EFFICIENT INSTANCE ANNOTATION IN MULTI-INSTANCE LEARNING

Anh T. Pham, Raviv Raich, and Xiaoli Z. Fern

School of EECS, Oregon State University, Corvallis, OR 97331-5501
{phaman, raich, xfern}@eecs.oregonstate.edu

ABSTRACT

The cost associated with manually labeling every individual instance in large datasets is prohibitive. Significant labeling efforts can be saved by assigning a collective label to a group of instances (a bag). This setup prompts the need for algorithms that allow labeling individual instances (instance annotation) based on bag-level labels. Probabilistic models in which instance-level labels are latent variables can be used for instance annotation. Brute-force computation of instance-level label probabilities is exponential in the number of instances per bag due to marginalization over all possible combinations. Existing solutions for addressing this issue include approximate methods such as sampling or variational inference. This paper proposes a discriminative probability model and an expectation maximization procedure for inference to address the instance annotation problem. A key contribution is a dynamic programming solution for exact computation of instance probabilities in quadratic time. Experiments on bird song, image annotation, and two synthetic datasets show a significant accuracy improvement by 4%-14% over a recent state-of-the-art rank loss SIM method.

***Index Terms*—** Multi-instance learning, discriminative model, expectation maximization, logistic regression, dynamic programming

1. INTRODUCTION

Multiple instance learning (MIL) is a framework for representing complex objects (bags) with multiple feature vectors (instances). For example, images are represented as a collection of segments and documents are represented as a collection of paragraphs. To reduce labeling effort, often a bag is provided with a bag-level label instead of instance-level labels. For example, instead of providing appropriate labels for all segments of an image, the image itself is tagged with the union of its instance labels. This setting gives rise to the study of the multiple instance multiple label learning (MIML) problem [1].

One problem in the MIML framework is how to predict bag-level labels for unseen bags. Another problem is how to learn a classifier that can predict instance-level labels, which is referred to as the instance annotation problem and is the focus of this paper. This problem introduces a challenge: the training data does not contain instance-level labels, only bag-level labels.

To address the instance annotation problem, M^{3}MIML [1] computes a bag-level score function by taking the maximum over the instance-level scores. This principle may ignore useful information from other instances in the bag. The rank loss support instance machine (rank loss SIM) [2] introduces a softmax score function that takes into account all instances with the weight corresponding to their normalized scores. However, both [1] and [2] have no mechanism to model the dependency between labels of instances given the label of their bag.

Some probabilistic graphical models have been proposed for instance annotation. To avoid computationally complex inference, approximate methods such as sampling [3] or variational inference [4] have been proposed. In [5], a *generative* model for the MIML problem and a tractable expectation maximization method are proposed. The authors postulate that when large amounts of labeled data are available, accuracy can be improved using discriminative models. To improve on the shortcomings of the aforementioned algorithms, we propose a *discriminative* probability model with exact inference method.

Our contribution in this paper is two-fold. First, we propose a discriminative model based on logistic regression for the instance annotation problem. Second, in the inference phase, we propose an expectation maximization framework to maximize the log-likelihood. To compute the posterior probability for the label of every instance, we propose a novel dynamic programming approach which reduces the run time from exponential to quadratic in the number of instances per bag thus enabling exact inference for the problem. Experiments on bird song, image annotation as well as synthetic datasets show a significant improvement, at around 4%-14%, in accuracy of our algorithm compared to the rank loss SIM method [2].

2. PROBLEM FORMULATION

We consider the instance annotation problem in the MIML framework. In this problem, the available training data consists of B bags of multiple instances, as $\{(\mathbf{X}_b, \mathbf{Y}_b)\}_{b=1}^{B}$. In this notation, $\mathbf{X}_b$ is a set consisting of n_b instances $\{\mathbf{x}_{b1}, \mathbf{x}_{b2}, \ldots, \mathbf{x}_{bn_b}\}$, where $\mathbf{x}_{bi} \in \mathbb{R}^d$ represents the feature vector of the ith instance in the bth bag. In addition, the bag-level label $\mathbf{Y}_b$ is a subset of $\mathcal{Y} = \{1, 2, \ldots, C\}$, the instance-level label set. Our goal is to train a classifier that maps an instance in $\mathbb{R}^d$ to a label in $\mathcal{Y}$ using only the aforementioned bag-level labeled multi-instance data.

2.1. The proposed model: ORed-logistic regression

We consider a discriminative probability model as in Fig. 1, which we denote as the ORed logistic regression (ORLR) model. We assume that the labeled bags $(\mathbf{X}_b, \mathbf{Y}_b)$ are generated independently, and that an instance label $\mathbf{y}_{bi}$ is generated based on the instance feature vector $\mathbf{x}_{bi}$ using the logistic regression model

$$p(\mathbf{y}_{bi}|\mathbf{x}_{bi}, \mathbf{w}) = \frac{\prod_{c=1}^{C} e^{I(\mathbf{y}_{bi}=c)\mathbf{w}_c^T \mathbf{x}_{bi}}}{\sum_{c=1}^{C} e^{\mathbf{w}_c^T \mathbf{x}_{bi}}}, \quad (1)$$

where $\mathbf{w}_c \in \mathbb{R}^d$ is the weight for the cth class and $\mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_C]$. Furthermore, we assume that the label of a bag is equal to

This work is partially supported by the National Science Foundation grants CCF-1254218 and IIS-1055113.

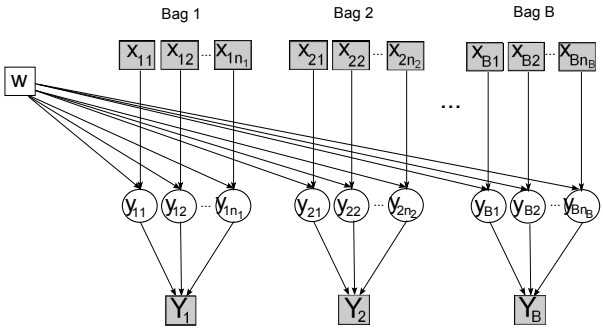

Fig. 1: Graphical model for our instance annotation problem where the observation is shaded.

the union of its instance labels. Consequently, the probability of the bag-level label given the instance-level labels of the bth bag is

$$p(\mathbf{Y}_b|\mathbf{y}_{b1}, \mathbf{y}_{b2}, \ldots, \mathbf{y}_{bn_b}) = \Psi(\mathbf{Y}_b, \mathbf{y}_{b1}, \ldots, \mathbf{y}_{bn_b}), \quad (2)$$

where $\Psi(\mathbf{Y}_b, \mathbf{y}_{b1}, \ldots, \mathbf{y}_{bn_b})$ is 1 if $\mathbf{Y}_b = \bigcup_{i=1}^{n_b} \mathbf{y}_{bi}$ and 0 otherwise.

2.2. Maximum Likelihood

We consider the maximum likelihood framework for inference. To simplify the notation, we use $(\mathbf{X}_D, \mathbf{Y}_D)$ to denote $\{(\mathbf{X}_b, \mathbf{Y}_b)\}_{b=1}^{B}$. Our goal is to estimate the model parameters $\mathbf{w}$ given observations $\mathbf{X}_D$ and $\mathbf{Y}_D$ by maximizing the likelihood given by $p(\mathbf{X}_D, \mathbf{Y}_D|\mathbf{w})$. Since $\mathbf{X}_D$ and $\mathbf{w}$ are independent, $p(\mathbf{X}_D|\mathbf{w}) = p(\mathbf{X}_D)$. Instead of maximizing $p(\mathbf{Y}_D, \mathbf{X}_D|\mathbf{w})$, we maximize $p(\mathbf{Y}_D|\mathbf{X}_D, \mathbf{w})$ since $p(\mathbf{Y}_D, \mathbf{X}_D|\mathbf{w}) = p(\mathbf{Y}_D|\mathbf{X}_D, \mathbf{w})p(\mathbf{X}_D)$ is dependent on the parameter $\mathbf{w}$ only through $p(\mathbf{Y}_D|\mathbf{X}_D, \mathbf{w})$. The probability of the observed bag-level label $\mathbf{Y}_D$ given the observed instance feature vector $\mathbf{X}_D$ and the unknown parameter $\mathbf{w}$ is computed as follows

$$\begin{aligned} p(\mathbf{Y}_D|\mathbf{X}_D, \mathbf{w}) &= \prod_{b=1}^{B} p(\mathbf{Y}_b|\mathbf{X}_b, \mathbf{w}) \quad (3) \\ &= \prod_{b=1}^{B} [\sum_{\mathbf{y}_{b1}=1}^{C} \cdots \sum_{\mathbf{y}_{bn_b}=1}^{C} p(\mathbf{Y}_b, \mathbf{y}_{b1}, \ldots, \mathbf{y}_{bn_b}|\mathbf{X}_b, \mathbf{w})] \\ &= \prod_{b=1}^{B} [\sum_{\mathbf{y}_{b1}=1}^{C} \cdots \sum_{\mathbf{y}_{bn_b}=1}^{C} \prod_{i=1}^{n_b} p(\mathbf{y}_{bi}|\mathbf{x}_{bi}, \mathbf{w}) \Psi(\mathbf{Y}_b, \mathbf{y}_{b1}, \ldots, \mathbf{y}_{bn_b})]. \end{aligned}$$

Note that the last step in (3) is based on two facts. First, all the instance-level labels $\mathbf{y}_{bi}$ are independent given their feature vectors $\mathbf{x}_{bi}$ and the parameter $\mathbf{w}$. Second, the label of the bag $\mathbf{Y}_b$ is deterministic given the labels of its instances $\{\mathbf{y}_{bi}\}_{i=1}^{n_b}$. Taking the logarithm on both sides of (3), we have the log-likelihood $L(\mathbf{Y}_D|\mathbf{X}_D, \mathbf{w}) = \log p(\mathbf{Y}_D|\mathbf{X}_D, \mathbf{w})$. Maximizing $L(\mathbf{Y}_D|\mathbf{X}_D, \mathbf{w})$, in the case of incomplete data where the labels of instances are unknown, is generally a difficult problem since, to the best of our knowledge, no closed-form solution exists. To maximize the log-likelihood, we propose an expectation maximization (EM) solution.

3. THE PROPOSED INFERENCE APPROACH

Expectation-maximization (EM) is an iterative algorithm for solving maximum likelihood (ML) [6]. Given the observed data $\mathbf{x}$, the ML estimator of the parameter θ is given by

$$\hat{\theta} = \operatorname*{argmax}_{\theta} l(\theta), \quad (4)$$

where $l(\theta) = \log p(\mathbf{x}|\theta)$. In EM [6], the hidden data $\mathbf{y}$ is introduced and the algorithm is based on the joint distribution of the hidden and observed data $p(\mathbf{x}, \mathbf{y}|\theta)$ as follows

- E-step: Compute $g(\theta, \theta^{(k)}) = E_{\mathbf{y}}[\log p(\mathbf{x}, \mathbf{y}|\theta)|\mathbf{x}, \theta^{(k)}]$
- M-step: $\theta^{(k+1)}$=$\operatorname{argmax}_{\theta}\, g(\theta, \theta^{(k)})$.

By using an auxiliary function $g(\theta, \theta')$ for the log-likelihood $l(\theta)$, the EM algorithm is guaranteed in each iteration to satisfy $l(\theta^{(k+1)}) \geq l(\theta^{(k)})$. In this paper, we use the generalized EM [7]. Instead of maximizing the auxiliary function $g(\theta, \theta^{(k)})$, we only require $\theta^{(k+1)}$ such that $g(\theta^{(k+1)}, \theta^{(k)}) \geq g(\theta^{(k)}, \theta^{(k)})$.

3.1. Expectation maximization for multi-instance learning

We proceed with the application of EM to our problem. The auxiliary function $g(\mathbf{w}, \mathbf{w}')$ is equal to $E_{\mathbf{y}}[\log p(\mathbf{Y}_D, \mathbf{y}|\mathbf{X}_D, \mathbf{w})|\mathbf{Y}_D, \mathbf{X}_D, \mathbf{w}']$. Based on the conditional rule and the i.i.d. assumption of the instance labels, we have

$$\begin{aligned} p(\mathbf{Y}_D, \mathbf{y}|\mathbf{X}_D, \mathbf{w}) &= p(\mathbf{y}|\mathbf{X}_D, \mathbf{w}) p(\mathbf{Y}_D|\mathbf{y}, \mathbf{X}_D, \mathbf{w}) \\ &= [\prod_{b=1}^{B} \prod_{i=1}^{n_b} p(\mathbf{y}_{bi}|\mathbf{x}_{bi}, \mathbf{w})] p(\mathbf{Y}_D|\mathbf{y}), \end{aligned} \quad (5)$$

where $p(\mathbf{Y}_D|\mathbf{y}) = \prod_{b=1}^{B} \Psi(\mathbf{Y}_b, \mathbf{y}_{b1}, \mathbf{y}_{b2}, \ldots, \mathbf{y}_{bn_b})$. Substituting (1) for $p(\mathbf{y}_{bi}|\mathbf{x}_{bi}, \mathbf{w})$ and taking the logarithm on both sides of (5), yields

$$\begin{aligned} \log p(\mathbf{Y}_D, \mathbf{y}|\mathbf{X}_D, \mathbf{w}) &= \sum_{b=1}^{B} \sum_{i=1}^{n_b} \sum_{c=1}^{C} I(\mathbf{y}_{bi} = c) \mathbf{w}_c^T \mathbf{x}_{bi} \quad (6) \\ &\quad - \sum_{b=1}^{B} \sum_{i=1}^{n_b} \log(\sum_{c=1}^{C} e^{\mathbf{w}_c^T \mathbf{x}_{bi}}) + \log p(\mathbf{Y}_D|\mathbf{y}). \end{aligned}$$

Finally, by taking the expectation based on the distribution of the hidden variables $\mathbf{y}_{bi}$ conditioned on the observed data $\mathbf{X}_D$ and $\mathbf{Y}_D$ given the parameter $\mathbf{w}'$, we obtain

$$\begin{aligned} g(\mathbf{w}, \mathbf{w}') &= E_{\mathbf{y}}[\log p(\mathbf{Y}_D, \mathbf{y}|\mathbf{X}_D, \mathbf{w})|\mathbf{Y}_D, \mathbf{X}_D, \mathbf{w}'] \quad (7) \\ &= \sum_{b=1}^{B} \sum_{i=1}^{n_b} [\sum_{c=1}^{C} p(\mathbf{y}_{bi} = c|\mathbf{Y}_b, \mathbf{X}_b, \mathbf{w}') \mathbf{w}_c^T \mathbf{x}_{bi} - \log(\sum_{c=1}^{C} e^{\mathbf{w}_c^T \mathbf{x}_{bi}})] + \zeta, \end{aligned}$$

where $\zeta = E_{\mathbf{y}}[\log p(\mathbf{Y}_D|\mathbf{y})|\mathbf{Y}_D, \mathbf{X}_D, \mathbf{w}']$ is a constant independent of $\mathbf{w}$. Based on (7), we obtain the following generalized EM iterations for the ORLR model:

- E-step: Compute $p(\mathbf{y}_{bi} = c|\mathbf{Y}_b, \mathbf{X}_b, \mathbf{w}^{(k)})$
- M-step: Find $\mathbf{w}^{(k+1)}$ such that $g(\mathbf{w}^{(k+1)}, \mathbf{w}^{(k)}) \geq g(\mathbf{w}^{(k)}, \mathbf{w}^{(k)})$.

3.1.1. The expectation step and our challenge

To compute $p(\mathbf{y}_{bi} = c|\mathbf{Y}_b, \mathbf{X}_b, \mathbf{w})$ from $p(\mathbf{y}_{bi} = c, \mathbf{Y}_b|\mathbf{X}_b, \mathbf{w})$, we apply the definition of conditional probability

$$p(\mathbf{y}_{bi} = c|\mathbf{Y}_b, \mathbf{X}_b, \mathbf{w}) = \frac{p(\mathbf{y}_{bi} = c, \mathbf{Y}_b|\mathbf{X}_b, \mathbf{w})}{\sum_{c=1}^{C} p(\mathbf{y}_{bi} = c, \mathbf{Y}_b|\mathbf{X}_b, \mathbf{w})}. \quad (8)$$

To compute $p(\mathbf{y}_{bi} = c, \mathbf{Y}_b|\mathbf{X}_b, \mathbf{w})$, we keep the label c for $\mathbf{y}_{bi}$ and

marginalize over all other instance labels as follows

$$p(\mathbf{y}_{bi} = c, \mathbf{Y}_b|\mathbf{X}_b, \mathbf{w}) \quad (9)$$
$$= \sum_{\mathbf{y}_{b1}=1}^{C} \cdots \sum_{\mathbf{y}_{bn_b}=1}^{C} [p(\mathbf{y}_{b1}, \ldots, \mathbf{y}_{bi} = c, \ldots, \mathbf{y}_{bn_b}|\mathbf{X}_b, \mathbf{w}) \times \Psi(\mathbf{Y}_b, \mathbf{y}_{b1}, \ldots, \mathbf{y}_{bi} = c, \ldots, \mathbf{y}_{bn_b})].$$

Note that in the summation, we exclude summing over $\mathbf{y}_{bi}$. Furthermore, computing $p(\mathbf{y}_{bi} = c, \mathbf{Y}_b|\mathbf{X}_b, \mathbf{w})$ is $O(C^{n_b-1})$, which is exponential in the number of instances per bag. Common approach to address such intractable problems in graphical models is by using approximate inference methods such as sampling [3] or variational inference [4]. However, to the best of our knowledge, a computationally efficient closed-form solution for (9) has not been proposed. We proceed with a computationally efficient dynamic programming method to solve this problem.

3.1.2. Dynamic programming forward algorithm for the E-step

To simplify the notation in this section, we only consider a bag $\mathbf{X}$ with its label $\mathbf{Y}$. Assuming a particular ordering of the instances in the bag, we denote a sub-bag containing from the 1st to the ith instance by $\mathbf{X}^i$, and the label of this sub-bag by $\mathbf{Y}^i$. For a bag label value $\mathbf{L}$, we use $\mathbf{L}_{\setminus c}$ to denote a label containing all of the labels in $\mathbf{L}$ except c.

In our model, $\mathbf{Y}^{i+1} = \bigcup_{j=1}^{i+1} \mathbf{y}_j = \mathbf{y}_{i+1} \bigcup \mathbf{Y}^i$. Therefore, for $c \notin \mathbf{L}$, $p(\mathbf{y}_{i+1} = c, \mathbf{Y}^{i+1} = \mathbf{L}|\mathbf{X}^{i+1}, \mathbf{w}) = 0$. For $c \in \mathbf{L}$, there are two mutually exclusive events that result in $(\mathbf{y}_{i+1} = c, \mathbf{Y}^{i+1} = \mathbf{L})$ including $(\mathbf{y}_{i+1} = c, \mathbf{Y}^i = \mathbf{L})$ and $(\mathbf{y}_{i+1} = c, \mathbf{Y}^i = \mathbf{L}_{\setminus c})$. Therefore, we have the following result when $c \in \mathbf{L}$

$$p(\mathbf{y}_{i+1} = c, \mathbf{Y}^{i+1} = \mathbf{L}|\mathbf{X}^{i+1}, \mathbf{w}) \quad (10)$$
$$= p(\mathbf{y}_{i+1} = c, \mathbf{Y}^i = \mathbf{L}|\mathbf{X}^{i+1}, \mathbf{w}) + p(\mathbf{y}_{i+1} = c, \mathbf{Y}^i = \mathbf{L}_{\setminus c}|\mathbf{X}^{i+1}, \mathbf{w}).$$

Since the labels of all the instances are independent, $\mathbf{y}_{i+1}$ and $\mathbf{Y}^i$ are independent given $\mathbf{X}^{i+1}$. Thus, we can rewrite (10) as follows

$$p(\mathbf{y}_{i+1} = c, \mathbf{Y}^{i+1} = \mathbf{L}|\mathbf{X}^{i+1}, \mathbf{w}) \quad (11)$$
$$= p(\mathbf{y}_{i+1} = c|\mathbf{x}_{i+1}, \mathbf{w})[p(\mathbf{Y}^i = \mathbf{L}|\mathbf{X}^i, \mathbf{w}) + p(\mathbf{Y}^i = \mathbf{L}_{\setminus c}|\mathbf{X}^i, \mathbf{w})].$$

Summing over all possible values of c in (11), we obtain

$$p(\mathbf{Y}^{i+1} = \mathbf{L}|\mathbf{X}^{i+1}, \mathbf{w}) = \sum_{c \in \mathbf{L}} p(\mathbf{y}_{i+1} = c|\mathbf{x}_{i+1}, \mathbf{w}) \times [p(\mathbf{Y}^i = \mathbf{L}|\mathbf{X}^i, \mathbf{w}) + p(\mathbf{Y}^i = \mathbf{L}_{\setminus c}|\mathbf{X}^i, \mathbf{w})]. \quad (12)$$

Note that (12) allows for an incremental computation of $p(\mathbf{Y}^{i+1} = \mathbf{L}|\mathbf{X}^{i+1}, \mathbf{w})$, $\forall \mathbf{L}$, given $p(\mathbf{Y}^i = \mathbf{L}|\mathbf{X}^i, \mathbf{w})$, $\forall \mathbf{L}$. This result is key to the efficient computation presented in the following. We now return to the original notation. Recall that for the bag $\mathbf{X}_b$, there are n_b instances and $|\mathbf{Y}_b|$ classes. Thus, for an arbitrary sub-bag $\mathbf{X}_b^{i+1}$ with $i < n_b$, there are $2^{|\mathbf{Y}_b|}$ possible labels which we denote as $\mathbf{L}^1$, $\mathbf{L}^2$,..., and $\mathbf{L}^{2^{|\mathbf{Y}_b|}}$. We proceed with Lemma 1.

Lemma 1. *Given the probability of the sub-bag $\boldsymbol{X}_b^i$ having label $\boldsymbol{L}^v$, $1 \leq \forall v \leq 2^{|\boldsymbol{Y}_b|}$, the computation of the sub-bag $\boldsymbol{X}_b^{i+1}$ having label $\boldsymbol{L}^u$, $1 \leq u \leq 2^{|\boldsymbol{Y}_b|}$, is $O(|\boldsymbol{Y}_b|)$.*

Proof. This directly follows from the number of terms $2|\mathbf{Y}_b|$ in the summation in (12). □

We derive a forward algorithm to compute the probability in (9). First, without loss of generality, we swap the position of the ith and n_bth instances. Next, we *incrementally* compute $p(\mathbf{Y}_b^{i+1} = \mathbf{L}^u|\mathbf{X}_b^{i+1}, \mathbf{w})$ from $p(\mathbf{Y}_b^i = \mathbf{L}^v|\mathbf{X}_b^i, \mathbf{w})$ based on (12), $\forall\, 1 \leq u, v \leq 2^{|\mathbf{Y}_b|}$. Using (11) for the last instance (i.e., $i = n_b$-1), we have the final result for $p(\mathbf{y}_{bn_b} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w})$.

Lemma 2. *The time complexity to compute $p(\boldsymbol{y}_{bi} = c, \boldsymbol{Y}_b = \boldsymbol{L}|\boldsymbol{X}_b, \boldsymbol{w})$ is $O(|\boldsymbol{Y}_b|2^{|\boldsymbol{Y}_b|}n_b)$.*

Proof. We build a $2^{|\mathbf{Y}_b|}$ by n_b table for each instance and the computation for each entry is $O(|\mathbf{Y}_b|)$. Due to space limitation, we omit the detailed proof. □

Lemma 2 suggests that the calculation of (9) per instance is linear in n_b the number of instances per bag and hence the computation of (9) for $i = 1, 2, \ldots, n_b$ is quadratic in n_b.

3.1.3. Maximization step

To increase objective (7), we consider gradient ascent. Specifically, we construct a backtracking [8] line search along the gradient to guarantee $g(\mathbf{w}^{(k+1)}, \mathbf{w}^{(k)}) \geq g(\mathbf{w}^{(k)}, \mathbf{w}^{(k)})$ where

$$\mathbf{w}_c^{(k+1)} = \mathbf{w}_c^{(k)} + \left.\frac{\partial g(\mathbf{w}, \mathbf{w}^{(k)})}{\partial \mathbf{w}_c}\right|_{\mathbf{w}=\mathbf{w}^{(k)}} \times \delta, \quad (13)$$

and the first derivative of $g(\mathbf{w}, \mathbf{w}^{(k)})$ w.r.t. $\mathbf{w}_c$ is computed as follows

$$\frac{\partial g(\mathbf{w}, \mathbf{w}^{(k)})}{\partial \mathbf{w}_c} = \sum_{b=1}^{B} \sum_{i=1}^{n_b} [p(\mathbf{y}_{bi} = c|\mathbf{Y}_b, \mathbf{X}_b, \mathbf{w}^{(k)})\mathbf{x}_{bi} - \frac{e^{\mathbf{w}_c^T \mathbf{x}_{bi}} \mathbf{x}_{bi}}{\sum_{c=1}^{C} e^{\mathbf{w}_c^T \mathbf{x}_{bi}}}]$$
$$= \sum_{b=1}^{B} \sum_{i=1}^{n_b} [p(\mathbf{y}_{bi} = c|\mathbf{Y}_b, \mathbf{X}_b, \mathbf{w}^{(k)}) - p(\mathbf{y}_{bi} = c|\mathbf{X}_b, \mathbf{w})]\mathbf{x}_{bi}. \quad (14)$$

The iterations will converge once (14) becomes zero. Intuitively, the inference approach fits a logistic regression model to ensure the average instance per class with and without the bag-level label information are the same.

3.2. Instance-level prediction

We consider two settings for instance-level prediction. The first setting is inductive prediction where the test bag-level label is unknown and the instance-level labels $\mathbf{y}_{ti}$ are predicted. Using the maximum-a-posteriori (MAP) approach, the label for the ith instance of the tth test bag is computed as follows

$$\hat{\mathbf{y}}_{ti} = \operatorname*{argmax}_k p(\mathbf{y}_{ti} = k|\mathbf{x}_{ti}, \mathbf{w}), \quad (15)$$

where $p(\mathbf{y}_{ti} = k|\mathbf{x}_{ti}, \mathbf{w})$ is as in (1). The second setting is transductive prediction where the test bag-level label is known. We estimate $\mathbf{y}_{ti}$ using the MAP approach as follows

$$\hat{\mathbf{y}}_{ti} = \operatorname*{argmax}_k p(\mathbf{y}_{ti} = k, \mathbf{Y}_t|\mathbf{X}_t, \mathbf{w}). \quad (16)$$

Note that since bag-level label information is available, the MAP prediction in (16) differs from (15) by the $\mathbf{Y}_t$ term. We use the dynamic programming technique in 3.1.2 to compute $p(\mathbf{y}_{ti} = k, \mathbf{Y}_t|\mathbf{X}_t, \mathbf{w})$.

4. EXPERIMENTS

We evaluate our approach on two real datasets: HJA bird song and MSCV2 and on two synthetic datasets: Carroll and Frost. Detailed information of these datasets can be found in [2, 9][1]. Table 1 shows

[1]We thank Dr. Forrest Briggs for his help with the dataset.

Table 1: Statistics of datasets in our experiments

Name	classes *(C)*	bags *(B)*	instances *(N)*	dimension *(d)*
HJA bird	13	548	4998	38
MSCV2	23	591	1758	48
Carroll	26	166	717	16
Frost	26	144	565	16

Table 2: Accuracy results for M-LR, M-RLSIM, S-LR, and S-SVM

Dataset	HJA bird	MSCV2	Carroll	Frost
M-LR-I	**0.701**	**0.557**	**0.624**	**0.645**
M-RLSIM-I	0.619	0.467	0.540	0.575
M-LR-T	**0.852**	**0.832**	**0.861**	**0.880**
M-RLSIM-T	0.817	0.697	0.745	0.775
S-LR	0.720	0.605	0.690	0.700
S-SVM	**0.772**	**0.638**	**0.772**	**0.753**

the number of classes, the number of bags, the number of instances, and the instance feature dimension for each dataset.

4.1. Baseline methods

We compare the performance of our proposed algorithm with that of the logistic regression trained in the single instance single label (SISL) setting, the rank loss SIM [2] (an SVM trained in the MIML setting), and the SVM trained in the SISL setting. Different from the MIML framework where only bag-level labels are given, in the SISL framework, a label is provided for each training instance. Consequently, we may expect a higher accuracy. Training our model in the SISL framework (i.e., when $\mathbf{y}_{ib}$ are provided) follows the standard logistic regression training. In the maximization step, the log-likelihood function follows (7) except that $p(\mathbf{y}_{bi} = c|\mathbf{Y}_b, \mathbf{X}_b, \mathbf{w}')$ is replaced by $I(\mathbf{y}_{bi} = c)$ as in (6).

4.2. Transductive and the inductive training

In the transductive setting, we use the entire dataset for training and test. Note that in this setting, we know the bag-level labels and our goal is to find the instance-level labels. In the inductive prediction, where labels of test bags are unknown, we use a 10-fold cross validation to separate the data into training and test. To provide a fair comparison to [2], we use the same 10-fold validation as in [2].

4.3. Results and discussion

We report the classification accuracy for our algorithm (M-LR) and for the logistic regression trained in the standard SISL setting (S-LR), and compare with the classification accuracy of the rank loss SIM (M-RLSIM) and the SVM trained in the standard SISL setting (S-SVM) in Table 2. From Table 2, we observe a significant improvement in accuracy of our algorithm compared to that of the rank loss SIM method. Our method is about 4-14% higher in accuracy compared to the rank loss method in almost all datasets. For the inductive prediction, our algorithm shows an improvement of 7-9% over rank loss SIM. In the transductive setting, the improvements relative to rank loss SIM are in the range of 4-14% (14% for MSCV2 and 4% for HJA bird). Furthermore, even though our algorithm learns from ambiguous bag-level labels, it approaches the performance of the logistic regression classifier trained in the SISL setting that learns from unambiguous instance-level labels. Especially for the HJA bird song dataset, the accuracy of our algorithm is just 2% lower than that of the logistic regression classifier trained in the SISL setting. In addition, note that even though the performance of SISL logistic regression is lower than that of SISL SVM, the accuracy of our algorithm is still higher than that of rank loss SIM. These results support the idea that carefully factoring in all the instances in each bag improves the classifier accuracy.

5. CONCLUSIONS

This paper addresses the instance annotation problem in the MIML setting. We propose a discriminative ORed-logistic regression model and develop a computationally efficient training algorithm. A key challenge is how to efficiently compute the posterior probability of the instance-level labels given their bag-level labels. Our algorithm avoids commonly used approximations such as sampling or variational Bayes and is well suited for cases where the number of instances per bag is large and the number of classes per bag is small. By defining sub-bags and dynamically computing the probability of their labels, we can compute all instance probabilities without approximation in quadratic time in the number of instances per bag. Experiments on bird song, image annotation, and two synthetic datasets show that the accuracy of our method is significant better, at around 7-9% in the inductive setting and 4-14% in the transductive setting, than that of the recent rank loss SIM method. We are currently working on linear run time probability computation.

6. REFERENCES

[1] M.L. Zhang and Z.H. Zhou, "M3MIML: A maximum margin method for multi-instance multi-label learning," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008, pp. 688–697.

[2] F. Briggs, X. Z. Fern, and R. Raich, "Rank-loss support instance machines for MIML instance annotation," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2012, KDD '12, pp. 534–542, ACM.

[3] C.T. Nguyen, D.C. Zhan, and Z.H. Zhou, "Multi-modal image annotation with multi-instance multi-label LDA," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*. AAAI Press, 2013, pp. 1558–1564.

[4] S.H. Yang, H. Zha, and B.G. Hu, "Dirichlet-Bernoulli alignment: A generative model for multi-class multi-label multi-instance corpora," in *Advances in neural information processing systems*, 2009, pp. 2143–2150.

[5] J. Foulds and P. Smyth, "Multi-instance mixture models and semi-supervised learning," in *SIAM International Conference on Data Mining*, 2011.

[6] T. Moon, "The expectation-maximization algorithm," *Signal processing magazine, IEEE*, vol. 13, no. 6, pp. 47–60, 1996.

[7] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*, vol. 382, John Wiley & Sons, 2007.

[8] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[9] L. Neal, F. Briggs, R. Raich, and X. Z. Fern, "Time-frequency segmentation of bird song in noisy acoustic environments," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2012–2015.