Semi-supervised Multi-label Learning by Constrained Non-negative Matrix Factorization

Yi Liu, Rong Jin and **Liu Yang**
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824, U.S.A.
{liuyi3, rongjin, yangliu1}@cse.msu.edu

Abstract

We present a novel framework for multi-label learning that explicitly addresses the challenge arising from the large number of classes and a small size of training data. The key assumption behind this work is that two examples tend to have large overlap in their assigned class memberships if they share high similarity in their input patterns. We capitalize this assumption by first computing two sets of similarities, one based on the input patterns of examples, and the other based on the class memberships of the examples. We then search for the optimal assignment of class memberships to the unlabeled data that minimizes the difference between these two sets of similarities. The optimization problem is formulated as a constrained Non-negative Matrix Factorization (NMF) problem, and an algorithm is presented to efficiently find the solution. Compared to the existing approaches for multi-label learning, the proposed approach is advantageous in that it is able to explore both the unlabeled data and the correlation among different classes simultaneously. Experiments with text categorization show that our approach performs significantly better than several state-of-the-art classification techniques when the number of classes is large and the size of training data is small.

Introduction

Multi-label learning refers to the classification problems where each example can be assigned to multiple different classes. It has found applications in many real-world problems. For example, text categorization is typically multi-labeled since each document can be assigned to several predefined topics; in bioinformatics, most genes are associated with more than one functional classes (e.g., metabolism, transcription and protein synthesis); automatic image annotation, can also be treated as a multi-label learning problem if we view each annotation word as a distinct class. A straightforward approach toward multi-label learning is to decompose it into a set of binary classification problems, one for each class. The drawback with this approach is that it does not explore the correlation among different classes, which often could be an important hint for deciding the class memberships. Many algorithms have been developed to incorporate the class correlation into multi-label learning, including (McCallum 1999; Elisseeff & JasonWeston 2002; Jin & Ghahramani 2003; Ueda & Saito 2003; Boutella *et al.* 2004; Gao *et al.* 2004; Ghamrawi & McCallum 2005; Kazawa *et al.* 2005; Zhu *et al.* 2005; Yu, Yu, & Tresp 2005; Tsochantaridis *et al.* 2004; Taskar, Chatalbashev, & Koller 2004; Crammer & Singer 2002). But most of these studies are limited to a relatively small number of classes and assume that the amount of training data is sufficient for training reliable classifiers. In contrast, the real-world application of multi-label learning often features a large number of classes and a relatively small size of training data. As a result, the amount of training data related to each class is often sparse and insufficient for learning a reliable classifier. To address this problem, we present a novel framework for multi-label learning that explicitly explores the correlation among different classes. Compared to the existing approaches for multi-label learning that also explore the class correlation, the proposed framework provides a natural means for exploring the unlabeled data and the class correlation simultaneously, thus effective for the learning scenarios with a large number of classes and a small size of training data.

The key assumption behind this work is that two examples tend to have large overlap in their assigned class memberships if they share high similarity in their input patterns. To be more specific, consider two examples $\mathbf{x}_1$ and $\mathbf{x}_2$ that are labeled by two sets of class labels $\mathbf{y}_1$ and $\mathbf{y}_2$, respectively. We can evaluate the similarity between these two examples in two different ways. The first one is based on the correlation between the input patterns of these two examples. The second one is based on the overlap between the class labels of these two examples. We denote the similarity based on the input patterns by $K_x(\mathbf{x}_1, \mathbf{x}_2)$, and the similarity based on the class labels by $K_y(\mathbf{y}_1, \mathbf{y}_2)$. If the assigned class labels $\mathbf{y}_1$ and $\mathbf{y}_2$ are appropriate for example $\mathbf{x}_1$ and $\mathbf{x}_2$, we would expect the two similarity measurements to be similar, namely $K_x(\mathbf{x}_1, \mathbf{x}_2) \approx K_y(\mathbf{y}_1, \mathbf{y}_2)$. Based on this assumption, we can determine the best assignment of class labels to the unlabeled data by minimizing the difference between the two sets of similarities. Clearly, this approach is able to effectively explore the unlabeled data because the assignment of class labels to each unlabeled example is dependent on the assignment of class labels of other unlabeled examples. This approach is also able to exploit the class correlation effectively through the kernel similarity function $K_y(\mathbf{y}_1, \mathbf{y}_2)$.

The rest of the paper is structured as follows: first, we briefly review the related work on multi-label learning and semi-supervised learning; second, we introduce the proposed framework for multi-label learning, and a formalization based on the constrained non-negative matrix factorization; third, we present an efficient algorithm to solve the related optimization problem that is based on the iterative bound optimization algorithm; fourth, we present the empirical study with a text categorization problem; finally, we conclude this study.

Related Work

We will first review the related work on multi-label learning, followed by the discussion of semi-supervised learning.

The simplest approach toward multi-label learning is to divide it into a number of binary classification problems (Yang 1999; Joachims 1998). There are a number of disadvantages with this approach. One is that it will not scale to a large number of classes since a different binary classifier has to be built for each class. Another disadvantage is that it treats each class independently, and therefore is unable to explore the correlation among different classes. The third disadvantage is that this approach often will suffer from the unbalanced data problem when the minority classes are given only a few training examples.

Another group of approaches toward multi-label learning is label ranking (Crammer & Singer 2002; Elisseeff & JasonWeston 2002; Schapire & Singer 2000). Instead of learning binary classifiers from labeled examples, these approaches learn a ranking function from the labeled examples that order class labels for a given test example according to their relevance to the example. Compared to the binary classification approaches, the label ranking approaches are advantageous in dealing with large numbers of classes because only a single ranking function is learned. However, similar to the binary classification approaches, the label ranking approaches are usually unable to exploit the class correlation information.

In the past, a number of studies have been devoted to exploring the class correlation within the context of multi-label learning. A generative model for multi-label learning was proposed in (Ueda & Saito 2003) to explicitly incorporate the pairwise correlation between any two class labels. A maximum entropy model is proposed in (Zhu *et al.* 2005) that capture the pairwise class correlation by constraints. Approaches based on latent variables were proposed in (McCallum 1999; Yu, Yu, & Tresp 2005) to capture the correlation among different classes. The study in (Rousu *et al.* 2004) exploited the class correlation information given the hierarchical structure of classes. Unlike the previous work on multi-label learning that only considers the correlation among different classes, in this paper, we present a novel framework that exploits the unlabeled data as well as the class correlation. This property will make the proposed approach more effective than the existing approaches for multi-label learning, particularly when the number of classes is large and the size of training data is small.

This work is also related to semi-supervised learning, in particular the label propagation approaches for semi-supervised learning. This is because by enforcing examples with similar input patterns to share similar sets of class labels, we essentially propagate the class labels through the similarity graph of examples, which is the key idea of the label propagation approaches. A number of machine learning methods have been developed recently for label propagation, including the Gaussian processes (Williams 1998), the harmonic functions (Zhu & Ghahramani 2003), and Green functions (Zhou, Schölkopf, & Hofmann 2005). Unlike most of the previous work on semi-supervised learning that is designed primarily for multi-class learning, this work is specifically targeted on the semi-supervised multi-label learning. It effectively explores the class correlation information when utilizing the unlabeled data. More discussion of semi-supervised learning can be found in (Seeger 2001; Zhu 2006).

Semi-Supervised Multi-label Learning by Constrained NMF

The following terminology and notations will be used throughout the rest of the paper. Let $\mathcal{D} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ denote the entire dataset, where n is the total number of examples, including both the labeled ones and the unlabeled ones. We assume that the first n_l examples are labeled ones, and their label information is presented in the binary matrix $\bar{\mathbf{T}} \in \{0,1\}^{n_l \times m}$ where m is the number of classes. Let the similarity of all the examples denoted by a matrix $\mathbf{A} = [A_{i,j}]_{n\times n}$, where element $A_{i,j} \geq 0$ represents the similarity between two examples based on their input patterns. We denote by $T_{i,k} \geq 0$ the confidence score of assigning the k-th class label to the i-th example, and by $\mathbf{t}_i = (T_{i,1}, T_{i,2}, \ldots, T_{i,m})^\top$ the confidence scores of assigning each class to the i-th example. Finally, the matrix $\mathbf{T} = [T_{i,k}]_{n\times m}$ denotes the confidence scores of assigning every class label to all examples.

A Framework for Multi-label Learning

The key assumption behind this work is that two examples tend to be assigned similar sets of class labels if they share high similarity in the input patterns. In order to utilize this assumption for predicting class labels, we evaluate the similarity of two examples in different ways, one by their input patterns and the other by their assigned class memberships. We refer to the former as the *input-based similarity*, and the latter as the *class-based similarity*. Then, if the class labels assigned to the examples are consistent with their input patterns, we would expect the class-based similarities to be close to the input-based similarities. Since the input-based similarities are already given by the matrix $\mathbf{A}$, the key question is how to compute the similarity of two examples based on their class memberships. The simplest approach is to compute the class-based similarity between examples $\mathbf{x}_i$ and $\mathbf{x}_j$ by the overlap between their classmemberships, or $\mathbf{t}_i^\top \mathbf{t}_j$. The problem with this similarity measurement is that it treats all the classes independently and therefore is unable to explore the correlation among them. In particular, it will give zero similarity whenever two examples share no common classes. However, two examples with no common shared classes can still be strongly related if their assigned classes have close relationship (e.g. the children-parent relationship in the hierarchy of class labels).

To capture the correlation among different classes, we introduce matrix $\mathbf{B} = [B_{k,l}]_{m \times m}$ for the class similarities. Each element $B_{k,l} \geq 0$ represents the similarity between two classes. Then, instead of computing the class-based similarity between two examples by the direct dot product, we compute it by a weighted dot product, i.e., $\mathbf{t}_i^\top \mathbf{B} \mathbf{t}_j$. Then, following the assumption stated above, we would expect $A_{i,j} \approx \mathbf{t}_i^\top \mathbf{B} \mathbf{t}_j$ if the class assignments $\mathbf{t}_i$ and $\mathbf{t}_j$ are appropriate for examples $\mathbf{x}_i$ and $\mathbf{x}_j$. This leads to the following optimization problem:

$$\begin{aligned} \arg\min_{\mathbf{T}} \quad & \sum_{i,j=1}^{n} \left(A_{i,j} - \sum_{k,l=1}^{m} T_{i,k} B_{k,l} T_{j,l} \right)^2 && (1) \\ \text{s. t.} \quad & T_{j,l} \geq 0,\, j = 1, \ldots, n,\, l = 1, \ldots, m \\ & T_{i,k} = \bar{T}_{i,k},\, i = 1, \ldots, n_l,\, k = 1, \ldots, m && (2) \end{aligned}$$

where the last set of constraints is to ensure that the estimated label confidences $T_{i,k}$'s are consistent with the assigned class labels $\bar{T}_{i,k}$'s for all the training examples.

Remark: It is interesting to see that the formalization in (1) can also be written as a non-negative matrix factorization problem under a linear constraint, if we ignore the constraints coming from the training examples, i.e.,

$$\begin{aligned} \arg\min_{\mathbf{T},\mathbf{H}} \quad & \|\mathbf{A} - \mathbf{T}\mathbf{H}\|_F \\ \text{s. t.} \quad & T_{j,l}, H_{j,l} \geq 0,\, j = 1, \ldots, n,\, l = 1, \ldots, m \\ & \mathbf{H} = \mathbf{B}\mathbf{T}^\top \end{aligned}$$

where $\|\cdot\|_F$ stands for the Frobenius norm. The above problem is similar to the standard Non-negative Matrix Factorization (NMF) problem except for the linear constraint that restricts the matrix $\mathbf{H}$ to be linearly dependent on the matrix $\mathbf{T}$. It is this constraint and furthermore the constraints arising from the labeled data that prevent the direct application of the NMF algorithm.

One problem with the formulation in (1) is that since the input-based similarity $A_{i,k}$ can be any positive value, it could be significantly larger than the elements in $\mathbf{B}$. As a result, the label confidence $T_{i,k}$ that minimizes the objective function in (1) will also be significantly larger than 1. But, to satisfy the constraints in (2), the label confidence $T_{i,k}$ should be restricted to 0 or 1 since the assigned class label $\bar{T}_{i,k}$ is binary. To resolve the conflicts between the minimizer of the objective function in (1) and the binary constraints, we introduce two sets of label confidences: the unnormalized label confidence $\{T_{i,k}\}$, and the normalized label confidence $\{\hat{T}_{i,k}\}$. The former can take any positive value, while the latter is positive and subject to the constraints of $\sum_{k=1}^m \hat{T}_{i,k} = 1$. We will on one hand, use the unnormalized label confidence to minimize the difference between the class-based similarity and the input-based similarity, and on the other hand, use the normalized label confidence to ensure that the predicted label confidence is consistent with the assigned class labels. Formally, we can summarize this idea into the following optimization problem:

$$\begin{aligned} \arg\min_{\mathbf{T},\hat{\mathbf{T}},\alpha} \quad & \sum_{i,j=1}^{n} \left(A_{i,j} - \sum_{k,l=1}^{m} T_{i,k} B_{k,l} T_{j,l} \right)^2 && (3) \\ & + C \sum_{j=1}^{n} \sum_{l=1}^{m} (T_{j,l} - \alpha_j \hat{T}_{j,l})^2 \\ \text{s. t.} \quad & T_{j,l}, \hat{T}_{j,l}, \alpha_j \geq 0,\, j = 1, \ldots, n,\, l = 1, \ldots, m \\ & \sum_{l=1}^{m} \hat{T}_{j,l} = 1,\, i = 1, \ldots, m \\ & \hat{T}_{i,k} = \frac{\bar{T}_{i,k}}{\sum_{k=1}^m \bar{T}_{i,k}},\, i = 1, \ldots, n_l,\, k = 1, \ldots, m \end{aligned}$$

Note that in the above formalization, we introduce the term $C \sum_{j=1}^n \sum_{l=1}^m (T_{j,l} - \alpha_j \hat{T}_{j,l})^2$ into the objective function to enforce that the two sets of label confidences are consistent and only differ by a scaling factor α_j for each example. Parameter C weights the importance of the second term against the first term, and is determined empirically.

Solving the Constrained NMF

An alternative optimization approach is adopted to solve the constrained NMF. In particular, we will solve the optimization problem by alternatively fixing one set of label confidences and finding the optimal solution for another set of label confidences.

More specifically, we first fix the normalized label confidence matrix $\hat{\mathbf{T}}$ and the scaling factors α_j's, and search for the unnormalized label confidence $T_{i,j}$ that optimizes (3). To this end, we upper-bound the term $\left(A_{i,j} - \sum_{k,l=1}^m T_{i,k} B_{k,l} T_{j,l} \right)^2$ as follows

$$\begin{aligned} & \left(A_{i,j} - \sum_{k,l=1}^{m} T_{i,k} B_{k,l} T_{j,l} \right)^2 \\ & \leq \sum_{k,l=1}^{m} \frac{\tilde{T}_{i,k} B_{k,l} \tilde{T}_{j,l}}{[\tilde{\mathbf{T}}\mathbf{B}\tilde{\mathbf{T}}^\top]_{i,j}} \left(A_{i,j} - [\tilde{\mathbf{T}}\mathbf{B}\tilde{\mathbf{T}}^\top]_{i,j} \frac{T_{i,k} B_{k,l} T_{j,l}}{\tilde{T}_{i,k} B_{k,l} \tilde{T}_{j,l}} \right)^2 \\ & = A_{i,j}^2 + \sum_{k,l=1}^{m} \left(\frac{[\tilde{\mathbf{T}}\mathbf{B}\tilde{\mathbf{T}}^\top]_{i,j}}{\tilde{T}_{i,k} B_{k,l} \tilde{T}_{j,l}} T_{i,k}^2 B_{k,l}^2 T_{j,l}^2 - 2 A_{i,j} T_{i,k} B_{k,l} T_{j,l} \right) \\ & \leq A_{i,j}^2 + \frac{1}{2} \sum_{k,l=1}^{m} [\tilde{\mathbf{T}}\mathbf{B}\tilde{\mathbf{T}}^\top]_{i,j} \tilde{T}_{i,k} B_{k,l} \tilde{T}_{j,l} \left(\frac{T_{i,k}^4}{\tilde{T}_{i,k}^4} + \frac{T_{j,l}^4}{\tilde{T}_{j,l}^4} \right) \\ & \quad - 2 \sum_{k,l=1}^{m} A_{i,j} \tilde{T}_{i,k} B_{k,l} \tilde{T}_{j,l} \left(1 + \log T_{i,k} + \log T_{j,l} \right. \\ & \quad \left. - \log \tilde{T}_{i,k} - \log \tilde{T}_{j,l} \right) \end{aligned}$$

In the above, $\tilde{\mathbf{T}}$ refers to the matrix $\mathbf{T}$ from the last iteration. We use the convexity of the quadratic function in the first step of the derivation, and the concaveness of the logarithm function in the third step of the derivation. Then, we can upper-bound the first term in the function (3) as

$$\begin{aligned}
&\sum_{i,j=1}^{n}\left(A_{i,j} - \sum_{k,l=1}^{m} T_{i,k}B_{k,l}T_{j,l}\right)^2 \\
&\leq \sum_{i,j=1}^{n}\Bigg\{A_{i,j}^2 + \sum_{l=1}^{m}[\tilde{\mathbf{T}}\mathbf{B}\tilde{\mathbf{T}}^\top]_{i,j}[\tilde{\mathbf{T}}\mathbf{B}]_{i,l}\frac{T_{j,l}^4}{\tilde{T}_{j,l}^3} \\
&\qquad - 4\sum_{l=1}^{m} A_{i,j}[\tilde{\mathbf{T}}\mathbf{B}]_{i,l}\tilde{T}_{j,l}\log T_{j,l} - 2A_{i,j}[\tilde{\mathbf{T}}\mathbf{B}\tilde{\mathbf{T}}^\top]_{i,j} \\
&\qquad +4\sum_{k=1}^{m} A_{i,j}\tilde{T}_{i,k}[\mathbf{B}\tilde{\mathbf{T}}^\top]_{k,j}\log\tilde{T}_{i,k}\Bigg\}
\end{aligned}$$

Similarly, we can upper-bound the second term in (3) as follows:

$$\begin{aligned}
&C\sum_{j=1}^{n}\sum_{l=1}^{m}(T_{j,l} - \alpha_j\hat{T}_{j,l})^2 \\
&= C\sum_{j=1}^{n}\sum_{l=1}^{m}(T_{j,l}^2 - 2\alpha_j\hat{T}_{j,l}T_{j,l} + \alpha_j^2\hat{T}_{j,l}^2) \\
&\leq C\sum_{j=1}^{n}\sum_{l=1}^{m}\left[T_{j,l}^2 - 2\alpha_j\hat{T}_{j,l}\tilde{T}_{j,l}(\log\frac{T_{j,l}}{\tilde{T}_{j,l}} + 1) + \alpha_j^2\hat{T}_{j,l}^2\right]
\end{aligned}$$

By combining the above two bounds, we have the upper bound for the objective function in (3). Taking the derivative of the bounding function with respect to $T_{j,l}$, we have

$$\begin{aligned}
&4\sum_{i=1}^{n}[\tilde{\mathbf{T}}\mathbf{B}\tilde{\mathbf{T}}^\top]_{i,j}[\tilde{\mathbf{T}}\mathbf{B}]_{i,l}\frac{T_{j,l}^3}{\tilde{T}_{j,l}^3} - 4\sum_{i=1}^{n}A_{i,j}[\tilde{\mathbf{T}}\mathbf{B}]_{i,l}\tilde{T}_{j,l}\frac{1}{T_{j,l}} \\
&+ C(2T_{j,l} - 2\alpha_j\hat{T}_{j,l}\tilde{T}_{j,l}\frac{1}{T_{j,l}}) = 0
\end{aligned}$$

which leads to the following solution

$$T_{j,l} = \left[\frac{-C\tilde{T}_{j,l}^3 + \sqrt{C^2 + 8U_{j,l}\tilde{T}_{j,l}^4(2V_{j,l} + C\hat{T}_{j,l}\alpha_j)}}{4U_{j,l}}\right]^{\frac{1}{2}} \tag{4}$$

where $U_{j,l} = [\tilde{\mathbf{T}}\mathbf{B}\tilde{\mathbf{T}}^\top\tilde{\mathbf{T}}\mathbf{B}]_{j,l}$ and $V_{j,l} = [\mathbf{A}\tilde{\mathbf{T}}\mathbf{B}]_{j,l}$.

In the second step, we fix the unnormalized label confidence $\mathbf{T}_{i,k}$ and search for the normalized label confidence $\tilde{T}_{i,k}$ that optimizes the problem in (3), which leads to the following optimal solution:

$$\hat{T}_{j,l} = \frac{T_{j,l}}{\sum_{l=1}^{m} T_{j,l}},\ j = n_l + 1, \ldots, n, l = 1, \ldots, m \tag{5}$$

$$\alpha_j = \sum_{l=1}^{m} T_{j,l},\ j = n_l, \ldots, n \tag{6}$$

In summary, the iterative steps solving the optimization problem (3) could be formulated as the following algorithm

```
Step 1  Randomly initialize $\mathbf{T}$ and $\hat{\mathbf{T}}$ subject to
        the constraints in (3)
Step 2  Until convergence, do
        1. Fix all $\alpha_j$'s and $\hat{\mathbf{T}}$, update $\mathbf{T}$
           using Equation (4)
        2. Fix $\mathbf{T}$, update $\hat{\mathbf{T}}$ using Equation (5)
        3. Fix $\mathbf{T}$, update all $\alpha_j$'s using Equation (6)
```

Experiments and Discussions

Our experiments are designed to evaluate our proposed multi-label learning framework in text categorization tasks, particularly in the case of a large number of classes and a small size of training data.

Experiment Setup

The dataset used in our study comes from the textual data of *The Eurovision St Andrews Photographic Collection (ESTA)* in ImageCLEF collection (ImageCLEF 2003). We randomly pick 3456 documents, and choose the top 100 most popular ones from all the categories those picked documents belong to. On average, each document is assigned to 4.5 classes. Documents are preprocessed by the SMART system with stop words removed and words stemmed, and each document is represented by a vector of weighted term frequency.

Our proposed framework is implemented in the following way. The document similarity $A_{i,j}$ is computed as the cosine similarity between the corresponding term frequency vectors. To compute the class similarity matrix $\mathbf{B}$, we first represent each class c by a binary vector whose elements are set to be one when the corresponding training documents belong to the class c and zero otherwise. We then compute the pairwise class similarity based on their vector representation using a normalized RBF kernel. Finally, the class assignment for each test document is made by the ranking of the label confidence scores that are obtained from the matrix $\mathbf{T}$. Every experiment is repeated 10 times by randomly re-splitting the dataset into the training and the testing sets. The parameter C in the objective function (3) is set to 100. We also varied the value of C from 20 to 200, and found that the results remain almost unchanged. For an easy reference, we will refer to the proposed algorithm as "**CNMF**".

Since our approach only produces a ranked list of class labels for a test document, in this study, we focus on evaluating the quality of class ranking. In particular, for each test document, we compute the precision/recall and the $F1$ measurement at each rank by comparing the ranked classes to the true class labels. Then, the precision/recall and the $F1$ measurement averaged over all the test documents is used as the final evaluation metric.

Three baseline models are used in our study. The first one is Spectral Graph Transducer ("SGT" for short) (Joachims 2003), which has been proved effective for exploring unlabeled data. An separated SGT classifier is built for each individual document category, and the probability values output by SGT are used to rank the class labels. The second

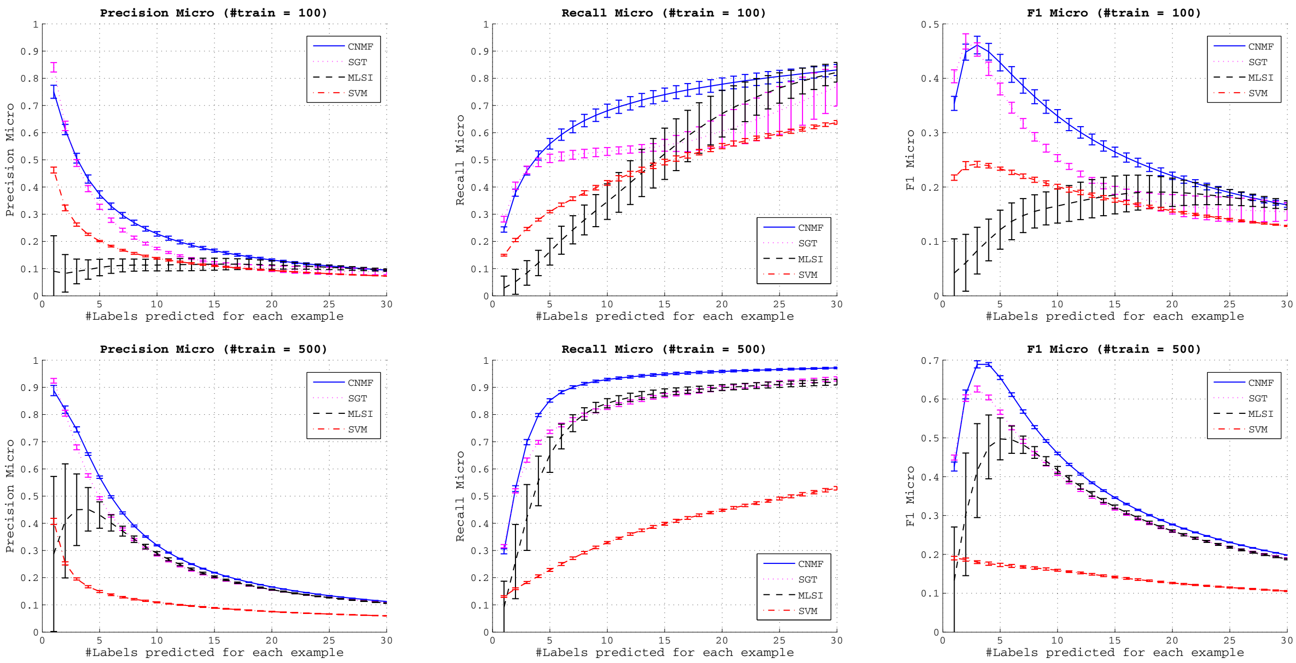

Figure 1: Classification performance when varying the number of predicted labels for each test example along the ranked list of class labels. From left to right, the three columns present evaluation figures based on the measurement of Precision Micro, Recall Micro and F1 Micro respectively. The upper panel is for the training set of 100 documents, and the lower panel is for the training set of 500 documents. Along the curves, we also plot the standard deviation.

baseline model is Multi-label Informed Latent Semantic Indexing ("MLSI" for short) (Yu, Yu, & Tresp 2005), which maps document vectors into a low-dimensional space that is strongly correlated with the class labels of the documents. It has been shown empirically that MLSI is effective for exploring both the unlabeled data and the correlation among classes. The last baseline model is Support Vector Machine ("SVM" for short). A linear SVM classifier based on the term frequency vectors of the documents is built for each category independently. All the baseline models are tested by a 10-fold experiment, using the same training/test split of the dataset as the proposed framework.

Experiment Results

Figure 1 shows the average precision, recall, and $F1$ measurement at different ranks, for both the proposed framework and the three baseline approaches. The upper panel of Figure 1 shows the results for 100 training documents, and the lower panel shows the results for 500 training documents.

A comparative analysis based on the results in Figure 1 lead to the following findings:

1. All four approaches show a same trend of decreasing precision and increasing recall, when the number of labels predicted for each document increases. This is in accordance with the usual precision-recall tradeoff. However, as a measurement balancing the precision and recall, each $F1$ curve clearly shows a peak. As can been seen from Figure (1), the $F1$ curve of **CNMF** reaches its climax when the number of predicted labels is around 3 to 4, which is close to the average number of labels per document (i.e., 4.5).

2. **CNMF** makes more significant improvement on the average recall than on the average precision when compared to the three baseline approaches. This is related to our task scenario, which focuses on multi-label learning with a large number of classes and a small size of training examples. Given such a scenario, we would expect a number of classes that are not provided with sufficient amount of training examples. As a result, we hypothesize that prediction on these classes will have to rely heavily on the correlation among classes. This hypothesis is partially justified by the comparison between the proposed approach, **CNMF**, that exploits class correlations, and SGT or SVM, which does not. Although **CNMF** and SGT achieve similar performance in precision, **CNMF** performs significantly better than the SGT in terms of the average recall.

3. More improvement by **CNMF** to the three baseline approaches is observed when the number of training documents is 100 than when the number of training documents is 500. This is partially due to the same reason mentioned above – the advantage of exploiting class correlations on sparse training data. It can also be attributed to the reason that our approach also makes use of the correlation among the unlabeled data, which has been proved by many studies in semi-supervised learning, for instance (Zhu & Ghahramani 2003; Seeger 2001;

Zhu 2006).

4. Although MLSI is able to explore the correlation among classes, its performance depends heavily on the appropriate choice of the number of latent variables and the tuning parameter determining how much the indexing should be biased by the outputs. These two parameters are usually determined by a cross validation approach and therefore could be problematic when the number of training examples is relatively small. This problem is directly reflected in the large variance in both precision and recall of the MLSI algorithm, which we believe it is due to the inappropriate choice of the aforementioned two parameters given the limited number of training examples.

Conclusions

In this paper, we propose a novel framework to meet the challenging multi-label learning problem when the number of classes is large and the size of training data is small. The advantage of our proposed framework is that it is able to exploit the correlation among classes and the unlabeled data. We also present an efficient algorithm to solve the related optimization problem. Experiments show that our proposed framework performs significantly better than the other three state-of-the-art multi-label learning techniques in text classification tasks.

References

Boutella, M. R.; Luob, J.; Shen, X.; and Browna, C. M. 2004. Learning multi-label scene classification. *Pattern Recognition* 37:1757–1771.

Crammer, K., and Singer, Y. 2002. A new family of online algorithms for category ranking. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in informaion retrieval*.

Elisseeff, A., and JasonWeston. 2002. A kernel method for multi-labelled classification. In Dietterich, T. G.; Becker, S.; and Ghahramani, Z., eds., *Advances in Neural Information Processing Systems 14*, 681–687. Cambridge, MA: MIT Press.

Gao, S.; Wu, W.; Lee, C.-H.; and Chua, T.-S. 2004. A MFoM learning approach to robust multiclass multi-label text categorization. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*.

Ghamrawi, N., and McCallum, A. 2005. Collective multi-label classification. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, 195–200. New York, NY, USA: ACM Press.

ImageCLEF. 2003. The CLEF Cross Language Image Retrieval Track (ImageCLEF), http://ir.shef.ac.uk/imageclef/.

Jin, R., and Ghahramani, Z. 2003. Learning with multiple labels. In S. Becker, S. T., and Obermayer, K., eds., *Advances in Neural Information Processing Systems 15*. Cambridge, MA: MIT Press. 897–904.

Joachims, T. 1998. Text categorization with suport vector machines: Learning with many relevant features. In *Proc European Conference on Machine Learning*.

Joachims, T. 2003. Transductive learning via spectral graph partitioning. In *Proceedings of the 20th International Conference on Machine Learning (ICML 2003)*.

Kazawa, H.; Izumitani, T.; Taira, H.; and Maeda, E. 2005. Maximal margin labeling for multi-topic text categorization. In Saul, L. K.; Weiss, Y.; and Bottou, L., eds., *Advances in Neural Information Processing Systems 17*. Cambridge, MA: MIT Press. 649–656.

McCallum, A. 1999. Multi-label text classification with a mixture model trained by EM. In *AAAI'99 Workshop on Text Learning*.

Rousu, J.; Saunders, C.; Szedmak, S.; and Shawe-Taylor, J. 2004. On maximum margin hierarchical multi-label classification. In *NIPS Workshop on Learning With Structured Outputs*.

Schapire, R. E., and Singer, Y. 2000. Boostexter: A boosting-based systemfor text categorization. *Machine Learning* 39(2-3).

Seeger, M. 2001. Learning with labeled and unlabeled data. Technical report, University of Edinburgh.

Taskar, B.; Chatalbashev, V.; and Koller, D. 2004. Learning associative markov networks. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, 102. New York, NY, USA: ACM Press.

Tsochantaridis, I.; Hofmann, T.; Joachims, T.; and Altun, Y. 2004. Support vector machine learning for interdependent and structured output spaces. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, 104. New York, NY, USA: ACM Press.

Ueda, N., and Saito, K. 2003. Parametric mixture models for multi-labeled text. In Saul, L. K.; Weiss, Y.; and Bottou, L., eds., *Advances in Neural Information Processing Systems 15*. Cambridge, MA: MIT Press. 649–656.

Williams, C. 1998. Computation with infinite neural networks. *Neural Computation* 10(5).

Yang, Y. 1999. An evaluation of statistical approaches to text categorization. *Information Retrieval* 1(1/2).

Yu, K.; Yu, S.; and Tresp, V. 2005. Multi-label informed latent semantic indexing. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in informaion retrieval*.

Zhou, D.; Schölkopf, B.; and Hofmann, T. 2005. Semi-supervised learning on directed graphs.

Zhu, X., and Ghahramani, Z. 2003. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning (ICML 2003)*.

Zhu, S.; Ji, X.; Xu, W.; and Gong, Y. 2005. Multi-labelled classification using maximum entropy method. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 274–281. New York, NY, USA: ACM Press.

Zhu, X. 2006. Semi-supervised learning literature survey. Technical Report TR 1530, Computer Sciences, University of Wisconsin - Madison.