SEMI-SUPERVISED LEARNING IN THE PRESENCE OF NOVEL CLASS INSTANCES

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

School of Electrical Engineering and Computer Science
Oregon State University, Corvallis, Oregon 97331-5501
{phaman,raich,xfern}@eecs.oregonstate.edu

ABSTRACT

In this paper, we present an approach for learning in the semi-supervised setting in the presence of novel class instances. In this setting, data consists of a labeled portion and an unlabeled portion that contains novel class instances along with unlabeled known class instances. Novel class instances are instances from concepts that do not have labeled training examples. This setting is appropriate for the case in which data is abundant and labeling the entire data is prohibitively expensive. We provide a model and an inference framework that allow for a direct control over the portion of novel class instances in the unlabeled data. Experiments on synthetic data demonstrate the usefulness of the proposed approach. Comparison to state-of-the-art approaches for learning in the presence of novel class instances using unlabeled data illustrates the advantage in using the proposed method in term of accuracy.

***Index Terms*—** Novelty detection, semi-supervised learning, sparsity learning, graphical model, dynamic programming

1. INTRODUCTION

In natural datasets, the number of classes in the data often increases with data size. Designing a learning algorithm which addresses the presence of novel classes in the data is essential. For example, in labeling bird songs, efforts are focused on a few species while the data may contain noise artifacts such as rain drops, moving cars, or other bird species. Since such artifacts are not labeled by the experts, they can be viewed as novel class instances. Moreover, due to the cost of labeling, unlabeled bird vocalizations are mixed among the novel class instances.

To the best of our knowledge, this setting is studied in several papers. A maximum margin approach for learning where training data contains both labeled instances from known classes and unlabeled instances from both known and unknown classes is proposed in [1]. The authors propose to tighten the boundary of known classes leaving the remaining space to be labeled as novel, as in Fig. 1. Theoretical analysis for learning in the presence of novel class instances with unlabeled data is considered in [2]. Other papers address this problem under the name of outlier detection [3], novel class [4] or unknown class detection [5] in the semi-supervised setting.

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

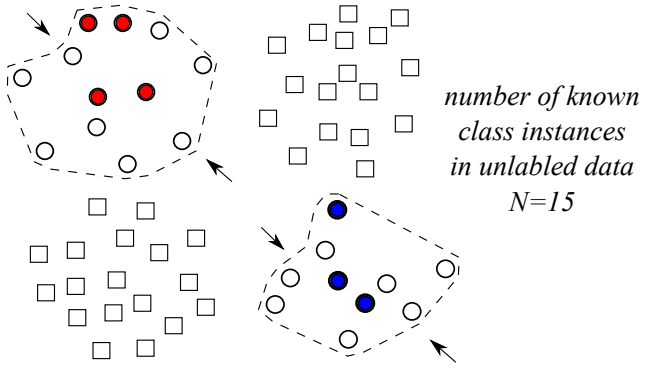


Fig. 1: Example of using unlabeled data for learning in the presence of novel class instances. Labeled instances from two classes are colored with red and blue, respectively. Unlabeled instances are in white. Unlabeled novel class instances are marked with a square and unlabeled known class instances are marked with a circle. Unlabeled data can help to tighten the boundary of known classes (dashed) leaving the remaining space to be labeled as novel.

In this paper, we propose a discriminative probabilistic framework that addresses the challenge of learning in the presence of novel class instances in a semi-supervised setting. We present a novel approach that is based on controlling the sparseness of the known class instances in the mixed unlabeled data pool. The challenging inference problem associated with the dependence structure created by the sparsity control is resolved using an efficient dynamic programming approach. Experiments on toy data and evaluations with state-of-the-art approaches on real world data illustrate the effectiveness of the proposed method.

2. PROBLEM FORMULATION

This paper considers the following setting for learning under novel class instances. We are given a set of labeled instances, represented by $\{\mathbf{x}_i, y_i\}_{i=1}^{L}$, where $\mathbf{x}_i \in \mathscr{X} = \mathbb{R}^d$ is the feature vector and $y_i \in \{1, 2, \ldots, C\}$ is the label of the ith labeled instance, with the total number of C known classes. The input also contains a set of unlabeled instances, represented by $\{\mathbf{x}_{L+i}\}_{i=1}^{U}$. Among U unlabeled instances, assume that there are N instances from known classes, $N \leq U$. Denote the novel class as 0. The task is to learn a classifier that maps an instance in $\mathscr{X}$ to a label in $\mathscr{Y} = \{0, 1, 2, \ldots, C\}$. Sparseness is often characterized by the number of nonzero entries in a given vector. In this paper, sparseness is related to the number of known class (nonzero class) instances contained in the unlabeled data.

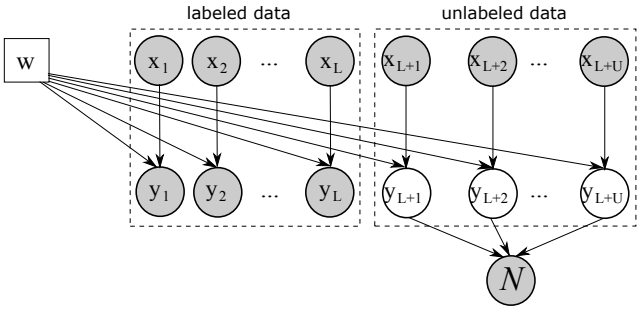


Fig. 2: Graphical model for learning in the presence of novel class instances using unlabeled data. Shaded nodes denote observed variables. N is used to control the number of known class instances in the unlabeled data.

3. PROPOSED MODEL

The proposed graphical model is shown in Fig. 2. Assume that instance labels are independent given instance features. The relation between instance feature $\mathbf{x}_i$ and instance label y_i, $1 \leq i \leq L+U$, is modeled by a multinomial logistic regression function as follows

$$p(y_i|\mathbf{x}_i, \mathbf{w}) = \frac{\prod_{c=0}^{C} e^{\mathbb{I}[y_i=c]\mathbf{w}_c^T\mathbf{x}_i}}{\sum_{c=0}^{C} e^{\mathbf{w}_c^T\mathbf{x}_i}}, \tag{1}$$

where $\mathbf{w} = [\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_C]$, such that $\mathbf{w}_k \in \mathbb{R}^d$ for $0 \leq k \leq C$, is the parameter to learn, and $\mathbb{I}[\cdot]$ is the indicator function. Additionally, the relation between the number of known class instances N and the hidden labels $y_{L+1}, y_{L+2}, \ldots, y_{L+U}$ is

$$p(N|y_{L+1}, y_{L+2}, \ldots, y_{L+U}) = \mathbb{I}[N = \sum_{i=1}^{U} \mathbb{I}[y_{L+i} \neq 0]]. \tag{2}$$

Even though N is unknown, we treat it as observed as shown in Fig. 2, which can be tuned as a hyperparameter.

4. MAXIMUM LIKELIHOOD FOR INFERENCE

Maximum likelihood is used to estimate the model parameters. For a simple representation, we denote the observed labeled data by $\{\mathbf{X}_L, \mathbf{y}_L\} \triangleq \{\mathbf{x}_i, y_i\}_{i=1}^{L}$ and the observed unlabeled data by $\mathbf{X}_U \triangleq \{\mathbf{x}_{L+i}\}_{i=1}^{U}$. Using the conditional rule and the independence assumption between the labeled and unlabeled portions of the data, the probability of the observations given the parameters is

$$p(\mathbf{X}_L, \mathbf{y}_L, \mathbf{X}_U, N|\mathbf{w}) = p(\mathbf{y}_L|\mathbf{X}_L, \mathbf{w})p(N|\mathbf{X}_U, \mathbf{w})p(\mathbf{X}_L, \mathbf{X}_U|\mathbf{w}). \tag{3}$$

Assuming that $\mathbf{X}_L$ and $\mathbf{X}_U$ are independent of $\mathbf{w}$, as in Fig. 2, $p(\mathbf{X}_L, \mathbf{X}_U|\mathbf{w})$ is a constant w.r.t. $\mathbf{w}$. Consequently, maximizing (3) is equivalent to maximizing the product of $p(\mathbf{y}_L|\mathbf{X}_L, \mathbf{w})p(N|\mathbf{X}_U, \mathbf{w})$ only. Taking the logarithm of (3), we obtain the log-likelihood

$$\mathbf{L}(\mathbf{w}) = \sum_{i=1}^{L} \log p(y_i|\mathbf{x}_i, \mathbf{w}) + \log p(N|\mathbf{X}_U, \mathbf{w}) + \zeta, \tag{4}$$

where $\zeta = \log p(\mathbf{X}_L, \mathbf{X}_U|\mathbf{w})$ is a constant w.r.t. $\mathbf{w}$. Note that (4) utilizes the independence assumption among labeled instances. We compute $p(N|\mathbf{X}_U, \mathbf{w})$ by marginalizing over the hidden labels

$$\begin{aligned} &p(N|\mathbf{X}_U, \mathbf{w}) \\ &= \sum_{y_{L+1}=0}^{C} \sum_{y_{L+2}=0}^{C} \cdots \sum_{y_{L+U}=0}^{C} p(N, y_{L+1}, \ldots, y_{L+U}|\mathbf{X}_U, \mathbf{w}) \\ &= \sum_{y_{L+1}=0}^{C} \sum_{y_{L+2}=0}^{C} \cdots \sum_{y_{L+U}=0}^{C} \Big(I[N = \sum_{i=1}^{U} I[y_{L+i} \neq 0]] \times \\ &\qquad \prod_{i=1}^{U} p(y_{L+i}|\mathbf{x}_{L+i}, \mathbf{w}) \Big), \end{aligned} \tag{5}$$

where the last equation follows the independence assumption among unlabeled instances and the relation between N and hidden labels of unlabeled instances in (2). Replacing (5) into (4) yields the explicit expression for the log-likelihood. To the best of our knowledge, no efficient closed-form solution for the direct maximization of the log-likelihood $\mathbf{L}(\mathbf{w})$ is available. Thus, we apply the Expectation Maximization [6] framework to maximize $\mathbf{L}(\mathbf{w})$.

4.1. Expectation maximization

Denote $\mathbf{y}_U$ as a vector of hidden variables $\{y_{L+i}\}_{i=1}^{U}$ and the observed data $\mathbb{D} \triangleq \{\mathbf{X}_L, \mathbf{y}_L, \mathbf{X}_U, N\}$. The surrogate function $g(\mathbf{w}, \mathbf{w}')$ for $\mathbf{L}(\mathbf{w})$ is computed as follows

$$\begin{aligned} g(\mathbf{w}, \mathbf{w}') &= E_{\mathbf{y}_U|\mathbb{D},\mathbf{w}'} \log p(\mathbf{y}_L, \mathbf{y}_U, N|\mathbf{X}_L, \mathbf{X}_U, \mathbf{w}) \\ &= E_{\mathbf{y}_U|\mathbb{D},\mathbf{w}'}[\log p(\mathbf{y}_L|\mathbf{X}_L, \mathbf{w}) + \log p(\mathbf{y}_U|\mathbf{X}_U, \mathbf{w})] \\ &\quad + E_{\mathbf{y}_U|\mathbb{D},\mathbf{w}'} \log p(N|\mathbf{y}_U, \mathbf{X}_U, \mathbf{w}). \end{aligned} \tag{6}$$

In (6), $E_{\mathbf{y}_U|\mathbb{D},\mathbf{w}'} \log p(N|\mathbf{y}_U, \mathbf{X}_U, \mathbf{w}) = E_{\mathbf{y}_U|\mathbb{D},\mathbf{w}'} \log p(N|\mathbf{y}_U) = \alpha$ is a constant w.r.t. $\mathbf{w}$. Using the independence assumption among labeled and unlabeled instances, $p(y_L|\mathbf{X}_L, \mathbf{w}) = \prod_{i=1}^{L} p(y_i|\mathbf{x}_i, \mathbf{w})$ and $p(y_U|\mathbf{X}_U, \mathbf{w}) = \prod_{i=1}^{U} p(y_{L+i}|\mathbf{x}_{L+i}, \mathbf{w})$. Hence, (6) becomes

$$\begin{aligned} &g(\mathbf{w}, \mathbf{w}') = \sum_{i=1}^{L} \log p(y_i|\mathbf{x}_i, \mathbf{w}) \\ &+ E_{\mathbf{y}_U|\mathbb{D},\mathbf{w}'} \Big(\sum_{i=1}^{U} \sum_{c=0}^{C} \mathbb{I}[y_{L+i} = c] \log p(y_{L+i} = c|\mathbf{x}_{L+i}, \mathbf{w}) \Big) + \alpha \\ &= \sum_{i=1}^{L} \log p(y_i|\mathbf{x}_i, \mathbf{w}) \\ &+ \sum_{i=1}^{U} \sum_{c=0}^{C} p(y_{L+i} = c|N, \mathbf{X}_U, \mathbf{w}') \log p(y_{L+i} = c|\mathbf{x}_{L+i}, \mathbf{w}) + \alpha. \end{aligned} \tag{7}$$

Replacing $p(y_i|\mathbf{x}_i, \mathbf{w})$ from (1) into (7), yields

$$\begin{aligned} &g(\mathbf{w}, \mathbf{w}') = \sum_{i=1}^{L} [\sum_{c=0}^{C} \mathbb{I}[y_i = c]\mathbf{w}_c^T\mathbf{x}_i - \log(\sum_{c=0}^{C} e^{\mathbf{w}_c^T\mathbf{x}_i})] + \\ &\sum_{i=1}^{U} \sum_{c=0}^{C} p(y_{L+i} = c|N, \mathbf{X}_U, \mathbf{w}')[\mathbf{w}_c^T\mathbf{x}_{L+i} - \log(\sum_{l=0}^{C} e^{\mathbf{w}_l^T\mathbf{x}_{L+i}})] + \alpha. \end{aligned} \tag{8}$$

We use generalized EM [7] to increase $\mathbf{L}(\mathbf{w})$ by increasing $g(\mathbf{w}, \mathbf{w}')$ rather than maximizing $g(\mathbf{w}, \mathbf{w}')$ at every iteration. As a result, we have the following EM update equations

- E-step: Compute the class membership probabilities for instances of the unlabeled data $p(y_{L+i} = c|N, \mathbf{X}_U, \mathbf{w}^{(k)})$, $1 \leq i \leq U$ and $0 \leq c \leq C$.
- M-step: Find $\mathbf{w}^{(k+1)}$ s.t. $g(\mathbf{w}^{(k+1)}, \mathbf{w}^{(k)}) \geq g(\mathbf{w}^{(k)}, \mathbf{w}^{(k)})$.

4.2. E-step

The class membership probability $p(y_{L+i} = c|N, \mathbf{X}_U, \mathbf{w})$ is computed from $p(y_{L+i} = c, N|\mathbf{X}_U, \mathbf{w})$ using the conditional rule as

$$p(y_{L+i} = c|N, \mathbf{X}_U, \mathbf{w}) = \frac{p(y_{L+i} = c, N|\mathbf{X}_U, \mathbf{w})}{\sum_{l=0}^{C} p(y_{L+i} = l, N|\mathbf{X}_U, \mathbf{w})}. \quad (9)$$

Due to the dependence between the y_{L+i}'s given N, the computation of $p(y_{L+i} = c, N|\mathbf{X}_U, \mathbf{w})$ is nontrivial. To simplify this derivation, we introduce an equivalent graphical model (see Fig. 3(a)). Denote $n_i = \sum_{k=1}^{i} \mathbb{I}[y_{L+k} \neq 0]$ as the number of known class instances from the 1st to the ith instances in the unlabeled data. Additionally, denote $n_U^{\setminus i}$ as the number of known class instances in the unlabeled data excluding the $(L+i)$th instance. Based on this notation, we compute $p(y_{L+i} = c, n_U = N|\mathbf{X}_U, \mathbf{w}), 0 \leq c \leq C, 1 \leq i \leq U$ as follows

• **Step 1**. Compute $p(n_U|\mathbf{X}_U, \mathbf{w})$ as follows
① Initialize $p(n_1|\mathbf{X}_U, \mathbf{w})$
$$\begin{cases} p(n_1 = 0|\mathbf{X}_U, \mathbf{w}) = & p(y_{L+1} = 0|\mathbf{x}_{L+1}, \mathbf{w}), \\ p(n_1 = 1|\mathbf{X}_U, \mathbf{w}) = & \sum_{c\neq 0} p(y_{L+1} = c|\mathbf{x}_{L+1}, \mathbf{w}). \end{cases}$$

② Compute $p(n_{i+1}|\mathbf{X}_U, \mathbf{w})$ from $p(n_i|\mathbf{X}_U, \mathbf{w})$ as follows

$$\begin{aligned} &p(n_{i+1} = k|\mathbf{X}_U, \mathbf{w}) = \quad (10) \\ &p(n_i = k|\mathbf{X}_U, \mathbf{w}) p(y_{L+i+1} = 0|\mathbf{x}_{L+i+1}, \mathbf{w}) + \\ &\mathbb{I}[k \geq 1] \times p(n_i = k-1|\mathbf{X}_U, \mathbf{w}) \sum_{c\neq 0} p(y_{L+i+1} = c|\mathbf{x}_{L+i+1}, \mathbf{w}), \end{aligned}$$

for $0 \leq k \leq N$. The intuition behinds (10) is that if there are k known class instances in the first $(i+1)$ instances of the unlabeled data, then two mutually exclusive events may exist. One event is when there are k known class instances in the first i instances meanwhile the $(i+1)$th instance is novel. Another event is when there are $(k-1)$ known class instances in the first i instances and the $(i+1)$th instance is also a known class instance.

• **Step 2**. From $p(n_U|\mathbf{X}_U, \mathbf{w})$, compute $p(n_U^{\setminus i}|\mathbf{X}_U, \mathbf{w})$ using forward substitution method [8] as follows
① Compute $p(n_U^{\setminus i} = 0|\mathbf{X}_U, \mathbf{w})$

$$p(n_U^{\setminus i} = 0|\mathbf{X}_U, \mathbf{w}) = \frac{p(n_U = 0|\mathbf{X}_U, \mathbf{w})}{p(y_{L+i} = 0|\mathbf{x}_{L+i}, \mathbf{w})}. \quad (11)$$

② Compute $p(n_U^{\setminus i} = k|\mathbf{X}_U, \mathbf{w}), 1 \leq k \leq N$ as follows

$$\begin{aligned} p(n_U^{\setminus i} = k|\mathbf{X}_U, \mathbf{w}) &= \frac{p(n_U = k|\mathbf{X}_U, \mathbf{w})}{p(y_{L+i} = 0|\mathbf{x}_{L+i}, \mathbf{w})} \\ &- \frac{p(n_U^{\setminus i} = k-1|\mathbf{X}_U, \mathbf{w}) \sum_{c\neq 0} p(y_{L+i} = c|\mathbf{x}_{L+i}, \mathbf{w})}{p(y_{L+i} = 0|\mathbf{x}_{L+i}, \mathbf{w})}. \quad (12) \end{aligned}$$

• **Step 3**. From $p(n_U^{\setminus i}|\mathbf{X}_U, \mathbf{w})$, compute $p(y_{L+i} = c, n_U = N|\mathbf{X}_U, \mathbf{w})$ as follows
$p(y_{L+i} = c, n_U = N|\mathbf{X}_U, \mathbf{w}) =$
$$\begin{cases} p(y_{L+i} = c|\mathbf{x}_{L+i}, \mathbf{w}) p(n_U^{\setminus i} = N-1|\mathbf{X}_U, \mathbf{w}), & \text{if } c \neq 0, \\ p(y_{L+i} = c|\mathbf{x}_{L+i}, \mathbf{w}) p(n_U^{\setminus i} = N|\mathbf{X}_U, \mathbf{w}). & \text{if } c = 0. \end{cases}$$

The graphical model for the E-step is shown in Fig. 3. The computational complexity of the E-step is $O(NU + CdU)$. Note that the idea of converting a V structure to a chain structure has been investigated for the OR relation [9] [10]. This paper introduces the dynamic programming technique for the addition relation instead.

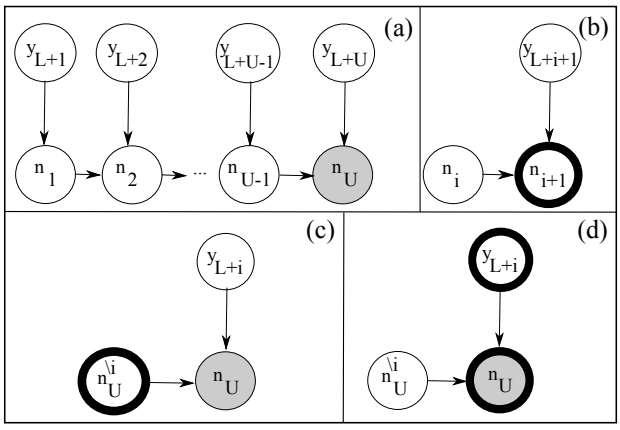


Fig. 3: Graphical models for the E-step. Bolded nodes denote nodes that are currently considered. Shaded nodes denote observed variables. (a) The graphical model with new variables. (b) **Step 1**: compute $p(n_U|\mathbf{X}_U, \mathbf{w})$ from $p(n_1|\mathbf{X}_U, \mathbf{w})$ using dynamic programming. (c) **Step 2**: compute $p(n_U^{\setminus i}|\mathbf{X}_U, \mathbf{w})$ from $p(n_U|\mathbf{X}_U, \mathbf{w})$ and $p(y_{L+i}|\mathbf{x}_{L+i}, \mathbf{w})$. (d) **Step 3**: compute $p(y_{L+i} = c, n_U = N|\mathbf{X}_U, \mathbf{w})$ from $p(n_U^{\setminus i}|\mathbf{X}_U, \mathbf{w})$ and $p(y_{L+i}|\mathbf{x}_{L+i}, \mathbf{w})$.

4.3. M-step

We use gradient ascent [11] to increase the surrogate function in (8). Specifically, the parameter $\mathbf{w}$ is updated as follows

$$\mathbf{w}_c^{(k+1)} = \mathbf{w}_c^{(k)} + \eta \times \frac{\partial g(\mathbf{w}, \mathbf{w}^{(k)})}{\partial \mathbf{w}_c}\bigg|_{\mathbf{w}=\mathbf{w}^{(k)}}, 0 \leq c \leq C, \quad (13)$$

where the formula for the gradient $\frac{\partial g(\mathbf{w}, \mathbf{w}^{(k)})}{\partial \mathbf{w}_c}$ is given by

$$\begin{aligned} \frac{\partial g(\mathbf{w}, \mathbf{w}^{(k)})}{\partial \mathbf{w}_c} &= \sum_{i=1}^{L} [\mathbb{I}[y_i = c] - p(y_i = c|\mathbf{x}_i, \mathbf{w})]\mathbf{x}_i + \quad (14) \\ &\sum_{i=1}^{U} [p(y_{L+i} = c|N, \mathbf{x}_{L+i}, \mathbf{w}^{(k)}) - p(y_{L+i} = c|\mathbf{x}_{L+i}, \mathbf{w})]\mathbf{x}_{L+i}, \end{aligned}$$

and η in (13) is computed using backtracking line search [11].

4.4. Prediction

The predicted label $\hat{y}_t$ of a test instance $\mathbf{x}_t$ is computed as follows

$$\hat{y}_t = \arg \max_{0 \leq c \leq C} p(y_t = c|\mathbf{x}_t, \mathbf{w}), \quad (15)$$

where $p(y_t = c|\mathbf{x}_t, \mathbf{w})$ is given in (1).

5. EXPERIMENTS

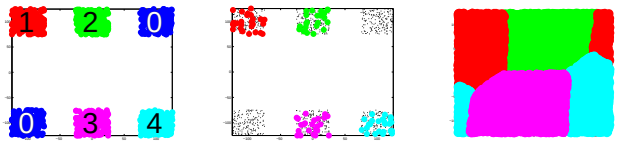
In this section, we evaluate the proposed approach on both synthetic and real world datasets.

5.1. Classification in the presence of novel class instances

Setting. We use a 2D toy example to illustrate the mechanism of the proposed approach. In the example, there are six regions representing five classes, as shown in Fig. 4(a). Specifically, red, green, pink, and cyan regions represent classes 1, 2, 3, and 4, respectively, and the

two blue regions represent class 0 (novel class). From that class distribution, we generate a dataset, as shown in Fig. 4(b). There are 100 labeled instances from the known classes (represented by red, green, pink, and cyan dots) and 1000 unlabeled instances (represented by small black dots). The instances are uniformly generated, therefore, the number of known class instances in the unlabeled data is around $(1000 \times 4)/6 \approx 666$. The proposed approach is repeated for N in the set $\{950, 800, 650, 500\}$. Note that since the data is not linearly separable, we use the kernel technique with random Fourier features that transforms $\mathbf{x}$ to $\phi(\mathbf{x})$ as in [12].

Results and analysis. The learned classifiers when N is 950, 800, 650, and 500 are shown in Fig. 4(c-f), respectively. When N decreases, (indicating that the number of novel class instances in the unlabeled data increases), the proposed method leaves more space for the novel class. For instance, when $N = 950 > 666$, there is no space for novel class. However, when $N = 500 < 666$, there is a larger space for novel class. When N is correctly tuned (around 666), partitions which correctly cover the original regions are found.



(a) (b) (c) $N = 950$

(d) $N = 800$ (e) $N = 650$ (f) $N = 500$

Fig. 4: Results on a toy dataset. (a) Class distribution, novel class instances are blue. (b) Labeled instances (red, pink, green, and cyan dots) and unlabeled instances (small black dots). (c-f) Learned boundary when $N = 950$, $N = 800$, $N = 650$, and $N = 500$, respectively.

5.2. Experiments on real world datasets

Setting. We compare the proposed logistic regression in the semi-supervised setting (LRSSS) with a state-of-the-art method for learning in the presence of novel class instances by using unlabeled data LACU [1]. LACU uses an SVM solver and several regularization parameters to approximately control the boundary between known classes and novel class. The parameter setting for LACU follows [1]. Specifically, η is searched from 1.1 to 1.5 with the step size of 0.05, λ is searched from 0.05 to 0.45 with the step size of 0.05, C is searched in the set $\{0.01, 0.1, 1, 10, 100\}$, and the ramp loss s is searched in the set $\{-0.7, -0.5, -0.3, -0.1\}$. Note that the proposed approach requires tuning of a single parameter, the number of known class instances in the unlabeled data N. Specifically, N is an integer searched from 0.01 to 0.7 (with the step size of 0.01) $\times$ the number of unlabeled instances.

Generation of partially labeled datasets. We perform our evaluation on HJA bird song, MSCV2 [13], 50Salad [14], and MNIST handritten datasets [15] and select some of the classes to represent the novel class. For MNIST, we apply PCA to reduce the feature dimension from 784 to 50. For each dataset, using the classes order of the original datasets, we select instances from 4 classes: from the 1st to 4th classes where the 4th class is selected as novel class and the 1st to 3rd classes are known classes. Moreover, 100 labeled instances and 400 unlabeled instances are randomly selected 10 times leading to 10 train-test evaluations. The small number of labeled and unlabeled instances is due to the memory limitation of LACU. The mean and standard deviation of prediction on unlabeled data over the 10 evaluations are reported.

Tuning N for LRSSS. To select N, we train LRSSS on different values of N to obtain classifier $h_N(\cdot)$ from $\mathscr{X}$ to $\mathscr{Y} = \{0, 1, 2, \ldots, C\}$ for each N. We then use labeled data as a validation set. The accuracy of each h_N on the validation set is acc_N. Consider the example in Fig. 4, with a large N, acc_N is very high ($\rightarrow 1$), where the boundary between known classes are correctly predicted. Even though the space for the novel class is designated for known classes, the validation data contains only known class instances, hence, there is no error. When N becomes smaller than the true number of known class instances, h_N classifies some known class instances as novel, as in Fig. 4(f), therefore reducing acc_N. The selected value for N is at the estimated knee of the acc_N curve.

Results and analysis. The accuracy results for the methods under consideration on the considered datasets are shown in Table 1. **LRSSS-opt** is LRSSS with the value of N giving the highest accuracy. **LRSSS-tune** is LRSSS with the value of N selected from aforementioned method. LR-L is a logistic regression classifier trained with labeled data only and unaware of novel class. **LRSSS-true** is LRSSS given the correct number of N in the unlabeled data. From Table 1, LRSSS-opt significantly outperforms LACU in term of accuracy. It is due to the fact that LRSSS-opt uses an exact method, without relaxation to control the boundary between known classes and novel class. The accuracy results of LRSSS-tune are comparable to those of LRSSS-opt except for MNIST dataset. The reason is due to the small number of instances in the validation set that cannot fully capture the true distribution of known classes. LR-L achieves lower accuracy results compared to LRSSS-opt since LR-L is unaware of the novel class. The accuracy results of LRSSS-true are comparable to those of LRSSS-opt except on MNIST, where the small amount of labeled instances may not guide the learner perfectly even with the true value of N.

Dataset	HJA bird	MSCV2	50Salad	MNIST
LRSSS-opt	**74.4±1.7**	**77.8±2.2**	**72.8±2.2**	**82.0±3.1**
LRSSS-tune	73.2±2.5	69.0±1.5	69.0±2.7	69.2±3.3
LACU	51.6±8.8	65.8±7.1	65.7±7.5	79.4±3.3
LR-L	54.1±2.1	73.2±1.7	42.1±1.9	67.7±2.4
LRSSS-true	74.3±1.3	78.0±2.1	70.1±2.5	78.5±3.3

Table 1: Accuracy results of the proposed approach and LACU. The proposed method and indistinguishable values using 95% confidence two-tailed paired t-tests with the highest values are bolded.

6. CONCLUSION

This paper proposed an approach for learning in the presence of novel class instances in the semi-supervised setting. A discriminative probabilistic model and the corresponding inference are proposed. The approach allows a direct control over the number of known class instances in the unlabeled data. Experiments on synthetic and real data illustrated the usefulness of the method. Compared with state-of-the-art approach on the same problem setting, the proposed method achieves higher accuracy in most cases. Future direction includes how to reduce the computational complexity of calculating the membership probability in the E-step as well as how to more effectively tune the N parameter.

7. REFERENCES

[1] Q. Da, Y. Yu, and Z.-H. Zhou, "Learning with augmented class by exploiting unlabeled data," in *AAAI Conference on Artificial Intelligence*, 2014, pp. 1760–1766.

[2] C. Scott and G. Blanchard, "Novelty detection: Unlabeled data definitely help," in *International Conference on Artificial Intelligence and Statistics*, 2009, pp. 464–471.

[3] D. J. Miller and J. Browning, "A mixture model and EM-based algorithm for class discovery, robust classification, and outlier rejection in mixed labeled/unlabeled data sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 11, pp. 1468–1483, 2003.

[4] F. Nie, S. Xiang, Y. Liu, and C. Zhang, "A general graph-based semi-supervised learning with novel class discovery," *Neural Computing and Applications*, vol. 19, no. 4, pp. 549–555, 2010.

[5] S. J. Frame and S. R. Jammalamadaka, "Generalized mixture models, semi-supervised learning, and unknown class inference," *Advances in Data Analysis and Classification*, vol. 1, no. 1, pp. 23–38, 2007.

[6] 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.

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

[8] G. H. Golub and C. F. Van Loan, *Matrix computations*, vol. 3, The Johns Hopkins University Press, 2012.

[9] 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.

[10] 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.

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

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

[13] 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.

[14] S. Stein and S. J. McKenna, "User-adaptive models for recognizing food preparation activities," in *Proceedings of the 5th international workshop on Multimedia for cooking & eating activities*. ACM, 2013, pp. 39–44.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.