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

IJCAI 2025

Background

Partial Multilabel Learning (PML)

Candidate labels

- cloud
- people
- mountain
- tree
- car
- sea
- beach

PML is a recently emerging paradigm of weakly supervised learning aiming to construct a multiclass 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].

Related work

To tackle the mentioned challenge, existing works mainly focus on disambiguation

PML-LC, PML-FP	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.
PARTICAL-MAP, PARTICAL-VLS	PARTICAL-MAP and PARTICAL-VLS elicit credible labels from the candidate label set for model induction.
PML-LFC	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.

Motivation

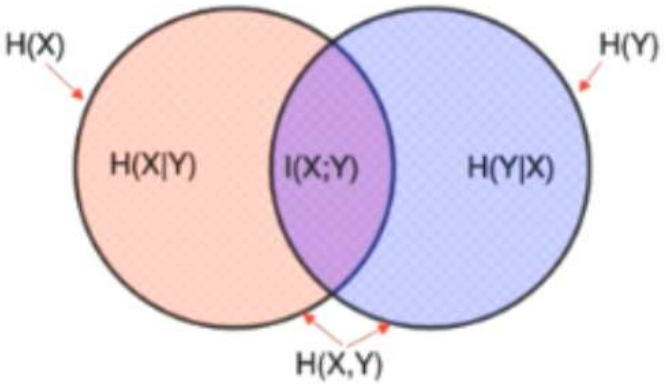
However, the above methods haven’t considered the problem of sparse positive labels in the dataset. In order to strengthen the identification ability of positive labels, we propose PML-FSMIR.

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 highdimensional information within it.
• Extensive experiments have been conducted on datasets in different fields, and the experimental results have demonstrated the superiority of the model.

Mutual Information

Mutual information measures the information shared by two random variables—how much the uncertainty about random variable Y is reduced by knowing random variable X (or vice versa), and is denoted by I(X;Y).

In information theory, entropy is a measure of uncertainty, represented by H. For a random variable X:



$$H(X) = -\sum_x p(x) \log p(x)$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Label Matrix Reconstruction

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

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

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, Then we can introduce the label reconstruction matrix T :

$$T = (YZ) \circ \operatorname{sign}(Y),$$

$$\operatorname{sign}\left(y_{ij}\right) = \begin{cases} 0, & y_{ij} = 0 \\ 1, & y_{ij} = 1 \end{cases}$$

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

The more labels in the candidate label set that are highly associated with it, the greater the value.

Reformed Low-rank Assumption

We adpot the reformed low-rank assumption to simultaneously remove noises and avoid redundancy and potential issues in low-rank assumption:

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

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 dimension of X is preserved, and the goal of removing noises is achieved without losing the highdimensional structural information of X. In this way, the lowrank 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 .

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

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

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.

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

Reformed Low-rank Assumption

Relax the $W_{2,1}$ into $Tr(W)^T Q(W)$ where 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:

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

Multiplicative gradient descent strategy is adopted to solve formula, in each iteration, each variable is updated while fixing other variables. By taking derivative of Formula based on KKT conditions, we have:

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

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

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

Weight Matrix Reconstruction

In partial multi-label datasets, positive labels are often more sparse but more important than negative labels. 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.

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

According to final value of $\|\mathrm{W}_{\mathrm{i}\,.}\|_2 (i = 1, \ldots, d)$ in a descending order, the top ranked features are obtained.

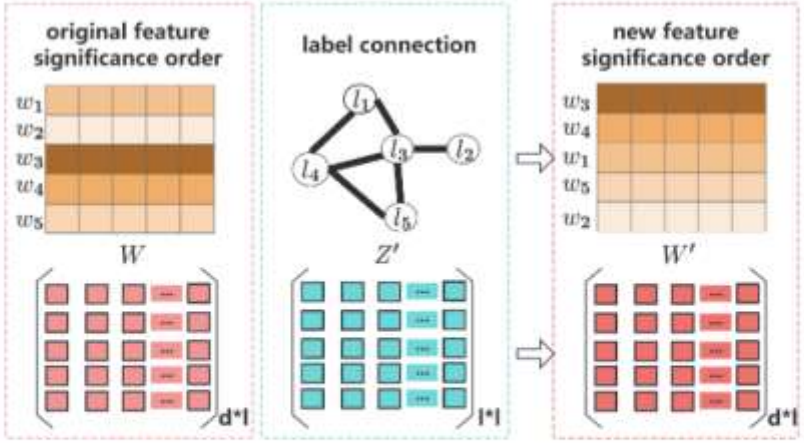


Figure 3: The weight matrix W is reconstructed through the mutual information matrix Z' representing the label connection, and the correct feature weight is obtained.

Method

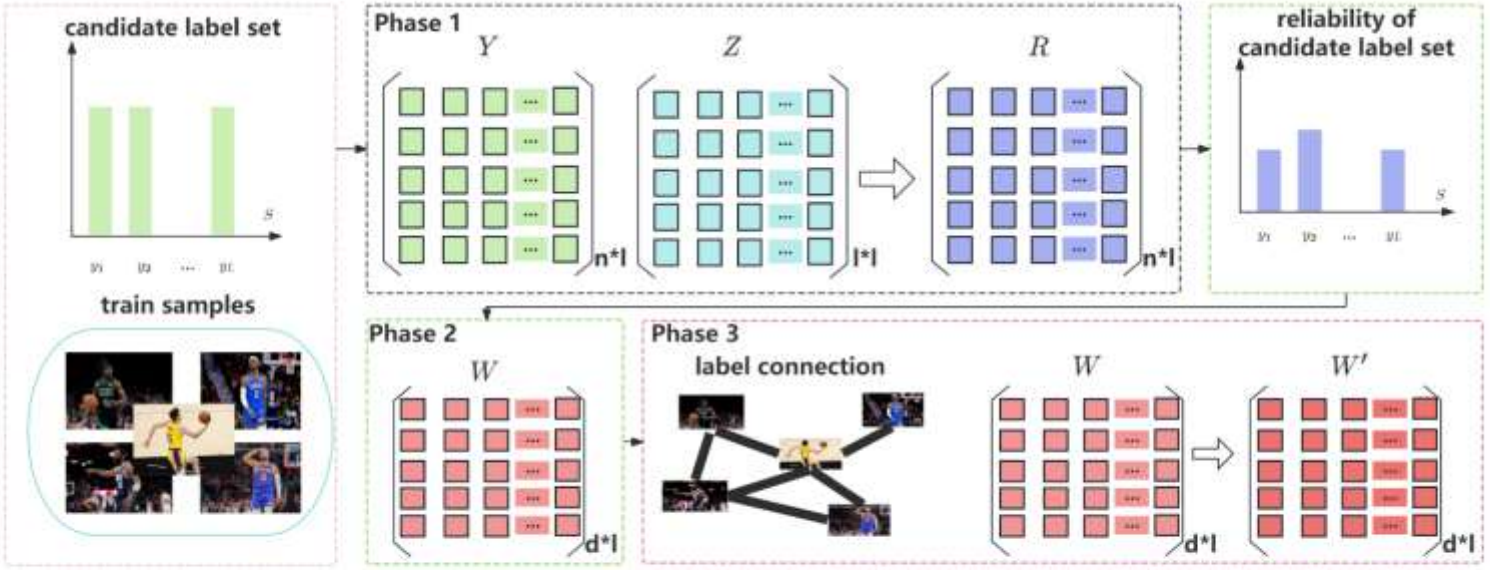


Figure 2: Illustration of PML-FSMIR. In the first stage, label matrix is reconstructed with mutual information matrix to get the reliability of the candidate label set. We reform low-rank assumption to avoid the potential issues in the second stage. Finally in the third stage, weight matrix is reconstructed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

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 $\|W_{i\cdot}\|_2$.

Experiment

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.

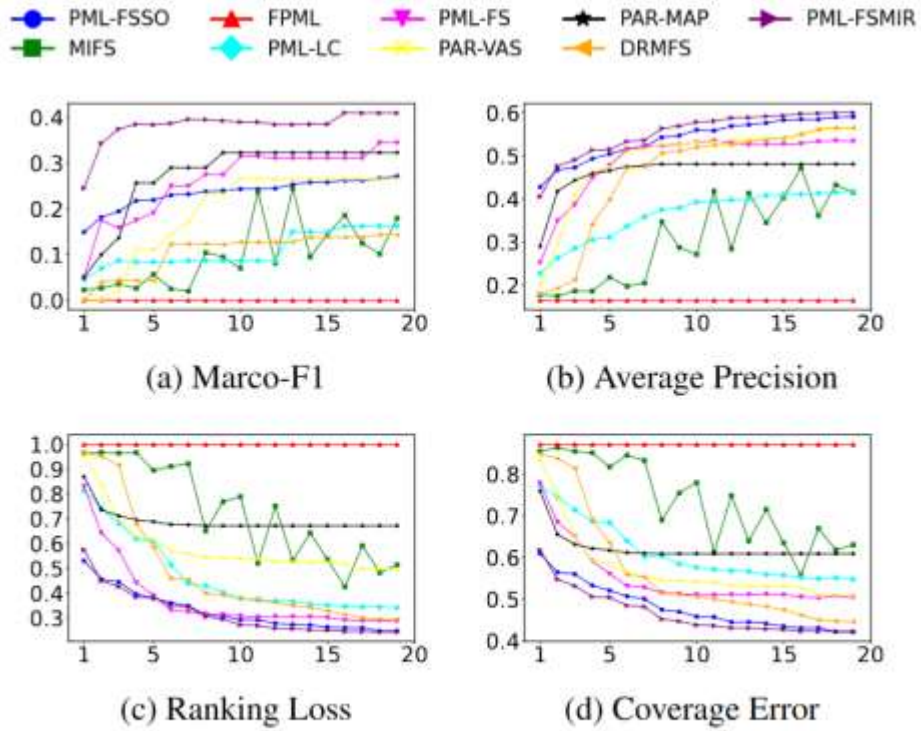
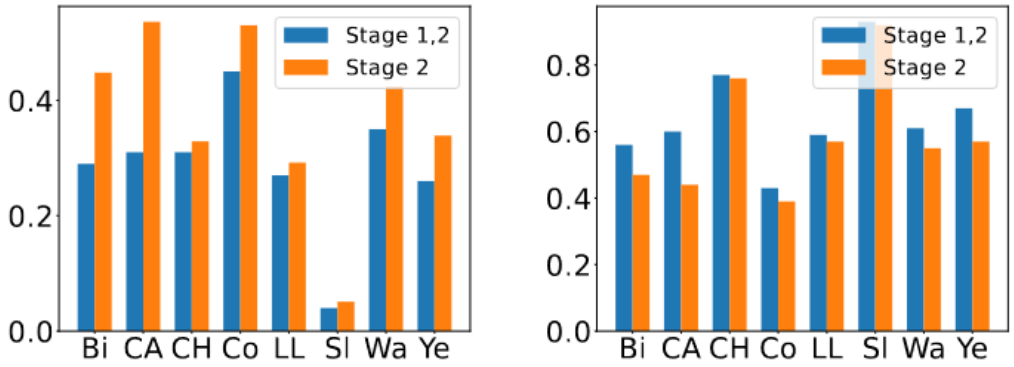


Figure 4: Nine methods on ***Birds*** in terms of Marco-F1, Average Precision, Ranking Loss and Coverage..

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

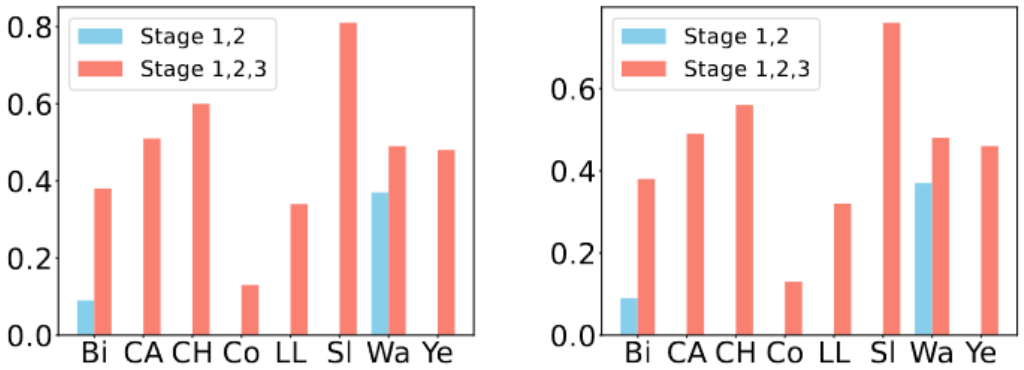
Table 4: Experimental results (mean ± std) in terms of Average Precision where the best performance is shown in boldface.

Experiment



(a) Ranking Loss (b) Average Precision

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



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

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

Thanks