Learning From Multi-Dimensional Partial Labels

Haobo Wang[1,2], **Weiwei Liu**[3*], **Yang Zhao**[2], **Tianlei Hu**[1,2], **Ke Chen**[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, chenk, htl, cg}@zju.edu.cn, liuweiwei863@gmail.com

Abstract

Multi-dimensional classification (MDC) has attracted much attention from the community. Though most studies consider fully annotated data, in real practice obtaining fully labeled data in MDC tasks is usually intractable. In this paper, we propose a novel learning paradigm: *Multi-Dimensional Partial Label Learning (MDPL)* where the ground-truth labels of each instance are concealed in multiple candidate label sets. We first introduce the *partial hamming loss* for MDPL that incurs a large loss if the predicted labels are not in candidate label sets, and provide an empirical risk minimization (ERM) framework. Theoretically, we rigorously prove the conditions for ERM learnability of MDPL in both independent and dependent cases. Furthermore, we present two MDPL algorithms under our proposed ERM framework. Comprehensive experiments on both synthetic and real-world datasets validate the effectiveness of our proposals.

1 Introduction

Multi-dimensional classification (MDC) aims to assign each instance to multiple classes, which has been seen in a variety of real-world applications, including but not limited to, text categorization [Ortigosa-Hernández *et al.*, 2012], gene function prediction [Barutçuoglu *et al.*, 2006] and image annotation [Read *et al.*, 2014; Batal *et al.*, 2013; Arias *et al.*, 2016]. In order to train an effective MDC model, it is typically desirable to obtain a large number of precisely annotated data. Unfortunately, obtaining fully labeled data in MDC tasks is usually intractable. As a result, it is non-trivial to learn multi-dimensional classifiers from partially labeled data.

Weakly-supervised learning has been explored to deal with partially labeled data in various settings. For example, semi-supervised learning (SSL) [Chapelle *et al.*, 2002] learns from both labeled and unlabeled data. In positive-unlabeled learning (PUL) [Denis, 1998; Kiryo *et al.*, 2017], there are only positive labeled data and unlabeled data available. In partial label learning (PLL) [Cour *et al.*, 2011; Liu and Dietterich, 2012; Wu and Zhang, 2018], the ground-truth label is concealed in a set of candidate labels. Recently, there are also some works that address the weakly-supervised learning problem in multiple-label setting, such as semi-supervised multi-label learning [Zhan and Zhang, 2017], partial multi-label learning (PML) [Fang and Zhang, 2019; Wang *et al.*, 2019] and semi-supervised multi-dimensional classification [Ortigosa-Hernández *et al.*, 2012].

In this work, we consider a new weakly-supervised MDC scenario where the ground-truth labels of each instance are concealed in multiple candidate label sets, i.e. *Multi-Dimensional Partial Label Learning (MDPL)*. Take the image [Khosla *et al.*, 2011] in Table **??** as an example, it is associated with four class variables {*Place, Tree, Dog Breeds, Weather*}. It is hard for the annotators to identify all the correct labels, but they can provide some candidate labels with much less effort. Label disambiguation and label correlation extraction pose the serious challenges in MDPL. The noisy information will decrease the generalization performance of MDPL. However, the label correlations will provide additional semantic information to disambiguate the noisy labels. For example, since there exist some trees in the image, it is more likely to be a Mountain instead of a Glacier.

Our main contributions in this paper are to formulate the MDPL problem and provide an empirical risk minimization (ERM) framework. In particular, we propose a *partial hamming loss* that incurs a large loss if the predicted labels are not included in candidate label sets. Theoretically, we rigorously present the conditions for ERM learnability of MDPL in both independent and dependent cases. Moreover, we instantiate two MDPL algorithms under our empirical risk minimization framework. Extensive experiments on both synthetic and real-world datasets demonstrate that our proposed methods can effectively handle MDPL tasks.

2 Related Work

2.1 Multi-Dimensional Classification

In multi-dimensional classification (MDC), each object is associated with multiple class variables. It is a generalization of multi-label learning [Liu and Tsang, 2017; Shen *et al.*, 2018; Liu *et al.*, 2019] that allows each class variable to have more than two values. Compared to MLL problems, the label correlations in MDC are more sophisticated, because the intra-

*Corresponding Author.

Type / Class	Multi-Dimensional	Multi-Dimensional Partial Labels
Place	Mountain	Mountain, Glacier
Tree	Yes	Yes
Dog Breeds	Malamute	Siberian Husky, Malamute
Weather	Sunny	Sunny, Snowy, Cloudy

Table 1: An example of MDPL task for image annotation. In MDC, we provide all the ground-truth labels. In MDPL, only some candidate labels are given but it takes much less time than precise annotation.

class labels are exclusive, while inter-class labels may still correlate to each other. One popular strategy for MDC is binary relevance (BR) [Read *et al.*, 2014] that decomposes the original problem into several multi-class classification problems. Despite its computational efficiency, BR neglects the label dependencies and hence the predictive performance is limited. To cope with this shortcoming, many works are proposed, including probabilistic graph model based algorithms [Batal *et al.*, 2013; Benjumeda *et al.*, 2018], classifier chains [Zaragoza *et al.*, 2011], instance-based approaches [Jia and Zhang, 2019] and so on. Nevertheless, all of them require the training data to be precisely labeled, which is demanding and time-consuming.

Consequently, some weakly-supervised multiple-label problems have been studied, such as semi-supervised multi-label learning [Zhao and Guo, 2015; Zhan and Zhang, 2017], partial multi-label learning (PML) [Fang and Zhang, 2019; Wang *et al.*, 2019], semi-supervised multi-dimensional classification [Ortigosa-Hernández *et al.*, 2012] and so on. However, most of these learning paradigms explore only multi-label setting where the labels are restricted to be binary and it is non-trivial to study the generalized weakly-supervised MDC problems.

2.2 Partial Label Learning

The partial label learning (PLL) setting is between fully supervised and unsupervised learning setting, but is quantitively different from SSL [Chapelle *et al.*, 2002; Zhan and Zhang, 2017] and PUL [Denis, 1998; Kiryo *et al.*, 2017]. In PLL, each instance is equipped with a set of candidate labels. The ground-truth label is guaranteed to be included and the remaining labels are termed as *distractor labels* or *false positive labels*. The biggest challenging issue in PLL is to disambiguate the ground-truth label from the distractor labels and many papers [Cour *et al.*, 2011; Liu and Dietterich, 2012; Wu and Zhang, 2018; Feng and An, 2019; Lv *et al.*, 2020] are presented to address this problem. There are also some works [Fang and Zhang, 2019; Wang *et al.*, 2019] studying partial multi-label learning, which extends PLL problem to the multiple-label learning field. Nonetheless, PML restricts the labels to be binary and thus is unpractical in many real-world tasks [Read *et al.*, 2014].

To bridge this gap, we propose a novel learning paradigm: multi-dimensional partial label learning where the ground-truth labels of each instance are concealed in multiple candidate label sets.

3 Learning Framework

We first formulate the problem of MDPL and introduce an empirical risk minimization framework.

Consider a standard setting of MDC problem with an input space $\mathcal{X} \subseteq \mathbb{R}^m$ and an output space $\mathcal{Y} = \mathcal{C}_1 \times \mathcal{C}_2 \times ... \times \mathcal{C}_d$. Here $\mathcal{C}_i = \{l_{i1}, l_{i2}, ..., l_{ik_i}\}$ is the i-th class space and $\mathcal{Y}$ is their Cartesian product. The ultimate goal of MDC is to induce a mapping from $\mathcal{X}$ to $\mathcal{Y}$ that captures the dependence of the outputs on inputs. To this end, based on a training dataset $Q = \{(\boldsymbol{x}_i, Y_i) | \boldsymbol{x}_i \in \mathcal{X}, Y_i \in \mathcal{Y}, 1 \leq i \leq n\}$, a learner chooses an optimal hypothesis h^* from a given hypothesis space $\mathcal{H}$ to minimize the prediction loss. Specifically, common choices of prediction loss (or *risk*) for MDC include *hamming loss* and *global loss* [Read *et al.*, 2014].

In MDPL tasks, we are interested in the case where the correct labels are adulterated by false positive labels. To be more specific, the ground-truth labels are invisible and only a collection of candidate label sets $S = \{s_1, s_2, ..., s_d\} \in \mathcal{S}$ is given, where $\mathcal{S} = (2^{\mathcal{C}_1} - \emptyset) \times (2^{\mathcal{C}_2} - \emptyset) \times ... \times (2^{\mathcal{C}_d} - \emptyset)$ is the candidate class space and $s_i \subseteq \mathcal{C}_i$ is the i-th candidate label set for the corresponding class space. We denote a complete example by $(\boldsymbol{x}, Y, S)$, where only instance vector $\boldsymbol{x}$ and candidate label collection S are accessible. The goal of MDPL is to learn a multi-dimensional classifier $h : \mathcal{X} \mapsto \mathcal{Y}$ from multi-dimensional partially labeled data by minimizing the *expected hamming loss*: $\mathcal{L}^H_{\mathcal{D}}(h) = E_{(\boldsymbol{x},Y,S)\sim\mathcal{D}}[\frac{1}{d}\sum_{i=1}^{d} \mathbb{I}(h^i(\boldsymbol{x}) \neq y_i)]$, where $\mathcal{D}$ is the underlying data distribution, $h^i(\boldsymbol{x})$ is the i-th predicted label and y_i denotes the i-th ground-truth label. Since the correct labels are invisible in the training dataset, we can not minimize the standard *hamming loss* directly. Inspired by *partial* $0/1$ *loss* [Cour *et al.*, 2011], we introduce a multi-dimensional version named *expected partial hamming loss*,

$$\mathcal{L}^P_{\mathcal{D}}(h) = E_{(\boldsymbol{x},Y,S)\sim\mathcal{D}}[\frac{1}{d}\sum_{i=1}^{d} \mathbb{I}(h^i(\boldsymbol{x}) \notin s_i)] \qquad (1)$$

An obvious observation is that *expected partial hamming loss* is an underestimate of the true *expected hamming loss*. Thus, it is not a surrogate and we have to explore some conditions where minimizing the partial loss can also bound the true loss. Moreover, a large loss will incur if the predicted labels are not included in candidate label sets. It motivates us to employ the VC-dimension of the inside-out set binary classification task as a bridge to complete our theoretical proof.

In summary, based on our proposed *expected partial hamming loss*, we propose an empirical risk minimization framework for MDPL, and each hypothesis h will be evaluated by *average partial hamming loss*,

$$\mathcal{L}_Z^P(h) = \frac{1}{nd}\sum_{i=1}^{n}\sum_{j=1}^{d}\mathbb{I}(h^j(\boldsymbol{x}_i) \notin s_j^i) \tag{2}$$

where $Z = \{(\boldsymbol{x}_i, S_i)|\boldsymbol{x}_i \in \mathcal{X}, S_i \in \mathcal{S}, 1 \leq i \leq n\}$ is the partially labeled training dataset and s_j^i is the j-th candidate label set of i-th training example.

4 Learnability of MDPL

In this section, we will discuss how to bound the true loss using *expected partial hamming loss*. In this paper, we only investigate the realizable case where an optimal hypothesis h^* makes the risk $\mathcal{L}_\mathcal{D}^H(h^*) = 0$.

4.1 Independent Case

We first consider the independent case, i.e. the labels are independent to each other. Then we can simply decompose the MDPL problem to a set of partial label learning problems.

Many works have studied PLL problems based on minimizing the upper-bound of risk $\mathcal{L}_{\mathcal{D}_p}$, usually, the *expected* $0/1$ *loss* [Cour *et al.*, 2011]: $\mathcal{L}_{\mathcal{D}_p}(h_p) = E_{(\boldsymbol{x},y,s)\sim\mathcal{D}_p}[\mathbb{I}(h_p(\boldsymbol{x}) \neq y)]$, where $\mathcal{D}_p$ is the underlying distribution of a PLL task. Based on this risk, we obtain the following lemma.

Lemma 1. *Assume that the labels in an MDPL problem are independent to each other. Then if a PLL problem adopts the expected* $0/1$ *loss as risk and it is PAC-learnable with sample complexity* $n_0(\mathcal{H}_p, \delta, \epsilon)$*, the MDPL problem is also PAC-learnable with sample complexity as follows,*

$$n_1(\mathcal{H}, \delta, \epsilon) = \max_i n_0(\mathcal{H}_p^i, \delta, \epsilon) \tag{3}$$

where $\mathcal{H}_p^i$ *is the i-th PLL hypothesis space.*

Proof. If a PLL problem is PAC-learnable [Shalev-Shwartz and Ben-David, 2014], then for every $\epsilon, \delta \in (0, 1)$, when the training set has size $n \geq n_0(\mathcal{H}_p, \delta, \epsilon)$, there exists an ERM learner $\mathcal{A}_p$ that returns a hypothesis $h_p \in \mathcal{H}_p$ with expected $0/1$ loss $\mathcal{L}_{\mathcal{D}_p}(h_p) \leq \epsilon$. Since the labels in this MDPL problem are independent to each other, the MDPL task can be decomposed to d PLL problems. By running an ERM learner $\mathcal{A}_p^i$ on each single PLL problem, and aggregating the hypotheses $h^i = \mathcal{A}_p^i(Z_p^i)$, we can obtain an MDPL classifier $h = [h^i]_d$, where Z_p^i is the i-th PLL training dataset. When the training set has size $n \geq n_1(\mathcal{H}, \delta, \epsilon) \geq n_0(\mathcal{H}_p^i, \delta, \epsilon)$, for every $\epsilon, \delta \in (0, 1)$, the following inequality holds with probability no less than $1 - \delta$,

$$\begin{aligned}\mathcal{L}_\mathcal{D}^H(h) &= \frac{1}{d}\sum_{i=1}^{d} E_{(\boldsymbol{x},y_i,s_i)\sim\mathcal{D}_p}[\mathbb{I}(h^i(\boldsymbol{x}) \neq y_i)] \\ &= \frac{1}{d}\sum_{i=1}^{d}\mathcal{L}_{\mathcal{D}_p}(h^i) \leq \frac{1}{d}\cdot d\epsilon = \epsilon\end{aligned} \tag{4}$$

We conclude that the MDPL problem is PAC-learnable with sample complexity $n_1(\mathcal{H}, \delta, \epsilon)$. □

The learnability of partial label learning can refer to many works [Cour *et al.*, 2011; Ishida *et al.*, 2017]. For instance, the small ambiguity degree condition, proposed by [Cour *et al.*, 2011], is one of the most popular assumptions in PLL problems.

4.2 Dependent Case

During the past decades, a variety of works [Zaragoza *et al.*, 2011; Read *et al.*, 2014; Shen *et al.*, 2018] have proved that neglecting label correlations may achieve degenerated performance in multiple-label problems. Thus, it is crucial to study the problem in what condition can MDPL tasks be learned in the dependent case.

Here we propose a sufficient condition for the PAC-learnability of MDPL tasks.

Theorem 1. *In an MDPL problem, if there exists a positive constant* $\gamma > 0$ *such that,*

$$\forall h \in \mathcal{H} : \mathcal{L}_\mathcal{D}^H(h) > 0, \quad \frac{\mathcal{L}_\mathcal{D}^P(h)}{\mathcal{L}_\mathcal{D}^H(h)} \geq \gamma \tag{5}$$

then in realizable case, the MDPL problem is PAC-learnable.

We first introduce an MDC algorithm, *Class Powerset* (CP) [Read *et al.*, 2014], into MDPL scenario. The basic idea is to transform the MDC problem to a multi-class classification problem by regarding each label combination as a new class. Then it learns a multi-class classifier $f : \mathcal{X} \mapsto \tilde{\mathcal{Y}}$ where $\tilde{\mathcal{Y}}$ is the new label space with size $\prod_{i=1}^{d} k_i$. Since all the label combinations are considered, we can fully explore the label correlations across the class space.

Specifically, we call an ERM learner $\mathcal{A}_{cp}$ that returns a hypothesis minimizing the multi-class risk, i.e. *expected* $0/1$ *loss*. Nonetheless, in MDPL setting, the learner does not have direct access to the precise data. To deal with this issue, we involve a surrogate loss called *global loss* $\mathcal{L}^G$ with corresponding *partial global loss* $\mathcal{L}^{GP}$, which are defined as,

$$\begin{aligned}\mathcal{L}_\mathcal{D}^G(h) &= E_{(\boldsymbol{x},Y,S)\sim\mathcal{D}}[\mathbb{I}(\exists i, h^i(\boldsymbol{x}) \neq y_i)], \\ \mathcal{L}_\mathcal{D}^{GP}(h) &= E_{(\boldsymbol{x},Y,S)\sim\mathcal{D}}[\mathbb{I}(\exists i, h^i(\boldsymbol{x}) \notin s_i)]\end{aligned} \tag{6}$$

We can immediately obtain their relation,

$$\begin{aligned}\frac{1}{d}\mathcal{L}_\mathcal{D}^G(h) \leq \mathcal{L}_\mathcal{D}^H(h) \leq \mathcal{L}_\mathcal{D}^G(h), \\ \frac{1}{d}\mathcal{L}_\mathcal{D}^{GP}(h) \leq \mathcal{L}_\mathcal{D}^P(h) \leq \mathcal{L}_\mathcal{D}^{GP}(h)\end{aligned} \tag{7}$$

The last step is to design a CP algorithm $\mathcal{A}_{cp}$ that minimizes the *empirical partial global loss*,

$$\begin{aligned}\mathcal{A}_{cp}(Z) &= \operatorname*{argmin}_{h\in\mathcal{H}_{mc}} \mathcal{L}_Z^{GP}(h) \\ &= \operatorname*{argmin}_{h\in\mathcal{H}_{mc}} \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(\exists i, h^i(\boldsymbol{x}) \notin s_i)\end{aligned} \tag{8}$$

In traditional MDC setting, it is a typical multi-class learning problem. Let $\mathcal{H}_b^{cp}$ be a binary hypothesis class with VC-dimension $\tau = \mathrm{VCdim}(\mathcal{H}_b^{cp})$, e.g., $\mathcal{H}_b^{cp}$ is linear with $\tau = m$.

Suppose that the multi-class hypothesis space $\mathcal{H}_{mc}$ is constructed above $\mathcal{H}_b^{cp}$ using one-versus-all strategy. According to Lemma 29.5 in [Shalev-Shwartz and Ben-David, 2014], the Natarajan dimension [Natarajan, 1989] of $\mathcal{H}_{mc}$ enjoys an upper-bound of,

$$\mathrm{Ndim}(\mathcal{H}_{mc}) \leq 3\tau \log(\tau \prod_{i=1}^{d} k_i) \prod_{i=1}^{d} k_i \tag{9}$$

where $\mathrm{Ndim}(\cdot)$ denotes the Natarajan dimension of a hypothesis space. Nevertheless, in MDPL setting, our learning problem is no longer a multi-class task and the Natarajan dimension can not directly yield the sample complexity.

Our strategy is to construct a binary classification task from the problem above. Given a partial example $(\boldsymbol{x}, S)$, the binary classifier should predict whether there exist some predicted labels y_i outside their corresponding candidate label sets, i.e. returning $\mathbb{I}(\exists i, y_i \notin s_i)$. We observe that the binary classification loss is the *partial global loss*. Therefore, we can design an ERM learner $\mathcal{A}_b$ that calls $\mathcal{A}_{cp}$ and then transforms the prediction to binary output space. Compared with class powerset method, it is unpractical but provides good theoretical results. Now our task is to explore the VC-dimension of the binary classifier.

Denote the hypothesis space of this binary classification task by $\mathcal{H}_b$. We have the following lemmas.

Lemma 2. *Let $K = \prod_{i=1}^{d} k_i$. The VC-dimension of $\mathcal{H}_b$ can be bounded by,*

$$\mathrm{VCdim}(\mathcal{H}_b) \leq \frac{3\tau K \log(\tau K)}{\log 2 - e^{-1}} (\log(3\tau K \log(\tau K)) + 2 \log K) \tag{10}$$

Proof. Let $\nu = \mathrm{VCdim}(\mathcal{H}_b)$ and $\mu = \mathrm{Ndim}(\mathcal{H}_{mc})$. Then, the maximum size of a set that $\mathcal{H}_b$ can shatter is ν. In other words, there are 2^ν different dichotomies (i.e., labelings) induced by $\mathcal{H}_b$ over these ν instances. Based on Lemma 29.4 in [Shalev-Shwartz and Ben-David, 2014], we can conclude that,

$$2^\nu \leq \nu^\mu K^{2\mu} \tag{11}$$

Taking the natural logarithm of both sides yields that,

$$\nu \log 2 \leq \mu \log \nu + 2\mu \log K \tag{12}$$

To bound ν, we involve a function $g(x) = e \log x - x$. Its maximum value $g(x) = 0$ is obtained when $g'(x) = 0$, i.e. $x = e$. Hence, $g(x) \leq 0$ holds for all $x > 0$. Choosing $x = \frac{\nu}{\mu}$ gives that $\log \nu \leq \frac{\nu}{e\mu} + \log \mu$. Hence,

$$\begin{aligned} \nu \log 2 &\leq \mu(\frac{\nu}{e\mu} + \log \mu) + 2\mu \log K \\ \nu &\leq \frac{\mu \log \mu + 2\mu \log K}{\log 2 - e^{-1}} \end{aligned} \tag{13}$$

Combining Eq. (9) and Eq (13), we obtain the desired result. □

Lemma 3. *For every $\delta, \epsilon \in (0, 1)$, every distribution $\mathcal{D}$ over $\mathcal{X}$, and the binary classification task defined above, if the realizable assumption holds, when running algorithm $\mathcal{A}_b$ on a training set of size n satisfying*

$$\begin{aligned} n \geq n_2(\mathcal{H}_b, \delta, \epsilon) = 4 \frac{32\nu}{\epsilon^2} \cdot \log(\frac{64\nu}{\epsilon^2}) \\ + \frac{8}{\epsilon^2} \cdot (8\nu \log(e/\nu) + 2\log(2/\delta)) \end{aligned} \tag{14}$$

then the algorithm returns a hypothesis h such that with probability of at least $1 - \delta$, $\mathcal{L}_{\mathcal{D}}^{GP}(h) \leq \epsilon$.

The proof can be found in [Shalev-Shwartz and Ben-David, 2014].

Now recalling Eq. (5) and Eq. (7), when $\mathcal{A}_b$ runs on a training set of size $n_2(\mathcal{H}_b, \delta, \frac{\epsilon}{\gamma})$, the hamming loss has the following bound,

$$\mathcal{L}_{\mathcal{D}}^{H}(h) \leq \gamma \mathcal{L}_{\mathcal{D}}^{P}(h) \leq \gamma \mathcal{L}_{\mathcal{D}}^{GP}(h) \leq \epsilon \tag{15}$$

Taking the corresponding multi-class hypothesis h as our solution, we obtain a provable algorithm that ensures the MDPL problems to be PAC-learnable with a finite sample complexity $n_2(\mathcal{H}_b, \delta, \frac{\epsilon}{\gamma})$. Therefore, Theorem 1 is proved.

4.3 Further Discussion

Remark of the Proposed Condition

Suppose we know the distribution of partial examples. We can design a Bayesian optimal classifier with zero partial hamming loss. In realizable case, our goal is to find a hypothesis h^* that satisfies $\mathcal{L}_{\mathcal{D}}^{H}(h^*) = 0$. Denote the optimal hypothesis set by $\mathcal{H}^*$. If there exists a hypothesis $\hat{h} \notin \mathcal{H}^*$ such that $\mathcal{L}_{\mathcal{D}}^{P}(\hat{h}) = 0$, even the Bayesian optimal classifier can not guarantee to return an optimal solution. Hence, our sufficient condition actually ensures the ERM learner to return an optimal hypothesis from $\mathcal{H}^*$.

Relation to PML

Another observation is that the recently popularized task of partial multi-label learning also benefits from our theoretical analysis. In a typical PML problem, the ground-truth binary labels are adulterated with some irrelevant labels. If we regard each candidate label as a two-element candidate label set, it can be categorized into MDPL problems. In independent case, each positive label is accompanied by a negative label. Thus, it should be treated as a positive-unlabeled learning problem instead of a PLL problem, whose learnability can be referred to [Denis, 1998]. In the dependent case, PML enjoys the same PAC-learnability as MDPL. In reality, PML is an untypical branch of MDPL, because only positive labels will be partially labeled.

Practical Implementation

According to our theoretical analysis, two MDPL algorithms are instantiated under our ERM framework. In independent case, we propose MDPL-BR that reduces an MDPL problem to multiple PLL tasks, which can be solved by any off-the-shelf PLL method. And we present the MDPL-CP method for dependent case. Note that by Eq. (7), partial hamming

Datasets	avg.#CLs[†]	MDPL-CP	MDPL-BR	MDPL-kNN	P-VLS	CoH
Puppy	1.3 1.1 1.4 1.4	**.603**±.047	.384±.033	.432±.043	.578±.084	.529±.073
Bridge	1 2 1 2 3	**.736**±.050	.367±.052	.461±.038	.659±.044	.473±.045
	1 2 2 2 3	**.673**±.041	.352±.033	.360±.025	.659±.053	.436±.024
	1 2 2 2 4	**.664**±.081	.359±.032	.364±.056	.646±.063	.418±.046
	1 2 2 2 5	**.609**±.023	.340±.062	.404±.041	.601±.034	.400±.045
Flare	3 2 1	**.942**±.013	.910±.010	.921±.006	.683±.125	.928±.008
	4 4 2	**.939**±.025	.862±.020	.906±.021	.331±.031	.922±.008
	6 4 2	**.928**±.051	.888±.011	.883±.015	.454±.070	.919±.009
	7 5 2	.902±.090	.837±.021	.868±.018	.576±.030	**.913**±.007
WQanimal	2 1 2 1 2 2 1	**.632**±.009	.614±.009	.604±.003	.622±.014	.601±.016
	2 2 2 2 2 2 2	**.631**±.016	.605±.005	.586±.008	.629±.018	.616±.009
	2 2 3 3 2 2 2	**.621**±.015	.577±.009	.566±.010	.616±.012	.578±.017
	3 3 3 3 3 3 3	.612±.011	.524±.022	.513±.010	**.622**±.018	.554±.015
WQplant	1 1 1 2 2 2 2	**.671**±.015	.644±.005	.638±.008	.659±.014	.615±.016
	2 2 2 2 2 2 2	**.660**±.006	.638±.003	.623±.005	.658±.008	.604±.018
	3 2 3 2 2 3 2	**.653**±.013	.601±.010	.579±.005	.643±.016	.596±.014
	3 3 3 3 3 3 3	**.648**±.010	.543±.007	.533±.013	.646±.013	.568±.011
Thyroid	2 2 1 1 1 1 1	.961±.002	**.962**±.001	.960±.001	.799±.014	.952±.016
	2 2 1 1 2 2 1	.960±.001	**.961**±.001	.960±.002	.717±.049	.954±.008
	3 3 1 1 2 2 1	**.959**±.001	.953±.003	.959±.001	.710±.022	.943±.004
	3 3 2 1 2 2 2	**.960**±.002	.896±.004	.958±.003	.746±.039	.949±.004

[†] Average number of candidate labels. Each configuration for a synthetic dataset demonstrates the average number of candidate labels on each dimension, respectively.

Table 2: Results of hamming accuracy on all datasets (mean±standard deviation). The best ones are in bold.

loss is also a surrogate loss to partial global loss. Therefore, we unify the two algorithms by minimizing the proposed partial hamming loss. To validate the theoretical results, we consider all the label combinations in the experiments. However, such a strategy may decrease the scalability of MDPL-CP. This problem can be alleviated by an ensemble technique [Tsoumakas *et al.*, 2011]. Due to the page limitation, we leave it for future work.

Datasets	N	m	d	K
Puppy	102	1,000	4	2-4
Bridges	108	7	5	2-6
Flare	1,066	10	3	3-8
WQanimal	1,060	16	7	4
WQplant	1,060	16	7	4
Thyroid	9,172	29	7	2-5

Table 3: Statistics of the experimental datasets.

5 Experiments

In this section, we evaluate the performance of our proposed methods on both synthetic and real-world dataset. All the computations are performed on a workstation with an i7-5930K CPU, a TITAN Xp GPU and 64GB main memory running Linux platform.

5.1 Dataset

Synthetic Datasets

We follow the experimental setting in [Wang *et al.*, 2019] and synthesize a total of 20 MDPL datasets from 5 real-world MDC datasets. The MDC datasets are collected from UCI repository [Dheeru and Karra Taniskidou, 2017]: 1) Bridges estimates bridge properties from specific constraints; 2) WQplant and WQanimals determine the plant and animal species in Slovenian rivers; 3) Flare predicts number of times that certain types of solar flare occurred within 24 hours period; 4) Thyroid determines types of thyroid problems based on patient information. For each class variable of an example, we randomly select some negative labels and aggregate them with the ground-truth label to obtain a candidate label set. Different configurations are controlled by the number of average candidate labels in each class space. The detailed information is reported in Table 3.

Puppy Dataset

Because the MDPL is a new learning setting, there is no publicly available MDPL dataset yet. To further boost our empirical studies, this paper builds one real-world MDPL dataset Puppy. A total of 102 dog images are collected and categorized to 4 class variables $\{$*Place, Tree, Dog Breeds, Weather*$\}$. We manually tagged all the data examples by ground-truth labels. The candidate label sets are collected by crowd-sourcing. Moreover, we extract 1000-dimensional fc-8 feature of these images using a pre-trained VGG-19 [Simonyan and Zisserman, 2015] model.

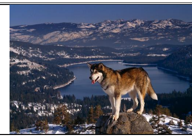
Images Examples		
Candidate Labels	Malamute/Husky No River Cloudy	Malamute/Husky Yes Mountain/River/Glacier Cloudy/Sunny
MDPL-CP Pred.	Malamute No River Sunny	Malamute Yes Mountain Sunny
MDPL-BR Pred.	Husky No River Sunny	Husky Yes Glacier Sunny
MDPL-kNN Pred.	Malamute No Glacier Sunny	Husky No Grassland Sunny
P-VLS Pred.	Samoyed No Grassland Sunny	Husky No Grassland Sunny
CoH Pred.	Malamute No River Sunny	Husky No Glacier Cloudy

Figure 1: Some MDPL image annotation examples on Puppy. For each image, we show the candidate labels, and the labels predicted by all the methods. The black labels denote the ground-truth or the correctly predicted ones. The red labels denote the distractor labels or wrongly predicted ones.

All the datasets are randomly split in to 80% training and 20% testing. We run five times on each dataset and the mean hamming accuracy with standard deviation are reported.

5.2 Baselines

We compared the proposed algorithms with three state-of-the-art baselines: 1) **P-VLS**: PARTICLE [Fang and Zhang, 2019] is an effective PML method that integrates the label propagation and calibrated label ranking techniques. By regarding each nominal label as a binary label, the MDPL problem can be transformed to a PML problem and solved by PARTICLE. In this work, we choose the virtual label splitting based version, i.e. P-VLS. 2)**CoH**: CoH [Shen *et al.*, 2018] is a label embedding based multi-label algorithm that jointly compresses the input and output to a latent space by co-hashing. We employ it to deal with MDPL tasks by treating all the candidate labels as valid ones. Note that P-VLS and CoH will return a group of positive labels in an uncertain size. Hence, we take the label with maximum score in its class space as our prediction. 3) **MDPL-kNN**: we induce a k-nearest neighbor model from MDPL data and the prediction is made by voting in each class space.

We use CLPL [Cour *et al.*, 2011] as the base PLL predictor in MDPL-BR method. For MDPL-CP, we choose linear classifier as the base hypothesis. In practice, we also adopt a naive convex surrogate loss in [Jin and Ghahramani, 2002] to implement MDPL-CP. For our methods, the empirical risk is optimized by stochastic gradient algorithm. We also add an l_2 regularization term. The learning rate and the regularization parameters are fine-tuned by cross-validation. The number of nearest neighbors is set as $k = 10$ for all the kNN-based approaches. Following the experimental setting in [Fang and Zhang, 2019], we set $thr = 0.9$ and $\alpha = 0.95$ for P-VLS. Finally, following [Shen *et al.*, 2018], the parameters of CoH are set as $\alpha = 100$ and $d = 10$.

5.3 Results

Table 2 lists the results of hamming accuracy of all the methods on Puppy and 20 synthetic MDPL datasets. Figure 1 shows some real predictive results on some test examples from Puppy dataset.

From the results, we can see that: 1) MDPL-CP algorithm achieves the best performance, which verifies our theoretical analysis. For instance, on Puppy dataset, MDPL-CP algorithm improves the best result of the baselines by 4.3%. By considering all the label combinations, it fully explores the label correlations with theoretically guaranteed disambiguation ability. 2) MDPL-BR works well on some datasets such as Thyroid. However, it generally underperforms MDPL-CP because of neglecting the correlations, which also further effects its disambiguating ability. 3) P-VLS and CoH are inferior to MDPL-CP method. Since they are designed for partial multi-label tasks, they will wrongly learn the label correlations of intra-class labels due to the false positive labels, which leads to degenerated performance. 4) Without considering both correlations and ambiguity, MDPL-kNN underperforms MDPL-CP and MDPL-BR. By these observations, we conclude that our proposed methods can effectively tackle the MDPL problems.

6 Conclusion

In this paper, we propose a novel learning paradigm named *Multi-Dimensional Partial Label Learning (MDPL)*, where each data instance is equipped with multiple candidate label sets. Based on our proposed *partial hamming loss*, we present an empirical risk minimization framework for MDPL. Theoretically, we rigorously prove the ERM learnability of MDPL in specific conditions. We further provide two effective MDPL algorithms under our ERM framework. In independent case, we propose MDPL-BR that decomposes the original task to a series of partial label learning problems. In dependent case, we propose MDPL-CP which fully explores the label correlations. Extensive experiments on both synthetic and real-world datasets validate our theoretical studies and the effectiveness of our proposed methods.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 61976161) and Key R&D Program of Zhejiang Province (Grant No. 2020C01024).

References

[Arias *et al.*, 2016] Jacinto Arias, José A. Gámez, Thomas D. Nielsen, and José Miguel Puerta. A scalable pairwise class interaction framework for multidimensional classification. *Int. J. Approx. Reasoning*, 68:194–210, 2016.

[Barutçuoglu *et al.*, 2006] Zafer Barutçuoglu, Robert E. Schapire, and Olga G. Troyanskaya. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22(7):830–836, 2006.

[Batal *et al.*, 2013] Iyad Batal, Charmgil Hong, and Milos Hauskrecht. An efficient probabilistic framework for multi-dimensional classification. In *CIKM*, pages 2417–2422, 2013.

[Benjumeda *et al.*, 2018] Marco Benjumeda, Concha Bielza, and Pedro Larrañaga. Tractability of most probable explanations in multidimensional bayesian network classifiers. *Int. J. Approx. Reasoning*, 93:74–87, 2018.

[Chapelle *et al.*, 2002] Olivier Chapelle, Jason Weston, and Bernhard Schölkopf. Cluster kernels for semi-supervised learning. In *NeurIPS*, pages 585–592, 2002.

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

[Denis, 1998] François Denis. PAC learning from positive statistical queries. In *ALT*, pages 112–126, 1998.

[Dheeru and Karra Taniskidou, 2017] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017.

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

[Feng and An, 2019] Lei Feng and Bo An. Partial label learning with self-guided retraining. In *AAAI*, pages 3542–3549. AAAI Press, 2019.

[Ishida *et al.*, 2017] Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. Learning from complementary labels. In *NeurIPS*, pages 5644–5654, 2017.

[Jia and Zhang, 2019] Bin-Bin Jia and Min-Ling Zhang. Multi-dimensional classification via knn feature augmentation. In *AAAI*, 2019.

[Jin and Ghahramani, 2002] Rong Jin and Zoubin Ghahramani. Learning with multiple labels. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *NeurIPS*, pages 897–904. MIT Press, 2002.

[Khosla *et al.*, 2011] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2, 2011.

[Kiryo *et al.*, 2017] Ryuichi Kiryo, Gang Niu, Marthinus Christoffel du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *NeurIPS*, pages 1674–1684, 2017.

[Liu and Dietterich, 2012] Li-Ping Liu and Thomas G. Dietterich. A conditional multinomial mixture model for superset label learning. In *NeurIPS*, pages 557–565, 2012.

[Liu and Tsang, 2017] Weiwei Liu and Ivor W. Tsang. Making decision trees feasible in ultrahigh feature and label dimensions. *J. Mach. Learn. Res.*, 18:81:1–81:36, 2017.

[Liu *et al.*, 2019] Weiwei Liu, Donna Xu, Ivor W. Tsang, and Wenjie Zhang. Metric learning for multi-output tasks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):408–422, 2019.

[Lv *et al.*, 2020] Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. *CoRR*, abs/2002.08053, 2020.

[Natarajan, 1989] B. K. Natarajan. On learning sets and functions. *Machine Learning*, 4:67–97, 1989.

[Ortigosa-Hernández *et al.*, 2012] Jonathan Ortigosa-Hernández, Juan Diego Rodríguez, Leandro Alzate, Manuel Lucania, Iñaki Inza, and José Antonio Lozano. Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers. *Neurocomputing*, 92:98–115, 2012.

[Read *et al.*, 2014] Jesse Read, Concha Bielza, and Pedro Larrañaga. Multi-dimensional classification with superclasses. *IEEE Trans. Knowl. Data Eng.*, 26(7):1720–1733, 2014.

[Shalev-Shwartz and Ben-David, 2014] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[Shen *et al.*, 2018] Xiaobo Shen, Weiwei Liu, Ivor W. Tsang, Quan-Sen Sun, and Yew-Soon Ong. Compact multi-label learning. In *AAAI*, pages 4066–4073, 2018.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[Tsoumakas *et al.*, 2011] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis P. Vlahavas. Random k-labelsets for multilabel classification. *IEEE Trans. Knowl. Data Eng.*, 23(7):1079–1089, 2011.

[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 Sarit Kraus, editor, *IJCAI*, pages 3691–3697. ijcai.org, 2019.

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

[Zaragoza *et al.*, 2011] Julio H. Zaragoza, Luis Enrique Sucar, Eduardo F. Morales, Concha Bielza, and Pedro Larrañaga. Bayesian chain classifiers for multidimensional classification. In *IJCAI*, pages 2192–2197, 2011.

[Zhan and Zhang, 2017] Wang Zhan and Min-Ling Zhang. Inductive semi-supervised multi-label learning with cotraining. In *KDD*, pages 1305–1314, 2017.

[Zhao and Guo, 2015] Feipeng Zhao and Yuhong Guo. Semi-supervised multi-label learning with incomplete labels. In *IJCAI*, pages 4062–4068, 2015.