Multi-instance multi-label learning in the presence of novel class instances

Anh T. Pham PHAMAN@EECS.OREGONSTATE.EDU
Raviv Raich RAICH@EECS.OREGONSTATE.EDU
Xiaoli Z. Fern XFERN@EECS.OREGONSTATE.EDU
School of Electrical Engineering and Computer Science, Corvallis, OR 97330-5501 USA

Jesús Pérez Arriaga JPEREZ@GTAS.DICOM.UNICAN.ES
Departamento de Ingeniería de Comunicaciones, Universidad de Cantabria, 39005 Santander, Spain

Abstract

Multi-instance multi-label learning (MIML) is a framework for learning in the presence of label ambiguity. In MIML, experts provide labels for groups of instances (bags), instead of directly providing a label for every instance. When labeling efforts are focused on a set of target classes, instances outside this set will not be appropriately modeled. For example, ornithologists label bird audio recordings with a list of species present. Other additional sound instances, e.g., a rain drop or a moving vehicle sound, are not labeled. The challenge is due to the fact that for a given bag, the presence or absence of novel instances is latent. In this paper, this problem is addressed using a discriminative probabilistic model that accounts for novel instances. We propose an exact and efficient implementation of the maximum likelihood approach to determine the model parameters and consequently learn an instance-level classifier for all classes including the novel class. Experiments on both synthetic and real datasets illustrate the effectiveness of the proposed approach.

1. Introduction

Multi-instance multi-label learning is a framework for learning in the presence of label ambiguity. In conventional single instance single label learning (SISL), a label is provided for every instance. In contrast, in MIML learning, instances are grouped into bags and labels are provided at the bag-level. MIML learning has many applications where data is complex and the labeling process is costly. For instance, in birdsong recognition, instead of providing a species label for each bird syllable, experts prefer to provide species labels for longer intervals, each containing multiple bird syllables (Briggs et al., 2012).

A common assumption in MIML is that the set of classes in which we are interested is a closed set and all instances we encounter in both training and testing are assumed to belong to this fixed set of classes. However, this assumption is frequently violated in real applications. For example, in birdsong recognition, experts label long audio intervals with a fix set of bird species. Other categories of sound such as rain or car sound are not included in the labeling process. Yet, such sounds are present in the data. Another example is image annotation, as shown in Fig. 1, where the annotator considers only a fixed set of tags and ignores 'grass' as it is not included in the tag set. We refer to this as the novel instance problem in MIML data. Properly modeling the novel instances in MIML data can have positive impacts on several applications. First, it enables effective detecting of novel instances in the data. Second, by modeling and recognizing the novel instances in MIML data, we can better model the instances of known classes, and consequently improve the ability to predict the class both at the bag-level and at the instance-level.

There are several lines of work that are related to the problem of MIML learning with novel instances. The problem of detecting instances from novel class has been studied under the SISL setting (Saligrama & Zhao, 2012), (Hsiao et al., 2012), (Da et al., 2014). However, a common assumption in the SISL setting is that we only observe instances from the known classes during training. In contrast, in MIML setting, we may in fact observe instances of the novel class during training; they are just not labeled. For example, consider the 2nd and 4th images in Fig. 1. Even though these bag labels, which are $\{$'building'$\}$ and $\{$'bird'$\}$ respectively, do not indicate the presence of novel instances from 'grass', however, two grass segments are

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

available during training.

(Yang et al., 2013) introduced the problem of MIML learning with weak labels, where the bag-level labels may be incomplete. Different from our setting, there the missing labels are from the fixed set of known classes. Finally, (Lou et al., 2013) proposed to solve the novelty instance detection problem for MIML by learning a score function for each known class and predicting an instance to belong to the novel class if it scores low for all known classes. Since there is no direct modeling mechanism for the novel class in training, the process of learning the class score function is not specific to novelty detection and various methods from the MIML literature can be used to accomplish this task, e.g., (Briggs et al., 2012), (Cour et al., 2011).

In this paper, we develop a discriminative probabilistic model that explicitly models the novel class instances under the MIML setting. Our contributions are as follows: (i) we develop a model that accounts for presence of novel class instances (ii) we present an exact and efficient inference method for the model. The advantages of the proposed framework are demonstrated by experiments on bag-level label prediction and novel class detection using both synthetic and real datasets.

2. Related work

Many multi-instance multi-label learning approaches follow the maximum margin principle (Huang et al., 2014), (Zhang & Zhou, 2008). Specifically, these approaches maximize the margin among classes and the score of a bag w.r.t. each class is computed from the score-maximizing instance in the bag. As a consequence, these methods utilize only a subset of the available instances in a given bag. (Briggs et al., 2012) addresses this issue by using the softmax score considering all instances in the bag. Learning from partial label is another framework that can be used for MIML learning, as in CMM-LSB (Liu & Dietterich, 2012). CMM-LSB discards the bag structure. Instead, the label of each instance is taken from its bag label. Thus, the method ignores the relationship among instances in each bag, which may then degrade the accuracy. Recently, ORLR (Pham et al., 2014), a probabilistic approach, is proposed for the MIML learning. ORLR considers the class membership for every instance in each bag and enforces the constraint that the bag-level label set is formed as union of the labels of all its instances. None of the aforementioned methods directly addresses the potential problem of novel class instances present in the data.

A number of approaches have been proposed for novel class detection in the SISL setting. One solution to detect novel class instances is to test whether it comes from a probability distribution of known instances (Markou & Singh, 2003). Another solution is finding the minimum set covering most of the training examples and consider remaining examples as novel instances such as (Zhao & Saligrama, 2009). In the MIML setting, we may observe instances from novel class during training, as in (Lou et al., 2013), where novel instance features are available in training. Specifically, the label of the bag only contains known instance labels and ignores the novel instances. For example, in image annotation, experts may only consider the popular labels, such as ‘sky’, ‘building’, and ‘car’, and label images containing segments from those labels. Experts may ignore segments coming from class such as ‘fence’ or ‘door’ depend on the granularity of the labeling process.

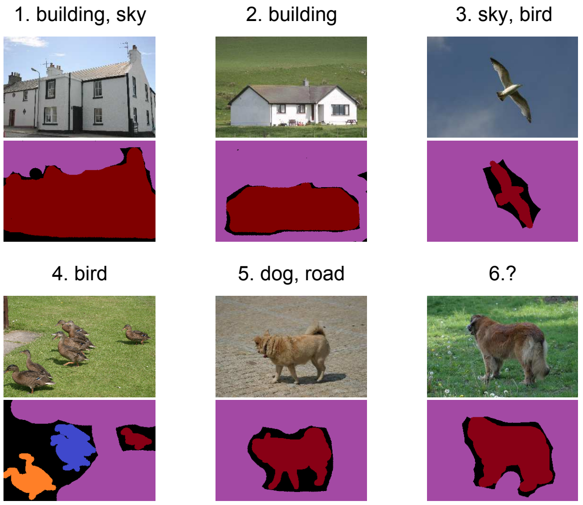


Figure 1. Example of images taken from MSCV2 dataset. Below each image, a segmentation into instances are provided. Segment ids in each image are denoted by pink, red, blue, and orange colors. On top of each training image (numbered from 1 to 5) is its bag label. The label for the test image (numbered 6) is unknown, denoted by ‘?’. The label ‘grass’ is novel since it is not used for any of the images but segments of this class appear in the 2nd and 4th images. In the test image, there is a ‘grass’ segment, which belongs to the novel class. The challenge is to correctly annotate each segment including the novel ‘grass’ segment.

In this paper, we propose to incorporate the novel class into our model so that it can be learned directly during training.

3. Problem Formulation

Setting. We consider the MIML setting in the presence of novel class instances. The training set contains B bags, denoted as $\{\mathbf{X}_D, \mathbf{Y}_D\} = \{(\mathbf{X}_b, \mathbf{Y}_b)\}_{b=1}^{B}$. $\mathbf{X}_b$ consists of n_b instances $\mathbf{x}_{b1}, \mathbf{x}_{b2}, \ldots,$ and $\mathbf{x}_{bn_b}$, $\mathbf{x}_{bi} \in \mathscr{X} = \mathbb{R}^d$. Each instance $\mathbf{x}_{bi}$ is associated with a latent instance label $y_{bi} \in \{0, 1, 2, \ldots, C\}$, where class 0 represents the novel class. The bag label $\mathbf{Y}_b$ is a subset of the set of known classes $\mathscr{Y} = \{1, 2, \ldots, C\}$. Note that the bag label does not include 0. Hence it does not provide information about the presence or absence of the novel class in the bag. In this setting, one can consider the following tasks: (i) Instance annotation: mapping an instance in $\mathscr{X}$ to a label in $\mathscr{Y} \bigcup \{0\}$, i.e., $\{0, 1, \ldots, C\}$. (ii) Novelty detection: mapping an instance in $\mathscr{X}$ to $\{\{0\}, \mathscr{Y}\}$, i.e., determining whether an instance belongs to the novel class or known classes. (iii) Bag level prediction: mapping a bag in $2^{\mathscr{X}}$ to a label in $2^{\mathscr{Y}}$.

Example: Consider the set of images (bags) taken from MSCV2 dataset (Winn et al., 2005) in the MIML format as in Fig. 1. Additionally assume that grass is not included in the set of tags used for annotation and hence segments (instances) from grass can be regarded as novel instances. Training images are numbered from $1 - 5$. For example, in the 1st image, the label is {'building', 'sky'} and there are two segments in the image. However, no information mapping segments to labels is available. Note that images 2 and 4, which contain the novel class ('grass'), do not indicate so in their bag labels. Our goal is to learn a classifier that can map a segment to one of the known classes, i.e., 'building', 'sky', 'bird', 'dog', 'road', or to the novel class. Note that since the classifier is never trained with the label 'grass', at best it is expected to map 'grass' segments to a novel class. The test image numbered 6 contains a segment from the known class 'dog' and a segment from the novel class 'grass'. A key challenge is to design a classifier that can correctly predict the label for segments from known classes and recognize 'grass' segments as belonging to a novel class.

4. Proposed approach

In this section, we introduce a probabilistic model for MIML data that includes unlabeled novel class instances. Additionally, we present a maximum likelihood algorithm for learning the parameters of the model.

4.1. Model

The proposed model addresses the MIML problem in the presence of novel class instances using two fundamental aspects: (i) a class ($c = 0$) is assigned to represent novel class instances. (ii) the bag-level label removes any evidence of the presence of the novel class from the union of instance labels. These features of the model are designed to allow learning a class for which no label is provided during training.

Our model is presented in Fig. 2. We assume that all bag labels and instance labels in each bag are independent. Specifically, we consider the relation between the instance label and feature vector, including novel class, as follows

$$p(y_{bi}|\mathbf{x}_{bi}, \mathbf{w}) = \frac{\prod_{c=0}^{C} e^{I(y_{bi}=c)\mathbf{w}_c^T \mathbf{x}_{bi}}}{\sum_{c=0}^{C} e^{\mathbf{w}_c^T \mathbf{x}_{bi}}}, \tag{1}$$

where $\mathbf{w} = [\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_C]$ and $\mathbf{w}_c \in \mathbb{R}^{d\times 1}$ is the weight for the cth class. For each bag, the union of its instance labels, including the novel label 0, is denoted as $\mathbf{Y}_b^{n_b} = \bigcup_{j=1}^{n_b} y_{bj}$. In addition, the relation between the observed bag label $\mathbf{Y}_b$ and $\mathbf{Y}_b^{n_b}$ is modeled as follows

$$p(\mathbf{Y}_b|\mathbf{Y}_b^{n_b}) = I(\mathbf{Y}_b = \mathbf{Y}_b^{n_b}) + I(\mathbf{Y}_b \bigcup \{0\} = \mathbf{Y}_b^{n_b}), \tag{2}$$

where $\mathbf{Y}_b \subseteq \{1, 2, \ldots, C\}$ and $\mathbf{Y}_b^{n_b} \subseteq \{0, 1, 2, \ldots, C\}$. This model implies that the bag label $\mathbf{Y}_b$ is obtained by removing the novel class label 0 from the union of the instance labels $\mathbf{Y}_b^{n_b}$ if it appears in the union. Hence $\mathbf{Y}_b$ cannot reveal information about the presence of novel class instances in the bag. We illustrate the use of our notation for the example in Fig. 1 using Table 1.

Bag 2	Bag 3	Bag 4
$n_2 = 2$	$n_3 = 2$	$n_4 = 4$
$y_{21} =$'grass' $y_{22} =$'building'	$y_{31} =$'sky' $y_{32} =$'bird'	$y_{41} =$'grass' $y_{42} =$'bird' $y_{43} =$'bird' $y_{44} =$'bird'
$\mathbf{Y}_2^2 =$ {'grass','building'}	$\mathbf{Y}_3^2 =$ {'sky','bird'}	$\mathbf{Y}_4^4 =$ {'grass','bird'}
$\mathbf{Y}_2 =$ {'building'}	$\mathbf{Y}_3 =$ {'sky','bird'}	$\mathbf{Y}_4 =$ {'bird'}

Table 1. Instance labels and bag labels of the proposed model for images 2, 3, and 4 in Fig. 1.

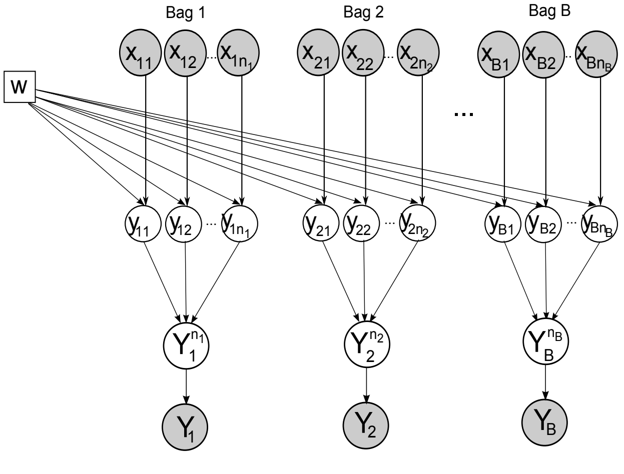


Figure 2. Graphical model for MIML learning in the presence of novel class instances (MIML-NC).

4.2. Maximum likelihood estimation using expectation maximization

To learn the model parameters, we consider maximum likelihood inference and proceed with the evaluation of the likelihood. Since $\mathbf{X}_D$ is independent of the parameters $\mathbf{w}$ and bag labels are independent given $\mathbf{X}_D$, the likelihood for the aforementioned model can be written as

$$p(\mathbf{Y}_D, \mathbf{X}_D|\mathbf{w}) = p(\mathbf{X}_D)\prod_{b=1}^{B} p(\mathbf{Y}_b|\mathbf{X}_b, \mathbf{w}), \qquad (3)$$

where $p(\mathbf{Y}_b|\mathbf{X}_b, \mathbf{w})$ is computed using the law of total probability as

$$p(\mathbf{Y}_b|\mathbf{X}_b, \mathbf{w}) = \sum_{y_{b1}=0}^{C} \cdots \sum_{y_{bn_b}=0}^{C} [\{I(\mathbf{Y}_b = \bigcup_{j=1}^{n_b} y_{bj}) + I(\mathbf{Y}_b \bigcup \{0\} = \bigcup_{j=1}^{n_b} y_{bj})\} \times \prod_{i=1}^{n_b} p(y_{bi}|\mathbf{x}_{bi}, \mathbf{w})]. \qquad (4)$$

Directly maximizing the logarithm of (3), i.e., the log-likelihood, is hard. Instead, we consider the expectation maximization (Dempster et al., 1977) solution for inference. Consider the instance labels $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_B\}$, where $\mathbf{y}_b = [y_{b1}, y_{b2}, \ldots, y_{bn_b}]$ for all $1 \leq b \leq B$, as hidden variables. In EM, the surrogate function $g(\mathbf{w}, \mathbf{w}') = E_{\mathbf{y}}[\log p(\mathbf{Y}_D, \mathbf{X}_D, \mathbf{y}|\mathbf{w})|\mathbf{Y}_D, \mathbf{X}_D, \mathbf{w}']$, i.e., the expectation w.r.t. $\mathbf{y}$ of the complete log-likelihood, is iteratively computed and maximized. The surrogate function for the proposed model is given by

$$g(\mathbf{w}, \mathbf{w}') = \sum_{b=1}^{B}\sum_{i=1}^{n_b}[\sum_{c=0}^{C} p(y_{bi} = c|\mathbf{Y}_b, \mathbf{X}_b, \mathbf{w}')\mathbf{w}_c^T\mathbf{x}_{bi} - \log(\sum_{c=0}^{C} e^{\mathbf{w}_c^T\mathbf{x}_{bi}})] + \zeta, \qquad (5)$$

where $\zeta = E_{\mathbf{y}}[\log p(\mathbf{Y}_D|\mathbf{y})|\mathbf{Y}_D, \mathbf{X}_D, \mathbf{w}'] + \log p(\mathbf{X}_D)$ is a constant w.r.t. $\mathbf{w}$. Detailed steps to obtain $g(\mathbf{w}, \mathbf{w}')$ are given in the supplementary material. The generalized expectation maximization framework consists of two steps as follows

- E-step: Compute $p(y_{bi} = c|\mathbf{Y}_b, \mathbf{X}_b, \mathbf{w}^{(k)})$ for $b = 1, \ldots, B$ and $i = 1, \ldots, n_b$.
- M-step: Find $\mathbf{w}^{(k+1)}$ such that $g(\mathbf{w}^{(k+1)}, \mathbf{w}^{(k)}) \geq g(\mathbf{w}^{(k)}, \mathbf{w}^{(k)})$.

A key challenge in this paper is to address the probability calculation of y_{bi} given $\mathbf{Y}_b$ in the presence of novel class instances. While the formulation of the problem in the EM setting appears straightforward, the computation in the E-step is nontrivial.

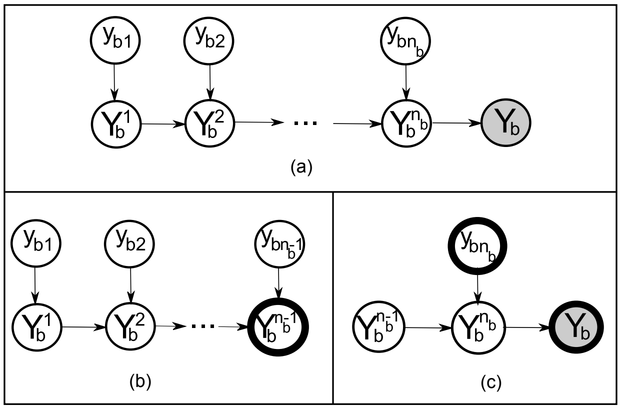


Figure 3. The process to compute $p(y_{bn_b} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w})$. Bolded nodes are nodes which are currently computed. (a) The variables for the bth bag. (b) Dynamically compute $p(\mathbf{Y}_b^{n_b-1}|\mathbf{X}_b, \mathbf{w})$. (c) Compute $p(y_{bn_b}, \mathbf{Y}_b|\mathbf{X}_b, \mathbf{w})$ using Proposition 1.

4.3. E-step

The probability $p(y_{bi} = c|\mathbf{Y}_b = \mathbf{L}, \mathbf{X}_b, \mathbf{w})$, for $c \in \mathbf{L} \cup \{0\}$, can be computed from $p(y_{bi} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w})$ using conditional rule as

$$p(y_{bi} = c|\mathbf{Y}_b = \mathbf{L}, \mathbf{X}_b, \mathbf{w}) = \frac{p(y_{bi} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w})}{\sum_{c\in\mathbf{L}\cup\{0\}} p(y_{bi} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w})}.$$

We proceed with an efficient alternative computation of $p(y_{bi} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w})$ as follows.

Define n_b random variables $\mathbf{Y}_b^1, \mathbf{Y}_b^2, \ldots, \mathbf{Y}_b^{n_b}$, where $\mathbf{Y}_b^i = \bigcup_{j=1}^{i} \mathbf{y}_{bj}$, is the union of labels of the first i instances in the bth bag. Note that this sub-bag label indicates the presence or absence of all classes including the novel class in the sub-bag. Using the newly introduced random variables $\mathbf{Y}_b^i$, we can replace the label portion of the graphical model in Fig. 2 for the bth bag with the chain structure in Fig. 3(a). The advantage of the new graphical model in Fig. 3(a) is that it allows for efficient computation of the desired probability $p(y_{bi} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w})$. Denote the power set of $\mathbf{L}\bigcup\{0\}$ excluding the empty set as $\mathbf{P}$. We derive the following procedure for computing the probability $p(y_{bi} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w})$.

Consider the case $i = n_b$ where $p(y_{bn_b} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w})$ is computed.

Step 1. Dynamically compute $p(\mathbf{Y}_b^{n_b-1}|\mathbf{X}_b, \mathbf{w})$ using the following recursion:

$$p(\mathbf{Y}_b^{j+1} = \mathbf{L}'|\mathbf{X}_b, \mathbf{w}) = \sum_{l\in\mathbf{L}'} p(y_{bj+1} = l|\mathbf{x}_{bj+1}, \mathbf{w}) \times [p(\mathbf{Y}_b^j = \mathbf{L}'_{\setminus l}|\mathbf{X}_b, \mathbf{w}) + p(\mathbf{Y}_b^j = \mathbf{L}'|\mathbf{X}_b, \mathbf{w})], \qquad (6)$$

where $\mathbf{L}' \subseteq \mathbf{P}$ and $\mathbf{L}'_{\backslash l} = \{c \in \mathbf{L}' | c \neq l\}$ (the proof of (6) is given in the supplementary material). This calculation is performed sequentially for every $\mathbf{Y}_b^j$ until $p(\mathbf{Y}_b^{n_b-1}|\mathbf{X}_b, \mathbf{w})$ is obtained as in Fig. 3(b). Note that the presence or absence of the novel class is taken into account when computing $p(\mathbf{Y}_b^{n_b-1} = \mathbf{L}'|\mathbf{X}_b, \mathbf{w})$ since $\mathbf{L}'$ may contain class 0.

Step 2. Compute $p(y_{bn_b}, \mathbf{Y}_b|\mathbf{X}_b, \mathbf{w})$ from $p(\mathbf{Y}_b^{n_b-1}|\mathbf{X}_b, \mathbf{w})$ based on the model in Fig. 3(c), as in the following proposition.

Proposition 1 *The probability $p(y_{bn_b} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w})$ for all $c \in \mathbf{L} \bigcup \{0\}$ can be computed as follows.*

- *If $c = 0$,*

$$p(y_{bn_b} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w}) = p(y_{bn_b} = c|\mathbf{x}_{bn_b}, \mathbf{w})\times$$
$$[p(\mathbf{Y}_b^{n_b-1} = \mathbf{L}|\mathbf{X}_b, \mathbf{w}) + p(\mathbf{Y}_b^{n_b-1} = \mathbf{L}\bigcup\{0\}|\mathbf{X}_b, \mathbf{w})].$$

- *Else if $c \neq 0$,*

$$p(y_{bn_b} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w}) = p(y_{bn_b} = c|\mathbf{x}_{bn_b}, \mathbf{w})\times$$
$$[p(\mathbf{Y}_b^{n_b-1} = \mathbf{L}|\mathbf{X}_b, \mathbf{w}) + p(\mathbf{Y}_b^{n_b-1} = \mathbf{L}\bigcup\{0\}|\mathbf{X}_b, \mathbf{w}) +$$
$$p(\mathbf{Y}_b^{n_b-1} = \mathbf{L}_{\backslash c}|\mathbf{X}_b, \mathbf{w}) + p(\mathbf{Y}_b^{n_b-1} = \mathbf{L}_{\backslash c}\bigcup\{0\}|\mathbf{X}_b, \mathbf{w})].$$

Proof. The detailed proof can be found in the supplementary material. □

Consider the remaining terms of the form $p(y_{bi} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w})$ where $i = 1, 2, \ldots, n_b - 1$. Define $\mathbf{Y}_b^{\backslash i} = \bigcup_{j=1 \neq i}^{n_b} y_{bj}$. For each i: First, swap y_{bi} and y_{bn_b} to have a new order of instances; Next, compute $p(\mathbf{Y}_b^{\backslash i}|\mathbf{X}_b, \mathbf{w})$ using Step 1 based on the new order; Finally, compute $p(y_{bi}, \mathbf{Y}_b|\mathbf{X}_b, \mathbf{w})$ using Step 2. By swapping the ith instance with the last instance, $p(y_{bi}, \mathbf{Y}_b|\mathbf{X}_b, \mathbf{w})$ is evaluated as the last instance when all instances have already been taken into account.

Pseudo code for computing the probability $p(y_{bi} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w})$ is provided in algorithm 1 in the supplementary material. The computational complexity of computing $p(y_{bi} = c|\mathbf{Y}_b, \mathbf{X}_b, \mathbf{w})$, for all $1 \leq i \leq n_b$ and $0 \leq c \leq C$ is $O((|\mathbf{Y}_b| + 1)2^{|\mathbf{Y}_b|+1}n_b^2)$. Although the exponential dependence on the number of labels per bag $2^{|\mathbf{Y}_b|+1}$ may appear as a limitation, in practice many real MIML datasets have a fairly low number of labels per bag (Huang et al., 2014, p. 4) despite a possible large number of classes. Bags with a large number of labels pose inherent challenges since they provide limited information.

4.4. M-step

We apply gradient ascent with backtracking line search to maximize $g(\mathbf{w}, \mathbf{w}')$ w.r.t. $\mathbf{w}$ as follows

$$\mathbf{w}_c^{(k+1)} = \mathbf{w}_c^{(k)} + \frac{\partial g(\mathbf{w}, \mathbf{w}^{(k)})}{\partial \mathbf{w}_c}\bigg|_{\mathbf{w}=\mathbf{w}^{(k)}} \times \eta, \qquad (7)$$

where the gradient w.r.t. $\mathbf{w}_c$, for all $c \in \{0, 1, 2, \ldots, C\}$, $\frac{\partial g(\mathbf{w}, \mathbf{w}^{(k)})}{\partial \mathbf{w}_c}$, is

$$\sum_{b=1}^{B}\sum_{i=1}^{n_b}[p(y_{bi} = c|\mathbf{Y}_b, \mathbf{X}_b, \mathbf{w}^{(k)})\mathbf{x}_{bi} - \frac{e^{\mathbf{w}_c^T\mathbf{x}_{bi}}\mathbf{x}_{bi}}{\sum_{c=0}^{C} e^{\mathbf{w}_c^T\mathbf{x}_{bi}}}]. \quad (8)$$

4.5. Prediction

We perform instance label prediction, bag label prediction, and novel class detection as follows.

Instance annotation. For an unlabeled test instance $\mathbf{x}_{ti}$, the predicted label $\hat{y}_{ti}$ is computed as follows

$$\hat{y}_{ti} = \arg\max_{0 \leq k \leq C} \mathbf{w}_k^T\mathbf{x}_{ti}. \qquad (9)$$

Bag label prediction. The bag label $\hat{\mathbf{Y}}_t$ of a test bag $\mathbf{X}_t$, is computed from its instance labels as $\hat{\mathbf{Y}}_t = (\bigcup_{i=1}^{n_t} \hat{y}_{ti})\backslash\{0\}$, where $\hat{y}_{ti}$ is computed from (9).

Novelty detection. An unlabeled test instance is detected as novel instance if $p(y_{ti} = 0|\mathbf{x}_{ti}, \mathbf{w}) \geq \theta$, where $0 \leq \theta \leq 1$ is a manually selected threshold.

5. Experiments

In this section, we compare our approach with related methods in MIML learning using MIML data that contains novel class instances.

We compare the proposed approach (MIML-NC) with the following *algorithms*: ORed Logistic Regression (ORLR) (Pham et al., 2014), kernel scoring (Lou et al., 2013), and SIM (Briggs et al., 2012) on both real and synthetic datasets. ORLR and SIM methods have been designed for making instance-level predictions for MIML data, but can also be used for bag-level prediction and novel class detection. In order to deal with the case where there are multiple novel classes and there is no fitted linear boundary to separate classes, as in Fig. 4(a), we also consider the kernel versions of the proposed framework and ORLR framework by applying techniques for kernel logistic regression in (Zhu & Hastie, 2001). We use Hamming loss as the *evaluation metric* to compare bag-level prediction results (Zhou et al., 2012). We use AUC (area under the curve) to compare novelty detection results.

5.1. Instance annotation in the novel class setting

In this experiment, we illustrate the rationale of our approach for *instance annotation* in the MIML setting with the presence of novel class instances on a toy dataset.

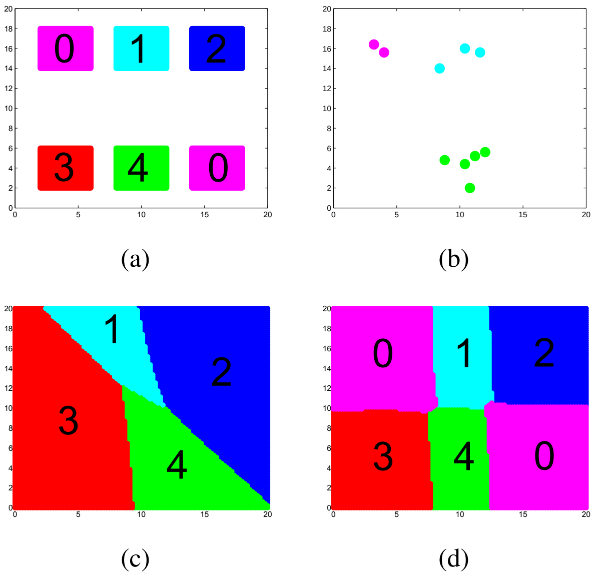
Setting. We generate $B = 100$ bags, each bag contains $n_b = 10$ instances, from six different regions, denoted by six rectangles, as shown in Fig. 4(a). In each bag, the number of instances belonging to each region is drawn from a Dirichlet distribution with equal parameters of 1/6 to allow for class sparseness. On average in each bag there are instances from only 2.5 regions. Note that we consider the two pink regions (top left rectangle and bottom right rectangle in Fig. 4(a)) as a novel class (labeled 0). The labels for 'cyan', 'blue', 'red', 'green' regions are 1, 2, 3, 4, respectively. Note that all bag labels do not contain label 0 and the presence or absence of instances from the pink regions is unavailable.

Results and analysis. The boundaries learned using kernel ORLR and the kernel version of our proposed approach are illustrated as in Fig. 4(c) and 4(d). Although ORLR may classify instances from known classes correctly, it combines the novel class with known classes. In contrast, the proposed method correctly separates both known and unknown classes.

5.2. Experiments on novelty detection

In this experiment, our goal is to show that the proposed algorithm can be effective in *detecting novel instances* in the MIML data. We compare the AUC of the kernel version of our proposed approach with that of the kernel scoring approach (Lou et al., 2013).

Datasets. We use MSCV2, Letter Carroll, and Letter Frost datasets (Briggs et al., 2012), MNIST handwritten dataset (Asuncion & Newman) in these experiments. MSCV2 is a MIML dataset containing 591 images from 23 classes. Each image (bag) contains multiple segments. Each segment is described by a 48-dimensional feature vector and is annotated with a class. The label for an image is the union of its segment labels. Letter Carroll and Letter Frost are also MIML datasets, each is taken from a poem (Briggs et al., 2012). Each word (bag) contains multiple letters. Each letter is described by a 16-dimensional feature vector and is annotated by one of 26 letter labels from 'a' to 'z'. The label for each word is the union of its letter labels. MNIST is a SISL dataset containing 70,000 samples, each sample is a 28 × 28 handwritten digit (from 0-9) image represented by a vector of 784 features. We apply PCA to reduce the dimension of instances from 784 to 20, as in (Lou et al., 2013).



(a) (b) (c) (d)

Figure 4. Toy datasets experiments. (a) The classes distribution. The labels for instances from 'cyan', 'blue', 'red', 'green' regions are 1, 2, 3, 4, respectively, and the label for 'pink' regions is 0 (novel). (b) An example training bag which is labeled $\{1, 4\}$. The label does not indicate the presence of novel instances from pink regions (class 0). (c) Prediction results of ORLR method. (d) Prediction results of the proposed method.

Setting. For MSCV2, Letter Frost, and Letter Carroll datasets, we consider different splits of the classes into known and unknown classes. Specifically, we consider the 1st to 4th classes as novel, the 1st to 8th classes as novel, and the 1st to 16th classes as novel. For each split, we remove the novel class labels from all bag labels, if they appear. From MNIST dataset, we generate MIML datasets containing 100 bags as follows. Each bag contains ten instances and the number of instances sampled from each digit is drawn from a Dirichlet distribution with parameter 1/10. This setting allows sparseness in the digits for each bag. On average, each bag contains 2.6 digits. For MNIST, we consider different splits of the classes into known and unknown. Specifically, we consider three different settings with changing number of novel classes {0, 1}, {0,1,2,3}, and {0,1,2,3,4,5}.

Parameter tuning. When comparing with kernel scoring, we consider the same form of regularization as in (Lou et al., 2013). Specifically, we add a regularization term $\lambda \sum_{c=0}^{C} \mathbf{w}_c^T \mathbf{K} \mathbf{w}_c$ to the objective function in (5), where $\mathbf{K}$ is a kernel matrix in which $\mathbf{K}(i,j) = e^{\frac{-||\mathbf{x}_i - \mathbf{x}_j||^2}{\delta}}$. Adding the regularization term results in adding $2\lambda \mathbf{K} \mathbf{w}_c$ to the gradient in (8). To tune δ, we compute $\bar{d}^2$, the mean square distance for every pair of instances in a dataset.

Datsets	MIML-NC	Kernel Scoring
MSCV2-1to4	**0.82 ± 0.05**	0.68 ± 0.07
MSCV2-1to8	**0.80 ± 0.04**	0.60 ± 0.04
MSCV2-1to16	**0.66 ± 0.07**	**0.64 ± 0.07**
LetterCarroll-1to4	**0.92 ± 0.03**	0.81 ± 0.05
LetterCarroll-1to8	**0.84 ± 0.04**	0.69 ± 0.08
LetterCarroll-1to16	**0.87 ± 0.05**	0.77 ± 0.08
LetterFrost-1to4	**0.94 ± 0.03**	0.84 ± 0.07
LetterFrost-1to8	**0.81 ± 0.05**	0.69 ± 0.08
LetterFrost-1to16	**0.86 ± 0.06**	**0.80 ± 0.09**
MNIST-0to1	**0.90 ± 0.03**	0.83 ± 0.06
MNIST-0to3	**0.88 ± 0.04**	0.78 ± 0.06
MNIST-0to5	**0.84 ± 0.06**	**0.79 ± 0.08**

Table 2. AUC results for the proposed method and kernel scoring. Values that are statistically indistinguishable using the two-tailed paired t-tests at 95% confidence level with the highest performances are bolded.

Then, we set $\delta = s \times \bar{d}^2$, where s is searched over the set $\{1, 2, 5, 10\}$. The parameter $\lambda_{MIML-NC}$ for the kernel version of the proposed approach is fixed at 10^{-5}. The parameter λ_{KS} for the kernel scoring approach is searched over $\{10^{-4}, 10^{-2}, 10^{-1}, 100\}$.

Results and analysis. The performance in terms of AUC of the proposed approach and kernel scoring is shown in Table 2. Note that for the MNIST dataset, we reproduced the experiments when novel class instances are 0,1 in Table 6 of (Lou et al., 2013, p. 5) and achieved similar accuracy. The AUC of the proposed approach is significantly higher than that of the kernel scoring approach. The reason is the kernel scoring approach only uses one maximizing-score instance to represent each bag w.r.t. a class. In contrast, the proposed approach takes into account all instances in each bag. Additionally, unlike the proposed method which directly models the novel class, kernel scoring does not directly model the novel class. Specifically, kernel scoring works as if there are no novel class instances in training. Then, an instance is classified as novel in testing if all of its scores w.r.t. known classes are small. The condition is violated when the novel class is considerably large in training.

5.3. Bag label prediction experiments

In this section, we examine the effectiveness of the proposed approach on *bag label prediction* when data contains novel class instances. We compare the Hamming loss performances of our proposed approach, the ORLR approach (Pham et al., 2014), and the SIM approach (Briggs et al., 2012). We also consider a dummy classifier where the predicted bag label consists of labels appearing in more than 50% of training bags. The dummy approach does not rely on instance features in prediction hence is used to obtain an upper bound on the Hamming loss of all other algorithms compared.

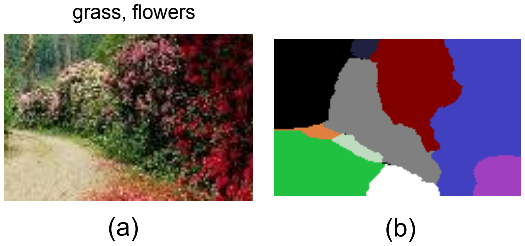


Figure 5. An example image (bag) with novel class instances in Corel5k-10. (a) is the image labeled as 'flowers' and 'grass'. (b) are segments from the image. The green segment is not from either 'flowers' or 'grass'.

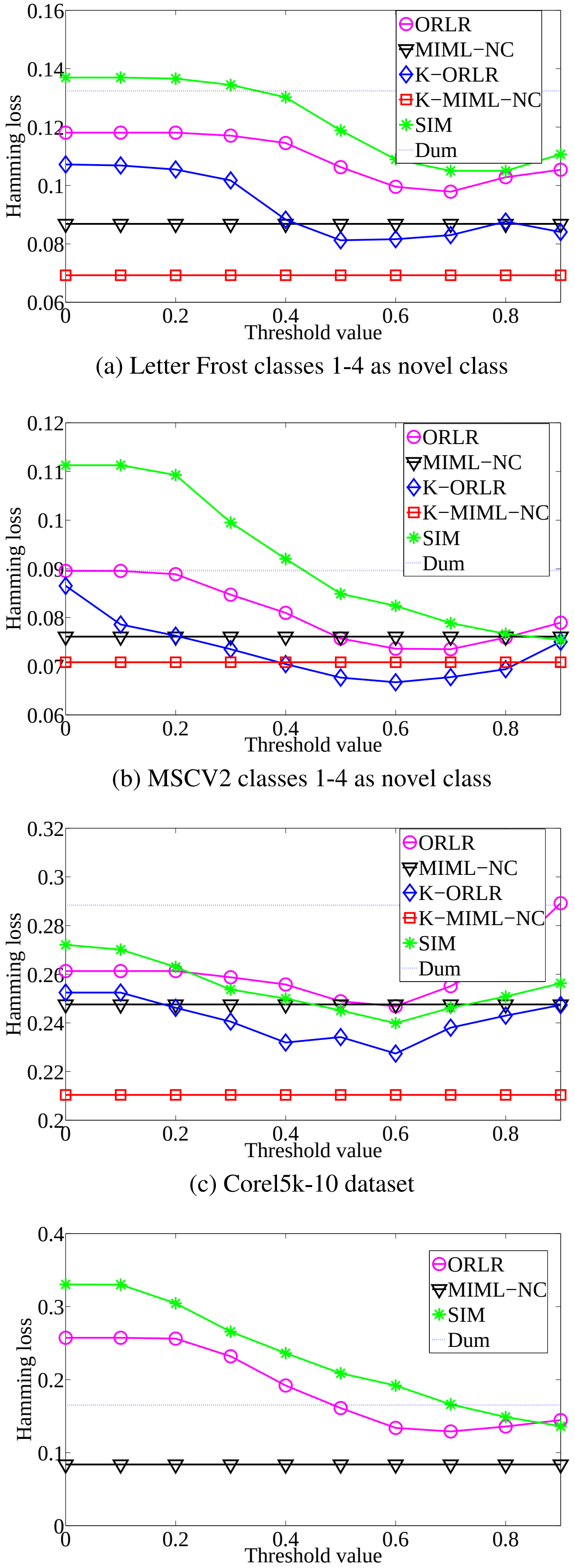
Datasets. We use MSCV2 and Letter Frost described in Section 5.2 for these experiments. Additionally, we consider two more datasets Corel5k and HJA which contain real novel class instances. Corel5k (Duygulu et al., 2002) is a MIML dataset where each image is a bag and its segments are instances. We sort classes based on the number of times they appear in all images. From the sorted list, we then select the top ten classes with highest values: 'water', 'sky', 'tree', 'people', 'grass', 'buildings', 'mountain', 'snow', 'flowers', 'clouds'. We remove bags whose bag label contains any class outside of these ten classes. As a result, instances outside the top ten classes are considered novel. An example image with novel class from the processed Corel5k dataset (Corel5k-10) is shown in Fig. 5. Another dataset is HJA (Briggs et al., 2012) which is a birdsong dataset in MIML format. Each bag is a ten second-recording labeled with a list of bird species singing in this time period. Each instance is a 38-dimensional feature vector extracted from a segment of the spectrogram of the ten second-recording. Instances other than bird sounds or instances which are too costly or difficult to label (Briggs et al., 2012) are unlabeled. Novel class instances such as rain, wind, or insect noises are included in the set of unlabeled instances. Out of a total of 10,232 instances there are only 4,998 instances whose labels are used in forming the bag-level labels. **Setting.** For ORLR and SIM, the bag label is predicted as the union of its instance labels. Since ORLR and SIM are unable to directly deal with novel instances, we use a threshold ϵ when predicting an instance label as novel. From ORLR model, the score of instance $\mathbf{x}_i$ w.r.t. class c is $p(y_i = c|\mathbf{x}_i, \mathbf{w})$ and for SIM model, the score of instance $\mathbf{x}_i$ w.r.t. class c is $\mathbf{w}_c^T \mathbf{x}_i$. We compute the mean and standard deviation for these scores for all instances then normalize these scores from 0 to 1. For each test instance, if the maximal confidence score is less than a threshold ϵ, the instance is considered novel. For ORLR, we search ϵ over the set of $\{0, 0.1, 0.2, \ldots, 0.9\}$ and report the Hamming loss for each value of ϵ. For kernel ORLR, for each ϵ, we search λ and δ as in Section 5.2 and report the lowest Hamming loss. For SIM, for each ϵ, we search the parameter λ (Briggs et al., 2012) over the set of $\{10^{-4}, 10^{-5}, \ldots, 10^{-9}\}$ and report the lowest Hamming loss. Since the matrix $\mathbf{K}$ for HJA is large, we do not run the kernel version of ORLR and MIML-NC on the HJA dataset.

Results and analysis. The Hamming loss of ORLR, MIML-NC, kernel ORLR (K-ORLR), kernel MIML-NC (K-MIML-NC), and SIM are presented in Fig. 6. We observe that the use of a threshold improves bag-level prediction for both ORLR and SIM when novel instances are present in the data. If the threshold ϵ is small, many of the novel class instances may be arbitrarily classified into any of the known classes which may lead to more false positives in the bag-level prediction. If ϵ is large, many instances whose scores are not sufficiently high will be considered novel and consequently their labels will not be included in the predicted bag label leading to many false negatives. The performance of ORLR and SIM can be improved by appropriately selecting the threshold ϵ. However, in some datasets, for example, Letter Frost and HJA, the optimally tuned ORLR or SIM is still outperformed by the proposed approach. The reason is that ORLR and SIM assume no novel class in their model. When the number of novel instances is considerably large, the assumption is violated. Instead, MIML-NC directly models the novel class that may lead to a lower Hamming loss performance. In contrast to SIM and ORLR, the performance of MIML-NC appears in straight lines since it is free of parameter tuning.

6. Conclusion

We presented a novel probabilistic model for MIML learning in the presence of novel class instances. We introduced a novel efficient computational approach for inference. We illustrated the use of the proposed framework for instance annotation on a toy dataset. Moreover, the experiments on MSCV2 image dataset, letter recognition and handwritten digit dataset show that the proposed method achieves a significant higher AUC compared to recent kernel scoring method in novelty detection. Experiments on datasets containing novel class including HJA bird song and Corel5k show that the accuracy in bag label prediction of the proposed approach is higher than those of recent state-of-the-art methods.

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



(a) Letter Frost classes 1-4 as novel class

(b) MSCV2 classes 1-4 as novel class

(c) Corel5k-10 dataset

(d) HJA dataset

Figure 6. Bag label prediction results.

References

Asuncion, A. and Newman, D. UCI machine learning repository.

Briggs, F., Fern, X. Z., and Raich, R. Rank-loss support instance machines for miml instance annotation. In *KDD*, pp. 534–542, 2012.

Cour, T., Sapp, B., and Taskar, B. Learning from partial labels. *JMLR*, 12:1501–1536, 2011.

Da, Q., Yu, Y., and Zhou, Z.-H. Learning with augmented class by exploiting unlabeled data. In *AAAI*, 2014.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

Duygulu, P., Barnard, K., de Freitas, J. F., and Forsyth, D. A. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV 2002*, pp. 97–112. 2002.

Hsiao, K.-J., Xu, K., Calder, J., and Hero, A. O. Multi-criteria anomaly detection using pareto depth analysis. In *NIPS*, pp. 845–853, 2012.

Huang, S.-J., Gao, W., and Zhou, Z.-H. Fast multi-instance multi-label learning. In *AAAI*, 2014.

Liu, L. and Dietterich, T. G. A conditional multinomial mixture model for superset label learning. In *NIPS*, pp. 557–565, 2012.

Lou, Q., Raich, R., Briggs, F., and Fern, X. Z. Novelty detection under multi-label multi-instance framework. In *Machine Learning for Signal Processing, IEEE International Workshop on*, pp. 1–6, 2013.

Markou, M. and Singh, S. Novelty detection: a review-part 1: statistical approaches. *Signal processing*, 83(12): 2481–2497, 2003.

Pham, A. T., Raich, R., and Fern, X. Z. Efficient instance annotation in multi-instance learning. In *Statistical Signal Processing, IEEE Workshop on*, pp. 137–140, 2014.

Saligrama, V. and Zhao, M. Local anomaly detection. In *AISTATS*, pp. 969–983, 2012.

Winn, J., Criminisi, A., and Minka, T. Object categorization by learned universal visual dictionary. In *ICCV*, pp. 1800–1807, 2005.

Yang, S.-J., Jiang, Y., and Zhou, Z.-H. Multi-instance multi-label learning with weak label. In *IJCAI*, pp. 1862–1868, 2013.

Zhang, M.-L. and Zhou, Z.-H. M3MIML: A maximum margin method for multi-instance multi-label learning. In *ICDM*, pp. 688–697, 2008.

Zhao, M. and Saligrama, V. Anomaly detection with score functions based on nearest neighbor graphs. In *NIPS*, pp. 2250–2258, 2009.

Zhou, Z.-H., Zhang, M.-L., Huang, S.-J., and Li, Y.-F. Multi-instance multi-label learning. *Artificial Intelligence*, 176(1):2291–2320, 2012.

Zhu, J. and Hastie, T. Kernel logistic regression and the import vector machine. In *NIPS*, pp. 1081–1088, 2001.