Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 10, No. 1, 2024

Multi-label Feature Selection based on Label-specific features and Manifold Learning

Wanzhu Wang, Yong Liu

Henan University of Science and Technology, 471000, Luoyang, China

Abstract: Each instance in multi-label data is associated with multiple labels, and there are irrelevant or redundant features in its feature space, which leads to the performance degradation of multi-label learning algorithms. Multi-label feature selection selects representative features from the feature space to improve the accuracy of the model. Due to the high cost of labels and the difficulty of data collection, there will be some missing labels in the data set, which affects the accuracy of feature selection. To solve this problem, a multi-label feature selection algorithm based on label-specific features and manifold learning is proposed. The algorithm uses the linear relationship between the features and labels in known label samples to build a linear regression model for learning label-specific features. By using the nonlinear relation between instances and the nonlinear relation between features, we can precisely learn the label-specific features. We use the Laplacian feature mapping method to construct the instance manifold model and the feature manifold model, which are also used as the regular term constraint weight matrix. The final model can not only complete the missing labels, but also select sparse and representative features. The feature selection is carried out by analyzing the weight of the feature given by the final model. Experiments were conducted to verify the effectiveness of the proposed algorithm under different label deletion rates on four evaluation indexes.

Keywords: Multi-label learning; Manifold learning; Label-specific features; Feature selection.

1. Introduction

Multi-label learning is characterized by high efficiency and flexibility, and has shown wide application prospects in many application fields such as machine learning, data mining and pattern recognition [1-3], and is also widely used in various practical tasks, such as text classification, music classification and gene function prediction. For example, in the text category, a news story can be classified as both social and technological; In the music category, a piece of music may belong to both light music and English songs. Multiple concepts can be expressed by labeling each instance differently. However, in such cases where multi-label classification is required, attaching only one label to each instance does not express the underlying complex semantics. Therefore, multi-label learning is achieved by assigning multiple labels to an instance at the same time.

Like single-label learning, the feature space of multi-label learning is often a set containing a large number of features, which may have up to thousands of dimensions [4]. In text classification, the complex semantic information of a text can be represented in detail by thousands of features. By extracting and utilizing these features effectively, we can understand the intrinsic meaning and context of the text more deeply. However, in the feature set, most of the features often present redundant or irrelevant features. These redundant and uncorrelated features not only increase the computational complexity, but also may negatively affect the performance of multi-label learning algorithms, and also lead to overfitting and other problems. In order to solve this problem, researchers proposed a multi-label feature selection algorithm, which is mainly used for dimensionality reduction of multi-label data. Literature [5] proposed a filter-based feature selection method for multi-label classification. Based on random forest as a learner, a filter feature selection method based on statistical integration method and label space division based on integration method were used. Literature [6] proposes a multi-label feature selection based on coarse-grained balls and label distribution. By using the pellet computing model and its proposed label enhancement method, logical labels are transformed into label distribution by exploring the similarities between instances and coarse-grained balls. The method measures feature significance by the consistency of labels in samples within the same information granularity. Based on mutual information theory, literature [7] designed a new multi-label feature selection method combining dynamic correlation change, label redundancy and interactive information.

In the model training of multi-label feature selection algorithm, the correlation between labels is a key information, which can be effectively used to optimize the feature selection process. By considering the correlation between labels, the algorithm can more accurately capture the correlation and dependency between different labels, so as to select the features that are closely related to multiple labels. Literature [8] uses the co-occurrence relationship between labels to evaluate the similarity relationship between samples. According to literature [9], label correlation and label-specific features are two important features of multi-label learning. The global and local label correlation are calculated by label co-occurrence and neighborhood information respectively, and the label-specific features is learned by l_1 norm. Literature [10] considers the influence of two types of label dependencies on the algorithm. This method considers that second-order label dependencies and higher-order label dependencies are both important and complementary to capture label information, so second-order label dependencies and higher-order label dependencies are considered at the same time.

Compared with traditional machine learning, multi-label labeling is very difficult. With the increasing complexity of multi-label tasks, there are also many challenges, among which the increase of sample dimension and data volume will significantly affect the labeling cost. First, because of the large number of sample instances, the process of labeling them one by one is time-consuming and labor-intensive. Therefore, in the labeling process, manual labeling tends to select labels that are more important to label and discard other labels. The second is that human taggers can only label what they know according to cognitive differences. For example, an article is labeled machine learning, Data mining, and artificial neural networks, and the human tagger does not know about the category of artificial neural networks and only labels the article with the remaining two labels. A variety of circumstances have led to the creation of missing labels. However, the accuracy of the model or the efficiency of classification can be affected by the multi-label feature selection or multi-label classification on the multi-label data set with missing labels. In order to solve the impact of missing labels, many multi-label feature selection algorithms have been proposed to deal with missing labels based on theories such as multi-label information entropy [11], linear regression [12] and label correlation. He et al. [13] simultaneously considered label relevance, missing labels, and feature selection through a multi-classification framework. Literature [14] proposes an embedded packaging approach that combines the generation of label-specific features with subsequent model induction to solve the multi-label classification problem of missing labels.

At the same time, it is more efficient to use the specific features of each label than to use all the features to classify. Label-specific features, refer to the subset of features most closely associated with each label. In the multi-label learning scenario, the use of label-specific features for feature selection can reveal the deep correlation between features and labels more efficiently. At the same time, from the point of view of the nonlinear relationship between data, consider the nonlinear relationship between instances and the nonlinear relationship between features, and study their influence on the relationship between labels and features. Based on the above considerations, this paper proposes a multi-label feature selection algorithm (ML-LSF) based on label-specific features and manifold learning that can deal with missing labels.

2. Proposed Models

In multi-label learning, $X = [x_1, x_2, \ldots, x_n]^T \in R^{n\times d}$ represents a multi-label data set with n instances, where d represents the feature dimension; $Y = [y_1, y_2, \ldots, y_n]^T \in R^{n\times l}$ represents the set of labels corresponding to n instances, where l represents the number of labels. The label values in set Y are set to 1, 0, and -1.When the label value is 1, the instance has the label. A label value of -1 indicates that the instance does not have the label. When the label value is 0, the instance is missing the label.

2.1. Build a label-specific feature model

In a multi-label classification task, each label in the label space has the most relevant and discriminating features for the label, which are called label-specific feature. First, the basic label-specific feature model is constructed by linear regression method. The purpose of adding l_1 norm is to ensure that the features selected by each label are sparse and representative. The objective function is as follows:

$$\min_W \frac{1}{2}\|XW - Y\|_F^2 + \lambda\|W\|_1 \tag{1}$$

Where X is the eigenmatrix of the instance, Y is the label matrix containing the missing label, and λ is the parameter. W is the coefficient matrix, w_i represents the i row in the matrix W, the row vector represents the feature vector, and the column vector represents the label vector. If the value of a certain place in W is 0, it means that there is no correlation between the feature and the label at that place.

2.2. Build the instance manifold and feature manifold models

The core idea of manifold learning is that data in high-dimensional space is actually distributed on a low dimensional manifold structure, that is, although data may appear complex and difficult to analyze in the original high-dimensional space, their essential structure can often be characterized by a low dimensional manifold. The characteristic makes manifold learning suitable for handling nonlinear relationships between data. The goal of manifold learning is to effectively map points on manifold M in an N-dimensional space to a low dimensional n-dimensional space, where n<N. Popular regularization refers to adding manifold related terms to machine learning problems, with the aim of constraining the coefficient matrix W through manifold regularization terms.

Two strongly related instances can share related subsets of each other [15]. If there are two strongly correlated instances, then the corresponding label sets or feature sets of the two instances are similar, and the prediction labels obtained by using the corresponding W are also very similar. The purpose of Laplacian Eigenmap in manifold learning is to ensure that two strongly correlated instances are as close as possible after dimensionality reduction, and to mine the correlation between the instances by using the manifold model between the instances. Add the instance correlation to the equation (1) to get the objective function:

$$\min_W \frac{1}{2}\|XW - Y\|_F^2 + \frac{\alpha}{2}tr(W^T X^T L_x XW) + \lambda\|W\|_1 \tag{2}$$

Where, α（$\alpha > 0$） is the parameter; $L_x = D_x - S_x$, L_x refers to the Laplacian matrix, D_x is a diagonal matrix, $S_x \in R^{d\times d}$ represents the instance similarity matrix.

For two related features of a label, the label has a high probability of having one feature and the other feature. Feature correlation considers the interaction between features. Add the feature correlation to the equation (2) to get the final objective function:

$$\min_W \frac{1}{2}\|XW - Y\|_F^2 + \frac{\alpha}{2}tr(W^T X^T L_x XW) + \frac{\beta}{2}tr(W^T L_y W) + \lambda\|W\|_1 \tag{3}$$

Where, β（$\beta > 0$） is a parameter; $L_y = D_y - S_y$, L_y refers to the Laplacian matrix, D_y is a diagonal matrix, $S_y \in R^{d\times d}$ represents the feature similarity matrix. In equation (1), the l_1 norm is used to regularize the coefficient matrix to make it sparse and exclude some redundant features in the feature space.

2.3. Fill missing labels

In the case of missing labels, it is difficult to learn the real mapping relationship between features and labels, which leads to the unsatisfactory classification effect of multi-label tasks. The missing label problem is dealt with by the method of restoring it. The iterative process of solving coefficient matrix W and the recovery process of missing label are carried out simultaneously.

According to literature [16], the known labels in the original label matrix are assigned to a new matrix A, then $A_{ij} = Y_{ij}$; And where the label is missing, $A_{ij} = 0$.

In each iteration process of solving coefficient matrix W, the missing label value in A is updated according to equation (4), where $A = XW$, and the updated A is continued for subsequent iterations.

$$A_{ij} = \begin{cases} -1, & A_{ij} \leq -1 \\ A_{ij}, & -1 < A_{ij} < 1 \\ 1, & A_{ij} \geq 1 \end{cases} \tag{4}$$

The continuous values in A are restored to discrete values, and the final matrix A can be obtained according to equation (5). At the same time, let $Y = A$.

$$A_{ij} = \begin{cases} -1, & A_{ij} < 0.1 \\ 1, & A_{ij} \geq 0.1 \end{cases} \tag{5}$$

2.4. Model Optimization

It can be seen that the objective function belongs to the non-smooth convex optimization problem, so the method of Accelerated Proximal Gradient can be used to deal with the optimization problem. Equation (3) is divided into two parts:

$$\min_W \{F(W) = f(W) + g(W)\} \tag{6}$$

Among them,

$$f(W) = \frac{1}{2}\|XW - Y\|_F^2 + \frac{\alpha}{2} tr(W^T X^T L_x XW) + \frac{\beta}{2} tr(WW^T L_y) \tag{7}$$

$$g(W) = \lambda \|W\|_1 \tag{8}$$

Each element in the coefficient matrix W is calculated using a soft threshold, which is:

$$W_{t+1} = prox_\varepsilon(G^{(t)}) \tag{9}$$

The method proposed by Lin et al[17] is adopted to accelerate the solution of W, which can be set

$$W^{(t)} = W_t + \frac{b_{t-1} - 1}{b_t}(W_t - W_{t-1}) \tag{10}$$

Significant runtime savings.

At the same time, when solving the optimization problem of equation (7), Lipschitz continuity is satisfied: $\|\nabla f(W_1) - \nabla f(W_2)\| \leq L_f \|W_1 - W_2\|$,$L_f$ is Lipschitz constant, and L_f can be obtained by proving Lipschitz continuity. The proof process is as follows:

The gradient of W in equation (7) is calculated as

$$\nabla f(W) = X^T XW - X^T Y + \alpha X^T L_x XW + \beta L_y W \tag{11}$$

Assuming that there are parameters W_1 and W_2, substituting parameters W_1 and W_2 into equation (11) respectively, we can get

$$\nabla f(W_1) = X^T XW_1 - X^T Y + \alpha X^T L_x XW_1 + \beta L_y W_1 \tag{12}$$

$$\nabla f(W_2) = X^T XW_2 - X^T Y + \alpha X^T L_x XW_2 + \beta L_y W_2 \tag{13}$$

According to equation (12) and equation (13), and $\Delta W = W_1 - W_2$, can be obtained

$$\begin{aligned} &\|\nabla f(W_1) - \nabla f(W_2)\|_F^2 = \\ &\quad \left\|X^T X\Delta W + \alpha X^T L_x X\Delta W + \beta L_y \Delta W\right\|_F^2 \leq \\ &\quad 2\|X^T X\Delta W\|_F^2 + 2\|\alpha X^T L_x X\Delta W\|_F^2 + 2\left\|\beta L_y \Delta W\right\|_F^2 \leq \\ &\quad 2\|X^T X\|_2^2 \|\Delta W\|_F^2 + 2\|\alpha X^T L_x X\|_2^2 \|\Delta W\|_F^2 + 2\left\|\beta L_y\right\|_2^2 \|\Delta W\|_F^2 = \\ &\quad 2\left(\|X^T X\|_2^2 + \|\alpha X^T L_x X\|_2^2 + \left\|\beta L_y\right\|_2^2\right)\|\Delta W\|_F^2 \end{aligned} \tag{14}$$

Therefore

$$\|\nabla f(W_1) - \nabla f(W_2)\|_F^2 \leq 2\left(\|X^T X\|_2^2 + \|\alpha X^T L_x X\|_2^2 + \left\|\beta L_y\right\|_2^2\right)\|\Delta W\|_F^2 \tag{15}$$

Therefore

$$L_f = \sqrt{2\left(\|X^T X\|_2^2 + \|\alpha X^T L_x X\|_2^2 + \left\|\beta L_y\right\|_2^2\right)} \tag{16}$$

ML-LSF algorithm as shown in algorithm 1.

Algorithm 1: Multi-label Feature Selection based on Label-specific Feature and Manifold Learning

Input: training matrix $X \in R^{n \times d}$, label matrix $Y \in R^{n \times l}$ with missing labels, parameters λ, α, β;

Output: coefficient matrix W^*

1）Initialization: $W_0 = W_1 = (X^T X + \gamma I)^{-1} X^T Y$, $b_0 = b_1 = 1$, $t = 1$

2）Repeat:

3） Calculate the Laplacian matrix L_x, L_y

4） Calculate L_f according to equation(16)

5） $W^{(t)} = W_t + \frac{b_{t-1} - 1}{b_t}(W_t - W_{t-1})$

6） $G^{(t)} = W^{(t)} - \frac{1}{L_f}\nabla f(W^{(t)})$

7） $W_{t+1} = prox_\varepsilon\left(W^{(t)} - \frac{1}{L_f}\nabla f(W^{(t)})\right)$

8） $W_{t-1} = W_t$

9） Update Y by $Y = XW_t$, according to equation (4)

10） $t = t + 1$

11） $b_{t+1} = \frac{1 + \sqrt{4b_t^2 + 1}}{2}$

12）until convergence

13）Restore the label matrix Y with missing labels according to equation (5)

14）$W^* = W_t$

15）Compute $\|w_i\|$, $i \in [1, d]$

16）Sort and select the first l features

2.5. Time Complexity

In step 1, W needs to be initialized and the time complexity is $O(ndl + nd^2 + d^3)$; In step 4, we need to compute L_f, and the time complexity is $O(d^3 + l^3)$; In step 6, we need to compute $\nabla f(W)$, the time complexity is $O(d^2 l + dl^2)$, and after t iterations, the time complexity is $O(t(d^2 l + dl^2))$. As a result, the algorithm's time complexity is $O(ndl + nd^2 + d^3) + O(d^3 + l^3) + O(t(d^2 l + dl^2))$.

3. Experimental Design and Results

3.1. Data set and experiment setup

In order to verify the effectiveness of ML-LSF, Mulan public data sets from multi-label library (http://mulan.sourceforge.net/datasets.html) to choose five classic data set, as shown in table 1.

Table 1. Multi-label Datasets

Data set	Domain	Sample	Feature	Class
Arts	Text	5000	462	26
Entertain	Text	5000	640	21
Emotions	Music	593	72	6
Reference	Text	5000	793	33
Scene	Image	2407	294	6

For the algorithm ML-LSF in this paper, there are three parameters λ , α , β , and the parameter λ is set to $\{10^{-6}, 10^{-5}, \ldots, 10^{3}\}$ range; Parameter $\alpha = \{2^{-6}, 2^{-5}, \ldots, 2^{6}\}$, parameter $\beta = \{2^{-6}, 2^{-5}, \ldots, 2^{6}\}$, K-nearest neighbor probability graph model $K = 10$ for instance correlation and feature correlation. Set the label missing rate to 0%, 10%, and 20%. As with all comparison algorithms, the classical ML-KNN classifier is used, where the parameters Num and Smooth are set to 10 and 1, respectively.

3.2. Experimental results and analysis

In order to verify the performance of ML-LSF algorithm, PMU, MFML, MLMLFS and LSF-CI algorithms are selected to compare with ML-LSF algorithm. The first three algorithms are popular multi-label feature selection algorithms and the last one is label-specific features multi-label learning algorithm. The best results are highlighted in bold in the table.

The experimental results from Table 2 to Table 5 were analyzed:

When the label missing rate is 10%, the performance of ML-LSF algorithm is significantly improved compared with the other four algorithms. When a label is missing, the method of first completing the missing label can effectively improve the performance of the algorithm. MFML algorithm uses feature interaction to select more valuable features that may be ignored due to incomplete label space. In this way, the multi-label feature selection problem under missing labels can be solved. MLMLFS algorithm uses linear regression model to recover missing labels, which improves the accuracy of its modeling. The experimental results show that these two algorithms can also alleviate the performance problems caused by missing labels. However, the PMU algorithm can effectively handle multi-label feature selection when the label space is complete. However, when there are missing labels, the label space is damaged to a certain extent, resulting in the algorithm being unable to accurately model the relationship between features and labels.

In general, the classification performance of ML-LSF algorithm is better than other algorithms.

Table 2. Comparison results under 10% missing labels

Data set	PMU	MFML	MLMLFS	LSF-CI	ML-LSF
Arts	0.0625	0.0590	0.0589	0.0610	0.0588
Entertain	0.0631	0.0587	0.0580	0.0617	0.0576
Emotions	0.2533	0.2458	0.2436	0.2564	0.2389
Reference	0.0312	0.0280	0.0283	0.0305	0.0278
Scene	0.1121	0.1097	0.1092	0.1103	0.1091
Average	0.1044	0.1002	0.0996	0.1040	0.0984

Table 3. Comparison results under 10% missing labels

Data set	PMU	MFML	MLMLFS	LSF-CI	ML-LSF
Arts	0.4630	0.4987	0.5012	0.5011	0.5197
Entertain	0.5512	0.5678	0.5519	0.5612	0.5803
Emotions	0.7236	0.6830	0.7326	0.7130	0.7401
Reference	0.6153	0.6055	0.5918	0.6011	0.6183
Scene	0.7989	0.6497	0.6501	0.7991	0.8137
Average	0.6304	0.6009	0.6055	0.6351	0.6544

Table 4. Comparison results under 10% missing labels

Data set	PMU	MFML	MLMLFS	LSF-CI	ML-LSF
Arts	5.5996	5.4869	5.5016	5.4756	5.3817
Entertain	3.4021	3.2659	3.2854	3.3520	3.1978
Emotions	2.3016	2.2356	2.2217	2.2231	2.0356
Reference	3.5325	3.5023	3.4958	3.5124	3.4726
Scene	0.7326	0.6989	0.6807	0.7012	0.6695
Average	3.1137	3.0379	3.0370	3.0529	2.9514

Table 5. Comparison results under 10% missing labels

Data set	PMU	MFML	MLMLFS	LSF-CI	ML-LSF
Arts	0.6658	0.6271	0.6411	0.6315	0.6154
Entertain	0.6119	0.5903	0.6123	0.6124	0.5189
Emotions	0.3772	0.4298	0.3967	0.4128	0.3121
Reference	0.5120	0.5109	0.5116	0.5056	0.4713
Scene	0.3210	0.5397	0.3530	0.3498	0.3115
Average	0.4976	0.5396	0.5029	0.5024	0.4458

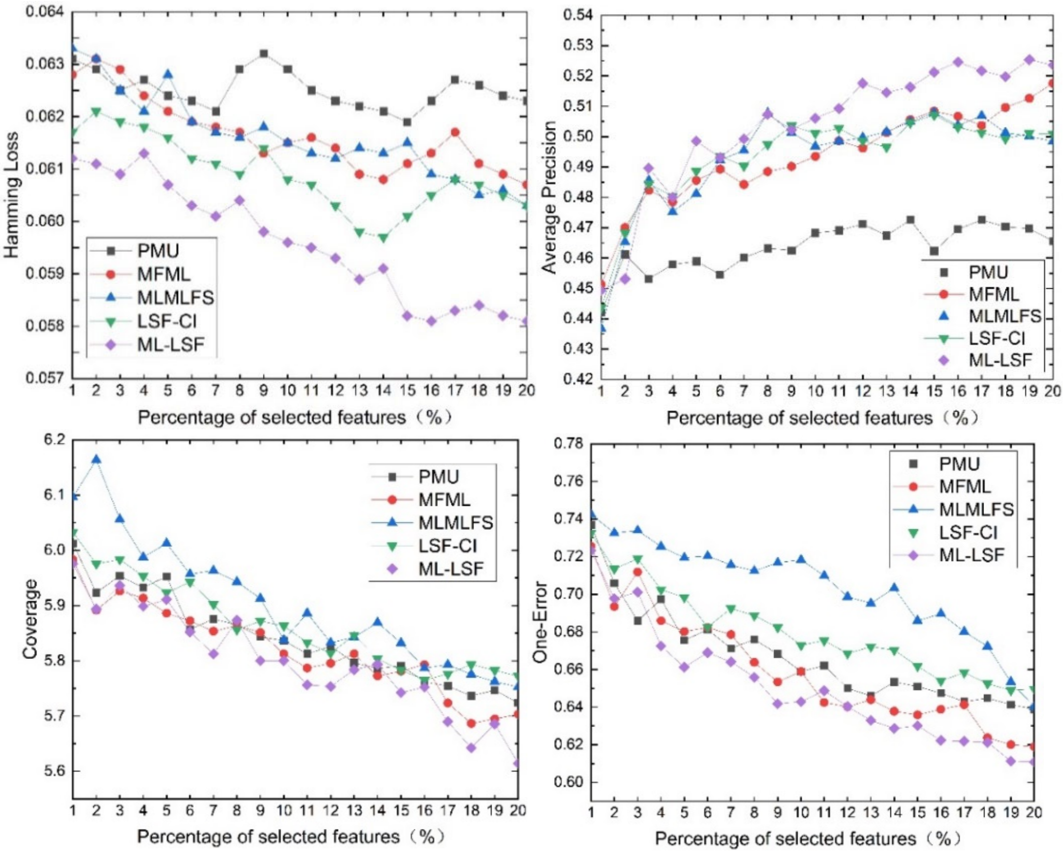

Figure 1. Results of four classification indicators when the label missing rate is 10% in Arts dataset

4. Conclusions

The missing label problem in the multi-label feature selection is analyzed, and a multi-label feature selection algorithm ML-LSF is proposed to deal with the missing label problem. Based on the linear relationship between label-specific features and labels and the nonlinear relationship between data, we learn the label-specific features and use the linear relationship between label-specific features and labels to fill the missing labels. The process of feature selection and label fill is iterative synchronously. The final solution is a non-smooth, convex optimization problem. The accelerated near-end gradient method is chosen to solve this problem. When comparing ML-LSF with other algorithms, the case of missing label rate of 10% is simulated, which can effectively verify the effectiveness of ML-LSF algorithm in processing multi-label data with missing label. Experimental results show that the proposed algorithm has good performance and is an effective multi-label feature selection algorithm.

References

[1] LIN Y J, HU Q H, LIU J H, et al. Streaming feature selection for multilabel learning based on fuzzy mutual information [J]. IEEE Transactions on Fuzzy Systems, 2017, 25(6): 1491-1507.

[2] WANG C X, LIN Y J, TANG L，et al. Multi-label feature selection based on information granulation[J]. Pattern Recognition and Artificial Intelligence, 2018, 31(2): 123-131.

[3] LIU J H, LI Y W, WENG W, et al. Feature selection for multilabel learning with streaming label [J]. Neurocomputing, 2020, 387: 268-278.

[4] ZHANG M L, ZHOU Z H. A review on multi-label learning algorithms [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8): 1819-1837.

[5] Shaikh R, Rafi M, Mahoto NA, Sulaiman A, Shaikh A. A filter-based feature selection approach in multilabel classification [J]. MACHINE LEARNING-SCIENCE AND TECHNOLOGY, 2023, 4(4).

[6] Qian WB, Xu FK, Qian J, Shu WH, Ding WP. Multi-label feature selection based on rough granular-ball and label distribution [J]. INFORMATION SCIENCES, 2023, 650.

[7] Ma XA, Jiang WT, Ling Y, Yang BL. Multi-label feature selection via maximum dynamic correlation change and minimum label redundancy [J]. ARTIFICIAL INTELLIGENCE REVIEW, 2023.

[8] Li Yuchen, Wei Wei, Bai Weiming, et al. Multi-label feature selection based on Label Co-occurrence Relationship [J]. Computer Engineering and Science, 2021, 43(11): 2049-2055.

[9] Weng Wei, Wei Bowen, Ke Wen, Fan Yuling, Wang Jinbo, Li Yuwen. Learning label-specific features with global and local

label correlation for multi-label classification[J]. APPLIED INTELLIGENCE, 2023, 53(3): 3017-3033.

[10] Li Yonghao, Hu Liang, Gao Wanfu. Label correlations variation for robust multi-label feature selection[J]. INFORMATION SCIENCES, 2022, 609: 1075-1097.

[11] WANG C X, LIN Y J, LIU J H. Feature selection for multi-label learning with missing labels[J]. Applied Intelligence, 2019, 49 (8): 3027-3042.

[12] ZHU P F, XU Q, HU Q H, et al. Multi-label feature selection with missing labels [J]. Pattern Recognition, 2018, 74: 488-502.

[13] Zhi-Fen He, Ming Yang, Yang Gao, Hui-Dong Liu, Yilong Yin. Joint multi-label classification and label correlations with missing labels and feature selection[J]. Knowledge-Based Systems, 2019, 163: 145-158.

[14] Zhao DW, Tan Y, Sun D, Gao QW, Lu YX, Zhu D. Multi-label learning of missing labels using label-specific features: an embedded packaging method [J]. APPLIED INTELLIGENCE, 2023.

[15] HAN H R, HUANG M X, ZHANG Y, et al. Multi-label Learning with Label Specific Features Using Correlation Information [J]. IEEE Access, 2019, 7: 11474-11484.

[16] Zhang Zhihao, Lin Yaojin, Lu Shun, et al. Multi-label feature selection based on label-specific features under missing labels [J]. Computer Applications, 2021, 41(10): 2849-2857.

[17] LIN Z C, GANESH A, WRIGHT J, et al. Fast Convex Optimization Algorithms for Exact Recovery of a Corrupted Low-Rank Matrix [C/ OL], 2020.