Feature-induced Partial Multi-label Learning

Guoxian Yu[1], Xia Chen[1,3], Carlotta Domeniconi[2], Jun Wang[1], Zhao Li[3], Zili Zhang[1,4], Xindong Wu[5]
[1]College of Computer and Information Sciences, Southwest University, China
[2]Department of Computer Science, George Mason University, USA
[3]Alibaba Group, Hangzhou, China
[4]School of Information Technology, Deakin University, Australia
[5]School of Computing and Informatics, University of Louisiana at Lafayette, USA
Email: [1]{gxyu, xchen, kingjun, zhangzl}@swu.edu.cn;
[2]carlotta@cs.gmu.edu; [3]lizhao.lz@alibaba-inc.com; [5]xwu@louisiana.edu

***Abstract*—Current efforts on multi-label learning generally assume that the given labels of training instances are noise-free. However, obtaining noise-free labels is quite difficult and often impractical, and the presence of noisy labels may compromise the performance of multi-label learning. Partial multi-label learning (PML) addresses the scenario in which each instance is annotated with a set of candidate labels, of which only a subset corresponds to the ground-truth. The PML problem is more challenging than partial-label learning, since the latter assumes that only one label is valid and may ignore the correlation among candidate labels. To tackle the PML challenge, we introduce a feature induced PML approach called fPML, which simultaneously estimates noisy labels and trains multi-label classifiers. In particular, fPML simultaneously factorizes the observed instance-label association matrix and the instance-feature matrix into low-rank matrices to achieve coherent low-rank matrices from the label and the feature spaces, and a low-rank label correlation matrix as well. The low-rank approximation of the instance-label association matrix is leveraged to estimate the association confidence. To predict the labels of unlabeled instances, fPML learns a matrix that maps the instances to labels based on the estimated association confidence. An empirical study on public multi-label datasets with injected noisy labels, and on archived proteomic datasets, shows that fPML can more accurately identify noisy labels than related solutions, and consequently can achieve better performance on predicting labels of instances than competitive methods.**

***Index Terms*—multi-label learning, low-rank matrix factorization, noisy labels, label correlations**

I. Introduction

Multi-label learning aims at learning from instances associated with multiple semantic labels and has attracted ever-increasing research interest in various domains [1], [2]. Recent years have witnessed the proliferation and success of multi-label learning in assigning a set of appropriate labels to unlabeled instances [3], [4], and replenishing missing labels for weakly-labeled instances [5], [6]. It's known that labels of multi-label instances are semantically correlated, and incorporating the correlations into multi-label learning can boost the performance [7], [8], [9], [6].

However, the performance of multi-label learning may be compromised by *noisy* (or incorrect) labels of training instances. Most existing multi-label learning methods, in fact, generally assume that the given labels are noise-free. But *noisy* labels often exist in practical applications, since multi-label instances are annotated by humans with a wide-range of expertise levels, different backgrounds and dedication [10]. For example, the multi-label image in Figure 1 is tagged as 'seaside', 'sunset', 'sky', 'cloud', 'sandbeach', 'ship', and 'people', but it should *not* be tagged as ~~'sandbeach'~~, ~~'ship'~~, and ~~'people'~~. In other words, 'sandbeach', 'ship', and 'people' are *noisy* (or irrelevant) labels of this image.

Fig. 1. An illustrative example of a multi-label instance with noisy labels. The multi-label image is tagged as 'seaside', 'sunset', 'sky', 'cloud', 'sandbeach', 'ship' and 'people', but the labels ~~'sandbeach'~~, ~~'ship'~~, and ~~'people'~~ do not apply.

Despite the vast progress achieved in multi-label learning, how to identify relevant labels in a candidate label set assigned to multi-label instances remains a largely *unexplored* topic. The identification task becomes much more challenging because the ground-truth labels are concealed in a set of candidate noisy labels, and the number of ground-truth labels is also unknown. Xie *et al.* [11] formalized this problem in a new *partial multi-label learning* (PML) framework, and proposed two approaches (PML-fp and PML-lc) to optimize the label confidence values and the relevance ordering of labels of each instance by exploiting structural information in feature and label spaces, and by minimizing the confidence weighted ranking loss.

Since labels are correlated, the label correlation and the *ground-truth* instance-label association matrices have a linear dependence structure, and thus they are low-rank [12], [6]. More importantly, the low-rank approximation of a *noisy* matrix is robust to noise [13], [14]. Thus, we seek the ground-truth instance-label association matrix via learning the low-rank approximation of the *observed* association matrix, which contains noisy associations. In addition, the labels of an instance

depend on its features, and thus the features of instances should be used to estimate noisy labels. As such, to tackle the PML problem, in this paper we introduce a novel *f*eature-induced *P*artial *M*ulti-label *L*earning approach called fPML, which leverages a low-rank matrix approximation and latent dependencies between labels and features to simultaneously identify the noisy labels and train a multi-label predictor. More specifically, fPML simultaneously factorizes the *observed* instance-label association matrix and the instance-feature matrix into low-rank matrices to achieve coherent low-rank matrices from label and feature spaces, and also a low-rank label correlation matrix. The low-rank approximation of the observed instance-label association matrix is leveraged to estimate the association confidence. Furthermore, to map input instances to output labels, fPML learns a matrix that maps instances to labels based on the estimated label confidence. An empirical study on public multi-label datasets with randomly injected noisy labels, as well as on archived proteomic datasets, demonstrates the effectiveness of fPML in identifying noisy labels of labeled instances, and in predicting labels of unlabeled instances.

II. Related Work

Partial multi-label learning deals with training instances annotated with a set of candidate labels, among which only a subset is relevant. It is closely related to partial-label learning, but the latter learning paradigm deals with the scenario in which only one label among the candidate labels is valid for the instance [15], [16], [17], [18], [19], [20], [21], [22], [23]. Another distinction is that partial multi-label learning should account for label correlation, whereas partial-label learning does not. Therefore, partial-label learning can be viewed as a special case of partial multi-label learning.

These partial-label learning approaches only work for single-label learning scenarios. They *neglect* the fact that multiple candidate labels might be all relevant for a given instance. Furthermore, they do *not* make concrete use of correlations among labels. Partial-label learning on multi-label instances was recently studied, and a framework called partial multi-label learning (PML) was proposed [11]. In [11], the authors assume that each candidate label has a confidence value of being the ground-truth label of an instance, and optimizes the classification model and the confidence values by minimizing the weighted confidence ranking loss in a unified framework. However, it has to simultaneously optimize multiple binary predictors and a very large number of confidence rankings of candidate label pairs; hence, suffers from heavy computational costs.

Label space dimensionality reduction (LSDR) is homologous to feature space dimensionality reduction and shares similar advantages: improving performance by removing irrelevant, redundant, or noisy information [24]. LSDR-based methods aim to tackle the multi-label classification problem with a large number of labels [25], [26]. These methods first compress the label space into a low-dimensional subspace, make prediction in the subspace, and then perform a mapping back into the original label space. Although these methods do not explicitly consider noisy labels assigned to training instances, they have the potential of removing noisy labels and make prediction based on annotated multi-label instances with noisy labels. In essence, these LSDR-based methods generally assume that the label correlation can be embedded and explored in the subspace spanned by a low-rank matrix [6], [4]. However, the widely-witnessed tail labels, which are assigned to only few samples but occupy the vast majority of labels, breakdown this assumption. To reuse the low-rank assumption to encode and explore label correlation, a robust extreme multi-label learning (REML) approach was introduced [4]. REML separately enforces sparsity on tail labels and low-rankness on non-tail labels for robust multi-label classification.

The label correlation matrix and the *ground-truth* instance-label association matrix are naturally low-rank [6], [12], and a low-rank approximation of a noisy matrix can be robust to noise [13], [14]. In addition, labels of instances depend on the features of instances. Given this, fPML simultaneously factorizes the *observed* instance-label association matrix with noisy associations and the instance-feature matrix into a coherent low-rank matrix and a low-rank label correlation matrix. fPML then estimates the label confidence based on the low-rank approximation of the observed instance-label association matrix, and learns the predictive matrix that maps the input instances to output labels via the estimated label confidence. Our empirical study shows that fPML not only can accurately identify irrelevant labels of labeled instances, but also assigns relevant labels to unlabeled instances more accurately than other competitive approaches [25], [26], [16], [21], [4], [11].

III. Proposed method

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ be the instance-feature data matrix for n multi-label instances in the d-dimensional feature space, and $\mathbf{Y} \in \{0,1\}^{q\times n}$ be the observed instance-label association matrix (or label matrix) with noisy labels, where each column corresponds to an instance and each row corresponds to a distinct label. If $\mathbf{x}_i$ is associated with the c-th label, $\mathbf{Y}_{ci} = 1$; otherwise, $\mathbf{Y}_{ci} = 0$. The goal of fPML is to identify noisy labels of labeled multi-label instances, and to predict labels of unlabeled instances.

Suppose $\widehat{\mathbf{Y}} \in \{0,1\}^{q\times n}$ is the ground-truth label matrix. Since labels are correlated in multi-label learning, the label matrix is often assumed to be low-rank [27], [12]. Intuitively, the low-rank ground-truth label matrix $\widehat{\mathbf{Y}}$ can be approximated as the product of two matrices:

$$\widehat{\mathbf{Y}} \simeq \mathbf{S}\mathbf{G}^T \tag{1}$$

where $\mathbf{S} \in \mathbb{R}^{q\times k}$ and $\mathbf{G} \in \mathbb{R}^{n\times k}$. $\mathbf{G}$ and $\mathbf{S}$ respectively encode the new representation of n instances, and q labels in the k-dimensional semantic space. Note that also $\mathbf{S}$ aims at encoding label correlations between the q labels and the k new semantic labels. Each of the original q labels may be affected by all the k new semantic labels, which implies a high-order one-to-all label correlation [28], [7].

To learn $\widehat{\mathbf{Y}}$, we minimize the reconstruction error between the observed label matrix $\mathbf{Y}$ and the product of $\mathbf{G}$ and $\mathbf{S}$ as follows:

$$\min_{\mathbf{S},\mathbf{G}} \|\mathbf{Y} - \mathbf{S}\mathbf{G}^T\|_F^2 \tag{2}$$

From the perspective of matrix factorization, Eq. (2) decomposes $\mathbf{Y}$ into two low-rank matrices $\mathbf{G}$ and $\mathbf{S}$. The product of $\mathbf{G}$ and $\mathbf{S}$ gives the low-rank approximation of $\mathbf{Y}$. A low-rank approximation ($\mathbf{S}\mathbf{G}^T$) of the original matrix ($\mathbf{Y}$) with a low reconstruction error is able to eliminate noisy entries of the original matrix [13]. In multi-label learning, common techniques to measure label correlations include co-occurrence rate [29], [11] and cosine distance [30]. However, these measures become unreliable in the presence of noisy labels, as the observed label distribution is different from the ground-truth. A unreliable label correlation measurement may even compromise the performance of multi-label learning. Given this, we use $\mathbf{S}$ to encode low-rank label correlations, and fix $k = q$.

The labels of a multi-label instance depend on the features of the instance. In addition, matrix factorization techniques have became popular in recent years for data representation in a semantic space [31]. Given these observations, to pursue a coherent low-rank representation of instances ($\mathbf{G}$) and to identify noisy entries of the association matrix, we collaboratively factorize $\mathbf{X}$ and $\mathbf{Y}$ into an identical k-dimensional semantic space by minimizing the objective function as follows:

$$\min_{\mathbf{S},\mathbf{F},\mathbf{G}} \|\mathbf{Y} - \mathbf{S}\mathbf{G}^T\|_F^2 + \lambda_1 \|\mathbf{X} - \mathbf{F}\mathbf{G}^T\|_F^2 \tag{3}$$

where $\mathbf{F} \in \mathbb{R}^{d\times k}$ explores and captures the interrelationships among features. $\lambda_1 > 0$ is a positive scalar parameter that balances the importance of the instance-feature data matrix and the instance-label association matrix. As such, noisy associations that are *inconsistent* with the latent relationship between features and labels are more likely to be identified as irrelevant labels. If $\lambda_1 = 0$, the low-rank representation of n instances is solely pursued by the observed label matrix, disregarding the feature information of these instances. This extreme setting is not expected, so $\lambda_1 > 0$. Different from $\mathbf{G}$ and $\mathbf{S}$ in Eq. (2), $\mathbf{G}$ in Eq. (3) coherently encodes the low-rank representation of n instances by simultaneously considering the instance-label association information and feature information, and $\mathbf{S}$ encodes the label correlation by additionally leveraging the latent dependency between labels and features. The coherence is pursued by sharing $\mathbf{G}$ with the data matrix and the association matrix.

To predict the relevant labels of unlabeled instances, we need to learn a matrix $\mathbf{W} \in \mathbb{R}^{d\times q}$ to map the instances to the labels. Intuitively, $\mathbf{W}$ can be optimized by minimizing the square loss as follows:

$$\min_{\mathbf{W}} \|\mathbf{Y} - \mathbf{W}^T\mathbf{X}\|_F^2 \tag{4}$$

However, Eq. (4) is biased because of the noisy associations which exist in $\mathbf{Y}$. Thus, we learn the mapping matrix $\mathbf{W}$ via minimizing the square loss between the instance-label mapping and the low-rank approximated instance-label matrix, which eliminates noisy labels. As a result, Eq. (4) is modified as follows:

$$\min_{\mathbf{W}} \|\mathbf{S}\mathbf{G}^T - \mathbf{W}^T\mathbf{X}\|_F^2 \tag{5}$$

To simultaneously identify the noisy labels and train the multi-label classifiers, the unified objective function of fPML is defined as follows:

$$\begin{aligned} \min_{\mathbf{S},\mathbf{F},\mathbf{G},\mathbf{W}} & \|\mathbf{Y} - \mathbf{S}\mathbf{G}^T\|_F^2 + \lambda_1 \|\mathbf{X} - \mathbf{F}\mathbf{G}^T\|_F^2 \\ & + \lambda_2 \|\mathbf{S}\mathbf{G}^T - \mathbf{W}^T\mathbf{X}\|_F^2 + \lambda_3 \|\mathbf{W}\|_1 \\ & s.t.\ \ \mathbf{S} \geq 0,\ \mathbf{G} \geq 0 \end{aligned} \tag{6}$$

where $\|\mathbf{W}\|_1$ controls the complexity of the induced prediction model. λ_1, λ_2, and λ_3 are tradeoff parameters. Eq. (6) simultaneously factorizes the observed instance-label association matrix and the instance-feature matrix into low-rank matrices to achieve the coherent low-rank matrix $\mathbf{G}$, and also a low-rank label correlation matrix $\mathbf{S}$. The low-rank approximation ($\mathbf{S}\mathbf{G}^T$) of the original instance-label association matrix reflects the label confidence, and is leveraged to learn the mapping matrix $\mathbf{W}$. If $\lambda_1 = 0$, we can also simultaneously achieve the identification of noisy labels and the multi-label predictor, but $\mathbf{G}$ is not pursued in a coherent feature and label space. Our investigation shows that $\lambda_1 > 0$ generally gives a more prominent performance for the identification of noisy labels. In other words, the feature induced information can boost the performance of partial multi-label learning.

The optimization problem in Eq. (6) is non-convex with respect to $\mathbf{S}$, $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{W}$ together. It is therefore unrealistic to expect to find the global optimal solutions for all the variables at the same time. In addition, the optimization of $\mathbf{W}$ is an l_1-norm regularization problem. From these observations, we use an iterative algorithm to optimize $\{\mathbf{G}, \mathbf{S}, \mathbf{F}, \mathbf{W}\}$. Particularly, $\{\mathbf{G}, \mathbf{S}, \mathbf{F}\}$ are optimized by leveraging the techniques used in standard NMF [32], and $\mathbf{W}$ is optimized by using the Accelerated Gradient Descent (AGD) algorithm [33], [34]. Leveraging the robustness of low-rank representations to noisy features [13], [32], [14], fPML reconstructs the numeric instance-label association matrix as: $\widetilde{\mathbf{Y}} = \mathbf{S}\mathbf{G}^T$. Each entry of $\widetilde{\mathbf{Y}}$ reflects the association confidence between a particular label and a particular instance. The associations available in $\mathbf{Y}$ but with low confidence values in $\widetilde{\mathbf{Y}}$ are more likely to be noisy labels, since they are not consistent with the latent relationship between features and labels of instances, and they are also not consistent with the correlation between labels. fPML then predicts the label distribution of unlabeled instances as $f(\mathbf{x}) = \mathbf{W}^T\mathbf{x}$.

IV. Experiments

A. Experimental setup

Datasets: For a comprehensive performance evaluation, we conduct experiments on six real-world datasets as listed in Table I. Enron and Yeast are two multi-label datasets collected

from Mulan[1]. Slashdot is a widely used multi-label text dataset[2]. Since there are no off-the-shelf multi-label datasets that can be directly used to validate the performance of identifying noisy labels in multi-label partial-label learning, we assume that the available labels of multi-label instances in these datasets are noise-free, and randomly inject additional 3 labels to each instance as noisy labels. The other three datasets, YeastBP, YeastCC, and YeastMF, are protein-protein interaction datasets collected from BioGrid[3]. We downloaded the functional annotations of Yeast proteins archived on different periods (historical: 2016-03-14, recent: 2017-03-13) from the Gene Ontology[4], and took the annotations available in history but absent in more recent times as *noisy* labels. Functional labels of proteins are divided in three orthogonal branches of the Gene Ontology: cellular component (CC), molecular function (MF), and biological process (BP). These functional labels are rather unbalanced. Many labels are associated to no more than 30 proteins, and few labels are associated to more than 300 proteins. To mitigate the imbalance impact, we consider labels that are associated to at least 100 proteins and at most 300 proteins for the experiments. As a result, we consider 50 CC labels, 39 MF labels, and 217 BP labels for YeastCC, YeastMF, and YeastBP, respectively. The numbers of noisy annotations of these three datasets are 260, 234, and 2385, respectively. These datasets are from different domains, have different feature representations, average numbers of labels per instance, and numbers of distinct labels.

TABLE I
CHARACTERISTICS OF THE DATASETS USED FOR THE EXPERIMENTS.

Dataset	Instances	Features	Labels	Avg	Noise
Enron	1702	1001	53	3.378	-
Slashdot	3782	1079	22	0.893	-
Yeast	2417	103	14	4.237	-
YeastBP	6139	6139	217	5.537	2385
YeastCC	6139	6139	50	1.348	260
YeastMF	6139	6139	39	1.005	234

Comparing methods: We compare fPML against PLST [25], ML-CSSP [26], REML [4], IPAL [16], PL-LEAF [21], and PML-fp [11]. PLST and ML-CSSP are LSDR-based multi-label learning methods that can be directly adopted to identify noisy labels of multi-label instances. REML is a robust multi-label learning method that uses the low-rankness assumption to explore label correlation. REML can also be adopted to estimate the association confidence between labels and instances, and thus to identify noisy labels. IPAL and PL-LEAF are representative partial-label learning methods. We extend them for multi-label learning by choosing the most confident labels as the ground-truth. PML-fp is a partial multi-label learning method introduced in [11]; it optimizes the ground-truth confidence values of candidate labels by exploiting the structure information from feature space. All these methods were introduced in Section II.

[1]http://mulan.sourceforge.net/datasets-mlc.html
[2]http://cse.seu.edu.cn/PersonalPage/zhangml/
[3]https://thebiogrid.org/
[4]http://www.geneontology.org

Evaluation metrics: We use three representative multi-label learning and partial-label learning evaluation metrics: RankingLoss (RankLoss), OneError, and AveragePrecision (AvgPrec)[1]. The smaller the values of RankLoss and OneError, the better the performance is. The larger the value of AvgPrec, the better the performance is. We report *1-RankLoss* and *1-OneError* in the following experiments. As such, a *larger* value implies a *better* performance.

B. Identification of noisy labels

Following the experimental protocol in [5], [16], we considered all instances in each dataset as both training and testing data, and performed experiments to investigate the performance of fPML on identifying noisy labels of labeled instances. Specifically, the labels training instances involve randomly injected noisy labels, while the labels of testing instances are the ground-truth labels without injected noisy labels.

The noisy labels of an instance are typically unknown in advance, thus the input parameters of fPML and of the competing methods cannot be tuned based on the unknown number of noisy labels. One possible tuning protocol is to assume that the multi-label instances with randomly injected noisy labels are noise-free, and inject an additional random label. Next, we can tune the optimal parameters, including values for λ_1, λ_2, and λ_3, over the task of identifying the added label, and then adopt the tuned parameter values for identifying the noisy labels. To simplify the implementation, λ_2 is fixed to 1 on all datasets for fPML. fPML may achieve a better performance on these datasets when the parameters are tuned. The input parameters of the competitive methods are fixed (or optimized) as suggested by the authors in their code, or respective papers. Table II reports the average results over 10 independent runs for all methods on the multi-label and archived proteomic datasets. Since PML-fp has a high time-complexity on datasets with a large number of labels, and the computation of PML-fp in one round on YeastBP could not complete after three days, its results on YeastBP cannot be reported.

From Table II, we can observe the following: (i) On all datasets, fPML significantly outperforms PLST and ML-CSSP across all evaluation metrics; (ii) fPML frequently outperforms REML, IPAL, PL-LEAF, and PML-fp on most cases; (iii) Although 1-OneError is biased towards partial-label learning, IPAL and PL-LEAF often lose to fPML when the 1-OneError measure is used. These observations show that correlations among labels should be exploited in partial multi-label learning and also demonstrate the effectiveness of fPML on identifying noisy labels.

C. Prediction of unlabeled instances

We performed another set of experiments to study the performance of fPML in predicting the labels of unlabeled instances. The parameter settings are kept the same as in the previous experiments. We independently repeat the experiments 10 times on each dataset and report the average results in Table III. A protein-protein interaction network is too sparse

TABLE II

PERFORMANCE FOR THE IDENTIFICATION OF NOISY LABELS AS THE NUMBER OF RANDOMLY INJECTED NOISY LABELS INCREASES. GROUND-TRUTH NOISY LABELS OF YEASTCC, YEASTMF, AND YEASTBP ARE KNOWN (NO LABELS ARE INJECTED FOR THESE DATASETS). $\bullet/\circ$ INDICATES WHETHER FPML IS STATISTICALLY (ACCORDING TO PAIRWISE t-TEST AT 95% SIGNIFICANCE LEVEL) SUPERIOR/INFERIOR TO THE OTHER METHOD.

Dataset	Metric	PLST	ML-CSSP	REML	IPAL	PL-LEAF	PML-fp	fPML
Enron	1-RankLoss	$0.811 \pm 0.006\bullet$	$0.675 \pm 0.092\bullet$	$0.989 \pm 0.000\bullet$	$0.981 \pm 0.000\bullet$	$0.987 \pm 0.000\bullet$	$0.993 \pm 0.000\bullet$	0.994 ± 0.000
	1-OneError	$0.930 \pm 0.008\bullet$	$0.768 \pm 0.070\bullet$	$0.937 \pm 0.004\bullet$	$0.822 \pm 0.005\bullet$	$0.939 \pm 0.003\bullet$	$0.957 \pm 0.002\circ$	0.954 ± 0.006
	AvgPrec	$0.788 \pm 0.007\bullet$	$0.621 \pm 0.090\bullet$	$0.917 \pm 0.002\bullet$	$0.802 \pm 0.003\bullet$	$0.875 \pm 0.003\bullet$	$0.922 \pm 0.000\bullet$	0.933 ± 0.004
Slashdot	1-RankLoss	$0.873 \pm 0.005\bullet$	$0.685 \pm 0.069\bullet$	$0.971 \pm 0.001\bullet$	$0.969 \pm 0.000\bullet$	$0.974 \pm 0.001\bullet$	$0.977 \pm 0.000\bullet$	0.979 ± 0.002
	1-OneError	$0.508 \pm 0.010\bullet$	$0.392 \pm 0.048\bullet$	0.769 ± 0.006	$0.681 \pm 0.004\bullet$	$0.732 \pm 0.007\bullet$	$0.739 \pm 0.004\bullet$	0.767 ± 0.009
	AvgPrec	$0.678 \pm 0.007\bullet$	$0.528 \pm 0.054\bullet$	$0.845 \pm 0.004\bullet$	$0.791 \pm 0.002\bullet$	$0.822 \pm 0.004\bullet$	$0.833 \pm 0.001\bullet$	0.850 ± 0.009
Yeast	1-RankLoss	$0.882 \pm 0.005\bullet$	$0.816 \pm 0.046\bullet$	$0.934 \pm 0.001\bullet$	$0.904 \pm 0.002\bullet$	$0.930 \pm 0.001\bullet$	$0.836 \pm 0.000\bullet$	0.945 ± 0.001
	1-OneError	$0.773 \pm 0.015\bullet$	$0.567 \pm 0.179\bullet$	$0.839 \pm 0.004\bullet$	$0.776 \pm 0.005\bullet$	$0.853 \pm 0.003\bullet$	$0.862 \pm 0.003\bullet$	0.898 ± 0.012
	AvgPrec	$0.760 \pm 0.007\bullet$	$0.673 \pm 0.068\bullet$	$0.860 \pm 0.003\bullet$	$0.808 \pm 0.004\bullet$	$0.860 \pm 0.001\bullet$	$0.829 \pm 0.000\bullet$	0.893 ± 0.003
YeastMF	1-RankLoss	$0.720 \pm 0.000\bullet$	$0.507 \pm 0.174\bullet$	$0.974 \pm 0.001\circ$	$0.978 \pm 0.000\circ$	$0.982 \pm 0.000\circ$	$0.980 \pm 0.000\circ$	0.964 ± 0.027
	1-OneError	$0.575 \pm 0.000\bullet$	$0.448 \pm 0.188\bullet$	$0.704 \pm 0.040\bullet$	$0.550 \pm 0.000\bullet$	$0.638 \pm 0.000\bullet$	$0.702 \pm 0.000\bullet$	0.717 ± 0.058
	AvegPrec	$0.645 \pm 0.000\bullet$	$0.467 \pm 0.162\bullet$	$0.835 \pm 0.018\circ$	$0.753 \pm 0.000\bullet$	$0.806 \pm 0.000\bullet$	$0.815 \pm 0.000\circ$	0.810 ± 0.046
YeastCC	1-RankLoss	$0.540 \pm 0.000\bullet$	$0.412 \pm 0.076\bullet$	0.971 ± 0.001	0.990 ± 0.000	0.990 ± 0.000	$0.984 \pm 0.000\bullet$	0.991 ± 0.001
	1-OneError	$0.623 \pm 0.000\bullet$	$0.396 \pm 0.126\bullet$	$0.732 \pm 0.025\bullet$	$0.736 \pm 0.000\bullet$	$0.717 \pm 0.000\bullet$	$0.763 \pm 0.000\bullet$	0.824 ± 0.029
	AvegPrec	$0.524 \pm 0.000\bullet$	$0.388 \pm 0.085\bullet$	$0.846 \pm 0.011\bullet$	$0.857 \pm 0.000\bullet$	$0.860 \pm 0.000\bullet$	$0.817 \pm 0.000\bullet$	0.892 ± 0.011
YeastBP	1-RankLoss	$0.282 \pm 0.000\bullet$	$0.196 \pm 0.039\bullet$	$0.978 \pm 0.000\bullet$	0.994 ± 0.000	0.994 ± 0.000	$--$	0.994 ± 0.001
	1-OneError	$0.346 \pm 0.000\bullet$	$0.309 \pm 0.042\bullet$	$0.768 \pm 0.007\bullet$	$0.768 \pm 0.000\bullet$	$0.811 \pm 0.000\circ$	$--$	0.790 ± 0.032
	AvegPrec	$0.258 \pm 0.000\bullet$	$0.196 \pm 0.034\bullet$	$0.812 \pm 0.004\bullet$	$0.818 \pm 0.000\bullet$	$0.839 \pm 0.000\circ$	$--$	0.833 ± 0.018

to be used as feature data matrix for prediction, so we do not use this kind of datasets for this experiment. To study the prediction performance of fPML under different ratios of training instances, we vary the ratios (TRatio) from 50% to 80%. The remaining 50% and 20% instances are testing data, respectively. The training instances are labeled, while the labels of the test data are only used for validation. We simulate a noisy label setting with $m = 3$ as before for the training instances, and predict a set of relevant labels for the test instances.

From Table III, we can observe the following: (i) fPML almost always outperforms PLST, ML-CSSP, REML, and IPAL on all datasets; (ii) fPML outperforms PML-fp on Enron and Slashdot datasets, but loses to PML-fp on Yeast dataset; (iii) Out of 18 cases (3 datasets $\times$ 3 metrics $\times$ 2 TRatios), fPML achieves a performance superior to PL-LEAF in 9 cases, and an inferior performance in 6 cases. PL-LEAF is a two-stage approach, and uses numeric label confidence values to train a predictive model, like fPML does. But unlike fPML, PL-LEAF does not consider correlations among labels. In addition, PL-LEAF estimates label confidence by encoding the manifold structure of the feature space into the label space, while fPML utilizes the instance-feature matrix and the instance-label association matrix for a coherent low-rank matrix approximation to estimate label confidences, which is robust to noisy labels. Therefore, PL-LEAF is outperformed by fPML in many cases. Although PL-LEAF achieves a superior performance in several cases, its time complexity is much higher than that of fPML. Like fPML, PML-fp also takes into account feature information, but it often loses to fPML. This observation further demonstrates the effectiveness of feature-induced low-rank matrix approximation for identifying noisy labels.

In addition, we investigated the benefit of identifying noisy labels. For this investigation, we introduce fPML-Y, a variant fPML, which directly optimizes $\mathbf{W}$ with an l_1 regularization norm based on the observed instance-label associations ($\mathbf{Y}$, instead of $\mathbf{SG}^T$), and then applies $\mathbf{W}$ to make predictions on unlabeled data. The experimental results on Enron and Slashdot (see Figure 2) show that fPML significantly outperforms fPML-Y. Thus, identifying noisy labels indeed improves the performance of multi-label learning.

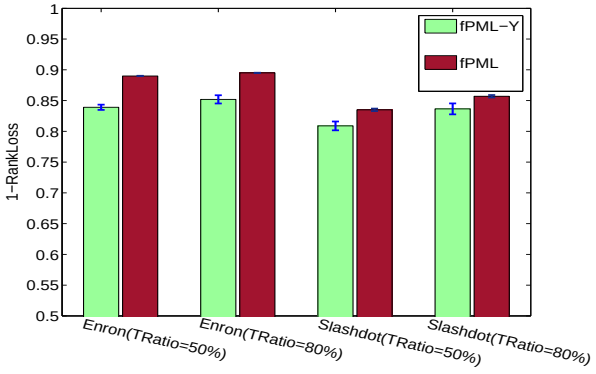

Fig. 2. *1-RankLoss* of fPML and its variant (fPML-Y) on predicting the labels of unlabeled instances on Enron and Slashdot. fPML-Y uses the observed instance-label associations for training without identifying noisy labels.

V. CONCLUSIONS

In this paper, we study an interesting but seldom explored variation of multi-label learning, called partial multi-label learning, which aims at identifying noisy labels of instances associated with multiple inter-correlated labels, and predicting labels of unlabeled instances using noisy labeled instances. To reach this goal, we introduce a unified model called fPML. Extensive experiments clearly support the superiority of fPML against competitive techniques. For simplicity and to automatically explore label correlation, fPML adopts a square label correlation matrix whose size is equal to the number of distinct labels. fPML can also work on datasets with a large label space by using a smaller sized non-square matrix. We will investigate the performance of fMPL on large-scale datasets, and on other types of noisy labels (i.e., false negative labels).

TABLE III

PERFORMANCE FOR PREDICTION WITH NOISY LABELS. $\bullet/\circ$ INDICATES WHETHER FPML IS STATISTICALLY (ACCORDING TO PAIRWISE t-TEST AT 95% SIGNIFICANCE LEVEL) SUPERIOR/INFERIOR TO THE OTHER METHOD.

	TRatio	PLST	ML-CSSP	REML	IPAL	PL-LEAF	PML-fp	fPML
				Enron				
1-RankLoss	50%	$0.719 \pm 0.008\bullet$	$0.715 \pm 0.014\bullet$	$0.875 \pm 0.005\bullet$	$0.690 \pm 0.014\bullet$	$0.853 \pm 0.005\bullet$	$0.874 \pm 0.003\bullet$	0.890 ± 0.006
	80%	$0.727 \pm 0.012\bullet$	$0.725 \pm 0.015\bullet$	$0.881 \pm 0.007\bullet$	$0.675 \pm 0.015\bullet$	$0.852 \pm 0.009\bullet$	$0.884 \pm 0.005\bullet$	0.895 ± 0.010
1-OneError	50%	$0.519 \pm 0.014\bullet$	$0.450 \pm 0.043\bullet$	$0.746 \pm 0.018\circ$	$0.719 \pm 0.016\bullet$	$0.745 \pm 0.013\circ$	$0.720 \pm 0.017\bullet$	0.738 ± 0.028
	80%	$0.515 \pm 0.029\bullet$	$0.484 \pm 0.024\bullet$	$0.743 \pm 0.017\bullet$	$0.720 \pm 0.021\bullet$	$0.741 \pm 0.025\bullet$	$0.728 \pm 0.016\bullet$	0.757 ± 0.022
AvgPrec	50%	$0.456 \pm 0.010\bullet$	$0.431 \pm 0.016\bullet$	$0.645 \pm 0.007\bullet$	$0.535 \pm 0.013\bullet$	$0.625 \pm 0.001\bullet$	$0.618 \pm 0.007\bullet$	0.659 ± 0.020
	80%	$0.467 \pm 0.013\bullet$	$0.460 \pm 0.014\bullet$	$0.650 \pm 0.009\bullet$	$0.532 \pm 0.017\bullet$	$0.626 \pm 0.014\bullet$	$0.635 \pm 0.007\bullet$	0.671 ± 0.017
				Slashdot				
1-RankLoss	50%	$0.742 \pm 0.008\bullet$	$0.729 \pm 0.012\bullet$	$0.818 \pm 0.005\bullet$	$0.711 \pm 0.011\bullet$	0.834 ± 0.009	$0.827 \pm 0.008\bullet$	0.835 ± 0.005
	80%	$0.791 \pm 0.006\bullet$	$0.776 \pm 0.009\bullet$	$0.828 \pm 0.012\bullet$	$0.722 \pm 0.016\bullet$	$0.849 \pm 0.008\bullet$	$0.835 \pm 0.005\bullet$	0.857 ± 0.008
1-OneError	50%	$0.337 \pm 0.012\bullet$	$0.282 \pm 0.057\bullet$	$0.481 \pm 0.010\bullet$	$0.467 \pm 0.012\bullet$	$0.545 \pm 0.013\circ$	$0.459 \pm 0.011\bullet$	0.516 ± 0.010
	80%	$0.421 \pm 0.016\bullet$	$0.370 \pm 0.040\bullet$	$0.516 \pm 0.014\bullet$	$0.491 \pm 0.017\bullet$	$0.578 \pm 0.013\circ$	$0.494 \pm 0.013\bullet$	0.565 ± 0.017
AvgPrec	50%	$0.466 \pm 0.008\bullet$	$0.419 \pm 0.046\bullet$	$0.576 \pm 0.008\bullet$	$0.563 \pm 0.011\bullet$	$0.628 \pm 0.011\circ$	$0.568 \pm 0.011\bullet$	0.609 ± 0.008
	80%	$0.538 \pm 0.009\bullet$	$0.494 \pm 0.031\bullet$	$0.602 \pm 0.013\bullet$	$0.491 \pm 0.015\bullet$	0.653 ± 0.011	$0.598 \pm 0.011\bullet$	0.652 ± 0.013
				Yeast				
1-RankLoss	50%	$0.786 \pm 0.007\bullet$	$0.766 \pm 0.015\bullet$	$0.791 \pm 0.003\bullet$	$0.742 \pm 0.007\bullet$	$0.794 \pm 0.004\bullet$	$0.821 \pm 0.003\circ$	0.803 ± 0.005
	80%	$0.800 \pm 0.011\bullet$	$0.785 \pm 0.015\bullet$	$0.792 \pm 0.006\bullet$	$0.747 \pm 0.010\bullet$	$0.801 \pm 0.011\bullet$	$0.816 \pm 0.013\circ$	0.809 ± 0.011
1-OneError	50%	$0.715 \pm 0.007\bullet$	$0.683 \pm 0.068\bullet$	$0.751 \pm 0.008\circ$	$0.732 \pm 0.012\bullet$	$0.762 \pm 0.007\circ$	$0.770 \pm 0.013\circ$	0.746 ± 0.015
	80%	$0.749 \pm 0.020\bullet$	$0.702 \pm 0.056\bullet$	$0.745 \pm 0.015\bullet$	$0.738 \pm 0.015\bullet$	$0.773 \pm 0.009\circ$	$0.750 \pm 0.016\bullet$	0.755 ± 0.018
AvgPrec	50%	$0.716 \pm 0.006\bullet$	$0.691 \pm 0.033\bullet$	$0.708 \pm 0.005\bullet$	$0.692 \pm 0.007\bullet$	$0.734 \pm 0.005\bullet$	$0.749 \pm 0.006\circ$	0.738 ± 0.008
	80%	$0.734 \pm 0.012\bullet$	$0.708 \pm 0.025\bullet$	$0.707 \pm 0.008\bullet$	$0.699 \pm 0.009\bullet$	0.744 ± 0.011	0.741 ± 0.015	0.745 ± 0.012

The code of fMPL and three proteomic datasets are available at http://mlda.swu.edu.cn/codes.php?name=fMPL.

VI. ACKNOWLEDGMENTS

This work is supported by Natural Science Foundation of China (61872300, 61741217 and 61873214), Natural Science Foundation of CQ CSTC (cstc2018jcyjAX0228 and cstc2016jcyjA0351), the Open Research Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing(KLIGIP-2017A05), and Chongqing Graduate Student Research Innovation Project (No. CYS18089).

REFERENCES

[1] M. Zhang and Z. Zhou, "A review on multi-label learning algorithms," *TKDE*, vol. 26, no. 8, pp. 1819–1837, 2014.

[2] E. Gibaja and S. Ventura, "A tutorial on multilabel learning," *ACM Computing Surveys*, vol. 47, no. 3, p. 52, 2015.

[3] P. Hou, X. Geng, and M.-L. Zhang, "Multi-label manifold learning," in *AAAI*, 2016, pp. 1680–1686.

[4] C. Xu, D. Tao, and C. Xu, "Robust extreme multi-label learning," in *KDD*, 2016, pp. 1275–1284.

[5] Y. Sun, Y. Zhang, and Z. Zhou, "Multi-label learning with weak label," in *AAAI*, 2010, pp. 593–598.

[6] L. Xu, Z. Wang, Z. Shen, Y. Wang, and E. Chen, "Learning low-rank label correlations for multi-label classification with missing labels," in *ICDM*, 2014, pp. 1067–1072.

[7] M.-L. Zhang and K. Zhang, "Multi-label learning by exploiting label dependency," in *KDD*, 2010, pp. 999–1008.

[8] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine Learning*, vol. 85, no. 3, p. 333, 2011.

[9] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Random k-labelsets for multilabel classification," *TKDE*, vol. 23, no. 7, pp. 1079–1089, 2011.

[10] N. Q. V. Hung, H. H. Viet, N. T. Tam, M. Weidlich, H. Yin, and X. Zhou, "Computing crowd consensus with partial agreement," *TKDE*, vol. 30, no. 1, pp. 1–14, 2018.

[11] M.-K. Xie and S.-J. Huang, "Partial multi-label learning," in *AAAI*, 2018, pp. 1–8.

[12] Y. Zhu, J. T. Kwok, and Z.-H. Zhou, "Multi-label learning with global and local label correlation," *TKDE*, vol. 30, no. 6, pp. 1081–1094, 2018.

[13] K. Konstantinides, B. Natarajan, and G. S. Yovanof, "Noise estimation and filtering using block-based singular value decomposition," *TIP*, vol. 6, no. 3, pp. 479–483, 1997.

[14] D. Meng and F. De La Torre, "Robust matrix factorization with unknown noise," in *ICCV*, 2013, pp. 1337–1344.

[15] T. Cour, B. Sapp, and B. Taskar, "Learning from partial labels," *JMLR*, vol. 12, no. 5, pp. 1501–1536, 2011.

[16] M. Zhang and F. Yu, "Solving the partial label learning problem: An instance-based approach." in *IJCAI*, 2015, pp. 4048–4054.

[17] F. Yu and M.-L. Zhang, "Maximum margin partial label learning," *Machine Learning*, vol. 104, no. 4, pp. 573–593, 2017.

[18] E. Hüllermeier and J. Beringer, "Learning from ambiguously labeled examples," *Intelligent Data Analysis*, vol. 10, no. 5, pp. 419–439, 2006.

[19] L. Liu and T. G. Dietterich, "A conditional multinomial mixture model for superset label learning." in *NIPS*, 2012, pp. 557–565.

[20] C.-Z. Tang and M.-L. Zhang, "Confidence-rated discriminative partial label learning," in *AAAI*, 2017, pp. 2611–2617.

[21] M. Zhang, B. Zhou, and X. Liu, "Partial label learning via feature-aware disambiguation," in *KDD*, 2016, pp. 1335–1344.

[22] C. Gong, T. Liu, Y. Tang, J. Yang, J. Yang, and D. Tao, "A regularization approach for instance-based superset label learning," *IEEE Transactions on Cybernetics*, vol. 48, no. 3, pp. 967–978, 2018.

[23] X. Wu and Z. Min-Ling, "Towards enabling binary decomposition for partial label learning," in *IJCAI*, 2018, pp. 2868–2874.

[24] Y.-N. Chen and H.-T. Lin, "Feature-aware label space dimension reduction for multi-label classification," in *NIPS*, 2012, pp. 1529–1537.

[25] F. Tai and H.-T. Lin, "Multilabel classification with principal label space transformation," *Neural Computation*, vol. 24, no. 9, pp. 2508–2542, 2012.

[26] W. Bi and J. Kwok, "Efficient multi-label classification with many labels," in *ICML*, 2013, pp. 405–413.

[27] Y. P. Wu and H. T. Lin, "Progressive random k-labelsets for cost-sensitive multi-label classification," *Machine Learning*, vol. 106, no. 5, pp. 671–694, 2017.

[28] J. Lee, S. Kim, G. Lebanon, and Y. Singer, "Local low-rank matrix approximation," in *ICML*, 2013, pp. 82–90.

[29] S. Diplaris, G. Tsoumakas, P. A. Mitkas, and I. Vlahavas, "Protein classification with multiple algorithms," in *In Proceedings of 10th Panhellenic Conference on Informatics*, 2005, pp. 448–456.

[30] H. Wang, H. Huang, and C. Ding, "Image annotation using bi-relational graph of images and semantic labels," in *CVPR*, 2011, pp. 793–800.

[31] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *TPAMI*, vol. 33, no. 8, pp. 1548–1560, 2011.

[32] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, 2001, pp. 556–562.

[33] Y. Nesterov, "Gradient methods for minimizing composite functions," *Mathematical Programming*, vol. 140, no. 1, pp. 125–161, 2013.

[34] J. Zhou, J. Chen, and J. Ye, "Malsar: Multi-task learning via structural regularization," *Arizona State University*, vol. 21, 2011.