Partial Multi-Label Learning with Noisy Label Identification

Ming-Kun Xie, Sheng-Jun Huang*
College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
MIIT Key Laboratory of Pattern Analysis and Machine Intelligence
Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, 211106
{mkxie, huangsj}@nuaa.edu.cn

Abstract

Partial multi-label learning (PML) deals with problems where each instance is assigned with a candidate label set, which contains multiple relevant labels and some noisy labels. Recent studies usually solve PML problems with the disambiguation strategy, which recovers ground-truth labels from the candidate label set by simply assuming that the noisy labels are generated randomly. In real applications, however, noisy labels are usually caused by some ambiguous contents of the example. Based on this observation, we propose a partial multi-label learning approach to simultaneously recover the ground-truth information and identify the noisy labels. The two objectives are formalized in a unified framework with trace norm and ℓ_1 norm regularizers. Under the supervision of the observed noise-corrupted label matrix, the multi-label classifier and noisy label identifier are jointly optimized by incorporating the label correlation exploitation and feature-induced noise model. Extensive experiments on synthetic as well as real-world data sets validate the effectiveness of the proposed approach.

Introduction

Multi-label learning (MLL) solves problems where each object is assigned with multiple class labels simultaneously (Zhang and Zhou 2014). For instance, an image may be annotated with labels *sea*, *sunset* and *beach*. A large number of recent works have witnessed the great successes that MLL has achieved in many research areas, *e.g.*, music emotion recognition (Trohidis et al. 2008), text categorization (Lin et al. 2018) and image annotation (Chen et al. 2019).

In traditional multi-label studies, a basic assumption is that each training instance has been precisely annotated with all of its relevant labels. However, in many real-world scenarios, it is difficult and costly to obtain precise annotations. Instead, it is more common that a set of candidate labels are roughly assigned by noisy annotators. In addition to the relevant labels, the candidate set may also contain some noisy labels, where the number of relevant or noisy labels is unknown. For example, in crowdsourcing image tagging (as shown in Figure 1), among the candidate labels annotated by annotators, only some of them are accurate ones owing to potential unreliable annotators. The scenario has been formalized as a learning framework called partial multi-label learning (PML) by (Xie and Huang 2018).

Figure 1: An example of partial multi-label learning. The image is partially labeled by noisy annotators in crowdsourcing. Among the candidate labels, house, tree, car, light and cloud are ground-truth labels while flower, cat and people are noisy labels.

To solve PML problems, one straightforward method is to simply treat all the candidate labels as relevant ones. Then the PML problem can be solved by standard multi-label learning algorithms, e.g., Binary Relevance (BR) (Boutell et al. 2004), ML-*k*NN (Zhang and Zhou 2007), CPLST(Chen and Lin 2012) and so on. However, such methods will be misled by the noisy labels in the candidate set, and fail to generalize well on future data.

In order to deal with the challenge, several PML techniques are proposed recently. Among them, the most commonly used strategy to learn from PML examples is *disambiguation*. It tires to recover ground-truth labeling information from candidate labels, by either introducing labeling confidences (Xie and Huang 2018; Fang and Zhang 2019) or employing low-rank and sparse decomposition scheme (Lijuan Sun and Jin 2019). Despite the advances these methods have achieved, a potential limitation is that they neglect the cause of noisy labels in the candidate set, which may be an important information for recovering the ground-truth labels. These methods typically assume that noisy labels are generated randomly, which may be not consistent with many

*This research was supported by NSFC(61876081, 61572252), the Aerospace Power Funds No. 6141B09050342 and the Fundamental Research Funds for the Central Universities, NO. NE2019104.

real-world scenarios. In practice, we observe that noisy labels are usually caused by some ambiguous contents of the example and there thus existing some relationships between the noisy labels and feature representations. For example, in crowdsourcing annotation scenario, annotators may be misled by some ambiguous contents associated with the example in specific tasks. Figure 1 illustrates an example in crowdsourcing image tagging, annotators provided the image with noisy labels *flower*, *cat* and *people* due to the misleading objects marked by the red, green and blue boxes. Similar cases also happen in other tasks, such as ambiguous words in the text categorization and ambiguous melody fragments in the music emotion recognition.

Based on the observations mentioned above, in this paper, we propose a new approach for Partial Multi-label Learning with Noisy label Identification (PML-NI), which recovers the ground-truth labeling information and identifies the noisy labels simultaneously. Specifically, the multi-label classifier and noisy label identifier are learned jointly under the supervision of the observed noise-corrupted label matrix. On one hand, the multi-label classifier is constrained to be low rank by trace norm regularization to capture the correlation among labels; on the other hand, the noisy label identifier with sparsity regularization is trained to model the feature-induced noise labels. Comprehensive experiments on synthetic as well as real-world data sets from diverse domains validate that the proposed approach consistently outperforms the compared methods.

The rest of this paper is organized as follows: Section 2 reviews some related works; Section 3 introduces our proposed PML-NI approach; experimental results are reported in Section 4, followed by the conclusion in Section 5.

Related Works

Partial multi-label learning is a powerful framework to deal with partially labeled data in multi-label setting. It is derived from two popular learning frameworks: multi-label learning and partial label learning.

There are plenty of literature on multi-label learning. Among them, Binary Relevance is the most simple approach which decomposes the task into a set of binary classification problems (Boutell et al. 2004). There are many studies trying to exploit the label correlations for enhancing the multi-label learning (Zhu, Kwok, and Zhou 2017; Huang, Yu, and Zhou 2012). Some of them focus on pairwise correlation (Li, Song, and Luo 2017), while some others consider high order correlation among all labels (Burkhardt and Kramer 2018; Read et al. 2011).

Partial label learning (PLL) is a framework for learning from partially labeled data for single label tasks (Grandvalet and Bengio 2004; Jin and Ghahramani 2002). In PLL problem, the partial label set consists of exactly one ground-truth label and some other noisy labels. The most common strategy applied in PLL methods is *disambiguation*, which tries to recover the ground-truth label from the candidate set (Feng Lei 2019; Zhang, Zhou, and Liu 2016). The disambiguation strategy are mostly implemented in two ways: one is to assume certain parametric model and the ground-truth label is regarded as the latent variable which can be iteratively refined by optimizing certain objectives, such as the maximum likelihood criterion (Grandvalet and Bengio 2004; Jin and Ghahramani 2002; Liu and Dietterich 2012) or the maximum margin criterion (Yu and Zhang 2017); the other one is to assume equal importance of each candidate label and then make prediction by averaging their modeling outputs. For parametric models, the averaged outputs for all candidate labels are distinguished from the outputs for candidate labels (Cour, Sapp, and Taskar 2011). For non-parametric models, the predicted label for unseen instance is determined by averaging the candidate labeling information from its neighboring examples in the PL training set (Hüllermeier and Beringer 2006; Zhang and Yu 2015). Compared to partial label learning, PML is much more challenging owing to the number of ground-truth labels in the candidate set is unknown.

To solve PML problems, the most intuitive method is to treat all candidate labels as relevant ones. In this case, PML problem can be solved by off-the-shelf multi-label learning algorithms. Nevertheless, such methods will be misled by the noisy labels in the candidate set, which may lead to degraded performances. In order to overcome this problem, some techniques are designed to solve PML problems recently. For example, (Xie and Huang 2018) propose two effective methods PML-*lc* and PML-*fp* by introducing a confidence value for each candidate label. The decomposition scheme is utilized to tackle PML data in (Lijuan Sun and Jin 2019). PARTICLE (Fang and Zhang 2019) identifies the credible labels with high labeling confidences by employing an iterative label propagation procedure. Despite the advances these methods have achieved, a potential limitation is that they do not consider the cause of noisy labels in the candidate set, which may be an essential information for solving PML problems.

The Proposed Approach

For each partially labeled training example, we denote by $\mathbf{x}_i \in \mathbb{R}^d$ a feature vector and its corresponding label vector $\mathbf{y} \in \{0,1\}^q$ with q class labels. Let $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n] \in \{0,1\}^{q\times n}$ denote the noise-corrupted label matrix. In this setting, $\mathbf{y}_{ji} = 1$ means the j-th label is a candidate label to the i-th instance. We further denote by $\widetilde{\mathbf{y}} \in \{0,1\}^q$ the unknown ground-truth label vector.

The PML-NI Framework

As mentioned in the above discussion, in many real-world scenarios, noisy labels are usually caused by some ambiguous contents of the example and there thus existing some relationships between noisy labels and feature contents. Here we model the noisy labels as the outputs of a linear mapping from the feature representations as follows:

$$\mathbf{y}_i - \widetilde{\mathbf{y}}_i = \widehat{\boldsymbol{V}}\mathbf{x}_i + \mathbf{s} = \boldsymbol{V}\boldsymbol{\phi}_i \tag{1}$$

where $\mathbf{y}_i - \widetilde{\mathbf{y}}_i$ represents the noisy label vector for instance $\mathbf{x}_i$. Here, $\widehat{\boldsymbol{V}} = [\widehat{\boldsymbol{v}}_1, \widehat{\boldsymbol{v}}_2, ..., \widehat{\boldsymbol{v}}_q]^\top$ is a weight matrix and $\mathbf{s} = [s_1, s_2, ..., s_q]^\top$ is a bias vector. For convenient describing, we set $\boldsymbol{V} = [\widehat{\boldsymbol{V}}, \mathbf{s}]$ and $\boldsymbol{\phi}_i = [\mathbf{x}_i; 1]$, in which we call $\boldsymbol{V}$

the noisy label identifier. Accordingly, the goal of our framework is to determine the optimal parameter $\boldsymbol{V}^*$ that can correctly identify the noisy labels given the feature vector $\boldsymbol{\phi}_i$. However, the ground-truth label $\widetilde{\mathbf{y}}_i$ here is unknown and the equation in Eq.(1) is thus intractable. To solve the problem, we propose a joint learning framework that can identify the noisy labels while training the multi-label classifier simultaneously:

$$\begin{aligned}\min_{\boldsymbol{W},\boldsymbol{U},\boldsymbol{V}} \mathcal{L}\left(\boldsymbol{W},\boldsymbol{\Phi},\mathbf{Y}\right)+\lambda R\left(\boldsymbol{W}\right)\\ s.t.\quad \boldsymbol{W}=\boldsymbol{U}+\boldsymbol{V}\end{aligned} \tag{2}$$

Here, $\boldsymbol{W}$ is the joint learning model that consists of the multi-label classifier $\boldsymbol{U}$ and noisy label identifier $\boldsymbol{V}$. The classifier $\boldsymbol{U}=[\widehat{\boldsymbol{U}},\boldsymbol{t}]$ tries to provide each training example $\boldsymbol{\phi}_i$ with its ground-truth label $\widetilde{\mathbf{y}}_i$, where $\widehat{\boldsymbol{U}}=[\widehat{\boldsymbol{u}}_1,\widehat{\boldsymbol{u}}_2,...,\widehat{\boldsymbol{u}}_q]^\top$ and $\boldsymbol{t}=[t_1,t_2,...,t_q]^\top$ are weight matrix and bias vector, respectively. $\mathcal{L}$ is the loss function to minimize empirical loss between modeling outputs $\boldsymbol{W}\boldsymbol{\Phi}$ and noise-corrupted label matrix $\mathbf{Y}$, where $\boldsymbol{\Phi}=[\boldsymbol{\phi}_1,\boldsymbol{\phi}_2,..,\boldsymbol{\phi}_n]$ is the feature matrix. R is a regularization term to control the model complexity, where λ is a balancing parameter. For simplicity, we choose the least square loss for model training and square Frobenius norm to control the model complexity, and then the optimization problem in eq.(2) can be re-written by:

$$\begin{aligned}\min_{\boldsymbol{W},\boldsymbol{U},\boldsymbol{V}} \tfrac{1}{2}\left\|\mathbf{Y}-\boldsymbol{W}\boldsymbol{\Phi}\right\|_{\mathrm{F}}^2+\tfrac{\lambda}{2}\left\|\boldsymbol{W}\right\|_{\mathrm{F}}^2\\ s.t.\quad \boldsymbol{W}=\boldsymbol{U}+\boldsymbol{V}\end{aligned} \tag{3}$$

However, the classifier and identifier here are unconstrained and their individual abilities, i.e., ground-truth label prediction and noisy label identification are hardly considered. To deal with the problem, in the following content, we will show how to capture their intrinsic property and potential structure information by employing different regularizers for each of $\boldsymbol{U}$ and $\boldsymbol{V}$. Therefore, the optimization problem in Eq.(3) can be firstly reformulated as follows:

$$\begin{aligned}\min_{\boldsymbol{W},\boldsymbol{U},\boldsymbol{V}} \tfrac{1}{2}\left\|\mathbf{Y}-\boldsymbol{W}\boldsymbol{\Phi}\right\|_{\mathrm{F}}^2+\tfrac{\lambda}{2}\left\|\boldsymbol{W}\right\|_{\mathrm{F}}^2+\beta\Omega\left(\boldsymbol{U}\right)\\ +\gamma\Psi\left(\boldsymbol{V}\right)\\ s.t.\quad \boldsymbol{W}=\boldsymbol{U}+\boldsymbol{V}\end{aligned} \tag{4}$$

Here, Ω and Ψ are regularizers to encourage the classifier and identifier to perform their individual abilities, where β and γ are balancing parameters.

In multi-label learning, a common assumption is that there existing the label correlations among different labels (Zhu, Kwok, and Zhou 2017; Huang and Zhou 2012; Lijuan Sun and Jin 2019) and the feature mapping matrix $\boldsymbol{U}$ is thus linearly dependent. The low-rank assumption is thus naturally used to capture this intrinsic property of the classifier. Therefore, the optimization problem can be defined by incorporating the label correlation exploitation:

$$\begin{aligned}\min_{\boldsymbol{W},\boldsymbol{U},\boldsymbol{V}} \tfrac{1}{2}\left\|\mathbf{Y}-\boldsymbol{W}\boldsymbol{\Phi}\right\|_{\mathrm{F}}^2+\tfrac{\lambda}{2}\left\|\boldsymbol{W}\right\|_{\mathrm{F}}^2+\beta\mathbf{rank}(\boldsymbol{U})\\ s.t.\quad \boldsymbol{W}=\boldsymbol{U}+\boldsymbol{V}\end{aligned}$$

Note that the goal of identifier $\boldsymbol{V}$ is to correctly identify noisy labels mixed up in the candidate set. On one hand, as aforementioned, noisy labels are usually caused by some specific contents, i.e., only a few of ambiguous feature, which are sparse among the observed feature matrix; on the other hand, noisy labels occur occasionally and tend to be sparse among observed candidate labels. In order to make full use of these two kinds of sparsity, we assume that the feature mapping matrix $\boldsymbol{V}$ also contains some sparsity and employ ℓ_0 norm regularizer as a feature-induced noise model to capture such structure information. Therefore, the optimization problem can be defined by incorporating the feature-induced noise model:

$$\begin{aligned}\min_{\boldsymbol{W},\boldsymbol{U},\boldsymbol{V}} \tfrac{1}{2}\left\|\mathbf{Y}-\boldsymbol{W}\boldsymbol{\Phi}\right\|_{\mathrm{F}}^2+\tfrac{\lambda}{2}\left\|\boldsymbol{W}\right\|_{\mathrm{F}}^2+\beta\mathbf{rank}(\boldsymbol{U})\\ +\gamma\left\|\boldsymbol{V}\right\|_0\\ s.t.\quad \boldsymbol{W}=\boldsymbol{U}+\boldsymbol{V}\end{aligned} \tag{5}$$

However, it is difficult to solve the optimization problem in Eq.(5) due to the rank and cardinality operators are highly nonconvex and computationally NP-hard (Fazel, Hindi, and Boyd 2004; Wright et al. 2008). Therefore, these operators are relaxed by their convex surrogates, i.e., the trace norm (Candès and Recht 2009) and ℓ_1-norm(Candes and Tao 2005). The final optimization problem can be re-written as follows:

$$\begin{aligned}\min_{\boldsymbol{W},\boldsymbol{U},\boldsymbol{V}} \tfrac{1}{2}\left\|\mathbf{Y}-\boldsymbol{W}\boldsymbol{\Phi}\right\|_{\mathrm{F}}^2+\tfrac{\lambda}{2}\left\|\boldsymbol{W}\right\|_{\mathrm{F}}^2+\beta\left\|\boldsymbol{U}\right\|_{\mathrm{tr}}\\ +\gamma\left\|\boldsymbol{V}\right\|_1\\ s.t.\quad \boldsymbol{W}=\boldsymbol{U}+\boldsymbol{V}\end{aligned} \tag{6}$$

The optimization problem in Eq.(6) can be solved effectively by alternating optimization.

Alternating Optimization

After substituting the constraint into Eq.(6), the optimization problem can be re-arranged as follows:

$$\begin{aligned}\min_{\boldsymbol{U},\boldsymbol{V}} \tfrac{1}{2}\left\|\mathbf{Y}-\left(\boldsymbol{U}+\boldsymbol{V}\right)\boldsymbol{\Phi}\right\|_{\mathrm{F}}^2+\tfrac{\lambda}{2}\left\|\boldsymbol{U}+\boldsymbol{V}\right\|_{\mathrm{F}}^2\\ +\beta\left\|\boldsymbol{U}\right\|_{\mathrm{tr}}+\gamma\left\|\boldsymbol{V}\right\|_1\end{aligned} \tag{7}$$

which can be effectively solved by alternatively optimizing $\boldsymbol{U}$ and $\boldsymbol{V}$.

When $\boldsymbol{V}$ is fixed, the optimization problem in Eq.(7) with respective to $\boldsymbol{U}$ can be reformulated as follows:

$$\min_{\boldsymbol{U}} \left\|\boldsymbol{U}\boldsymbol{\Phi}-\boldsymbol{E}\right\|_{\mathrm{F}}^2+\lambda\left\|\boldsymbol{U}+\boldsymbol{V}\right\|_{\mathrm{F}}^2+\beta\left\|\boldsymbol{U}\right\|_{\mathrm{tr}} \tag{8}$$

where $\boldsymbol{E}=\mathbf{Y}-\boldsymbol{V}\boldsymbol{\Phi}$. The accelerated proximal gradient descend has been proved to be an effective optimization technique for trace norm minimization to solve this problem (Huang et al. 2018). Let

$$g\left(\boldsymbol{U}\right)=\frac{1}{2}\left\|\boldsymbol{U}\boldsymbol{\Phi}-\boldsymbol{E}\right\|_{\mathrm{F}}^2+\frac{\lambda}{2}\left\|\boldsymbol{U}+\boldsymbol{V}\right\|_{\mathrm{F}}^2$$

and

$$h\left(\boldsymbol{U},\boldsymbol{Z}\right)=g\left(\boldsymbol{Z}\right)+\left\langle\nabla g\left(\boldsymbol{Z}\right),\boldsymbol{U}-\boldsymbol{Z}\right\rangle+\beta\left\|\boldsymbol{U}\right\|_{\mathrm{tr}}$$

where

$$\nabla g\left(\boldsymbol{Z}\right)=\left(\boldsymbol{Z}\boldsymbol{\Phi}-\boldsymbol{E}\right)\boldsymbol{\Phi}^\top+\lambda\left(\boldsymbol{U}+\boldsymbol{V}\right)$$

The main steps are summarized as follows:

- Choose $\theta_0 = \theta_{-1} \in (0, 1]$, $L > 1$, $\boldsymbol{U}_0 = \boldsymbol{U}_{-1}$, $\eta > 1$. Set $k = 0$.
- In the k-th iteration,
  - Set $\boldsymbol{Z}_k = \boldsymbol{U}_k + \theta_k(\theta_{k-1}^{-1} - 1)(\boldsymbol{U}_k - \boldsymbol{U}_{k-1})$
  - Set $\boldsymbol{U}_{k+1} = \operatorname{argmin}_{\boldsymbol{U}}\{h(\boldsymbol{U}, \boldsymbol{Z}_k) + \frac{L}{2}\|\boldsymbol{U} - \boldsymbol{Z}_k\|_{\mathrm{F}}^2\}$
  - while $g(\boldsymbol{U}_{k+1}) + \beta\|\boldsymbol{U}_{k+1}\|_{\mathrm{tr}} > h(\boldsymbol{U}_{k+1}, \boldsymbol{Z}_k) + \frac{L}{2}\|\boldsymbol{U}_{k+1} - \boldsymbol{Z}_k\|_{\mathrm{F}}^2$:
    - Increase $L = \eta L$
    - $\boldsymbol{U}_{k+1} = \operatorname{argmin}_{\boldsymbol{U}}\{h(\boldsymbol{U}, \boldsymbol{Z}_k) + \frac{L}{2}\|\boldsymbol{U} - \boldsymbol{Z}_k\|_{\mathrm{F}}^2\}$
  - Set $\theta_{k+1} = \sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2/2$
  - Update $k = k + 1$

The iteration continues until convergence. In the above steps, we have omitted the procedures for obtaining $\boldsymbol{U}_{k+1}$ and next we will show this. The problem can be rewritten as

$$\min_{\boldsymbol{U}} \langle \nabla g(\boldsymbol{Z}_k), \boldsymbol{U} - \boldsymbol{Z}_k \rangle + \frac{L}{2}\|\boldsymbol{U} - \boldsymbol{Z}_k\|_{\mathrm{F}}^2 + \beta\|\boldsymbol{U}\|_{\mathrm{tr}}$$

which is equivalent to

$$\min_{\boldsymbol{U}} \frac{L}{2}\left\|\boldsymbol{U} - \left(\boldsymbol{Z}_k - \frac{1}{L}\nabla g(\boldsymbol{Z}_k)\right)\right\|_{\mathrm{F}}^2 + \beta\|\boldsymbol{U}\|_{\mathrm{tr}}$$

This can be solved by Singular Value Thresholding (SVT) (Cai, Candès, and Shen 2010), which performs singular value decomposition on $\boldsymbol{Z}_k - \frac{1}{L}\nabla g(\boldsymbol{Z}_k) = \widetilde{\mathbf{U}}\boldsymbol{\Sigma}\widetilde{\mathbf{V}}$. Let $\widetilde{\Sigma}_{ii} = \max(0, \Sigma_{ii} - \frac{\beta}{L})$ and then the solution is given by $\widetilde{\mathbf{U}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\mathbf{V}}$.

When $\boldsymbol{U}$ is fixed, the optimization problem in Eq.(7) with respective to $\boldsymbol{V}$ can be reformulated as follows:

$$\min_{\boldsymbol{V}} \|\boldsymbol{V}\boldsymbol{\Phi} - \boldsymbol{\Lambda}\|_{\mathrm{F}}^2 + \lambda\|\boldsymbol{V} + \boldsymbol{U}\|_{\mathrm{F}}^2 + \gamma\|\boldsymbol{V}\|_1$$

where $\boldsymbol{\Lambda} = \mathbf{Y} - \boldsymbol{U}\boldsymbol{\Phi}$. The problem can be solved effectively by employing the shrinkage operator (Lin, Chen, and Ma 2010).

Let

$$f(\boldsymbol{V}) = \frac{1}{2}\|\boldsymbol{V}\boldsymbol{\Phi} - \Lambda\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\|\boldsymbol{U} + \boldsymbol{V}\|_{\mathrm{F}}^2$$

and we have

$$\mathbf{prox}_{\gamma} = \operatorname{argmin}_{\boldsymbol{V}} \gamma\|\boldsymbol{V}\|_1 + f(\boldsymbol{V})$$

which is called shrinkage operator and the optimal parameter $\boldsymbol{V}^* = \mathbf{prox}_{\gamma}$. Let $\boldsymbol{H} = \boldsymbol{V} - \frac{1}{L}\nabla f(\boldsymbol{V})$, where L is the Lipschitz constant. And then the closed solution can be obtained (Combettes and Wajs 2005):

$$\forall k \in [q], i \in [d] \quad V_{ki}^* = \begin{cases} H_{ki} - \gamma/L, & H_{ki} > \gamma/L \\ 0, & |H_{ki}| \le \gamma/L \\ H_{ki} + \gamma/L, & H_{ki} < \gamma/L \end{cases}$$

where V_{ki} and H_{ki} are i-th dimensional values of k-th label for $\boldsymbol{V}$ and $\boldsymbol{H}$.

The entire optimization procedure will be terminated when the overall loss converges.

Experiment

Experimental Setting

We perform experiments on totally ten data sets including synthetic as well as real-world PML data sets [1]. These data sets spanned a broad range of applications: *image* and *corel16k* for image annotation, *music_emotion*, *music_style* and *birds* for music recognition, *genbase* for protein classification as well as *medical*, *enron*, *bibtex* and *tmc2007* for text categorization. Table 1 illustrates the *number of instances*, *number of class labels*, *cardinality* and *domain* for each data set. We also did some pre-processing to facilitate the partially labeling as in (Xie and Huang 2018; Fang and Zhang 2019). Specifically, for data sets with too many class labels (more than 100 in our experiments), their rare labels are filtered out to keep under 15 labels, and instances without any relevant labels are filtered out.

There are different criteria for evaluating the performances of multi-label learning. In our experiments, we employ five commonly used criteria including *ranking loss*, *hamming loss*, *one error*, *coverage* and *average precision*. More detail about these evaluation metrics can be found in (Zhang and Zhou 2014). For the *ranking loss*, *hamming loss*, *one error* and *coverage* metrics, the smaller value, the better the performance. For the *average precision* metric, the larger the value, the better the performance.

To validate the effectiveness of the proposed PML-NI method, we compare with three state-of-the-art PML algorithms and two well-established MLL approaches as follows:

- PARTICLE (Fang and Zhang 2019). It transforms the PML task into a multi-label problem through a label propagation procedure. Then a calibrated label ranking model is induced to instantiate two PML methods PAR-VLS and PAR-MAP.
- PML-LRS (PML-LRS) (Lijuan Sun and Jin 2019). It utilizes low-rank and sparse decomposition scheme to capture the ground-truth label matrix and irrelevant label matrix from the observed candidate label matrix.
- ML-kNN (Zhang and Zhou 2007). It is a nearest neighbor based multi-label classification method. ML-kNN is a very popular baseline method in multi-label learning literature owing to its simplicity.
- CPLST (Chen and Lin 2012). It is a typical label embedding approach in MLL, which integrates the concepts of principal component analysis and canonical correlation analysis.

For the comparing methods, parameters are set as suggested in the original paper, i.e., PAR-VAL and PAR-MAP: balancing parameter $\alpha = 0.95$ and credible label elicitation threshold $thr = 0.9$; PML-LRS: balancing parameters are set as $\gamma = 0.01$, $\beta = 0.1$ and $\eta = 1$. For CPLST, we take the first 5 principle components following the experimental setting in (Wang et al. 2019). k is set as 10 for all the nearest neighbor based algorithms. Libsvm (Chang and Lin 2011) is used as

[1] Publicly available at http://mulan.sourceforge.net/datasets.html and http://meka.sourceforge.net/#datasets

Table 1: Characteristics of the experimental data sets.

Data set	# Instances	# Features	# Class Labels	Cardinality	Domain
music_emotion	6833	98	11	2.42	music
music_style	6839	98	10	1.44	music
birds	654	260	19	2.402	music
genbase	662	1186	27	1.252	biology
medical	978	1449	45	1.245	text
enron	1702	1001	53	3.378	text
image	2000	294	5	1.23	image
bibtex	7395	1836	159	2.402	text
corel16k	13811	500	161	2.867	image
tmc2007	21519	500	22	2.158	text

Table 2: Experimental results of each comparing approach in terms of *ranking loss*, where •/○ indicates whether PML-NI is superior/inferior to the other method.

Data	$\alpha\%$	PML-NI	PAR-VAL	PAR-MAP	PML-LRS	ML-kNN	CPLST
music_emotion		$.251 \pm .009$	$.265 \pm .008$•	$.253 \pm .008$•	$.256 \pm .002$•	$.257 \pm .006$•	$.364 \pm .009$•
music_style		$.141 \pm .003$	$.157 \pm .002$•	$.164 \pm .004$•	$.148 \pm .006$•	$.157 \pm .005$•	$.232 \pm .006$•
birds	50%	$.190 \pm .014$	$.438 \pm .058$•	$.285 \pm .021$•	$.302 \pm .018$•	$.324 \pm .040$•	$.252 \pm .012$•
	100%	$.207 \pm .019$	$.400 \pm .046$•	$.298 \pm .017$•	$.323 \pm .028$•	$.322 \pm .019$•	$.283 \pm .031$•
	150%	$.236 \pm .028$	$.466 \pm .066$•	$.307 \pm .026$•	$.330 \pm .014$•	$.331 \pm .030$•	$.293 \pm .013$•
genbase	50%	$.003 \pm .001$	$.025 \pm .013$•	$.012 \pm .006$•	$.017 \pm .004$•	$.008 \pm .004$•	$.050 \pm .010$•
	100%	$.004 \pm .002$	$.059 \pm .030$•	$.010 \pm .004$•	$.017 \pm .003$•	$.011 \pm .004$•	$.063 \pm .018$•
	150%	$.010 \pm .003$	$.017 \pm .008$•	$.011 \pm .004$•	$.031 \pm .008$•	$.027 \pm .007$•	$.075 \pm .016$•
medical	50%	$.023 \pm .005$	$.157 \pm .034$•	$.071 \pm .015$•	$.048 \pm .013$•	$.047 \pm .008$•	$.089 \pm .008$•
	100%	$.023 \pm .007$	$.155 \pm .035$•	$.074 \pm .017$•	$.049 \pm .008$•	$.047 \pm .008$•	$.097 \pm .010$•
	150%	$.025 \pm .005$	$.147 \pm .029$•	$.073 \pm .013$•	$.053 \pm .005$•	$.049 \pm .005$•	$.102 \pm .015$•
enron	50%	$.175 \pm .013$	$.318 \pm .070$•	$.188 \pm .047$•	$.163 \pm .021$○	$.180 \pm .007$•	$.301 \pm .019$•
	100%	$.176 \pm .012$	$.376 \pm .088$•	$.216 \pm .048$•	$.168 \pm .012$○	$.190 \pm .011$•	$.294 \pm .011$•
	150%	$.178 \pm .013$	$.366 \pm .077$•	$.209 \pm .047$•	$.171 \pm .021$○	$.196 \pm .011$•	$.297 \pm .017$•
image	50%	$.175 \pm .005$	$.195 \pm .045$•	$.267 \pm .102$•	$.187 \pm .010$•	$.186 \pm .016$•	$.189 \pm .019$•
	100%	$.178 \pm .009$	$.198 \pm .042$•	$.267 \pm .099$•	$.182 \pm .014$•	$.190 \pm .012$•	$.189 \pm .010$•
	150%	$.183 \pm .006$	$.205 \pm .059$•	$.265 \pm .139$•	$.185 \pm .015$•	$.212 \pm .013$•	$.196 \pm .013$•
bibtex	50%	$.038 \pm .003$	$.080 \pm .002$•	$.057 \pm .001$•	$.042 \pm .002$•	$.115 \pm .008$•	$.115 \pm .010$•
	100%	$.032 \pm .002$	$.095 \pm .006$•	$.062 \pm .004$•	$.035 \pm .004$•	$.136 \pm .019$•	$.138 \pm .002$•
	150%	$.033 \pm .003$	$.098 \pm .007$•	$.064 \pm .004$•	$.035 \pm .003$•	$.143 \pm .011$•	$.151 \pm .006$•
corel16k	50%	$.211 \pm .002$	$.288 \pm .002$•	$.236 \pm .003$•	$.214 \pm .003$•	$.264 \pm .007$•	$.229 \pm .004$•
	100%	$.224 \pm .004$	$.334 \pm .008$•	$.262 \pm .005$•	$.226 \pm .004$•	$.273 \pm .002$•	$.239 \pm .005$•
	150%	$.224 \pm .006$	$.326 \pm .007$•	$.258 \pm .003$•	$.228 \pm .001$•	$.275 \pm .007$•	$.237 \pm .005$•
tmc2007	50%	$.046 \pm .001$	$.087 \pm .014$•	$.057 \pm .008$•	$.046 \pm .001$•	$.075 \pm .004$•	$.080 \pm .002$•
	100%	$.047 \pm .001$	$.082 \pm .014$•	$.057 \pm .009$•	$.047 \pm .002$•	$.079 \pm .002$•	$.081 \pm .001$•
	150%	$.050 \pm .002$	$.107 \pm .023$•	$.060 \pm .010$•	$.050 \pm .002$•	$.082 \pm .001$•	$.086 \pm .001$•

the base learner to instantiate PAR-VLS and PAR-MAP. For PML-NI, balancing parameters are set as $\lambda = 1$, $\beta = 1$ and $\gamma = 0.5$.

For the last 8 data sets, to construct partial multi-label assignments for the training data, we simulate the annotation process by using a svm classifier trained on original supervised multi-label data sets as the human annotator. Specifically, a svm classifier is firstly trained on the multi-label data set. Then, for each instance $\mathbf{x}_i$ of the data set, we add the irrelevant noisy labels of $\mathbf{x}_i$ with $\alpha\%$ number of ground-truth labels according to their probabilities to be relevant labels predicted by the svm classifier and the $\alpha\%$ is varied in the range $\{50\%, 100\%, 150\%\}$. To examine the performance of the proposed approaches, we performed experiments with all possible percentages of the noisy labels. In the following content, we will show details of three groups of experiments on these totally 26 data sets including 24 synthetic and 2 real-world data sets.

Comparison Results

Due to the page limit, we follow the setting in (Fang and Zhang 2019) to only report detailed results of each com-

Table 3: Experimental results of each comparing approach in terms of *average precision*, where •/○ indicates whether PML-NI is superior/inferior to the other method.

Data	$\alpha\%$	PML-NI	PAR-VAL	PAR-MAP	PML-LRS	ML-kNN	CPLST
music_emotion		.598 ± .010	.607 ± .010○	.611 ± .011○	.589 ± .006•	.595 ± .007•	.506 ± .009•
music_style		.731 ± .003	.713 ± .004•	.710 ± .007•	.714 ± .008•	.717 ± .011•	.658 ± .009•
birds	50%	.507 ± .019	.413 ± .034•	.395 ± .024•	.371 ± .030•	.370 ± .037•	.451 ± .015•
	100%	.466 ± .013	.416 ± .042•	.386 ± .024•	.352 ± .033•	.366 ± .037•	.410 ± .033•
	150%	.419 ± .026	.392 ± .033•	.369 ± .023•	.344 ± .031•	.352 ± .017•	.387 ± .040•
genbase	50%	.980 ± .005	.895 ± .022•	.968 ± .020•	.860 ± .022•	.948 ± .011•	.738 ± .028•
	100%	.971 ± .010	.819 ± .039•	.965 ± .019•	.851 ± .025•	.920 ± .055•	.723 ± .030•
	150%	.922 ± .022	.897 ± .042•	.960 ± .010○	.785 ± .049•	.773 ± .069•	.612 ± .020•
medical	50%	.819 ± .010	.703 ± .021•	.737 ± .029•	.738 ± .034•	.737 ± .014•	.592 ± .027•
	100%	.809 ± .017	.680 ± .020•	.714 ± .031•	.724 ± .020•	.734 ± .014•	.568 ± .027•
	150%	.758 ± .016	.673 ± .013•	.675 ± .018•	.665 ± .014•	.664 ± .032•	.498 ± .031•
enron	50%	.563 ± .013	.297 ± .132•	.432 ± .068•	.528 ± .022•	.450 ± .017•	.350 ± .004•
	100%	.494 ± .017	.271 ± .129•	.398 ± .081•	.474 ± .019•	.412 ± .016•	.346 ± .013•
	150%	.474 ± .014	.264 ± .120•	.397 ± .058•	.453 ± .021•	.395 ± .017•	.326 ± .022•
image	50%	.780 ± .007	.770 ± .055•	.734 ± .076•	.765 ± .013•	.767 ± .015•	.766 ± .019•
	100%	.782 ± .006	.767 ± .051•	.735 ± .077•	.772 ± .016•	.763 ± .016•	.769 ± .007•
	150%	.772 ± .007	.760 ± .068•	.709 ± .150•	.770 ± .016•	.732 ± .009•	.757 ± .015•
bibtex	50%	.890 ± .008	.810 ± .009•	.831 ± .006•	.888 ± .007•	.748 ± .009•	.733 ± .017•
	100%	.889 ± .007	.763 ± .010•	.816 ± .011•	.874 ± .013•	.708 ± .028•	.621 ± .008•
	150%	.888 ± .006	.761 ± .010•	.816 ± .009•	.873 ± .006•	.697 ± .019•	.598 ± .015•
corel16k	50%	.511 ± .006	.473 ± .003•	.484 ± .003•	.511 ± .004○	.456 ± .010•	.500 ± .003•
	100%	.483 ± .007	.453 ± .006•	.454 ± .007•	.481 ± .007•	.436 ± .004•	.476 ± .005•
	150%	.487 ± .006	.458 ± .004•	.455 ± .009•	.479 ± .005•	.433 ± .009•	.475 ± .007•
tmc2007	50%	.804 ± .002	.731 ± .033•	.783 ± .022•	.803 ± .006•	.746 ± .008•	.747 ± .002•
	100%	.803 ± .003	.737 ± .035•	.785 ± .021•	.802 ± .005•	.729 ± .004•	.738 ± .005•
	150%	.793 ± .003	.676 ± .033•	.760 ± .036•	.792 ± .005•	.710 ± .005•	.721 ± .002•

Table 4: Friedman statistics F_F in terms of each evaluation metric and the critical value at 0.05 significance level (# comparing algorithms $k = 6$, # data sets $N = 24$).

Evaluation metric	F_F	critical value
Hamming Loss	30.1256	
Ranking loss	37.9784	
One Error	14.8082	2.2932
Coverage	38.7910	
Average Precision	23.5169	

paring methods in terms of *ranking loss* and *average precision* in Table 2 and 3, while similar results can be observed in terms of other evaluation metrics (the detailed results in terms of *hamming loss*, *one error* and *coverage* are reported on supplementary materials). When compare PML-NI approach with other methods, our algorithm shows significant superiority. It achieves the best performance in most cases. Among the five comparing approaches, PML-LRS shows some superiority, and is better than PML-NI in three cases on *enron* in terms of *ranking loss* and one case on *corel16k* in terms of *average precision*, while losses for other cases. PAR-MAP outperforms PML-NI in one case on *genbase* in terms of *average precision*, while losses for other cases.

To validate the effectiveness of PML-NI for real applications, we also perform experiments on real-world PML data sets *music_emotion* and *music_style*. The results show that PML-NI achieves the best results in almost all cases except for the data set *music_emotion* where PAR-VAL and PAR-MAP outperform PML-NI in terms of *average precision*.

Furthermore, we also use *Friedman test* (Demsar 2006; Zhang, Zhong, and Zhang 2018; Lijuan Sun and Jin 2019) as the statistical test to analyze the relative performance among the comparing approaches. Table 4 reports the Friedman statistics F_F and the corresponding critical value with respective to each evaluation metric. For each evaluation metric, the null hypothesis of indistinguishable performance among the comparing algorithm is rejected at 0.05 significance level.

Then, the post-hoc *Bonferroni-Dunn test* (Demsar 2006; Zhang, Zhong, and Zhang 2018; Lijuan Sun and Jin 2019) is utilized to illustrate the relative performance among comparing approaches. Here, PML-NI is regarded as the control method whose average rank difference against the comparing algorithm is calibrated with the *critical difference* (CD). Accordingly, PML-NI is deemed to have significantly different performance to one comparing algorithm if their average ranks differ by at least one CD (CD = 1.3912 in our experiment: # comparing algorithms $k = 6$, # data sets $N = 8 \times 3 = 24$). Figure 2 shows the CD diagrams ((Demsar 2006)) on each evaluation metric, where the average rank

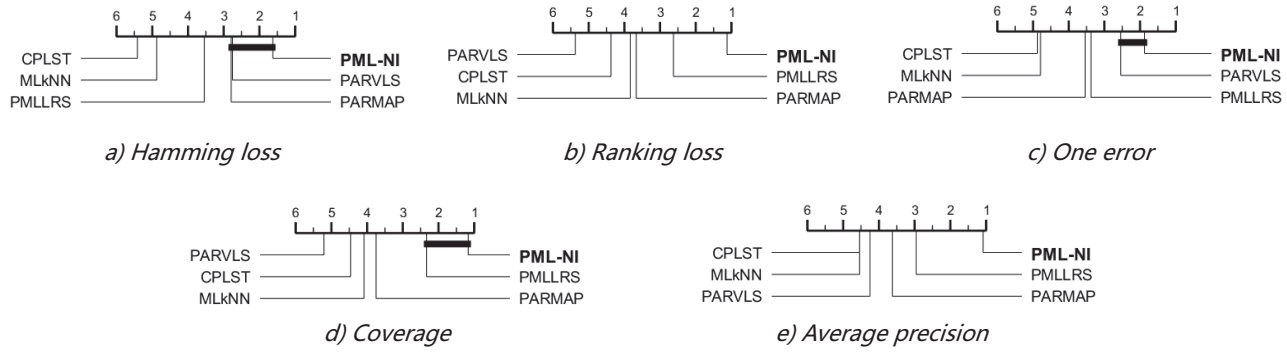
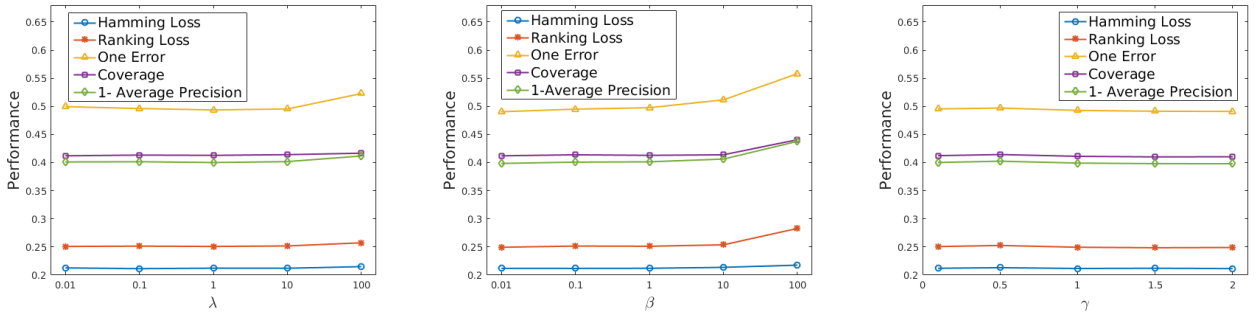


Figure 2: Comparison of PML-NI (control algorithm) against five comparing algorithms with the Bonferroni-Dunn test. Algorithms not connected with PML-NI in the CD diagram are considered to have a significantly different performance from the control algorithm (CD = 1.5510 at 0.05 significance level).



(a) performance curve with λ changes (b) performance curve with β changes (c) performance curve with γ changes

Figure 3: Results of PML-NI with varying value of trade-off parameters on *music_emotion*.

of each comparing algorithm is marked along the axis (lower ranks to the right). In each subfigure, any comparing algorithms whose average rank is within one CD to that of PML-NI is interconnected to each other with a thick line. It can be observed that PML-NI achieves the best (lowest) average rank in terms of all evaluation metrics. These experimental results convincingly validate the significance of the superiority for our PML-NI approach.

Sensitive Analysis

In this section, we study the influences of three balancing parameters, λ, β and γ for the proposed approach on the real-world data sets. We conducted experiments by varying one parameter while keeping the other two parameters fixed. Due to the page limit, we only show the experimental results which are measured by the five evaluation metrics on real-world data set *music_emotion* in Figure 3, while the results on real-world data set *music_style* are reported on supplementary materials. As we can see, in general, performance is not sensitive to the parameters except for the parameter β, whose performance will be significantly degraded when the value of β is too large (approximates to 100 in the experiment). Therefore we can safely set them in a wide range in practice.

Conclusion

In this paper, we disclose the phenomenon that noise labels are usually caused by some ambiguous contents of the example. Based on this observation, we propose to learn partial multi-label problems in a novel strategy by exploiting the potential connections between noisy labels and feature contents. Under the supervision of the observed label matrix, the proposed PML-NI approach jointly learn the multi-label classifier and noisy label identifier by incorporating the label correlation exploitation and feature-induced noise model. Experiments results validate that the proposed approaches are superior to state-of-the-art approaches. In the future, we plan to improve the PML-NI method by considering various forms of noisy labels and utilizing more powerful learning models.

References

Boutell, M. R.; Luo, J.; Shen, X.; and Brown, C. M. 2004. Learning multi-label scene classification. *Pattern recognition* 37(9):1757–1771.

Burkhardt, S., and Kramer, S. 2018. Online multi-label dependency topic models for text classification. *Machine Learning* 107(5):859–886.

Cai, J.-F.; Candès, E. J.; and Shen, Z. 2010. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization* 20(4):1956–1982.

Candès, E. J., and Recht, B. 2009. Exact matrix completion via convex optimization. *Foundations of Computational mathematics* 9(6):717.

Candes, E., and Tao, T. 2005. Decoding by linear programming. *arXiv preprint math/0502327*.

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.

Chen, Y.-N., and Lin, H.-T. 2012. Feature-aware label space dimension reduction for multi-label classification. In *Advances in Neural Information Processing Systems*, 1529–1537.

Chen, Z.-M.; Wei, X.-S.; Wang, P.; and Guo, Y. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5177–5186.

Combettes, P. L., and Wajs, V. R. 2005. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation* 4(4):1168–1200.

Cour, T.; Sapp, B.; and Taskar, B. 2011. Learning from partial labels. *Journal of Machine Learning Research* 12:1501–1536.

Demsar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* 7:1–30.

Fang, J., and Zhang, M. 2019. Partial multi-label learning via credible label elicitation. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, (AAAI-19)*.

Fazel, M.; Hindi, H.; and Boyd, S. 2004. Rank minimization and applications in system theory. In *Proceedings of the 2004 American control conference*, volume 4, 3273–3278. IEEE.

Feng Lei, B. A. 2019. Partial label learning with self-guided retraining. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, (AAAI-19)*.

Grandvalet, Y., and Bengio, Y. 2004. Learning from partial labels with minimum entropy. *Cirano Working Papers*.

Huang, S.-J., and Zhou, Z.-H. 2012. Multi-label learning by exploiting label correlations locally. In *Twenty-sixth AAAI conference on artificial intelligence*.

Huang, S.-J.; Xu, M.; Xie, M.-K.; Sugiyama, M.; Niu, G.; and Chen, S. 2018. Active feature acquisition with supervised matrix completion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1571–1579. ACM.

Huang, S.; Yu, Y.; and Zhou, Z. 2012. Multi-label hypothesis reuse. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, 525–533.

Hüllermeier, E., and Beringer, J. 2006. Learning from ambiguously labeled examples. *Lecture Notes in Computer Science* 10(5):419–439.

Jin, R., and Ghahramani, Z. 2002. Learning with multiple labels. In *Advances in Neural Information Processing Systems*, 897–904.

Li, Y.; Song, Y.; and Luo, J. 2017. Improving pairwise ranking for multi-label image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 1837–1845.

Lijuan Sun, Songhe Feng, T. W. C. L., and Jin, Y. 2019. Partial multi-label learning by low-rank and sparse decomposition. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, (AAAI-19)*.

Lin, J.; Su, Q.; Yang, P.; Ma, S.; and Sun, X. 2018. Semantic-unit-based dilated convolution for multi-label text classification. In *Proceedings of the 2018 Conference on Empirical Methods in NaturalLanguage Processing*, 4554–4564.

Lin, Z.; Chen, M.; and Ma, Y. 2010. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*.

Liu, L., and Dietterich, T. G. 2012. A conditional multinomial mixture model for superset label learning. In *Proceedings of 26th Annual Conference on Neural Information Processing Systems*, 557–565.

Read, J.; Pfahringer, B.; Holmes, G.; and Frank, E. 2011. Classifier chains for multi-label classification. *Machine Learning* 85(3):333–359.

Trohidis, K.; Tsoumakas, G.; Kalliris, G.; and Vlahavas, I. P. 2008. Multi-label classification of music into emotions. In *Proceedings of 9th International Conference on Music Information Retrieval*, 325–330.

Wang, H.; Liu, W.; Zhao, Y.; Zhang, C.; Hu, T.; and Chen, G. 2019. Discriminative and correlative partial multi-label learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, 3691–3697.

Wright, J.; Yang, A. Y.; Ganesh, A.; Sastry, S. S.; and Ma, Y. 2008. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence* 31(2):210–227.

Xie, M., and Huang, S. 2018. Partial multi-label learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 4302–4309.

Yu, F., and Zhang, M. 2017. Maximum margin partial label learning. *Machine Learning* 106(4):573–593.

Zhang, M., and Yu, F. 2015. Solving the partial label learning problem: An instance-based approach. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 4048–4054.

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.

Zhang, Q.; Zhong, Y.; and Zhang, M. 2018. Feature-induced labeling information enrichment for multi-label learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, 4446–4453.

Zhang, M.; Zhou, B.; and Liu, X. 2016. Partial label learning via feature-aware disambiguation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1335–1344.

Zhu, Y.; Kwok, J. T.; and Zhou, Z.-H. 2017. Multi-label learning with global and local label correlation. *IEEE Transactions on Knowledge and Data Engineering* 30(6):1081–1094.