DISCRIMINATIVE PROBABILISTIC FRAMEWORK FOR GENERALIZED MULTI-INSTANCE LEARNING

Anh T. Pham, Raviv Raich, Xiaoli Z. Fern, Weng-Keen Wong, and Xinze Guan

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

ABSTRACT

Multiple-instance learning is a framework for learning from data consisting of bags of instances labeled at the bag level. A common assumption in multi-instance learning is that a bag label is positive if and only if at least one instance in the bag is positive. In practice, this assumption may be violated. For example, experts may provide a noisy label to a bag consisting of many instances, to reduce labeling time. Here, we consider generalized multi-instance learning, which assumes that the bag label is non-deterministically determined based on the number of positive instances in the bag. The challenge in this setting is to simultaneous learn an instance classifier and the unknown bag-labeling probabilistic rule. This paper addresses the generalized multi-instance learning using a discriminative probabilistic graphical model with exact and efficient inference. Experiments on both synthetic and real data illustrate the effectiveness of the proposed method relative to other methods including those that follow the traditional multiple-instance learning assumption.

Index Terms— Multi-instance learning, graphical model, expectation-maximization, dynamic programming

1. INTRODUCTION

Multiple-instance learning (MIL) is a framework for learning with label ambiguity. In MIL, the objects of interest are referred to as bags and each bag consists of one or more parts called instances. Instances can be either positive or negative. If all instances from a bag are negative, the bag is labeled as negative. Otherwise, if at least one instance in the bag is positive, the bag is labeled as positive. This setting is referred to as the presence-based assumption [1]. Multiple-instance learning has many applications, e.g., in image processing [2], drug activity prediction [3], bird song species prediction [4], document categorization [5], and activity recognition [6].

In practice, the labeling process is often noisy and imprecise, violating the presence-based assumption. For example, experts may only label a bag as positive if they feel that sufficient number positive instances are present in the bag. This setting is referred to as generalized multi-instance learning [7]. For example, an image is labeled as 'forest' if and only if it contains more than ten 'tree' segments.

The classical MIL framework, presents two type of classification problems: (i) instance level prediction and (ii) bag level prediction. The presence-based assumption provides a framework to derive a bag level classification rule based on the learned instance level classifier. Generalized multi-instance learning presents an additional challenge because the relation between the bag label and instance labels is unknown and must be learned in addition to the instance level classifier. In this paper, we propose a discriminative probabilistic graphical model for generalized multi-instance learning. Our contributions are as follows. We introduce a probabilistic model that relates the bag label to the number of positive instance in the bag and derive an Expectation-Maximization update rule to facilitate parameter estimation of the proposed model. We demonstrate that the instance level classifier learning and the bag labeling rule learning can be solved separately in each iteration. Experiments on real and synthetic data illustrate the effectiveness of the proposed solution in comparison to the current state-of-the-art for generalized MIL, and methods that follows the presence-based assumption.

This work is partially supported by the National Science Foundation grants CCF-1254218, DBI-1356792, and IIS-1055113.

2. RELATED WORK

Generalized multi-instance learning has been considered using two-level classification (TLC) scheme [1]. In TLC, authors consider three different assumptions in MIL: the presence assumption, the threshold assumption, and the count assumption where each represents a different rule for generating the bag label based on instance labels. Specifically, these assumptions are the bag is labeled positive if the number of instances in considered concepts is greater than one, greater than an unknown threshold, or in a range, respectively. A two-level decision tree is used to find the solution. Another line of work is constructive clustering ensemble (CCE) multi-instance learning [8]. In CCE, instances are first groups into clusters. Then, bags are featured by the number of instances in these clusters. Support Vector Machine (SVM) is used to learn the relation among bag features and bag labels.

Another line of research named generalized multi-instance learning is in [9] and [10]. Specifically, k concept points are defined and a bag is positive if there are r out of k concept points such that for each of these concept points, there is at least one instance in the bag close to it. [11] considers the use of cardinality potentials for MIL by defining potential function between instance labels, instance features, and bag labels then solves using an SVM approach which is used frequently for MIL. In contrast, the proposed framework uses a probabilistic graphical model approach with logistic relation among instance labels, instance features, and bag labels. Note that probabilistic graphical models with exact inference often achieve higher accuracy than SVM-based approaches in multi-instance multi-label setting [12] [13].

3. PROBLEM FORMULATION AND MODEL

Data in MIL is modeled as $\{\mathbb{X}, \mathbb{Y}\} = \{\mathbf{X}_b, Y_b\}_{b=1}^{B}$ where B indicates the number of bags. We denote the bth bag of instance feature vectors by $\mathbf{X}_b = \{\mathbf{x}_{b1}, \mathbf{x}_{b2}, \ldots, \mathbf{x}_{bn_b}\}$ where $\mathbf{x}_{bi} \in \mathcal{X} = \mathbb{R}^d$ denotes the ith instance feature and d is the number of features. The bag

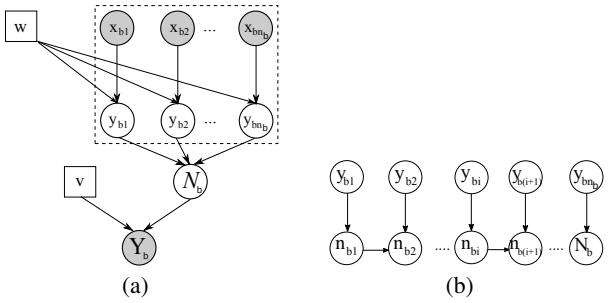


Fig. 1. (a) The proposed graphical model Soft OR Logistic Regression SORLR for the bth bag. Observed nodes are shaded. (b) Graphical model to compute instance membership probability $p(y_{bi} = 1|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}')$ and probability of the number of positive instances $p(N_b = n|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}')$.

label is denoted as $Y_b \in \{0, 1\}$ in which 0 and 1 represent negative and positive labels, respectively. Hidden instance labels are denoted as $\mathbf{y}_b = \{y_{b1}, y_{b2}, \ldots, y_{bn_b}\}$. Our goal is twofold: (i) we would like to find a bag label classifier $g : 2^{\mathcal{X}} \rightarrow \{0, 1\}$ based on training data of the form $\{\mathbb{X}, \mathbb{Y}\}$ and (ii) we would like to find an instance level classifier $g : \mathcal{X} \rightarrow \{0, 1\}$ based on training data of the form $\{\mathbb{X}, \mathbb{Y}\}$. The focus in this paper is to achieve the two goals under the assumption of generalized MIL setting.

3.1. Proposed model

We assume that the B bags in our training data $\mathbf{X}_1, \ldots, \mathbf{X}_B$ are independent. The graphical model for the bth bag is presented in Fig. 1(a). We introduce a discriminative probability model that relates instance feature vectors to the corresponding instance labels using an *instance level classifier* followed by a *bag labeler* relating the bag level label to set of instance level labels.

The instance level classifier. We assume that instance level labels are generated independently given the collection of instance level features: $p(y_{b1}, \ldots, y_{bn_b}|\mathbf{X}_b) = \prod_{i=1}^{B} p(y_{bi}|\mathbf{X}_b)$ and that instance labels depend on $\mathbf{X}_b$ through their corresponding feature vector, i.e., $p(y_{bi}|\mathbf{X}_b) = p(y_{bi}|\mathbf{x}_{bi})$. Additionally, the relation between the instance label $y_{bi} \in \{0, 1\}$ and instance feature $\mathbf{x}_{bi}$ is modeled using logistic regression function as follows

$$p(y_{bi} = 1|\mathbf{x}_{bi}, \mathbf{w}) = \frac{e^{\mathbf{w}^T \mathbf{x}_{bi}}}{1 + e^{\mathbf{w}^T \mathbf{x}_{bi}}}, \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the instance level classifier parameter.

The bag labeler. To produce a bag level label Y_b given the collection of instance level labels $\{y_{b1}, \ldots, y_{bi}\}$ we assume a cardinality based model. In this approach, the key assumption is that the bag label is related to the instance labels through the number of positively labeled instances. The number of positive instances in the bth bag denoted by N_b is modeled according to

$$p(N_b = n|\mathbf{y}_b) = \mathbb{I}\big[\sum_{i=1}^{n_b} y_{bi} = n\big], \tag{2}$$

where $\mathbb{I}[\cdot]$ is the indicator function taking the value of one for true argument and zero otherwise. The relation between the bag label $Y_b \in \{0, 1\}$ and the number of positive instances N_b is modeled as follows

$$p(Y_b = 1|N_b = n, \mathbf{v}) = \phi(n; \mathbf{v}), \tag{3}$$

where $\mathbf{v}$ is the bag labeler parameter vector and $\phi : \{0\} \cup \mathbb{Z}^+ \times \mathbb{R}^{\dim(\mathbf{v})} \rightarrow [0, 1]$. We consider three cases for the function:
Case (1):

$$\phi(n; \mathbf{v}) = \begin{cases} v_n & \text{if } n = 0, 1, \ldots, m, \\ v_m & \text{if } n > m. \end{cases} \tag{4}$$

where $\mathbf{v} = [v_0, v_1, \ldots, v_m] \in \mathbb{R}^{m+1}$ is an unknown bag labeler parameter vector with m is the maximum training bag size. This model assigns bag label probability for each possible value of the number of positive instances in an independent fashion. Note that the model can be restricted to enforce monotonicity on the v_n coefficients such that $0 \leq v_1 \leq v_2 \leq \ldots, v_m$.
Case (2):

$$\phi(n; \mathbf{v}) = \frac{e^{v_0 + n \times v_1}}{1 + e^{v_0 + n \times v_1}}, \tag{5}$$

where $\mathbf{v} = [v_0, v_1] \in \mathbb{R}^2$ is an unknown bag labeler parameter vector. The model follows a logistic regression model with an input vector of $[1, n]$. This model guarantees the monotonicity of the bag label probability as a function of the number of positive instances. With $v_1 > 0$, the probability of positively labeling a bag increase with the number of positive instance labels.
Case (3):

$$\phi(n; \mathbf{v}) = \mathbb{I}[n \geq v_0], \tag{6}$$

where $\mathbf{v} = [v_0] \in \mathbb{R}^1$ is an unknown bag labeler parameter. This model offers a deterministic assignment of a positive label to a bag if the number of positive instances in the bag is greater or equal v_0. If $v_0 = 1$, the setting of this case is similar to that of multi-instance learning.

Three models are ordered by the flexibility level. The first model allows arbitrary relation among number of positive instances and bag label. The second model enforces monotonic constraint. The last model is a threshold model. If the labelers are very sure that the bag label is positive if there are at least three positive instances, then, the third model is the most suitable model.

Based on the aforementioned general model, our original goal in the beginning of this section translate to learning the instance classifier parameter $\mathbf{w}$ vector and the bag labeler parameter vector $\mathbf{v}$.

4. INFERENCE FOR THE PROPOSED MODEL

We consider a maximum likelihood approach for the inference of the aforementioned model parameters given training data $\{\mathbb{X}, \mathbb{Y}\}$. The likelihood $p(\mathbb{Y}, \mathbb{X}|\mathbf{w}, \mathbf{v})$ is computed using $p(\mathbb{Y}, \mathbb{X}|\mathbf{w}, \mathbf{v}) = p(\mathbb{Y}|\mathbb{X}, \mathbf{w}, \mathbf{v})p(\mathbb{X})$ since $P(\mathbb{X})$ is independent of $\mathbf{w}$ and $\mathbf{v}$. Consequently, the likelihood can be maximized by maximizing $p(\mathbb{Y}|\mathbb{X}, \mathbf{w}, \mathbf{v})$. Following the independence assumption among bags and marginalizing over the number of positive instances N_b and instance labels $\mathbf{y}_b$ in each bag, the log-likelihood $\mathbf{L}(\mathbb{Y}|\mathbb{X}, \mathbf{w}, \mathbf{v})$ is computed as $\sum_{b=1}^{B} \log[\sum_{N_b=0}^{n_b} \sum_{\mathbf{y}_b \subset M_b} p(Y_b, N_b, \mathbf{y}_b|\mathbf{X}_b, \mathbf{w}, \mathbf{v})]$, where $M_b \in \{0, 1\}^{n_b}$ is the set of all possible instance labels $\mathbf{y}_b$ for the bth bag. To the best of our knowledge, no closed-form solution is available for the maximization of $\mathbf{L}(\mathbb{Y}|\mathbb{X}, \mathbf{w}, \mathbf{v})$. Consequently, we proceed with the Expectation Maximization framework [14] for implementing maximum likelihood estimation.

4.1. Expectation maximization for generalized multi-instance learning

We consider the training data $\{\mathbb{X}, \mathbb{Y}\}$ as the observed data and the positive instance counts and instance labels $\{\mathbf{N}_D, \mathbf{y}_D\}$ where $\mathbf{N}_D = \{N_b\}_{b=1}^{B}$ and $\mathbf{y}_D = \{\mathbf{y}_b\}_{b=1}^{B}$ as the hidden data. Consequently, the

complete data ia given by $\{\mathbb{X}, \mathbb{Y}, \mathbf{N}_D, \mathbf{y}_D\}$. The conditional version of the complete data log-likelihood follows

$$\begin{aligned}&\mathbf{L}(\mathbb{Y}, \mathbf{N}_D, \mathbf{y}_D|\mathbb{X}, \mathbf{w}, \mathbf{v}) \quad (7)\\ &= \sum_{b=1}^{B}[\sum_{i=1}^{n_b}\Big(\mathbb{I}[y_{bi}=1]\mathbf{w}^T\mathbf{x}_{bi} - \log(1+e^{\mathbf{w}^T\mathbf{x}_{bi}})\Big)+\\ &\sum_{n=0}^{n_b}\mathbb{I}[N_b=n]\Big(\log p(Y_b|N_b=n,\mathbf{v}) + \log p(N_b=n|\mathbf{y}_b)\Big)].\end{aligned}$$

Taking the expectation of the complete log-likelihood w.r.t. $\mathbf{N}_D$ and $\mathbf{y}_D$ given $\mathbb{Y}, \mathbb{X}, \mathbf{w}', \mathbf{v}'$ and noting that the expectation of the second term is independent of $\mathbf{w}$ and $\mathbf{v}$, we obtain the surrogate for log-likelihood, i.e., the auxiliary function, as follows

$$\mathcal{Q}(\mathbf{w}, \mathbf{v}; \mathbf{w}', \mathbf{v}') = \mathcal{Q}_1(\mathbf{w}; \mathbf{w}', \mathbf{v}') + \mathcal{Q}_2(\mathbf{v}; \mathbf{w}', \mathbf{v}') \quad (8)$$

where $\mathcal{Q}_1(\mathbf{w}; \mathbf{w}', \mathbf{v}') = \sum_{b=1}^{B}\sum_{i=1}^{n_b}[p(y_{bi}=1|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}')\mathbf{w}^T\mathbf{x}_{bi} - \log(1+e^{\mathbf{w}^T\mathbf{x}_{bi}})]$, and $\mathcal{Q}_2(\mathbf{v}; \mathbf{w}', \mathbf{v}') = \sum_{b=1}^{B}\sum_{n=0}^{n_b} p(N_b = n|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}')\Big(Y_b \log\phi(n;\mathbf{v}) + (1-Y_b)\log(1-\phi(n;\mathbf{v}))\Big)$. The premise of EM, is that by using the update rule $[\mathbf{w}^{k+1}, \mathbf{v}^{k+1}] = \arg\max_{\mathbf{w},\mathbf{v}} \mathcal{Q}(\mathbf{w}, \mathbf{v}; \mathbf{w}^k, \mathbf{v}^k)$ where $\mathcal{Q}$ is given by (8), a series of non-decreasing incomplete data log-likelihood values can be obtained. Based on the auxiliary function in (8), the expectation maximization update procedure is

E-step:
- Compute instance label probability $p(y_{bi}=1|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}'), \forall 1 \le i \le n_b$
- Compute the number of positive instances probability $p(N_b = n|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}'), \forall 0 \le n \le n_b$.

M-step:
- Estimate the instance label classifier parameter $\mathbf{w}$ using $\max_{\mathbf{w}} \mathcal{Q}_1(\mathbf{w}; \mathbf{w}', \mathbf{v}')$
- Estimate the bag labeler parameter $\mathbf{v}$ using $\max_{\mathbf{v}} \mathcal{Q}_2(\mathbf{v}; \mathbf{w}', \mathbf{v}')$.

4.2. E-step

Our goal in the E-step is to compute $p(y_{bi}=1|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}')$ and $p(N_b = n|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}')$. Brute-force computations require marginalizing over hidden variables $y_{b1}, \ldots, y_{bn_b}$ with the cost of $O(2^{n_b})$. To address this problem, we propose a reformulation of the graphical model as a chain. Additionally, we present a forward-backward message passing algorithm that allows for a $O(n_b^2)$ computation of the aforementioned probabilities.

Assume an order of instances in the bth bag. Denote the total number of positive instances from the 1st to the ith instance by n_{bi}: $n_{bi} = \sum_{j=1}^{i} y_{bj}$. The new variable n_{bi} admits a chain structure due to the recursion $n_{b(i+1)} = n_{bi} + y_{b(i+1)}$, which is initialized with $n_{b1} = y_{b1}$. Following the new notation, the number of positive instances in a bag N_b satisfies $N_b = n_{nn_b}$. We adapt the technique in [15] of converting a tree to a chain structure to our particular case. We then derive the forward and backward method to compute $p(y_{bi}=1|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}')$ and $p(N_b = n|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}')$ as follows.

1. *The forward message.* Define $\alpha_i(k) \triangleq p(n_{bi} = k|\mathbf{X}_b, \mathbf{w}')$. Then, $\alpha_{i+1}(k)$ is computed forwardly for $i = 1, 2, \ldots, n_b - 1$ as follows

$$\begin{aligned}\alpha_{i+1}(k) = &\, p(y_{b(i+1)}=1|\mathbf{x}_{b(i+1)}, \mathbf{w}')\alpha_i(k-1)\\ &+ p(y_{b(i+1)}=0|\mathbf{x}_{b(i+1)}, \mathbf{w}')\alpha_i(k). \quad (9)\end{aligned}$$

2. *The backward message.* Define $\beta_i(k) \triangleq p(Y_b|n_{bi} = k, \mathbf{X}_b, \mathbf{w}')$. Then, $\beta_i(k)$ is computed backwardly for $i = n_b - 1, \ldots, 2, 1$ as follows

$$\begin{aligned}\beta_i(k) = &\, p(y_{b(i+1)}=1|\mathbf{x}_{b(i+1)}, \mathbf{w}')\beta_{i+1}(k+1)\\ &+ p(y_{b(i+1)}=0|\mathbf{x}_{b(i+1)}, \mathbf{w}')\beta_{i+1}(k). \quad (10)\end{aligned}$$

3. The first required E-step probability $p(y_{bi}=1|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}')$ can be computed from α and β as follows

$$p(y_{bi}=1|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}') = \frac{\pi_1}{\pi_1 + \pi_0}, \quad (11)$$

where $\pi_1 = \sum_{k=0}^{n_b}\alpha_{i-1}(k)\beta_i(k+1)p(y_{bi}=1|\mathbf{x}_{bi}, \mathbf{w}')$ and $\pi_0 = \sum_{k=0}^{n_b}\alpha_{i-1}(k)\beta_i(k)p(y_{bi}=0|\mathbf{x}_{bi}, \mathbf{w}')$.

4. The second required E-step probability $p(n_{bi} = n|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}')$ can be computed from α and β as follows

$$p(n_{bi}=n|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}') = \frac{\alpha_i(n)\beta_i(n)}{\sum_{k=0}^{n_b}\alpha_i(k)\beta_i(k)}. \quad (12)$$

Setting $i = n_b$ in (12), we obtain $p(N_b = n|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}')$.
The E-step computation requires traversing the graphical model in Fig. 1(b) back and forth in $O(n_b)$ steps. The number of values each of the states n_{bi} takes is $O(n_b)$ and each term computation is $O(1)$. Hence, the computational complexity of the E-step for the bth bag is $O(n_b^2)$. The detailed derivations of the update rules for $\alpha_i(k)$, $\beta_i(k)$, and computation of $p(y_{bi}=1|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}')$ and $p(N_b = n|Y_b, \mathbf{X}_b, \mathbf{w}', \mathbf{v}')$ are provided in the supplementary material [16].

4.3. M-step

Recall the auxiliary function from (8). Since the objective function is separable in $\mathbf{w}$ and $\mathbf{v}$, we derive the optimization process for them separately. Denote $\theta' = [\mathbf{w}', \mathbf{v}']$.

Instance level classifier update. We apply Newton method with backtracking line search [17] to update $\mathbf{w}$ in order to solve the minimization of $\mathcal{Q}_1$: $\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \eta \times \mathbf{H}_w^{-1}\mathbf{d}_w\Big|_{\mathbf{w}=\mathbf{w}^{(k)}}$, where the gradient $\mathbf{d}_w = \sum_{b=1}^{B}\sum_{i=1}^{n_b}[p(y_{bi}=1|Y_b, \mathbf{X}_b, \theta') - p(y_{bi}=1|\mathbf{x}_{bi}, \mathbf{w})]\mathbf{x}_{bi}$, and the Hessian $\mathbf{H}_w = -\sum_{b=1}^{B}\sum_{i=1}^{n_b} p(y_{bi}=1|\mathbf{x}_{bi}, \mathbf{w})p(y_{bi}=0|\mathbf{x}_{bi}, \mathbf{w})\mathbf{x}_{bi}\mathbf{x}_{bi}^T$.

Bag labeler update. In the following, we present the update rule for $\mathbf{v}$ for the three bag labeler models.

Case ①, when $p(Y_b = 1|N_b = n, \mathbf{v})$ is modeled by (4), then by setting the gradient of $\mathcal{Q}_2$ to 0, v_n can be computed as $v_n = \frac{\delta_n}{\delta_n + \tau_n}$, where $\delta_n = \sum_{b=1}^{B} Y_b p(N_b = n|Y_b = 1, \mathbf{X}_b, \theta')$ and $\tau_n = \sum_{b=1}^{B}(1-Y_b)p(N_b = n|Y_b = 0, \mathbf{X}_b, \theta')$.

Case ②, we use Newton method with backtracking line search for updating $\mathbf{v}$. Specifically, $\mathbf{v}^{(k+1)} = \mathbf{v}^{(k)} - \eta \times \mathbf{H}_v^{-1}\mathbf{d}_v\Big|_{\mathbf{v}=\mathbf{v}^{(k)}}$, where the gradient is given by $\mathbf{d}_v = \sum_{b=1}^{B}\sum_{n=0}^{n_b}\rho_1(b,n)[n \;\; 1]^T$, where $\rho_1(b,n) = [Y_b - p(Y_b = 1|N_b = n, \mathbf{v})]p(N_b = n|Y_b, \mathbf{X}_b, \theta')$ and the Hessian is $\mathbf{H}_v = \sum_{b=1}^{B}\sum_{n=0}^{n_b}\rho_2(b,n)(-1)[1 \;\; n]^T[1 \;\; n]$, where $\rho_2(b,n) = p(Y_b = 1|N_b = n, \mathbf{v}) \times p(Y_b = 0|N_b = n, \mathbf{v})p(N_b = n|Y_b, \mathbf{X}_b, \theta')$.

Case ③, $\mathbf{v} = [v_0]$ is found by $\min_{v_0}\sum_{b=1}^{B}\big[Y_b p(N_b < v_0|Y_b = 1, \mathbf{X}_b, \theta') + (1-Y_b)p(N_b \ge v_0|Y_b = 0, \mathbf{X}_b, \theta')\big]$. Note that the search over v_0 can be restricted to the set $\{0, 1, \ldots, n_{max}+1\}$, where $n_{max} = \max_{b=1}^{B} n_b$.

4.4. Prediction

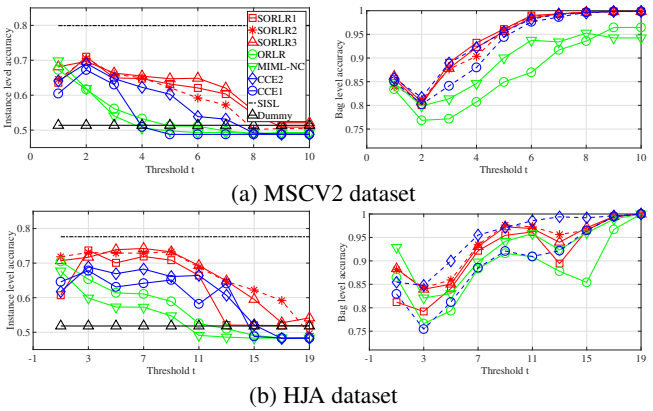
Instance label prediction: the label for the ith instance in the tth test bag is predicted as $\hat{y}_{ti} = \mathbb{I}[\mathbf{w}^T \mathbf{x}_{bi} > 0]$. **Bag label prediction**: the label for the tth test bag is predicted as $p(Y_t = 1|\mathbf{X}_t, \mathbf{w}, \mathbf{v}) = \sum_{n=0}^{n_t} p(Y_t = 1, N_{n_t} = n|\mathbf{X}_t, \mathbf{w}, \mathbf{v}) = \sum_{n=0}^{n_t} \phi(n; \mathbf{v})\alpha_{n_t}(n)$, where $\alpha_{n_t}(n)$ is computed using the forward approach. The predicted bag label is 1 if $p(Y_t = 1|\mathbf{X}_t, \mathbf{w}, \mathbf{v}) \geq 0.5$.

5. EXPERIMENTS

Setting. We compare three different versions of the proposed method Soft ORed Logistic Regression SORLR1-3 corresponding to the three aforementioned cases with MIML-NC [12], a multi-instance multi-label learning algorithm. MIML-NC is considered as a baseline algorithm for two reasons. First, it is a discriminative graphical model that has been shown to outperform SVM-based methods [12]. Second, the setting of MIML-NC is similar to the MIL setting. MIML-NC works in multi-instance multi-label setting (MIML), a general setting for multi-instance learning (MIL), and can deal with novel class. Specifically, if a bag has novel class instances, the novel class is removed from the bag label. Similarly in MIL, for bags with several negative instances and at least one positive instance, the negative label is removed from the bag label. MIML-NC facilitates EM framework to learn an instance level classifier for both known classes and novel class. We further consider an additional setting of the proposed method for case 3, which is called ORed Logistic Regression ORLR, where v_0 is fixed at 1. We expect this method to perform similarly as MIML-NC. We also compared the proposed method with CCE [8]. In CCE, instances are group into clusters. Bags are featurized using presence or count of is instances in each cluster and an SVM classifier is used to learn from bags. We consider CCE1, which is based on the presence assumption and CCE2, which uses the count assumption. Note that while CCE separates the clustering instances process with the learning the bag-clusters relation process, SORLR combines two steps into a single probabilistic framework. Since CCE is an SVM-based approach, we use the RBF kernel with parameter γ searched in the set $\{0.001, 0.01, 0.1, \ldots, 1000\}$ and C searched over the set of $\{0.001, 0.01, 0.1, \ldots, 1000\}$. The kernel version of SORLR1-3 and MIML-NC is implemented based on the random Fourier transform technique as in [18] with the RBF kernel width is selected post-hoc in the set $\{0.01, 0.02, 0.05, 0.1, 0.2, \ldots, 5, 10\}$. We also consider logistic regression trained under the single instance single label (SISL) setting where the instance labels are known. Since instance labels are provided in the SISL setting, we expect the LR model trained in this setting to outperform other methods. Additionally, we consider a Dummy classifier. For every instance, the Dummy classifier predicts the label of the most frequent class in the training data. The Dummy classifier is expected to be outperformed by methods that consider the instance features in prediction.

Evaluation metrics. We consider instance-level prediction accuracy and bag-level prediction accuracy, which is the ratio between the correctly predicted and the total number of bags, as measures for all algorithms. Note that CCE methods are designed to make bag-level predictions. To generate instance-level prediction, we use CCE1 and CCE2 to train bag-level classifiers and then simply create a bag for each instance at testing stage to make prediction.

Datasets. We consider the following MIML datasets: HJA bird song and MSCV2 image annotation [4] datasets. We form two-class datasets by considering several classes as positive and the remaining classes as negative. Specifically, classes $1-5$ and $1-6$ in HJA and MSCV2, respectively, are chosen as positive while keeping the resulting datasets balanced. Each bag is labeled positive if it has no less than t positive instances where t is varied from 1 to 19 with step size of 2 for HJA, and from 1 to 10 for MSCV2 due to its smaller average bag size.



(a) MSCV2 dataset

(b) HJA dataset

Fig. 2. Instance-level accuracy (left) and bag-level accuracy (right) as a function of the threshold t of the proposed models (SORLR1-3) and baseline methods.

Results and analysis. From Fig. 2, we first note that the three methods that are based on the presence assumption, namely ORLR, MIML-NC, and CCE1, performed consistently worse than other methods as we increase the threshold t. In fact, as t increases, their performance eventually approaches the performance of the Dummy classifier. This is consistent with our expectation because these methods do not correctly model the bag labeler. CCE2, the current state-of-the-art method that considers the concept count, performs very competitively for the bag-level accuracy. However, when considering the instance-level prediction accuracy, CCE2 is generally inferior to the proposed methods (SORLR1-3). For example, if the threshold is 7, the accuracy of SORLR3 is 9% and 6% higher than those of CCE2 on MSCV2 and HJA, respectively. When the threshold increases a little, the presence assumption for algorithms is violated. Therefore, both bag and instance level accuracy decrease. However, when the threshold becomes very big, all training bags and test bags are negative. Hence, for most classifiers, all instances in spaces are learnt to be negative. Thus, the instance accuracy decreases. However, test bags are negative leading to the bag accuracy is remaining high. SORLR2 and SORLR3 are quite consistent in term of performance whereas SORLR1 may have overfitting problem, especially in HJA where a large number of parameters are used in vector $\mathbf{v}$.

6. CONCLUSION

This paper considered generalized multi-instance learning in which the bag label depends on the number of positive instances. We presented a discriminative graphical model taking into account both the instance level classifier and the bag labeler model. An efficient and exact inference framework was introduced. Experiments on real datasets illustrate that the proposed method can effectively adapt to bag labeling that does not follow the presence-based assumption. Moreover, the proposed approach was demonstrated to be competitive compared to state-of-the-art techniques for the generalized MIL setting.

7. REFERENCES

[1] N. Weidmann, E. Frank, and B. Pfahringer, "A two-level learning method for generalized multi-instance problems," in *European Conference on Machine Learning*, pp. 468–479. 2003.

[2] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang, "Joint multi-label multi-instance learning for image classification," in *Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[3] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

[4] F. Briggs, X. Z. Fern, and R. Raich, "Rank-loss support instance machines for MIML instance annotation," in *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 534–542.

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

[6] M. Stikic, D. Larlus, S. Ebert, and B. Schiele, "Weakly supervised recognition of daily life activities with wearable sensors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2521–2537, 2011.

[7] J. Foulds and E. Frank, "A review of multi-instance learning assumptions," *The Knowledge Engineering Review*, vol. 25, no. 01, pp. 1–25, 2010.

[8] Z.-H. Zhou and M.-L. Zhang, "Solving multi-instance problems with classifier ensemble based on constructive clustering," *Knowledge and Information Systems*, vol. 11, no. 2, pp. 155–170, 2007.

[9] S. Scott, J. Zhang, and J. Brown, "On generalized multiple-instance learning," *International Journal of Computational Intelligence and Applications*, vol. 5, no. 01, pp. 21–35, 2005.

[10] Q. Tao, S. Scott, N. Vinodchandran, and T. T. Osugi, "SVM-based generalized multiple-instance learning via approximate box counting," in *Proceedings of the International Conference on Machine Learning*, 2004, p. 101.

[11] H. Hajimirsadeghi, J. Li, G. Mori, M. Zaki, and T. Sayed, "Multiple instance learning by discriminative training of markov networks," in *Uncertainty in Artificial Intelligence*, 2013, p. 262.

[12] A. T. Pham, R. Raich, X. Z. Fern, and J. P. Arriaga, "Multi-instance multi-label learning in the presence of novel class instances," in *Proceedings of the International Conference on Machine Learning*, 2015, pp. 2427–2435.

[13] A. T. Pham, R. Raich, and X. Z. Fern, "Efficient instance annotation in multi-instance learning," in *Proceedings of the IEEE Workshop on Statistical Signal Processing*, 2014, pp. 137–140.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society B*, pp. 1–38, 1977.

[15] D. Heckerman and J. S. Breese, "A new look at causal independence," in *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 1994, pp. 286–292.

[16] A. T. Pham, R. Raich, X. Z. Fern, W. K. Wong, and Guan X., "GMIL-Supplementary Material," `web.engr.oregonstate.edu/~phaman/GMIL_supplementary_material.pdf`.

[17] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

[18] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, 2007, pp. 1177–1184.