Noise-Resistant Label Reconstruction Feature Selection for Partial Multi-Label Learning

Wanfu Gao[1,2], **Hanlin Pan**[1,2], **Qingqi Han**[1,2] and **Kunpeng Liu**[3*]
[1]College of Computer Science and Technology, Jilin University, China
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China
[3]Department of Computer Science, Portland State University, Portland, OR 97201 USA
gaowf@jlu.edu.cn, panhl23@mails.jlu.edu.cn, hanqq22@mails.jlu.edu.cn, kunpeng@pdx.edu

Abstract

The "Curse of dimensionality" is prevalent across various data patterns, which increases the risk of model overfitting and leads to a decline in model classification performance. However, few studies have focused on this issue in Partial Multi-label Learning (PML), where each sample is associated with a set of candidate labels, at least one of which is correct. Existing PML methods addressing this problem are mainly based on the low-rank assumption. However, low-rank assumption is difficult to be satisfied in practical situations and may lead to loss of high-dimensional information. Furthermore, we find that existing methods have poor ability to identify positive labels, which is important in real-world scenarios. In this paper, a PML feature selection method is proposed considering two important characteristics of dataset: label relationship's noise-resistance and label connectivity. Our proposed method utilizes label relationship's noise-resistance to disambiguate labels. Then the learning process is designed through the reformed low-rank assumption. Finally, representative labels are found through label connectivity, and the weight matrix is reconstructed to select features with strong identification ability to these labels. The experimental results on benchmark datasets demonstrate the superiority of the proposed method.

1 Introduction

PML [Xie and Huang, 2018] is a recently emerging paradigm of weakly supervised learning aiming to construct a multi-class classifier with uncertain data. Specifically, PML attempts to learn the model from partially labeled samples: a sample is assigned with a candidate label set, and at least one label in the candidate label set is truly related to the sample but the total number of truly related labels is unknown [Sun *et al.*, 2019; Yu *et al.*, 2018]. Compared to multi-label learning [Li *et al.*, 2023a; Gao *et al.*, 2023], PML can better handle situations where labels are missing or ambiguous, which is quite common in real-world scenarios.

Currently, methods for PML can be broadly classified into two categories [Xie and Huang, 2021; Wang *et al.*, 2019; Durand *et al.*, 2019]. The first ones treat all labels in candidate set as correct annotations, ignore noises in labels and directly utilize models for learning such as ML-KNN [Zhang and Zhou, 2007]. The second ones consider the noises in candidate label set and reduce the influence of noises through matrix decomposition and low-rank confidence matrix approximation [Sun *et al.*, 2019].

Figure 1: An example of PML. Among the candidate set of seven labels, only five of them are valid ones (in red).

However, the formers are highly affected by noises, while the latters often rely on the low-rank assumption (i.e. assuming that the low-rank matrix generated after dimension reduction is consistent with the original high-order matrix) and reduce the dimension of the feature/label matrix to reduce the influence of noises. While previous studies have shown that the requirements of the low-rank assumption are relatively strict and often not fully applicable in practical situations [Liu *et al.*, 2015; Xu *et al.*, 2016]. Compared to the original matrix, a low-rank matrix inevitably loses some high-order structural information. In addition, the consistency between the low-rank matrix and the original high-order matrix is also difficult to maintain because of the structural complexity of the original high-dimensional data space.

Another challenge arises from the inherent sparsity of positive labels within multi-label datasets [Liu *et al.*, 2021]. This characteristic of datasets poses a significant hurdle, diminishing the efficacy of many multi-label methods in identifying

*corresponding author

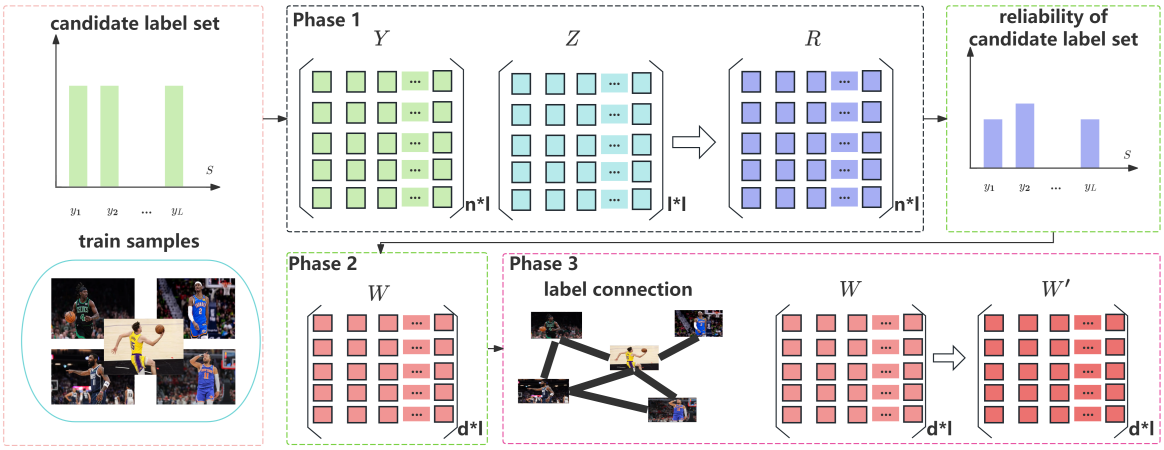


Figure 2: Illustration of PML-FSMIR. In the first stage, label matrix is reconstructed with mutual information matrix to get the reliability of the candidate label set. We reform low-rank assumption to avoid the potential issues in the second stage. Finally in the third stage, weight matrix is reconstructed with the label connection to find the representative labels.

positive labels due to the lack of samples with positive labels. Consequently, the performance of the methods in recognizing positive labels tends to be suboptimal. However, accurate identification of positive labels is important in prediction tasks in various domains such as disease diagnosis [Pham *et al.*, 2022], sentiment analysis [Yilmaz *et al.*, 2021], and gene detection [Wang *et al.*, 2021]. Thus it is necessary for methods capable of addressing this challenge. The sparsity of positive labels demands that we make the most of the available information. Feature selection can enhance the representativeness of the selected features with respect to positive labels, making it easier for the model to capture the underlying pattern.

To address these two challenges, we propose a novel PML feature selection framework based on two crucial characteristics of the datasets. The first one is the noise-resistance of the label relationship: The relationship of the label space is relatively stable and won't be affected by small amount of noise in it. This is because a group-level perspective can smooth out anomalies or noises in individual data points, thereby providing more stable and trustworthy information. Leveraging the relationship information of label space can effectively reduce noisy labels. Another is the label connectivity of label space: Labels in the dataset are not isolated entities but part of a network of interrelations that can significantly influence their importance in the dataset. In this network, the importance of a label is associated with the number of labels it is connected to and the intensity of their connections. The more numerous and stronger these connections, the higher the perceived importance of the labels. If the model can accurately identify an important label, then the model could also identify the highly connected labels accurately. Thus identifying these important labels can effectively enhance the model's precision.

To make use of these two characteristics, mutual information is introduced to measure this structural stability as it can effectively quantifies both linear and nonlinear relationships between variables [Shannon, 1948]. We propose a Partial Multi-label Feature Selection model based on Mutual Information Reconstruction (PML-FSMIR). The whole process is illustrated in Figure 2. Consequently, the of noise-resistance of the label relationship is utilized to reconstruct the label set, numerical values derived from mutual information matrix are employed to represent the reliability of candidate labels. Subsequently, the reconstructed label set is adopted as input for model learning, ensuring that the dimensions of the feature and label spaces remain unchanged throughout the learning process at the same time. Then with the label connectivity, this method reconstructs the learned weight matrix through mutual information matrix. Through this step, we identify important labels and increase their weights to enhance the performance of the model. Experimental validations on datasets from diverse domains affirm the method's efficacy. Our main contributions are:

- We have creatively proposed a method that couples mutual information and sparse learning to make use of collective label characteristics.
- This method breaks free the low-rank assumption commonly used in the past for PML, maintaining the dimensions of the sample space and preserving the high-dimensional information within it.
- Extensive experiments have been conducted on datasets in different fields, and the experimental results have demonstrated the superiority of the model.

2 Related Work

2.1 Multi-label Feature Selection

Over the past years, many traditional multi-label feature selection methods have been proposed. They can be broadly categorized into three main types: filters, wrappers and embeddings. Filter methods select important features by designing a metric, and this process is independent of the subsequent task. Some methods employ mutual information as the metric

[Gao *et al.*, 2018; Zhou *et al.*, 2022]. MDMR combines mutual information with a max-dependency and min-redundancy method to select superior feature subset for multi-label learning [Lin *et al.*, 2015]. Others utilize structural similarity as the metric [Zhang *et al.*, 2021a; Zhang and Gao, 2021; Li *et al.*, 2024]. Others design new evaluation criteria based on existing metrics. Zhang et. al. propose a method that distinguishes three types of label relationships (independence, redundancy and supplementation) and considers changes of label relationships based on different features [Zhang *et al.*, 2021b]. GMM proposes geometric mean to aggregate the mutual information of multiple labels [Gonzalez-Lopez *et al.*, 2020].

Wrapper methods evaluate the performance of feature subsets by training a specific model [Chandra and Bedi, 2021; Rigatti, 2017]. Due to their time-consuming nature and the need for specific model, they are not used as comparative methods in this paper.

Embedding methods assess features by designing specific models and integrate the process of feature selection with learning method. A typical idea is to design a function to carry out the process of learning and feature selection at the same time and then optimize this function [Li *et al.*, 2023b]. MIFS decomposes multi-label information into a low-dimensional space and then employ the reduced space to steer feature selection process [Jian *et al.*, 2016a]. Hu et. al. propose a method utilizes Coupled Matrix Factorization (CMF) to extract the shared common mode between feature matrix and label matrix, considering the comprehensive data information in two matrices [Hu *et al.*, 2020b].

However, the above-mentioned works are unable to select optimal features in partially multi-label datasets as they are built on the assumption that all observed labels are correct.

2.2 Partial Multi-label learning

Partial Multi-label Learning (PML) addresses the issue in which each example is assigned to a candidate label set, and only a part of them is considered correct. To handle PML problem, one approach is to treat the candidate set as ground truth labels directly so that multi-label learning methods can be applied in dealing with PML problem. Nevertheless, these methods are susceptible to being misled by false positive labels concealed within the candidate label set. Hence, there have been proposals for methods specifically tailored to address the challenges of PML. Two effective methods PML-LC and PML-FP are first proposed by estimating a confidence value for each candidate label and training a classifier by optimizing the label ranking confidence matrix [Xie and Huang, 2018]. These methods identify noisy labels by the relationship between labels and features [Li *et al.*, 2021]. Yu et. al. propose a method utilizes a low-rank matrix approximation and latent dependencies between labels and features to identify noisy labels and train a multi-label classifier [Yu *et al.*, 2018]. PARTICAL-MAP and PARTICAL-VLS elicit credible labels from the candidate label set for model induction [Zhang and Fang, 2020]. From the other view, Xu et. al. propose a method leverages the topological information of the feature space and the correlations among the labels to recover label distributions [Xu *et al.*, 2020]. Another view is how to exploit the negative information like the non-candidate set. Li and Wang propose a method that recovers the ground-truth labels by estimating the ground-truth confidences from the label enrichment, which is composed of the relevance degrees of candidate labels and irrelevance degrees of non-candidate labels [Li and Wang, 2020]. PML-LFC estimates the confidence values of relevant labels for each instance using the similarity from both the label and feature spaces, and trains the desired predictor with the estimated confidence values [Yu *et al.*, 2020].

However, the above methods haven't considered the problem of sparse positive labels in the dataset. In order to strengthen the identification ability of positive labels, we propose PML-FSMIR.

3 The Proposed Method

The proposed method consists of three stages: (1) label matrix reconstruction stage aims to reduce noises in labels; (2) reformed low-rank assumption stage aims to exploit cleaned information for feature selection with reformed low-rank assumption; (3) weight matrix reconstruction stage aims to enhance sensitivity to key labels for final feature selection.

3.1 Label Matrix Reconstruction

In the first stage, to reduce the influence of noises in the label set, the label mutual information matrix is introduced to reconstruct the label matrix. Label relationship is more stable than a single label, which means it is less likely to be affected by the noises in the label set. Thus label mutual information matrix which reflects the degree of the label relationship is adopted to reduce the noises.

The input consists of two parts: the feature matrix $X = [x_1, x_2, \dots x_n] \in R^{n\times d}$ where d is the dimension of feature vector and n is the number of training instances and label matrix $Y = [y_1, y_2, \dots y_n] \in [0, 1]^{n\times q}$, where q is the dimension of label vector. In label matrix, $y_{ij} = 1$ means that the j-th label is a candidate label of the i-th instance.

Initially, we compute the mutual information between labels and form them into a q-dimensional square matrix Z:

$$Z_{ij} = I\left(y_{:i}, y_{:j}\right). \tag{1}$$

Z_{ij} denotes the mutual information between i-th and j-th label. As mutual information can quantify relationship between two variables and structure of the label is more stable and less likely to be influenced by noises in single sample, the mutual information between labels can be employed with candidate labels to determine the reliability of candidate labels. Then we can introduce the label reconstruction matrix T:

$$\begin{aligned} T &= (YZ) \circ \operatorname{sign}(Y), \\ \operatorname{sign}\left(y_{ij}\right) &= \begin{cases} 0, & y_{ij} = 0 \\ 1, & y_{ij} = 1 \end{cases} \end{aligned} \tag{2}$$

Simply using the dot product of Y and Z would causes non-candidate labels to be assigned with non-zero values, the $sign(Y)$ is further employed so that non-candidate labels will remain as zero. In this matrix, T_{ij} can be expressed as:

$$T_{ij} = y_{ij} \sum_{k=1}^{q} y_{ik} I\left(y_{ij}, y_{ik}\right) \tag{3}$$

Compared to y_{ij}, T_{ij} comprehensively considers the relationship between labels and the candidate label set in the sample: the more labels in the candidate label set that are highly associated with it, the greater the value. For candidate label y_{ij}, if the labels that are structurally similar to it in the label space are included in the candidate set, then T_{ij} will be large, indicating that the possibility of it being noise is low. Conversely, if the labels similar to it are not in the candidate set, then T_{ij} will be small, suggesting that it is very likely to be noise. After normalization, this matrix T can replace the original matrix Y to the subsequent stages. The effectiveness of this step would be demonstrated in ablation experiments.

3.2 Reformed Low-rank Assumption

In the second stage, we hope to construct the objective function while avoiding the possible problems of the traditional low-rank assumption. For the label matrix Y , we have reconstructed the label matrix T in the first stage for disambiguation, thus no operations is needed in this stage. However, the noises and redundancy issues that exist in X still exist. So we adpot the reformed low-rank assumption to simultaneously remove noises and avoid redundancy and potential issues in low-rank assumption:

$$\min_{U,V,W} \|UVW - T\|_F^2 + \alpha\|X - UV\|_F^2. \tag{4}$$

Where $U \in R^{n\times k}$, $V \in R^{k\times d}$ and $W \in R^{d\times q}$ represent cluster matrix, cluster weight matrix, and feature weight matrix respectively. The traditional low-rank assumption uses matrix decomposition to remove redundancy and noises in the original matrix. We retain this matrix decomposition term, but adopt UV instead of the decomposed U to replace the original X. The dimension of X is preserved, and the goal of removing noises is achieved without losing the high-dimensional structural information of X. In this way, the low-rank assumption is reformed.

To ensure the weight matrix W have the same structure as the original data points. We apply a manifold regularization term to W. Specifically, the mutual information matrix Z' of T is employed to replace the affinity matrix as this matrix is less sensitive to noises. It can be expressed as:

$$Z'_{ij} = I\left(T_{:i}, T_{:j}\right). \tag{5}$$

So the formula is changed as follow:

$$\min_{U,V,W} \|UVW - T\|_F^2 + \alpha\|X - UV\|_F^2 + \beta\,\mathrm{Tr}(W)L_T(W)^T. \tag{6}$$

Where $L_T = A - Z'$ is graph laplacian matrix of T and A is a diagonal matrix. Finally, to achieve feature selection, we further add a $l_{2,1}$-norm of W to Formula 6 [Nie *et al.*, 2010]:

$$\min_{U,V,W} \|UVW - T\|_F^2 + \alpha\|X - UV\|_F^2 + \beta\,\mathrm{Tr}(W)L_T(W)^T + \gamma\|W\|_{2,1}. \tag{7}$$

For optimization, this formula involves three variables W, U and V. After relax the $W_{2,1}$ into $\mathrm{Tr}(W)^T Q(W)$ where

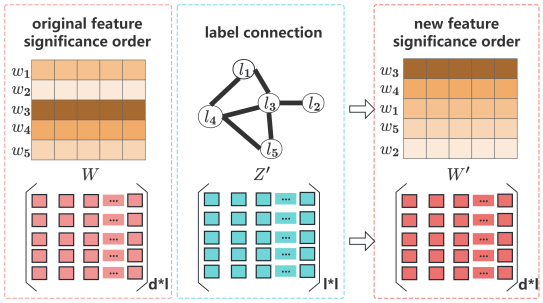


Figure 3: The weight matrix W is reconstructed through the mutual information matrix Z' representing the label connection, and the correct feature weight is obtained.

Q is a diagonal matrix: $Q_{ii} = \frac{1}{2\sqrt{W_i^T W_i + \epsilon}}, (\epsilon \to 0)$. The objective function can be rewritten as:

$$\begin{aligned}\Theta(U,V,W) = &\,\mathrm{Tr}\left((UVW - T)^T(UVW - T)\right) + \\ &\alpha\,\mathrm{Tr}\left((X - UV)^T(X - UV)\right) + \beta\,\mathrm{Tr}\,WL_TW^T + \\ &\gamma\,\mathrm{Tr}\left(W^TQW\right).\end{aligned} \tag{8}$$

Multiplicative gradient descent strategy is adopted to solve Formula 8 [Beck and Teboulle, 2009], in each iteration, each variable is updated while fixing other variables. By taking derivative of Formula 8 based on KKT conditions, we have:

$$U_{ij}^{t+1} \leftarrow U_{ij}^t \frac{\left(TW^TV^T + \alpha XV^T\right)_{ij}}{\left(UVWW^TV^T + \alpha UVV^T\right)_{ij}}. \tag{9}$$

$$V_{ij}^{t+1} \leftarrow V_{ij}^t \frac{\left(U^TTW^T + \alpha U^TX\right)_{ij}}{\left(U^TUVWW^T + \alpha U^TUV\right)_{ij}}. \tag{10}$$

$$W_{ij}^{t+1} \leftarrow W_{ij}^t \frac{\left(V^TU^TT\right)_{ij}}{\left(V^TU^TUVW + \beta W + \delta QW\right)_{ij}}. \tag{11}$$

In real implementation, a small positive value is added to the denominator to avoid zero values.

Algorithm 1 Pseudo code of PML-FSMIR

Input: Feature matrix X and label matrix Y, regularization parameters α, β, and γ.
Output: Return the ranked features.

```
Construct mutual information matrix Z of Y;
Calculate T by Formula 2;
Construct mutual information matrix Z' and graph laplacian matrix of T;
while not coverage do
    Calculate Q;
    Update U by Formula 9 with other variables fixed;
    Update V by Formula 10 with other variables fixed;
    Update W by Formula 11 with other variables fixed;
end while
Reconstruct W by Formula 12;
return Return features according to ‖W_i·‖_2.
```

3.3 Weight Matrix Reconstruction

In partial multi-label datasets, positive labels are often more sparse but more important than negative labels [Liu *et al.*, 2006]. To further enhance the predicting ability of selected features of positive labels, we design this stage using mutual information matrix Z' to reconstruct the weight matrix.

To this end, the keypoint is to identify some representative labels which are related to more labels compared to other labels. If we can find those labels that are highly connected with a large number of labels and select these features that can identify these labels, the experimental results will significantly improve. As shown in Figure 3, since features is considered as equal after the end of the second stage, the feature weights are not correctly ranked as expected. Therefore, the weight matrix needs to be reconstructed using label connection, so that the reconstructed weight matrix can better identify the features that are more important to the representative labels. In weight matrix W, W_{ij} is the weight of the i-th feature to the j-th label, so the sum of the row W_i can be regarded as the importance of the i-th feature to the label set. To enhance the identification ability of the selected feature to the representative labels, we utilize the mutual information between labels to reconstruct W:

$$W_{ij} = \sum_{k=1}^{q} W_{ik} Z'_{kj}. \tag{12}$$

Through this equation, W_{ij}, the weight of i-th feature to j-th label is updated according to the relevance of this labels to other labels. After this operation, those features that have higher weight to representative labels would be selected. According to final value of $\|W_{i\cdot}\|_2 \, (i = 1, \ldots, d)$ in a descending order, the top ranked features are obtained. The pseudo code is presented in Algorithm 1. Subsequent ablation experiments have demonstrated the necessity of this stage.

4 Experiments

4.1 Datasets

We perform experiments on eight datasets from a broad range of applications: ***Birds*** for audio, ***CAL500*** for music classification, ***Corel5K*** for image annotation, ***LLOG_F*** and ***Slashdot*** for text categorization, ***Water*** for chemistry, ***Yeast*** for gene function prediction, and ***CHD49*** for medicine. We keep the noisy level of every dataset at 20%. On each dataset, ten-fold cross-validation is performed where the mean metric values as well as standard deviations are recorded for each compared method. Detailed information is shown in Table 1.

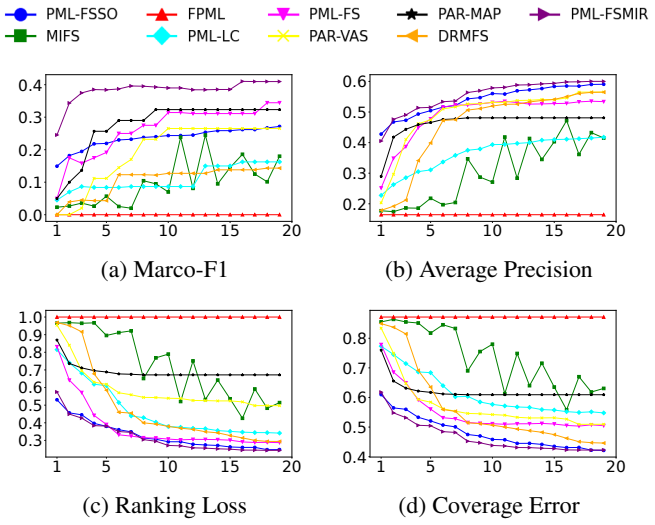
Name	Domain	#Instances	#Features	#Labels
Birds [Briggs *et al.*, 2013]	audio	645	260	19
CAL [Turnbull *et al.*, 2008]	music	555	49	6
CHD_49 [Shao *et al.*, 2013]	medicine	555	49	6
Corel5K [Duygulu *et al.*, 2002]	image	5000	499	374
LLOG_F [Read, 2010]	text	1460	1004	75
Slashdot [Read, 2010]	text	3782	1079	22
Water [Blockeel *et al.*, 1999]	chemistry	1060	16	14
Yeast [Elisseeff and Weston, 2001]	biology	2417	103	14

Table 1: Characteristics of experimental datasets.

4.2 Experimental Setup

Evaluation Metrics: We adopt five widely used multi-label metrics including Ranking Loss, Coverage, Average Precision, Marco-F1, and Micro-F1. For Ranking Loss and Coverage, the smaller value, the better the performance. For Average Precision, Marco-F1, and Micro-F1, the larger the value, the better the performance.

Baselines: We implement eight state-of-the-art methods of multi-label learning and PML for comparison and record the feature selection result with twenty points of different percentages. They are one Partial Multi-label Feature Selection methods (PML-FSSO [Hao *et al.*, 2023]), five Partial Multi-label Learning methods (PML-LC [Xie and Huang, 2018], PML-FP [Xie and Huang, 2018], PAR-VLS [Zhang and Fang, 2020], PAR-MAP [Zhang and Fang, 2020] and FPML [Yu *et al.*, 2018]) and two Multi-label Feature Selection methods (MIFS [Jian *et al.*, 2016b] and DRMFS[Hu *et al.*, 2020a]). Due to the lack of feature selection method in partial multi-label learning, the weight matrix is extracted from model to reflect the importance of features. We adopt ten-fold cross-validation to train these models and the selected features are compared on SVM classifier[1].



(a) Marco-F1 (b) Average Precision

(c) Ranking Loss (d) Coverage Error

Figure 4: Nine methods on ***Birds*** in terms of Marco-F1, Average Precision, Ranking Loss and Coverage..

4.3 Results

The tables below show the detail of the experiment result, for all datasets except ***Water***, we select specified number of features according to the importance as descending order from one to twenty percent for each percentage of the features, the five used metrics for each dataset are recorded in form of mean and standard deviation among different percentages. We also illustrate one dataset for the detail of all metrics in Figure 4 to show our performance clearly. As for ***Water*** only has 16 features, we select 1 to 16 features. From overall results, we make following observations:

[1]The code is available at https://github.com/typsdfgh/PML-FSMIR

Datasets	PML-FSMIR	PML-LC	PML-FP	PAR-VLS	PAR-MAP	FPML	PML-FSSO	MIFS	DRMFS
Birds	**0.32±0.09**	0.47±0.15	0.38±0.15	0.59±0.12	0.69±0.05	1.00±0.00	0.33±0.08	0.73±0.19	0.48±0.23
CAL	**0.31±0.07**	0.63±0.12	0.61±0.15	0.69±0.13	0.72±0.07	0.68±0.09	0.39±0.10	0.58±0.26	0.57±0.17
CHD_49	**0.29±0.11**	0.46±0.16	0.47±0.13	0.34±0.14	0.61±0.06	0.72±0.08	0.31±0.19	0.35±0.12	0.33±0.17
Corel5K	**0.49±0.19**	0.65±0.12	0.64±0.12	0.84±0.08	0.90±0.06	0.92±0.02	0.63±0.019	0.76±0.15	0.71±0.16
LLOG_F	**0.31±0.10**	0.76±0.02	0.69±0.02	0.81±0.05	0.67±0.11	0.53±0.18	0.45±0.11	0.50±0.21	0.63±0.20
Slashdot	**0.05±0.04**	0.49±0.22	0.47±0.21	0.44±0.27	0.43±0.27	1.00±0.00	0.05±0.07	0.34±0.25	0.60±0.23
Water	**0.36±0.01**	0.46±0.06	0.44±0.06	0.48±0.03	0.44±0.03	0.44±0.03	0.39±0.06	0.44±0.03	0.42±0.07
Yeast	**0.24±0.09**	0.43±0.01	0.45±0.02	0.34±0.14	0.62±0.01	0.34±0.12	0.35±0.14	0.24±0.10	0.37±0.16

Table 2: Experimental results (mean ± std) in terms of Ranking Loss where the best performance is shown in boldface.

Datasets	PML-FSMIR	PML-LC	PML-FP	PAR-VLS	PAR-MAP	FPML	PML-FSSO	MIFS	DRMFS
Birds	**0.47±0.05**	0.61±0.07	0.55±0.08	0.57±0.09	0.62±0.04	0.87±0.00	0.48±0.06	0.74±0.10	0.56±0.14
CAL	**0.63±0.02**	0.81±0.01	0.79±0.01	0.74±0.03	0.71±0.02	0.71±0.03	0.66±0.04	0.74±0.06	0.72±0.04
CHD_49	0.50±0.04	0.62±0.04	0.62±0.03	0.66±0.02	0.66±0.02	0.67±0.01	0.49±0.07	0.53±0.04	**0.49±0.07**
Corel5K	**0.37±0.08**	0.51±0.03	0.48±0.03	0.55±0.03	0.59±0.01	0.57±0.01	0.43±0.08	0.51±0.06	0.47±0.07
LLOG_F	**0.53±0.01**	0.51±0.03	0.48±0.03	0.59±0.00	0.56±0.01	0.55±0.02	0.59±0.02	0.57±0.02	0.59±0.00
Slashdot	**0.06±0.01**	0.23±0.06	0.23±0.06	0.19±0.09	0.19±0.08	0.37±0.00	0.07±0.02	0.16±0.08	0.24±0.08
Water	**0.72±0.01**	0.76±0.01	0.75±0.01	0.73±0.00	0.73±0.01	0.74±0.00	0.73±0.04	0.74±0.01	0.75±0.03
Yeast	**0.50±0.03**	0.56±0.08	0.58±0.02	0.56±0.05	0.78±0.00	0.53±0.04	0.58±0.05	0.52±0.07	0.58±0.06

Table 3: Experimental results (mean ± std) in terms of Coverage where the best performance is shown in boldface.

Datasets	PML-FSMIR	PML-LC	PML-FP	PAR-VLS	PAR-MAP	FPML	PML-FSSO	MIFS	DRMFS
Birds	**0.59±0.03**	0.36±0.06	0.49±0.08	0.49±0.10	0.46±0.05	0.16±0.00	0.54±0.05	0.31±0.10	0.43±0.04
CAL	**0.59±0.03**	0.39±0.01	0.41±0.01	0.48±0.05	0.55±0.04	0.55±0.05	0.51±0.04	0.47±0.08	0.43±0.04
CHD_49	**0.77±0.02**	0.68±0.01	0.66±0.01	0.76±0.02	0.71±0.01	0.65±0.02	0.77±0.04	0.75±0.02	0.77±0.03
Corel5K	**0.42±0.08**	0.30±0.03	0.32±0.04	0.29±0.04	0.24±0.02	0.27±0.02	0.37±0.08	0.28±0.05	0.33±0.07
LLOG_F	0.59±0.02	0.47±0.01	0.48±0.01	0.48±0.01	0.57±0.03	**0.60±0.04**	0.46±0.01	0.54±0.03	0.47±0.00
Slashdot	**0.93±0.03**	0.42±0.11	0.43±0.11	0.64±0.23	0.65±0.22	0.17±0.00	0.93±0.05	0.70±0.19	0.49±0.18
Water	0.60±0.01	0.52±0.01	0.53±0.01	0.60±0.01	0.59±0.02	**0.61±0.02**	0.58±0.04	0.53±0.03	0.55±0.05
Yeast	**0.70±0.04**	0.51±0.03	0.52±0.03	0.61±0.04	0.57±0.01	0.63±0.04	0.57±0.04	0.68±0.08	0.58±0.05

Table 4: Experimental results (mean ± std) in terms of Average Precision where the best performance is shown in boldface.

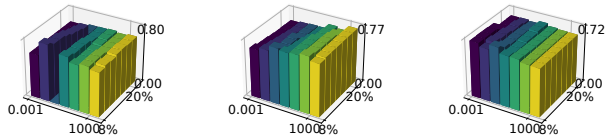
Datasets	PML-FSMIR	PML-LC	PML-FP	PAR-VLS	PAR-MAP	FPML	PML-FSSO	MIFS	DRMFS
Birds	**0.38±0.04**	0.11±0.04	0.26±0.08	0.19±0.01	0.28±0.08	0.00±0.00	0.24±0.03	0.01±0.07	0.11±0.05
CAL	**0.49±0.08**	0.02±0.03	0.07±0.06	0.02±0.00	0.08±0.08	0.10±0.06	0.00±0.00	0.00±0.00	0.00±0.00
CHD_49	**0.56±0.19**	0.27±0.13	0.23±0.11	0.16±0.12	0.24±0.12	0.01±0.01	0.19±0.13	0.13±0.10	0.08±0.11
Corel5K	**0.13±0.08**	0.06±0.02	0.05±0.03	0.04±0.04	0.03±0.01	0.00±0.00	0.00±0.00	0.00±0.00	0.02±0.02
LLOG_F	**0.32±0.03**	0.00±0.00	0.00±0.00	0.03±0.00	0.23±0.11	0.22±0.06	0.10±0.04	0.15±0.06	0.02±0.01
Slashdot	**0.76±0.10**	0.21±0.10	0.14±0.10	0.34±0.21	0.37±0.20	0.00±0.00	0.00±0.00	0.13±0.09	0.05±0.05
Water	**0.48±0.11**	0.32±0.15	0.34±0.17	0.10±0.06	0.17±0.11	0.11±0.07	0.13±0.09	0.44±0.02	0.07±0.09
Yeast	**0.46±0.16**	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.03±0.03	0.02±0.03	0.20±0.15	0.03±0.01

Table 5: Experimental results (mean ± std) in terms of Marco-F1 where the best performance is shown in boldface.

Datasets	PML-FSMIR	PML-LC	PML-FP	PAR-VLS	PAR-MAP	FPML	PML-FSSO	MIFS	DRMFS
Birds	**0.38±0.03**	0.12±0.04	0.27±0.08	0.20±0.10	0.29±0.09	0.00±0.00	0.33±0.09	0.11±0.07	0.11±0.05
CAL	**0.51±0.06**	0.02±0.03	0.07±0.05	0.02±0.00	0.08±0.08	0.12±0.07	0.00±0.00	0.00±0.00	0.00±0.00
CHD_49	**0.60±0.19**	0.31±0.15	0.24±0.12	0.20±0.14	0.26±0.12	0.01±0.01	0.18±0.14	0.16±0.12	0.09±0.13
Corel5K	**0.13±0.08**	0.08±0.02	0.07±0.05	0.05±0.06	0.04±0.02	0.00±0.00	0.00±0.00	0.00±0.00	0.04±0.04
LLOG_F	**0.34±0.02**	0.00±0.00	0.00±0.00	0.03±0.00	0.23±0.12	0.32±0.09	0.08±0.04	0.19±0.07	0.02±0.07
Slashdot	**0.81±0.09**	0.26±0.12	0.21±0.14	0.40±0.23	0.44±0.20	0.00±0.00	0.00±0.00	0.18±0.12	0.06±0.07
Water	**0.49±0.11**	0.33±0.15	0.35±0.18	0.09±0.05	0.15±0.11	0.11±0.07	0.19±0.14	0.46±0.02	0.06±0.09
Yeast	**0.48±0.16**	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.03±0.03	0.23±0.16	0.02±0.01

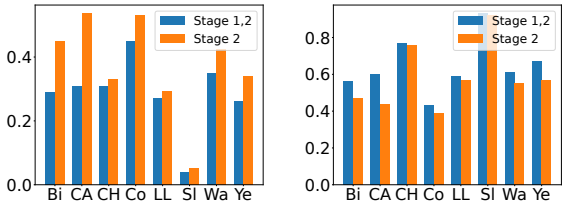
Table 6: Experimental results (mean ± std) in terms of Mirco-F1 where the best performance is shown in boldface.

- On eight datasets across all evaluation metrics, PML-FSMIR ranks first in all cases except Coverage on ***CHD_49*** and Average Precision on ***LLOG_F***, while in these two cases PML-FSMIR all ranks second. These results fully demonstrate the superiority of PML-FSMIR.
- For the superiority in Ranking Loss, Coverage, and Average Precision. We attribute it to the reconstruction of the label matrix through mutual information matrix in the first stage. This step effectively reduces the noises in the labels, making the selected features more helpful for subsequent work. The subsequent ablation experiments further confirmed its function.
- For the superiority in Mirco-F1 and Marco-F1. We attribute it to the reconstruction of the weight matrix through mutual information matrix in the third stage. This step improves the weight of features with strong ability to determine key labels by reconstructing the weight matrix through mutual information matrix. The results in Table 4 and 5 show that this step significantly improves the F1-score, which improves the best results of the baselines at least by 15.2%. The subsequent ablation experiments further confirmed this opinion.



(a) α (b) β (c) γ

Figure 5: Parameter sensitivity studies on the ***CAL*** in terms of Coverage.

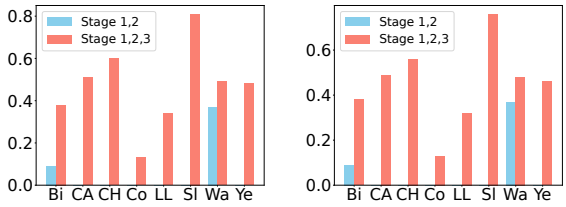


(a) Ranking Loss (b) Average Precision

Figure 6: The results of the comparison experiment between the first two stage and the second stage.

4.4 Parameter Analysis

In PML-FSMIR, there are three parameters α, β, and γ that affect experimental results. Figure 5 shows how these three parameters affect the performance of the model on the ***LLOG_F*** in terms of Ranking Loss. Each parameter is independently tuned from 0.001 to 1000, and we selected the performance of the model when selecting 8% of the features to 20% of the features. From the figure, it can be seen that the model is obviously insensitive to the parameters, which shows the robustness of the model.



(a) Mirco-F1 (b) Marco-F1

Figure 7: The results of the comparison experiment between the first two stage and the whole process.

4.5 Ablation Study

To prove the necessity of reconstructing objective matrix using mutual information matrix in first stage and third stage, we set up two ablation experiments: (1) we employ the first two stage compared to directly conducting the second stage to verify the superiority of the first stage; (2) we employ whole procedure compared to conducting the first two stage to verify the superiority of the third stage. In the first experiments we employ Ranking Loss, Coverage, and Average Precision as evaluate metrics. In the second experiments Mirco-F1 and Marco-F1 are used as evaluate criteria. The result are shown in Figures 6, 7. The Figure 6 proves the effectiveness of first stage. In all the datasets, the results of the first two stage are better than the second stage, which indicates that using mutual information matrix to reconstruct label matrix can effectively reduce the influence of noises in the labels.

The Figure 7 proves the effectiveness of the third stage. It clearly shows the significant improvement of weight matrix reconstruction. The method with the first two stage even cannot identify positive labels in six out of eight datasets while after the third stage it can do so in all datasets. This experiment enhances the correctness of our core idea: there are some labels in the label set that are more important than others. Finding features related to these labels can effectively improve the model's ability to identify positive labels.

5 Conclusion

In this paper, we tackle the challenges of label noise and sparsity in partial multi-label data by introducing a novel three-stage feature selection method PML-FSMIR). This method first reduce the noises in the label sets by reconstructing label matrix using mutual information matrix. Then it trains a weight matrix under a reformed low-rank assumption, which helps to prevent overfitting and ensures a more accurate reflection of the data's underlying structure. Finally the weight matrix is reconstructed to enhance the ability of the selected features in effectively identifying key labels, thereby improving the model's overall performance. Extensive experimental results validate the superiority of our method. In the future, we plan to explore how to exploit mutual information and other methods to reduce noises and improve the identification of positive labels furthermore.

Acknowledgments

This work was supported by the Science Foundation of Jilin Province of China under Grant YDZJ202501ZYTS286, and in part by Changchun Science and Technology Bureau Project under Grant 23YQ05.

References

[Beck and Teboulle, 2009] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[Blockeel *et al.*, 1999] Hendrik Blockeel, Sašo Džeroski, and Jasna Grbović. Simultaneous prediction of multiple chemical parameters of river water quality with tilde. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 32–40. Springer, 1999.

[Briggs *et al.*, 2013] Forrest Briggs, Yonghong Huang, Raviv Raich, Konstantinos Eftaxias, Zhong Lei, William Cukierski, Sarah Frey Hadley, Adam Hadley, Matthew Betts, Xiaoli Z Fern, et al. The 9th annual mlsp competition: New methods for acoustic classification of multiple simultaneous bird species in a noisy environment. In *2013 IEEE international workshop on machine learning for signal processing (MLSP)*, pages 1–8. IEEE, 2013.

[Chandra and Bedi, 2021] Mayank Arya Chandra and SS Bedi. Survey on svm and their application in image classification. *International Journal of Information Technology*, 13(5):1–11, 2021.

[Durand *et al.*, 2019] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 647–657, 2019.

[Duygulu *et al.*, 2002] Pinar Duygulu, Kobus Barnard, Joao FG de Freitas, and David A Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Computer Vision—ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part IV 7*, pages 97–112. Springer, 2002.

[Elisseeff and Weston, 2001] André Elisseeff and Jason Weston. A kernel method for multi-labelled classification. *Advances in neural information processing systems*, 14, 2001.

[Gao *et al.*, 2018] Wanfu Gao, Liang Hu, and Ping Zhang. Class-specific mutual information variation for feature selection. *Pattern Recognition*, 79:328–339, 2018.

[Gao *et al.*, 2023] Wanfu Gao, Pingting Hao, Yang Wu, and Ping Zhang. A unified low-order information-theoretic feature selection framework for multi-label learning. *Pattern Recognition*, 134:109111, 2023.

[Gonzalez-Lopez *et al.*, 2020] Jorge Gonzalez-Lopez, Sebastián Ventura, and Alberto Cano. Distributed multi-label feature selection using individual mutual information measures. *Knowledge-Based Systems*, 188:105052, 2020.

[Hao *et al.*, 2023] Pingting Hao, Liang Hu, and Wanfu Gao. Partial multi-label feature selection via subspace optimization. *Information Sciences*, 648:119556, 2023.

[Hu *et al.*, 2020a] Juncheng Hu, Yonghao Li, Wanfu Gao, and Ping Zhang. Robust multi-label feature selection with dual-graph regularization. *Knowledge-Based Systems*, 203:106126, 2020.

[Hu *et al.*, 2020b] Liang Hu, Yonghao Li, Wanfu Gao, Ping Zhang, and Juncheng Hu. Multi-label feature selection with shared common mode. *Pattern Recognition*, 104:107344, 2020.

[Jian *et al.*, 2016a] Ling Jian, Jundong Li, Kai Shu, and Huan Liu. Multi-label informed feature selection. In *IJCAI*, volume 16, pages 1627–33, 2016.

[Jian *et al.*, 2016b] Ling Jian, Jundong Li, Kai Shu, and Huan Liu. Multi-label informed feature selection. In *International Joint Conference on Artificial Intelligence*, 2016.

[Li and Wang, 2020] Ximing Li and Yang Wang. Recovering accurate labeling information from partially valid data for effective multi-label learning. *arXiv preprint arXiv:2006.11488*, 2020.

[Li *et al.*, 2021] Ziwei Li, Gengyu Lyu, and Songhe Feng. Partial multi-label learning via multi-subspace representation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 2612–2618, 2021.

[Li *et al.*, 2023a] Yonghao Li, Liang Hu, and Wanfu Gao. Multi-label feature selection via robust flexible sparse regularization. *Pattern Recognition*, 134:109074, 2023.

[Li *et al.*, 2023b] Yonghao Li, Liang Hu, and Wanfu Gao. Robust sparse and low-redundancy multi-label feature selection with dynamic local and global structure preservation. *Pattern Recognition*, 134:109120, 2023.

[Li *et al.*, 2024] Yonghao Li, Liang Hu, and Wanfu Gao. Multi-label feature selection with high-sparse personalized and low-redundancy shared common features. *Information Processing & Management*, 61(3):103633, 2024.

[Lin *et al.*, 2015] Yaojin Lin, Qinghua Hu, Jinghua Liu, and Jie Duan. Multi-label feature selection based on max-dependency and min-redundancy. *Neurocomputing*, 168:92–103, 2015.

[Liu *et al.*, 2006] Yi Liu, Rong Jin, and Liu Yang. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *AAAi*, volume 6, pages 421–426, 2006.

[Liu *et al.*, 2015] Meng Liu, Yong Luo, Dacheng Tao, Chao Xu, and Yonggang Wen. Low-rank multi-view learning in matrix completion for multi-label image classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[Liu *et al.*, 2021] Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor W Tsang. The emerging trends of multi-label learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7955–7974, 2021.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding. Efficient and robust feature selection via joint l2, 1-norms minimization. *Advances in neural information processing systems*, 23, 2010.

[Pham *et al.*, 2022] Thuan Pham, Xiaohui Tao, Ji Zhang, Jianming Yong, Yuefeng Li, and Haoran Xie. Graph-based multi-label disease prediction model learning from medical data and domain knowledge. *Knowledge-based systems*, 235:107662, 2022.

[Read, 2010] Jesse Read. *Scalable multi-label classification*. PhD thesis, University of Waikato, 2010.

[Rigatti, 2017] Steven J Rigatti. Random forest. *Journal of Insurance Medicine*, 47(1):31–39, 2017.

[Shannon, 1948] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[Shao *et al.*, 2013] Huan Shao, GuoZheng Li, GuoPing Liu, and YiQin Wang. Symptom selection for multi-label data of inquiry diagnosis in traditional chinese medicine. *Science China Information Sciences*, 56:1–13, 2013.

[Sun *et al.*, 2019] Lijuan Sun, Songhe Feng, Tao Wang, Congyan Lang, and Yi Jin. Partial multi-label learning by low-rank and sparse decomposition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5016–5023, 2019.

[Turnbull *et al.*, 2008] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):467–476, 2008.

[Wang *et al.*, 2019] Haobo Wang, Weiwei Liu, Yang Zhao, Chen Zhang, Tianlei Hu, and Gang Chen. Discriminative and correlative partial multi-label learning. In *IJCAI*, pages 3691–3697, 2019.

[Wang *et al.*, 2021] Qianqian Wang, Jiafeng Cheng, Quanxue Gao, Guoshuai Zhao, and Licheng Jiao. Deep multi-view subspace clustering with unified and discriminative learning. *IEEE Transactions on Multimedia*, 23:3483–3493, 2021.

[Xie and Huang, 2018] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Xie and Huang, 2021] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning with noisy label identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3676–3687, 2021.

[Xu *et al.*, 2016] Chang Xu, Dacheng Tao, and Chao Xu. Robust extreme multi-label learning. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1275–1284, 2016.

[Xu *et al.*, 2020] Ning Xu, Yun-Peng Liu, and Xin Geng. Partial multi-label learning with label distribution. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6510–6517, 2020.

[Yilmaz *et al.*, 2021] Selim F Yilmaz, E Batuhan Kaynak, Aykut Koç, Hamdi Dibeklioğlu, and Suleyman Serdar Kozat. Multi-label sentiment analysis on 100 languages with dynamic weighting for label imbalance. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[Yu *et al.*, 2018] Guoxian Yu, Xia Chen, Carlotta Domeniconi, Jun Wang, Zhao Li, Zili Zhang, and Xindong Wu. Feature-induced partial multi-label learning. In *2018 IEEE international conference on data mining (ICDM)*, pages 1398–1403. IEEE, 2018.

[Yu *et al.*, 2020] Tingting Yu, Guoxian Yu, Jun Wang, and Maozu Guo. Partial multi-label learning with label and feature collaboration. In *Database Systems for Advanced Applications: 25th International Conference, DASFAA 2020, Jeju, South Korea, September 24–27, 2020, Proceedings, Part I 25*, pages 621–637. Springer, 2020.

[Zhang and Fang, 2020] Min-Ling Zhang and Jun-Peng Fang. Partial multi-label learning via credible label elicitation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3587–3599, 2020.

[Zhang and Gao, 2021] Ping Zhang and Wanfu Gao. Feature relevance term variation for multi-label feature selection. *Applied Intelligence*, 51:5095–5110, 2021.

[Zhang and Zhou, 2007] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.

[Zhang *et al.*, 2021a] Ping Zhang, Wanfu Gao, Juncheng Hu, and Yonghao Li. A conditional-weight joint relevance metric for feature relevancy term. *Engineering Applications of Artificial Intelligence*, 106:104481, 2021.

[Zhang *et al.*, 2021b] Ping Zhang, Guixia Liu, Wanfu Gao, and Jiazhi Song. Multi-label feature selection considering label supplementation. *Pattern recognition*, 120:108137, 2021.

[Zhou *et al.*, 2022] Hongfang Zhou, Xiqian Wang, and Rourou Zhu. Feature selection based on mutual information with correlation coefficient. *Applied intelligence*, 52(5):5457–5474, 2022.