Partial Multi-label Learning via Large Margin Nearest Neighbour Embeddings

Xiuwen Gong [†, ‡], **Dong Yuan** [†], **Wei Bao** [†]

[†] Faculty of Engineering, The University of Sydney
[‡] Hunan Huishiwei Intelligent Technology Co., Ltd.
{xiuwen.gong, dong.yuan, wei.bao}@sydney.edu.au

Abstract

To deal with ambiguities in partial multi-label learning (PML), existing popular PML research attempts to perform disambiguation by direct ground-truth label identification. However, these approaches can be easily misled by noisy false-positive labels in the iteration of updating the model parameter and the latent ground-truth label variables. When labeling information is ambiguous, we should depend more on underlying structure of data, such as label and feature correlations, to perform disambiguation for partially labeled data. Moreover, large margin nearest neighbour (LMNN) is a popular strategy that considers data structure in classification. However, due to the ambiguity of labeling information in PML, traditional LMNN cannot be used to solve the PML problem directly. In addition, embedding is an effective technology to decrease the noise information of data. Inspried by LMNN and embedding technology, we propose a novel PML paradigm called Partial Multi-label Learning via Large Margin Nearest Neighbour Embeddings (PML-LMNNE), which aims to conduct disambiguation by projecting labels and features into a lower-dimension embedding space and reorganize the underlying structure by LMNN in the embedding space simultaneously. An efficient algorithm is designed to implement the proposed method and the convergence rate of the algorithm is analyzed. Moreover, we present a theoretical analysis of the generalization error bound for the proposed PML-LMNNE, which shows that the generalization error converges to the sum of two times the Bayes error over the labels when the number of instances n goes to infinity. Comprehensive experiments on artificial and real-world datasets demonstrate the superiorities of the proposed PML-LMNNE.

Introduction

Partial multi-label learning (PML) (Xie and Huang 2018) is a weakly supervised learning problem (Goldberg et al. 2010), (Vasisht et al. 2014), where each instance is associated with a set of candidate labels, but only a part of them are the ground-truth labels while others are false positive labels. In recent years, many real-world applications are arising as partially labeled data are much easier and less costly to obtain and the demand for identifying ground-truth labels from partially labeled data is growing. For example, in crowdsourcing image annotation, a web image might be annotated online by a potential unreliable annotator with many specific labels, such as, *cloud*, *tree*, *bird*, *grass*, *river*, *dog*, *people* and *building* as shown in Fig. 1, but only the labels with red dots are the ground-truth labels.

The candidate label set is {**cloud, tree, bird, grass, dog, river, people, building**}, but only labels with a red dot are the ground-truth labels.

Figure 1: An example of crowdsourcing image annotation for PML.

PML aims to train a classifier from partially labeled data so as to predict the ground-truth labels for an unseen instance automatically. The main challenge is how to deal with the ambiguities caused by false positive labels in candidate label set. One straightforward way is to simply treat all candidate labels equally as the ground-truth labels, and then solve the PML problem by off-the-shelf multi-label classification methods (Liu 2019), (Boutell et al. 2004). However, these methods can be easily misled by the noisy false-positive labels in the candidate set, and fail to generalize well in testing. As a result, the state-of-the-art PML methods attempt to perform disambiguation by identifying the ground-truth labels directly from candidate label set, which becomes a popular and effective disambiguation strategy. However, this kind of approaches can also be misled by the noisy false-positive labels in the iteration of updating the model parameter and latent variables of ground-truth labels. When labeling information is ambiguous in the partially labeled data, we should depend more on the underlying data structure, such as the label and feature inter-dependencies, to perform disambiguation. Xie and Huang (2018) proposed PML-lc and PML-fp to implement disambiguation by either considering label correlations or feature correlations instead of both. Sun et al. (2019) developed PML-LRS which considers both

the feature and label correlations, but only in the projection of feature matrix rather than the projection of both feature and label matrix. Besides, these methods have no theoretical analysis to guarantee the assumptions. Motivated by this, we would like to develop a method which considers the feature and label correlations in the projection of both feature and label matrix, as well as theoretical guarantee.

Moreover, large margin nearest neighbour (LMNN) (Domeniconi, Gunopulos, and Peng 2005), (Weinberger, Blitzer, and Saul 2005) is a popular strategy in classification of supervised learning, which takes instance and class correlations into consideration. The main idea is to learn a metric by constraining that k-nearest neighbours are classified to the same class and instances from different classes are seperated by a large margin. However, due to the ambiguity of labeling information in partial multi-label learning (PML), it is difficult to precisely identify whether two instances belong to the same class. Thus, traditional LMNN cannot be used to solve the PML problem directly. Moreover, embedding (Adosoglou, Lombardo, and Pardalos 2021) is an effective technology to deal with noisy data, which can be implemented by the projection matrix. Inspried by LMNN and embedding technology, we propose a novel PML paradigm called Partial Multi-label Learning via Large Margin Nearest Neighbour Embeddings (PML-LMNNE) in this paper.

The main contribution of this paper is summarized as follows:

- We propose a new insight into partial multi-label learning (PML) from LMNN and embedding perspective.
- We develop a novel method called Partial Multi-label Learning via Large Margin Nearest Neighbour Embeddings (PML-LMNNE), which aims to perform disambiguation by projecting labels and features into a lower-dimension embedding space and reorganize the underlying structure by LMNN in the embedding space simultaneously.
- We design an efficient algorithm to implement the proposed method. In addition, convergence rate of the algorithm is analyzed in this paper.
- Moreover, we present a theoretical analysis of the generalization error bound for the proposed PML-LMNNE. The results show that the generalization error converges to the sum of two times the Bayes error over the labels when the number of instances n goes to infinity.
- To thoroughly evaluate the effectiveness of the proposed method, we conduct extensive experiments on four synthetic datasets as well as six real-world PML datasets of different scales, which demonstrate the superior performance of the proposed method.

Related Work

Partial multi-label learning (PML) (Xie and Huang 2018) differs from partial label learning (Zhang, Zhou, and Liu 2016; Gong, Yuan, and Bao 2021a,b; Xu, Lv, and Geng 2019; Gong et al. 2021; Zhou, He, and Gu 2017) and multi-label learning (Zhang and Zhou 2007; Liu and Tsang 2017; Liu, Tsang, and Müller 2017; Liu et al. 2019). In PML, each instance is associated with a set of candidate labels, but only part of them are the ground-truth labels while others are false positive labels. The state-of-the art research attempts to perform disambiguation by ground-truth label identification methods. Fang and Zhang (2019) propose PARTICLE to extract credible labels with high confidence via propagation matrix and use the identified labels to train multi-label classifiers. Wang et al. (2019) propose DRAMA to get the reliable labels with high confidence by employing the feature manifold, and then use the identified labels to train the multi-label clssifier. Moreover, Xie and Huang (2018) propose PML-lc and PML-fp to optimize the label ranking confidence matrix in training classifiers which considers the label correlations and the feature prototype respectively. Yu et al. (2018) develop fPML to optimize the label confidentce matrix by considering feature and label correlations. Sun et al. (2019) propose PML-LRS to get label ranking which utilizes the low-rank and sparse decomposition to train classifiers while considering the feature and label interdependencies, the whole process of which is conducted within one projection of feature matrix and decomposes the label matrix into a ground-truth label matrix and an irrelevant label matrix, where the feature mapping matrix and the ground-truth label matrix are constrained to be low rank while the irrelevant label matrix is constrained to be sparse. Li, Lyu, and Feng (2020) develop MUSER to train classifiers by decreasing the feature noise and label redundancy via mapping with orthogonality constraint and graph Laplacian regularization, which considers the feature correlation and label correlation simultaneously. Xu, Liu, and Geng (2020) propose PML-LD to learn from partial multi-label examples via label enhancement, which attempts to recover the label distributionsby exploiting the topological information from the feature space and label correlations from the label space, and then induces a predictive model by fitting the recovered label distributions. Recently, Xie, Sun, and Huang (2021) develop PML-MD, which tries to disambiguate alternatively with a meta-learning strategy. Specifically, multi-label classifier utilizing the supervised information according to the label quality is trained by minimizing a confidence-weighted ranking loss, and the confidence for each candidate label is adaptively estimated with its performance on a small validation set.

When the labeling information is ambiguous in PML, we should depend more on the underlying structure of data to perform disambiguation for the partially labeled data. Large margin nearest neighbour (LMNN), is a popular strategy of classification in supervised learning, which takes instance and class inter-relationships into consideration. Moreover, embedding is an effective technology to deal with noisy data, which can be implemented by projection matrix. Inspried by LMNN and embedding technology, we propose a novel PML paradigm called Partial Multi-label Learning via Large Margin Nearest Neighbour Embeddings (PML-LMNNE).

the Proposed Method

In this section, we propose a novel paradiagm called Partial Multi-label Learning via Large Margin Nearest Neighbour Embeddings (PML-LMNNE).

Let $\{(x_i, Y_i)\}_{i=1}^n$ be the PML training dataset of n training examples, where $x_i \subseteq \mathbb{R}^{p\times 1}$ is the i-th instance of p dimensions (features); $Y_i \subseteq \mathbb{R}^{q\times 1}$ is the candidate label set of q dimensions corresponding to x_i; Let $X \in \mathbb{R}^{n\times p}$ be the input matrix and $Y \in \{0,1\}^{n\times q}$ be the output matrix.

Inspried by LMNN and embedding technology, we conduct disambiguation by projecting labels and features into a lower-dimension embedding space, while reorganizing the underlying data structure by constraining that features of an instance are close to its own labels in the embedding space than the labels of its nearest neighbour. As a result, the projections are embedded into the LMNN framework, which can be achieved by the following two steps.

As features and labels are of different dimensions, we first need to project features into the label space by learning an embedding matrix $V \in \mathbb{R}^{p\times q}$, and then features and labels can be compared. The matrix V can be learned by optimizing the following formulation:

$$\min_{V\in R^{p\times q}} \frac{1}{2}||V^T X^T - Y^T||_F^2 \tag{1}$$

where $\|\cdot\|_F$ represents the Frobenius norm and the superscript ^{T} denotes the transpose of a vector or matrix.

In order to facilitate disambiguation as well as reorganize the underlying structure, we continue to project the labels as well as the projected features into a lower-dimension embedding space by learning another embedding matrix $W \in \mathbb{R}^{q\times d}$.

The projection matrix W can be learned by optimizing the following formulation:

$$\begin{aligned}\min_{W\in R^{q\times d}} & \frac{1}{2}||W||_F^2 + \frac{C}{n}\sum_{i=1}^n \xi_i^2 \\ s.t. \quad & ||W^T V^T x_i - W^T \hat{Y}||_2^2 - ||W^T V^T x_i - W^T Y_i||_2^2 \\ & \geq \Delta(\hat{Y}, Y_i) - \xi_i, \quad \xi_i \geq 0, \quad \forall i \in \{1, \cdots, n\}\end{aligned} \tag{2}$$

where C is a trade-off parameter. $\hat{Y}$ is the label vector of x_i's nearest neighbour. $\Delta(\hat{Y}, Y_i)$ is the margin, defined as $||\hat{Y} - Y_i||_1$ and $||\cdot||_1$ is the l_1 norm. ξ_i is the slack variable.

Here, we define a metric called *embedding distance* to evaluate the correlation between features and labels in the embedding space. The embedding distance between x_i and its label Y_i can be denoted as the square of l_2 norm, i.e., $||W^T V^T x_i - W^T Y_i||_2^2$. Similarly, the embedding distance between x_i and its nearest neighbour's label $\hat{Y}$ is denoted as $||W^T V^T x_i - W^T \hat{Y}||_2^2$.

Therefore, constraints in Eq. (2) guarantee that the embedding distance between x_i and its label Y_i is smaller than that of x_i and its nearest neighour's label $\hat{Y}$ by at least $\Delta(\hat{Y}, Y_i) - \xi_i$, where ξ_i endows our model with more robustness. As a result, the underlying structure (i.e., correlation) between features and labels is reorganized by the constraints, which is retained in W.

We define $P = WW^T$ to be a $q \times q$ symmetric positive semidefinite matrix S_q^+; define $\phi(x_i, Y_i) = V^T x_i - Y_i$. Therefore, Eq. (2) can be transformed to the following formulation:

$$\begin{aligned}\min_{P\in S_q^+} & \frac{1}{2}trace(P) + \frac{C}{n}\sum_{i=1}^n \xi_i^2 \\ s.t. \quad & \phi(x_i, \hat{Y})^T P \phi(x_i, \hat{Y}) - \phi(x_i, Y_i)^T P \phi(x_i, Y_i) \\ & \geq \Delta(\hat{Y}, Y_i) - \xi_i, \quad \xi_i \geq 0, \quad \forall i \in \{1, \cdots, n\}\end{aligned} \tag{3}$$

Moreover, Eq. (3) can be reformulated to the following optimization problem,

$$\begin{aligned}\min_{P\in S_q^+} & \frac{1}{2}trace(P) + \frac{C}{n}\sum_{i=1}^n \mathfrak{L}_i^2 \\ \text{s.t. } \mathfrak{L}_i = & \max\{0, \max_{\hat{Y}\in N(i)} (\Delta(\hat{Y}, Y_i) - ((\phi(x_i, \hat{Y})^T P \phi(x_i, \hat{Y}) \\ & - \phi(x_i, Y_i)^T P \phi(x_i, Y_i)))\}\end{aligned} \tag{4}$$

where $\mathfrak{L}_i$ is the hinge loss.

In addition, define $g(P) = \frac{1}{2}trace(P)$ and $f(P) = \frac{C}{n}\sum_{i=1}^n \mathfrak{L}_i^2$, thus, Eq. (4) can be rewritten as:

$$\min_{P\in S_q^+} F(P), \quad F(P) = f(P) + g(P) \tag{5}$$

where $F(P)$ is the objective function.

Optimization

As f is convex smooth and ∇f is Lipschitz continuous with respect to some positive scalar L_f, we consider to optimize the quadratic approximation of $F(P)$ in Eq. (5).

For any $Z \in S_q^+$, we denote the quadratic approximation at Z as $\hat{F}_\lambda(P, Z)$. Then, we have the following formulation:

$$\begin{aligned}\hat{F}_\lambda(P, Z) &= f(Z) + < \nabla f(Z), P - Z > \\ &\quad + \frac{\lambda}{2}\|P - Z\|_F^2 + g(P) \\ &= \frac{\lambda}{2}\|P - (Z - \frac{1}{\lambda}\nabla f(Z))\|_F^2 + g(P) \\ &\quad + f(Z) - \frac{1}{2\lambda}\|\nabla f(Z)\|_F^2\end{aligned} \tag{6}$$

where λ is a positive constant. To minimize $\hat{F}_\lambda(P, Z)$ in Eq. (6) with respect to P, it is reduced to solve the following optimization problem:

$$\min_{P\in S_q^+} \frac{\lambda}{2}||P - G||_F^2 + g(P) \tag{7}$$

where $G = Z - \frac{1}{\lambda}\nabla f(Z)$.

To solve Eq. (7), we take the derivative of the objective function with respect to P in Eq. (7): $\lambda(P - G) + \frac{1}{2}I = 0$, then $P = G - \frac{1}{2\lambda}I$. We take the singular value decomposition (SVD) of G as $G = U\overline{G}U^T$, and $P = U\overline{G}U^T - \frac{1}{2\lambda}UU^T$, then $P = U(\overline{G} - \frac{1}{2\lambda}I)U^T$. We replace the negative entries in $\overline{G} - \frac{1}{2\lambda}I$ with zeros. Finally, we obtain the

Algorithm 1: PML-LMNNE Algorithm

Input: Let $\eta \in (0,1)$ be a given constant.
Output: Get the optimal solution to Eq. (7).

1: Set $Z_1 = P_1 = P_0 \in S_q^+$, $\alpha_1 = \alpha_0 = 1$ and $\lambda_0 = L_f$.
2: **for** $t = 1, 2, \ldots,$ **do**
3: Set $\hat{\lambda}_0 = \eta\lambda_{t-1}$
4: **for** $i = 0, 1, 2, \ldots,$ **do**
5: Set $G_t = Z_t - \frac{1}{\hat{\lambda}_i}\nabla f(Z_t)$, compute $S_{\hat{\lambda}_i}(G_t)$ from the SVD of G_t.
6: **if** $F(S_{\hat{\lambda}_i}(G_t)) \leq \hat{F}_{\hat{\lambda}_i}(S_{\hat{\lambda}_i}(G_t), Z_t)$, **then**
7: Set $\lambda_t = \hat{\lambda}_i$, $S_{\lambda_t}(G_t) = S_{\hat{\lambda}_i}(G_t))$, stop;
8: **else**
9: Set $\hat{\lambda}_{i+1} = \min\{\frac{1}{\eta}\hat{\lambda}_i, L_f\}$;
10: **end if**
11: **end for**
12: Set $P_{t+1} = S_{\lambda_t}(G_t)$
13: Compute $\alpha_{t+1} = \frac{1+\sqrt{1+4(\alpha_t)^2}}{2}$. Let $t = t + 1$
14: Set $Z_{t+1} = P_t + \frac{\alpha_t - 1}{\alpha_{t+1}}(P_t - P_{t-1})$
15: Quit if $\frac{F(P_t)-F(P_{t+1})}{F(P_t)} \leq \epsilon$ where ϵ is a moderately small tolerance.
16: **end for**

symmetric positive semidefinite matrix solution of Eq. (7), denoted by $S_\lambda(G)$.

Assume ∇f is L_f-Lipschitz and continuous on S_q^+, then $f(P) \leq f(Z) + < \nabla f(Z), P - Z > + \frac{L_f}{2}\|P - Z\|_F^2$. Given this, we can get $f(P) + g(P) \leq f(Z) + < \nabla f(Z), P - Z > + \frac{L_f}{2}\|P - Z\|_F^2 + g(P)$, i.e., $F(P) \leq f(Z) + < \nabla f(Z), P - Z > + \frac{L_f}{2}\|P - Z\|_F^2 + g(P)$. Comparing this inequality with Eq. (6), we can find that, for any $\lambda \geq L_f$, $F(P) \leq \hat{F}_\lambda(P, Z)$. Similarly, we can get

$$F(S_\lambda(G)) \leq \hat{F}_\lambda(S_\lambda(G), Z) \tag{8}$$

Implementation

We design a novel algorithm called PML-LMNNE Algorithm to solve the optimization problem of Eq. (7). The complete procedures are shown in Algorithm 1. We initialize the regularization parameter C and estimate the Lipschitz constant to be $L_f = 0.01nC$.

We first do some initializations (Step 1). In the whole loop (Step 2 – Step 16), the linesearch-like acceleration strategy is incorporated to update λ at iteration t by setting $\lambda_t = \hat{\lambda}_i$ and the small loop will continue to update $\hat{\lambda}_i$ by $\hat{\lambda}_{i+1} = \min\{\frac{1}{\eta}\hat{\lambda}_i, L_f\}$ until the inequality condition is met (i.e. $F(S_{\hat{\lambda}_i}(G_t)) \leq \hat{F}_{\hat{\lambda}_i}(S_{\hat{\lambda}_i}(G_t), Z_t)$) (Step 4 – 11). Compute the symmetric positive semidefinite matrix solution $S_\lambda(G)$ at round t and assign it to P as the value at $(t+1)$-th iteration (Step 12). We set the updating rule for parameter α by $\alpha_{t+1} = \frac{1+\sqrt{1+4(\alpha_t)^2}}{2}$ (Step 13). After that, we set the rule to update Z at t-th iteration based on $Z_t = P_t + \frac{\alpha_{t-1}-1}{\alpha_t}(P_t - P_{t-1})$ (Step 14). The optimization problem in Eq. (4) is convex with no constraint, and the optimal solution can be achieved when $\nabla F(P) = 0$, but it is usually very time-consuming to achieve the optimal solution. In practice, we seek for an ϵ-optimal solution instead and set the stopping condition to be $\frac{F(P_t)-F(P_{t+1})}{F(P_t)} \leq \epsilon$ where ϵ is a small tolerance value and we set it to 10^{-3} in practice (Step 15).

After W is figured out from P, we make prediction based on kNN in the embedding space. Specifically, We first compute the embedding distances between a testing instance x_t and all the training instances x_i $(i = 1, \cdots, n)$ by $||(W^T V^T x_t - W^T V^T x_i)||_2^2$ and then find the nearest neighbour in the embedding space. After that, the nearest neighbour's labels in the embedding space are endowed to the testing instance.

Convergence Analysis

The convergence rate of Algorithm 1 is guaranteed in the following Theorem 1. Before deriving our results, we first present the following lemmas.

Lemma 1. *Since inequality* (8) *is satisfied for* $\lambda \geq L_f$, *where* L_f *is the Lipschitz constant of* ∇f, *it follows from Step 6 – 9 of Algorithm 1 that* $\lambda_t \leq \frac{1}{\eta}L_f$. *Overall,* $L_f \leq \lambda_t \leq \frac{1}{\eta}L_f$.

Lemma 2. *Let* $Z \in S_q^+$ *and* $\lambda > 0$ *be such that*

$$F(S_\lambda(G)) \leq \hat{F}_\lambda(S_\lambda(G), Z) \quad where \quad G = Z - \frac{1}{\lambda}\nabla f(Z)$$

Then for any $X \in S_q^+$,

$$F(X) - F(S_\lambda(G)) \geq \frac{\lambda}{2}\|S_\lambda(G) - Z\|^2 + \lambda\langle Z - X, S_\lambda(G) - Z\rangle$$

Proof. The proof of this lemma can be found in the supplementary material. □

Lemma 3. *The sequence* $\{P_t\}$ *generated via Algorithm 1 satisfies for every* $t \geq 1$,

$$\frac{2}{\lambda_t}\alpha_t^2 v_t - \frac{2}{\lambda_{t+1}}\alpha_{t+1}^2 v_{t+1} \geq \|u_{t+1}\|_F^2 - \|u_t\|_F^2$$

where $v_t = F(P_t) - F(P^*)$, $u_t = \alpha_t P_t - (\alpha_t - 1)P_{t-1} - P^*$ *and* $\alpha_{t+1} = \frac{1+\sqrt{1+4(\alpha_{t-1})^2}}{2}$.

Proof. The proof of this lemma can be found in the supplementary material. □

Lemma 4. *Let* $\{a_t, b_t\}$ *be positive sequences of reals satisfying* $a_t - a_{t+1} \geq b_{t+1} - b_t \quad \forall t \geq 1$, *with* $a_1 + b_1 \leq c$, $c \geq 0$. *Then,* $a_k \leq c$ *for all* $t \geq 1$.

Lemma 5. *The positive sequence* α_t *generated in Algorithm 1 via* $\alpha_{t+1} = \frac{1+\sqrt{1+4(\alpha_t)^2}}{2}$ *with* $\alpha_1 = 1$ *satisfies* $\alpha_t \geq (t+1)/2$ *for all* $t \geq 1$.

Theorem 1. *Let $\{P_t\}$ be the sequences generated by Algorithm 1 and L_f be the Lipschitz constant of ∇f, then for any $t \geq 1$, we have*

$$F(P_t) - F(P^*) \leq \frac{2L_f \|P_0 - P^*\|_F^2}{\eta(t+1)^2} \tag{9}$$

where $P^* = \arg\min_P F(P)$.

Proof. The proof of this theorem can be found in the supplementary material. □

Generalization Error Bound

This section analyzes the generalization error bound of the proposed PML-LMNNE.

Assume $S = \{(x_i, \mathcal{Y}_i)\}_{i=1}^n$ is drawn i.i.d from distribution $\mathcal{D}$, where $x_i \subseteq \mathbb{R}^p$ denotes the p-dimension vector of i-th instance, and $\mathcal{Y}_i \subseteq \{0,1\}^q$ denotes the ground-truth label vector of q dimensions.

Let $h_j^S(x)$ represent the j-th predicted label of an input x using our model trained from S; let y_j represent the true value of j-th label. The overall performance of PML-LMNNE can be measured in terms of generalization error, which is denoted as the expected loss on a new example $(x, \mathcal{Y})$ drawn from the distribution $\mathcal{D}$:

$$E(\sum_{j=1}^{q} \ell(y_j, h_j^S(x))) \tag{10}$$

where $\ell(y_j, h_j^S(x))$ represents the loss function. We further define the loss function as the following form for analysis.

$$\ell(y_j, h_j^S(x)) = P(y_j \neq h_j^S(x)) \tag{11}$$

We define the following function for the j-th label

$$f_z^j(x) = P(y_j = z|x), z \in \{0,1\} \tag{12}$$

Then, the Bayes optimal classifier b^* for j-th label can be defined as

$$b_j^*(x) = \arg\max_{z \in \{0,1\}} f_z^j(x) \tag{13}$$

Before deriving our results, we first present some important definitions.

Definition 1 (Covering Numbers, (Shawe-Taylor et al. 1998))**.** *Let $(\mathcal{X}, d)$ be a metric space, A be a subset of $\mathcal{X}$ and $\varepsilon > 0$. Another subset B of $\mathcal{X}$ is an ε-cover for A, if for every $a \in A$, there exists $b \in B$ such that $d(a,b) < \varepsilon$. The ε-covering number of A, $\mathbb{N}(\varepsilon, A, d)$ is the minimal cardinality of an ε-cover for A (if there is no such finite cover then it is defined as ∞).*

Definition 2 (Doubling Dimension, (Krauthgamer and Lee 2004))**.** *The doubling dimension of a metric space $(\mathcal{X}, d)$ denoted by $ddim(\mathcal{X})$ is the minimum value ρ such that every set in $\mathcal{X}$ can be covered by 2^ρ sets of half the diameter. The diameter of a set $A \subseteq \mathcal{X}$ is $\sup\{d(x,y) : x, y \in A\}$. Define the closed ball of radius r about x in $A \subseteq \mathcal{X}$ to be $B_A(x,r) = \{y \in A : d(x,y) \leq r\}$ and the minimum value ρ such that every ball in $\mathcal{X}$ can be covered by 2^ρ balls of half the radius.*

Lemma 6 (Doubling Metric, (Kontorovich and Weiss 2014))**.** *A metric is doubling when its doubling dimension is bounded. Let $(\mathcal{X}, d)$ be a metric space, let every ball in $\mathcal{X}$ be covered by ρ balls of half the radius, the doubling dimension of $\mathcal{X}$ is $ddim(\mathcal{X}) = \log_2 \rho$. The ε-covering number, $\mathbb{N}(\varepsilon, \mathcal{X}, d)$, is the smallest number of balls of radius ε, is bounded by*

$$\mathbb{N}(\varepsilon, \mathcal{X}, d) \leq (\frac{2diam(\mathcal{X})}{\varepsilon})^{ddim(X)} \tag{14}$$

where $diam(\mathcal{X}) = \sup\{d(x,y) : x, y \in \mathcal{X}\}$ is the diameter of $\mathcal{X}$.

Theorem 2 (Generalization Error Bound for PML-LMNNE)**.** *Given a metric space $(\mathcal{X}, d_{pro})$, where d_{pro} is the embedding distance, assume function $f^j(x) : \mathcal{X} \to \{0,1\}^q$ is L-Lipschitz with respect to the sup-norm for each label. Suppose $\mathcal{X}$ has a finite doubling dimension: $ddim(\mathcal{X}) = \mathbb{N} < \infty$ and $diam(\mathcal{X}) = 1$. Thus, we have*

$$\begin{aligned}\mathbb{E}(\sum_{j=1}^{q} P(y_j \neq h_j^S(x))) \leq & \sum_{j=1}^{q} 2P(b_j^*(x) \neq y_j) \\ & + \frac{3qL(\|V\|_F + \|W\|_F)}{n^{1/(\mathbb{N}+1)}}\end{aligned} \tag{15}$$

Proof. The proof of this theorem can be found in the supplementary material. □

Remark. From the above results, we claim that the error of PML-LMNNE with generalized regularization converges to the sum of two times the Bayes error over the labels when n goes to infinity.

Experiments

In this section, we conduct experiments to evaluate the classification performance of the proposed method and compare it with six state-of-the-art PML methods.

Datasets

Experiments are conducted on six synthetic PML datasets[1] and four real-world PML datasets (i.e. YeastBP (Yu et al. 2018), Music-emotion (Huiskes and Lew 2008), Music-style (Huiskes and Lew 2008), MIRFlickr (Huiskes and Lew 2008) of different scales, which are summarized in Table 1 and Table 2 respectively. For synthetic datasets, given the configuration strategy over multi-label datasets in (Xie and Huang 2018), (Fang and Zhang 2019), we construct the candidate label set by choosing some irrelevant labels together with the ground-truth labels for each instance. Specifically, considering the averaging number of labels in each multi-label dataset, we configure the corresponding candidate label set with different number of labels. In this paper, we generate twenty-eight synthetic PML datasets accordingly. For brevity, we report the detailed results of two configurations for each dataset, i.e. Candidate Labels being 7 and 11 for Enron, Corel5k, Eurlex-sm; 9 and 13 for Eurlex-ed and Mediamill; 45 and 65 for CAL500.

[1] http://mulan.sourceforge.net/datasets-mlc.html

Table 1: Statistics of synthetic PML datasets.

Datasets	#Instances	#Features	#Classes	#Ground-truth Labels (avg.)	#Candidate Labels (avg.)	Domain
Enron	1702	1001	53	3.38	5, 7, 9, 11,13	text
Corel5k	5000	449	374	3.52	5, 7, 9, 11,13	image
Eurlex-sm	19348	5000	201	2.21	5, 7, 9, 11,13	text
Eurlex-ed	19348	5000	3993	5.31	7, 9, 11,13,15	text
CAL500	502	68	174	26.04	35, 45, 55, 65, 75	music
Mediamill	43907	120	101	4.38	7, 9,11,13,15	video

Table 2: Statistics of real-world PML datasets.

Datasets	#Instances	#Features	#Classes	#Labels (avg.)
YeastBP	560	5548	217	30.43
Music-emotion	6833	98	11	5.29
Music-style	6839	98	10	6.04
MIRFlickr	10433	100	7	3.35

Baselines

We compare the proposed PML-LMNNE method with six state-of-the-art PML approaches.

- *PARTICLE* (Fang and Zhang 2019): An identifying method, which tries to extract credible labels with high-confidence values by label propagation procedure, and then trains classifiers by applying two exisiting multi-label models, which are PAR-VLS and PAR-MAP for short. Here, we choose PAR-VLS for comparison.
- *DRAMA* (Wang et al. 2019): An identifying method, which tries to get the reliable labels with high-confidence by conisdering the structure of feature space, and then induces a gradient boosting model to train classifiers.
- *PML-fp and PML-lc* (Xie and Huang 2018): An embedding method, which attempts to figure out the label confidence by minimizing the ranking loss and exploiting data structure information with two models: one considering feature prototype (i.e., PML-fp) and the other considering label correlations (i.e., PML-lc). Here, we choose PML-lc for comparison.
- *fPML* (Yu et al. 2018): An embedding method, which figures out the label confidence by adopting a feature and label coherent matrix to factorize the original matrix for prediction.
- *PML-LRS* (Sun et al. 2019): An embedding method, which utilizes low-rank and sparse decomposition to capture the ground-truth label matrix and irrelevant label matrix from the observed candidate label matrix.
- *MUSER* (Li, Lyu, and Feng 2020): An embedding method, which considers redundant labels together with noisy features and figures out the label confidence via optimizing correlation matrix.

For all PML baselines, we set the trade-off parameters as suggested in the original papers. i.e., PAR-VLS and PAR-MAP: trade-off parameter $\alpha = 0.95$, credible label elicitation threshold $thr = 0.9$ and the number of neighours $k = 10$; DRAMA: $\delta_1 = 0.01$ and $\delta_2 = 1/0.5$; PML-fp and PML-lc: $C_1 = 1$, C_2 is chose from $\{1, 2, \ldots, 10\}$ and C_3 is chose from $\{1, 10, \ldots, 100\}$; fPML: $\lambda_2 = 1$; PML-LRS: trade-off parameters are set as $\gamma = 0.01$, $\beta = 0.1$ and $\eta = 1$; MUSER: α, β, γ are chosen from $\left\{10^{-3}, \ldots, 10^{3}\right\}$ with a grid search manner. Libsvm (Chang and Lin 2011) is used as the binary learning algorithm for PARTICLE.

For PML-LMNNE, we initialize the regularization parameter C by 10-fold cross-validation over the range $\left\{10^{-2}, \ldots, 10^{2}\right\}$ and use Euclidean metric to find the nearest neighbour in initializing the training process.

Evaluation metrics: We employ five widely-used multi-label metrics including ranking loss, hamming loss, one error, coverage, and average precision. More details about these evaluation metrics can be found in (Fürnkranz et al. 2008), (Zhang and Zhou 2014), (Zhou and Zhang 2017).

For the ranking loss, hamming loss, one error and coverage metrics, the smaller value means the better performance. For the average precision metric, the larger value means the better performance.

Experimental Results

We report the performance of the proposed PML-LMNNE and six state-of-the-art PML methods on six synthetic datasets and four real-world datasets in terms of ranking loss, hamming loss, one error, coverage and average precision. As the results of ranking loss, one error and coverage are similar to that of hamming loss and average precision, we only report hamming loss and average precision in Tables 3 and 4 respectively. The other results can be found in the supplementary material. From the overall results, we make the following observations:

- The proposed PML-LMNNE consistently outperforms all baselines on most real-world datasets, like Music-emotion, Music-style and MIRFlickr datasets, while is comparable to the best performance on YeastBP dataset. For example, PML-LMNNE is comparable to MUSER in terms of ranking loss, hamming loss, one error, average precision, and comparable to DRAMA in terms of coverage.
- PML-LMNNE is superior to all baseslines on most synthetic datasets, like Enron, Corel5k, Eurlex-sm, Eurlex-ed and Mediamill while is comparable to the best performance on CAL500 dataset. Specifically, PML-LMNNE is comparable to MUSER in terms of ranking loss and coverage, while it is comparable to DRAMA and MUSER in terms of hamming loss and average precision, to fPML and MUSER in terms of one error.
- From the above results, we can see PML-LMNNE performs the best on most real-world and synthetic datasets,

Table 3: Experimental results of the proposed PML-LMNNE with six state-of-the-art PML baselines on real-world as well as synthetic PML datasets in terms of **hamming loss**. The best result (the smaller the better) is *in bold*.

Dataset	Candidate Labels	PML-LMNNE	PAR-VLS	DRAMA	PML-lc	fPML	PML-LRS	MUSER
YeastBP	30.43	.161±.014	.236±.012	.227±.021	.218±.014	.214±.015	.182±.009	**.158±.012**
Music-emotion	5.29	**.281±.010**	.360±.012	.318±.013	.354±.013	.452±.025	.381±.028	.284±.023
Music-style	6.04	**.162±.011**	.173±.021	.169±.031	.167±.013	.338±.027	.379±.023	.173±.016
MIRFlickr	3.35	**.145±.016**	.193±.017	.219±.014	.216±.018	.223±.022	.237±.012	.193±.056
Enron	7	**.097±.012**	.286±.005	.183±.022	.320±.014	.115±.017	.207±.021	.108±.003
	11	**.121±.011**	.303±.005	.209±.013	.331±.012	.128±.018	.209±.014	.123±.014
Corel5k	7	**.007±.002**	.015±.006	.013±.012	.011±.003	.009±.002	.008±.006	.009±.003
	11	**.011±.004**	.038±.006	.021±.005	.023±.014	.018±.013	.019±.002	.012±.005
Eurlex-sm	7	**.110±.003**	.168±.014	.151±.007	.179±.017	.113±.072	.115±.006	.112±.005
	11	**.166±.005**	.767±.009	.363±.016	.194±.016	.668±.016	.597±.008	.169±.003
Eurlex-ed	9	**.043±.003**	.061±.013	.058±.008	.085±.016	.098±.009	.078±.008	.047±.003
	13	**.147±.007**	.769±.025	.371±.013	.199±.018	.812±.024	.717±.010	.149±.002
CAL500	45	.260±.013	.271±.024	**.235±.017**	.286±.008	.268±.015	.282±.013	.279±.024
	65	.285±.016	.357±.032	.327±.036	.375±.014	.288±.015	.327±.026	**.283±.014**
Mediamill	9	**.059±.013**	.098±.014	.101±.020	.096±.004	.065±.007	.072±.023	.087±.021
	13	**.122±.010**	.145±.018	.201±.027	.218±.025	.513±.021	.191±.017	.183±.012

Table 4: Experimental results of the proposed PML-LMNNE with six state-of-the-art PML baselines on real-world as well as synthetic PML datasets in terms of **average precision**. The best result (the larger the better) is *in bold*.

Dataset	Candidate Labels	PML-LMNNE	PAR-VLS	DRAMA	PML-lc	fPML	PML-LRS	MUSER
YeastBP	30.43	.152±.023	.082±.031	.083±.017	.140±.035	.096±.021	.085±.023	**.154±.031**
Music-emotion	5.29	**.611±.017**	.527±.006	.582±.012	.541±.023	.538±.015	.516±.014	.598±.033
Music-style	6.04	**.726±.024**	.717±.031	.693±.013	.627±.010	.659±.017	.716±.018	.718±.013
MIRFlickr	3.35	**.831±.012**	.685±.017	.707±.014	.743±.018	.731±.015	.796±.012	.801±.016
Enron	7	**.779±.010**	.601±.006	.613±.002	.679±.003	.751±.012	.782±.011	.771±.003
	11	**.694±.006**	.587±.006	.556±.012	.660±.004	.670±.006	.683±.007	.681±.005
Corel5k	7	**.289±.010**	.205±.036	.235±.014	.253±.023	.264±.017	.237±.013	.280±.003
	11	**.282±.005**	.196±.012	.218±.025	.226±.013	.258±.015	.217±.011	.276±.015
Eurlex-sm	7	**.755±.019**	.741±.024	.744±.017	.718±.037	.685±.022	.699±.016	.751±.025
	11	**.752±.014**	.721±.016	.728±.013	.716±.006	.628±.027	.615±.017	.748±.023
Eurlex-ed	9	**.757±.013**	.735±.013	.727±.016	.719±.024	.686±.010	.696±.015	.755±.023
	13	**.755±.028**	.728±.015	.725±.023	.715±.017	.668±.014	.681±.016	.752±.012
CAL500	45	.615±.021	.446±.024	.563±.027	.581±.018	.531±.025	.516±.023	**.620±.014**
	65	.480±.011	.432±.012	**.481±.015**	.434±.015	.412±.022	.448±.014	.479±.018
Mediamill	9	**.765±.018**	.756±.018	.687±.017	.685±.025	.695±.017	.689±.010	.716±.012
	13	**.733±.016**	.699±.024	.698±.014	.685±.019	.674±.018	.686±.013	.702±.021

except on YeastBP and CAL500, which is maybe because the instance number of YeastBP and CAL500 is small that decreases the prediction performance. This aligns with our theoretical analysis that PML-LMNNE can generalize well with the increasing number of instances.

The overall results validate the superior performance of the proposed method.

Conclusion

This paper provides a new insight into partial multi-label learning problem from LMNN and embedding perspectives. In order to perform disambiguation for PML, we propose a novel method called Partial Multi-label Learning via Large Margin Nearest Neighbour Embeddings (PML-LMNNE), which aims to conduct disambiguation by projecting labels and features into a lower-dimension embedding space and reorganize the underlying structure by LMNN in the embedding space simultaneously. An efficient algorithm is designed to implement the proposed method and the convergence rate of the algorithm is analyzed in this paper. Moreover, we present a theoretical analysis of the generalization error bound for the proposed PML-LMNNE. The results show that the generalization error converges to the sum of two times the Bayes error over the labels when the number of instances n goes to infinity. To thoroughly evaluate the effectiveness of PML-LMNNE, we conduct extensive experiments on four synthetic datasets as well as six real-world PML datasets. The results clearly demonstrate the superiorities of the proposed PML-LMNNE compared with the state-of-the-art PML methods.

References

Adosoglou, G.; Lombardo, G.; and Pardalos, P. M. 2021. Neural Network Embeddings on Corporate Annual Filings for Portfolio selection. *Expert Systems with Applications*, 164: 114053.

Boutell, M. R.; Luo, J.; Shen, X.; and Brown, C. M. 2004. Learning Multi-label Scene Classification. *Pattern Recognition*, 37(9): 1757–1771.

Chang, C.; and Lin, C. 2011. LIBSVM: A library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3): 27:1–27:27.

Domeniconi, C.; Gunopulos, D.; and Peng, J. 2005. Large Margin Nearest Neighbor Classifiers. *IEEE Transactions on Neural Networks and Learning Systems*, 16(4): 899–909.

Fang, J.; and Zhang, M. 2019. Partial Multi-Label Learning via Credible Label Elicitation. In *AAAI*, 3518–3525.

Fürnkranz, J.; Hüllermeier, E.; Mencía, E. L.; and Brinker, K. 2008. Multilabel Classification via Calibrated Label Ranking. *Machine Learning*, 73(2): 133–153.

Goldberg, A. B.; Zhu, X.; Recht, B.; Xu, J.; and Nowak, R. D. 2010. Transduction with Matrix Completion: Three Birds with One Stone. In *NeurIPS*, 757–765.

Gong, X.; Yang, J.; Yuan, D.; and Bao, W. 2021. Generalized Large Margin kNN for Partial Label Learning. *IEEE Transactions on Multimedia*.

Gong, X.; Yuan, D.; and Bao, W. 2021a. Discriminative Metric Learning for Partial Label Learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Gong, X.; Yuan, D.; and Bao, W. 2021b. Top-k Partial Label Machine. *IEEE Transactions on Neural Networks and Learning Systems*.

Huiskes, M. J.; and Lew, M. S. 2008. The MIR Flickr Retrieval Evaluation. In *SIGMM*, 39–43.

Kontorovich, A.; and Weiss, R. 2014. Maximum Margin Multiclass Nearest Neighbors. In *ICML*, 892–900.

Krauthgamer, R.; and Lee, J. R. 2004. Navigating Nets: Simple Algorithms for Proximity Search. In *SODA*, 798–807.

Li, Z.; Lyu, G.; and Feng, S. 2020. Partial Multi-Label Learning via Multi-Subspace Representation. In *IJCAI 2020*, 2612–2618.

Liu, W. 2019. Copula Multi-label Learning. In *NeurIPS*, 6334–6343.

Liu, W.; and Tsang, I. W. 2017. Making Decision Trees Feasible in Ultrahigh Feature and Label Dimensions. *Journal of Machine Learning Research*, 18: 81:1–81:36.

Liu, W.; Tsang, I. W.; and Müller, K. 2017. An Easy-to-hard Learning Paradigm for Multiple Classes and Multiple Labels. *Journal of Machine Learning Research*, 18: 94:1–94:38.

Liu, W.; Xu, D.; Tsang, I. W.; and Zhang, W. 2019. Metric Learning for Multi-Output Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 408–422.

Shawe-Taylor, J.; Bartlett, P. L.; Williamson, R. C.; and Anthony, M. 1998. Structural Risk Minimization Over Data-Dependent Hierarchies. *IEEE Transactions on Information Theory*, 1926–1940.

Sun, L.; Feng, S.; Wang, T.; Lang, C.; and Jin, Y. 2019. Partial Multi-Label Learning by Low-Rank and Sparse Decomposition. In *AAAI*, 5016–5023.

Vasisht, D.; Damianou, A. C.; Varma, M.; and Kapoor, A. 2014. Active Learning for Sparse Bayesian Multilabel Classification. In *SIGKDD*, 472–481.

Wang, H.; Liu, W.; Zhao, Y.; Zhang, C.; Hu, T.; and Chen, G. 2019. Discriminative and Correlative Partial Multi-Label Learning. In *IJCAI*, 3691–3697.

Weinberger, K. Q.; Blitzer, J.; and Saul, L. K. 2005. Distance Metric Learning for Large Margin Nearest Neighbor Classification. In *NeurIPS*, 1473–1480.

Xie, M.; and Huang, S. 2018. Partial Multi-Label Learning. In *AAAI*, 4302–4309.

Xie, M.; Sun, F.; and Huang, S. 2021. Partial Multi-Label Learning with Meta Disambiguation. In *SIGKDD*, 1904–1912.

Xu, N.; Liu, Y.; and Geng, X. 2020. Partial Multi-Label Learning with Label Distribution. In *AAAI*, 6510–6517.

Xu, N.; Lv, J.; and Geng, X. 2019. Partial Label Learning via Label Enhancement. In *AAAI*, 5557–5564.

Yu, G.; Chen, X.; Domeniconi, C.; Wang, J.; Li, Z.; Zhang, Z.; and Wu, X. 2018. Feature-Induced Partial Multi-label Learning. In *ICDM*, 1398–1403.

Zhang, M.; Zhou, B.; and Liu, X. 2016. Partial Label Learning via Feature-Aware Disambiguation. In *SIGKDD*, 1335–1344.

Zhang, M.; and Zhou, Z. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7): 2038–2048.

Zhang, M.; and Zhou, Z. 2014. A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8): 1819–1837.

Zhou, Y.; He, J.; and Gu, H. 2017. Partial Label Learning via Gaussian Processes. *IEEE Transactions on Cybernetics*, 47(12): 4443–4450.

Zhou, Z.; and Zhang, M. 2017. Multi-label Learning. In *Encyclopedia of Machine Learning and Data Mining*, 875–881. Springer.