Transductive Multi-Label Learning via Label Set Propagation

Xiangnan Kong, Michael K. Ng, and Zhi-Hua Zhou, *Senior Member, IEEE*

***Abstract*— The problem of multi-label classification has attracted great interest in the last decade, where each instance can be assigned with a set of multiple class labels simultaneously. It has a wide variety of real-world applications, *e.g.*, automatic image annotations and gene function analysis. Current research on multi-label classification focuses on supervised settings which assume existence of large amounts of labeled training data. However, in many applications, the labeling of multi-labeled data is extremely expensive and time-consuming, while there are often abundant unlabeled data available. In this paper, we study the problem of transductive multi-label learning and propose a novel solution, called TRAM, to effectively assign a set of multiple labels to each instance. Different from supervised multi-label learning methods, we estimate the label sets of the unlabeled instances effectively by utilizing the information from both labeled and unlabeled data. We first formulate the transductive multi-label learning as an optimization problem of estimating label concept compositions. Then we derive a closed-form solution to this optimization problem and propose an effective algorithm to assign label sets to the unlabeled instances. Empirical studies on several real-world multi-label learning tasks demonstrate that our TRAM method can effectively boost the performance of multi-label classification by using both labeled and unlabeled data.**

***Index Terms*— Data mining, machine learning, multi-label learning, transductive learning, semi-supervised learning, unlabeled data.**

I. INTRODUCTION

Conventional classification approaches assume that each instance is associated with only *one* class label within a number of candidate classes. However, many real-world applications often involve the scenario where each instance can be assigned with a set of *multiple* labels. For example, in image annotation, one image can be tagged with a set of multiple words, such as *urban*, *building* and *road*, indicating the contents of the image [6], [27]. In bioinformatics, one gene sequence can be associated with a set of multiple functions, such as *metabolism* and *protein synthesis* indicating the functions of the gene sequence within a cell's life circle [10]. In text categorization, one news article can cover multiple aspects of an event, thus being assigned with a set of multiple topics, such as *economics* and *politics* [24], [28]. An effective classification model for these real-world data should be able to adopt the multiple labels of each training example and predict a label set, instead of one single label, for each testing example. Motivated by these challenges, the problem of multi-label learning has received considerable attention in the last decade.

X. Kong and Z.-H. Zhou are with the National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China (e-mail: {kongxn, zhouzh}@lamda.nju.edu.cn).

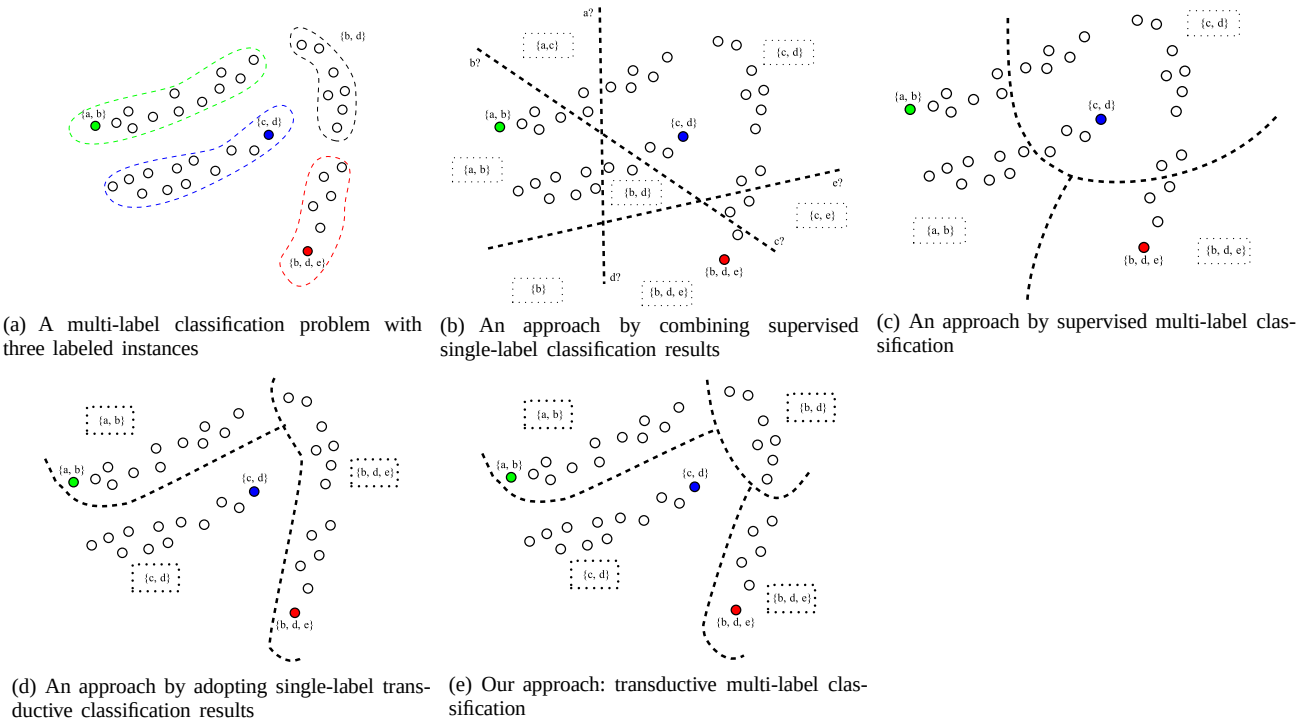
M. Ng is with the Center for Mathematical Imaging and Vision and Department of Mathematics, Hong Kong Baptist University, Hong Kong, China (e-mail: mng@math.hkbu.edu.hk).

In the literature, multi-label learning has been extensively studied [30]. Conventional approaches focus on supervised settings, which require a sufficiently large amount of labeled examples in order to train an accurate model. However, in many real world applications, the labeling process is extremely expensive and time-consuming, especially with multi-label data. Creating a large training dataset, where each example is labeled with a set of multiple labels within the candidate classes, is usually infeasible in practice. For example, in image annotation, human experts have to go through the entire list of all candidate words in order to decide the set of all possible tags for an image. It requires time, efforts and excessive resources to manually tag each image with all its labels, and hence only a limited amount of labeled images can be obtained in practice. Moreover, there are often copious amounts of unlabeled images available from various sources. Thus it is much desired that the large amount of unlabeled data can be effectively utilized together with the limited amount of labeled data to improve the multi-label classification performances. Transductive learning [32] is a type of approaches to exploit unlabeled data in classification processes. Transductive learning assumes all the testing data are available, and the goal is to achieve better performances on these testing data by exploiting the unlabeled testing data in the classification process. It has been shown useful in many single-label classification tasks [17], [32].

Formally, the transductive multi-label classification problem corresponds to predicting the label sets of a group of unlabeled instances simultaneously by utilizing the information from both labeled and unlabeled data. Transductive learning is particularly challenging in multi-label settings. The reason is that, in the single-label case, conventional transductive learning methods can be applied to propagate class labels among the unlabeled data and predict each unlabeled instance with the class label which has the highest confidence. But in multi-label cases, each instance contains multiple label concepts and the transductive classification task corresponds to finding a label set for each unlabeled instance within the space of label sets, *i.e.*, the *power set* of all labels. The number of possible label sets is exponential to the number of candidate labels, which is extremely large even with a small number of candidate labels.

If we consider the transductive learning and multi-label classification as a whole, the major research challenges on transductive multi-label classification can be summarized as follows:

1. *Lack of labeled data*: One fundamental problem in transductive multi-label classification lies in the labeling cost of the training data. Conventional multi-label classification approaches focuses on supervised settings [30]. The training of classification models strictly follows the assumption that there exists a large amount of labeled data. However, many real-world multi-label classification problems usually suffer from a lack of training data due to the labeling costs. Thus it is ineffective to only use the limited training data and



(a) A multi-label classification problem with three labeled instances

(b) An approach by combining supervised single-label classification results

(c) An approach by supervised multi-label classification

(d) An approach by adopting single-label transductive classification results

(e) Our approach: transductive multi-label classification

Fig. 1. An illustrative example for transductive multi-label classification problem

directly adopt existing multi-label classification approaches. For example, in Fig. 1, we show an illustrative example on multi-label classification. In Fig. 1(a), we have three labeled instances with a large number of unlabeled instances. Fig. 1(b) and Fig. 1(c) show that supervised classification methods, either based upon combining single-label methods or multi-label approaches, can only make use of the information from labeled instances to make predictions on the unlabeled data, where the predictions are not quite effective when the number of labeled data is small. To cope with this issue, it is deemed that the information within the unlabeled data should be exploited to facilitate multi-label classification.

2. *Multiple labels*: Another problem in transductive multi-label classification lies in the multiple labels of each instance. Conventional transductive learning approaches focus on single-labeled classification problems [7], [38], [41]. The classification strategy strictly follows the assumption that each instance has only one label. However in multi-label classification problem, each instance can be associated with a *set* of labels within the power set of all labels. Directly adopting conventional single-label transductive approaches may not be effective for multi-label classification. For example, in Fig. 1(d), we directly adopt a single-label transductive classification approach by treating each type of label set as a "class" (*i.e.*, we directly convent a multi-label classification problem to a single-label classification problem with three classes). Since we only have a limited number of labeled instances, not every ground-truth label set has a representative instance being labeled in the training set, *e.g.*, the label set {b, d}. Thus the trivial application of single-label transductive classification method will not be able to predict new label sets like {b, d} in the unlabeled data.

In this paper, we study the problem of transductive multi-label classification and propose a novel solution, called TRAM (TRAnsductive Multi-label classification), to effectively assign multiple labels to each instance using both labeled and unlabeled data. Different from supervised multi-label classification methods, we estimate the label sets of the unlabeled instances effectively by utilizing the information from both labeled and unlabeled data. We first formulate the transductive multi-label classification as an optimization problem of estimating label concept compositions. Then we derive a closed-form solution to this optimization problem and propose an effective algorithm to assign label sets to the unlabeled instances. Empirical studies on several real-world multi-label classification tasks demonstrate that our TRAM method can effectively boost the performance of multi-label classification by using both labeled and unlabeled data.

The rest of this paper is organized as follows. Section II gives a brief summary of related work on multi-label classification and transductive learning. In Section III we formulate transductive multi-label classification as an optimization problem, and then derive a closed-form solution. Section IV introduces label set prediction methods. Evaluation metrics used in multi-label classification are then briefly introduced and experiments of TRAM on real-world multi-label classification tasks are reported in Section VI. Finally, we give some concluding remarks in Section VII.

II. RELATED WORK

A. Multi-Label Classification

Multi-label classification deals with the problem where each example can belong to multiple different classes simultaneously. Traditional two-class and multi-class problems can both be cast

as special cases of multi-label classification problem. Thus multi-label problems are inevitably more difficult and complicated to solve than traditional single-label problems (i.e., two-class or multi-class problems). Until now, multi-label classification problem has been studied by a lot of researchers and many algorithms have been developed to solve different real-world application tasks, such as text categorization [8], [13], [20], [24], [28], [31], bioinformatics [10], [34], scene classification [4], image or video annotation [27].

Some multi-label learning algorithms are derived from traditional learning techniques. One famous approach proposed by Schapire and Singer, BOOSTEXTER [28], is extended from the popular ensemble learning method ADABOOST [11]. In the training phase, BOOSTEXTER maintains a set of weights over both training examples and their labels, which will be incrementally enlarged if examples or labels are hard to be predicted correctly. Elisseeff and Weston [10] presented a kernel method RANK-SVM for multi-label classification, by minimizing a loss function named *ranking loss*. Experimental results on the Yeast gene functional classification problem demonstrate its effectiveness. Zhang and Zhou [35] extended the lazy learning algorithm, kNN, to a multi-label version, ML-KNN. It employs label prior probabilities gained from each example's *k* nearest neighbors and use maximum *a posteriori* (MAP) principle to determine labels. Extension of other traditional learning techniques have also been studied, such as probabilistic generative models [24], [31], decision trees [8], neural networks [34], maximal margin methods [15], [20], maximum entropy methods [14], [40] and ensemble methods [12].

Unlike the previous works that only consider the correlations among different categories, Liu et al. [22] presents a semi-supervised multi-label classification method to exploit unlabeled data as well as category correlations. This approach is based on constrained non-negative matrix factorization. Generally, in comparison with supervised methods, semi-supervised methods can efficiently make use of the information provided by unlabeled instances. Zhou et al. [39] proposed the MIML framework which deals with multi-label examples each is represented as a set of instances. Sun et al. [29] employed hypergraph spectral learning to solve multi-label classification problems.

B. Transductive Learning

The use of unlabeled data has been increasingly popular these years in machine learning society. As in many practical learning problems, we usually need to handle situations when a small size of labeled data with a large amount of unlabeled data are available. The unlabeled data are usually much easier to obtain but quite expensive to identify their labels. Roughly speaking, there are three main paradigms of approaches to utilize unlabeled data [38], that is, semi-supervised learning, transductive learning and active learning. Semi-supervised learning approaches attempt to automatically exploit unlabeled data usually assuming the testing data are different from the unlabeled data; transductive learning approaches attempt to automatically exploit unlabeled data where the testing data are exactly the unlabeled data; active learning approaches query an *oracle* for the labels of specific instances in the input space, in order to get better models while minimizing the number of required queries.

In this paper, we focus on transductive learning. Transductive learning was proposed by Vladimir Vapnik [32] in the 1990's where all unlabeled points belong to the testing set. Many transductive learning approaches have been proposed. One famous approach is Transductive SVMs, introduced by [32] and applied to text classification by [17]. They exploit the structure in both training and testing data for better positioning the maximum margin hyperplane. Another type of approaches are graph-based methods, which define a graph with the nodes representing both labeled and unlabeled instances, and edges reflect the similarity of instances (*e.g.* [1], [37], [41]). Graph-based approaches usually assumes label smoothness over the graph. One example is to exploit the structure of the entire dataset in search for mincuts [3] or for min average cuts [18] on the graph.

III. PROBLEM FORMULATION

A. Transductive Multi-Label Classification

Before presenting the transductive multi-label classification model, we first introduce the notations that will be used throughout this paper. Let $\mathcal{D} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ denote the entire dataset, which consists of n instances ($\boldsymbol{x}_i \in \mathbb{R}^d$). The data set includes both labeled and unlabeled instances. Without loss of generality, we assume the first $n_l (n_l \ll n)$ instances within $\mathcal{D}$ are labeled by $\{Y_1, \cdots, Y_{n_l}\}$, where $Y_i \subseteq \mathcal{C}$ denotes the set of multiple labels assigned to $\boldsymbol{x}_i$. Here $\mathcal{C} = \{l_1, \cdots, l_m\}$ is the set of all possible label concepts. For convenience, we also denote $\mathcal{L} = \{1, \cdots, n_l\}$ as the index set for the labeled instances and $\mathcal{U} = \{n_l+1, \cdots, n\}$ for the unlabeled instances ($n = n_l + n_u$). The multi-label classification task corresponds to finding an optimal label set Y_i for each unlabeled instance $\boldsymbol{x}_i$ in the space of label sets $\mathcal{P}(\mathcal{C})$, *i.e.* the power set of $\mathcal{C}$.

As reviewed in Section I, previous approaches in multi-label classification are focused on supervised settings. In this paper, we address the multi-label classification problem under the transductive setting. Our goal is to find a simple and efficient way to improve the performance of multi-label classification by exploiting both labeled and unlabeled data.

The key issue of transductive multi-label classification is how to predict *a set* of multiple labels for each unlabeled instance based on a limited number of labeled examples and a large number of unlabeled examples, which is a non-trivial task due to the following problems:

(P1) How to properly estimate the composition of label concepts within the label set of an unlabeled instance based upon information from both labeled instances and all the other unlabeled instances? Intuitively, all the unlabeled instances should be estimated simultaneously and similar instances should contain similar label concepts in their label set. The question is how to jointly and effectively estimate the compostion of label concepts on each instance within the unlabeled dataset.

(P2) How to predict the label set for each unlabeled instance based on the estimated label concept composition with only a limited number of training examples? Some types of the label sets may not even have any representative labeled data in the training set. The question is how to predict new label sets based upon only limited examples of label sets in the training dataset.

In the following sections, we will introduce the optimization framework for transductive multi-label classification. Then we will derive our closed form solution to the optimization problem and propose an effective algorithm to predict multiple labels for each unlabeled instance.

B. Basic Idea

We address problem (P1) discussed as in Section III-A by defining transductive multi-label classification as an optimization problem of estimating the label composition for each unlabeled instance. Our target is to first effectively estimate the label concept composition for each unlabeled instance and then make the multi-label predictions based upon the estimated concept compositions. Here we define the *label concept composition* for a multi-label instance as follows: Suppose we have a multi-label instance $\boldsymbol{x}_i$, and its label set Y_i contains a set of multiple label concepts. For example, if we have a text document with 20% of the paragraphs writing about the label concept "politics" (l_1), 50% of the paragraphs writing about "economics" (l_2) and the rest about "culture" (l_3). Now we can say the label set for $\boldsymbol{x}_i$ is $\{l_1, l_2, l_3\}$ and the label concept composition is $(\ l_1 : 0.2,\ l_2 : 0.5,\ l_3 : 0.3,\ l_4 : 0, \cdots, l_m : 0)$. Here the label concept composition means that in the text document, only 20% of the paragraphs were writing about concept l_1. Of course this is just an extreme example, since in most cases there is no clear 'fraction' of the instance belonging to different labels. Indeed, the label concept composition expresses the typicality of the belongingness of the example to the labels, or the probability for the example to have different labels.

Formally, we denote the concept composition for instance $\boldsymbol{x}_i$ as $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{im})^\top$, where α_{ij} represents the fraction of label concept l_j in instance $\boldsymbol{x}_i$. Here we assume $\alpha_{ij} \geq 0$ and $\boldsymbol{\alpha}_i^\top \mathbf{1} = 1$ $(\forall i)$. For convenience of representation, we denote $\boldsymbol{\alpha}_{(j)} = (\alpha_{1j}, \cdots, \alpha_{nj})^\top$ and illustrate our notations as follow:

$$\begin{matrix} \boldsymbol{\alpha}_{(1)} & \cdots & \boldsymbol{\alpha}_{(j)} & \cdots & \boldsymbol{\alpha}_{(m)} \\ \downarrow & & \downarrow & & \downarrow \end{matrix}$$
$$\begin{bmatrix} \alpha_{11} & \cdots & \alpha_{1j} & \cdots & \alpha_{1m} \\ \vdots & & \vdots & & \vdots \\ \alpha_{i1} & \cdots & \alpha_{ij} & \cdots & \alpha_{im} \\ \vdots & & \vdots & & \vdots \\ \alpha_{n1} & \cdots & \alpha_{nj} & \cdots & \alpha_{nm} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}_1^\top \\ \vdots \\ \boldsymbol{\alpha}_i^\top \\ \vdots \\ \boldsymbol{\alpha}_n^\top \end{bmatrix} \begin{matrix} \\ \\ \leftarrow \boldsymbol{x}_1 \\ \\ \leftarrow \boldsymbol{x}_i \\ \\ \leftarrow \boldsymbol{x}_n \end{matrix}$$

In multi-label classification problems, we only know the label set of each training instance. There is no concept composition information available explicitly. We can only assume that, in a labeled training instance, all label concepts in its label set have equal weights or importance for concept composition, *i.e.*, the ground-truth concept composition $\bar{\boldsymbol{\alpha}}_i = (\bar{\alpha}_{i1}, \cdots, \bar{\alpha}_{im})^\top$ for a labeled instance $\boldsymbol{x}_i$ is defined as follow:

$$\bar{\alpha}_{ij} = \begin{cases} \frac{1}{|Y_i|}, & \text{if } l_j \in Y_i, \\ 0, & \text{otherwise.} \end{cases} \qquad (i \in \mathcal{L})$$

And our target is to estimate the concept compositions of all the unlabeled instances based upon both labeled and unlabeled data.

We assume that the optimal estimation of concept compositions should have the following property: *smoothness*, *i.e.*, similar instances should have similar concept compositions within their label sets. If an unlabeled instance $\boldsymbol{x}_i$ is similar to a labeled instance $\boldsymbol{x}_j$, the $\boldsymbol{\alpha}_i$ should be similar to $\boldsymbol{\alpha}_j = \bar{\boldsymbol{\alpha}}_j$. Moreover, if two unlabeled instances are similar to each other, their concept compositions should also be similar. Thus it is deemed that we need the estimate the concept compositions for all the unlabeled instances jointly/simultaneously in order to find optimal solutions on all the unlabeled data.

C. Optimization

In order to characterize the relation between similar instances, we build a weighted neighborhood graph $G = (V, E)$ on both labeled and unlabeled instances. Each vertex corresponds to an instance $\boldsymbol{x}_i$, an edge is put between $\boldsymbol{x}_i$ and $\boldsymbol{x}_z$, iff $\boldsymbol{x}_i$ is among the k nearest neighbors of $\boldsymbol{x}_z$ or $\boldsymbol{x}_z$ is among the k nearest neighbors of $\boldsymbol{x}_i$.

In order to reduce computational cost of kNN search among labeled and unlabeled instances, we use kd-tree to efficiently search for approximate k nearest neighbors for each instance. Since kd-trees suffer seriously from the curse of dimensionality which will degenerate to linear search in high dimensions [33], in our work a multi-label dimensionality reduction approach (MDDM [36]) is used before using kd-tree to construct kNN graphs, which finds a linear subspace from the original features to maximize the dependence between the label information and the subspace.

After the kNN search, we define a sparse $n \times n$ matrix W indicating the similarities among neighboring instances:

$$W_{iz} = \begin{cases} \frac{1}{Z_i} \exp\left(-\frac{||\boldsymbol{x}_i - \boldsymbol{x}_z||^2}{2\sigma^2}\right), & \text{if } z \in \mathcal{N}_i, \\ 0, & \text{otherwise.} \end{cases} \qquad (1)$$

where $\mathcal{N}_i$ is the index set of i-th instance's k nearest neighbors. Typically, $||\cdot||$ refers to the Euclidean distance. And parameter σ is empirically estimated as the average distance between instances. $Z_i = \sum_{z \in \mathcal{N}_i} \exp\left(-\frac{||\boldsymbol{x}_i - \boldsymbol{x}_z||^2}{2\sigma^2}\right)$, thus $\sum_z W_{iz} = 1$ for instances.

Thus based on the *smoothness* assumption in the previous subsection, we propose the following general optimization framework to estimate the optimal alpha values for unlabeled instances:

$$\begin{aligned} \min_{\boldsymbol{\alpha}_{n_l+1}, \cdots, \boldsymbol{\alpha}_n} & \sum_{i \in \mathcal{U}} \sum_{j=1}^{m} \left(\alpha_{ij} - \sum_{z \in \mathcal{N}_i} W_{iz} \alpha_{zj} \right)^2 \\ s.t.\ & \begin{matrix} \alpha_{ij} \geq 0, & \sum_{j=1}^{m} \alpha_{ij} = 1 \\ \alpha_{ij} = \bar{\alpha}_{ij} & (\forall i \in \mathcal{L}) \end{matrix} \end{aligned} \qquad (2)$$

Here the $\bar{\alpha}_{ij}$ is defined as

$$\bar{\alpha}_{ij} = \begin{cases} \frac{1}{|Y_i|}, & \text{if } l_j \in Y_i, \\ 0, & \text{otherwise.} \end{cases} \qquad (i \in \mathcal{L})$$

The optimization objective is to minimize the weighted differences among the concept compositions of similar/neighboring instances. As for the labeled instances, the concept compositions are "known", and hence we put constraints $\alpha_{ij} = \bar{\alpha}_{ij}$ in the optimization. In an optimal solution to the above problem, it guarantees that the estimated concept compositions of any pair of instances, that are closely connected in the weighted neighborhood graph G, will be similar. Intuitively, the estimation process corresponds to the propagation of concept compositions among instances along the graph G.

To simplify the optimization, we have

$$\begin{aligned} & \sum_{i \in \mathcal{U}} \sum_{j=1}^{m} \left(\alpha_{ij} - \sum_{z \in \mathcal{N}_i} W_{iz} \alpha_{zj} \right)^2 \\ = & \sum_{j=1}^{m} \left\| D_u \left(\boldsymbol{\alpha}_{(j)} - W \boldsymbol{\alpha}_{(j)} \right) \right\|^2 \end{aligned}$$

where $D_u = \begin{pmatrix} 0 & 0 \\ 0 & I_u \end{pmatrix}_{(n\times n)}$, and the vector $\boldsymbol{\alpha}_{(j)} = (\alpha_{1j}, \cdots, \alpha_{nj})^\top = \begin{bmatrix} \boldsymbol{\alpha}_{\mathcal{L}j} \\ \boldsymbol{\alpha}_{\mathcal{U}j} \end{bmatrix}$. Then, the optimization problem in Eq. 2 can be simplified into matrix form as

$$\begin{aligned} &\min_{\boldsymbol{\alpha}_{(1)},\cdots,\boldsymbol{\alpha}_{(m)}} \sum_{j=1}^{m} \left\| D_u(I-W)\boldsymbol{\alpha}_{(j)} \right\|^2 \\ &s.t. \begin{cases} \boldsymbol{\alpha}_{(j)} \geq \mathbf{0}, \sum_{j=1}^{m} \boldsymbol{\alpha}_{(j)} = \mathbf{1} \\ \boldsymbol{\alpha}_{\mathcal{L}j} = \bar{\boldsymbol{\alpha}}_{\mathcal{L}j} \end{cases} \end{aligned} \tag{3}$$

D. A Closed-Form Solution

We note that the objective function and the constraints in Eq.3 are convex. Therefore a global minimizer exists [25]. Let $A = I - W$ in Eq.3. We partition the matrix A and $\boldsymbol{\alpha}_{(j)}$ vectors into blocks according to the labeled and unlabeled data,

$$A = \begin{bmatrix} A_{\mathcal{LL}} & A_{\mathcal{LU}} \\ A_{\mathcal{UL}} & A_{\mathcal{UU}} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\alpha}_{(j)} = \begin{bmatrix} \boldsymbol{\alpha}_{\mathcal{L}j} \\ \boldsymbol{\alpha}_{\mathcal{U}j} \end{bmatrix}, (j = 1, \cdots, m)$$

By ignoring the constraints $\boldsymbol{\alpha}_{(j)} \geq \mathbf{0}$, the Lagrange function for Eq. 3 becomes

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) =& \frac{1}{2} \sum_{j=1}^{m} \left\| D_u A \boldsymbol{\alpha}_{(j)} \right\|^2 \\ & - \boldsymbol{\beta}^\top (\sum_{j=1}^{m} \boldsymbol{\alpha}_{\mathcal{U}j} - \mathbf{1}) - \sum_{j=1}^{m} {\boldsymbol{\gamma}_j}^\top (\boldsymbol{\alpha}_{\mathcal{L}j} - \bar{\boldsymbol{\alpha}}_{\mathcal{L}j}) \end{aligned}$$

where $\boldsymbol{\beta} \geq \mathbf{0}$ and $\boldsymbol{\gamma}_j \geq \mathbf{0}$. The optimal condition for $\boldsymbol{\alpha}_{(j)}$ is

$$\frac{\partial L}{\partial \boldsymbol{\alpha}_{(j)}} = A^\top D_u^\top D_u A \boldsymbol{\alpha}_{(j)} - \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\beta} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\gamma}_j \\ \mathbf{0} \end{bmatrix} = \mathbf{0} \tag{4}$$

By summing over the optimal conditions in Eq.4 for all $\alpha_{(j)}$ $(j = 1, \cdots, m)$, we have

$$\sum_{j=1}^{m} (A^\top D_u^\top D_u A \boldsymbol{\alpha}_{(j)}) = \begin{bmatrix} \sum_{j=1}^{m} \boldsymbol{\gamma}_j \\ m\boldsymbol{\beta} \end{bmatrix}.$$

Then using the constraints $\sum_{j=1}^{m} \boldsymbol{\alpha}_{(j)} = \mathbf{1}$, we have

$$A^\top D_u^\top D_u A \mathbf{1} = \begin{bmatrix} \sum_{j=1}^{m} \boldsymbol{\gamma}_j \\ m\boldsymbol{\beta} \end{bmatrix}.$$

Notice that the $A\mathbf{1} = (I - W)\mathbf{1} = \mathbf{1} - W\mathbf{1} = \mathbf{0}$. So, the following equations can be derived, $\boldsymbol{\beta} = \mathbf{0}$, $\sum_{j=1}^{m} \boldsymbol{\gamma}_j = \mathbf{0}$ and then we substitute them into Eq. 4,

$$\begin{bmatrix} A_{\mathcal{UL}}^\top A_{\mathcal{UL}} & A_{\mathcal{UL}}^\top A_{\mathcal{UU}} \\ A_{\mathcal{UU}}^\top A_{\mathcal{UL}} & A_{\mathcal{UU}}^\top A_{\mathcal{UU}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_{\mathcal{L}j} \\ \boldsymbol{\alpha}_{\mathcal{U}j} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\gamma}_j \\ \mathbf{0} \end{bmatrix}.$$

Therefore we get

$$A_{\mathcal{UU}}^\top (A_{\mathcal{UL}} \boldsymbol{\alpha}_{\mathcal{L}}^j + A_{\mathcal{UU}} \boldsymbol{\alpha}_{\mathcal{U}}^j) = \mathbf{0} \tag{5}$$

Here $A_{\mathcal{UU}}^\top$ is guaranteed to be nonsingular for a connected graph [2]. By substituting the constraints $\boldsymbol{\alpha}_{\mathcal{L}j} = \bar{\boldsymbol{\alpha}}_{\mathcal{L}j}$ into Eq. 5, the optimal alpha values of unlabeled instances for class j i.e., $\boldsymbol{\alpha}_{\mathcal{U}j}$) can be calculated by the following linear equation:

$$A_{\mathcal{UU}} \boldsymbol{\alpha}_{\mathcal{U}j} = -A_{\mathcal{UL}} \bar{\boldsymbol{\alpha}}_{\mathcal{L}j} \tag{6}$$

which is a sparse, symmetric linear system. The number of equations equals to n_u and the number of nonzero entries is less than $(k+1) \times n_u$. Here, the solution $\boldsymbol{\alpha}_{\mathcal{U}j}$ is guaranteed to exist and be unique with values guaranteed to lie between 0 and 1. The proofs can be found in [25], we put them in the Appendix section to make the paper self-contained.

After the optimal alpha values are solved in Eq. 6, we will show how to use the optimal alpha values to predict a set of labels for each unlabeled instance in the following section.

IV. Label Set Prediction

In this section, we address Problem (P2) as discussed in Section III-A to predict a set of labels for each unlabeled instance based on the optimal alpha values. We propose a supervised version of label set prediction method, and a transductive version of label set prediction method. The differences between these two versions are as follows: (1) In the supervised version, we only make use of the labeled instances to learn a *threshold* function and directly predict a label set based upon the estimated alpha values. (2) In the transductive version, we make use of both labeled and unlabeled instances to estimate the *cardinality* of the label set for each unlabeled instance. After the label set cardinality is estimated, we sort all the labels based on instance's concept composition (*i.e.* the estimated alpha values), and predict the label set with the top ranked labels with the estimated label set cardinality.

A. Supervised Label Set Prediction via Linear Regression

In this subsection, we propose a supervised label set predicting mechanism based on the optimal alpha values on unlabeled instances. More precisely, a label set predicting function $f(\alpha(\boldsymbol{x}))$ is modeled by a linear function $f(\alpha(\boldsymbol{x})) = P\alpha(\boldsymbol{x})$, where $\alpha(\boldsymbol{x}) = (\alpha_1(\boldsymbol{x}), \cdots, \alpha_m(\boldsymbol{x}))$ is the m-dimensional vector of the optimal alpha values for unlabeled instance $\boldsymbol{x}$, and P is a $m \times m$ linear transformation matrix. The procedure used to learn the optimal linear transformation matrix P is described as follows:

We perform the leave-one-out process using Eq. 6 on the training set to calculate the estimated optimal alpha values on each training instance, denoted by $\hat{\alpha}_{ij}$'s. By combining $\hat{\alpha}_{ij}, (i \in \mathcal{L})$ into a vector, the estimated alpha outputs on every training instance can be solved by the following equation:

$$\hat{\boldsymbol{\alpha}}_{\mathcal{L}j} = (I - A_{\mathcal{LL}})\boldsymbol{\alpha}_{\mathcal{L}j} = W_{\mathcal{LL}} \boldsymbol{\alpha}_{\mathcal{L}j} \qquad (j = 1, \cdots, m) \tag{7}$$

Suppose the output vector for instance i is $\hat{\boldsymbol{\alpha}}_i = (\hat{\alpha}_{i1}, \hat{\alpha}_{i2}, \cdots \hat{\alpha}_{im})^\top$ $(i \in \mathcal{L})$. The ground-truth labels for instance i are known, i.e., $Y_i \subseteq \mathcal{C}$. Here for convenience of prediction, we denote the vector of ground-truth labels as $\tilde{\boldsymbol{y}}_i \in \{-1, 1\}^m$. Then, transformation matrix P can be calculated by minimizing the following sum-of-squares error function with a regular term,

$$P = \arg\min_{P} \sum_{i \in \mathcal{L}} \|\tilde{\boldsymbol{y}}_i - P\hat{\boldsymbol{\alpha}}_i\|_2^2 + \lambda \sum\nolimits_j \|P_{j\cdot}\|_2^2$$

where $P_{j\cdot}$ denotes the j-th row of matrix P. Then the solution is

$$P = \tilde{\boldsymbol{y}}_{\mathcal{L}} \hat{\boldsymbol{\alpha}}^\top (\hat{\boldsymbol{\alpha}} \hat{\boldsymbol{\alpha}}^\top + \lambda I)^{-1}. \tag{8}$$

Here λ is used to avoid the singularity of the linear system in (8). In practice, we set λ as a very small number (it is set to be 1×10^{-7} in the experiment). Then, with the linear transforms matrix P, we can predict label vector for unlabeled instances from their optimal alpha values by

$$\boldsymbol{y}_i = sign(P\boldsymbol{\alpha}_i) \quad (\forall i \in \mathcal{U}).$$

Where $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{im})^\top$. Then the outputted label set for the i-th instance is $Y_i = \{l_j : y_{ij} = 1\}$.

B. Transductive Label Set Prediction

In this subsection, we propose a transductive label set predicting method based on the optimal alpha values. Different from the supervised method in the previous subsection, the transductive label set prediction method can utilize information from both labeled and unlabeled data.

As we have already found the optimal alpha values for any unlabeled instance $\boldsymbol{x}_i$. A sorted list of all potential labels for $\boldsymbol{x}_i$ can be find by ranking all candidate labels using their alpha values in descending order. The larger the alpha value is, the more likely $\boldsymbol{x}_i$ will have the corresponding label. For example, suppose there are three class labels l_1, l_2, l_3, and the optimal alpha values $\boldsymbol{x}_i$ are $(\alpha_{i1} = 0.25, \alpha_{i2} = 0.4, \alpha_{i3} = 0.35)$. The sorted list for instance $\boldsymbol{x}_i$ is (l_2, l_3, l_1). Now the only problem is how to decide how many labels should be predicted into the label set of $\boldsymbol{x}_i$ using both labeled and unlabeled data. As long as the number of labels on instance $\boldsymbol{x}_i$ is decided, say θ_i, we can predict the top θ_i labels on the sorted list as the label set of instance $\boldsymbol{x}_i$.

Let θ_i denote the number of labels in the label set for instance $\boldsymbol{x}_i$. The θ_i values on the labeled instances are fixed according to the ground truth of their label sets, *i.e.* $\theta_i = |Y_i|$ $(i \in \mathcal{L})$. For unlabeled data, the number of labels (θ_i) should be a non-negative integer, here we can relax the $\theta_i \in \mathbb{R}$ and $\theta_i \geq 0$ $(i \in \mathcal{U})$. Then by using similar $smoothness$ assumption in Section III-B, we assume similar instances should have similar number of labels.

Then the optimal θ_i values can be solved by the following optimization problem:

$$\min_{\theta_1,\cdots,\theta_n} \sum_{i\in\mathcal{U}} \left(\theta_i - \sum_{z\in\mathcal{N}_i} W_{iz}\theta_z \right)^2$$
$$s.t.\ \ \theta_i = |Y_i| \ (\forall i \in \mathcal{L}) \tag{9}$$

Similar to the optimization problem in Section III-D, optimal solutions of the Eq. 9 can be found by solving the following linear equation:

$$A_{\mathcal{U}\mathcal{U}}\boldsymbol{\theta}_{\mathcal{U}} = -A_{\mathcal{U}\mathcal{L}}\boldsymbol{\theta}_{\mathcal{L}} \tag{10}$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)^\top = \begin{bmatrix} \boldsymbol{\theta}_{\mathcal{L}} \\ \boldsymbol{\theta}_{\mathcal{U}} \end{bmatrix}$. We can now use the optimal solutions (θ_i^*) on each unlabeled data to predict its label set. The number of labels for unlabeled instance $\boldsymbol{x}_i$ is predicted as the closest integer to θ_i^*.

The TRAM method is briefly summarized in Figure 2. Note the default label set prediction method in TRAM is the transductive version described in Section IV-B. The TRAM method using supervised version of label set prediction in Section IV-A is denoted as TRAM_S.

V. COMPUTATIONAL COMPLEXITY

In this section, we briefly analyze the computational complexity of TRAM as follows. Beyond the computational cost of MDDM dimensionality reduction $(O(m \cdot n))$ in the training step and the neighborhood graph searched by kd-tree $(O(n \log n))$ in the testing step, the alpha solutions and the label learning procedure of TRAM involve the following costs: In the worst case, the least squares solution of the linear systems in Eq.6 requires $O(n_u^3 + n_l \cdot n_u))$ operations when all data points are connected in a full graph (i.e., $k = n$). However, this cost can be significantly reduced using a k-nearest neighbor graph $(k \ll n)$ which leads directly to a sparse matrix $(A_{\mathcal{U}\mathcal{U}})$. Thus the linear systems are large, sparse and symmetric, many good solvers can be employed, *e.g.*, direct methods (*e.g.*, LU factorizations), or iterative solvers [16]. In practice, "the cost of computing the sparse LU factorization depends in a complicated way on the size of $A_{\mathcal{U}\mathcal{U}}$, the number of nonzero elements, its sparsity pattern, but is often dramatically smaller than the cost of a dense LU factorization. In many cases the cost grows approximately linearly with n_u, when n_u is large. This means that when $A_{\mathcal{U}\mathcal{U}}$ is sparse, we can solve $A_{\mathcal{U}\mathcal{U}}\boldsymbol{\alpha}_{\mathcal{U}j} = \boldsymbol{b}$ very efficiently, often with an order approximately n_u" [5].

For simplicity, we have used QR factorization designed for sparse matrix in MATLAB to compute the R factor very cheaply, which avoids the expensive computation of an explicit Q, details are described in [23]. Then for label learning procedure of TRAM, the computation of $\hat{\boldsymbol{\alpha}}_{\mathcal{L}}^j$ and transforms matrix P costs respectively $O(m \cdot n_l)$ and $O(n_l \cdot m + m^3)$.

The computational complexity of RANK-SVM [10] is currently of the order $O(m \cdot {n_l}^2)$ in each iteration for training. ML-KNN [35] as a lazy learning algorithm requires $(O(n_l^2 + n_l \cdot m))$ for training, and $O(n_l \cdot n_u + n_u \cdot m)$ for testing. BOOSTEXTER [28] requires $O(n_l \cdot m)$ for each iteration round in training with additional cost for the training of base learners. CNMF [22] as a transductive learning method requires $O(n^2)$ for similarity calculation between samples and $O(m \cdot n_u)$ in each iteration for testing.

$(Y_U, \boldsymbol{\alpha}_{\mathcal{U}})$ = TRAM(X, $Y_{\mathcal{L}}$)

Input:
$X : (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ encoding features of the whole data set
$Y_{\mathcal{L}}$: $(Y_1, \cdots, Y_l)$ encoding labels of training set
Process:
1 Construct kNN graph among instances.
2 Initialize the similarities on each edge as $W_{iz} = \exp(-\frac{||\boldsymbol{x}_i - \boldsymbol{x}_z||^2}{2\sigma^2})$ and normalize to $\sum_z W_{iz} = 1$;
3 Determine the $\boldsymbol{\alpha}_{\mathcal{U}}^j$ values for all unlabeled data by solving the linear system in Eq.6;
\# Supervised version:
4 Compute the label set prediction matrix P by solving Eq.8;
5 Predict label set for each unlabeled instance by $\boldsymbol{y}_i = sign(P\boldsymbol{\alpha}_i) \quad (\forall i \in \mathcal{U})$.
\# Transductive version:
4 Compute sorted label list on each unlabeled instance using optimal alpha values in Step 3;
5 Determine the optimal number of labels on each instance by solving the linear equation in Eq. 10.
Output:
$Y_{\mathcal{U}}$: the predicted labels for unlabeled instances.
$\boldsymbol{\alpha}_{\mathcal{U}}$: the alpha value outputs for unlabeled instances.

Fig. 2. The TRAM algorithm

VI. EXPERIMENTS

In this section, we show the performance of TRAM on several real-world multi-label classification tasks. Table I summarizes the characteristics of the data sets used. For comparison, we also compare with several general-purpose multi-label classification algorithms, including CNMF [22], BOOSTEXTER [28], RANK-SVM [10] and ML-KNN [35], which are applicable to various multi-label problems, and represent the state-of-the-art techniques in multi-label classification:

1. TRAM: The proposed algorithm TRAM, *i.e.* a transductive multi-label classification algorithm via label set propagation (implementation in MATLAB). For label set prediction step,

the default setting is using transductive version of label set prediction. TRAM with supervised version of label set prediction is also compared, denoted by TRAM_S;

2. CNMF: The CNMF [22] is a semi-supervised multi-label classification algorithm by constrained non-negative matrix factorization. The key assumption behind CNMF is that two instances tend to have large overlap in their assigned class memberships if they share high similarity in their input patterns. By minimizing the difference between inputs similarity with class label overlaps, CNMF can determine the labels of unlabeled data;
3. BOOSTEXTER: The BOOSTEXTER [28] (implementation in C) is a Boosting style multi-label ranking system, which has been shown with excellent performance in previous studies, especially on text categorization tasks;
4. RANK-SVM: The RANK-SVM [10] (implementation in MATLAB) is an SVM style multi-label classification algorithm which minimizes ranking loss directly and has also exhibited excellent performance in previous studies;
5. ML-KNN: The ML-KNN [35] (implementation in MATLAB) is a kNN style multi-label classification algorithm which often outperforms other existing multi-label algorithms.

Parameters are used in their default settings unless otherwise specified. For BOOSTEXTER[1], the number of boosting rounds is set to 500 because on all data sets studied in this paper, the performance of BOOSTEXTER will not significantly change after the specified boosting rounds; For RANK-SVM the best parameters reported in the literature [10] are used; For CNMF, the best parameters in [22] are used.

Our TRAM implementation is in MATLAB and the size of neighbors k is 10. Moreover, the influence of TRAM's parameters will be discussed in Section VI-G.

A. Evaluation Metrics

Multi-label classification systems require much more complicated evaluation criteria than traditional single-label systems. In this section we briefly summarize the criteria used for performance evaluation from various perspectives. Since our approach not only produces a ranked list of class labels, but also produces a predicted label set, in this paper we employ two sets of evaluation metrics to evaluate the performance of label ranking as well as the label set prediction. Adopting the same notations as used in Section III, for a test set $\mathcal{D}_{\mathcal{U}} = \{(\boldsymbol{x}_{l+1}, Y_{l+1}), \cdots, (\boldsymbol{x}_n, Y_n)\}$, the following multi-label evaluation criteria are used in this paper, which have been used in [10], [28], [34], [35].

Label Set Prediction Performances: The first set of evaluation criteria are concerning algorithm's performance on label set prediction for each instance. It is based on multi-label classifier's label set prediction function $h : \mathbb{R}^d \to \mathcal{P}(\mathcal{C})$, assume $h(\boldsymbol{x}_i)$ be the set of labels predicted by a multi-label classifier for instance $\boldsymbol{x}_i$.

1) *MicroF1*: evaluates both micro average of Precision and micro average of Recall with equal importance.

$$MicroF1 = \frac{2 \times \sum_{i \in \mathcal{U}} |h(\boldsymbol{x}) \cap Y_i|}{\sum_{i \in \mathcal{U}} |h(\boldsymbol{x})| + \sum_{i \in \mathcal{U}} |Y_i|}$$

The bigger the value, the better the performance. This criterion has been used in [19], [22].

[1]http://www.cs.princeton.edu/~schapire/boostexter.html

2) *Hamming loss*: evaluates how many times an instance-label pair is misclassified.

$$HammingLoss(h, \mathcal{D}_{\mathcal{U}}) = \frac{1}{|\mathcal{D}_{\mathcal{U}}|} \sum\nolimits_{i \in \mathcal{U}} \frac{1}{m} |h(\boldsymbol{x}_i) \Delta Y_i|$$

where Δ stands for the symmetric difference of two sets. The smaller the value, the better the performance.

Label Ranking Performances: The second group of evaluation criteria are concerning algorithm's label ranking performance for each instance, they are based on the real-valued output function $f : \mathbb{R}^d \times \mathcal{C} \to \mathbb{R}$ of each algorithm. For TRAM method, the optimal alpha values are used as the real-valued outputs.

3) *Ranking loss*: evaluates the average fraction of label pairs that are not correctly ordered.

$$RankLoss(f, \mathcal{D}_{\mathcal{U}}) = \frac{1}{|\mathcal{D}_{\mathcal{U}}|} \sum_{i \in \mathcal{U}} \frac{1}{|Y_i||\overline{Y_i}|} |\{(y_1, y_2) \in Y_i \times \overline{Y_i} | f(\boldsymbol{x}_i, y_1) \leq f(\boldsymbol{x}_i, y_2)\}|$$

Where the $\overline{Y_i}$ denotes the complementary set of Y_i in $\mathcal{C}$. The performance is perfect when $RankLoss(f) = 0$. The smaller the value, the better the performance.

4) *Average Precision*: evaluates the average fraction of labels ranked above a particular label $y \in Y_i$ which actually is in Y_i.

$$AvePrec(f, \mathcal{D}_{\mathcal{U}}) = \frac{1}{|\mathcal{D}_{\mathcal{U}}|} \sum_{i \in \mathcal{U}} \frac{1}{|Y_i|} \sum_{y \in Y_i} \frac{|\{y' \in Y_i | r_f(\boldsymbol{x}_i, y') \leq r_f(\boldsymbol{x}_i, y)\}|}{r_f(\boldsymbol{x}_i, y)}$$

The bigger the value, the better the performance.

Note that all the criteria evaluate the performance of multi-label classification systems from different aspects. Usually few algorithms could outperform another algorithm on all those criteria. In order to make our evaluation criteria more comprehensive, we will use the value of $1 - AvePrec$ and $1 - MicroF1$ to replace the original *Average Precision* and *MicroF1*. Thus under all evaluation criteria, smaller values are always indicating better performances.

B. Application to Automatic Image Annotation

We test the automatic image annotation task on Corel dataset used in [9]. The original data set contains $5,000$ images each was segmented into several regions and tagged with several words. The regions of similar features are clustered into 500 clusters, known as blobs [9]. Then, each image is represented by a binary vector of these 500 blobs. The average annotated words for each image is 3.5. We remove the words that occur less than 100 times, and obtain $4,800$ images and 43 annotation words.

This data set is partitioned randomly into labeled/unlabeled data sets according to certain ratios. In detail, we randomly draw from 1% to 9 % of the data as labeled training examples and randomly selection 50% of the data from the remaining as unlabeled examples. For instance, assuming the data set contains 4,800 examples and the label rate is 1%, we randomly draw 48 examples as labeled training examples; and $2,400$ examples from the remaining data set as unlabeled testing examples. Thirty runs of experiments are conducted under every label rate; in each run, algorithms are evaluated on random data set partitions. We also compared against the RANK-SVM algorithm [10], but on the Image Annotation dataset alone, the algorithm did not get good results.

TABLE I
SUMMARY OF EXPERIMENTAL DATA SETS

Task Studied	Data Set	# Instances	# Attributes	# Labels
Automatic Image Annotation	annotation	4,800	500	43
Gene Functional Analysis	yeast	2,417	103	14
Web Page Categorization	yahoo (11 subsets)	5,000	(462 $\sim$ 1,047)	(21 $\sim$ 40)
Text Categorization	RCV1-v2	6,000	662	54
Natural Scene Classification	scene	2,407	294	6

The results of multi-label classification on image annotation task are shown in Figure 3[2]. In label set prediction performances, TRAM with transductive version of label set prediction gets much better performances on MicroF1 than other algorithms including the supervised version of TRAM on label set prediction (i.e., TRAM$_S$). It is not strange that the classic multi-label classification methods such as ML-KNN could not work well in this setting since they were designed for supervised scenarios where there are lots of labeled training examples. When the number of labeled data is extremely small, the supervised version of TRAM becomes unstable in MicroF1 performance, since TRAM$_S$ only use labeled data to train the label set prediction function, and the supervised information in labeled data can be weak in these cases. Although TRAM$_S$ gets better performance in Hamming Loss than TRAM, this may be explained by the fact that Hamming Loss treats two types of misclassification errors (false alarm and missing prediction) equally, which is quite similar to the sum-of-squares error function in TRAM$_S$'s label set prediction step. In image annotation task each image usually has a small number of labels compared with the large number of classes. In other words, the label distribution on each class is quite imbalanced. Classification methods like TRAM$_S$ with better Hamming Loss and bad MicroF1 are biased to avoid predicting any label for each instance. TRAM$_S$ obtains bad Micro-Recall performance and good Micro-Precision performance. Since MicroF1 is treating both Micro-Precision and Micro-Recall equally, MicroF1 can better evaluated the label set prediction performances in this case.

On evaluation criteria concerning label ranking, i.e., ranking loss and average precision, TRAM's performances are better than other methods. TRAM can make use of both labeled and unlabeled data to get an optimal set of alpha values on each unlabeled instance, which may significantly help to improve the ranking performance especially when there are not sufficient but reasonable size of training data.

C. Application to Yeast Gene Functional Analysis

The task of the yeast gene functional analysis has been studied as a multi-label classification problem in many works (e.g., [10] and [26]). Following [10], we aim at predicting the functional classes in the gene of yeast *Saccharomyces cerevisiae*. These functional classes are structured into 4 levels of hierarchies[3]. As in [10], only top level hierarchy is considered. The whole data set has 2,417 instances of genes and 14 possible class labels. Each of the gene is represented by a 103-dimensional vector and the average number of class labels is 4.24 ± 1.57 for each instance.

The data set is partitioned randomly into labeled/unlabeled data sets according to certain ratios, the same setup as in the automatic image annotation task. Thirty runs of experiments are conducted under every label rate; in each run, algorithms are evaluated on random data set partitions and the average performance is recorded.

The results of multi-label classification on Yeast Gene Functional Analysis are shown in Figure 4. For label set prediction performances, TRAM gets better performances than the other methods on MicroF1, while getting comparable performances with other methods on Hamming Loss. For label ranking performances, TRAM outperforms the other methods on all evaluation criteria and all label rates.

D. Application to Automatic Web Page Categorization

The web page categorization task has been studied in [20], [31], [35]. In this experiment, our task is to classify web pages in a collection of eleven data subsets[4]. The web pages were collected from the "yahoo.com" domain, represented by the form of "*Bag-of-Words*", *i.e.* each dimension of the feature vector represents the number of times a word appearing in the web page. Each data subset corresponds to a top-level category (e.g. "Entertainment", "Education", etc.), which contains 2,000 web pages in the training set and 3,000 web pages in the test set. Each web page is assigned to several second-level categories and may belongs to multiple categories simultaneously.

The web page data subsets are briefly summarized in Table II. Details of these data subsets can also be found in [35]. Comparing with the data sets used in previous tasks, the number of instances and size of vocabulary size in these 11 data subsets are much larger. Furthermore, a larger percentage of instances (about $30\% \sim 40\%$) are assigned to multiple labels. Thus, the data subsets used in automatic web page categorization tasks are more difficult to learn from.

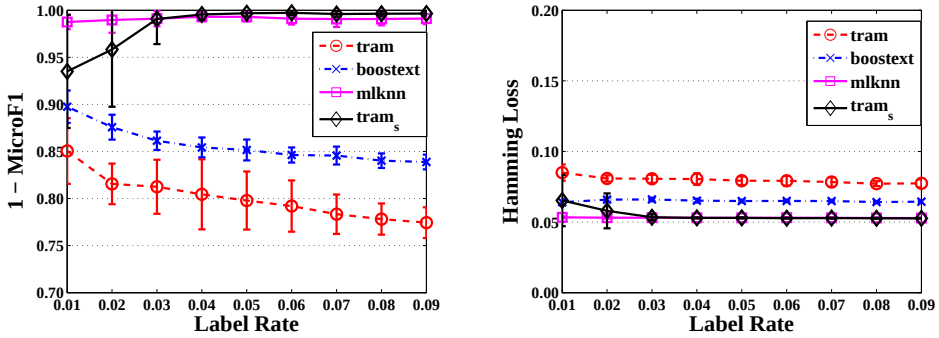
The same experiment settings are used to randomly partition the data subset into labeled/unlabeled sets according to different label rates. To make a more meaningful comparison among 11 data subsets, we used the geometrical means of the evaluation values across the 11 data subsets instead of simply using the average values. Such that, only the algorithms that have good performances over all 11 data subsets can have good performance values after the geometrical means.

The results of multi-label classification on automatic web page categorization task are shown in Figure 5. For label set prediction performances, TRAM has better MicroF1 results after the geometrical mean over 11 data subsets on this task, in other words, TRAM achieves better performances on average over 11

[2]Evaluation results of *Hamming Loss* and *MicroF1* are not available for CNMF.

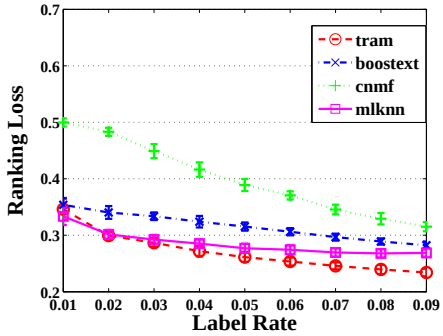
[3]Details in http://mips.gsf.de/proj/yeast/catalogues/funcat/.

[4]Data set available at http://www.kecl.ntt.co.jp/as/members/ueda/yahoo.tar.gz

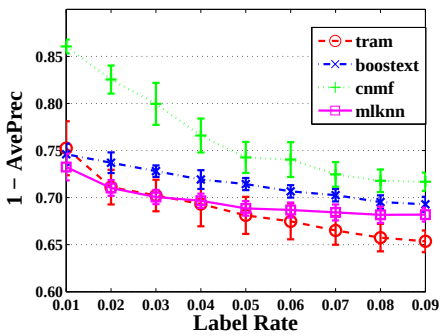


(a) 1- MicroF1.

(b) Hamming Loss.

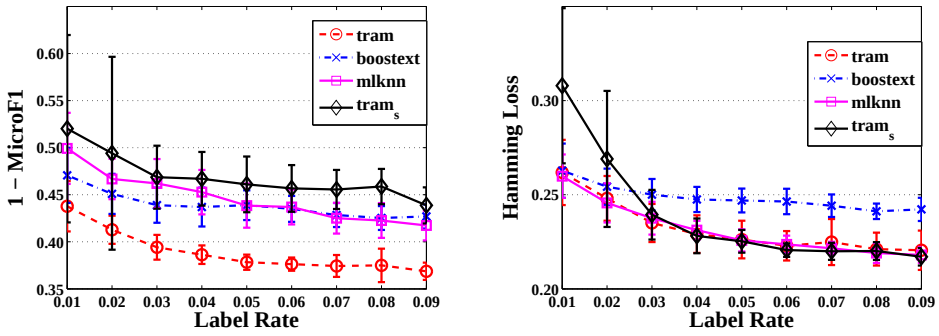


(c) Ranking Loss.



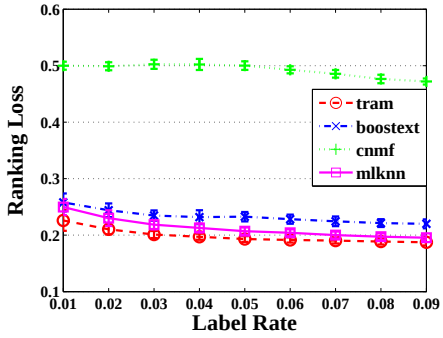
(d) 1 - AvePrec.

Fig. 3. Results on automatic image annotation task under different label rates. The lower the value, the better the performance. Along with the curves, we also plot the mean ± std on each point for different random data set partitions.

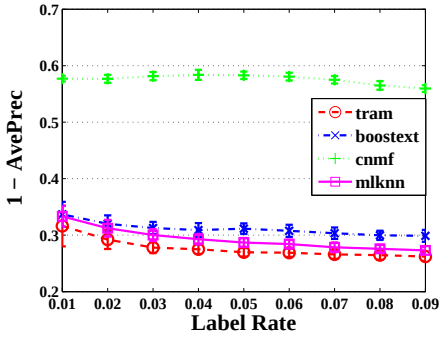


(a) 1- MicroF1.

(b) Hamming Loss.



(c) Ranking Loss.



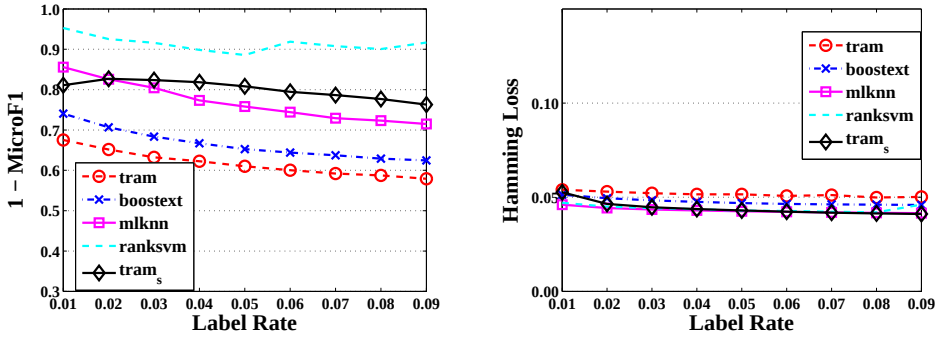
(d) 1 - AvePrec.

Fig. 4. Results on yeast gene function analysis task with different label rates. The lower the value, the better the performance. Along with the curves, we also plot the mean ± std on each point for different random data set partitions.

TABLE II

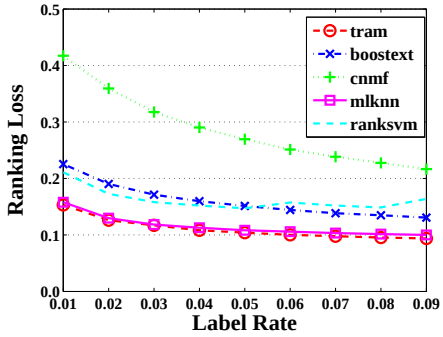
DATA SUBSETS USED IN THE AUTOMATIC WEB PAGE CATEGORIZATION TASK. "*MDoc%*" DENOTES THE PERCENTAGE OF WEB PAGES BELONGING MULTIPLE CATEGORIES, AND "*#AveLabel*" REPRESENTS THE AVERAGE NUMBER OF LABELS FOR EACH WEB PAGE.

	Number of	Vocabulary	Training Set		Test Set	
Data Subset	Labels	Size	*MDoc%*	*#AveLabel*	*MDoc%*	*#AveLabel*
Arts&Humanities	26	462	44.50%	1.627	43.63%	1.642
Business&Economy	30	438	42.20%	1.590	41.93%	1.586
Computers&Internet	33	681	29.60%	1.487	31.27%	1.522
Education	33	550	33.50%	1.465	33.73%	1.458
Entertainment	21	640	29.30%	1.426	28.20%	1.417
Health	32	612	48.05%	1.667	47.20%	1.659
Recreation&Sports	22	606	30.20%	1.414	31.20%	1.429
Reference	33	793	13.75%	1.159	14.60%	1.177
Science	40	743	34.85%	1.489	30.57%	1.425
Social&Science	39	1,047	20.95%	1.274	22.83%	1.290
Society&Culture	27	636	41.90%	1.705	39.97%	1.684

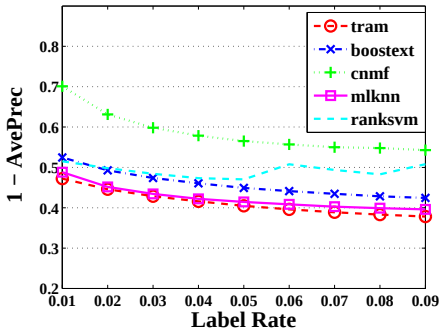


(a) 1- MicroF1.

(b) Hamming Loss.



(c) Ranking Loss.



(d) 1 - AvePrec.

Fig. 5. Results on automatic web page categorization task with different label rates. Note that the values in each figure are reported as the geometrical means across the 11 data subsets.

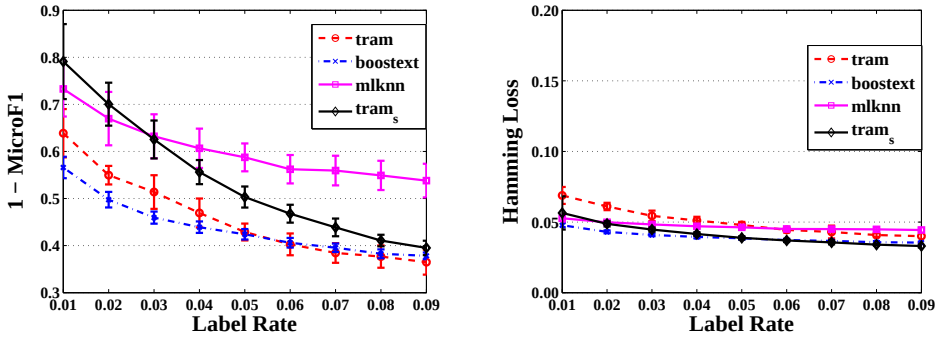
data subsets. On web page categorization task, the average number of labels on each webpage is much smaller than the number of classes. Thus, TRAM's performance on Hamming Loss is not as good as TRAM$_S$, but the difference is not quite significant. For label ranking performances, TRAM gets better or comparable performances than other methods after the geometrical mean on 11 data subsets.

E. Application to Text Categorization

In this Section, we perform text categorization using RCV1-v2 dataset [21]. The original data set has 804,414 documents, and 47,236 features. We use a benchmark subset, rcv1v2 (topics;subset)[5], which contains 6,000 documents. We removed the words that occur less than 200 times and topics with less than 50 positive examples, thus obtain 662 words and 54 topics. Note that the number of examples in this subset (6,000) is much larger than in the previous tasks in this paper. Here the dimensionality (662) is also very high.

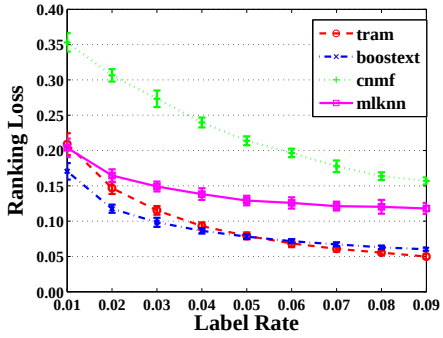
The results of multi-label classification on automatic text categorization task are reported in Figure 6. The performance of TRAM and BOOSTEXTER get best performances on label set prediction and label ranking. BOOSTEXTER is originally designed and one of the state-of-the-art multi-label classification methods on text data. Although on some label rates, BOOSTEXTER gets

[5]Data set available at http://www.csie.ntu.edu.tw/∼cjlin/libsvmtools/datasets/multi-label.html

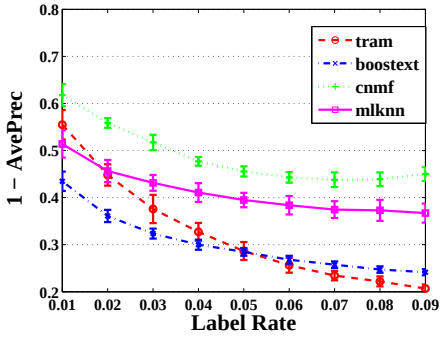


(a) 1- MicroF1.

(b) Hamming Loss.

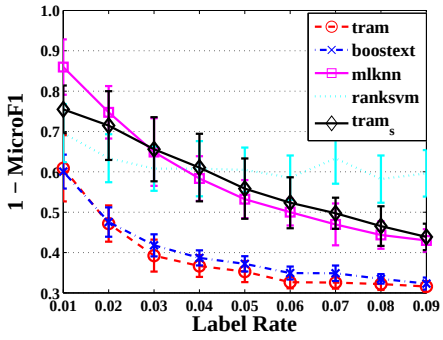


(c) Ranking Loss.

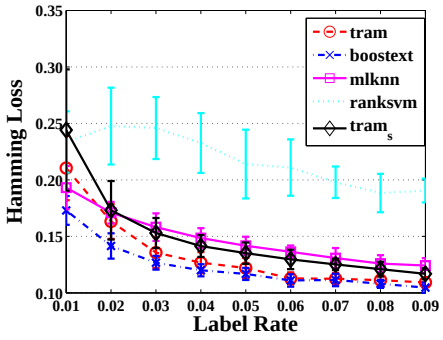


(d) 1 - AvePrec.

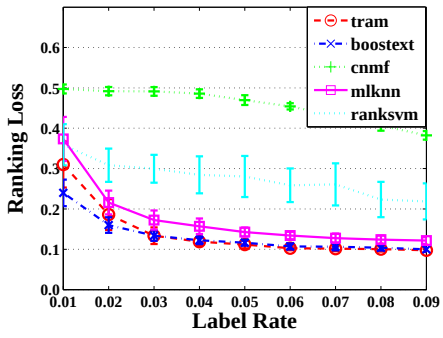
Fig. 6. Results on text categorization task under different label rates. The lower the value, the better the performance. Along with the curves, we also plot the mean $\pm$ std on each point for different random data set partitions.



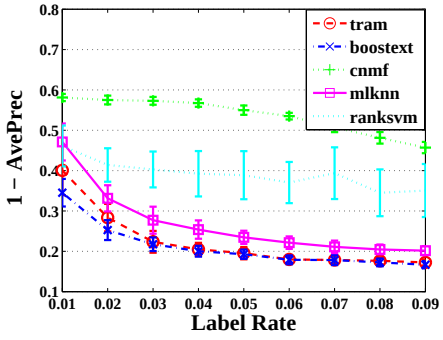
(a) 1- MicroF1.



(b) Hamming Loss.



(c) Ranking Loss.



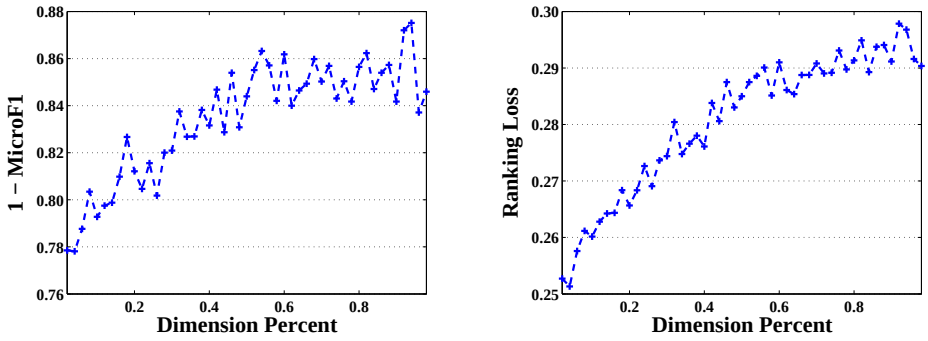
(d) 1 - AvePrec.

Fig. 7. Results on natural scene classification task with different label rates. The lower the value, the better the performance. Along with the curves, we also plot the mean $\pm$ std on each point for different random data set partitions.

TABLE III

RESULTS (MEAN±STD.) OF TRAM WITH DIFFERENT NUMBER OF NEAREST NEIGHBORS CONSIDERED IN THE INSTANCE GRAPH CONSTRUCTION STEP ON AUTOMATIC IMAGE ANNOTATION TASK ("↓" INDICATES "THE SMALLER THE BETTER", AND "↑" INDICATES "THE LARGER THE BETTER").

Evaluation	Number of Nearest Neighbors Considered				
Criterion	k=8	k=9	k=10	k=11	k=12
MicroF1 ↑	0.2075±0.0203	**0.2077±0.0215**	0.2066±0.0256	0.2049±0.0219	0.2031±0.0286
Hamming Loss ($\times 10^{-1}$)↓	**0.7860±0.0200**	**0.7860±0.0210**	0.787±0.025	0.788±0.021	0.791±0.028
Ranking Loss ↓	**0.2590±0.0080**	**0.2590±0.0080**	0.2601±0.0058	0.2604±0.0079	0.2605±0.0061
Average Precision ↑	**0.3240±0.0138**	0.3239±0.0146	0.3216±0.0206	0.3217±0.0141	0.3184±0.0225



(a) 1 - MicroF1. (b) Ranking Loss.

Fig. 8. Performances of TRAM with different percentages of dimensions in MDDM step on automatic image annotation task.

better performances than TRAM, but TRAM is still getting better performances than the other comparing methods on MicroF1, Ranking Loss and Average Precision.

F. Application to Natural Scene Classification

The last multi-label task studied in this paper is natural scene classification. The data set is relatively small, and consists of 2,400 natural scene images belonging to different classes, which is also used in [4]. Following [4], we convert each color image to the CIE Luv space, where the Euclidean distances closely correspond to the color differences perceived by human. Then the image is divided into 7×7 blocks using grids of equal width, and in each block the first and second moments of each color band are calculated, which is equal to resizing the image to a low-resolution and calculating simple texture features. Thus, each image is represented as a feature vector with $7 \times 7 \times 3 \times 2 = 294$-dimensions. The percentage of images that have multiple labels is over 22%. The same setting as in the previous experiments are used to randomly partition the data set into labeled/unlabeled sets according to different label rates.

The results of multi-label classification on natural scene classification task are reported in Figure 7. TRAM is among the most accurate methods on both label set prediction and label ranking. Since this data set is relatively small, the number of labeled data set is smaller than all the other tasks. The TRAM's performances are still stable as the labeled instances decrease to small label rates.

G. The Influence of Parameters

As observed in previous sections, when TRAM is used with the same parameters in all the multi-label tasks, it can all achieve satisfactory classification performances as accurate as the others. In this section, we analyze the influence of parameters in TRAM.

The first exploration is about the number of nearest neighbors during the instance graph construction. The experiment is based on automatic image annotation task. We randomly partition the dataset into labeled and unlabeled data with 5% label rate. The experiment result of TRAM is reported in Table III, when the number of nearest neighbor during the graph construction varies from 8 to 12. The value following "±" gives the standard deviation and the best result on each metric is shown in bold face. With respect to above configurations, Table III shows that the number of nearest neighbors used in graph construction step does not significantly affect TRAM's performance. Therefore, all the results of TRAM shown in this paper are obtained with the parameter k set to be the moderate value of 10.

Besides the number of nearest neighbor, another parameter is about the number of dimensions in the subspace used by MDDM. Note that due to the curse of the dimensionality, the similarities directly calculated based on distances between instances in the input space may be unreliable, especially when these similarities are the key parameters for the TRAM model. A simple, but often very effective, way of dealing with high-dimensional data is to reduce the number of dimensions, by finding a subspace from the input features that is most relevant to label information. Therefore, we need to utilize MDDM before the graph construction among instances. In order to verify this assumption, the results under different percentage of dimensions in the pre-process stage are reported in Figure 8. The experiment is based on automatic image annotation task, and results on other tasks are similar to the case in this task.

Figure 8 shows that on automatic image annotation task, the *MicroF1* and *Ranking Loss* of TRAM are significantly improved by introducing the dimensionality reduction (MDDM) before constructing the instance graph. TRAM's best performance are more likely to appear at the relatively low percentage of dimensions. Nonetheless, the number of dimensions does not have to be

pre-specified, which can automatically be determined by setting MDDM's threshold parameter thr as preserving 99.99% of the eigenvalues.

VII. Conclusion

In this paper, we propose TRAM, a transductive multi-label classification method by label set propagation. At first, we formulate the task as an optimization problem which is able to exploit unlabeled data to obtain an effective model for assigning appropriate multiple labels to instances. Then, we develop an efficient algorithm which has a closed-form solution for this optimization problem. Empirical studies on a broad range of real-world tasks demonstrate that our TRAM method can effectively boost the performance of multi-label classification by using unlabeled data in addition to labeled data.

Appendix

Here, we study the properties of the linear systems solutions for Eq.5 and Eq.6. For convenience of study, we combine the Eq.5 with the constrains for labeled data as:

$$A_{\mathcal{U}\mathcal{U}}\boldsymbol{\alpha}_{\mathcal{U}j} + A_{\mathcal{U}\mathcal{L}}\boldsymbol{\alpha}_{\mathcal{L}j} = \mathbf{0} \tag{11}$$

$$\boldsymbol{\alpha}_{\mathcal{L}j} = \bar{\boldsymbol{\alpha}}_{\mathcal{L}j} \tag{12}$$

which is equivalent to:

$$\tilde{A}\boldsymbol{\alpha}_{(j)} = \boldsymbol{b}_{(j)}, \quad j = 1, \cdots, m \tag{13}$$

where

$$\tilde{A} = \begin{bmatrix} A_{\mathcal{U}\mathcal{U}} & A_{\mathcal{U}\mathcal{L}} \\ 0 & I \end{bmatrix} \text{ and } \boldsymbol{b}_{(j)} = \begin{bmatrix} \mathbf{0} \\ \bar{\boldsymbol{\alpha}}_{\mathcal{L}j} \end{bmatrix}$$

Then we show that the solution of $\tilde{A}\boldsymbol{\alpha}_{(j)} = \boldsymbol{b}_{(j)}$ automatically satisfies the bilateral constrains $\mathbf{0} \le \boldsymbol{\alpha}_{(j)} \le \mathbf{1}$.

Instance i and z are connected by an edge if and only if they are a neighbor of each other, and W_{iz} and W_{zi} are both positive. Let $\alpha = (\alpha_i)$ be a discrete function defined on $\mathcal{U} \bigcup \mathcal{L}$, then the (*strong*) *discrete maximum principle* says that α can only attain its maximum in $\mathcal{L}$, unless α is constant in $\mathcal{U} \bigcup \mathcal{L}$. It is similar for the minimum principle. If there are more than one connected components in $\mathcal{U}$, we can apply the principle to each component independently. We also assume that each point in $\mathcal{L}$ is a neighbor of some instance in $\mathcal{U}$.

THEOREM *1:* The solution to $\tilde{A}\boldsymbol{\alpha} = \boldsymbol{b}$ satisfies the *discrete maximum principle*.

Proof: Suppose that the maximum of α can be attained at an interior point $i_0 \in \mathcal{U}$. Then the i_0-th equation of Eq.13 is $(\tilde{A}\boldsymbol{\alpha})_{i_0} = 0$ since $b_{i_0} = 0$. Notice that the i_0-th row of $\tilde{A}$ is the same as the i_0-th row of $A = I - W$. Therefore,

$$(\tilde{A}\boldsymbol{\alpha})_{i_0} = \alpha_{i_0} - \sum_{z \in \mathcal{N}_{i_0}} W_{i_0 z}\alpha_z = 0$$

or

$$\alpha_{i_0} = \sum_{z \in \mathcal{N}_{i_0}} W_{i_0 z}\alpha_z$$

Note that $W_{i_0 z} > 0$ for $z \in \mathcal{N}_{i_0}$ and $\sum_{z \in \mathcal{N}_{i_0}} W_{i_0 z} = 1$, which means the maximum value α_{i_0} equals a weighted average of $\{\alpha_z : z \in \mathcal{N}_{i_0}\}$, thus for all $z \in \mathcal{N}_{i_0}$, α_z is also the maximum. Similarly, since the domain $\mathcal{U}$ is connected, we can conclude that the values of α in $\mathcal{U}$ and the neighbor of $\mathcal{U}$ which covers $\mathcal{L}$ are all maximum. This shows that if α has an interior maximum, then α is constant in $\mathcal{U} \bigcup \mathcal{L}$. ■

COROLLARY *1:* The solution to $\tilde{A}\boldsymbol{\alpha} = \boldsymbol{b}$ satisfies the the bilateral constraints $\mathbf{0} \le \boldsymbol{\alpha} \le \mathbf{1}$, if $\{\alpha_i = 0 : i \in \mathcal{L}\}$ and $\{\alpha_i = 1 : i \in \mathcal{L}\}$ are non-empty sets.

Proof: According to maximum principle, $\alpha_z \le \max_{i \in \mathcal{L}} \alpha_i = 1$ for all $z \in \mathcal{U}$. Similarly, we have $\alpha \ge \min_{i \in \mathcal{L}} \alpha_i = 0$. Therefore, $0 \le \alpha_z \le 1$ for all $z \in \mathcal{U}$. ■

Acknowledgment

The authors wish to thank the editor and anonymous reviewers for their helpful comments and suggestions, and Yu-Feng Li, Jieping Ye, Yang Yu, De-Chuan Zhan and Yin Zhang for reading a draft of the paper. This work was supported by the National Fundamental Research Program of China (2010CB327900), the National Science Foundation of China (61073097), the Jiangsu Science Foundation (BK2008018), the Jiangsu 333 High-Level Talent Cultivation Program, the Hong Kong Baptist University Faculty Research Grants and the Hong Kong Research Grant Council (201508).

References

[1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: a geometric framework for learning from examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[2] N. Biggs. *Algebraic Graph Theory*. Cambridge University Press, Cambridge, UK, 1974.

[3] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the 18th International Conference on Machine Learning*, pages 19–26, Williamstown, MA, 2001.

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

[5] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, 2004.

[6] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):394–410, 2007.

[7] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[8] F. D. Comité, R. Gilleron, and M. Tommasi. Learning multi-label altenating decision tree from texts and data. In *Proceedings of the 3rd International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 35–49, Leipzig, Germany, 2003.

[9] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision*, pages 97–112, Copenhagen, Denmark, 2002.

[10] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 681–687. Cambridge, MA: MIT Press, 2002.

[11] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

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

[13] S. Gao, W. Wu, C. H. Lee, and T.-S. Chua. A MFoM learning approach to robust multiclass multi-label text categorization. In *Proceedings of the 21th International Conference on Machine Learning*, pages 329–336, Banff, Canada, 2004.

[14] N. Ghamrawi and A. McCallum. Collective multi-label classification. In *Proceedings of the 14th International Conference on Information and Knowledge Management*, pages 195–200, Bremen, Germany, 2005.

[15] 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*, pages 22–30, Sydney, Australia, 2004.

[16] W. Hackbusch. Iterative solution of large sparse systems of equations. *Mathematics of Computation*, 64(212):1759–1761, 1995.

[17] T. Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the 16th International Conference on Machine Learning*, pages 200–209, Bled, Slovenia, 1999.

[18] T. Joachims. Transductive learning via spectral graph partitioning. In *Proceedings of the 20th International Conference on Machine Learning*, pages 290–297, Washington, DC, 2003.

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

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

[21] D. D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

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

[23] P. Matstoms. Sparse QR factorization in MATLAB. *ACM Transactions on Mathematical Software*, 20(1):136 – 159, 1994.

[24] A. McCallum. Multi-label text classification with a mixture model trained by EM. In *Working Notes of the AAAI'99 Workshop on Text Learning*, Orlando, FL, 1999.

[25] M. Ng, G. Qiu, and A. Yip. A study of interactive multiple class image segmentation problems. Technical Report 07-51, UCLA CAM, 2007.

[26] P. Pavlidis, J. Weston, J. Cai, and W. N. Grundy. Combining microarray expression data and phylogenetic profiles to learn functional categories using support vector machines. In *Proceedings of the 5th International Conference on Computational Biology*, pages 242–248, Montréal, Canada, 2001.

[27] G. J. Qi, X. S. Hua, Y. Rui, J. Tang, T. Mei, and H. J. Zhang. Correlative multi-label video annotation. In *Proceedings of the 15th International Conference on Multimedia*, pages 17–26, Augsburg, Germany, 2007.

[28] R. E. Schapire and Y. Singer. Boostexter: a boosting-based system for text categorization. *Machine Learning*, 39(2-3):135–168, 2000.

[29] L. Sun, S.-W. Ji, and J.-P. Ye. Hypergraph spectral learning for multi-label classification. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 668–676, Las Vegas, NV, 2008.

[30] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007.

[31] N. Ueda and K. Saito. Parametric mixture models for multi-labeled text. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 721–728. Cambridge, MA: MIT Press, 2003.

[32] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, 1998.

[33] R. Weber, H.-J. Schek, and S. Blott. A quantitative analysis and performance study for similaritysearch methods in high-dimensional spaces. In *Proceedings of 24th International Conference on Very Large Data Bases*, pages 194–205, New York, NY, 1998.

[34] M.-L. Zhang and Z.-H. Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1479–1493, 2006.

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

[36] Y. Zhang and Z.-H. Zhou. Multilabel dimensionality reduction via dependence maximization. *ACM Transactions on Knowledge Discovery from Data*, 4(3):Article 14, 2010.

[37] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 321–328. Cambridge, MA: MIT Press, 2003.

[38] Z.-H. Zhou and M. Li. Semi-supervised learning by disagreement. *Knowledge and Information Systems*, 24(3):415–439, 2010.

[39] Z.-H. Zhou and M.-L. Zhang. Multi-instance multi-label learning with application to scene classification. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1609–1616. Cambridge, MA: MIT Press, 2006.

[40] S. Zhu, X. Ji, W. Xu, and Y. Gong. Multi-labelled classification using maximum entropy method. In *Proceedings of the 28th International Conference on Research and Development in Information Retrieval*, pages 274–281, Salvador, Brazil, 2005.

[41] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, 2006.

Xiangnan Kong received his bachelor and master degrees in computer science from Nanjing University, China, in 2006 and 2009, respectively. He has been working in data mining and machine learning, particularly in multi-label learning and semi-supervised learning. He is currently pursuing PhD degree in the Department of Computer Science, University of Illinois at Chicago.

Michael Ng is a Professor in the Department of Mathematics at the Hong Kong Baptist University. He obtained his B.Sc. degree in 1990 and M.Phil. degree in 1992 at the University of Hong Kong, and Ph.D. degree in 1995 at Chinese University of Hong Kong. He was a Research Fellow of Computer Sciences Laboratory at Australian National University (1995-1997), and an Assistant/Associate Professor (1997-2005) of the University of Hong Kong before joining Hong Kong Baptist University. As an applied mathematician, Michael's main research areas include Bioinformatics, Data Mining, Operations Research and Scientific Computing. Michael has published and edited 5 books, published more than 200 journal papers. He currently serves on several editorial boards of international journals.

Zhi-Hua Zhou (S'00-M'01-SM'06) received the BSc, MSc and PhD degrees in computer science from Nanjing University, China, in 1996, 1998 and 2000, respectively, all with the highest honors. He joined the Department of Computer Science & Technology at Nanjing University as an assistant professor in 2001, and is currently Cheung Kong Professor and Director of the LAMDA group. His research interests are mainly in artificial intelligence, machine learning, data mining and pattern recognition. In these areas he has published over 80 papers in leading international journals or conference proceedings, and holds 11 patents. Dr. Zhou has been awarded with various honors such as the Fok Ying Tung Young Scholar First-Grade Award (2010), the National Science & Technology Award for Young Scholars of China (2006), the Microsoft Young Professorship Award (2006), the Award of National Science Fund for Distinguished Young Scholars of China (2003), etc. He is an Associate Editor-in-Chief of the *Chinese Science Bulletin*, Associate Editor of the *IEEE Transactions on Knowledge and Data Engineering* and *ACM Transactions on Intelligent Systems and Technology*, and on the editorial boards of various other journals. He is the Founding Steering Committee Co-Chair of ACML and Steering Committee member of PAKDD and PRICAI. He serves/ed as Program Committee Chair/Co-Chair for PAKDD'07, PRICAI'08 and ACML'09, Vice Chair or Area Chair or Senior PC of various conferences such as KDD, ICDM, SDM, ECMLPKDD, AAAI, IJCAI, CIKM, ICPR, etc., and chaired various native conferences in China. He is the Chair of the Machine Learning Technical Committee of the Chinese Association of Artificial Intelligence, Vice Chair of the Artificial Intelligence & Pattern Recognition Technical Committee of the China Computer Federation and the Chair of the IEEE Computer Society Nanjing Chapter. He is a senior member of the ACM and senior member of the IEEE.