Discriminative and Correlative Partial Multi-Label Learning

Haobo Wang[1,2] , **Weiwei Liu**[3] , **Yang Zhao**[2] , **Chen Zhang**[2] , **Tianlei Hu**[1,2*] and **Gang Chen**[1,2]

[1]Key Lab of Intelligent Computing Based Big Data of Zhejiang Province, Zhejiang University
[2]College of Computer Science and Technology, Zhejiang University
[3]School of Computer Science, Wuhan University
{wanghaobo, awalk, zc99, htl, cg}@zju.edu.cn, liuweiwei863@gmail.com

Abstract

In partial multi-label learning (PML), each instance is associated with a candidate label set that contains multiple relevant labels and other false positive labels. The most challenging issue for the PML problem is that the training procedure is prone to be affected by the labeling noise. We observe that state-of-the-art PML methods are either powerless to disambiguate the correct labels from the candidate labels or incapable of extracting the label correlations sufficiently. To fill this gap, a two-stage *DiscRiminative and correlAtive partial Multi-label leArning* (DRAMA) algorithm is presented in this work. In the first stage, a confidence value is learned for each label by utilizing the feature manifold, which indicates how likely a label is correct. In the second stage, a gradient boosting model is induced to fit the label confidences. Specifically, to explore the label correlations, we augment the feature space by the previously elicited labels on each boosting round. Extensive experiments on various real-world datasets clearly validate the superiority of our proposed method.

1 Introduction

In multi-label learning (MLL) tasks, an object can be associated with multiple labels simultaneously. A lot of recent works have witnessed the rapid development of MLL in many research areas, *e.g.* text categorization [Lin *et al.*, 2018], image/video annotation [Yang *et al.*, 2018], music emotion recognition [Trohidis *et al.*, 2008], and gene function prediction [Fodeh and Tiwari, 2018].

A common assumption in traditional MLL tasks is that the training instances are precisely annotated. Unfortunately, in many real-world applications, it is difficult to obtain noisy-free labels but alternatively, a set of candidate labels are accessible. This scenario is referred as to partial multi-label (PML) learning which is formalized by [Xie and Huang, 2018]. Formally, let $\mathcal{X} = \mathbb{R}^p$ be the d-dimensional feature space and $\mathcal{Y} = \{y_1, y_2, ...y_q\}$ be the q-dimensional label space. Given a PML training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, Y_i)|1 \leq i \leq n\}$ where $\boldsymbol{x}_i \in \mathcal{X}$ is the instance vector and $Y_i \subseteq \mathcal{Y}$ is the candidate label set, the goal of PML is to learn a multi-label predictor $f : \mathcal{X} \mapsto \mathcal{Y}$ from $\mathcal{D}$. PML makes a basic assumption that the ground-truth label set $\hat{Y}_i$ of an instance $\boldsymbol{x}_i$ is concealed in its candidate label set, i.e. $\hat{Y}_i \subseteq Y_i$, and is invisible to the predictor.

The most intuitive way is to regard all the candidate labels as valid ones. Then the PML problem can be solved by any off-the-shelf multi-label learning algorithms, e.g. Binary Relevance (BR) [Zhang and Zhou, 2014], Classifier Chains [Liu *et al.*, 2017], CPLST [Chen and Lin, 2012] and so on. However, such a strategy neglects the false positive labels in the candidate label set, which may lead to insufficient label correlation extraction and in turn the performance degenerates.

In order to tackle this problem, a few PML techniques are proposed. Some methods focus on disambiguation by assigning a confidence value for each candidate label to estimate how likely it is a correct label. For example, Xie and Huang [2018] propose two effective approaches PML-*lc* and PML-*fp* where the confidence scores are calculated by minimizing a confidence-weighted ranking loss. Nonetheless, when the proportion of false positive labels is high, the algorithms are error-prone due to the alternative optimization strategy. PARTICLE [Fang and Zhang, 2019] utilizes the nearest neighbors in the feature space to identify the credible labels with high labeling confidences through an iterative label propagation procedure. Then it applies the pair-wise label ranking technique to induce a multi-label predictor. However, it only extracts second-order label correlations and hence may achieve degenerated performance on complex datasets. fPML [Yu *et al.*, 2018] is another popular PML approach that concentrates on exploring label correlations. It follows the classic label embedding paradigm and can only handle those datasets whose label spaces are highly sparse. We observe that existing PML methods pay attention to either candidate label set disambiguation or label correlation extraction. As a result, their predictive performance is limited.

To bridge this gap, we propose a novel two-stage PML approach named *DiscRiminative and correlAtive partial Multi-label leArning* (DRAMA). In the first stage, we generate a real-valued label confidence matrix under the guidance of feature manifold and the candidate label set. To achieve the goal of disambiguation, we make the *smoothness assumption* [Zhu *et al.*, 2005] that the examples close to each other in the fea-

[*]Corresponding Author.

ture space are prone to share the same labels. Based on the our assumption, the feature and label spaces share a similar local topological structure which could be captured by the sparse reconstruction relationships among each instance and its nearest neighbors. In the second stage, based on the label confidences, we present a gradient boosting algorithm for our proposed multi-output regression problem. In each boosting round, the learned labels are used to augment the feature space, and thus the label correlations can be effectively extracted to improve the generalization performance. Extensive experimental results demonstrate that our proposed method outperforms other state-of-the-art partial multi-label learning algorithms.

The rest of the paper is organized as follows. The next section briefly discusses some related works on partial multi-label learning. The technical details of the proposed DRAMA are presented in the Section 3. Section 4 reports our experimental results on various real-world datasets. Finally, concluding remarks are provided in the last section.

2 Related Work

The goal of partial multi-label learning is to deal with the imprecise tagging problem in multi-label learning. Hence, PML integrates two popular learning framework: multi-label learning and partial label learning.

Partial label learning (PLL) is a weak-supervised multi-class learning framework where each instance is tagged by a set of candidate labels. Note that the relevant label is guaranteed to be contained in the candidate label set and the remaining labels are termed as *false positive labels* or *distractor labels*. To learn from these ambiguous examples, one intuitive strategy is to aggregate the output to optimize the objective such as margin [Nguyen and Caruana, 2008] or likelihood [Jin and Ghahramani, 2002; Cour *et al.*, 2011] over training examples. Problem transformation is another popular strategy and is adopted by many algorithms. For instance, [Zhang *et al.*, 2017; Wu and Zhang, 2018] construct multiple binary label datasets from the original partial label dataset, and then the PLL task is decomposed to a set of binary learning tasks. Other data transformation methods learn label structure from feature space [Zhang and Yu, 2015] to obtain a new numerical dataset whose labels are confidence vectors.

In multi-label learning (MLL), an object can be associated with multiple labels. It has attracted huge attention of researchers from different domains [Yang *et al.*, 2018; Fodeh and Tiwari, 2018]. Binary Relevance [Zhang and Zhou, 2014] is one of the most straightforward solutions for MLL, which aims to decompose the MLL task to a series of independent single label classification problems. Despite its computational efficiency, BR ignores the correlations between labels and thus generally underperforms. To tackle this problem, many MLL algorithms are proposed, such as tree-based [Liu and Tsang, 2017] methods, embedding-based [Yeh *et al.*, 2017; Chen and Lin, 2012] approaches, and augmentation-based [Liu *et al.*, 2017] algorithms.

Obviously, the biggest challenging issue for partial multi-label learning is how to simultaneously disambiguate the correct labels and exploit the label correlations. To address this issue, we propose an effective PML algorithm DRAMA, which marries the concepts of problem transformation in PLL and feature augmentation in MLL.

3 The Proposed Method

In this section, we present the details of DRAMA.

3.1 Candidate Label Set Disambiguation

To explore the underlying structure of feature space, we first build a weighted graph $\mathcal{G} = (V, E, \boldsymbol{W})$ from the given training dataset $\mathcal{D}$. Here $V = \{\boldsymbol{x}_i | 1 \leq i \leq n\}$ denotes the vertex set and $E = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) | i \neq j, \boldsymbol{x}_j \in k\text{NN}(\boldsymbol{x}_i)\}$ denotes the edge set where $k\text{NN}(\boldsymbol{x}_i)$ is the set of $\boldsymbol{x}_i$'s k-nearest neighbors. $\boldsymbol{W} = [w_{ij}]_{n\times n}$ is a non-negative weight matrix where $w_{ij} = 0$ if $(\boldsymbol{x}_i, \boldsymbol{x}_j) \neq E$. For each example $\boldsymbol{x}_i$, we assume that it can be linearly reconstructed from its nearest neighbors and the sparse reconstruction error can be expressed as follows,

$$\mathcal{E}(\boldsymbol{W}) = \sum_{i=1}^{n} ||\boldsymbol{x}_i - \sum_{j\neq i} w_{ij}\boldsymbol{x}_j||_2^2 + ||\boldsymbol{W}||_1 \tag{1}$$

It is noteworthy that the l_1-regularization can actually be regarded as n constraints $\mathbf{1}^\top \boldsymbol{w}_j = 1$ $(1 \leq j \leq n)$. Moreover, since columns in $\boldsymbol{W}$ are independent to one another, we can obtain a series of standard constrained least square programming problems to minimize $\mathcal{E}(\boldsymbol{W})$,

$$\begin{aligned} & \min \boldsymbol{w}_j^\top \boldsymbol{G}^j \boldsymbol{w}_j \\ \text{s.t.} \quad & \mathbf{1}^\top \boldsymbol{w}_j = 1, \quad 1 \leq j \leq n, \\ & W_{ij} \geq 0 \quad (\forall (\boldsymbol{x}_i, \boldsymbol{x}_j) \in E), \\ & W_{ij} = 0 \quad (\forall (\boldsymbol{x}_i, \boldsymbol{x}_j) \notin E) \end{aligned} \tag{2}$$

where $\boldsymbol{w}_j$ is the j-th column vector of $\boldsymbol{W}$. Here $\boldsymbol{G}^j$ is the local Gram matrix for $\boldsymbol{x}_j$ with $G^j_{ab} = (\boldsymbol{x}_j - \boldsymbol{x}_a)^\top(\boldsymbol{x}_j - \boldsymbol{x}_b)$. By solving Eq. (2), the local topological information is embodied in the graph, which can be utilized to disambiguate the partially labeled data. Note that $\boldsymbol{W}$ is usually asymmetric because the importance of $\boldsymbol{x}_i$ in reconstructing $\boldsymbol{x}_j$ is generally different from the inverse case.

According to the smoothness assumption, the feature and label spaces are prone to share the same topological structure. Hence, the feature manifold can be transferred to a numerical label space. Formally, we can reconstruct the labels through the following minimization problem,

$$\begin{aligned} & \min_{\boldsymbol{U}} \sum_{i=1}^{n} ||\boldsymbol{u}_i - \sum_{j\neq i} w_{ij}\boldsymbol{u}_j||_2^2 \\ \text{s.t.} \quad & \sum_{j=1}^{q} \max\{u_{ij}, 0\} \geq 1 \quad (\forall 1 \leq \ i \leq n), \\ & u_{ij} \geq -\delta_1 \quad (\forall 1 \leq i \leq n, y_j \in Y_i), \\ & u_{ij} \leq -\delta_2 \quad (\forall 1 \leq i \leq n, y_j \notin Y_i) \end{aligned} \tag{3}$$

where $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_n] = [u_{ij}]_{n\times n}$ is the transformed label matrix, i.e. the confidence matrix. Here δ_1 and δ_2 are positive constants.

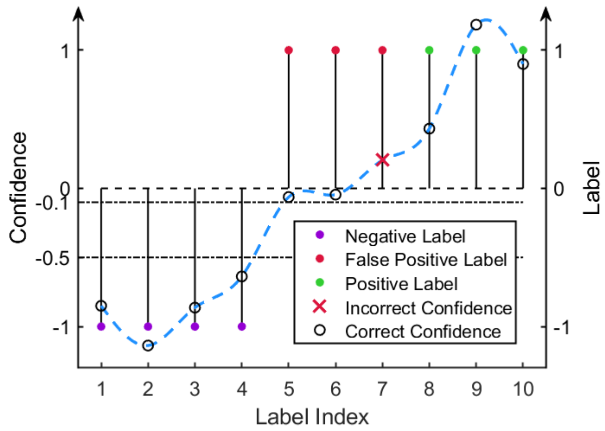

Figure 1: An example of a confidence distribution with 4 negative labels, 3 false positive labels (in red), and 3 positive labels. Here δ_1, δ_2 are set as 0.1, 0.5. Two of the false positive labels obtain a correct confidence because our constraints give those labels with low confidences a chance to be negative. Only one distractor label has a wrong sign of confidence, but the magnitude is relatively small.

It is worth pointing out that Eq. (3) is meticulously designed. The first constraint comes from the multi-label setting, i.e. there might be multiple relevant labels. Then, the original label information is preserved through the subsequent constraints. The sign of each element in $\boldsymbol{U}$ indicates whether the corresponding label is relevant or irrelevant and the magnitude reflects the relative confidence of the relevance. The threshold parameters also play an important role in our algorithm. As illustrated in Figure 1, with a small δ_1, the ground-truth labels can obtain relatively high positive confidences and the false positive labels generally get low or even negative confidences. Moreover, with a relative bigger δ_2, the confidences of irrelevant labels are forced to be negative enough.

Remark that we do not aim to perform dimension reduction [Gao *et al.*, 2018; 2019] in the feature space. The instance matrix and the confidence matrix have different semantic information and are in two independent spaces that merely share the same local topological structure. Furthermore, our disambiguation strategy has three main advantages: 1) all the optimization problems are standard quadratic programming (QP) problems and can be efficiently solved; 2) the logical labels are extended to numerical labels, which helps enrich the original label space; 3) the labels are treated in an unequal manner which prevents the ground-truth labels being overwhelmed by distractor labels.

3.2 Correlative Multi-Label Predictor Inducing

In the second stage, we begin with transforming the training dataset to its disambiguated counterpart $\tilde{\mathcal{D}}_0 = \{(\boldsymbol{x}_i, \boldsymbol{u}_i) | 1 \leq i \leq n\}$. Since our new labels are numerical, we have to treat the learning problem as a multi-output regression problem now. There are a number of effective algorithms proposed, such as multi-regression support vector machines [Chung *et al.*, 2015], metric learning based regressor [Liu *et al.*, 2019] and so on. Nevertheless, most of existing multi-output regressors ignore the correlations among the labels and hence achieve degenerated performance. Consequently, we introduce a novel gradient boosting based model that manipulates the feature space in each boosting round.

We first induce a simple BR regression model $f_0 : \mathcal{X} \mapsto \hat{\mathcal{Y}}$ from $\tilde{\mathcal{D}}_0$ where $\hat{\mathcal{Y}}$ denotes the numerical label space. Any off-the-shelf base regressor can be utilized, such as classification and regression tree (CART) [Liaw *et al.*, 2002], support vector regressor (SVR) [Chung *et al.*, 2015] and many others. Then we can obtain a predicted label matrix $\hat{\boldsymbol{U}}^0 = f_0(\boldsymbol{X})$ where $\boldsymbol{X}$ is the instance matrix. Note that such a simple BR model usually underperforms for two reasons: 1) the generalization performance of BR is restricted to the base classifiers; 2) the label correlations are neglected. Thus, we boost it by incorporating a set of correlation aware regressors.

Formally, our goal is to induce a regressor F that minimizes the following loss function,

$$\mathcal{L}(F(\boldsymbol{X}), \boldsymbol{U}) = \frac{1}{2} ||F(\boldsymbol{X}) - \boldsymbol{U}||_{\mathcal{F}}^2 \tag{4}$$

where $|| \cdot ||_{\mathcal{F}}$ is the Frobenius norm. In order to find the optimal solution, we adopt an ensemble model by adding weak learners to the BR model using a gradient descent like procedure. Specifically, in t-th boosting round, a weak regressor f_t is trained to fit the negative gradient of $\mathcal{L}$ at $F_{t-1}(\boldsymbol{X})$,

$$\boldsymbol{R}^t = - \frac{\partial \mathcal{L}(F(\boldsymbol{X}), \boldsymbol{U})}{\partial F(\boldsymbol{X})} \bigg|_{F=F_{t-1}} = \boldsymbol{U} - F_{t-1}(\boldsymbol{X}) \tag{5}$$

where $\boldsymbol{R}^t = [\boldsymbol{r}_1^t, \boldsymbol{r}_2^t, ..., \boldsymbol{r}_n^t]$. To exploit the label correlations, we further augment the original feature space using previously learned labels $\hat{\boldsymbol{U}}^{t-1} = [\hat{\boldsymbol{u}}_1^{t-1}, \hat{\boldsymbol{u}}_2^{t-1}, ..., \hat{\boldsymbol{u}}_n^{t-1}] = F_{t-1}(\boldsymbol{X})$. In other words, each weak regressor $f_t(\boldsymbol{X}, \hat{\boldsymbol{U}}^{t-1})$ is induced from the following training dataset,

$$\tilde{\mathcal{D}}_t = \{(\tilde{\boldsymbol{x}}_i^t, \boldsymbol{r}_i^t) | 1 \leq i \leq n\}, \quad \tilde{\boldsymbol{x}}_i^t = [\boldsymbol{x}_i, \hat{\boldsymbol{u}}_i^{t-1}] \tag{6}$$

Finally, we can sum up all the weak learners to obtain a robust model,

$$F_T(\boldsymbol{X}) = f_0(\boldsymbol{X}) + \sum_{t=1}^{T-1} \lambda_t f_t(\boldsymbol{X}, \hat{\boldsymbol{U}}^{t-1}) \tag{7}$$

where T is the number of iterations. Here λ_t is the learning rate and can be calculated by,

$$\lambda_t = \operatorname*{argmin}_{\lambda} \frac{1}{2} ||\boldsymbol{U}^{t-1} + \lambda f_t(\boldsymbol{X}, \hat{\boldsymbol{U}}^{t-1}) - \boldsymbol{U}||_{\mathcal{F}}^2 \tag{8}$$

In this work, we choose CART as our boosting weak learners. Since CART is a non-linear model, even complex label correlations can be explored. When an unseen instance x^* is given, we can feed it in to F and then take the sign of the real-valued outputs to get logical labels.

It is worth noting that our model has two main superiorities: 1) the learned weak regressors can help improve the generalization ability of our simple BR model; 2) the boosting procedure is in a coarse-to-fine prediction manner and thus the label correlations can be effectively exploited as the iterations proceed.

Datasets	#Examples	#Features	#Labels	#avg.CLs[†]	#avg.GLs[‡]	Domain
Cal500	500	68	174	27,30,60,90	26.04	music
Emotions	593	72	6	2,3,4	1.87	music
Image	2,000	294	5	2,3,4	1.24	image
Scene	2,407	294	6	2,3,4	1.07	image
Slashdot	3,782	1,079	22	2,3,4	1.18	text

[†] Average number of candidate labels. Each number corresponds to a synthesized PML dataset.
[‡] Average number of ground-truth labels.

Table 1: Characteristics of the experimental datasets.

4 Experiments

4.1 Settings

Datasets

Since PML is a newly proposed framework and there are no public partial multi-label learning datasets available yet, we synthesized a number of PML datasets from 5 multi-label datasets. These datasets are collected from various real-world tasks: Image [Fang and Zhang, 2019] and Scene [Boutell *et al.*, 2004] for image annotation, Slashdot [Read *et al.*, 2011] for text categorization, Cal500 [Turnbull *et al.*, 2008] and Emotions [Trohidis *et al.*, 2008] for music classification. For each example, we randomly select some irrelevant labels and aggregate them with ground-truth labels to obtain a candidate label set. Finally, a total of 16 datasets are synthesized whose characteristics are reported in Table 1. All the datasets are randomly partitioned to 80% for training and the rest for testing. We conduct all the experiments for 5 times and the mean metric values with standard deviations are reported.

Comparison Approaches

We compare our proposed method with two well-established multi-label classifiers and three partial multi-label methods:

- Binary Relevance (BR) [Zhang and Zhou, 2014]: BR is a classical one-vs-all MLL approach that breaks the original problem into binary classification tasks.
- CPLST [Chen and Lin, 2012]: it is a typical label embedding approach in MLL, which integrates the concepts of principal component analysis and canonical correlation analysis.
- PARTICLE [Fang and Zhang, 2019]: it transforms the PML task to a multi-label problem through a label propagation procedure. Then a calibrated label ranking model is induced to instantiate two effective PML methods P-VLS and P-MAP.
- PML-kNN: we further induce a k-nearest neighbor model from the PML datasets and an averaging strategy is adopted to obtain the predictions.

To validate the superiority of our gradient boosting procedure, we further propose a Naïve-DRAMA (N-DRAMA) method that uses only a simple BR model in the second stage.

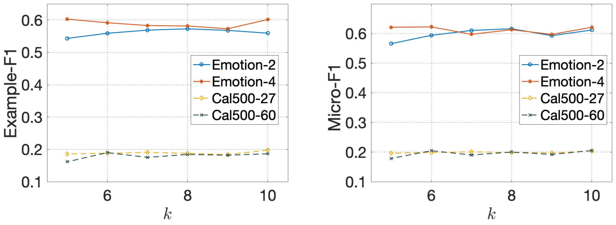

Figure 2: Performance of DARAMA changes as the number of nearest neighbors k changes from 5 to 10 on 4 datasets.

In this paper, δ_1, δ_2 are empirically set as 0.01, 1 for Cal500 and 0.01, 0.5 for other datasets. The QP problems are solved by interior-point method. For BR and our models, we employ Scikit-learn's [Pedregosa *et al.*, 2011] implementation of SVM (for BR), SVR (for f_0), and CART (for gradient boosting) as the base learners. Specifically, Gaussian kernel is equipped by SVM and SVR to handle non-linear separable cases. The number of boosting rounds is fixed to 10. k is set as 10 for all the nearest neighbor based algorithms. For CPLST, we take the first 5 principal components. Following the experimental setting in [Fang and Zhang, 2019], we set $thr = 0.9$ and $\alpha = 0.95$ for PARTICLE.

Evaluation Metrics

Inspired by [Trohidis *et al.*, 2008; Lin *et al.*, 2018], we consider three widely-used multi-label learning metrics to evaluate the predictive performance of all the comparing methods, i.e. Micro-F1, Macro-F1 and Example-F1.

4.2 Results

Table 3 lists the performance of each comparing algorithm in terms of Micro-F1, Macro-F1 and Example-F1. Figure 2 illustrates the parameter sensitivity of DRAMA w.r.t. the number of nearest neighbor k. From the results, we conclude that:

- The proposed DRAMA generally achieves the best performance. For instance, on Emotions with 2 average candidate labels, in terms of Micro-F1, Macro-F1, Example-F1, DRAMA improves the best results of the baselines (except N-DRAMA) by 5.52%, 3.45%, 9.83% respectively. The results listed above validate the superiority of the proposed method.
- BR and PML-kNN is inferior to other methods due to the ignorance of label correlations. CPLST addresses this issue by label embedding and obtains excellent results on Emotions and CAL500 datasets. However, it is prone to be misled by false positive labels and thus underperforms on other datasets.
- N-DRAMA works well on many datasets, which proves the effectiveness of our disambiguation strategy. Nev-

Datasets	avg.#CLs	Micro-F1						
		DRAMA	N-DRAMA	P-VLS	P-MAP	CPLST	PML-kNN	BR
Image	2	**.7489**±.0037	.6480±.0061	.6386±.0045	.6616±.0087	.6591±.0036	.5766±.0026	.7328±.0055
	3	**.6658**±.0055	.6272±.0040	.6301±.0045	.6251±.0100	.4976±.0019	.5168±.0122	.5129±.0046
	4	**.5691**±.0092	.4761±.0063	.4522±.0154	.5584±.0029	.4074±.0014	.3967±.0068	.4070±.0022
Scene	2	**.8250**±.0021	.6659±.0037	.7666±.0044	.8023±.0058	.7004±.0081	.7144±.0050	.7463±.0019
	3	.7469±.0041	.6810±.0043	.7345±.0096	**.7633**±.0085	.5006±.0036	.6760±.0067	.5399±.0038
	4	**.6954**±.0172	.5352±.0036	.6683±.0057	.6873±.0139	.3489±.0011	.6054±.0077	.3899±.0018
Slashdot	2	.7337±.0023	.6145±.0011	.1721±.0863	.3169±.0304	.6073±.0035	.1711±.0029	**.7411**±.0034
	3	**.6635**±.0041	.6188±.0068	.1797±.0827	.3050±.0435	.5722±.0036	.1751±.0190	.6035±.0032
	4	.6122±.0086	**.6218**±.0061	.1808±.0902	.3206±.0129	.5348±.0053	.1681±.0075	.5075±.0031

Datasets	avg.#CLs	Macro-F1						
		DRAMA	N-DRAMA	P-VLS	P-MAP	CPLST	PML-kNN	BR
Image	2	**.7528**±.0047	.6490±.0069	.6415±.0039	.6658±.0089	.6596±.0035	.5785±.0026	.7345±.0059
	3	**.6672**±.0068	.6290±.0039	.6346±.0054	.6285±.0106	.4956±.0018	.5159±.0129	.5117±.0042
	4	**.5691**±.0097	.4747±.0063	.4607±.0235	.5582±.0057	.4063±.0015	.3940±.0061	.4059±.0021
Scene	2	**.8305**±.0022	.6508±.0040	.7757±.0050	.8109±.0057	.7071±.0073	.7220±.0048	.7474±.0019
	3	.7503±.0040	.6870±.0047	.7441±.0100	**.7707**±.0088	.5004±.0034	.6809±.0063	.5387±.0038
	4	**.7048**±.0182	.5392±.0032	.6850±.0089	.6829±.0152	.3487±.0013	.6048±.0090	.3888±.0017
Slashdot	2	.5332±.0013	.3196±.0075	.0936±.0555	.1933±.0177	.1884±.0012	.0390±.0021	**.5591**±.0022
	3	**.4705**±.0010	.3296±.0135	.0737±.0310	.1910±.0352	.1755±.0027	.0368±.0087	.4499±.0029
	4	**.4360**±.0120	.3454±.0079	.0764±.0419	.1909±.0201	.1652±.0052	.0363±.0053	.3796±.0020

Datasets	avg.#CLs	Example-F1						
		DRAMA	N-DRAMA	P-VLS	P-MAP	CPLST	PML-kNN	BR
Image	2	**.7392**±.0087	.5903±.0088	.6035±.0058	.6784±.0071	.6492±.0044	.5879±.0024	.7344±.0091
	3	**.6781**±.0067	.6433±.0033	.6107±.0076	.6425±.0102	.5051±.0030	.5250±.0126	.5237±.0058
	4	**.5909**±.0084	.4835±.0064	.4455±.0251	.5723±.0044	.4028±.0015	.4012±.0067	.4018±.0019
Scene	2	**.8164**±.0058	.5480±.0032	.7357±.0056	.8131±.0064	.6951±.0078	.7191±.0053	.7838±.0025
	3	.7663±.0056	.7159±.0055	.7060±.0084	**.7790**±.0093	.5165±.0036	.6805±.0065	.5770±.0043
	4	**.7223**±.0164	.5700±.0045	.6607±.0093	.6945±.0161	.3529±.0014	.6090±.0076	.4043±.0018
Slashdot	2	.7285±.0089	.5159±.0058	.1425±.1052	.3626±.0381	.4996±.0041	.1797±.0032	**.7561**±.0037
	3	**.6651**±.0076	.5279±.0093	.1777±.0905	.3512±.0475	.4831±.0055	.1838±.0198	.6311±.0040
	4	**.6175**±.0082	.5369±.0065	.1761±.0949	.3693±.0152	.4704±.0066	.1764±.0076	.5414±.0044

Table 2: Transductive performance (mean±standard deviation) of all the methods on 9 synthesized datasets. The best ones are in bold.

ertheless, DRAMA is better than N-DRAMA, because it further considers high-order label correlations.

- It is worth noting that P-VLS and P-MAP are competitive to other methods, but underperform our DRAMA. Since PARTICLE only models second-order label dependencies, its generalization ability is limited.
- Our method is relatively stable with varying values of k.

4.3 Further Analysis

In addition, we evaluate all the methods on six synthesized PML datasets in the transductive setting, i.e. there is no testing data. The transductive performance of a PML algorithm manifests its disambiguating ability to the candidate label sets. The experimental results are reported in Table 2. We can observe that our method is the most successful on all datasets. For example, on Image with 3 average candidate labels, in terms of Micro-F1, Macro-F1, Example-F1, DRAMA improves the best results of the baselines (except N-DRAMA) by 5.7%, 5.1%, 5.4% respectively.

5 Conclusion

This paper focuses on the challenging problem of disambiguation and label correlation extraction in partial multi-label learning. We propose a novel two-stage *DiscRiminative and correlAtive partial Multi-label leArning* (DRAMA) algorithm that marries the concepts of problem transformation in PLL and feature augmentation in MLL. The resultant algorithm firstly disambiguates the candidate label sets by exploring the feature and label manifolds. Then we induce a gradient boosting regressor to utilize the elicited label information. In each boosting round, the original feature space is augmented by the elicited labels such that the high-order label correlations are exploited. Our empirical studies on a range of real-world datasets demonstrate that DRAMA can effectively handle PML tasks.

Acknowledgments

This work is supported by National Key R&D Program of China (2017YFB1201001). It is completed when the authors Yang Zhao and Chen Zhang are in inAI Lab, ZJUINA.

Datasets	avg.#CLs	Micro-F1						
		DRAMA	N-DRAMA	P-VLS	P-MAP	CPLST	PML-kNN	BR
Cal500	27	**.4027**±.0118	.4025±.0054	.0027±.0016	.0726±.0041	.3576±.0157	.0576±.0010	.3434±.0144
	30	**.3846**±.0205	.3845±.0152	.0088±.0007	.0694±.0061	.3602±.0083	.0565±.0010	.3340±.0108
	60	.3715±.0849	.3688±.0622	.0006±.0017	.0716±.0022	**.4529**±.0200	.0498±.0038	.2885±.0079
	90	**.3948**±.0033	.3908±.0091	.0011±.0005	.0672±.0029	.3723±.0048	.0445±.0034	.3551±.0040
Emotions	2	**.6860**±.0141	.6812±.0158	.5903±.0418	.5996±.0314	.6432±.0128	.4938±.0237	.6501±.0247
	3	**.6400**±.0188	.6293±.0213	.5788±.0424	.6082±.0091	.6274±.0120	.4734±.0208	.5834±.0277
	4	**.6212**±.0232	.6207±.0234	.5867±.0136	.5943±.0253	.5324±.0236	.4501±.0227	.5437±.0123
Image	2	**.6223**±.0170	.6101±.0087	.5513±.0762	.6137±.0233	.5311±.0140	.5669±.0150	.6016±.0310
	3	**.5878**±.0143	.5696±.0161	.5132±.0905	.5843±.0311	.4609±.0143	.5220±.0176	.4677±.0107
	4	.4956±.0211	.4406±.0114	.3690±.1548	**.5319**±.0109	.4023±.0045	.3946±.0167	.3962±.0054
Scene	2	**.7228**±.0142	.7130±.0124	.6736±.0473	.7140±.0173	.6058±.0093	.7053±.0098	.6101±.0143
	3	**.6692**±.0178	.6396±.0161	.6333±.0457	.6660±.0277	.4583±.0143	.6628±.0303	.4663±.0118
	4	.5759±.0168	.5028±.0115	.5828±.0724	**.6331**±.0154	.3438±.0052	.5923±.0170	.3546±.0029
Slashdot	2	.5262±.0164	**.5461**±.0256	.3950±.0655	.2658±.0075	.4350±.0094	.1706±.0262	.4086±.0143
	3	**.5250**±.0124	.4649±.0219	.0814±.0690	.2810±.0072	.4183±.0185	.1588±.0065	.3060±.0127
	4	**.5195**±.0136	.5186±.0189	.1434±.0787	.2626±.0140	.3752±.0079	.1657±.0203	.2674±.0114

Datasets	avg.#CLs	Macro-F1						
		DRAMA	N-DRAMA	P-VLS	P-MAP	CPLST	PML-kNN	BR
Cal500	27	**.1852**±.0113	.1809±.0033	.0009±.0005	.0070±.0012	.0819±.0074	.0087±.0004	.1657±.0098
	30	**.1809**±.0118	.1697±.0130	.0021±.0002	.0065±.0016	.0851±.0039	.0084±.0007	.1682±.0083
	60	**.1996**±.0840	.1994±.0343	.0003±.0007	.0069±.0008	.1489±.0111	.0092±.0008	.1903±.0045
	90	.1901±.0043	.1228±.0057	.0010±.0006	.0059±.0005	.2251±.0050	.0108±.0015	**.2301**±.0055
Emotions	2	**.6559**±.0189	.6491±.0206	.5561±.0519	.5619±.0376	.6303±.0140	.4279±.0225	.6340±.0268
	3	**.6213**±.0226	.6072±.0245	.5506±.0460	.5650±.0276	.6128±.0160	.4213±.0190	.5727±.0270
	4	**.6152**±.0214	.6143±.0202	.5726±.0254	.5602±.0314	.5265±.0225	.4106±.0253	.5364±.0132
Image	2	**.6244**±.0176	.6151±.0106	.3561±.1273	.6171±.0247	.5304±.0140	.5652±.0151	.6028±.0311
	3	**.5867**±.0130	.5673±.0156	.3502±.1108	.5862±.0302	.4581±.0136	.5202±.0173	.4641±.0108
	4	.4973±.0214	.4410±.0112	.2972±.1260	**.5344**±.0095	.4011±.0043	.3945±.0160	.3948±.0049
Scene	2	**.7324**±.0168	.7075±.0133	.4119±.0367	.7226±.0159	.6143±.0106	.7148±.0065	.6163±.0168
	3	**.6791**±.0154	.6415±.0137	.4037±.0353	.6679±.0319	.4589±.0148	.6672±.0266	.4664±.0123
	4	.5811±.0151	.5052±.0106	.3863±.0290	**.6415**±.0150	.3412±.0053	.5918±.0173	.3531±.0028
Slashdot	2	**.3428**±.0098	.2655±.0033	.2635±.0082	.1201±.0131	.1335±.0053	.0315±.0076	.2317±.0125
	3	**.2875**±.0102	.2031±.0091	.0248±.0089	.1636±.0056	.1273±.0077	.0281±.0086	.2099±.0101
	4	**.3002**±.0112	.2223±.0053	.0325±.0154	.1478±.0124	.1232±.0078	.0275±.0073	.1901±.0085

Datasets	avg.#CLs	Example-F1						
		DRAMA	N-DRAMA	P-VLS	P-MAP	CPLST	PML-kNN	BR
Cal500	27	**.4033**±.0106	.4033±.0058	.0028±.0016	.0759±.0047	.3566±.0144	.0596±.0013	.3380±.0135
	30	**.3816**±.0178	.3815±.0144	.0103±.0005	.0721±.0062	.3590±.0096	.0586±.0016	.3316±.0109
	60	.3709±.0778	.3669±.0606	.0004±.0010	.0742±.0018	**.4499**±.0205	.0519±.0040	.2870±.0082
	90	.3949±.0036	**.3968**±.0084	.0011±.0005	.0702±.0034	.3705±.0049	.0462±.0031	.3524±.0039
Emotions	2	**.6806**±.0158	.6761±.0173	.5355±.0505	.5954±.0258	.5974±.0179	.4989±.0305	.6197±.0221
	3	**.6195**±.0291	.6064±.0311	.5115±.0605	.6090±.0136	.6137±.0139	.4778±.0203	.5688±.0311
	4	**.6018**±.0260	.6012±.0273	.5454±.0271	.5850±.0319	.5262±.0258	.4542±.0216	.5329±.0108
Image	2	**.6428**±.0225	.6145±.0055	.5251±.0959	.6273±.0216	.5176±.0163	.5758±.0168	.5979±.0277
	3	**.6021**±.0155	.5835±.0162	.4869±.1154	.5970±.0346	.4617±.0139	.5311±.0162	.4716±.0104
	4	.5096±.0211	.4398±.0136	.3504±.1669	**.5451**±.0120	.3965±.0035	.3988±.0166	.3896±.0054
Scene	2	**.7324**±.0098	.6734±.0177	.6268±.0563	.7228±.0169	.5985±.0088	.7114±.0096	.6222±.0157
	3	**.7019**±.0187	.6673±.0171	.5961±.0522	.6773±.0302	.4722±.0170	.6675±.0310	.4875±.0121
	4	.6142±.0210	.5270±.0157	.5587±.0888	**.6446**±.0183	.3466±.0062	.5960±.0179	.3623±.0038
Slashdot	2	**.5326**±.0064	.4694±.0058	.4040±.0756	.2933±.0152	.3588±.0054	.1802±.0271	.3847±.0197
	3	**.5041**±.0087	.3509±.0096	.0644±.0781	.3164±.0125	.3660±.0161	.1661±.0066	.3321±.0137
	4	**.5203**±.0049	.4154±.0073	.1396±.0855	.2989±.0193	.3455±.0119	.1736±.0194	.2994±.0160

Table 3: Prediction performance (mean±standard deviation) of all the methods on 16 synthesized datasets. The best ones are in bold.

References

[Boutell *et al.*, 2004] Matthew R. Boutell, Jiebo Luo, Xipeng Shen, and Christopher M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

[Chen and Lin, 2012] Yao-Nan Chen and Hsuan-Tien Lin. Feature-aware label space dimension reduction for multi-label classification. In *NIPS*, pages 1538–1546, 2012.

[Chung *et al.*, 2015] Wooyong Chung, Jisu Kim, Heejin Lee, and Euntai Kim. General dimensional multiple-output support vector regressions and their multiple kernel learning. *IEEE Trans. Cybernetics*, 45(11):2572–2584, 2015.

[Cour *et al.*, 2011] Timothée Cour, Benjamin Sapp, and Ben Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12:1501–1536, 2011.

[Fang and Zhang, 2019] Jun-Peng Fang and Min-Ling Zhang. Partial multi-label learning via credible label elicitation. In *AAAI*, 2019.

[Fodeh and Tiwari, 2018] Samah Jamal Fodeh and Aditya Tiwari. Exploiting MEDLINE for gene molecular function prediction via NMF based multi-label classification. *Journal of Biomedical Informatics*, 86:160–166, 2018.

[Gao *et al.*, 2018] Quanxue Gao, Lan Ma, Yang Liu, Xinbo Gao, and Feiping Nie. Angle 2dpca: A new formulation for 2dpca. *IEEE Trans. Cybernetics*, 48(5):1672–1678, 2018.

[Gao *et al.*, 2019] Quanxue Gao, Sai Xu, Fang Chen, Chris Ding, Xinbo Gao, and Yunsong Li. ${R}_1$ -2-dpca and face recognition. *IEEE Trans. Cybernetics*, 49(4):1212–1223, 2019.

[Jin and Ghahramani, 2002] Rong Jin and Zoubin Ghahramani. Learning with multiple labels. In *NIPS*, pages 897–904, 2002.

[Liaw *et al.*, 2002] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[Lin *et al.*, 2018] Junyang Lin, Qi Su, Pengcheng Yang, Shuming Ma, and Xu Sun. Semantic-unit-based dilated convolution for multi-label text classification. In *EMNLP*, pages 4554–4564, 2018.

[Liu and Tsang, 2017] Weiwei Liu and Ivor W. Tsang. Making decision trees feasible in ultrahigh feature and label dimensions. *Journal of Machine Learning Research*, 18:81:1–81:36, 2017.

[Liu *et al.*, 2017] Weiwei Liu, Ivor W. Tsang, and Klaus-Robert Müller. An easy-to-hard learning paradigm for multiple classes and multiple labels. *Journal of Machine Learning Research*, 18:94:1–94:38, 2017.

[Liu *et al.*, 2019] Weiwei Liu, Donna Xu, Ivor W. Tsang, and Wenjie Zhang. Metric learning for multi-output tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):408–422, 2019.

[Nguyen and Caruana, 2008] Nam Nguyen and Rich Caruana. Classification with partial labels. In *KDD*, pages 551–559, 2008.

[Pedregosa *et al.*, 2011] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[Read *et al.*, 2011] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359, 2011.

[Trohidis *et al.*, 2008] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis P. Vlahavas. Multi-label classification of music into emotions. In *ISMIR 2008*, pages 325–330, 2008.

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

[Wu and Zhang, 2018] Xuan Wu and Min-Ling Zhang. Towards enabling binary decomposition for partial label learning. In *IJCAI*, pages 2868–2874, 2018.

[Xie and Huang, 2018] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning. In *AAAI*, pages 4302–4309, 2018.

[Yang *et al.*, 2018] Erkun Yang, Cheng Deng, Chao Li, Wei Liu, Jie Li, and Dacheng Tao. Shared predictive cross-modal deep quantization. *IEEE Transactions on Neural Networks and Learning Systems*, 29(11):5292–5303, 2018.

[Yeh *et al.*, 2017] Chih-Kuan Yeh, Wei-Chieh Wu, Wei-Jen Ko, and Yu-Chiang Frank Wang. Learning deep latent space for multi-label classification. In *AAAI*, pages 2838–2844, 2017.

[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 *ICDM*, pages 1398–1403, 2018.

[Zhang and Yu, 2015] Min-Ling Zhang and Fei Yu. Solving the partial label learning problem: An instance-based approach. In *IJCAI*, pages 4048–4054, 2015.

[Zhang and Zhou, 2014] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.*, 26(8):1819–1837, 2014.

[Zhang *et al.*, 2017] Min-Ling Zhang, Fei Yu, and Cai-Zhi Tang. Disambiguation-free partial label learning. *IEEE Trans. Knowl. Data Eng.*, 29(10):2155–2167, 2017.

[Zhu *et al.*, 2005] Xiaojin Zhu, John Lafferty, and Ronald Rosenfeld. *Semi-supervised learning with graphs*. PhD thesis, 2005.