Partial Multi-Label Optimal Margin Distribution Machine

Nan Cao , **Teng Zhang**[*] and **Hai Jin**
National Engineering Research Center for Big Data Technology and System
Services Computing Technology and System Lab, Cluster and Grid Computing Lab
School of Computer Science and Technology, Huazhong University of Science and Technology, China
{nan_cao, tengzhang, hjin}@hust.edu.cn

Abstract

Partial multi-label learning deals with the circumstance in which the ground-truth labels are not directly available but hidden in a candidate label set. Due to the presence of other irrelevant labels, vanilla multi-label learning methods are prone to be misled and fail to generalize well on unseen data, thus how to enable them to get rid of the noisy labels turns to be the core problem of partial multi-label learning. In this paper, we propose the *Partial Multi-Label Optimal margin Distribution Machine* (PML-ODM), which distinguishs the noisy labels through explicitly optimizing the distribution of ranking margin, and exhibits better generalization performance than minimum margin based counterparts. In addition, we propose a novel feature prototype representation to further enhance the disambiguation ability, and apply the non-linear kernels to handle the linearly inseparable data. Extensive experiments on real-world data sets validates the superiority of our proposed method.

1 Introduction

Multi-label learning is a supervised learning framework to cope with the problems where each instance is associated with more than one class label [Boutell *et al.*, 2004; Zhang and Zhou, 2013; Gibaja and Ventura, 2015]. Existing multi-label learning methods heavily rely on the high quality labeled data, which can be hardly satisfied in many real-world applications, because precise annotations are quite expensive and even impossible in some restricted scenarios. Instead, a set of noisy candidate labels are usually accessible. To deal with such imprecisely labeled data, the partial multi-label learning [Xie and Huang, 2018; Chen *et al.*, 2020], a unified framework of multi-label learning and partial label learning [Cour *et al.*, 2011; Feng and An, 2019], comes into being.

A trivial way to partial multi-label learning is to directly apply the vanilla multi-label learning methods [Zhang and Zhou, 2007; Feng *et al.*, 2019], but due to the presence of false positive labels, these methods are prone to be misled and generalize poorly on unseen data. One more sensible way is to adopt some elaborated disambiguation strategies to recover the ground-truth labels during training, which can be further divided into two groups. One is the low-rank models, which assumes the ground-truth label matrix is intrinsically low-rank, and the noisy label matrix is sparse [Sun *et al.*, 2019], or shares some low-rank subspace with feature matrix [Yu *et al.*, 2018]. Besides the low-rank and sparse decomposition on label matrix, the same assumption has also been made on the learner side [Xie and Huang, 2020], which leads to a joint learning of a low-rank multi-label classifier and a sparse noisy label identifier. However, the low-rank requirement sometimes is too rigorous to fulfill, and all these methods can hardly incorporate the non-linear kernels. The other is the confidence learning models, whose main idea is to evaluate the possibility of each label being a ground-truth. For example in [Zhang and Fang, 2020], an iterative label propagation is performed at first and only those with confidence value above the threshold are treated as credible labels to train the subsequent multi-label classifier; while in [Xie and Huang, 2018; Wang *et al.*, 2019], some smooth assumptions that highly correlated labels share similar confidence values or closest instances have similar ground-truth labels are adopted, and the confidence is learned through utilizing the local topological structure.

In this paper, we propose the *Partial Multi-Label Optimal margin Distribution Machine* (PML-ODM), which recovers the ground-truth labels via explicit optimization of the distribution of ranking margin. It is the generalization of ODM [Zhang and Zhou, 2019], a newly proposed learning framework rooting in margin theory [Gao and Zhou, 2013], thus it inherits the superiority and performs significantly better than the minimum margin based counterparts. Besides, we propose a novel feature prototype representation which is adaptively updated during training to further enhance the disambiguation ability. Extensive empirical studies show that PML-ODM can effectively identify the ground-truth labels, and achieve excellent generalization performance. Our method belongs to the second group mentioned above, but enjoys three advantages compared with previous works:

- It is the first attempt to introduce the powerful margin distribution into partial multi-label learning, and achieves significantly better generalization performance.
- It proposes a novel method to adaptively update the fea-

[*] Contact Author

ture prototype during training, rather than heuristically fixing it as constant ahead of time.

- It incorporates non-linear kernels to further enhance the generalization performance for linearly inseparable data.

The rest of paper is organized as follows. We first briefly introduce some preliminaries, and then detail the proposed method. After that, we derive the optimization procedure, followed by the empirical studies. Finally we conclude the paper with future work.

2 Preliminaries

For convenience, we first introduce some notations and terminologies used throughout the paper. We denote scalars with lower case letters (e.g., y), and vectors / matrix with boldface letters (e.g., $\boldsymbol{x}$ / $\mathbf{X}$). Sets are designated by upper case letters (e.g., Y), and in particular $[n] \triangleq \{1, 2, \ldots, n\}$.

Let X, Y, H denote the instance, label, hypothesis space respectively, and $D = \{(\boldsymbol{x}_1, \hat{Y}_1), (\boldsymbol{x}_2, \hat{Y}_2), \ldots, (\boldsymbol{x}_m, \hat{Y}_m)\}$ is a training set of size m drawn identically and independently (i.i.d.) according to some unknown distribution over $X \times Y$. The feature mapping associated with some positive definite kernel $\kappa(\cdot, \cdot)$ is denoted by $\phi : X \mapsto \mathbb{H}$. The hypothesis $h \in H$ is parameterized with n weight vectors $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_n \in \mathbb{H}$, each of which induces a scoring function $\boldsymbol{x} \mapsto \boldsymbol{w}_k^\top \phi(\boldsymbol{x})$ for $k \in [n]$.

2.1 Optimal Margin Distribution Learning

Margin is one of the most essential concepts in machine learning. Roughly speaking, it indicates the confidence of learning results. The larger the margin, the more confidence we have on the learner, and a negative margin signifies a wrong prediction. Formally, it is a function mapping from $X \times Y \times H$ to $\mathbb{R}$. Let $\bar{Y} = Y \setminus \hat{Y}$ denotes the irrelevant label set. For single-label learning problems where $\hat{Y} = \{y\}$ is a singleton, it is defined as

$$\gamma(\boldsymbol{x}, \hat{Y}, h) = \boldsymbol{w}_y^\top \phi(\boldsymbol{x}) - \max_{l \in \bar{Y}} \boldsymbol{w}_l^\top \phi(\boldsymbol{x}) \tag{1}$$

that is the smallest difference of scores between the ground-truth label and irrelevant labels. Particularly for binary classification problem where $Y = \{1, 2\}$, Eqn. (1) reduces to a more common form $\gamma(\boldsymbol{x}, \{y\}, h) = \hat{y}\boldsymbol{w}^\top \phi(\boldsymbol{x})$ where $\hat{y} = -2y + 3 \in \{\pm 1\}$ and $\boldsymbol{w} = \boldsymbol{w}_1 - \boldsymbol{w}_2$. For multi-label learning problem where $|\hat{Y}| \geq 2$, the following ranking margin is commonly considered,

$$\gamma(\boldsymbol{x}, \hat{Y}, h) = \boldsymbol{w}_k^\top \phi(\boldsymbol{x}) - \boldsymbol{w}_l^\top \phi(\boldsymbol{x}),\ \forall (k, l) \in \hat{Y} \times \bar{Y} \tag{2}$$

that is the difference of scores between each relevant and irrelevant label pair.

Recent studies on margin theory [Gao and Zhou, 2013] demonstrate an upper bound characterizing the relationship between the generalization and margin distribution, and the subsequent study on lower bound [Grønlund *et al.*, 2019] further proves that the upper bound is almost optimal up to a logarithmic factor, which means margin distribution is actually the essence of generalization. Inspired by these insightful works, a novel statistical learning framework named *Optimal margin Distribution Machine* (ODM) has appeared. Formally, it explicitly optimizes the distribution of margin by maximizing mean and minimizing variance simultaneously, that is

$$\begin{aligned} \min_{h, \xi_i, \epsilon_i} \ & \Omega(h) - \alpha\bar{\gamma} + \frac{\lambda}{2m} \sum_{i \in [m]} (\xi_i^2 + \epsilon_i^2) \\ \text{s.t.}\ & \bar{\gamma} - \xi_i \leq \gamma(\boldsymbol{x}_i, \hat{Y}_i, h) \leq \bar{\gamma} + \epsilon_i,\ \forall i \in [m] \end{aligned}$$

where $\Omega(h)$ is a regularization term to control model complexity, $\bar{\gamma}$ is the mean of margin, and α, λ are trading-off hyper-parameters. Note that the slack variables ξ_i and ϵ_i are deviations from margin mean, thus the summation in the last term is exactly the margin variance.

Due to the excellent generalization performance shown on both binary and multi-class classification tasks [Zhang and Zhou, 2014; Zhang and Zhou, 2017], many works extend ODM to some general learning settings. For example in [Tan *et al.*, 2020], a multi-label version of ODM is proposed. In addition to the cardinality, i.e., $|\hat{Y}|$, the accuracy of supervision is also commonly considered, and for most inaccurate supervised learning settings, ODM has already owned corresponding adaptions, just to name a few, clustering in which no supervision is provided [Zhang and Zhou, 2018a], semi-supervised learning in which only a fraction of instances have labels [Zhang and Zhou, 2018b], and multi-instance learning in which one label is just associated with a collection of instances [Zhang and Jin, 2020; Luan *et al.*, 2020]. Partial label learning [Xu *et al.*, 2019] is also a kind of inaccurate supervised learning, in which supervision is redundant with irrelevant labels. By further mixing up with the cardinality leads to the partial multi-label learning, which is much more difficult since the number of ground-truth labels is also unknown.

3 Proposed Method

In this section, we detail the PML-ODM. Following the same strategy in [Wang *et al.*, 2019], we incorporate a confidence matrix to distinguish the ground-truth labels. To be specific, let $p_{ik} \in [0, 1]$ denote the confidence that label k is a ground-truth label of $\boldsymbol{x}_i$, and $\mathbf{P} \in \mathbb{R}^{m \times n}$ denote the confidence matrix with p_{ik} as the (i, k)-th entry. Obviously, if k is an irrelevant label of $\boldsymbol{x}_i$, i.e., $k \in \bar{Y}_i$, we have $p_{ik} = 0$.

Different from multi-label learning in which ranking margin is only calculated for each relevant and irrelevant label pair, here we have two kinds of label pairs: 1) $(k, l) \in \hat{Y} \times \bar{Y}$, i.e., one is a candidate label and the other is irrelevant; 2) $(k, l) \in \hat{Y} \times \hat{Y}$, i.e., both are candidate labels. By putting the two kinds of label pairs together, and substituting the ranking margin into the formulation of ODM, we obtain:

$$\begin{aligned} \min_{\boldsymbol{w}, \boldsymbol{\xi}_i, \boldsymbol{\epsilon}_i} \ & \Omega(\boldsymbol{w}) - \alpha\bar{\gamma} + \frac{\lambda}{2m} \sum_{i \in [m]} \sum_{(k,l) \in Z_i} \frac{\tilde{p}_{ikl}(\xi_{ikl}^2 + \epsilon_{ikl}^2)}{|Z_i|} \\ \text{s.t.}\ & \boldsymbol{w}_k^\top \phi(\boldsymbol{x}_i) - \boldsymbol{w}_l^\top \phi(\boldsymbol{x}_i) \geq \bar{\gamma} - \xi_{ikl} \\ & \boldsymbol{w}_k^\top \phi(\boldsymbol{x}_i) - \boldsymbol{w}_l^\top \phi(\boldsymbol{x}_i) \leq \bar{\gamma} + \epsilon_{ikl},\ \forall (k, l) \in Z_i \end{aligned} \tag{3}$$

where $Z_i = \hat{Y}_i \times Y$ is the label pair set associated with $\boldsymbol{x}_i$, $\tilde{p}_{ikl} = \max(0, p_{ik} - p_{il})$ is the difference of confidence between label k and l, and the larger this value, the more important this label pair. From this perspective, $\tilde{p}_{ikl}$ is actually an attention which can filter out unimportant label pairs.

If the confidence matrix $\mathbf{P}$ is given by users in advance or can be summarized from domain knowledge, Eqn. (3) turns to a convex optimization problem and can be solved similarly as ODM. However in general $\mathbf{P}$ is unavailable, thus we treat it as a variable and learn it jointly with the multi-label classifier. Following the formulation in [Xie and Huang, 2018], let $\boldsymbol{q}_k$ denotes the feature prototype for the k-th class, we can obtain the following optimization problem:

$$\begin{aligned} \min_{\mathbf{P}} \ & \sum_{i\in[m]} \sum_{k\in\hat{Y}_i} p_{ik} \|\boldsymbol{x}_i - \boldsymbol{q}_k\| \\ \text{s.t.} \ & \sum_{k\in\hat{Y}_i} p_{ik} \geq 1, \ 0 \leq p_{ik} \leq 1, \ p_{il} = 0 \\ & \forall i \in [m], k \in \hat{Y}_i, l \in \bar{Y}_i \end{aligned} \tag{4}$$

where the constraint $\sum_{k\in\hat{Y}_i} p_{ik} \geq 1$ is to avoid the trivial solution $\mathbf{P} = \mathbf{0}$. In [Xie and Huang, 2018], the feature prototype $\boldsymbol{q}_k$ is simply set as the average of all instances with label k and kept unchanged in the subsequent training, which will undoubtedly hurt the generalization performance since the initialization involves noisy labels. Instead in this paper, we iteratively refine it. Specificly, given the current multi-label classifier, we can obtain the latest prediction $\boldsymbol{y}_i = [y_{i1}, y_{i2}, \ldots, y_{in}] \in \{0,1\}^n$ for $\boldsymbol{x}_i$, and according to which we collect all instances with label k and calculate the feature prototype as a weighted sum of them:

$$\boldsymbol{q}_k = \sum_{i\in V_k} \frac{c_{ik}}{\sum_{j\in V_k} c_{jk}} \boldsymbol{x}_i \tag{5}$$

where $V_k = \{i \in [m] \mid y_{ik} = 1\}$ is the index set of instances with label k, and the weight coefficient c_{ik} is determined by

$$c_{ik} = \left(y_{ik} + \sum_{t\in N_i} y_{tk} \frac{d_{\max} - d_t}{d_{\max} - d_{\min}} \right) / (|N_i| + 1) \tag{6}$$

where N_i is the index set of $\boldsymbol{x}_i$'s nearest neighbors, d_t denotes the distance between $\boldsymbol{x}_i$ and the neighbor $\boldsymbol{x}_t$, $d_{\max} = \max_{k\in N_i}\{d_k\}$ and $d_{\min} = \min_{k\in N_i}\{d_k\}$. The intuition behind Eqn. (6) is that the closer the instances, the more likely they share similar labels, which can also be regarded as the smooth constraint being able to avoid over-fitting.

By combining with Eqns. (3)-(4), we obtain the final formulation of PML-ODM:

$$\begin{aligned} \min_{\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{\epsilon},\mathbf{P}} \ & \frac{1}{2} \sum_{s\in[n]} \|\boldsymbol{w}_s\|_{\mathbb{H}}^2 + \frac{\lambda_1}{2m} \sum_{i\in[m]} \sum_{(k,l)\in Z_i} \frac{\tilde{p}_{ikl}(\xi_{ikl}^2 + \mu\epsilon_{ikl}^2)}{|Z_i|} \\ & + \lambda_2 \sum_{i\in[m]} \sum_{k\in\hat{Y}_i} p_{ik} \|\boldsymbol{x}_i - \boldsymbol{q}_k\| \\ \text{s.t.} \ & \boldsymbol{w}_k^\top \phi(\boldsymbol{x}_i) - \boldsymbol{w}_l^\top \phi(\boldsymbol{x}_i) \geq 1 - \theta - \xi_{ikl} \\ & \boldsymbol{w}_k^\top \phi(\boldsymbol{x}_i) - \boldsymbol{w}_l^\top \phi(\boldsymbol{x}_i) \leq 1 + \theta + \epsilon_{ikl} \\ & \sum_{k\in\hat{Y}_i} p_{ik} \geq 1, \ 0 \leq p_{ik} \leq 1, \ p_{il} = 0 \\ & \forall i \in [m], k \in \hat{Y}_i, l \in \bar{Y}_i \end{aligned} \tag{7}$$

Here we follow the same processing as ODM, i.e., scaling $\boldsymbol{w}$ to make the mean $\bar{\gamma}$ as 1, introducing hyper-parameters μ and θ to enhance the model's capability and flexibility, and using the ℓ_2 regularization to control the model complexity.

Once we obtain the solution of Eqn. (7), i.e., $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$, we can calculate the scores for any unseen data, but to predict whether a label is relevant or irrelevant, an extra threshold is required. Here we determine the threshold t_k for any label k by minimizing the overall misclassification rate. To be specific, for any training instance $(\boldsymbol{x}_i, \hat{Y}_i)$, if $k \in \hat{Y}_i$ yet $\boldsymbol{w}_k^\top \phi(\boldsymbol{x}_i) < t_k$, or $k \in \bar{Y}_i$ yet $\boldsymbol{w}_k^\top \phi(\boldsymbol{x}_i) > t_k$, a misclassification occurs. We determine t_k by minimizing overall misclassification rate, which can be solved by sorting $\{\boldsymbol{w}_k^\top \phi(\boldsymbol{x}_i)\}_{i\in[m]}$. Without loss of generality, let us assume $\boldsymbol{w}_k^\top \phi(\boldsymbol{x}_1) \leq \boldsymbol{w}_k^\top \phi(\boldsymbol{x}_2) \leq \cdots \leq \boldsymbol{w}_k^\top \phi(\boldsymbol{x}_m)$, then t_k can be determined by checking the middle value of $m-1$ intervals $[\boldsymbol{w}_k^\top \phi(\boldsymbol{x}_i), \boldsymbol{w}_k^\top \phi(\boldsymbol{x}_{i+1})]_{i\in[m-1]}$.

4 Optimization

Due to the coupling of $\mathbf{P}, \boldsymbol{\xi}, \boldsymbol{\epsilon}$ in the second term, Eqn. (7) is difficult to optimize directly, thus we resort to the alternating optimization strategy, i.e., in each iteration, we first solve $\boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\epsilon}$ by fixing $\mathbf{P}$, then update $\boldsymbol{q}$ as described previously, and finally solve $\mathbf{P}$ with the latest $\boldsymbol{\xi}, \boldsymbol{\epsilon}, \boldsymbol{q}$. As for initialization, we can first calculate c_{ik} as Eqn. (6) and appropriately scale it as p_{ik}; on the other hand $\boldsymbol{q}_k$ can be initialized by Eqn. (5). The above steps are performed iteratively until convergence or the maximum number of iterations is reached. Algorithm 1 summarizes the whole procedure.

4.1 Subproblem

When fixing $\mathbf{P}$, Eqn. (7) turns to a quadratic programming problem. Due to the underlying infinite dimensional feature mapping $\phi(\cdot)$, it is usually cast in the dual form. By introducing the dual variables α_{ikl} and β_{ikl}, the Lagrangian function

Algorithm 1 PML-ODM

```
Input: data set $D = \{(\boldsymbol{x}_i, \hat{Y}_i)\}_{i\in[m]}$, hyper-parameters $\mu, \theta, \lambda_1, \lambda_2$, nearest neighbors number $|N|$, maximum iteration number $T$.
Initialize: confidence matrix $\mathbf{P}$, feature prototype $\boldsymbol{q}$, $t \leftarrow 0$.
while $t < T$ and not converge do
    Optimize $\boldsymbol{w}$, $\boldsymbol{\xi}$ and $\boldsymbol{\epsilon}$ by fixing $\mathbf{P}$;
    Obtain the indicator vector $\boldsymbol{y}_i$ according to $\boldsymbol{w}$;
    Calculate weight coefficient $\boldsymbol{c}_i$ as Eqn. (6);
    Update $\boldsymbol{q}_k$ as Eqn. (5);
    Optimize $\mathbf{P}$ by fixing $\boldsymbol{w}$, $\boldsymbol{\xi}$ and $\boldsymbol{\epsilon}$;
end while
Output: $\boldsymbol{w}_s$ for $s \in [n]$.
```

can be written as:

$$\begin{aligned} L = &\frac{1}{2}\sum_{s\in[n]}\|\boldsymbol{w}_s\|_{\mathbb{H}}^2 + \frac{\lambda_1}{2m}\sum_{i\in[m]}\sum_{(k,l)\in Z_i}\frac{\tilde{p}_{ikl}(\xi_{ikl}^2+\mu\epsilon_{ikl}^2)}{|Z_i|} \\ &-\sum_{i\in[m]}\sum_{(k,l)\in Z_i}\alpha_{ikl}((\boldsymbol{w}_k-\boldsymbol{w}_l)^\top\phi(\boldsymbol{x}_i)-1+\theta+\xi_{ikl}) \\ &+\sum_{i\in[m]}\sum_{(k,l)\in Z_i}\beta_{ikl}((\boldsymbol{w}_k-\boldsymbol{w}_l)^\top\phi(\boldsymbol{x}_i)-1-\theta-\epsilon_{ikl}) \end{aligned}$$

The partial derivatives of L w.r.t. $\boldsymbol{w}_s, \xi_{ikl}, \epsilon_{ikl}$ are

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{w}_s} &= \boldsymbol{w}_s - \sum_{i\in[m]} 1_{s\in\hat{Y}_i}\sum_{l\in Y}(\alpha_{isl}-\beta_{isl})\phi(\boldsymbol{x}_i) \\ &\quad + \sum_{i\in[m]}\sum_{k\in\hat{Y}_i}(\alpha_{iks}-\beta_{iks})\phi(\boldsymbol{x}_i) \\ &= \boldsymbol{w}_s - \sum_{i\in[m]}\sum_{(k,l)\in Z_i}(\alpha_{ikl}-\beta_{ikl})d_{ikl}^{(s)}\phi(\boldsymbol{x}_i), \end{aligned}$$

$$\frac{\partial L}{\partial \xi_{ikl}} = \frac{\lambda_1\tilde{p}_{ikl}\xi_{ikl}}{m|Z_i|} - \alpha_{ikl}, \quad \frac{\partial L}{\partial \epsilon_{ikl}} = \frac{\lambda_1\tilde{p}_{ikl}\mu\epsilon_{ikl}}{m|Z_i|} - \beta_{ikl}$$

where $d_{ikl}^{(s)} = 1_{k=s} - 1_{l=s}$. By setting these partial derivatives to zero, we have

$$\begin{aligned} \boldsymbol{w}_s &= \sum_{i\in[m]}(\boldsymbol{\alpha}_i-\boldsymbol{\beta}_i)^\top\boldsymbol{d}_i^{(s)}\phi(\boldsymbol{x}_i) \\ \xi_{ikl} &= \frac{m|Z_i|}{\lambda_1\tilde{p}_{ikl}}\alpha_{ikl}, \quad \epsilon_{ikl} = \frac{m|Z_i|}{\lambda_1\tilde{p}_{ikl}\mu}\beta_{ikl} \end{aligned}$$

where $\boldsymbol{\alpha}_i = [\alpha_{ikl}]_{(k,l)\in Z_i}$, $\boldsymbol{\beta}_i = [\beta_{ikl}]_{(k,l)\in Z_i}$, $\boldsymbol{d}_i^{(s)} = [d_{ikl}^{(s)}]_{(k,l)\in Z_i}$, By substituting into the Lagrangian function, we can obtain the dual problem:

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \ &\frac{1}{2}\sum_{s\in[n]}\sum_{i,j\in[m]}(\boldsymbol{\alpha}_i-\boldsymbol{\beta}_i)^\top\boldsymbol{d}_i^{(s)}(\boldsymbol{\alpha}_j-\boldsymbol{\beta}_j)^\top\boldsymbol{d}_j^{(s)}\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) \\ &+\sum_{i\in[m]}((\theta-1)\boldsymbol{\alpha}_i^\top\boldsymbol{e}_i+(\theta+1)\boldsymbol{\beta}_i^\top\boldsymbol{e}_i) \\ &+\sum_{i\in[m]}\frac{m|Z_i|}{2\lambda_1}\left(\boldsymbol{\alpha}_i^\top\tilde{\mathbf{P}}\boldsymbol{\alpha}_i+\frac{\boldsymbol{\beta}_i^\top\tilde{\mathbf{P}}\boldsymbol{\beta}_i}{\mu}\right) \\ \text{s.t. } &\boldsymbol{\alpha}_i \geq \mathbf{0}, \ \boldsymbol{\beta}_i \geq \mathbf{0}, \ \forall i\in[m] \end{aligned}$$

where $\boldsymbol{e}_i \in \mathbb{R}^{|Z_i|}$ is the all one vector, and $\tilde{\mathbf{P}} \in \mathbb{R}^{|Z_i|\times|Z_i|}$ is the diagonal matrix with $1/\tilde{p}_{ikl}$ as the diagonal element.

The dual problem is a convex quadratic programming. Notice that all the variables are decoupled and only have a lower bound constraint, therefore it can be efficiently solved by the coordinate descent method [Hsieh *et al.*, 2008]. To be specific, in each iteration, only one variable is selected to minimize while other variables are kept as constants, and a closed-form solution can be achieved, finally this procedure is repeated until convergence.

When fixing $\boldsymbol{w}$, $\boldsymbol{\xi}$ and $\boldsymbol{\epsilon}$, Eqn. (7) turns to a linear programming problem, whose solution has already matured attribute to the intensive studies for decades. Thus we simply invoke the off-the-shelf solver Mosek, which integrates the effective interior point method.

5 Experiments

In this section, we empirically evaluate the effectiveness of our proposed method.

5.1 Setting

We conduct the experiments on eight real-world multi-label data sets[1] which come from a broad range of field. Since partial multi-label learning is a recently proposed learning framework, the customized public data sets are not available yet, thus for each data set, we randomly add $\eta \in \{1, 2, 3\}$ noisy labels to candidate label set and repeat five times to performe the experiments. The average values as well as the standard deviations are recorded. All the data sets with their basic statistics are listed in Table 1.

Data sets	#Ins.	#Fea.	#Lab.	avg#CL	avg#GL
Birds	645	260	19	2,3,4	1.01
Emotions	593	72	6	2,3,4	1.87
Scene	2407	294	6	2,3,4	1.07
Flags	194	10	7	4,5,6	3.39
Yeast	2417	103	14	5,6,7	4.24
Genbase	662	1186	27	2,3,4	1.25
Medical	978	1449	45	2,3,4	1.25
Enron	1702	1001	53	4,5,6	3.38

Table 1: Experimental data sets with their basic statistics, avg#CL and avg#GL indicate the average number of candidate labels and ground-truth labels respectively

To demonstrate the superiority of the proposed PML-ODM, we compare it with two confidence learning models PAR-VLS and PAR-MAP [Zhang and Fang, 2020], and one low-rank model PML-LRS [Sun *et al.*, 2019]. We also compare PML-ODM with the mlODM [Tan *et al.*, 2020], the state-of-the-art multi-label learning method which also optimizes the margin distribution.

The parameters of PAR-VAL and PAR-MAP are set as suggested in its paper, i.e., balancing parameter $\alpha = 0.95$ and credible label elicitation threshold $t = 0.9$. For our method $k = 10$ and the width of RBF kernel is selected from the set $\{2^{-10}, 2^{-9}, \ldots, 2^3\}$. For the ODM based method, the parameters μ and θ are selected from the set $\{0.1, 0.2, \ldots, 0.9\}$. All the parameters are selected by 5-fold cross validation.

We take the *Ranking Loss* , *Hamming Loss* , *Coverage* , *Average Precision* , *Macro-F1* and *Micro-F1* as the performance evaluation metrics.

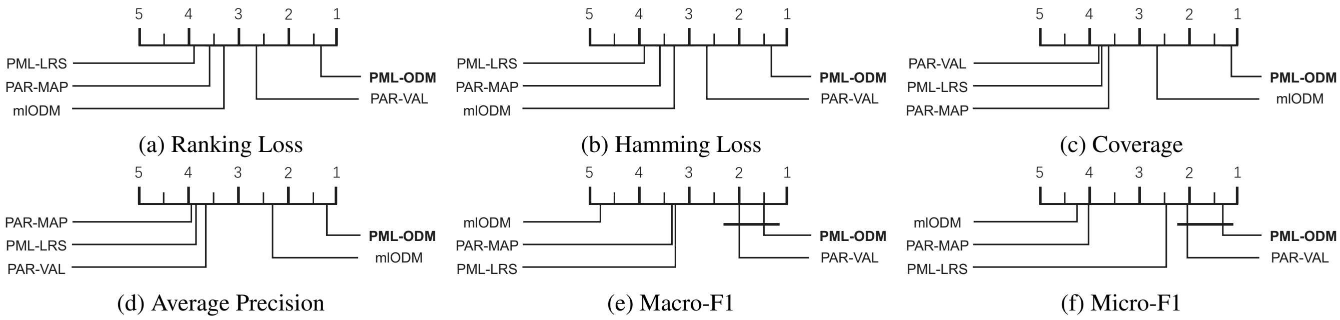
5.2 Results

Table 2 summarizes the performance of five methods on eight data sets in terms of six evaluation metrics. From the results we can see that our proposed PML-ODM outperforms most of the baselines and achieves best performance in most cases. Among the five methods, PML-ODM beats other methods on three data sets Emotions, Scene, Genbase in terms of

[1] http://palm.seu.edu.cn/zhangml/ and http://mulan.sourceforge.net/datasets-mlc.html

Data sets	PML-ODM	PAR-VAL	PAR-MAP	PML-LRS	mlODM	PML-ODM	PAR-VAL	PAR-MAP	PML-LRS	mlODM
			$\eta = 1$					$\eta = 2$		
					Ranking Loss ↓					
Birds	.217±.011	**.197±.035**	.332±.027●	.317±.024●	.227±.017	**.243±.007**	.282±.023●	.289±.024●	.351±.014●	.244±.013
Emotions	**.177±.009**	.181±.019	.189±.011●	.304±.015●	.181±.011	**.212±.003**	.261±.007●	.281±.014●	.313±.027●	.227±.014●
Scene	**.079±.008**	.097±.023●	.201±.018●	.091±.013	.103±.015●	**.117±.023**	.169±.018●	.282±.023●	.119±.008	.119±.011
Flags	**.219±.004**	.283±.021●	.229±.024	.302±.022●	.258±.009●	**.225±.009**	.385±.021●	.246±.019	.303±.019●	.331±.017●
Yeast	**.169±.004**	.193±.012●	.183±.016●	.371±.009●	.189±.018●	**.168±.009**	.194±.009●	.182±.021	.267±.011●	.184±.017●
Genbase	**.002±.001**	.039±.014●	.061±.004●	.002±.011	.003±.002	**.004±.002**	.043±.013●	.067±.009●	.005±.002	.007±.005
Medical	**.081±.005**	.115±.027●	.112±.008●	.094±.018●	.091±.006●	.113±.014	.137±.011●	.132±.018●	.112±.012	**.107±.009**
Enron	.146±.014	.231±.018●	.169±.017●	.179±.008●	.161±.019●	.133±.017	.253±.014●	.143±.018●	.189±.015●	.172±.018●
w/t/l		6/2/0	7/1/0	6/2/0	5/3/0		8/0/0	6/2/0	5/3/0	4/4/0
					Hamming Loss ↓					
Birds	**.078±.004**	.115±.037●	.127±.019●	.088±.012	.089±.008	**.094±.004**	.118±.016●	.123±.014●	.125±.021●	.112±.006●
Emotions	**.274±.013**	.277±.009	.291±.029●	.322±.014●	.287±.017●	.392±.005	.419±.013●	.424±.016●	**.301±.017**○	.479±.013●
Scene	**.151±.007**	.178±.012●	.162±.009	.314±.018●	.159±.011	**.167±.017**	.177±.013	.266±.008●	.279±.013●	.496±.021●
Flags	**.334±.009**	.479±.017●	.357±.012●	.371±.011●	.351±.015●	**.424±.013**	.473±.019●	.454±.013	.462±.014●	.478±.016●
Yeast	.261±.023	.279±.026●	.263±.017	.324±.023●	**.228±.014**○	.274±.013	.297±.012●	.294±.021●	.288±.027●	**.261±.011**
Genbase	**.014±.003**	.017±.004	.045±.016●	.032±.009●	.021±.007●	.043±.017	.062±.023●	.062±.011●	.041±.011	**.033±.012**
Medical	**.027±.005**	.041±.019●	.084±.017●	.094±.019●	.052±.011●	**.028±.012**	.073±.026●	.088±.017●	.121±.009●	.092±.009●
Enron	**.059±.008**	.072±.016●	.081±.013●	.075±.011●	.091±.027●	**.008±.004**	.113±.008●	.117±.006●	.132±.014●	.109±.017●
w/t/l		6/2/0	6/2/0	7/1/0	5/2/1		7/1/0	7/1/0	6/1/1	6/2/0
					Coverage ↓					
Birds	**.119±.024**	.121±.033	.157±.019●	.176±.011●	.153±.012●	**.142±.003**	.143±.021	.163±.022●	.193±.024●	.171±.018●
Emotions	**.342±.005**	.349±.011	.367±.012●	.441±.018●	.348±.019	**.350±.007**	.401±.011●	.401±.018●	.441±.015●	.372±.009●
Scene	**.112±.009**	.132±.034●	.173±.018●	.117±.014	.113±.008	**.119±.014**	.122±.017	.238±.011●	.119±.018	.121±.017
Flags	.544±.007	.668±.018●	.559±.019	.486±.021	.564±.013●	**.551±.017**	.742±.014●	.589±.006●	.631±.023●	.649±.014●
Yeast	**.455±.017**	.482±.019●	.479±.014●	.659±.011●	.475±.021●	**.457±.009**	.481±.009●	.478±.021●	.647±.024●	.481±.019●
Genbase	**.012±.007**	.078±.013●	.097±.004●	.044±.002●	.022±.009●	**.016±.006**	.087±.014●	.183±.019●	.048±.019●	.031±.012●
Medical	.121±.005	.132±.012	.133±.007	.146±.018●	.124±.022	**.118±.009**	.142±.018●	.137±.009●	.142±.014●	.127±.011●
Enron	.419±.023	.461±.011●	.425±.016	.454±.022●	.437±.024●	.319±.011	.451±.015●	.324±.014	.373±.021●	.392±.023●
w/t/l		5/3/0	5/3/0	6/2/0	5/3/0		6/2/0	7/1/0	7/1/0	7/1/0
					Average Precision ↑					
Birds	.513±.009	.412±.032●	.409±.031●	.397±.022●	**.521±.014**	.480±.013	.389±.032●	.382±.018●	.381±.017●	**.509±.013**
Emotions	**.791±.012**	.771±.012●	.773±.019●	.648±.009●	.788±.003	**.760±.002**	.734±.021●	.681±.013●	.647±.018●	.744±.007●
Scene	**.894±.023**	.893±.017	.718±.024●	.827±.011●	.837±.012●	.818±.024	.768±.016●	.601±.026●	.796±.023	**.823±.019**
Flags	**.831±.014**	.811±.022●	.816±.019●	.796±.017●	.778±.011●	**.795±.011**	.728±.022●	.783±.019	.753±.021●	.741±.013●
Yeast	**.771±.011**	.755±.013●	.738±.018●	.587±.021●	.741±.016●	**.762±.009**	.751±.017	.738±.017●	.601±.009●	.735±.015●
Genbase	**.994±.005**	.965±.012●	.849±.003●	.981±.006	.982±.014	**.987±.007**	.956±.013●	.841±.008●	.982±.007	.983±.021
Medical	**.703±.005**	.569±.024●	.504±.016●	.682±.018●	.662±.024●	**.644±.004**	.595±.012●	.525±.012●	.618±.017●	.623±.024●
Enron	.559±.021	.488±.019●	.531±.011●	.509±.021●	.544±.019●	**.629±.005**	.451±.014●	.524±.018	.479±.011●	.531±.011●
w/t/l		7/1/0	8/0/0	7/1/0	5/3/0		7/1/0	6/2/0	6/2/0	5/3/0
					Macro-F1 ↑					
Birds	**.531±.011**	.523±.016●	.275±.021●	.443±.016●	.297±.024●	**.485±.014**	.456±.022●	.199±.019●	.332±.013●	.323±.022●
Emotions	**.697±.007**	.684±.008	.491±.013●	.677±.009●	.672±.015●	**.639±.009**	.616±.011●	.474±.017●	.607±.021●	.578±.017●
Scene	.821±.004	**.834±.017**	.718±.019●	.803±.007●	.505±.014●	.772±.015	**.797±.011**○	.614±.019●	.701±.025●	.419±.021●
Flags	**.691±.011**	.645±.019●	.667±.013●	.632±.013●	.625±.017●	**.677±.013**	.597±.024●	.651±.021	.619±.018●	.613±.019●
Yeast	.773±.016	.808±.022○	**.845±.026**○	.639±.019●	.508±.019●	.644±.012	.627±.021	**.648±.026**	.601±.021●	.479±.018●
Genbase	.924±.009	.929±.013	.911±.017●	**.937±.008**	.884±.011●	**.889±.007**	.878±.016	.885±.009	.837±.012●	.814±.015●
Medical	**.673±.013**	.659±.012●	.494±.013●	.379±.022●	.198±.028●	**.585±.006**	.556±.011●	.462±.015●	.469±.016●	.272±.022●
Enron	.701±.022	**.726±.018**	.418±.021●	.438±.026●	.218±.021●	.667±.019	**.704±.015**●	.512±.018●	.497±.019●	.218±.023●
w/t/l		3/4/1	7/0/1	7/1/0	8/0/0		5/2/1	5/3/0	8/0/0	8/0/0
					Micro-F1 ↑					
Birds	**.537±.008**	.381±.014●	.237±.019●	.344±.018●	.341±.023●	**.471±.011**	.337±.018●	.179±.015●	.322±.019●	.329±.019●
Emotions	.711±.013	**.724±.005**●	.531±.009●	.541±.011●	.679±.021●	**.677±.013**	.668±.014	.503±.013●	.599±.017●	.579±.019●
Scene	**.793±.006**	.785±.013	.424±.017●	.705±.014●	.521±.018●	**.712±.017**	.708±.017	.464±.024●	.695±.024●	.417±.022●
Flags	**.747±.011**	.561±.023●	.544±.027●	.619±.019●	.712±.015●	**.721±.016**	.689±.021●	.609±.017●	.603±.026●	.659±.021●
Yeast	.698±.015	**.707±.017**	.471±.021●	.418±.024●	.677±.019	.683±.009	**.694±.023**	.473±.013●	.651±.018●	.652±.017●
Genbase	**.954±.009**	.922±.015●	.847±.009●	.929±.011	.891±.019●	.907±.008	**.916±.013**●	.844±.011●	.901±.014	.854±.014●
Medical	**.709±.008**	.572±.019●	.466±.016●	.419±.023●	.169±.022●	**.664±.009**	.622±.014●	.426±.014●	.417±.018●	.211±.025●
Enron	**.471±.025**	.448±.016●	.352±.013●	.331±.022●	.164±.025●	.457±.021	**.464±.019**	.416±.023●	.356±.017●	.158±.021●
w/t/l		6/2/0	8/0/0	7/1/0	7/1/0		4/4/0	8/0/0	7/1/0	8/0/0

Table 2: Results on eight data sets in terms of six evaluation metrics, the ↓ indicates that the smaller the value, the better the performance, and vice versa. The best result on each data set is bolded. ●/○ indicates the performance of PML-ODM is significantly better/worse than the compared method (pairwise t-test at 0.05 significance level). The win/tie/loss counts for PML-ODM are summarized in the last row.



(a) Ranking Loss (b) Hamming Loss (c) Coverage

(d) Average Precision (e) Macro-F1 (f) Micro-F1

Figure 1: Bonferroni-Dunn test of PML-ODM. Methods not connected with ours are considered to have significant difference with PML-ODM (CD = 1.1401 at 0.05 significance level).

nearly all the evaluation metrics, and on two data sets Medical and Flags in terms of *Average Precision*, *Macro-F1* and *Micro-F1*. Compared to mlODM, PML-ODM is much more robust in resisting the noisy labels, and achieves better performance in most of the experiments.

To verify the superiority of our proposed feature prototype, we compare PML-ODM with its variant PML-ODM^f, which fixs the feature prototype as the average of all the instances. Experiments are performed on data sets Emotions, Scene, Flags, each of which contains 1 noisy label, and we take the *Average Precision* (AP) and *One Error* (OE) as evaluation metrics. As shown in Table 3, we can find that PML-ODM achieves higher *Average Precision* (AP) and lower *One Error* (OE) than PML-ODM^f, which validates the superiority of our proposed feature prototype.

Prototype	Emotions		Scene		Flags	
	AP	OE	AP	OE	AP	OE
PML-ODM^f	.717	.411	.782	.368	.782	.247
PML-ODM	.771	.322	.803	.336	.815	.184

Table 3: Feature prototype comparison test

To analyze their relative performance, we also conduct the Friedman test. Table 4 shows the Friedman statistical results as well as the corresponding critical value with respect to six evaluation criteria, from the results we can find that all the Friedman statistical values are greater than the critical value which indicates that the performance of PML-ODM is remarkably different from other methods. For each criterion, the null hypothesis of distinguishable performance within the four baselines is rejected at 0.05 significance level.

Furthermore, the post-hoc Bonferroni-Dunn test is utilized to analyze the relative performance of our PML-ODM with other baselines. The control method is our PML-ODM, and the difference of average ranking between control method and other methods will be calibrated with the *Critical Difference* (CD). The PML-ODM is believed to have prominently different performance to another baselines if the average ranking gap is larger than one CD. Figure 1 displays the results of Bonferroni-Dunn test on six evaluation metrics. For each baselines, the average ranking in terms of each evaluation metric is marked on the axis, and for the methods whose distance to the control method are less than one CD, we use the thick line to connect them with PML-ODM. Obviously, PML-ODM significantly outperforms most of the baselines.

Evaluation Metric	Friedman Statistics	Critical Value
Hamming Loss	15.7653	2.4982
Ranking Loss	20.3861	
Coverage	25.3154	
Average Precision	29.1573	
Macro-F1	46.9472	
Micro-F1	55.1361	

Table 4: Friedman statistics in terms of evaluation metric and the critical value at 0.05 significance level (baselines $k = 5$, data sets $N = 24$)

6 Conclusion

In this paper, we propose the *Partial Multi-Label Optimal margin Distribution Machine* (PML-ODM). To sum up, our contributions are threefold: 1) we first introduce the powerful margin distribution into partial multi-label learning; 2) we propose an ingenious method to adaptively refine the feature prototype during the training process rather than fixing it as constant; 3) we leverage non-linear kernels to further improve the generalization performance for linearly inseparable data. In the future, we will make some theoretical analysis on our proposed method.

Acknowledgments

This work was supported by NSFC (62006088).

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 *et al.*, 2020] Ze-Sen Chen, Xuan Wu, Qing-Guo Chen, Yao Hu, and Min-Ling Zhang. Multi-view partial multi-label learning with graph-based disambiguation. In

Proceedings of the 34th AAAI Conference on Artificial Intelligence, pages 3553–3560, New York, NY, 2020.

[Cour *et al.*, 2011] Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *The Journal of Machine Learning Research*, 12(4):1501–1536, 2011.

[Feng and An, 2019] Lei Feng and Bo An. Partial label learning by semantic difference maximization. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2294–2300, Macao, China, 2019.

[Feng *et al.*, 2019] Lei Feng, Bo An, and Shuo He. Collaboration based multi-label learning. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pages 3550–3557, Hawaii, HI, 2019.

[Gao and Zhou, 2013] Wei Gao and Zhi-Hua Zhou. On the doubt about margin explanation of boosting. *Artificial Intelligence*, 203:1–18, 2013.

[Gibaja and Ventura, 2015] Eva Gibaja and Sebastián Ventura. A tutorial on multilabel learning. *ACM Computing Surveys*, 47(3):1–38, 2015.

[Grønlund *et al.*, 2019] Allan Grønlund, Lior Kamma, Kasper Green Larsen, Alexander Mathiasen, and Jelani Nelson. Margin-based generalization lower bounds for boosted classifiers. In *Advances in Neural Information Processing Systems 32*, pages 11963–11972, Vancouver, Canada, 2019.

[Hsieh *et al.*, 2008] Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S. Sathiya Keerthi, and Sellamanickam Sundararajan. A dual coordinate descent method for large-scale linear svm. In *Proceedings of the 25th International Conference on Machine Learning*, pages 408–415, Helsinki, Finland, 2008.

[Luan *et al.*, 2020] Tianxiang Luan, Tingjin Luo, Wenzhang Zhuge, and Chenping Hou. Optimal representative distribution margin machine for multi-instance learning. *IEEE Access*, 8:74864–74874, 2020.

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

[Tan *et al.*, 2020] Zhi-Hao Tan, Peng Tan, Yuan Jiang, and Zhi-Hua Zhou. Multi-label optimal margin distribution machine. *Machine Learning*, 109(3):623–642, 2020.

[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 *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3691–3697, Macao, China, 2019.

[Xie and Huang, 2018] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, page 4302–4309, New Orleans, LA, 2018.

[Xie and Huang, 2020] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning with noisy label identification. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 6454–6461, New York, NY, 2020.

[Xu *et al.*, 2019] Ning Xu, Jiaqi Lv, and Xin Geng. Partial label learning via label enhancement. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pages 5557–5564, Hawaii, HI, 2019.

[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 *Proceedings of the 18th IEEE International Conference on Data Mining*, pages 1398–1403, Singapore, 2018.

[Zhang and Fang, 2020] Min-Ling Zhang and Jun-Peng Fang. Partial multi-label learning via credible label elicitation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(1):99, 2020.

[Zhang and Jin, 2020] Teng Zhang and Hai Jin. Optimal margin distribution machine for multi-instance learning. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, pages 2383–2389, Yokohama, Japan, 2020.

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

[Zhang and Zhou, 2013] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2013.

[Zhang and Zhou, 2014] Teng Zhang and Zhi-Hua Zhou. Large margin distribution machine. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 313–322, New York, NY, 2014.

[Zhang and Zhou, 2017] Teng Zhang and Zhi-Hua Zhou. Multi-class optimal margin distribution machine. In *Proceedings of the 34th International Conference on Machine Learning*, pages 4063–4071, Sydney, Australia, 2017.

[Zhang and Zhou, 2018a] Teng Zhang and Zhi-Hua Zhou. Optimal margin distribution clustering. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 4474–4481, New Orleans, LA, 2018.

[Zhang and Zhou, 2018b] Teng Zhang and Zhi-Hua Zhou. Semi-supervised optimal margin distribution machines. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 3104–3110, Stockholm, Schweden, 2018.

[Zhang and Zhou, 2019] Teng Zhang and Zhi-Hua Zhou. Optimal margin distribution machine. *IEEE Transactions on Knowledge and Data Engineering*, 32(6):1143–1156, 2019.