Large-Scale Multi-Label Learning with Incomplete Label Assignments

Xiangnan Kong[*] Zhaoming Wu [†] Li-Jia Li[‡] Ruofei Zhang[§] Philip S. Yu [¶]
Hang Wu[‖] Wei Fan[**]

Abstract

Multi-label learning deals with the classification problems where each instance can be assigned with multiple labels simultaneously. Conventional multi-label learning approaches mainly focus on exploiting label correlations. It is usually assumed, explicitly or implicitly, that the label sets for training instances are fully labeled without any missing labels. However, in many real-world multi-label datasets, the label assignments for training instances can be incomplete. Some ground-truth labels can be missed by the labeler from the label set. This problem is especially typical when the number instances is very large, and the labeling cost is very high, which makes it almost impossible to get a fully labeled training set. In this paper, we study the problem of large-scale multi-label learning with incomplete label assignments. We propose an approach, called Mpu, based upon positive and unlabeled stochastic gradient descent and stacked models. Unlike prior works, our method can effectively and efficiently consider missing labels and label correlations simultaneously, and is very scalable, that has linear time complexities over the size of the data. Extensive experiments on two real-world multi-label datasets show that our Mpu model consistently outperform other commonly-used baselines.

1 Introduction

Multi-label learning has drawn much attention in recent years, where each data example can be assigned with multiple labels simultaneously. Multi-label learning has a wide range of real-world applications. For example, in text categorization, one text document can belong to multiple categories, such as *sports* and *entertainment*; Researchers in text mining are interested in learning a model that can automatically predict a set of categories for each text document. Similarly, in image annotation tasks, one image can contain multiple objects or tags, and researchers in computer vision are interested in automatically predicting tags/objects for unlabeled image collections.

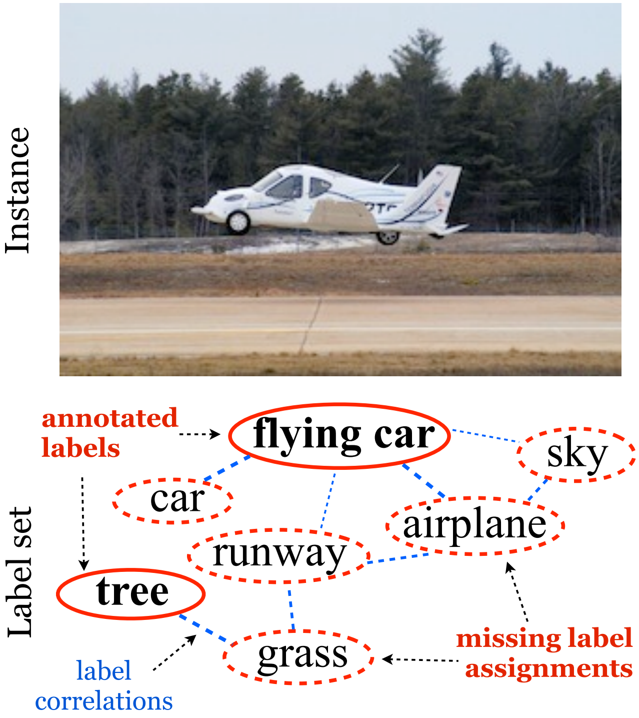

Figure 1: An example of multi-label learning with incomplete label assignments. The labels annotated by labelers are highlighted by bold font, while the other ground-truth labels are missed by the labelers, but can be inferred by exploiting label correlations.

Conventional approaches for multi-label learning [22, 10, 14, 26, 20] mainly focus on utilizing the correlations among different class labels to facilitate the learning process. It is usually assumed, explicitly or implicitly, that all the label sets for training instances are fully labeled, *i.e.*, all the labels within a label set of an instance should be completely annotated by a la-

[*] Computer Science Department, University of Illinois at Chicago, USA. xkong4@uic.edu
[†] Computer Science Department, Tsinghua University, China
[‡] Yahoo! Research, USA. lijiali@yahoo-inc.com
[§] Microsoft, USA.
[¶] Computer Science Department, University of Illinois at Chicago, USA. psyu@cs.uic.edu
[‖] Computer Science Department, Tsinghua University, China
[**] Huawei Noah's Ark Lab, Hong Kong, China.

beler without any missing labels. However, in many real-world multi-label learning tasks, it is very hard or expensive to get a fully labeled dataset, especially when the number of classes and/or instances are very large. It is usually much easier to get a set of partially labeled data, where some ground-truth labels for training instances can be missed by the labeler. For example, in Figure 1, we should an image that contain many tags in its ground-truth label set. It usually happens that the labeler may only want to annotate a small number of the labels for the image, instead of going through all possible labels in the vocabulary. In this example, many true labels of the image are missed by the labeler. If we directly use existing multi-label learning methods on such datasets, the missing labels in the training data will be treated as negative examples, and the performances of multi-label classification will degenerate greatly due to the simple treatment.

In this paper, we study the problem of large-scale multi-label learning with incomplete label assignments, as shown in Figure 2. Despite its value and significance, large-scale multi-label learning with incomplete label assignments is a much more challenging task due to the specific characteristics of the task. The reasons are listed as follows.

- *Incomplete Label Assignments.* Most existing multi-label learning methods, such as Ml-knn [25] and Rank-svm [5], assume that the training data are fully labeled. However, in most real-world applications, the label assignments for training instances can be incomplete. Thus we cannot simply treat the missing labels as negative examples. A label that is not annotated by the labeler can still belong the ground-truth label set. We need to consider the incomplete label assignments explicitly while building our models on training data.

- *Label Correlations.* Positive and Unlabeled learning methods [6, 16] can usually handle the cases when the label assignments are partially missing under single label settings. However, in multi-label learning problems, each instance can be assigned with multiple labels. Different class labels are correlated with each other, instead of being independent. We need to exploit the label correlations to facility the learning process of multi-label classification.

- *Scalability.* Previous approaches on multi-label learning with incomplete label assignments [21, 2, 18] are mainly designed for small/moderate-sized datasets. However, many real-world problems involve a large number of instances. In large-scale datasets, it is even more typical to encounter the incomplete labeling issues due to the huge cost of labeling. In these cases, it is even more important that the learning method can handle large-scale datasets, with time complexities that are linear to both the number of instances and the number of classes.

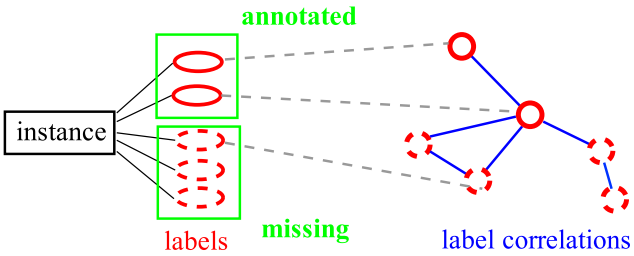

Figure 2: The framework for multi-label learning with incomplete label assignments.

In order to solve the above issues, we propose a novel solution, called Mpu, to learn from partially labeled training instances and can exploit label correlations to facilitate the learning process. Different from previous work, the proposed Mpu can scale to large-scale problems with time complexity that is linear in both the number instances and number of classes. Empirical studies on real-world datasets show that the proposed method can significantly boost the performance of multi-label classification by considering missing labels and incorporating label correlations.

The rest of the paper is organized as follows. We start by introducing the preliminary concepts, giving the problem analysis and present the Mpu approach in Section 2. Then Section 3 reports the experiment results. We briefly review on related works of multi-label learning and learning from missing labels in Section 4. In Section 5, we conclude the paper.

2 Problem Formulation

Given n data points of the form $(\mathbf{x}_i, \mathbf{y}_i)$, where $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^D$ is a D dimensional vector denoting the features of the i-th instance. $\mathbf{y}_i = (y_i^1, \cdots, y_i^q)^\top \in \{-1, 1\}^q$ corresponds to its set of ground-truth labels within a fixed dictionary of q possible labels. $y_i^k = 1$ if the k-th label is in the label set of instance i, otherwise $y_i^k = -1$. In many real-world multi-label learning tasks, the training data are usually not fully labeled, where each instance can be labeled with a subset of the ground-truth labels. We call such settings as multi-label learning with incomplete label assignments or weak labeling problems, following the definition in [21]. Specially, in the training set, the ground-truth label set

Table 1: Important Notations.

Symbol	Definition
$\mathbf{x}_1, \cdots, \mathbf{x}_n$	the feature vectors for training instances
$\mathbf{y}_1, \cdots, \mathbf{y}_n$	the set of variables for label sets of the training instances
$\mathbf{s}_1, \cdots, \mathbf{s}_n$	the set of annotated labels for the training instances
$\mathbf{y}_i = \left(y_i^1, \cdots, y_i^q\right)^\top$	the vector of variables for the label set of instance $\mathbf{x}_i$, and $y_i^k \in \{-1, 1\}$
$\mathbf{s}_i = \left(s_i^1, \cdots, s_i^q\right)^\top$	the vector of annotated labels for instance $\mathbf{x}_i$, and $s_i^k \in \{-1, 1\}$, $s_i^k \leq y_i$
$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{s}_i)\}_{i=1}^n$	the training set for multi-label learning with incomplete label assignments

$\mathbf{y}_i$ for each instance is not available. Instead, we only know a set of labels that are annotated by the labeler, denoted as $\mathbf{s}_i = (s_i^1, \cdots, s_i^q)^\top \in \{-1, 1\}^q$, where $s_i^k \leq y_i^k$ $(\forall 1 \leq i \leq n, \forall 1 \leq k \leq q)$. Thus, $\mathbf{s}_i$ only provides partial labels for instance $\mathbf{x}_i$. When $s_i^k = 1$, $y_i^k = 1$ is certain.

$$\Pr\left(s_i^k = 1 | \mathbf{x}_i, y_i^k = -1\right) = 0, \ \forall i, k$$

But when $s_i^k = -1$, either $y_i^k = -1$ or $y_i^k = 1$ can be true. In multi-label learning with incomplete label assignments, the training set is a set of partially labeled instances $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{s}_i, \mathbf{y}_i)\}_{i=1}^n$, where $\mathbf{y}_i$'s are unknown. The learning task is to learn a prediction model $f(\cdot, \cdot)$ from $\mathcal{D}$, such that for any unseen test data $\mathbf{x}_i$, the prediction $f(\mathbf{x}_i, \cdot)$ should be close to the ground-truth, *i.e.*, for any $\mathbf{z} \in \{-1, 1\}^q$, $\mathbf{f}(\mathbf{x}_i, \mathbf{z}) = \Pr(\mathbf{y}_i = \mathbf{z} | \mathbf{x}_i)$ as close as possible.

The key issues of multi-label learning with incomplete label assignments are as follows:

- How can we estimate the true label sets for the training data and use them to facilitate the training process of multi-label learning?
- How can we exploit the correlations among multiple labels to improve the multi-label classification performances?

In the following sections, we will first introduce a model to estimate missing labels in the training set based upon PU (*i.e.*, Positive and Unlabeled) stochastic gradient descent. Next we will describe our framework for incororating the correlations among labels based upon stack models.

2.1 Handling Missing Labels via PU Stochastic Gradient Descent In this subsection, we first address the problem of missing label assignments while assuming all labels are independent from each other. Then in the next subsection, we will show further how to extend the model to consider label correlations.

One simple solution for multi-label learning is to one-vs-all decomposition by treating a multi-label classification problem as multiple binary classification problems (one for each label):

$$\Pr(\boldsymbol{y}_i | \boldsymbol{x}_i) = \prod_{k=1}^{q} \Pr\left(y_i^k = 1 | \boldsymbol{x}_i\right)$$

Inspired by the positive and unlabeled learning in single-label classification [6], we propose a method, called PU Stochastic Gradient Descent, which can handle large-scale datasets with missing label assignments. Let $\{\mathbf{w}_k\}_{k=1}^q$ be a set of parameters of our classifier, where $\mathbf{w}_k \in \mathbb{R}^D$. According to the principle of maximum likelihood, we need to find the optimized $\mathbf{w}_k^*$ to maximize the likelihood of y_i^k's.

$$\mathbf{w}_k{}^* = \arg\max_{\mathbf{w}_k} \ \log\left(\prod_{i=1}^{n} \Pr\left(y_i^k = 1 | \mathbf{x}_i, \mathbf{w}_k\right)\right)$$

In this method, we extend logistic regression to classification problems with incomplete label assignments as follows. Assume that y_i^k satisfies a Bernouli distribution, and

$$\Pr\left(y_i^k = 1 | \mathbf{x}_i, \mathbf{w}_k\right) = \frac{1}{1 + \exp(-\mathbf{w}_k{}^\top \mathbf{x}_i)}$$

Following the assumption in [6], we assume that annotated labels are randomly sampled from the ground-truth label set with a constant rate c, where the sampling process is totally independent everything else, such as the feature of the instance. Assume that the probability that a label is not missing by the labeler is an unknown constant c.

$$c = \Pr\left(s_i^k = 1 | y_i^k = 1\right) = \Pr\left(s_i^k = 1 | y_i^k = 1, \mathbf{x}_i, \mathbf{w}_k\right)$$

Here c can be directly estimated from the training set using cross validation process in [6]. With Bayes' theorem, we have

$$\Pr\left(y_i^k = 1 | \mathbf{x}_i, \mathbf{w}_k\right) = \frac{\Pr\left(s_i^k = 1 | \mathbf{x}_i, \mathbf{w}_k\right)}{\Pr\left(s_i^k = 1 | y_i^k = 1, \mathbf{x}_i, \mathbf{w}_k\right)}$$

```
Input:
Given a multi-label training set with missing labels $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{s}_i)\}_{i=1}^n$. The number of stacking levels $L$.
The base learner $A$, i.e., the PU stochastic gradient descent algorithm in Section 2.1
Learning:
  - When $\ell = 0$, $\forall\ k = 1, \cdots, q$, train a model $f_k^{(0)}$:
    1. Construct a training set $\mathcal{D}_k^{(0)} = \big\{(\mathbf{x}_i, s_i^k)\big\}_{i=1}^n$
    2. Train a model $f_k^{(0)} = A\Big(\mathcal{D}_k^{(0)}\Big)$, let $\mathbf{x}_i^{(0)} = \mathbf{x}_i$.
  - Learn the stacked models, for $\ell = 1, \cdots, L$ :
    1. Construct estimated predictions $\hat{\boldsymbol{y}}_i^{(\ell-1)}$ for $\mathcal{D}_k^{(\ell-1)}$ using cross-validation in Figure 4
    2. $\forall\ k = 1, \cdots, q$, train a model $f_k^{(\ell)}$:
       Construct a extended training set $\mathcal{D}_k^{(\ell)} = \Big\{(\mathbf{x}_i^{(\ell)}, s_i^k)\Big\}$
       where $\mathbf{x}_i^{(\ell)} = \Big(\mathbf{x}_i^{(\ell-1)}, \hat{\mathbf{y}}_i^{(\ell-1)}\Big)$
       Let $f_k^{(\ell)} = A\Big(\mathcal{D}_k^{(\ell)}\Big)$ be the model trained on $\mathcal{D}_k^{(\ell)}$.
Inference: given a test instance $\mathbf{x}$
  1. $\hat{\mathbf{y}}^{(0)} = \Big(f_1^{(0)}(\mathbf{x}), \cdots, f_q^{(0)}(\mathbf{x})\Big)$, let $\mathbf{x}^{(0)} = \mathbf{x}$
  2. for $\ell = 1, \cdots, L$ :
     Construct the extended testing instance $\mathbf{x}^{(\ell)} = \big(\mathbf{x}^{(\ell-1)}, \hat{\mathbf{y}}^{(\ell-1)}\big)$
     $\hat{\mathbf{y}}^{(\ell)} = \Big(f_1^{(\ell)}(\mathbf{x}^{(\ell)}), \cdots, f_q^{(\ell)}(\mathbf{x}^{(\ell)})\Big)$
Output:
  $\hat{\mathbf{y}}^{(L)}$ :   the label set prediction for the test instance.
```

Figure 3: The Mpu Algorithm based upon Stacked Graphical Learning and Inference

Then, we have

$$\begin{aligned}
&\Pr\big(s_i^k = 1 \mid \mathbf{x}_i, \mathbf{w}_k\big) = c \cdot \Pr\big(y_i^k = 1 | \mathbf{x}_i^k, \mathbf{w}_k\big)\\
&\quad = \frac{c}{1 + \exp(-\mathbf{w}_k{}^\top \mathbf{x}_i)}\\
&\quad = \frac{c}{1 + \exp\big(-s_i^k \mathbf{w}_k{}^\top \mathbf{x}_i\big)} + \frac{(1 - s_i^k)(1 - c)}{2}\\
&\Pr\big(s_i^k = -1 \mid \mathbf{x}_i, \mathbf{w}_k\big) = 1 - c \cdot \Pr\big(y_i^k = 1 | \mathbf{x}_i^k, \mathbf{w}_k\big)\\
&\quad = \frac{c}{1 + \exp\big(-s_i^k \mathbf{w}_k{}^\top \mathbf{x}_i\big)} + \frac{(1 - s_i^k)(1 - c)}{2}
\end{aligned}$$

Thus, we can get

$$\begin{aligned}
\mathbf{w}_k{}^* &= \arg\max_{\mathbf{w}_k} \sum_{i=1}^n \log \Pr\Big(s_i^k | \mathbf{x}_i, \mathbf{w}_k\Big)\\
&= \arg\max_{\mathbf{w}_k} \sum_{i=1}^n \log \left(\frac{c}{1 + \exp(-\mathbf{w}_k{}^\top \mathbf{x}_i)}\right)\\
&= \arg\max_{\mathbf{w}_k} \sum_{i=1}^n \log \left(\frac{c}{1 + \exp\big(-s_i^k \mathbf{w}_k{}^\top \mathbf{x}_i\big)} + \frac{(1 - s_i^k)(1 - c)}{2}\right)
\end{aligned}$$

In order to scale to large-scale problems, we use stochastic gradient descent to solve the above logistic regression problem efficiently. Different from conventional stochastic gradient descent approaches, which assume all the labels are availabel, we consider the incomplete label assignments and define the loss function as follows:

$$l(\mathbf{w}_k, \mathcal{D}) = -\sum_{i=1}^n \log \left(\frac{c}{1 + \exp\big(-s_i^k \mathbf{w}_k{}^\top \mathbf{x}_i\big)} + \frac{(1 - s_i^k)(1 - c)}{2}\right)$$

2.2 Handling Label Correlations via Stacked Graphical Models In the previous subsection, we discussed how can we deal with incomplete label assignments in multi-label learning. Now we show how can we use the previous model to further consider label correlations.

Inspired by the collective classification methods in [15, 7] based on stacking, we proposed a multi-label learning method called Mpu. Mpu can consider the label correlations effectively using stacked graphical model, which does not rely on joint inference for all labels. Stacking [23] is one type of ensemble methods which build a chain of models. Each model in the stacking uses the outputs of previous models as the inputs.

```
Given a training set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{s}_i)\}_{i=1}^n$ and a PU
learning algorithm $A$. Construct a corss-validation
prediction $\hat{\mathbf{y}}_i$ for each instance $\mathbf{x}_i$ as follows:
1. Convert $\mathcal{D}$ into $\{\mathcal{D}_1, \cdots, \mathcal{D}_q\}$, $\mathcal{D}_k = \{(\mathbf{x}_i, s_i^k)\}_{i=1}^n$
2. $\forall$ $k$, partition $\mathcal{D}^k$ into $m$ disjoint subsets with
     equal-size, denoted as $\mathcal{D}_k^1, \cdots, \mathcal{D}_k^m$
3. $\forall j = 1, \cdots, m, k = 1, \cdots, q$,
     train one model $f_k^j = A\Big(\mathcal{D}_k - \mathcal{D}_k^j\Big)$;
     $\forall$ $\mathbf{x}_i \in \mathcal{D}_k^j$, $\hat{y}_i^k = f_k^j(\mathbf{x}_i)$;
4. $\forall i = 1, \cdots, n$, we have $\hat{\mathbf{y}}_i = \left(\hat{y}_i^1, \cdots, \hat{y}_i^q\right)^\top$
     Return $\{\hat{\mathbf{y}}_i\}_{i=1}^n$
```

Figure 4: Cross-validation to obtain predictions for training instances

A stacked multi-label model allows inferences about one label to influence inferences about other labels but uses a different mechanism than other approaches to multi-label ensemble [4]. Rather than using joint inference to propagate inferences among labels, the stacked model uses one base model to predict the class labels for each label and uses those inferred labels as input to another stacked model.

Learning: The learning algorithm to learn the stacked model is shown in Figure 3. First, we use the PU stochestic gradient descent method to learn a base model to predict the labels of the instances using the instance features. This base model is then used to infer labels for each of the instances. In order to avoid overfitting, or any bias from applying the base model on the same data from which it was trained, we use a cross-validation procedure (shown in Figure 4) to infer the estimated outputs of the base model as the inputs for the next stacked model.

Next, we then construct an extended feature set to learn a stacked model using both features and estimated outputs of previous models as the inputs. In this way, we can build many levels of base models in a stacked model, where each subsequent base model uses the estimated predictions of class labels from the previous levels. In most cases, one single level of stacking is sufficient for multi-label learning, that can consider complex label correlations.

Inference: During inference process, we take turns to apply the base models from different levels one by one. The outputs of the model in previous level is directly used as the inputs of the next level in the stacking. Then the base models in the last level produces the final predictions. Different from other multi-label ensemble methods that learn on true labels, which are not known at the inference time, we learn the stacked model on the inferred labels, where all input features are known at inference time. Such design can permit exact inference, while other ensemble methods require approximate inference techniques [4], such as classifier chains.

Table 2: Properties of the multi-label datasets.

	Datasets	
property	**RCV1 small**	**RCV1 large**
# instance	6,000	804,414
# label	101	101
# feature	47,236	47,236

3 Experiments

3.1 Data Collection In order to evaluate the performances of the proposed approach for multi-label classification with missing labels, we tested our algorithm on two datasets as summarized in Table 2.

1) **RCV1 Small (Topics Subset)**: The first dataset we used in this paper is a medium scale dataset, *i.e.*, RCV1v2 Topics subset[1], in order to test the performances of different multi-label learning methods on medium scale problems. The dataset consists of 6,000 news articles which are categorized into 101 classes. For each news article, the bag-of-words features are extracted resulting in 47,236 features.

2) **RCV1 Large (Topics)**: The second dataset we used is a large-scale dataset, *i.e.*, RCV1v2 Topic full set, in order to test the scalability of different multi-label learning methods on large-scale problems. The dataset consists of 804,414 news articles which are categorized into 101 classes. The same number of bag-of-words features are extracted as the previous small dataset.

3.2 Evaluation Metric In order to evaluate the performance of multi-label learning by the models, we follow previous works [9, 13, 17] by using *Micro-F1* as the performance measure. Suppose that a multi-label dataset is $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, which consists of n multi-label instances. $\mathbf{y}_i \in \{0, 1\}^q$ $(i = 1, \cdots, n)$. Let $h(\cdot)$ be a multi-label classifier, and its predicted label set for $\mathbf{x}_i$ is $h(\mathbf{x}_i)$.

Micro-F1 is the harmonic mean of *Micro-Precision* and *Micro-Recall*. The *Micro-Precision* is the Micro-

[1] `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multilabel.html`

Table 3: Summary of compared methods.

Method	Type of Classification	Properties Considered	Publication
M3L	Large-Scale Multi-Label Learning	① Label Correlations ② Large-Scale Data	[12]
Elkan08	Positive & Unlabeled Learning	③ Missing Label Assignments	[6]
Well	Multi-Label Learning with Missing Labels	① Label Correlations ③ Missing Label Assignments	[21]
Mpu	Large-Scale Multi-Label Learning with Missing Labels	① Label Correlations ② Large-Scale Data ③ Missing Label Assignments	This paper

average of precision over all labels . Similarly, *Micro-Recall* is the Micro-average of recall overal all possible labels.

$$\text{Micro-F1}(h, \mathcal{D}) = \frac{2 \times \sum_{i=1}^{n} \|h(\mathbf{x}_i) \cap \mathbf{y}_i\|_1}{\sum_{i=1}^{n} \|h(\mathbf{x}_i)\|_1 + \sum_{i=1}^{n} \|\mathbf{y}_i\|_1}$$

The larger the value of *Micro-F1*, the better the performance of multi-label classification model.

All experiments are conducted on machines with Intel Xeon$^{\text{TM}}$Quad-Core CPUs of 2.26 GHz and 36 GB RAM.

3.3 Comparative Methods In order to study the effectiveness of the proposed approach, we compare our method with different baseline methods, including a large scale multi-label learning method, a multi-label learning method with missing labels and a positive and an unlabeled learning method. The compared methods are summarized as follows:

- *Large-scale Multi-label Learning (M3L)* [12]: The first baseline method is a multi-label classification method for large scale problems. We compared the M3L methods with two different kernels, *i.e.*, Linear kernel (denoted as M3L Linear) and RBF kernel (denoted as M3L RBF), separately.

- *Positive and Unlabeled Learning (Elkan08)* [6]: The second baseline method is a PU learning method which handle the cases where some positive instances can be missed by the labeler. This method is originally designed for binary classification problems. We use the binary decomposition method to solve multi-label classification problems by training one model over each class [1].

- *Multi-label Learning with Missing Labels (*Well*)* [21]: we also compare with another baseline which are designed for multi-label learning with incomplete label assignments. The method can train a model on weakly labeled multi-label instances, and predict complete label sets on testing data. It can handle moderate-size datasets, but cannot scale to large-scale datasets. We use default parameter settings for this method.

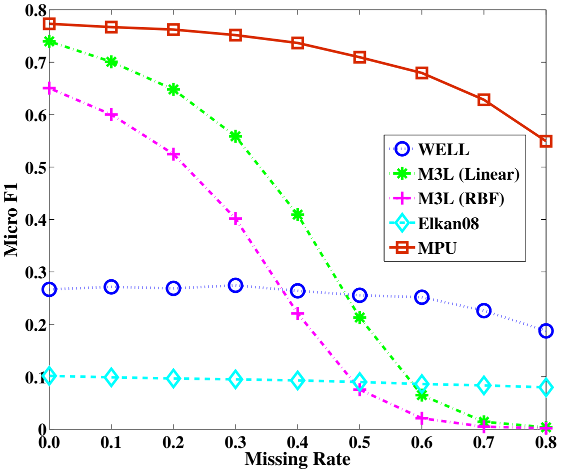

Figure 5: Classification performances on RCV1 small dataset (RCV1 topics subset).

- *Large-scale Multi-label Learning with Missing Labels (*Mpu*)*: the proposed method in this paper for large-scale multi-label learning with incomplete label assignments. Mpu can explicitly consider label correlations to facilitate multi-label learning process when a part of the ground-truth labels are missing in the label set. For simipicity, we set the number of stacking levels to the minium value 2.

3.4 Performances on Small Dataset We first study the effectiveness of the proposed Mpu method on multi-label classification with incomplete label assignment. In our experiment, 10-fold cross validation is performed on the small data set to evaluate the multi-label classification performances. In each round of the cross validation, the instances are partitioned into two groups: 9 folds are used as training data, the remaining fold is used as testing data. In order to simulate the incomplete label assignments, we randomly sample and remove a subset of the labels from each of label sets in training data according to a ratio, called *missing rate*. For example, if the missing rate is 20%, we randomly sample 20% of the labels from the ground-truth labels of training instances and remove them from the train set. The higher the missing rate, the more ground-truth labels are missed by the labeler in the training set. We report the average results of 10-fold cross validation on the dataset.

We study the performance of the proposed Mpu method on multi-label classification with different number of missing rates: 0%, 10%, 20%, *etc.* When the missing rate is 0%, it represents the setting of conventional multi-label learning where the training instances are fully labeled. In real-world multi-label learning, we can usually only have a small number of labels annotated in the label set.

The results of all compared methods are shown in Figure 5. Firstly, we can observe that when the training set is fully labeled (*i.e.*, missing rate is 0%), all multi-label learning methods outperform the single-label learning method, Elkan08. In general, these results support the importance of exploiting correlations among different class labels in multi-label learning problems. Because the method Elkan08 is based upon one-vs-all decomposition, where different labels are assumed to be independent, thus it cannot work well in multi-label learning tasks when different labels are correlated.

We also observe that when the missing rate increases, the performances of M3L decrease very quickly. While all the other methods that can consider missing labels are relatively stable in their performances. This is because M3L is designed for supervised multi-label learning problems, which assumes all the training instances are fully labeled. When a label l is missed in the label set of a training instance, M3L will consider the instance as a negative example for the label l. This result can support the importance of considering missing labels in the training data, by assuming that some labels in the ground-truth label set can be missed by the labeler. Thus, for each single class label, the 'negative' examples are no longer pure negative examples, but are mixture of both positive examples and negative examples.

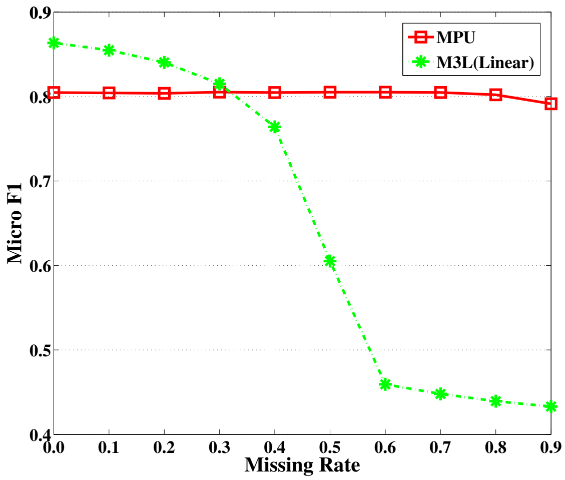

Figure 6: Classification performances on RCV1 large dataset (topics full set).

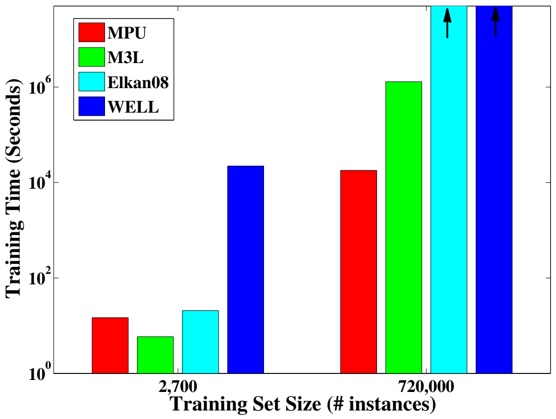

Figure 7: Running time performances.

We further observe that our proposed method Mpu outperforms all the other methods on all missing rates. Mpu can estimate the missing rate of the training data automatically, and can utilize label correlations to improve the classification performances. This result can support the importance of considering both label correlations and missing labels at the same time.

3.5 Performances on Large Dataset In our second experiment, we evaluate the efficiency of different multi-label learning methods on a large-scale dataset, which consists of around 800K instances. In this section, we compare both the Micro-F1 performances and running time in the training step. We reported the performances of Micro-F1 in Figure 6. We only showed the methods that can finish running within a week.

The methods show similar properties to those in the small datset. The M3L method initially has good performances when the training set is fully labeled. However when the missing rate increases, the performances of M3L drop very fast. The performances of Mpu method is quite stable. When the missing rate increases, the performances are still pretty good. Mpu can outperform M3L when the missing rate is greater than 30%.

We also show the training time of all methods in both dataset in Figure 7. We can observe that both Mpu and M3L can scale well to large-scale datasets. While the remaining methods such as Well cannot scale to datasets in this size. It is because the time complexity of Well is $O(n^2)$ in the number of training instances (n). Both Mpu and M3L are linear $O(n)$ in the number of instances.

4 Related Work

To the best of our knowledge, this paper is the first work addressing the problem of large-scale multi-label classification with incomplete label assignments. Our work is related to both multi-label learning techniques and positive and unlabeled learning. We briefly discuss both of them.

Multi-label learning corresponds to the classification problem where each instance can be assigned to multiple labels simultaneously [22, 10, 14, 8, 11]. The key challenge of multi-label learning is the large space of all label sets, *i.e.* the power set of all labels. In order to tackle this challenging task, many multi-label learning approaches focus on utilizing the labels correlations to facilitate the learning process. Conventional multi-label learning approaches can be roughly categorized as follows: (a) *one-vs-all* approaches: This type of approaches treat different labels as independent by converting the multi-label problem into multiple binary classification problems (one for each label) [1]. Zhang and Zhou[25] proposed Ml-knn, a binary method by extending the kNN algorithm to a multi-label problems using *maximum a posteriori* (MAP) principle to determine the label set predictions. (b) *pairwise* correlations: This type of approaches mainly use the pairwise relationships among different labels to facilitate the learning process [9]. Elisseeff and Weston [5] proposed Rank-svm, a kernel method by minimizing *ranking loss* to rank label pairs. (c) *High-order* correlations [24, 24]: This type of approaches can utilize higher order relationships among different labels. Examples include random subset ensemble approaches [19, 20], Bayesian network approach [24] and full-order approaches [3, 4].

The work in [12] studied the large-scale multi-label leanring problems, and proposed an efficient approach M3L. In addition, multi-label learning with incomplete label assignment has also been studied on small/moderate-size datasets [21, 2, 18]. The work in [21] proposed an approach Well to infer missing labels in multi-label learning under transductive settings.

Our work is also related to another line of research, called positive and unlabeled learning, or PU learning [16, 6]. PU learning corresponds to the binary classification problems where some positive samples can be mislabeled. Thus in the training set, only positive and unlabeled examples are available, no reliable negative examples are given in the training set. Many previous works on PU learning focus on estimating reliable negative examples from the unlabeled dataset, and utilize the estimated labels to improve the classification performances, The work in [16] proposed a method based upon biased SVM. Initially all the unlabeled instances are treated as negative examples. But the cost of classifying an unlabeled example with positive label is lower than that of predicting positive examples with negative labels. Elkan and Noto [6] provided a statistic model for positive and unlabeled learning. It is assumed that the ground-truth labels are randomly sampled into the training set according to a constant factor. The random sampling process is assumed to be independent from everything else, such as features of the instances. The constant factor can be estimated using cross-validation process on the training data. Then most conventional classification models can be modified according to the factor to consider the missing labels.

5 Conclusion

In this paper, we have described and studied the problem of large-scale multi-label learning with incomplete label assignments. We have studied two real-world datasets, one small and one large to evaluate the performance of our proposed method. Different from previous works in multi-label learning, we consider missing labels in the training set and label correlations simultaneously. By explicitly consider the missing labels using positive and unlabeled learning model, and label correlations using stacking models, our method can effectively boost the performance of multi-label classification with partially labeled training data.

6 Acknowledgements

This work is supported in part by NSF through grants CNS-1115234, DBI-0960443, and OISE-1129076, US Department of Army through grant W911NF-12-1-0066, and Huawei Grant.

References

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

[2] S. Bucak, R. Jin, and A. K. Jain. Multi-label learning with incomplete class assignments. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition*, pages 2801–2808, Colorado Springs, CO, 2011.

[3] W. Cheng and E. Hüllermeier. Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*, 76(2-3):211–225, 2009.

[4] K. Dembczyński, W. Cheng, and E. Hüllermeier. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proceedings of the 27th International Conference on Machine Learning*, pages 279–286, Haifa, Israel, 2010.

[5] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *NIPS*, pages 681–687. MIT Press, Cambridge, MA, 2002.

[6] C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data*, pages 213–220, Las Vegas, NV, 2008.

[7] A. Fast and D. Jensen. Why stacked models perform effective collective classification. In *Proceedings of the 8th IEEE International Conference on Data Mining*, Pisa, Italy, 2008.

[8] I. V. G. Tsoumakas. Random k-labelsets: An ensemble method for multilabel classification. In *Proceedings of 18th European Conference on Machine Learning*, pages 406–417, Warsaw, Poland, 2007.

[9] N. Ghamrawi and A. McCallum. Collective multi-label classification. In *Proceedings of ACM CIKM International Conference on Information and Knowledge Management*, pages 195–200, Bremen, Germany, 2005.

[10] S. Godbole and S. Sarawagi. Discriminative methods for multi-labeled classification. In *Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2004.

[11] Y. Guo and D. Schuurmans. Adaptive large margin training for multilabel classification. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, San Francisco, CA, 2011.

[12] B. hariharan, L. Zelnik-Manor, S. V. N. Vishwanathan, and M. Varma. Large scale max-margin multi-label classificaiton with priors. In *Proceedings of the 27th International Conference on Machine Learning*, pages 423–430, Haifa, Israel, 2010.

[13] F. Kang, R. Jin, and R. Sukthankar. Correlated label propagation with application to multi-label learning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1719–1726, New York, NY, 2006.

[14] H. Kazawa, T. Izumitani, H. Taira, and E. Maeda. Maximal margin labeling for multi-topic text categorization. In *NIPS*. 2005.

[15] Z. Kou and W. Cohen. Stacked graphical models for ecient inference in markov random fields. In *Proceedings of the 7th SIAM International Conference on Data Mining*, Minneapolis, MI, 2007.

[16] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu. Building text classifiers using positive and unlabeled examples. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, pages 19–22, Melbourne, FL, 2003.

[17] Y. Liu, R. Jin, and L. Yang. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *The 21st National Conference on Artificial Intelligence*, pages 421–426, Boston, MA, 2006.

[18] Z. Qi, M. Yang, Z. Zhang, and Z. Zhang. Mining partially annotated images. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, 2011.

[19] J. Read, B. Pfahringer, and G. Holmes. Multi-label classification using ensembles of pruned sets. In *Proceedings of the 8th IEEE International Conference on Data Mining*, pages 995–1000, Pisa, Italy, 2008.

[20] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. In *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 254–269, Bled, Slovenia, 2009.

[21] Y. Sun, Y. Zhang, and Z. Zhou. Multi-label learning with weak label. In *The 24th AAAI Conference on Artificial Intelligence*, pages 593–598, Atlanta, GA, 2010.

[22] N. Ueda and K. Saito. Parametric mixture models for multi-labeled text. In *NIPS*, pages 721–728. MIT Press, Cambridge, MA, 2003.

[23] D. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[24] M.-L. Zhang and K. Zhang. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 999–1008, Washington, DC, 2010.

[25] M.-L. Zhang and Z.-H. Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.

[26] M.-L. Zhang and Z.-H. Zhou. Multi-label learning by instance differentiation. In *Proceedings of the 21nd AAAI Conference on Artificial Intelligence*, pages 669–674, Vancouver, Canada, 2007.