58.99	64.38	65.73	**66.37**
Yeast	75.15	74.76	74.82	75.23	73.07	73.12	75.14	**75.63**
Scene	78.58	76.52	79.02	78.97	70.41	77.27	74.74	**81.68**
Computers	60.59	60.75	61.17	60.92	61.22	60.59	61.07	**62.11**

Experimental results and analysis

In this section, it is mainly verified through charts that the SSLR algorithm is superior to other comparative algorithms. Tables 2 and 3 compare the average accuracy of different algorithms for each dataset labeled with 20% and 40% samples in Table 1, respectively. The algorithm with the highest average accuracy is represented in bold, while the suboptimal algorithm is represented by an underline. Tables 2 and 3 show the optimal results of the algorithm under optimal parameters. When labeled 20% of the samples, it is easy to see from Table 2 that our algorithm (LRSS) has higher average accuracy than the other seven comparison algorithms in all six comparison algorithms. Especially on the Emotion and Scene datasets, our algorithm outperforms the suboptimal algorithm significantly. It can be easily seen from Table 3 that when 40% of the samples are labeled, the average accuracy of our algorithm (LRSS) is higher than the other seven comparison algorithms among the six comparison algorithms, especially on the Scene dataset, where our algorithm is significantly better than the suboptimal algorithm 3-3FS.

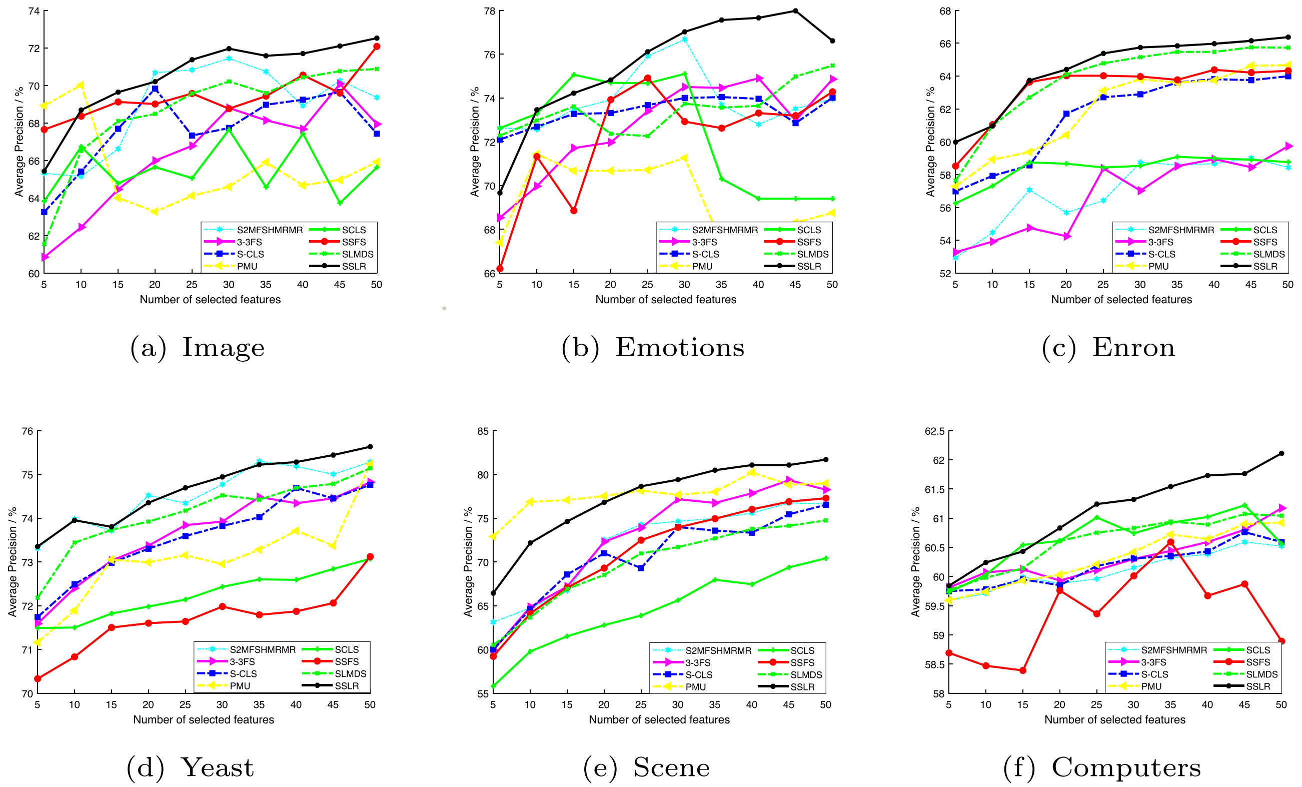
Figure 3 and Fig. 4, respectively, show the relationship between the average precision of the seven comparison algorithms, namely S2MFSHMRMR, S-CLS, 3-3FS, PMU, SCLS, SSFS, SLMDS and our own algorithm (SSLR), and the number of different feature selection under the samples marked with 20% and 40% of the six data sets in Table 1. It can be easily seen from the Fig. 3 that the algorithm of this paper largely outperforms the other comparison algorithms on all six datasets when 20% of the samples are labeled. In particular, on the datasets Enron and Computer, this paper's algorithm outperforms the other comparison algorithms regardless of the number of feature selections taken.It can be easily seen from Fig. 4 that the algorithm in this paper outperforms the other compared algorithms on the datasets Enron and Computer when 40% of the samples are labeled, regardless of the number of feature selections taken. While on the dataset Emotion the results are a bit worse, when the number of features is 5–20, the algorithm of this paper is slightly worse than S2MFSHMRMR.

Figure 5 shows the comparison of algorithm SSLR with other comparison algorithms in the evaluation indicators of hamming loss, ranking loss, coverage, and one error on the dataset Scene and Yeast (the lower the result, the better). From the graph, we can see that under other commonly used evaluation indicators, SSLR performs better than other comparison algorithms.

As shown in Fig. 6, the horizontal axis represents the ranking of the SSLR algorithm under each indicator; From left to right, the performance of the algorithm is getting better and better, with the best performing algorithm located at the far right of Fig. 6. Meanwhile, we report the results of the Bonferroni–Dunn test ($\alpha = 0.1$) in the form of an average rank graph, and the algorithm groups with no significant differences are connected. If the difference in average ranking reaches the critical value (CD) of the difference, there is a significant difference. The calculation formula for CD is: $CD = q_\alpha\sqrt{\frac{K(K+1)}{6N}}$.

In the formula: $q_\alpha = 3.1640\,(K = 8, \alpha = 0.1)$; K denotes the number of experimental algorithms; N denotes the number of experimental datasets; $CD = 4.4746$ $(K = 8, N = 6)$.

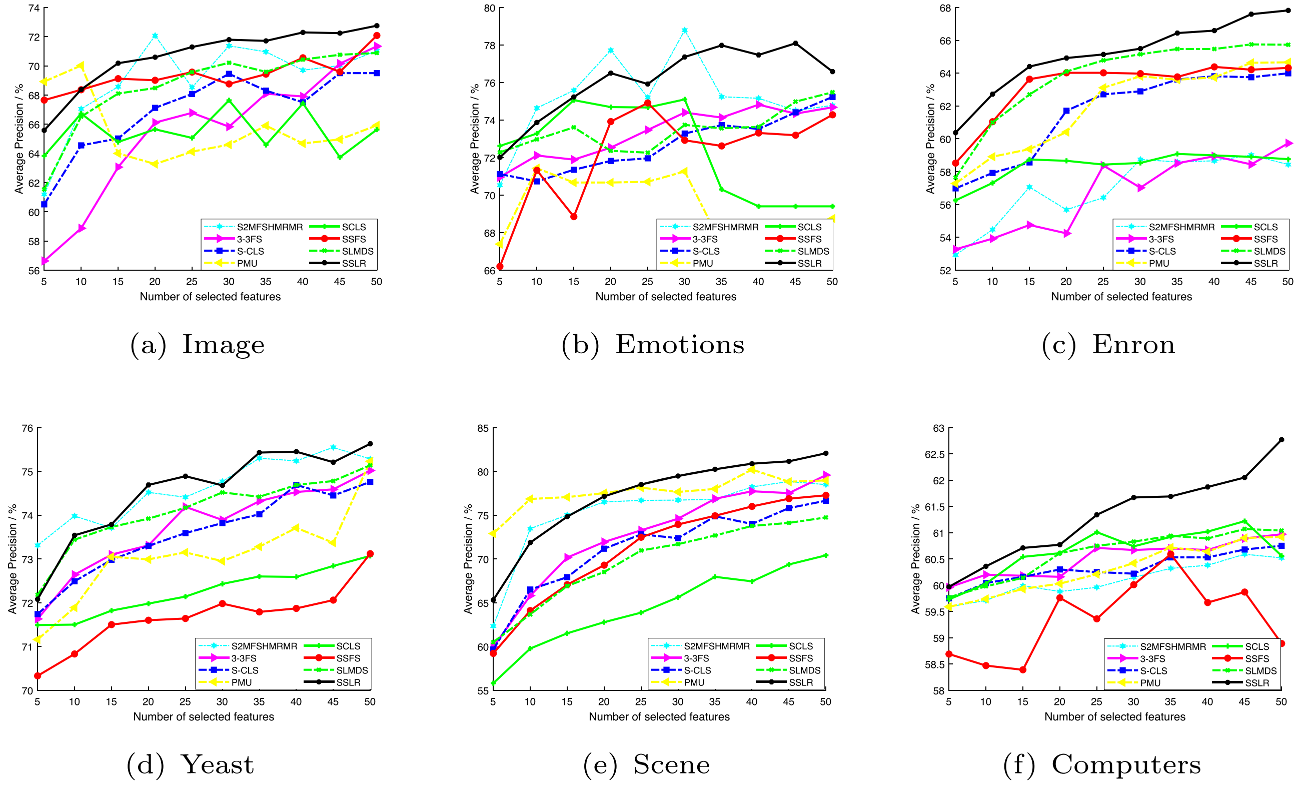
By observing Fig. 6, we found that SSLR did not show significant differences in various indicators compared to algorithms SLMDS and S2MFDHMRMR, but showed significant differences compared to PMU and SSFS.



(a) Image (b) Emotions (c) Enron

(d) Yeast (e) Scene (f) Computers

Fig. 3 Comparison of average precision between different algorithms on each dataset (scale of labeled data is 20%)



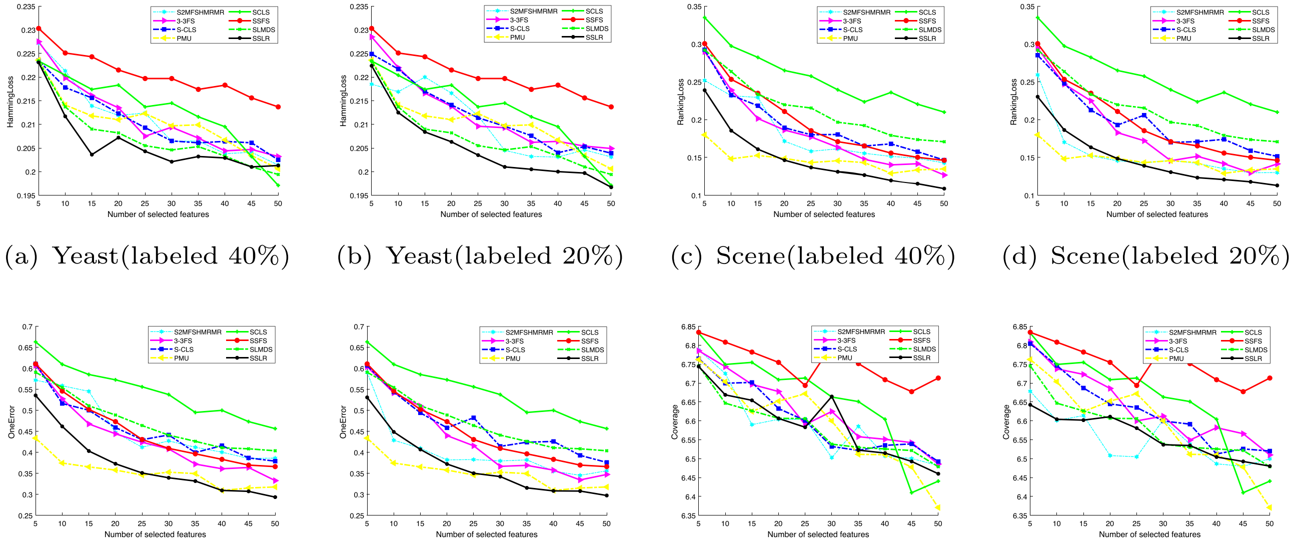
(a) Image (b) Emotions (c) Enron

(d) Yeast (e) Scene (f) Computers

Fig. 4 Comparison of average precision between different algorithms on each dataset (scale of labeled data is 40%)

Table 3 Comparison of average accuracy (%) of different algorithms under each dataset (scale of labeled data is 40%)

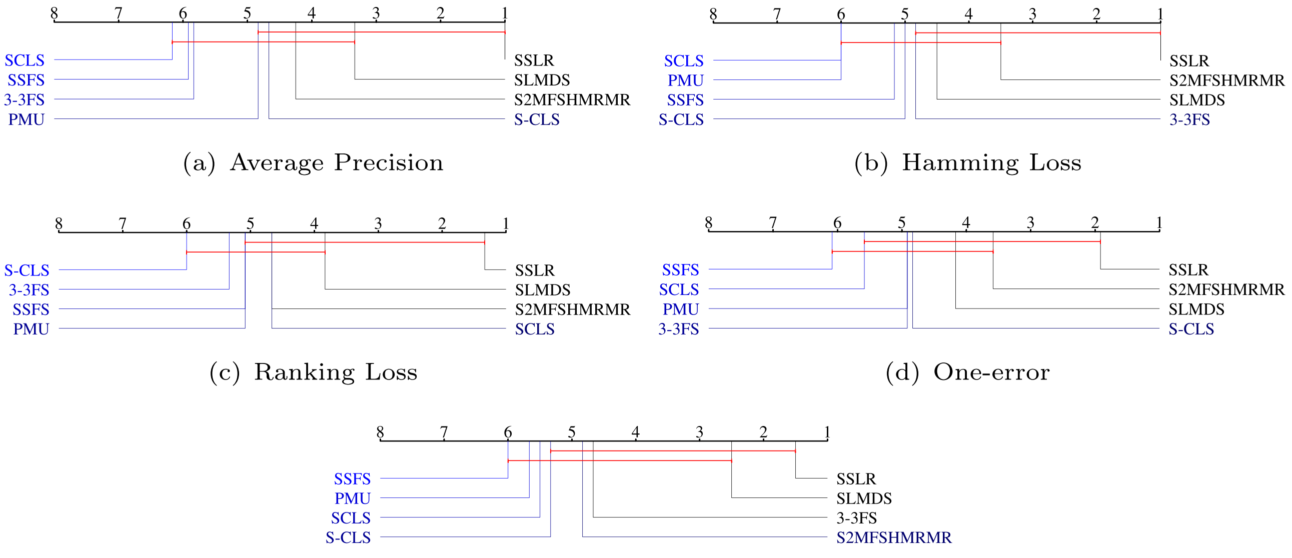
Datesets	S2MFSHMRMR	S-CLS	3-3FS	PMU	SCLS	SSFS	SLMDS	SSLR
Image	72.07	69.50	71.34	70.02	67.64	72.08	70.88	**72.75**
Emotions	78.79	75.23	74.82	71.43	75.1	74.91	75.48	**78.86**
Enron	59.73	63.77	59.73	64.66	58.99	64.38	65.73	**67.82**
Yeast	75.55	74.98	75.02	75.23	73.07	73.12	75.14	**75.63**
Scene	76.55	76.65	80.21	78.97	70.41	77.27	74.74	**82.07**
Computers	60.59	60.76	61.21	60.92	61.22	60.59	61.07	**62.77**



(a) Yeast(labeled 40%) (b) Yeast(labeled 20%) (c) Scene(labeled 40%) (d) Scene(labeled 20%)

(e) Scene(labeled 40%) (f) Scene(labeled 20%) (g) Yeast(labeled 40%) (h) Yeast(labeled 20%)

Fig. 5 Comparison of SSLR and contrast algorithm on hamming loss, ranking loss,coverage and one error (the lower the result, the better)

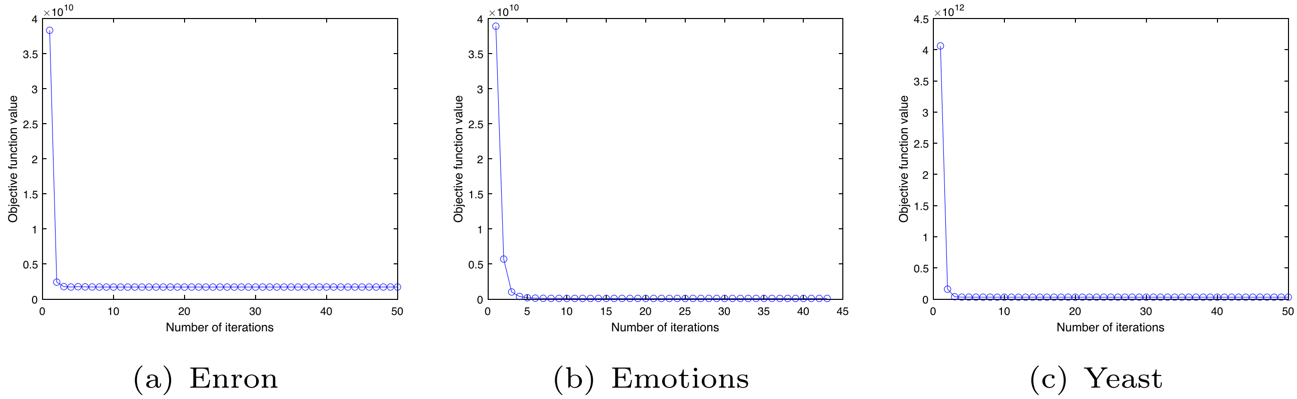


(a) Average Precision (b) Hamming Loss

(c) Ranking Loss (d) One-error

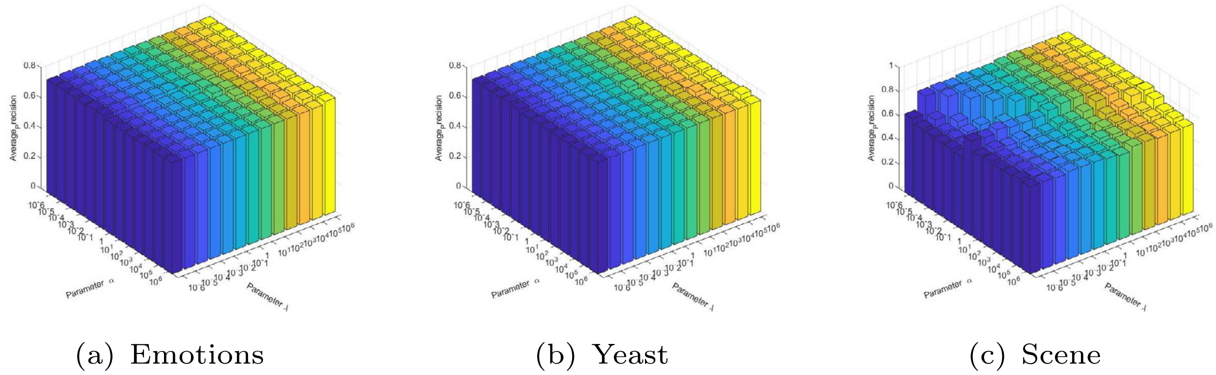
(e) Coverage

Fig. 6 The form of average rank graph of Bonferroni–Dunn test results



(a) Enron (b) Emotions (c) Yeast

Fig. 7 Convergence of SSLR algorithm under different datasets. (The proportion of labeled data was 40%)



(a) Emotions (b) Yeast (c) Scene

Fig. 8 Influence of different parameters α and λ on SSLR algorithm

Convergence analysis

In this section, we used experiments to verify the convergence of the algorithm on some datasets, labeled 40% of instances, and selected a feature count of 50. We fixed the number of nearest neighbors to 10. The results are shown in Fig. 7. From the results, we can see that our algorithm is effective and can converge within 10 iterations of each dataset, and the rate of convergence is very fast.

Parameter sensitivity analysis

In this section, we investigated the parameter sensitivity of the SSLR algorithm. Our feature selection model includes three algorithm parameters, namely α, β and λ. We use the “grid search" strategy. α and λ fix the search area to $[10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}, 10^{6}]$ and set other parameters to $\beta = 10^{4}$. Figure 8 shows the sensitivity analysis of parameters α and λ on some datasets in Table 1 for the algorithm SSLR, where our algorithm tag is 40%, the number of features selected is 50, and parameter β is fixed. From Fig. 8, we can easily see that SSLR is relatively stable on the dataset Yeast, and the average accuracy changes less with changes in parameters. However, the dataset Emotions and Scene are more sensitive to parameter α and λ in algorithm SSLR, and the optimal value of parameter α on the dataset Emotions is$[10^{3}, 10^{6}]$. On the dataset Scene, the optimal value for parameter α is $[10^{-6}, 10^{-4}]$. Different parameters have different impacts on the SSLR algorithm and varying degrees of sensitivity, which also proves that each regularization term improves the performance of the SSLR algorithm from different aspects.

Ablation experiment

In this section, to verify the effectiveness of the algorithm proposed in this paper by combining latent representation learning, dynamic manifold regularization, and sparse learning, a ablation experiment was designed to compare the effectiveness of our algorithm with the removal of latent representation learning, dynamic manifold learning, and sparse regularization terms, respectively. The impact of these four experiments on the dataset Images was compared in Table 4, and the results were displayed as the average accuracy of the evaluation indicators. The first row of the table shows the average accuracy of the LRSS algorithm in this paper, the second row shows the average accuracy of removing potential representation terms, the third row shows the average

Table 4 Comparison of average precision of ablation experiments under the number of feature selection on the dataset Image (the proportion of marked data is 0.4)

Number of features	5	10	15	20	25	30	35	40	45	50
LRSS	65.58	68.38	70.18	70.59	71.29	71.79	71.71	72.29	72.42	72.15
Latent representation	64.05	66.50	69.07	70.08	70.34	71.25	70.53	70.67	71.00	71.64
Dynamic manifold terms	63.31	67.27	68.52	69.44	70.45	70.09	70.64	71.71	71.11	71.02
Sparse regularization terms	64.65	67.01	68.90	69.56	70.06	70.26	71.23	71.03	71.56	72.00

accuracy of removing dynamic popular regularization terms, and the fourth item shows the average accuracy of removing sparse regularization terms.From Table 4, it can be seen that the LRSS algorithm performs better than the ablation experiment.

Conclusion

This paper proposes a sparse semi-supervised multi-label feature selection algorithm based on latent representation. First, semi-supervised information is added to construct the initial similarity matrix to learn a more ideal similarity matrix. Second, the latent information hidden in the optimal similarity graph is mined and decomposed into the product of the pseudo-label matrix and its transposition. And it can complete the update of dynamic manifold diagrams. Finally, use a dynamic manifold graph to constrain the manifold structure of the weight matrix. And compared with other comparative algorithms on multiple datasets, experimental results have proven the effectiveness of this algorithm.

The next step is to combine the model with non convex optimization, so that it can be applied to data sets in more fields, and more representative feature subsets can be selected.

Acknowledgements This work is supported by the Natural Science Foundation of China (61976130, 12101478), Natural Science Basic Research Program of Shaanxi (Program No. 2024JC-YBMS-015), Foreign Expert Program of China (Grant No. DL20230410021).

Data Availability The real multi-label datasets are obtained from http://mulan.sourceforge.net/datasets.html.

Declarations

Conflict of interest The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a Conflict of interest in connection with the work submitted.

Ethical and informed consent for data used This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Wang S, Jiang L, Li C (2015) Adapting naive bayes tree for text classification. Knowl Inform Syst 44:77–89
2. Jiang L, Li C, Wang S, Zhang L (2016) Deep feature weighting for naive bayes and its application to text classification. Eng Appl Artif Intell 52:26–39 (https://api.semanticscholar.org/CorpusID:7521864)
3. Schietgat L et al (2010) Predicting gene function using hierarchical multi-label decision tree ensembles. BMC Bioinform 11:2–2 (https://api.semanticscholar.org/CorpusID:18865766)
4. Weston J, Bengio S, Usunier N (2011) Wsabie: Scaling up to large vocabulary image annotation . https://api.semanticscholar.org/CorpusID:1337776
5. Zhang Y, cang Ma Y (2022) Sparse multi-label feature selection via dynamic graph manifold regularization. Int J Mach Learn Cybern 14, 1021–1036. https://api.semanticscholar.org/CorpusID:252798850
6. Gu Q, Li ZJ, Han J (2011) Correlated multi-label feature selection. https://api.semanticscholar.org/CorpusID:977205
7. Hu J, Li Y, Xu G, Gao W (2021) Dynamic subspace dual-graph regularized multi-label feature selection. Neurocomputing 467:184–196 (https://api.semanticscholar.org/CorpusID:241026167)
8. Hashemi A, Dowlatshahi MB, Nezamabadi-pour H (2020) Mgfs: a multi-label graph-based feature selection algorithm via pagerank centrality. Expert Syst Appl 142. https://api.semanticscholar.org/CorpusID:208115788
9. Fan Y (2021) et al. Multi-label feature selection with constraint regression and adaptive spectral graph. Knowl Based Syst 212, 106621. https://api.semanticscholar.org/CorpusID:229389731
10. Sun L, Kudo M, Kimura K (2017) Reader: Robust semi-supervised multi-label dimension reduction. IEICE Trans Inf Syst 100-D, 2597–2604. https://api.semanticscholar.org/CorpusID:30671335
11. Zhang Y, Zhou Z-H (2008) Multilabel dimensionality reduction via dependence maximization . https://api.semanticscholar.org/CorpusID:8189944

12. Sheikhpour R, Sarram MA, Gharaghani S, Chahooki MAZ (2017) A survey on semi-supervised feature selection methods. Pattern Recognit 64:141–158 (https://api.semanticscholar.org/CorpusID:8010228)
13. Zhang P, Liu G, Song J (2023) Mfsjmi: multi-label feature selection considering join mutual information and interaction weight. Pattern Recognit 138:109378 (https://api.semanticscholar.org/CorpusID:256590800)
14. Lyu Y, Feng Y, Sakurai K (2023) A survey on feature selection techniques based on filtering methods for cyber attack detection. Information 14:191 (https://api.semanticscholar.org/CorpusID:257613641)
15. Zhang Y, Ma Y (2022) Non-negative multi-label feature selection with dynamic graph constraints. Knowl Based Syst 238:107924 (https://www.sciencedirect.com/science/article/pii/S0950705121010728)
16. Lai J, Chen H, Li W, Li T, Wan J (2022) Semi-supervised feature selection via adaptive structure learning and constrained graph learning. Knowl Based Syst 251:109243 (https://api.semanticscholar.org/CorpusID:249716426)
17. Lai J, Chen H, Li T, Yang X (2022) Adaptive graph learning for semi-supervised feature selection with redundancy minimization. Inf Sci 609:465–488 (https://api.semanticscholar.org/CorpusID:250938272)
18. Xing Z, cang Ma Y, Yang X, Nie F (2021) Graph regularized nonnegative matrix factorization with label discrimination for data clustering. Neurocomputing 440, 297–309. https://api.semanticscholar.org/CorpusID:233874032
19. Liu K et al (2019) Rough set based semi-supervised feature selection via ensemble selector. Knowl Based Syst 165:282–296 (https://api.semanticscholar.org/CorpusID:58005379)
20. Tao H, Qiu J, Chen Y, Stojanovic V, Cheng L (2023) Unsupervised cross-domain rolling bearing fault diagnosis based on time-frequency information fusion. J Franklin Inst 360:1454–1477 (https://www.sciencedirect.com/science/article/pii/S0016003222008055)
21. Shen L, Tao H, Ni Y, Wang Y, Stojanovic V (2023) Improved yolov3 model with feature map cropping for multi-scale road object detection. Measur Sci Technol 34:045406. https://doi.org/10.1088/1361-6501/acb075
22. Zhu X (2005) Semi-supervised learning literature survey. https://api.semanticscholar.org/CorpusID:2731141
23. Zhao M, Chow TWS, Wu Z, Zhang Z, Li B (2015) Learning from normalized local and global discriminative information for semi-supervised regression and dimensionality reduction. Inf Sci 324:286–309 https://api.semanticscholar.org/CorpusID:205462596
24. Basu S, Davidson I, Wagstaff KL (2008) Constrained clustering: advances in algorithms, theory, and applications. https://api.semanticscholar.org/CorpusID:58294616
25. Zhang Z, Chow TWS, Zhao M (2013) Trace ratio optimization-based semi-supervised nonlinear dimensionality reduction for marginal manifold visualization. IEEE Trans Knowl Data Eng 25:1148–1161 (https://api.semanticscholar.org/CorpusID:11058995)
26. Guo B, Tao H, Hou C, yun Yi D (2019) Semi-supervised multi-label feature learning via label enlarged discriminant analysis. Knowl Inform Syst **62**, 2383–2417. https://api.semanticscholar.org/CorpusID:204711899
27. 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. https://api.semanticscholar.org/CorpusID:53034807
28. Zhang Y, cang Ma Y, Yang X, jun Zhu H, Yang T (2021) Semi-supervised multi-label feature selection with local logic information preserved. Adv Comput Intell 1. https://api.semanticscholar.org/CorpusID:239627976
29. Lv S, Shi S, Wang H, Li F (2021) Semi-supervised multi-label feature selection with adaptive structure learning and manifold learning. Knowl Based Syst 214:106757 (https://api.semanticscholar.org/CorpusID:232022520)
30. Chen G, Song Y, Wang F, Zhang C (2008) In: Semi-supervised multi-label learning by solving a sylvester equation. https://api.semanticscholar.org/CorpusID:16797771
31. Kraus V, Benabdeslem K, Canitia B (2020) Laplacian-based semi-supervised multi-label regression. In: 2020 International Joint Conference on Neural Networks (IJCNN) 1–8. https://api.semanticscholar.org/CorpusID:221659516
32. Tang C et al (2019) Unsupervised feature selection via latent representation learning and manifold regularization. Neural Netw Off J Int Neural Netw Soc 117:163–178 (https://api.semanticscholar.org/CorpusID:174817440)
33. Peng Y et al (2022) Ogssl: a semi-supervised classification model coupled with optimal graph learning for EEG emotion recognition. IEEE Trans Neural Syst Rehab Eng 30:1288–1297 (https://api.semanticscholar.org/CorpusID:248832239)
34. Nie F, Wang X, Huang H (2014) Clustering and projected clustering with adaptive neighbors. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. https://api.semanticscholar.org/CorpusID:5226376
35. Wu X, Chen H, Li T, Chen H, Luo C (2021) Semi-supervised multi-label feature selection using hessian energy based on maximum relevance and minimum redundancy. In: 2021 16th International Conference on Intelligent Systems and Knowledge Engineering (ISKE) 242–248. https://api.semanticscholar.org/CorpusID:248248580
36. Alalga A, Benabdeslem K, Taleb N (2016) Soft-constrained laplacian score for semi-supervised multi-label feature selection. Knowl Inform Syst 47:75–98 (https://api.semanticscholar.org/CorpusID:14472875)
37. Alalga A, Benabdeslem K, Mansouri DEK (2021) 3–3fs: ensemble method for semi-supervised multi-label feature selection. Knowl Inform Syst 63:2969–2999 (https://api.semanticscholar.org/CorpusID:240162883)
38. Lee J, Lim H, Kim D-W (2012) Approximating mutual information for multi-label feature selection. Electron Lett 48:929–930 (https://api.semanticscholar.org/CorpusID:120361202)
39. Lee J-S, Kim D-W (2017) Scls: multi-label feature selection based on scalable criterion for large label set. Pattern Recognit 66:342–352 (https://api.semanticscholar.org/CorpusID:5157040)
40. Gao W, Li Y, Hu L (2021) Multilabel feature selection with constrained latent structure shared term. IEEE Trans Neural Netw Learn Syst 34:1253–1262 (https://api.semanticscholar.org/CorpusID:237315011)
41. Li Y, Hu L, Gao W (2022) Robust sparse and low-redundancy multi-label feature selection with dynamic local and global structure preservation. Pattern Recognit 134:109120 (https://api.semanticscholar.org/CorpusID:253105009)

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