Unsupervised Transfer Classification: Application to Text Categorization

Tianbao Yang, Rong Jin, Anil K. Jain, Yang Zhou, Wei Tong
Department of Computer Science and Engineering
Michigan State University, East Lansing, MI, 48824, USA
{yangtia1, rongjin, jain, zhouyang, tongwei}@cse.msu.edu

ABSTRACT

We study the problem of building the classification model for a target class in the absence of any labeled training example for that class. To address this difficult learning problem, we extend the idea of transfer learning by assuming that the following side information is available: (i) a collection of labeled examples belonging to other classes in the problem domain, called the auxiliary classes; (ii) the class information including the prior of the target class and the correlation between the target class and the auxiliary classes. Our goal is to construct the classification model for the target class by leveraging the above data and information. We refer to this learning problem as **unsupervised transfer classification**. Our framework is based on the generalized maximum entropy model that is effective in transferring the label information of the auxiliary classes to the target class. A theoretical analysis shows that under certain assumption, the classification model obtained by the proposed approach converges to the optimal model when it is learned from the labeled examples for the target class. Empirical study on text categorization over four different data sets verifies the effectiveness of the proposed approach.

Categories and Subject Descriptors

I.5 [**Pattern Recognition**]: Design Methodology—*Classifier design and evaluation*; H.1 [**Models and Principles**]: Miscellaneous

General Terms

Algorithms, Experimentation

Keywords

Unsupervised Transfer Classification, Text Categorization, Generalized Maximum Entropy

KDD'10, July 25–28, 2010, Washington, DC, USA.

1. INTRODUCTION

Semi-supervised learning is designed to reduce the number of labeled examples for building accurate classification model by utilizing unlabeled data. Many semi-supervised learning techniques ([31] and references therein) have been developed and successfully applied to text categorization. In this work, we examine the problem of learning the classification model for a given class, called the target class, without a single labeled training example for that class. This can be viewed as an extreme case of semi-supervised learning. In order to address this difficult learning problem, we extend the idea of transfer learning by assuming that the following side information is available

- a collection of labeled training examples for the classes other than the target class, called the *auxiliary classes*, and
- the *class information*, including the prior for the target class and the conditional probabilities between the target class and the auxiliary classes.

Our goal is to construct the classification model for the target class by effectively transferring the label information of the auxiliary classes to the target class. We refer to the above problem as **unsupervised transfer classification** in order to distinguish from most studies in transfer learning for classification where some labeled examples are available for the target class.

Unsupervised transfer classification is particularly useful when the target class does not have any labeled example. This scenario is encountered in many applications. For instance, in automatic image annotation [12], given the limited size of the vocabulary that is used for training, we often encounter the problem of how to annotate images with a keyword outside the training vocabulary. A similar problem arises in social tagging [14] when some of the tags are so rare that it becomes difficult to collect any useful example for these tags. The unsupervised transfer classification can be applied to these problems by automatically learning an annotation model for the new keyword (or rare tag) even if it does not have a single labeled example.

We address the problem of unsupervised transfer classification by effectively transferring the label information of the auxiliary classes to the target class. We propose a framework based on the generalized maximum entropy model that effectively leverages the class information as well as the training examples for the auxiliary classes. Our analysis shows that under certain assumption, the classification model found by the proposed approach will converge to the optimal model when it is learned from the labeled examples for the target class. An empirical study on text categorization over four different data sets verifies the efficacy of the proposed approach. The contributions of this paper are summarized as follows:

- We propose a generalization of the traditional maximum entropy method for classification.
- We present a framework for unsupervised transfer classification based on the generalized maximum entropy model.
- We provide a consistency analysis of the proposed approach under certain assumption.
- We design an efficient algorithm to solve the optimization problem.

The remainder of this paper is organized as follows. In section 2, we review some related work. In section 3, we present the framework for unsupervised transfer classification. We present experimental results in section 4, and finally conclude in section 5.

2. RELATED WORK

Our work is related to transfer learning, multi-label learning, and maximum entropy learning.

Transfer Learning The objective of transfer learning is to transfer the knowledge from the source domain to the target domain. The transferred knowledge can take various forms, such as knowledge about training examples [7, 11, 17], knowledge of feature representation [2, 3, 8], and knowledge of model parameters [16, 4, 27] and many others [21]. Since the objective of our work is to transfer the label information of the auxiliary classes to the target class, it is closely related to the study of transfer learning. Unlike most studies in transfer learning for classification that require labeled examples for the target class, we assume no labeled example is available for the target class, making it a more challenging and realistic problem in many applications.

Multi-label Learning Several multi-label learning algorithms [26, 29, 30, 13, 10] have been designed to exploit the correlation between classes for constructing classification models. [26] proposes a generative model for multi-label learning to incorporate the pairwise class correlation information; [29] exploits the class correlation by introducing a common prior shared by all classes; Zhu et al. [30] propose a maximum entropy model for multi-label learning that exploits the class correlation information. [13] explores the class correlation in a label propagation framework. In [10], a common subspace is assumed to be shared by all the labels. Our work is related to these studies in exploring the class correlation for classification, but differs from these studies in that while they are focused on supervised learning, the objective of this work is to build a classification model without a single labeled example for the target class.

Maximum Entropy Learning Maximum entropy principle has been successfully applied to natural language processing [6, 24, 23] and text categorization [20, 30, 18]. The proposed maximum entropy model generalizes the traditional model by introducing (i) inequality constraints into the model to replace the original equality constraints, and (ii) different ways for estimating the sufficient statistics. Note that the proposed generalized maximum entropy model is closely related to [1] in which inequality constraints are introduced into the framework of divergence minimization. Our generalized maximum entropy model differs from [1] in that we introduce a regularization term for the errors related to the inequality constraints. The regularization term is particularly important for maximum entropy model since the dual problem of maximum entropy model is in general not strongly convex. Additionally, we want to point out that the generalized maximum entropy model is also presented in the work of Yang et al. [28], but we emphasize that this work differs from [28] in that we solve the problem of unsupervised transfer classification rather than learning from noisy side information.

Finally, our work is also closely related to [15, 22]. Similar to our problem, the label information is not given explicitly in these studies, and the goal is to build classification models from multiple sets of unlabeled data for which only the class proportion information is available. Unlike these two studies, in our work, the class information is utilized to assist the transfer of label information of the auxiliary classes to the target class.

3. UNSUPERVISED TRANSFER CLASSIFICATION

In this section, we first present the problem of unsupervised transfer classification. We then present a generalized maximum entropy model, and a framework for unsupervised transfer classification based on the generalized maximum entropy model. The optimization algorithm and the consistency analysis of the proposed method are also presented. The issues of how to obtain the class information and the specific assumption made by the proposed framework are addressed at the end of this section.

3.1 Problem Definition

Let $\mathcal{D} = \{\mathbf{x}_i \in \mathcal{X}, i = 1, \ldots, n\}$ be a set of training examples that are assigned to the auxiliary classes $\mathcal{C} = \{c_1, \ldots, c_K\}$, where K is the number of the auxiliary classes. We use $y_i^k \in \{0, 1\}, k = 1, \ldots, K$ to indicate the assignment of class c_k to example $\mathbf{x}_i$. Note that each example can be assigned to multiple classes. By using a standard supervised learning method (e.g., logistic regression model), we can learn a binary prediction function, denoted by $p(y^k = 1|\mathbf{x})$, that outputs the likelihood of assigning $\mathbf{x}$ to class c_k. Our objective however is to learn a binary prediction function $p(y^t|\mathbf{x})$ for a target class $c_t \notin \mathcal{C}$ that does not have a single labeled example. In order to learn $p(y^t|\mathbf{x})$ for the target class c_t, we need to transfer the label information from the auxiliary classes in $\mathcal{C}$ to the target class. To this end, we assume the following class information is available: (i) the prior for the target class c_t, i.e., $p(y^t = 1)$, and (ii) the conditional probabilities between the target class and the auxiliary classes in $\mathcal{C}$, i.e., $p(y^t = 1|y^k = 1), k = 1, \ldots, K$. We refer to this learning problem as **unsupervised transfer classification**.

A straightforward approach for this problem is to construct the prediction function $p(y^t|\mathbf{x})$ by a weighted combination of the prediction functions for the auxiliary classes, i.e.,

$$p(y^t = 1|\mathbf{x}) = \frac{\sum_{k=1}^{K} p(y^k = 1|\mathbf{x}) p(y^t = 1|y^k = 1)}{\sum_{k'=1}^{K} p(y^t = 1|y^{k'} = 1)} \qquad (1)$$

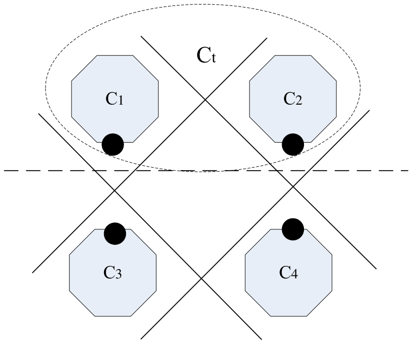


Figure 1: An illustrative example showing the limitation of the combination approach in (1).

We refer to the method in (1) as the combination of classification models, or **cModel** for short.

The major shortcoming of the combination approach is that it imposes a strong constraint in constructing the prediction function for the target class c_t. In particular, if a training example is not a support vector[1] for any of the auxiliary classes in $\mathcal{C}$, it will never be a support vector for the target class c_t, leading to a serious limitation in building classification model for c_t.

To illustrate this limitation, consider the problem in Figure 1, in which we have four auxiliary classes c_1, c_2, c_3, c_4 that are highlighted by four shaded octagons. The decision boundaries that distinguish each class from the other three classes are highlighted by four solid lines. Assume SVM is used to learn the individual classifiers. Note that since the training examples in the four small solid circles are not the closest to any of the decision boundaries, they will not be the support vectors for the four auxiliary classes. Suppose the target class c_t is essentially a combination of c_1 and c_2 whose decision boundary is highlighted by the horizonal dashed line. In the case of supervised learning, since the examples in the small solid circles are the closest to the dashed line, they should be support vectors for c_t. Unfortunately, in the prediction function generated by the combination approach, none of these examples will be support vectors for c_t, leading to a suboptimal model for c_t.

3.2 Generalized Maximum Entropy Model

Before we present the generalized maximum entropy model, we first motivate the proposed approach by explaining why maximum entropy could be an attractive approach for unsupervised transfer classification. The formulation of the traditional maximum entropy model for classification is given as follows:

$$\begin{aligned} \max \quad & -\sum_{i=1}^{n}\sum_{y} p(y|\mathbf{x}_i)\log p(y|\mathbf{x}_i) \\ s.t. \quad & \frac{1}{n}\sum_{i=1}^{n} p(y|\mathbf{x}_i) f_j(\mathbf{x}_i) = \frac{1}{n}\sum_{i=1}^{n}\delta(y_i, y) f_j(\mathbf{x}_i), \forall y, j \end{aligned} \tag{2}$$

where $f_j(\cdot)$ is the $j^{\text{th}} \in \{1, \cdots, d\}$ feature function defined on $\mathcal{X}$ and $\delta(y_i, y)$ is the Kronecker delta function that outputs 1 if $y_i = y$ and zero, otherwise. It is important to note that in order to train a maximum entropy model, we only need to know the quantity $\sum_{i=1}^{n}\delta(y_i, y) f_j(\mathbf{x}_i)/n$ (i.e., the sufficient statistics), not the class assignments of individual examples. In fact, if we have means to approximately compute $\sum_{i=1}^{n}\delta(y_i, y) f_j(\mathbf{x}_i)/n$, denoted by $\widehat{f}_j(y)$, without knowing the class label of each example, we can modify the maximum entropy model in (2) by replacing $\sum_{i=1}^{n}\delta(y_i, y) f_j(\mathbf{x}_i)/n$ with $\widehat{f}_j(y)$, i.e.,

$$\begin{aligned} \max \quad & -\sum_{i=1}^{n}\sum_{y} p(y|\mathbf{x}_i)\log p(y|\mathbf{x}_i) \\ s.t. \quad & \frac{1}{n}\sum_{i=1}^{n} p(y|\mathbf{x}_i) f_j(\mathbf{x}_i) = \widehat{f}_j(y), \forall y, j \end{aligned} \tag{3}$$

The key idea of the proposed approach is to estimate $\widehat{f}_j(y)$ using the class information, which will be elaborated later.

Although it appears intuitively correct, the formulation suggested in (3) could be problematic. First, for an arbitrarily approximate estimate $\widehat{f}_j(y)$, the problem in (3) may not even be feasible. For instance, in the case of binary classification, i.e., $y \in \{0, 1\}$, by adding the constraints for $\widehat{f}_j(y = 1)$ and $\widehat{f}_j(y = 0)$, we will have the following implicit constraint

$$\widehat{f}_j(y = 0) + \widehat{f}_j(y = 1) = \frac{1}{n}\sum_{i=1}^{n} f_j(\mathbf{x}_i). \tag{4}$$

If two arbitrary estimates $\widehat{f}_j(y = 0)$ and $\widehat{f}_j(y = 1)$ do not satisfy the constraint in (4), they will lead to an infeasible optimization problem for (3). More importantly, using the equality constraints in the maximum entropy model in (2) is by itself problematic. Note that both sides of the equality constraint in (2) can be interpreted as empirical estimates of the expectation $\mathrm{E}_{X,Y}[\delta(Y, y) f_j(X)]$. As indicated by the following theorem, the two quantities are identical only when the number of examples n goes to infinity, and could be significantly different when n is small.

THEOREM 1. **Concentration of MaxEnt's Constraint** *Assume $(\mathbf{x}_i, y_i)$ are i.i.d. samples from an unknown distribution $P(X, Y)$. The equality constraint in (2) for any j and y holds with probability 1 when the number of examples n approaches infinity. However, with finite n, the following inequality holds for any $\epsilon > 0$*

$$\begin{aligned} & \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} p(y|\mathbf{x}_i) f_j(\mathbf{x}_i) - \frac{1}{n}\sum_{i=1}^{n}\delta(y_i, y) f_j(\mathbf{x}_i)\right| \geq \epsilon\right) \\ & \leq 4\exp\left(-\frac{\epsilon^2 n}{8R_j^2}\right) \end{aligned}$$

where $R_j = \max_{X \in \mathcal{X}} |f_j(X)|$.

The proof for the theorem can be found in Appendix A. Based on the above discussion, we relax the equality constraint in the maximum entropy model into inequality constraint,

$$\begin{aligned} \max \quad & -\frac{1}{n}\sum_{i=1}^{n}\sum_{y} p(y|\mathbf{x}_i)\log p(y|\mathbf{x}_i) - \frac{1}{2\gamma}\sum_{y}\|\epsilon_y\|^2 \\ s.t. \quad & \frac{1}{n}\sum_{i=1}^{n} p(y|\mathbf{x}_i) f_j(\mathbf{x}_i) \geq \widehat{f}_j(y) - \epsilon_{yj}, \forall y, j \end{aligned} \tag{5}$$

[1] Here, we slightly abuse the terminology of support vectors. For a non-SVM classifier, support vector is referred to any training example that is heavily weighted by the classifier.

where $\epsilon_y = (\epsilon_{y1}, \dots, \epsilon_{yd})$ and $\|\epsilon_y\|$ measures the norm of vector ϵ_y. Note that in the above formulation, to account for the difference between the two estimates, we introduce dummy variables ϵ_{yj}. In addition, we introduce a regularization term $\|\epsilon\|^2/(2\gamma)$ into the objective for these dummy variables so that they can be determined automatically. γ is a regularization parameter that will be determined empirically. We refer to the formulation in (5) as the **Generalized Maximum Entropy Model**. It includes two distinguished features compared to the traditional maximum entropy model: (i) it allows different ways for estimating $\mathrm{E}_{X,Y}[\delta(Y,y)f_j(X)]$ (i.e., $\widehat{f}_j(y)$) that could potentially avoid the requirement of knowing the class assignments of all training examples, and (ii) it uses the inequality constraint to allow the mismatch between data and the prediction function $p(y|\mathbf{x})$. With the inequality constraint, we essentially have

$$\widehat{f}_j(y) - \epsilon_{yj} \leq \frac{1}{n}\sum_{i=1}^{n} p(y|\mathbf{x}_i) f_j(\mathbf{x}_i) \leq \widehat{f}_j(y) + \epsilon_{\bar{y}j}$$

where $\bar{y} = 1 - y$ and the upper bound is derived from the following implicit constraint

$$\frac{1}{n}\sum_{y}\sum_{i=1}^{n} p(y|\mathbf{x}_i) f_j(\mathbf{x}_i) = \sum_{y} \widehat{f}_j(y) = \frac{1}{n}\sum_{i=1}^{n} f_j(\mathbf{x}_i)$$

The following proposition shows the relationship between the generalized maximum entropy model in (5) and the *regularized* logistic regression model.

Proposition 1. *When* $\widehat{f}_j(y) = \sum_{i=1}^{n} \delta(y, y_i) f_j(\mathbf{x}_i)/n$, *and* $\|\cdot\| = \|\cdot\|_2$, *the dual problem of (5) is equivalent to the regularized logistic regression model, i.e.,*

$$\min_{\lambda \in \mathbb{R}^d} \frac{\gamma}{2}\|\lambda\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp\left(-\sum_{j=1}^{d} \tilde{y}_i \lambda_j f_j(\mathbf{x}_i)\right)\right)$$

where $\tilde{y}_i = 1$ *if* $y_i = 1$ *and* $\tilde{y} = -1$ *if* $y_i = 0$.

Proposition 1 follows directly the result in (10) that will be presented later.

3.3 Estimating $\widehat{f}_j(y)$ Using Class Information

In order to build a classification model for the target class c_t , we will apply the generalized maximum entropy model in (5). The key question is how to compute $\widehat{f}_j(y^t)$, an estimate of the expectation $\mathrm{E}_{X,Y^t}[\delta(Y^t, y^t) f_j(X)]$ for the target class c_t, using the class information.

First, notice that using the class prior information $p(y^t)$, we could write the expectation of feature functions as

$$\mathrm{E}_{X,Y^t}[\delta(Y^t, y^t) f_j(X)] = p(y^t)\mathrm{E}_{X|Y^t=y^t}[f_j(X)]$$

Thus, $\widehat{f}_j(y^t)$ can be computed as

$$\widehat{f}_j(y^t) \simeq p(y^t) u_{\mathbf{x}|y^t}[f_j(\mathbf{x})] \tag{6}$$

where $u_{\mathbf{x}|y^t}[f_j(\mathbf{x})]$ is the estimate of the conditional expectation $\mathrm{E}_{X|Y^t=y^t}[f_j(X)]$ based on the finite number of training examples. Therefore, our goal is to compute $u_{\mathbf{x}|y^t}[f_j(\mathbf{x})]$.

Second, note that for all the auxiliary classes $c_k \in \mathcal{C}$, $u_{\mathbf{x}|y^k}[f_j(\mathbf{x})]$ can be simply computed as

$$u_{\mathbf{x}|y^k}[f_j(\mathbf{x})] = \frac{\sum_{i=1}^{n} \delta(y_i^k, y^k) f_j(\mathbf{x}_i)}{\sum_{i=1}^{n} \delta(y_i^k, y^k)} \tag{7}$$

To compute $u_{\mathbf{x}|y^t}[f_j(\mathbf{x})]$ for the target class, an intuitive approach is to approximate it by a linear combination of its counterparts for the auxiliary classes. As a result, we need to establish the relationship between $u_{\mathbf{x}|y^t}[f_j(\mathbf{x})]$ and $\{u_{\mathbf{x}|y^k}[f_j(\mathbf{x})], k = 1, \dots, K\}$. To this end, we make the following assumption

Assumption 1 (**A1**). *The following relationship holds for* $\Pr(X|Y^k = y^k)$ *for any auxiliary class* c_k

$$\begin{aligned}\Pr(X|Y^k = y^k) &= \Pr(X|Y^t = 1)\Pr(Y^t = 1|Y^k = y^k) \\ &\quad + \Pr(X|Y^t = 0)\Pr(Y^t = 0|Y^k = y^k)\end{aligned}$$

Note that assumption **A1** essentially assumes that X is conditionally independent of Y^k given Y^t, i.e. $\Pr(X|Y^t, Y^k) = \Pr(X|Y^t)$, which may not be true in real-world applications. Later on we will discuss how to relax this assumption. Given the assumption **A1**, we have the following relations for the conditional expectation $\mathrm{E}_{X|Y=y}[f_j(X)]$:

$$\begin{aligned}\mathrm{E}_{X|Y^k=y^k}[f_j(X)] &= \mathrm{E}_{X|Y^t=1}[f_j(X)]\Pr(Y^t = 1|Y^k = y^k) \\ &\quad + \mathrm{E}_{X|Y^t=0}[f_j(X)]\Pr(Y^t = 0|Y^k = y^k)\end{aligned}$$

which leads to the following regression relations for $u_{\mathbf{x}|y^k}[f_j(\mathbf{x})]$

$$\begin{aligned}u_{\mathbf{x}|y^k}[f_j(\mathbf{x})] &\simeq u_{\mathbf{x}|y^t=1}[f_j(\mathbf{x})]p(y^t = 1|y^k) \\ &\quad + u_{\mathbf{x}|y^t=0}[f_j(\mathbf{x})]p(y^t = 0|y^k) + \varepsilon\end{aligned}$$

where ε is the error term. By putting the regression relations for all the auxiliary classes into the matrix form, we have the following regression system:

$$\widehat{\mathbf{A}} \simeq \mathbf{W}_{K\times 2} U_{\mathbf{x}|y^t} + \varepsilon \tag{8}$$

where $\mathbf{W}$ and $U_{\mathbf{x}|y^t}$ and $\widehat{\mathbf{A}}$ are defined as follows

$$\begin{aligned}\mathbf{W} &= \begin{pmatrix} p(y^t = 1|y^1 = 1) & p(y^t = 0|y^1 = 1) \\ \cdots & \cdots \\ p(y^t = 1|y^K = 1) & p(y^t = 0|y^K = 1) \end{pmatrix} \\ U_{\mathbf{x}|y^t} &= \begin{pmatrix} u_{\mathbf{x}|y^t=1}[f_1(\mathbf{x})] & \dots & u_{\mathbf{x}|y^t=1}[f_d(\mathbf{x})] \\ u_{\mathbf{x}|y^t=0}[f_1(\mathbf{x})] & \dots & u_{\mathbf{x}|y^t=0}[f_d(\mathbf{x})] \end{pmatrix} \\ \widehat{\mathbf{A}} &= \begin{pmatrix} u_{\mathbf{x}|y^1=1}[f_1(\mathbf{x})] & \dots & u_{\mathbf{x}|y^1=1}[f_d(\mathbf{x})] \\ \cdots & & \cdot \\ u_{\mathbf{x}|y^K=1}[f_1(\mathbf{x})] & \dots & u_{\mathbf{x}|y^K=1}[f_d(\mathbf{x})] \end{pmatrix}\end{aligned}$$

By solving the regression system in (8) under the implicit constraint in (4), i.e. $U_{\mathbf{x}|y^t}^\top \mathbf{p}_t = \mathbf{u}_{\mathbf{x}}[f(\mathbf{x})]$, we have the following solution for $U_{\mathbf{x}|y^t}$

$$\begin{aligned}U_{\mathbf{x}|y^t} &= (\mathbf{W}^\top\mathbf{W})^{-1}\cdot \\ &\left(\frac{\mathbf{p}_t \mathbf{u}_{\mathbf{x}}^\top}{\mathbf{p}_t^\top(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{p}_t} + \left(\mathbf{I} - \frac{\mathbf{p}_t\mathbf{p}_t^\top(\mathbf{W}^\top\mathbf{W})^{-1}}{\mathbf{p}_t^\top(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{p}_t}\right)\mathbf{W}^\top\widehat{\mathbf{A}}\right)\end{aligned} \tag{9}$$

where

$$\begin{aligned}\mathbf{u}_{\mathbf{x}}[f(\mathbf{x})] &= (u_{\mathbf{x}}[f_1(\mathbf{x})], \dots, u_{\mathbf{x}}[f_d(\mathbf{x})])^\top \\ u_{\mathbf{x}}[f_j(\mathbf{x})] &= \frac{1}{n}\sum_{i=1}^{n} f_j(\mathbf{x}_i),\ \mathbf{p}_t = \begin{pmatrix} p(y_t = 1) \\ p(y_t = 0) \end{pmatrix}\end{aligned}$$

3.4 Optimization and Consistency Analysis

Using the method described in the previous subsection, we will be able to compute $\widehat{f}_j(y^t)$ using the class information. In this section, we present the optimization algorithm and consistency analysis.

Maximum entropy model is usually solved via its dual problem. Here we show the dual problem for (5) in (10). The derivation is skipped due to space limitations.

$$\max_{\substack{\lambda_1\in\mathbb{R}^d_+\\ \lambda_0\in\mathbb{R}^d_+}} \; -L(\lambda) = \left(\lambda_1^\top,\lambda_0^\top\right)\begin{pmatrix}\widehat{\mathbf{f}}_1\\ \widehat{\mathbf{f}}_0\end{pmatrix} - \frac{\gamma}{2}\left(\|\lambda_1\|_*^2+\|\lambda_0\|_*^2\right) \qquad (10)$$

$$-\frac{1}{n}\sum_i \log\left(\exp[\lambda_1^\top f(\mathbf{x}_i)] + \exp[\lambda_0^\top f(\mathbf{x}_i)]\right)$$

where

- λ_1 and λ_0 are the dual variables, $\lambda^\top = (\lambda_1^\top, \lambda_0^\top)$,
- $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$,
- $\widehat{\mathbf{f}}_1 = p(y^t = 1)U_{\mathbf{x}|y^t=1}^\top$ and $\widehat{\mathbf{f}}_0 = p(y^t = 0)U_{\mathbf{x}|y^t=0}^\top$, and
- $f(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_d(\mathbf{x}))^\top$.

The dual problem can be solved efficiently by using Nesterov method [19] with details given in Algorithm 1. One of the key steps in running the Nesterov method in Algorithm 1 is to solve the constrained optimization problem in (11). It is easy to verify that the optimal solution to (11) is

$$T(C,\mathbf{a}_i) = \Pi_{\mathbb{R}^{2d}_+}\left(\mathbf{a}_i - \frac{1}{C}\nabla L(\mathbf{a}_i)\right)$$

where $\Pi_{\mathbb{R}^{2d}_+}$ is the operator that projects the elements in a vector into the positive orthant. The convergence rate of the Nesterov method is $\mathcal{O}\left(\frac{1}{N^2}\right)$, where N is the total number of iterations. The time complexity of the optimization algorithm is $\mathcal{O}(N(n+d))$. With the computed dual variables λ_1 and λ_0, the prediction function for the target class c_t is given by

$$p(y^t = 1|\mathbf{x}) = \frac{\exp(\lambda_1^\top f(\mathbf{x}))}{\exp(\lambda_1^\top f(\mathbf{x})) + \exp(\lambda_0^\top f(\mathbf{x}))}$$

Next, we present the consistency analysis and show that under assumption **A1** the solution obtained by the proposed approach will converge to the optimal one trained by the labeled examples for the target class. Since our approach depends on $u_{\mathbf{x}|y^t}[f_j(\mathbf{x})]$, which is an estimate for $\mathrm{E}_{X|Y^t}[f_j(X)]$, in the first step of consistency analysis, we bound the difference between these two quantities via the following theorem.

THEOREM 2. *Assume bounded feature function $f_j(\mathbf{x})$, i.e. $|f_j(\mathbf{x})| \le R, j = 1, \dots, d$, and the prior for all the auxiliary classes are significantly large, i.e., there exists some positive constant $\rho > 0$ such that $p(y^k = 1) \ge \rho, k = 1, \dots, K$. Under the assumption **A1**, for any $\delta > 10Kd\exp(-n\rho^2/4)$, with probability at least $1-\delta$, we have*

$$\left\|\mathrm{E}_{X|Y^t}[f(X)] - U_{\mathbf{x}|y^t}\right\|_F \le$$
$$\frac{4\sqrt{2}R(1+\kappa^{3/2})}{\rho\sqrt{\sigma_{\min}}}\sqrt{\frac{Kd}{n}\ln\left(\frac{10Kd}{\delta}\right)} + \frac{\kappa R}{\|\mathbf{p}_t\|_2}\sqrt{\frac{2d}{n}\ln\left(\frac{4d}{\delta}\right)}$$

where $\sigma_{\max}, \sigma_{\min}$ are the maximum and minimum eigenvalue of $\mathbf{W}^\top\mathbf{W}$, and $\kappa = \sigma_{\max}/\sigma_{\min}$.

The proof can be found in Appendix B. As revealed by Theorem 2, the difference between the estimate $U_{\mathbf{x}|y^t}$ and $\mathrm{E}_{X|Y^t}[f(X)]$ will essentially diminish when the number of examples n goes to infinity. In the next theorem, we show how the difference in the estimates $\widehat{f}_j(y)$ will affect the solution to the dual problem in (10).

Algorithm 1 Solving the dual problem in (10)

```
1. **Input** the number of iterations or convergence rate
2. Initialize the approximate solution $\mathbf{b}_1 \in \mathbb{R}^{2d}$, search
   point $\mathbf{a}_0 \in \mathbb{R}^{2d}$, auxiliary point $\mathbf{q}_0 \in \mathbb{R}^{2d}$ and positive
   reals $t_0, C_0 \in \mathbb{R}_+$:
       $\mathbf{b}_1 = 0, \mathbf{a}_0 = 0, \mathbf{q}_0 = 0, t_0 = 1, C_0$
3. In the $i^{\text{th}} (i \ge 1)$ step, given $\mathbf{b}_i, \mathbf{a}_{i-1}, \mathbf{q}_{i-1}, t_{i-1}, C_{i-1}$,
   act as follows:
     • Set  $\mathbf{a}_i = \mathbf{b}_i - \frac{1}{t_{i-1}}(\mathbf{b}_i + \mathbf{q}_{i-1})$
       compute $L(\mathbf{a}_i), \nabla L(\mathbf{a}_i)$
     • Testing sequentially the values $C = 2^j C_{i-1}, j = 0, 1 \cdots$, find the first value of $C$ such that
           $L(T(C,\mathbf{a}_i)) \le L_{C,\mathbf{a}_i}(T(C,\mathbf{a}_i))$
       where
           $T(C,\mathbf{a}_i) = \arg\min_{\mathbf{b}\in\mathbb{R}^{2d}_+} L_{C,\mathbf{a}_i}(\mathbf{b})$     (11)
               $= L(\mathbf{a}_i) + \nabla L(\mathbf{a}_i)^\top(\mathbf{b}-\mathbf{a}_i) + \frac{C}{2}\|\mathbf{b}-\mathbf{a}_i\|_2^2$
       Set $C_i$ to the resulting value of $C$.
     • Set  $\mathbf{b}_{i+1} = T(C_i, \mathbf{a}_i)$
            $\mathbf{q}_i = \mathbf{q}_{i-1} + t_{i-1}(\mathbf{a}_i - T(C_i, \mathbf{a}_i))$
       Set $t_i$ as the largest root of the equation $t^2 - t = t_{i-1}^2$
4. repeat Step 3 until the input number of iterations is
   exceeded or convergence rate is satisfied.
5. **Output** $(\lambda_1; \lambda_0) = \mathbf{b}$
```

THEOREM 3. *Assume ℓ_2 norm is in the generalized maximum entropy model, i.e., $\|\cdot\| = \|\cdot\|_2$. Let $\lambda^{*\top} = (\lambda_1^{*\top}, \lambda_0^{*\top})$ be the solution to the optimization problem in (10) with $\widehat{\mathbf{f}}^* = \left(\widehat{\mathbf{f}}_1^*, \widehat{\mathbf{f}}_0^*\right)^\top$, and $\lambda^{o\top} = (\lambda_1^{o\top}, \lambda_0^{o\top})$ be the solution with $\widehat{\mathbf{f}}^o = \left(\widehat{\mathbf{f}}_1^o, \widehat{\mathbf{f}}_0^o\right)^\top$. We have*

$$\|\lambda^* - \lambda^o\|_2 \le \frac{2}{\gamma}\|\widehat{\mathbf{f}}^* - \widehat{\mathbf{f}}^o\|_F$$

The proof can be found in Appendix C. Combining the results in Theorems 2 and 3, we have the following consistency result for the solution obtained by the proposed approach, which verifies the model obtained by the proposed approach converges to the optimal model learned in the presence of labeled examples for the target class.

THEOREM 4. *Assume (i) ℓ_2 norm is used in the generalized maximum entropy model, (ii) $|f_j(\mathbf{x})| \le R, j = 1, \dots, d$ for any $\mathbf{x}$, and (iii) $p(y^k = 1) \ge \rho, k = 1, \dots, K$ for some $\rho > 0$. As the number of training examples goes to infinite, under the assumption **A1**, the optimal solution $\lambda^\top = (\lambda_1^\top, \lambda_0^\top)$ to (10) with $\widehat{f}_j(y^t) = p(y^t)u_{\mathbf{x}|y^t}[f_j(\mathbf{x})]$ will converge to $\lambda^{*\top} = (\lambda_1^{*\top}, \lambda_0^{*\top})$ with probability 1, where λ^* is the optimal solution to (10) with $\widehat{f}_j^*(y^t) = p(y^t)\mathrm{E}_{X|Y^t=y^t}[f_j(X)]$.*

3.5 Implementation Issues

In this section, we discuss two issues: (i) how to obtain the class information in real-world applications, and (ii) how to relax the assumption **A1**.

Obtaining Class Information The class information includes $p(y^t = 1|y^k = 1)$ and $p(y^t = 1)$. One approach is to derive the class information from a different domain that shares the same set of classes as the target domain. For example, in the case of text categorization of language a, we could derive the class information from the labeled documents that are in a different language b. The class information can also be obtained by querying external sources such as a web search engine or a particular web site. For instance, to classify research articles into predefined topics, we can measure the correlation between two research topics by simply counting the number of returned URLs after querying a search engine with the conjunction of the two topics. Evidently, these methods may not obtain an accurate estimate of the class information. However, as will be shown in our empirical study, even with such possibly inaccurate estimate of the class information, the proposed approach could still yield reasonably accurate prediction for text categorization.

Relaxing Assumption A1 As we discussed before, assumption **A1** is equivalent to having the independence between X and Y^k given Y^t, i.e. $\Pr(X|Y^t, Y^k) = \Pr(X|Y^t)$, which may not be true for real-world applications. We can relax this assumption by approximating $\Pr(X|Y^t, Y^k)$ with a linear combination of $\Pr(X|Y^t)$ and $\Pr(X|Y^k)$. For instance, we could have the following approximations for the probability $\Pr(X|Y^t, Y^k)$:

$$\Pr(X|Y^t = 1, Y^k = 0) = \Pr(X|Y^t = 1)$$

$$\Pr(X|Y^t = 0, Y^k = 1) = \Pr(X|Y^k = 1)$$

$$\Pr(X|Y^t = y, Y^k = y) = \frac{1}{2}\left[\Pr(X|Y^t = y) + \Pr(X|Y^k = y)\right]$$

With the above approximations, a similar result can be derived using the procedure described in Section 3.3.

4. EXPERIMENTS WITH TEXT CATEGORIZATION

4.1 Data Sets and Baselines

We verify the efficacy of the proposed algorithm for unsupervised transfer classification on the problem of text categorization. Four text data sets are used for evaluation: (i) "tmc2007" [5] data set is used in 2007 SIAM text mining workshop for text mining competition, (ii) "enron" [2] data set includes email messages from about 150 users, mostly senior management of Enron, (iii) "bibtex" data set, processed by Katakis et al. [14], contains the metadata for the bibtex items such as title and authors of papers and (iv) "delicious" [25] data set extracted from the delicious social bookmarking site on April 7 2007; it contains textual description of each web page along with its annotated tags. In our study, we follow the default partition of training data and testing data specified by the authors of the data sets. The statistics of the four data sets are summarized in Table 1. The last column gives the percentage of the training examples. To preprocess the documents, we normalize the attributes of a document by first dividing them by the sum of all attributes of the document and then taking the square root of the ratios [9]. Each normalized attribute is used as a different feature function $f_j(\mathbf{x})$.

Table 1: Statistics of Data sets

name	#examples	#attributes	#class	%training
tmc2007	28596	49060	22	75%
enron	1702	1001	53	67%
bibtex	7395	1836	159	67%
delicious	16105	500	983	80%

To test the capability of the proposed algorithm in building the classification model for a target class without a single labeled example, we follow the paradigm of "leave one class out" cross validation by choosing one class as the target class and using the remaining classes as the auxiliary classes. The proposed algorithm is applied to learn a classification model for the target class when we only have (i) the assignments of training documents to the auxiliary classes and (ii) the class information. We evaluated the proposed algorithm on both training data[3] and testing data. We repeat the same procedure for every class in each data set, and the result averaged over all the classes in the data set is reported in this study. The Area under ROC curve (AUC) is used as the evaluation metric in our study. Compared to the other evaluation metrics (such as F_1), AUC is advantageous in that it does not require the classifier to make explicit binary decisions and therefore avoids the bias in evaluation caused by the choice of the threshold. For all the experiments, the regularization parameter γ is set to $\gamma = 0.01/n$, where n is size of the training set. This choice of γ usually yields good classification performance.

Besides the combination approach in (1) (i.e., **cModel**), we introduce the following two baseline approaches in our study. These two baselines use the same generalized maximum entropy model as the proposed approach. They differ from the proposed approach in how to compute $\widehat{f}_j(y)$. In the first baseline, we estimate $\mathbf{u}_{\mathbf{x}|y^t}[f(\mathbf{x})]$ by a weighted combination of $\mathbf{u}_{\mathbf{x}|y^k}[f(\mathbf{x})]$, i.e.,

$$\mathbf{u}_{\mathbf{x}|y^t=1}[f(\mathbf{x})] = \frac{1}{K} \sum_{y^k \in \{0,1\}} p(y^k|y^t = 1)\mathbf{u}_{\mathbf{x}|y^k}[f(\mathbf{x})],$$

and compute $\widehat{f}_j(y)$ using (6) and (4). In the second approach, we first predict the assignments of the target class c_t for the training examples by a weighted combination of the auxiliary classes in $\mathcal{C}$, i.e.,

$$\widehat{y}_i^t = I\left(\frac{\sum_{k=1}^K y_i^k p(y^t = 1|y^k = 1)}{\sum_{k'=1} p(y^t = 1|y^k = 1)} > p(y^t = 1)\right), i = 1, \ldots, n$$

where $I(z)$ is an indicator function that outputs 1 if z is true and zero, otherwise. We then compute $\widehat{f}_j(y)$ based on the predictions $\widehat{y}_i^t, i = 1, \ldots, n$, i.e.,

$$\widehat{f}_j(y) = \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i)\delta(\widehat{y}_i^t, y)$$

We refer to the first approach as the generalized maximum entropy model that estimates the expectation by average, or

[2] http://bailando.sims.berkeley.edu/enron_email

[3] Note that we do not have the assignments of the training documents to the target class

Table 2: Comparison with baselines (AUC)

		tmc2007	enron	bibtex	delicious
Training	GME-Reg	**0.8270**	**0.7376**	**0.8832**	**0.7552**
	GME-avg	0.4379	0.4379	0.4740	0.4779
	cModel	0.7506	0.6512	0.8500	0.6307
	cLabel	0.6311	0.6280	**0.8771**	0.7086
Testing	GME-Reg	**0.8092**	**0.6741**	**0.8625**	**0.7244**
	GME-avg	0.4501	0.4326	0.4831	0.4776
	cModel	0.7273	0.6096	0.8266	0.6171
	cLabel	0.6307	0.5783	**0.8583**	0.6923
All	GME-Reg	**0.8224**	**0.6985**	**0.8760**	**0.7492**
	GME-avg	0.4421	0.4660	0.4775	0.4778
	cModel	0.7437	0.5779	0.8417	0.6280
	cLabel	0.6312	0.6072	**0.8705**	0.7052

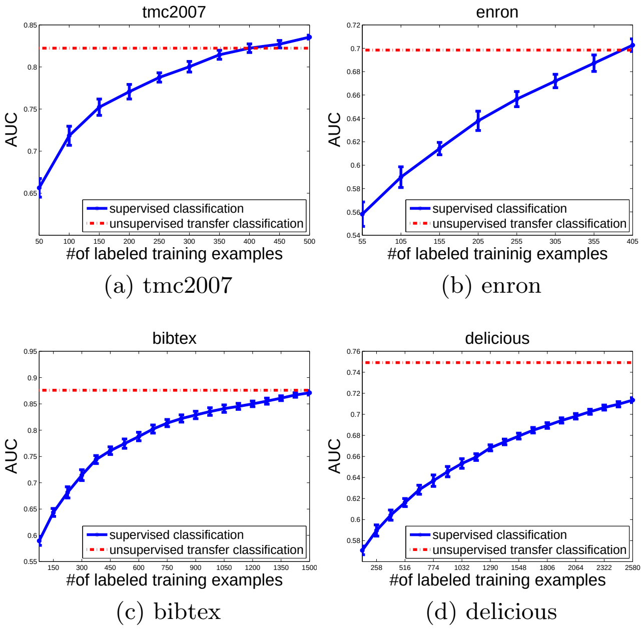
GME-avg for short, and the second one as the combination of class labels, or **cLabel** for short. Since these two baselines differ from the proposed approach only in computing $\widehat{f_j}(y)$, a comparison to these two baselines will show if the proposed approach for computing $\widehat{f_j}(y)$ is effective for unsupervised transfer classification. Finally, we refer to our method as the generalized maximum entropy model that estimates the expectation by regression, or **GME-Reg** for short.

4.2 Comparison with Baselines

We compare our method with the three baseline methods. The class information, i.e. the conditional probabilities $p(y^t = 1|y^k = 1)$ and the class prior $p(y^t = 1)$, are estimated from the training data. Table 2 summarizes the result of AUC for training data, testing data, and for all the data that includes both training data and testing data. Note that since we only have the assignments of the auxiliary classes for the training data, it is therefore valuable to evaluate the classification accuracy of the target class for the training data. It is not surprising to observe that for all methods in comparison, their performance for training data is in general better than that for testing data. We also observe that for all the cases, the proposed method **GME-Reg** outperforms the baseline methods significantly (student-t test at 95% significance level) except for **cLabel** on "bibtex". Our result also reveals that the proposed algorithm is computationally efficient: on a 2.0GHz CPU, 2.0GB memory linux server, the averaged running time of the optimization algorithm is 23 seconds for "tmc", 0.97 seconds for "enron", 11 seconds for "bibtex", and 15 seconds for "delicious" when the convergence accuracy is set as 10^{-4}.

4.3 Comparison with Supervised Classification

In this experiment, we compare the unsupervised transfer classification to the *supervised classification* using the generalized maximum entropy model. The objective of this comparison is to measure the amount of label information transferred from the auxiliary classes to the target class. In particular, to find the number of labeled examples that are needed to achieve the same performance as the unsupervised transfer classification, we increase the number of labeled examples for the target class in supervised classification. Figure 2 shows the result of AUC for *all* data (i.e., training data + testing data) for both the supervised and the unsupervised classification approaches. To ensure the robustness of our result, for supervised classification, we repeat each experiment five times and report AUC averaged over five runs. We observe that the label information transferred from the auxiliary classes is indeed significant: for data sets "tmc2007" and "enron", the amount of information transferred from the auxiliary classes is equivalent to a few hundred labeled examples; for "bibtex" and "delicious", it is more valuable and is equivalent to a few thousand labeled examples.



(a) tmc2007

(b) enron

(c) bibtex

(d) delicious

Figure 2: Comparison of unsupervised transfer classification to supervised classification with increasing number of labeled training examples

4.4 Estimating Class Information Using External Sources

In this experiment, we evaluate the proposed approach with the class information estimated from external sources. We choose "bibtex" and "delicious" data sets for evaluation since the class names are given in these two data sets. By removing the classes in these data sets that are not meaningful, such as "2005", "2006", "and", "of" , "?", "??" etc, we finally obtain a total of 123 classes in "bibtex" and 920 classes in "delicious". Two external sources are used for obtaining the class information for data set "bibtex": (i) the social bookmark and publication sharing web site bibsonomy [4], referred to as **bib.org** for short, and (ii) ACM digital library [5], referred to as **acm.org** for short. The external source for obtaining class information for "delicious" data set is the delicious social bookmarking [6] web site, referred to as **deli.com** for short. We obtain the class information by sending queries to the external sources that consist of the class name(s), and computing the class information based on the number of returned entities. The conditional probability is computed as $p(y^t = 1|y^k = 1) = \#ENT(c_t, c_k)/\#ENT(c_k)$, where $\#ENT(c_t, c_k)$ is the number of entities tagged by

[4] `http://www.bibsonomy.org/tags/`

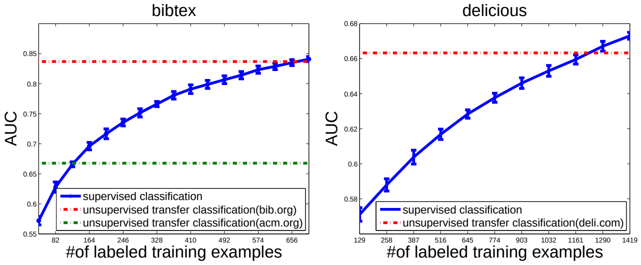
[5] `http://portal.acm.org/`

[6] `http://delicious.com/tag/`

Table 3: Classification Accuracy(AUC) using external class information for "bibtex" and "delicious"

	bibtex			delicious	
source	data	bib.org	acm.org	data	deli.com
GME-Reg	**0.8662**	**0.8369**	**0.6677**	**0.7525**	**0.6633**
GME-avg	0.3995	0.4727	0.4700	0.4755	0.4770
cModel	0.8148	0.7949	0.6044	0.6245	0.5936
cLabel	0.8290	0.7884	0.5427	0.7033	0.5449

both class c_t and c_k and $\#ENT(c_k)$ is the number of entities tagged by c_k. The class prior $p(y^t = 1)$ is estimated as $\#ENT(c_t)/\#ENT$, where $\#ENT$ is the total number of entities in the external source. However, since the total number of entities is unavailable for bib.org and deli.com, we replace $\#ENT$ with the sum of entities in all the classes, i.e., $\#ENT(c_t)+\sum_{k=1}^{K}\#ENT(c_k)$. The AUC for *all* (training+testing) examples is shown in Table 3. For the convenience of comparison, we also include in Table 3 the AUC results using the class information estimated from the data set itself. We observe that the proposed approach still yields reasonably accurate prediction even using class information estimated from the external sources. It is not surprising that using the class information estimated from bib.org, the proposed approach yields better performance on "bibtex" than using the class information estimated from acm.org because "bibtex" is actually collected from bib.org. By comparing to the result of supervised classification in Figure 3, we find that for "bibtex", the amount of information transferred from the auxiliary classes is equivalent to 600 labeled examples when using the class information estimated from bib.org, and 100 labeled examples when using the class information estimated from acm.org. For "delicious", the equivalent number of labeled examples is over $1,000$ when using deli.com as the external source to estimate the class information. These results further confirm the value of unsupervised transfer classification even with rough estimates of the class information from external sources.



(a) bibtex

(b) delicious

Figure 3: Comparison of unsupervised transfer classification to supervised classification with increasing number of labeled training examples

5. CONCLUSIONS

We have considered the challenging problem of unsupervised transfer classification whose goal is to build the classification model for a target class not by its labeled examples but by leveraging the label information of auxiliary classes. We propose a framework based on the generalized maximum entropy model that effectively transfers the label information of the auxiliary classes to the target class. We present efficient algorithm for solving the related optimization problem and consistency analysis for the solution obtained by the proposed approach. Extensive empirical studies show the promising performance of the framework for unsupervised transfer classification. In the future, we plan to investigate the performance of the proposed approach on different tasks such as image annotation and with different means of estimating the class information such as using the WordNet.

Acknowledgement

This work was supported in part by National Science Foundation (IIS-0643494), Office of Naval Research (N00014-09-1-0663), and Army Research Office (W911NF-09-1-0421). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, ONR and ARO. Part of this research was supported by WCU(World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology(R31-2008-000-10008-0) to Korea University.

6. REFERENCES

[1] Y. Altun and A. J. Smola. Unifying divergence minimization and statistical inference via convex duality. In *COLT*, 2006.

[2] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *NIPS*, 2007.

[3] A. Argyriou, C. A. Micchelli, M. Pontil, and Y. Ying. A spectral regularization framework for multi-task structure learning. In *NIPS*, 2007.

[4] E. Bonilla, K. M. Chai, and C. Williams. Multi-task gaussian process prediction. In *NIPS*, 2008.

[5] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41:1–58, 2009.

[6] S. F. Chen and R. Rosenfeld. A gaussian prior for smoothing maximum entropy models. Technical report, 1999.

[7] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. Boosting for transfer learning. In *ICML*, 2007.

[8] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. Self-taught clustering. In *ICML*, 2008.

[9] T. Jebara, R. Kondor, A. Howard, K. Bennett, and N. Cesa-bianchi. Probability product kernels. *Journal of Machine Learning Research*, 5:819–844, 2004.

[10] S. Ji, L. Tang, S. Yu, and J. Ye. Extracting shared subspace for multi-label classification. In *SIGKDD*, 2008.

[11] J. Jiang and C. Zhai. Instance weighting for domain adaptation in nlp. In *ACL*, 2007.

[12] Y. Jin, L. Khan, L. Wang, and M. Awad. Image annotations by combining multiple evidence & wordnet. In *ACMMM*, 2005.

[13] F. Kang, R. Jin, and R. Sukthankar. Correlated label propagation with application to multi-label learning. In *CVPR*, 2006.

[14] I. Katakis, G. Tsoumakas, and I. Vlahavas. Multilabel text classification for automated tag suggestion. In *ECML/PKDD Discovery Challenge*, 2008.

[15] H. Kück and N. de Freitas. Learning about individuals from group statistics. In *UAI*, 2005.

[16] N. D. Lawrence and J. C. Platt. Learning to learn with the informative vector machine. In *ICML*, 2004.

[17] X. Liao, Y. Xue, and L. Carin. Logistic regression with an auxiliary data source. In *ICML*, 2005.

[18] A. Mikheev and R. Mooney. Feature lattices and maximum entropy models, 1998.

[19] A. Nemirovski. Efficient methods in convex programming. 1994.

[20] K. Nigam, J. Lafferty, and A. Mccallum. Using maximum entropy for text classification, 1999.

[21] S. J. Pan and Q. Yang. A survey on transfer learning. Technical report, 2008.

[22] N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le. Estimating labels from label proportions. In *ICML*, 2008.

[23] A. Ratnaparkhi, J. Reynar, and S. Roukos. A maximum entropy model for prepositional phrase attachment. In *ARPA Workshop on Human Language Technology*, 1994.

[24] R. Rosenfeld. A maximum entropy approach to adaptive statistical language modeling. *Computer, Speech and Language*, 10:187–228, 1996.

[25] G. Tsoumakas, I. Katakis, and I. Vlahavas. Effective and efficient multilabel classification in domains with large number of labels. In *ECML/PKDD 2008 Workshop on Mining Multidimensional Data*, 2008.

[26] N. Ueda and K. Saito. Parametric mixture models for multi-labeled text. In *NIPS*, 2003.

[27] Q. Yang, Y. Chen, G. Xue, W. Dai, and Y. Yu. Heterogeneous transfer learning for image clustering via the social web. In *ACL*, 2009.

[28] T. Yang, R. Jin, and A. K. Jain. Learning from noisy side information by generalized maximum entropy model. In *ICML*, 2010.

[29] K. Yu, S. Yu, and V. Tresp. Multi-label informed latent semantic indexing. In *SIGIR*, 2005.

[30] S. Zhu, X. Ji, W. Xu, and Y. Gong. Multi-labelled classification using maximum entropy method. In *SIGIR*, 2005.

[31] X. Zhu. Semi-supervised learning literature survey, 2006.

APPENDIX

A. PROOF OF THEOREM 1

PROOF. The theorem can be proved by using McDiarmid's inequality. Considering the quantity of $\delta(Y,y)f_j(X)$, the expectation of the quantity is equal to

$$
\begin{aligned}
\mathrm{E}_{Y,X\sim P(Y,X)}\left[\delta(y,Y)f_j(X)\right] &= \mathrm{E}_X\mathrm{E}_{Y|X}[\delta(y,Y)f_j(X)] \\
&= \mathrm{E}_X[\Pr(y|X)f_j(X^1,X^2)]
\end{aligned}
$$

So we can see that $\frac{1}{n}\sum_{i=1}^{n}\delta(y,y_i)f_j(\mathbf{x}_i)$, $\frac{1}{n}\sum_{i=1}^{n}p(y|\mathbf{x}_i)f_j(\mathbf{x}_i)$ are the empirical estimates of above expectations, and under i.i.d. assumption we have

$$
\begin{aligned}
\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\delta(y_i,y)f_j(\mathbf{x}_i)\right] &= \mathrm{E}\left[\delta(Y,y)f_j(X)\right] \\
\mathrm{E}\left[\frac{1}{n}\sum_{i=1}p(y|\mathbf{x}_i)f_j(\mathbf{x}_i)\right] &= \mathrm{E}\left[\Pr(Y=y|X)f_j(X)\right] \\
&= \mathrm{E}[\delta(Y,y)f_j(X)]
\end{aligned}
$$

Following McDiarmid's inequality, we have

$$
\begin{aligned}
&\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\delta(y_i,y)f_j(\mathbf{x}_i)-\mathrm{E}\left[\delta(Y,y)f_j(X)\right]\right|\geq\epsilon\right) \\
&\leq 2\exp\left(-\frac{\epsilon^2 n}{2R_j^2}\right) \\
&\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}p(y|\mathbf{x}_i)f_j(\mathbf{x}_i)-\mathrm{E}\left[\delta(Y,y)f_j(X)\right]\right|\geq\epsilon\right) \\
&\leq 2\exp\left(-\frac{\epsilon^2 n}{2R_j^2}\right)
\end{aligned}
$$

which imply

$$
\begin{aligned}
&\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\delta(y_i,y)f_j(\mathbf{x}_i)-\frac{1}{n}\sum_{i=1}^{n}p(y|\mathbf{x}_i)f_j(\mathbf{x}_i)\right|\geq 2\epsilon\right) \\
&\leq 4\exp\left(-\frac{\epsilon^2 n}{2R_j^2}\right)
\end{aligned}
$$

Replacing ϵ with $\frac{1}{2}\epsilon$ we complete the proof. □

B. PROOF OF THEOREM 2

Due to limited space, we sketch the proof for theorem 2. First, we bound $U_{\mathbf{x}|y^t}$ from $\mathrm{E}_{X|Y^t}$ in terms of $\|\mathrm{E}_X[f(X)]-\mathbf{u}_\mathbf{x}[f(\mathbf{x})]\|_2$ and $\|\bar{\mathbf{A}}-\widehat{\mathbf{A}}\|_F$, where $\bar{\mathbf{A}}$ is the true expectations of estimates in $\widehat{\mathbf{A}}$, as stated in the following lemma.

LEMMA 1. *Under assumption **A1**, we have*

$$
\begin{aligned}
&\|\mathrm{E}_{X|Y^t}[f(X)]-U_{\mathbf{x}|y^t}\|_F \leq \\
&\frac{\sqrt{2}(1+\kappa^{3/2})}{\sqrt{\sigma_{\min}}}\|\bar{\mathbf{A}}-\widehat{\mathbf{A}}\|_F+\frac{\kappa}{\|\mathbf{p}_t\|_2}\|\mathrm{E}_X[f(X)]-\mathbf{u}_\mathbf{x}[f(\mathbf{x})]\|_2
\end{aligned}
$$

PROOF. First under assumption **A1**, we have similar solution for $\mathrm{E}_{X|Y^t}[f(X)]$

$$
\begin{aligned}
&\mathrm{E}_{X|Y^t}[f(X)] = (\mathbf{W}^\top\mathbf{W})^{-1}\cdot \\
&\left(\frac{\mathbf{p}_t\mathrm{E}_X^\top[f(X)]}{\mathbf{p}_t^\top(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{p}_t}+(\mathbf{I}-\frac{\mathbf{p}_t\mathbf{p}_t^\top(\mathbf{W}^\top\mathbf{W})^{-1}}{\mathbf{p}_t^\top(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{p}_t})\mathbf{W}^\top\bar{\mathbf{A}}\right)
\end{aligned}
$$

Then we have

$$\begin{aligned}
&\|\mathrm{E}_{X|Y^t}[f(X)] - U_{\mathbf{x}|y^t}\|_F \leq \\
&\left\|(\mathbf{W}^\top\mathbf{W})^{-1}\frac{\mathbf{p}_t(\mathrm{E}_X[f(X)] - \mathbf{u}_\mathbf{x}[f(\mathbf{x})])^\top}{\mathbf{p}_t^\top(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{p}_t}\right\|_F \\
&+ \left\|(\mathbf{W}^\top\mathbf{W})^{-1}(\mathbf{I} - \frac{\mathbf{p}_t\mathbf{p}_t^\top(\mathbf{W}^\top\mathbf{W})^{-1}}{\mathbf{p}_t^\top(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{p}_t})\mathbf{W}^\top(\bar{\mathbf{A}} - \widehat{\mathbf{A}})\right\|_F \\
&\leq \frac{\kappa}{\|\mathbf{p}_t\|_2}\|\mathrm{E}_X[f(X)] - \mathbf{u}_\mathbf{x}[f(\mathbf{x})]\|_2 \\
&+ \left(\sqrt{2\sigma_{\min}^{-1}} + \frac{\sigma_{\min}^{-2}}{\sigma_{\max}^{-1}}\sqrt{2\sigma_{\max}}\right)\|\bar{\mathbf{A}} - \widehat{\mathbf{A}}\|_F \\
&= \frac{\sqrt{2}(1+\kappa^{3/2})}{\sqrt{\sigma_{\min}}}\|\bar{\mathbf{A}} - \widehat{\mathbf{A}}\|_F + \frac{\kappa}{\|\mathbf{p}_t\|_2}\|\mathrm{E}_X[f(X)] - \mathbf{u}_\mathbf{x}[f(\mathbf{x})]\|_2
\end{aligned}$$

where we use the fact $\|\mathbf{p}_t\|_2^2\sigma_{\max}^{-1} \leq \|\mathbf{p}_t(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{p}_t\| \leq \|\mathbf{p}_t\|_2^2\sigma_{\min}^{-1}$, $\|\mathbf{W}\|_F \leq \sqrt{2\sigma_{\max}}$, and $\|\mathbf{W}^{-1}\|_F \leq \sqrt{2\sigma_{\min}^{-1}}$. □

Lemma 2. *Assume bounded feature function $f_j(\mathbf{x})$, i.e., $|f_j(\mathbf{x})| \leq R, j = 1, \ldots, d$, with at least probability $1-\delta$, we have*

$$\|\mathrm{E}_X[f(X)] - \mathbf{u}_\mathbf{x}[f(\mathbf{x})]\|_2 \leq \sqrt{\frac{2dR^2}{n}\ln\left(\frac{2d}{\delta}\right)}$$

This lemma can be proved by McDiarmid's inequality and union bound.

Lemma 3. *Assume bounded feature function $f_j(\mathbf{x})$, i.e., $|f_j(\mathbf{x})| \leq R, j = 1, \ldots, d$, and the prior for all auxiliary classes are significantly large, i.e. there exists some positive constant $\rho > 0$ such that $p(y^k = 1) \geq \rho, k = 1, \ldots, K$, for any $\delta > 5Kd\exp(-n\rho^2/4)$, with probability at least $1-\delta$, we have*

$$\left\|\bar{\mathbf{A}} - \widehat{\mathbf{A}}\right\|_F \leq \frac{4R}{\rho}\sqrt{\frac{Kd}{n}\ln\left(\frac{5Kd}{\delta}\right)}$$

Proof. To prove this lemma, we first show the bound for each element in $\widehat{\mathbf{A}}$, i.e. $\mathbf{u}_{\mathbf{x}|y^k=1}[f_j(\mathbf{x})]$ as defined in (7) from the true conditional expectation $\mathrm{E}_{X|Y^k=1}[f_j(X)]$. Note that

$$\begin{aligned}
&|\mathbf{u}_{\mathbf{x}|y^k=1}[f_j(\mathbf{x})] - \mathrm{E}_{X|Y^k=1}[f_j(X)]| \\
&= \left|\frac{\sum_i \delta(y_i^k, 1)f_j(\mathbf{x}_i) - \sum_i \delta(y_i^k, 1)\mathrm{E}_{X|Y^k=1}[f_j(X)]}{\sum_i \delta(y_i^k, 1)}\right|
\end{aligned}$$

We can bound both the numerator denoted by d_{uE}^k and the denominator in the above equation with McDiarmid's inequality,

$$\begin{aligned}
&\Pr\left(\left|\frac{1}{n}d_{uE}^k\right| \geq 2R\epsilon\right) \leq 4\exp\left(-2\epsilon^2 n\right) \\
&\Pr\left(\frac{1}{n}\sum_i \delta(y_i^k, 1) \leq p(y^k = 1) - \epsilon\right) \leq \exp(-2\epsilon^2 n)
\end{aligned}$$

Then with union bound, we have

$$\begin{aligned}
&\Pr\left(|\mathbf{u}_{\mathbf{x}|y^k=1}[f_j(\mathbf{x})] - \mathrm{E}_{X|Y^k=1}[f_j(X)]| \leq \frac{2R\epsilon}{p(y_k = 1) - \epsilon}\right) \\
&\geq 1 - 5\exp(-\epsilon^2 n)
\end{aligned}$$

Since $p(y^k = 1) \geq \rho$, with $\epsilon \leq \rho/2$, we have

$$\begin{aligned}
&\Pr\left(|\mathbf{u}_{\mathbf{x}|y^k=1}[f_j(\mathbf{x})] - \mathrm{E}_{X|Y^k=1}[f_j(X)]| \leq \frac{4R\epsilon}{\rho}\right) \\
&\geq 1 - 5\exp(-\epsilon^2 n)
\end{aligned}$$

Again applying the union bound, we have

$$\Pr\left(\|\bar{\mathbf{A}} - \widehat{\mathbf{A}}\|_F^2 \leq Kd\left(\frac{4R\epsilon}{\rho}\right)^2\right) \geq 1 - 5Kd\exp(-\epsilon^2 n)$$

Let $\delta = 5Kd\exp(-\epsilon^2 n)$, then with probability $1-\delta$, we have the lemma 3. □

Combining the above lemmas together, we complete the proof for theorem 2.

C. PROOF OF THEOREM 3

Let

$$\begin{aligned}
L(\lambda) &= \frac{1}{n}\sum_i \log(\exp(\lambda_1^\top f(\mathbf{x}_i)) + \exp(\lambda_0^\top f(\mathbf{x}_i))) \\
&\quad - \left(\lambda_1^\top, \lambda_0^\top\right)\begin{pmatrix}\widehat{\mathbf{f}}_1^* \\ \widehat{\mathbf{f}}_0^*\end{pmatrix} + \frac{\gamma}{2}\lambda_1^\top\lambda_1 + \frac{\gamma}{2}\lambda_0^\top\lambda_0 \\
&= g(\lambda) - \lambda^\top\begin{pmatrix}\widehat{\mathbf{f}}_1^* \\ \widehat{\mathbf{f}}_0^*\end{pmatrix} + \frac{\gamma}{2}\|\lambda\|_2^2
\end{aligned}$$

where $\lambda = \begin{pmatrix}\lambda_1 \\ \lambda_0\end{pmatrix}$, $g(\lambda)$ is the sum of log-exponential function of λ, which is convex in λ. Assume λ^* is the optimal solution to minimizing $L(\lambda)$, λ^o is the optimal solution to minimizing $L(\lambda)$ with $\begin{pmatrix}\widehat{\mathbf{f}}_1^* \\ \widehat{\mathbf{f}}_0^*\end{pmatrix}$ replaced by $\begin{pmatrix}\widehat{\mathbf{f}}_1^o \\ \widehat{\mathbf{f}}_0^o\end{pmatrix}$, then we have

$$\begin{aligned}
L(\lambda^o) &\geq L(\lambda^*) + \nabla L(\lambda^*)^\top(\lambda^o - \lambda^*) + \frac{\gamma}{2}\|\lambda^o - \lambda^*\|_2^2 \\
&\geq L(\lambda^*) + \frac{\gamma}{2}\|\lambda^o - \lambda^*\|_2^2
\end{aligned}$$

where we use the fact that $L(\cdot)$ is a c_r-strongly convex function, and the optimality criterion that $\nabla L(\lambda^*)^\top(\lambda^o - \lambda^*) \geq 0$. Then

$$\begin{aligned}
L(\lambda^o) &= g(\lambda^o) - \lambda^{o\top}\begin{pmatrix}\widehat{\mathbf{f}}_1^* \\ \widehat{\mathbf{f}}_0^*\end{pmatrix} + \frac{\gamma}{2}\|\lambda^o\|_2^2 \\
&= g(\lambda^o) - \lambda^{o\top}\begin{pmatrix}\widehat{\mathbf{f}}_1^o \\ \widehat{\mathbf{f}}_0^o\end{pmatrix} + \frac{\gamma}{2}\|\lambda^o\|_2^2 + \lambda^{o\top}\begin{pmatrix}\widehat{\mathbf{f}}_1^o - \widehat{\mathbf{f}}_1^* \\ \widehat{\mathbf{f}}_0^o - \widehat{\mathbf{f}}_0^*\end{pmatrix} \\
&\leq g(\lambda^*) - \lambda^{*\top}\begin{pmatrix}\widehat{\mathbf{f}}_1^o \\ \widehat{\mathbf{f}}_0^o\end{pmatrix} + \frac{\gamma}{2}\|\lambda^*\|_2^2 + \lambda^{o\top}\begin{pmatrix}\widehat{\mathbf{f}}_1^o - \widehat{\mathbf{f}}_1^* \\ \widehat{\mathbf{f}}_0^o - \widehat{\mathbf{f}}_0^*\end{pmatrix} \\
&\leq g(\lambda^*) - \lambda^{*\top}\begin{pmatrix}\widehat{\mathbf{f}}_1^* \\ \widehat{\mathbf{f}}_0^*\end{pmatrix} + \frac{\gamma}{2}\|\lambda^*\|_2^2 + (\lambda^o - \lambda^*)^\top\begin{pmatrix}\widehat{\mathbf{f}}_1^o - \widehat{\mathbf{f}}_1^* \\ \widehat{\mathbf{f}}_0^o - \widehat{\mathbf{f}}_0^*\end{pmatrix} \\
&\leq L(\lambda^*) + \|\lambda^o - \lambda^*\|_2\|\widehat{\mathbf{f}}^* - \widehat{\mathbf{f}}^o\|_F
\end{aligned}$$

Coming the above two bounds together, we have

$$\frac{\gamma}{2}\|\lambda^o - \lambda^*\|_2^2 \leq \|\lambda^o - \lambda^*\|_2\|\widehat{\mathbf{f}}^* - \widehat{\mathbf{f}}^o\|_F$$

i.e.,

$$\|\lambda^* - \lambda^o\|_2 \leq \frac{2}{\gamma}\|\widehat{\mathbf{f}}^* - \widehat{\mathbf{f}}^o\|_F$$