Binary Classification from Multiple Unlabeled Datasets via Surrogate Set Classification

Nan Lu 1 2, Shida Lei 1, Gang Niu 2, Issei Sato 1 2, Masashi Sugiyama 2 1

Abstract

To cope with high annotation costs, training a classifier only from weakly supervised data has attracted a great deal of attention these days. Among various approaches, strengthening supervision from completely unsupervised classification is a promising direction, which typically employs class priors as the only supervision and trains a binary classifier from unlabeled (U) datasets. While existing risk-consistent methods are theoretically grounded with high flexibility, they can learn only from two U sets. In this paper, we propose a new approach for binary classification from m U sets for m ≥ 2. Our key idea is to consider an auxiliary classification task called surrogate set classification (SSC), which is aimed at predicting from which U set each observed sample is drawn. SSC can be solved by a standard (multi-class) classification method, and we use the SSC solution to obtain the final binary classifier through a certain linear-fractional transformation. We built our method in a flexible and efficient end-to-end deep learning framework and prove it to be classifier-consistent. Through experiments, we demonstrate the superiority of our proposed method over state-of-the-art methods.

1. Introduction

Deep learning with large-scale supervised training data has shown great success on various tasks (Goodfellow et al., 2016). However, in practice, obtaining strong supervision, e.g., the complete ground-truth labels, for big data is very costly due to the expensive and time-consuming manual annotations (Zhou, 2018). Thus, it is desirable for machine learning techniques to work with weaker forms of supervision, such as noisy labels (Natarajan et al., 2013; Patrini et al., 2017; Van Rooyen & Williamson, 2018; Han et al., 2018; 2020; Fang et al., 2020; Xia et al., 2020), partial labels (Cour et al., 2011; Ishida et al., 2017; 2019; Feng et al., 2020; Lv et al., 2020), and pairwise comparison information (Bao et al., 2018; Xu et al., 2019; Feng et al., 2021).

This paper focuses on a challenging setting which we call $U^m$ classification: the goal is to learn a binary classifier from m (m ≥ 2) sets of U data with different class priors, i.e., the proportion of positives in each U set. Such a learning scheme can be conceivable in many real-world scenarios. For example, U sets with different class priors can be naturally collected from spatial or temporal differences. Considering morbidity rates, they can be potential patient data collected from different areas (Croft et al., 2018). Likewise, considering approval rates, they can be unlabeled voter data collected in different years (Newman, 2003). In such cases, individual labels are often not available due to privacy reasons, but the corresponding class priors of U sets, i.e., the morbidity rates or approval rates in the aforementioned examples, can be obtained from related medical reports or pre-existing census (Quadrianto et al., 2009; Ardehaly & Culotta, 2017; Tokunaga et al., 2020), and is the unique weak supervision that will be leveraged in this work.

Breakthroughs in $U^m$ classification research were brought by Menon et al. (2015) and Lu et al. (2019) in proposing the risk-consistent methods given two U sets. Recently, Scott & Zhang (2020) extended them to incorporate multiple U sets by two steps: firstly, pair all the U sets so that they are sufficiently different in each pair; secondly, linearly combine the unbiased balanced risk estimators obtained from each pair. Although this method is advantageous since it is compatible with any model and stochastic optimizer, and is statistically consistent, there are several issues that may limit its potential for practical use: first, the computational complexity for the optimal pairing strategy is $O(m^3)$ for m U sets (Edmonds & Karp, 1972), which cannot work efficiently with a large number of U sets; second, the optimal combination weights are proved with strong model assumptions and thus remaining difficult to be tuned in practice.

Now, a natural question arises: can we propose a computationally efficient method for $U^m$ classification with both flexibility on the choice of models and optimizers and theo-
retical guarantees? The answer is affirmative.

In this paper, we provide a new approach for $U^m$ classification by solving a Surrogate Set Classification task ($U^m$-SSC). More specifically, we regard the index of each U set as a surrogate-set label and consider the supervised multi-class classification task of predicting the surrogate-set labels given observations. The difficulty is how to link our desired binary classifier with the learned surrogate multi-class classifier. To solve it, we theoretically bridge the original and surrogate class-posterior probabilities with a linear-fractional transformation, and then implement it by adding a transition layer to the neural network so that the trained model is guaranteed to be a good approximation of the original class-posterior probability. Our proposed $U^m$-SSC scheme is built within an end-to-end framework, which is computationally efficient, compatible with any model architecture and stochastic optimization, and naturally incorporates multiple U sets. Our contributions can be summarized as follows:

- Theoretically, we prove that the proposed $U^m$-SSC method is classifier-consistent (Patrini et al., 2017; Lv et al., 2020), i.e., the classifier learned by solving the surrogate set classification task from multiple sets of U data converges to the optimal classifier learned from fully supervised data under mild conditions. Then we establish an estimation error bound of our method.
- Practically, we propose an easy-to-implement, flexible, and computationally efficient method for $U^m$ classification, which is shown to outperform the state-of-the-art methods in experiments. We also verify the robustness of the proposed method by simulating $U^m$ classification in the wild, e.g., on varied set sizes, set numbers, noisy class priors, and the results are promising.

Our method provides new perspectives of solving the $U^m$ classification problem, and is more suitable to be applied in practice given its theoretical and practical advantages.

2. Problem Setup and Related Work

In this section, we introduce some notations, formulate the $U^m$ classification problem, and review the related work.

2.1. Learning from Fully Labeled Data

Let $X$ be the input feature space and $Y = \{+1, -1\}$ be a binary label space, $x \in X$ and $y \in Y$ be the input and output random variables following an underlying joint distribution $D$. Let $f : X \to \mathbb{R}$ be an arbitrary binary classifier, and $\ell_b(t, y) : \mathbb{R} \times Y \to \mathbb{R}_+$ be a loss function such that the value $\ell_b(t, y)$ means the loss by predicting $t$ when the ground-truth is $y$. The goal of binary classification is to train a classifier $f$ that minimizes the risk defined as

$$R(f) = \mathbb{E}_{(x, y) \sim D}[\ell_b(f(x), y)]$$

where $\mathbb{E}$ denotes the expectation. For evaluation, $\ell_b$ is often chosen as $\ell_{01}(t, y) = (1 - \text{sign}((t - \frac{1}{2}) \cdot y))/2$ and then the risk $R$ becomes the standard performance measure for classification, a.k.a. the classification error. For training, $\ell_{01}$ is replaced by a surrogate loss,$^1$ e.g., the logistic loss $\ell_{\text{log}}(t, y) = \ln(1 + \exp(-t \cdot y))$, since $\ell_{01}$ is discontinue and therefore difficult to optimize (Ben-David et al., 2003).

In most cases, $R$ cannot be calculated directly because the joint distribution $D$ is unknown to the learner. Given the labeled training set $X = \{(x_i, y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} D$ with $n$ samples, empirical risk minimization (ERM) (Vapnik, 1998) is a common practice that computes an approximation of $R$ by

$$\hat{R}(f) = \frac{1}{n} \sum_{i=1}^n \ell_b(f(x_i), y_i).$$

2.2. Learning from Multiple Sets of U Data

Next, we consider $U^m$ classification. We are given $m(m \geq 2)$ sets of unlabeled samples drawn from $m$ marginal densities $\{p_{\text{tr}}^j(x)\}_{j=1}^m$, where

$$p_{\text{tr}}^j(x) = \pi_j p_p(x) + (1 - \pi_j) p_n(x),$$

each $p_{\text{tr}}^j(x)$ is seen as a mixture of the positive and negative class-conditional densities $(p_p(x), p_n(x)) = (p(x|y = +1), p(x|y = -1))$, and $\pi_j = p_{\text{tr}}^j(y = +1)$ denotes the class prior of the $j$-th U set. Note that given only U data, it is theoretically impossible to learn the class priors without any assumptions (Menon et al., 2015), so we assume all necessary class priors are given, which are the only weak supervision we will leverage.$^2$ To make the problem mathematically solvable, among the $m$ sets of U data, we also assume that at least two of them are different, i.e., $\exists j, j' \in \{1, \ldots, m\}$ such that $j \neq j'$ and $\pi_j \neq \pi_{j'}$.

In contrast to supervised classification where we have a fully labeled training set $X$ directly drawn from $D$, now we only have access to $m$ sets of U data $X_{\text{tr}} = \{X_{\text{tr}}^j\}_{j=1}^m$, where

$$X_{\text{tr}}^j = \{ x_{\text{tr}}^j_1, \ldots, x_{\text{tr}}^j_{n_j} \} \overset{i.i.d.}{\sim} p_{\text{tr}}^j(x),$$

and $n_j$ denotes the sample size of the $j$-th U set. But our goal is still the same as supervised classification: to obtain a binary classifier that generalizes well with respect to $D$, despite the fact that it is unobserved.

2.3. Related Work

Here, we review some related works for $U^m$ classification.

---

$^1$The surrogate loss $\ell_b$ should be classification-calibrated so that the predictions can be the same for classifiers learned by using $\ell_\infty$ and $\ell_{01}$ (Bartlett et al., 2006).

$^2$By introducing the mutually irreducible assumption (Scott et al., 2013), the class priors become identifiable and can be estimated in some cases, see Menon et al. (2015), Liu & Tao (2016), Jain et al. (2016), and Yao et al. (2020) for details.
Clustering methods Learning from only U data is previously regarded as 
discriminative clustering (Xu et al., 2004; Gomes et al., 2010). However, these methods are often 
suboptimal since they rely on a critical assumption that one cluster exactly correspon-
ds to one class, and hence even perfect clustering may still result in poor classification. As 
a consequence, we prefer ERM to clustering.

Proportion risk methods The U\(^m\) classification setting is also related to 
learning with label proportions (LLP), with a subtle difference in the experimental design.\(^3\) However, most LLP methods are not ERM-based, but based on the 
following empirical proportion risk (EPR) (Yu et al., 2014):
\[
\hat{R}_{\text{prop}}(f) = \sum_{j=1}^{m} d_{\text{prop}}(\pi_j, \hat{\pi}_j),
\]
where \(\pi_j\) and \(\hat{\pi}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} (1 + \text{sign}(f(x^j_i) - 1/2))/2\) are the true and predicted label proportions for the j-th U 
set \(X^j\), and \(d_{\text{prop}}\) is a distance function. State-of-the-art 
method in this line combined EPR with consistency regularization and proposed the following learning 
objective:
\[
\hat{R}_{\text{prop-c}}(f) = \hat{R}_{\text{prop}}(f) + \alpha \ell_{\text{cons}}(f),
\]
where \(\ell_{\text{cons}}(f) = d_{\text{cons}}(f, \hat{f}(\hat{x}))\) is the consistency loss 
given a distance function \(d_{\text{cons}}\) and \(\hat{x}\) is a perturbed input 
from the original one \(x\) (Tsai & Lin, 2020).

Classification risk methods A breakthrough of the ERM-based 
method for U\(^m\) classification is Lu et al. (2019) which 
assumed \(m = 2\) and \(\pi_1 > \pi_2\), and proposed an equivalent 
expression of the classification risk (1):
\[
R_{U^2}(f) = \mathbb{E}_{x \sim p^U_1} c_1^+ \ell_b(f(x), +1) - \mathbb{E}_{x \sim p^U_2} c_2^+ \ell_b(f(x), -1) + R_{U^2-p}(f)
\]
\[
-\mathbb{E}_{x \sim p^U_1} c_1^- \ell_b(f(x), -1) + \mathbb{E}_{x \sim p^U_2} c_2^- \ell_b(f(x), +1),
\]
where \(c_1^+ = \frac{(1-\pi_2)\pi_D}{\pi_1-\pi_2}\), \(c_1^- = \frac{\pi_2(1-\pi_D)}{\pi_1-\pi_2}\), \(c_2^+ = \frac{(1-\pi_1)\pi_D}{\pi_1-\pi_2}\), \(c_2^- = \frac{\pi_1(1-\pi_D)}{\pi_1-\pi_2}\), and \(\pi_D\) denotes the class prior of the test 
set. If \(\pi_D\) is assumed to be 0.5 in \(R_{U^2}(f)\), the obtained 
\(R_{U^2-b}(f)\) (Menon et al., 2015) corresponds to the balanced risk, a.k.a. the balanced error (Brodersen et al., 2010):
\[
R_b(f) = \frac{1}{2} \mathbb{E}_{x \sim p_b} [\ell_b(f(x), +1)] + \frac{1}{2} \mathbb{E}_{x \sim p_b} [\ell_b(f(x), -1)],
\]
where \(\ell_b\) is \(\ell_{01}\). Note that \(R_b(f) = R(f)\) for any \(f\) if and only if \(\pi_D = 0.5\), which means that it definitely biases 
learning when \(\pi_D \approx 0.5\) is not the case. Given \(\mathcal{X}_b\) and \(\mathcal{X}_b^2\), \(R_{U^2}(f)\) and \(R_{U^2-b}(f)\) can be approximated by their 
empirical counterparts \(\hat{R}_{U^2}(f)\) and \(\hat{R}_{U^2-b}(f)\).

It is shown in Lu et al. (2020) that the empirical training risk 
\(\hat{R}_{U^2}(f)\) can take negative values which causes overfitting, 
so they proposed a corrected learning objective that wraps 
the empirical risks of the positive class \(\hat{R}_{U^2-p}(f)\) and 
the negative class \(\hat{R}_{U^2-n}(f)\) into some non-negative correction function \(f_c\), such that \(f_c(x) = x\) for all \(x \geq 0\) and \(f_c(x) > 0\) for all \(x < 0\): 
\(\hat{R}_{U^2-c}(f) = f_c(\hat{R}_{U^2-p}(f)) + f_c(\hat{R}_{U^2-n}(f)).\) 
Note that \(\hat{R}_{U^2-c}\) is biased with finite samples, but Lu et al. (2020) 
showed its risk-consistency, i.e., it converges to the 
original risk \(R\) in (1) if \(n_1, n_2 \to \infty\).

Although these risk-consistent methods are advantageous in 
terms of flexibility and theoretical guarantees, they are limited 
to 2 U sets. Recently, Scott & Zhang (2020) extended 
the previous method for the general \(m(m \geq 2)\) setting. More specifically, they assumed the number of sets \(m = 2k\) 
and proposed a pre-processing step that finds \(k\) pairs of 
the U sets by solving a maximum weighted matching problem 
(Edmonds, 1965). Then they linearly combine the unbiased 
balanced risk estimator of each pair.\(^4\) The resulted weighted 
learning objective is given by
\[
\hat{R}_{U^m}(f) = \sum_{j=1}^{k} \omega_j \hat{R}_{U^2-b}(f).
\]
This method is promising but has some practical issues:

\(^3\)The majority of LLP papers use uniform sampling for bag 
generation, which may result in the same label proportion for all 
the U sets and make the LLP problem computationally intractable (Scott & Zhang, 2020). Our simulation in Sec. 4 avoids the issue.

\(^4\)In Scott & Zhang (2020), it is assumed that \(\pi_1 \neq \pi_2\) in each 
pair, which is a stronger assumption than ours.
the pairing step is computationally very inefficient and the weights are hard to tune in practice.

A comparison of the previous works with our proposed method that will be introduced in Sec. 3 is given in Table 1.

3. U\textsuperscript{m} classification via Surrogate Set Classification

In this section, we propose a new ERM-based method for learning from multiple U sets via a surrogate set classification task and analyze it theoretically. All the proofs are given in Appendix A.

3.1. Surrogate Set Classification Task

The main challenge in the U\textsuperscript{m} classification problem is that we have no access to the ground-truth labels of the training examples so that the empirical risk (2) in supervised binary classification cannot be computed directly. Our idea is to consider a surrogate set classification task that could be tackled easily from the given U sets. It serves as a proxy and gives us a classifier-consistent solution to the original binary classification problem.

Specifically, denote by \( \tilde{y} \in \{1, 2, \ldots, m\} \) the index of the U set, i.e., the index of the corresponding marginal density. By treating \( \tilde{y} \) as a surrogate label, we formulate the surrogate set classification task as the standard multi-class classification. Let \( \mathcal{D} \) be the joint distribution for the random variables \( x \in \mathcal{X} \) and \( \tilde{y} \in \tilde{Y} = \{1, 2, \ldots, m\} \). Any \( \mathcal{D} \) can be identified via the class priors \( \{p(j = j)\}_{j=1}^{m} \) and the class-conditional densities \( \{p(x \mid \tilde{y} = j) = p_{j}^{\tilde{y}}(x)\}_{j=1}^{m} \), where \( \rho_{j} \) can be estimated by \( \rho_{j} = \frac{1}{\sum_{j=1}^{n} nx_{j}} \).

The goal of surrogate set classification is to train a classifier \( g(x) : \mathcal{X} \to \mathbb{R}^{m} \) that minimizes the following risk:

\[
R_{\text{surr}}(g) = \mathbb{E}_{(x, \tilde{y}) \sim \mathcal{D}}[\ell(g(x), \tilde{y})],
\]

where \( \ell(g(x), \tilde{y}) : \mathbb{R}^{m} \times \tilde{Y} \to \mathbb{R}_{+} \) is a proper loss for \( m \)-class classification, e.g., the cross-entropy loss:

\[
\ell_{ce}(g(x), \tilde{y}) = -\sum_{j=1}^{m} 1(\tilde{y} = j) \log(g_{j}(x)) = -\log(g_{\tilde{y}}(x)),
\]

where \( 1(\cdot) \) is the indicator function, \( g_{j}(x) \) is the \( j \)-th element of \( g(x) \), and is a score function that estimates the true class-posterior probability \( \hat{\pi}_{j}(x) = p(\tilde{y} = j \mid x) \). Typically, the predicted label \( \hat{y}_{\text{pred}} \) takes the form \( \hat{y}_{\text{pred}} = \arg\max_{\tilde{y} \in [m]} g_{\tilde{y}}(x) \).

Now the unlabeled training sets given by (4) for the binary classification can be seen as a labeled training set \( \mathcal{X}_{tr} = \{(x_{i}, y_{i})\}_{i=1}^{n_{tr}} \overset{i.i.d.}{\sim} \mathcal{D} \) for the \( m \)-class classification, where \( n_{tr} = \sum_{j=1}^{m} n_{j} \) is the total number of U data. We can use \( \mathcal{X}_{tr} \) to approximate the risk \( R_{\text{surr}} \) by

\[
\hat{R}_{\text{surr}}(g) = \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \ell(g(x_{i}), \tilde{y}_{i}).
\]

3.2. Bridge Two Posterior Probabilities

Let \( \eta(x) = p(y = +1 \mid x) \) be the class-posterior probability for class +1 in the original binary classification problem, and \( \tilde{\eta}_{j}(x) = p(\tilde{y} = j \mid x) \) be the class-posterior probability for class \( j \) in the surrogate set classification problem. We theoretically bridge them by the following theorem.

Theorem 1. By the definitions of \( \mathcal{D} \), \( \eta(x) \), \( \tilde{\mathcal{D}} \), and \( \tilde{\eta}_{j}(x) \), we have

\[
\tilde{\eta}_{j}(x) = T_{j}(\eta(x)), \quad \forall j = 1, \ldots, m,
\]

where

\[
T_{j}(\eta(x)) = \frac{a_{j} \cdot \eta(x) + b_{j}}{c \cdot \eta(x) + d},
\]

\( a_{j} = \rho_{j}(\pi_{j} - \pi_{D}), \quad b_{j} = \rho_{j}\pi_{D}(1 - \pi_{j}), \quad c = \sum_{j=1}^{m} \rho_{j}(\pi_{j} - \pi_{D}), \quad \text{and} \quad d = \sum_{j=1}^{m} \rho_{j}\pi_{D}(1 - \pi_{j}) \).

Such a relationship has been previously studied by Menon et al. (2015) in the context of corrupted label learning for a specific \( 2 \times 2 \) case, i.e., 2 clean classes are transformed to 2 corrupted classes, and they used \( T_{j}(\cdot) \) to post-process the threshold of the score function learned from corrupted data. Our proposal can be regarded as its extension to a general \( 2 \times m \) case and \( T_{j}(\cdot) \) is used to connect the original binary classifier with the surrogate multi-set-class classifier.

Let \( T(\cdot) : \mathbb{R} \to \mathbb{R}^{m} \) be a vector form of the transition function \( T(\cdot) = [T_{1}(\cdot), \ldots, T_{m}(\cdot)]^{\top} \). Note that the coefficients in \( T_{j}(\cdot) \) are all constants and \( T(\cdot) \) is deterministic. Next, we study properties of the transition function \( T(\cdot) \) in the following lemma, which implies the feasibility of approximating \( \eta(x) \) by means of estimating \( \tilde{\eta}_{j}(x) \).

Lemma 2. The transition function \( T(\cdot) \) is an injective function in the domain \([0, 1]\).

3.3. Classifier-consistent Algorithm

Given the transition function \( T(\cdot) \), we have two choices to obtain \( \eta(x) \) from \( \tilde{\eta}_{j}(x) \). First, one can estimate \( \tilde{\eta}_{j}(x) \), then calculate \( \eta(x) \) via the inverse function \( T_{j}^{-1}(\tilde{\eta}_{j}(x)) \). Second, one can encode \( \eta(x) \) as a latent variable into the computation of \( \tilde{\eta}_{j}(x) \) and obtain both of them simultaneously. We prefer the latter for three reasons.

- Computational efficiency: the latter is a one-step solution and avoids additional computations of the inverse functions, which provides computational efficiency and easiness for implementation.
- Robustness: since the coefficients of \( T_{j}(\cdot) \) may be perturbed by some noise in practice, its inversion in
the former method may enlarge the noise by orders of magnitude, making the learning process less robust.

- Identifiability: calculating $T_j^{-1}(\tilde{y}_j(x))$ in the former method for all $j = 1, \ldots, m$ induces $m$ estimates of $\eta(x)$, and they are usually non-identical due to the estimation error of $\tilde{y}_j(x)$ from finite samples or noisy $T_j(\cdot)$, causing a new non-identifiable problem.

Therefore, we choose to embed the estimation of $\eta(x)$ into the estimation of $\tilde{y}_j(x)$. More specifically, let $f(x)$ be the model output that estimates $\eta(x)$, then we make use of the transition function $T_j(\cdot)$ and model $g_j(x) = T_j(f(x))$.

Based on it, we propose to learn with the following modified loss function:

$$\ell(g(x), \bar{y}) = \ell(T(f(x)), \bar{y}), \quad (12)$$

where $T(f(x)) = [T_1(f(x)), \ldots, T_m(f(x))]^\top$. Then the corresponding risk for the surrogate task can be written as

$$R_{\text{surr}}(f) = E_{(x, \bar{y})\sim \mathcal{D}}[\ell(T(f(x)), \bar{y})] = E_{(x, \bar{y})\sim \mathcal{D}}[\ell(g(x), \bar{y})] = R_{\text{surr}}(g), \quad (13)$$

and an equivalent expression of the empirical risk (10) is given by

$$\bar{R}_{\text{surr}}(f) = \frac{1}{n_{\text{tr}}} \sum_{i=1}^{n_{\text{tr}}} \ell(T(f(x_i)), \bar{y}_i). \quad (14)$$

In order to prove that this method is classifier-consistent, we introduce the following lemma.

**Lemma 3.** Let $\tilde{\eta}(x) = [\tilde{\eta}_1(x), \ldots, \tilde{\eta}_m(x)]^\top$ and $g^*(x) = \arg\min_{g} R_{\text{surr}}(g; \ell)$ be the optimal classifier of (9). Provided that a proper loss function, e.g., the cross-entropy loss or mean squared error, is chosen for $\ell$, we have $g^*(x) = \tilde{\eta}(x)$.

Since $g(x) = T(f(x))$ and $T(\cdot)$ is deterministic, when considering minimizing $R_{\text{surr}}(f)$ that takes $f$ as the argument, we can prove the following classifier-consistency.

**Theorem 4 (Identification of the optimal binary classifier).** Assume that the cross-entropy loss or mean squared error is used for $\ell$ and $\ell_{\text{tr}}$, and the model $\mathcal{G}$ used for learning $g$ is very flexible, e.g., deep neural networks, so that $g^* \in \mathcal{G}$. Let $f^*_{\text{surr}}$ be the U$^m$-SSC optimal classifier induced by $g^*$, and $f^* = \arg\min_{f} R(f; \ell_{\text{tr}})$ be the optimal classifier of (1). We have $f^*_{\text{surr}} = f^*$.

So far, we have proved that the optimal classifier for the original binary classification task can be identified by the U$^m$-SSC learning scheme. Its algorithm is described in Algorithm 1 and its implementation is illustrated in Figure 1.

We implement $T(\cdot)$ by adding a transition layer following the sigmoid function of the neural network (NN). At the training phase, a sample $(x_{\text{tr}}, y_{\text{tr}})$ is fetched to the network. A sigmoid function $f_{\text{sig}}(x) = \frac{1}{1 + e^{-x}}$ is used to map the output of NN to the range $[0,1]$ such that the output $f(x)$ is an estimate of $\eta(x)$. Then $f(x)$ is forwarded to the transition layer and a vector output $g(x) = T(f(x))$ is obtained. The loss computed on the output $g(x)$ and the surrogate label $\bar{y}_{\text{tr}}$ by (10) is then used for updating the NN weights $w$. Note that the transition layer is fixed and only the weights in the base network are learnable. At the test phase, for any test sample $x_{\text{te}}$, we compute $f(x_{\text{te}})$ using only the trained base network and sigmoid function. The test sample is classified by using the sign function, i.e., $\text{sign}(f(x_{\text{te}}) - \frac{1}{2})$. Our proposed method is model-agnostic and can be easily trained with a stochastic optimization algorithm, which ensures its scalability to large-scale datasets.

### 3.4. Theoretical Analysis

In what follows, we upper-bound the estimation error of our proposed method. Let $f^*_{\text{surr}} = \arg\min_{f \in \mathcal{F}} R_{\text{surr}}(f)$ be our empirical classifier, where $\mathcal{F} = \{f : \mathcal{X} \rightarrow \mathbb{R}\}$ is a class of measurable functions, and $f^*_{\text{surr}} = \arg\min_{f \in \mathcal{F}} R_{\text{surr}}(f)$ be the optimal classifier, the estimation error is defined as the gap between the risk of $f^*_{\text{surr}}$ and that of $f^*_{\text{surr}}$, i.e., $R_{\text{surr}}(f^*_{\text{surr}}) - R_{\text{surr}}(f^*_{\text{surr}})$. To derive the estimation error bound, we firstly investigate the Lipschitz continuity of the transition function $T(f(x))$.
Table 2. Specification of datasets and corresponding models.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Train</th>
<th># Test</th>
<th># Features</th>
<th>$\pi_D$</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST (LeCun et al., 1998)</td>
<td>60,000</td>
<td>10,000</td>
<td>784</td>
<td>0.49</td>
<td>5-layer MLP</td>
</tr>
<tr>
<td>Fashion-MNIST (Xiao et al., 2017)</td>
<td>60,000</td>
<td>10,000</td>
<td>784</td>
<td>0.8</td>
<td>5-layer MLP</td>
</tr>
<tr>
<td>Kuzushiji-MNIST (Cluanwat et al., 2018)</td>
<td>60,000</td>
<td>10,000</td>
<td>784</td>
<td>0.3</td>
<td>5-layer MLP</td>
</tr>
<tr>
<td>CIFAR-10 (Krizhevsky, 2009)</td>
<td>50,000</td>
<td>10,000</td>
<td>3,072</td>
<td>0.7</td>
<td>ResNet-32</td>
</tr>
</tbody>
</table>

**Lemma 5.** Assume that among the $m$ sets of $U$ data, at least two of them are different, i.e., $\exists j, j' \in \{1, \ldots, m\}$ such that $j \neq j'$ and $\pi_j \neq \pi_{j'}$, and $0 \leq f(x) \leq 1, \forall x \in \mathcal{X}$, e.g., $f(x)$ is mapped to $[0, 1]$ by the sigmoid function. Then, for any $\delta > 0$, we have with probability at least $1 - \delta$,

$$
\alpha = \min \left( \sum_{j=1}^{m} \rho_j \pi_j (1 - \pi_D), \sum_{j=1}^{m} \rho_j \pi_D (1 - \pi_j) \right).
$$

Then we analyze the estimation error as follows.

**Theorem 6 (Estimation error bound).** Assume that the loss $\ell(T(f), y)$ is upper-bounded by $M_L$ and $\ell$ is $L_\ell$-Lipschitz continuous w.r.t. $f(x)$ with a Lipschitz constant $2/\alpha^2$, where

$$
R_{\text{surr}}(\tilde{f}_{\text{surr}}) - R_{\text{surr}}(f_{\text{surr}}^*) \leq \frac{8\sqrt{2mL_\ell}}{\alpha^2} R_{\mathbb{R}_{\mathcal{U}^m}}(\mathcal{F}) + 2M_L \sqrt{\frac{\ln(2/\delta)}{2m_L}}.
$$

Theorem 6 demonstrates that as the number of training samples goes to infinity, the risk of $f_{\text{surr}}^*$ converges to the risk of $f_{\text{surr}}$ since $R_{\mathbb{R}_{\mathcal{U}^m}}(\mathcal{F}) \to 0$ for all parametric models with a bounded norm. Moreover, the coefficient $\alpha$ implies that a tighter error bound could be obtained when the class priors $\pi_j$ are close to 0 or 1. This conclusion agrees with our intuition that purer $U$ sets (containing almost only positive/negative examples) lead to better performance.

## 4. Experiments

In this section, we experimentally analyze the proposed method and compare it with state-of-the-art methods in the $U^m$ classification setting.\(^5\)

**Datasets** We train on widely adopted benchmarks MNIST, Fashion-MNIST, Kuzushiji-MNIST, and CIFAR-10. Table 2 briefly summarizes the benchmark datasets. Since the four datasets contain 10 classes originally, we manually corrupt them into binary classification datasets. More details about the datasets are in Appendix B.1.

In the experiments, unless otherwise specified, the number of training data contained in all $U$ sets are the same and fixed as $n_j = n_{tr}/m$ for all benchmark datasets; and the class priors $\{\pi_j\}_{j=1}^m$ of all $U$ sets are randomly sampled from the range $[0, 1, 0.9]$ under the constraint that the sampled class priors are not all identical, ensuring that the problem is mathematically solvable. Given $\{n_j\}_{j=1}^m$ and $\{\pi_j\}_{j=1}^m$, we generate $m$ sets of $U$ training data following (4). Note that in most LLP papers, each $U$ set is uniformly sampled from the shuffled $U$ training data, therefore the label proportions of all the $U$ sets are the same in expectation. As the set size increases, all the proportions converge to the same class prior, making the LLP problem computationally intractable (Scott & Zhang, 2020). As shown above, our experimental scheme avoids this issue by determining valid class priors before sampling each $U$ set.

**Models** The models are optimizers used are also described in Table 2, where MLP refers to multi-layer perceptron, ResNet refers to residual networks (He et al., 2016), and their detailed architectures are in Appendix B.2. As a common practice, we use Adam (Kingma & Ba, 2015) with the cross-entropy loss for optimization. We train 300 epochs for all the experiments, and the classification error rates at the test phase are reported. All the experiments are repeated 3 times and the mean values with standard deviations are recorded for each method.

**Baselines** We compare the proposed method with state-of-the-art methods based on the classification risk (Scott & Zhang, 2020) and the empirical proportion risk (Tsai & Lin, 2020) for the $U^m$ classification problem. Recall that the proposed learning objective $\hat{R}_{U^m}(f)$ in Scott & Zhang (2020) is a combination of the unbiased balanced risk estimators $\hat{R}_{U^m:b}(f)$, which are shown to underperform the unbiased risk estimator $\hat{R}_{U^m}(f)$ in Lu et al. (2019). So we improve the baseline method of Scott & Zhang (2020) by combining $\hat{R}_{U^m}(f)$ instead of $\hat{R}_{U^m:b}(f)$. As shown in Lu et al. (2020), the empirical risks $\hat{R}_{U^m}(f)$ can go negative during training which may cause overfitting, so we further improve the baseline by combining the corrected non-negative risk estimators $\hat{R}_{U^m+}(f)$. The baselines are summarized as follows:

\(^5\)Our implementation of $U^m$-SSC is available at https://github.com/leishida/Um-Classification.
Figure 2. Experimental results of learning from 10 and 50 sets of U data. Solid curves are the test errors (in percentage) and dashed curves are the empirical training risks. Dark colors show the mean errors (risks) of 3 trials and light colors show the standard deviations.
Table 3. Mean errors (standard deviations) over 3 trials in percentage for the proposed U\textsuperscript{m} -SSC method tested on different set sizes. The uniform set size \( n_j \) is shifted to \( \tau \cdot n_j \) (smaller \( \tau \), larger shift). Random means uniformly sample a set size from range \([0, T]\).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sets</th>
<th>( n_j )</th>
<th>( \tau = 0.8 )</th>
<th>( \tau = 0.6 )</th>
<th>( \tau = 0.4 )</th>
<th>( \tau = 0.2 )</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>10</td>
<td>6000</td>
<td>2.83 (0.18)</td>
<td>2.91 (0.04)</td>
<td>3.2 (0.2)</td>
<td>3.19 (0.35)</td>
<td>2.66 (0.08)</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>1200</td>
<td>2.46 (0.1)</td>
<td>2.58 (0.08)</td>
<td>2.76 (0.13)</td>
<td>2.97 (0.11)</td>
<td>3.0 (0.12)</td>
</tr>
<tr>
<td>Fashion-MNIST</td>
<td>10</td>
<td>6000</td>
<td>7.88 (0.21)</td>
<td>7.68 (0.36)</td>
<td>7.84 (0.17)</td>
<td>7.89 (0.24)</td>
<td>7.04 (0.13)</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>1200</td>
<td>8.29 (0.19)</td>
<td>8.91 (0.29)</td>
<td>7.61 (0.55)</td>
<td>8.8 (0.31)</td>
<td>8.62 (0.24)</td>
</tr>
<tr>
<td>Kuzushiji-MNIST</td>
<td>10</td>
<td>6000</td>
<td>8.98 (0.52)</td>
<td>9.83 (0.57)</td>
<td>9.43 (0.41)</td>
<td>10.03 (0.81)</td>
<td>8.38 (0.31)</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>1200</td>
<td>9.35 (0.33)</td>
<td>9.53 (0.64)</td>
<td>9.89 (0.72)</td>
<td>11.08 (0.61)</td>
<td>10.34 (0.73)</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>10</td>
<td>5000</td>
<td>12.55 (0.61)</td>
<td>12.25 (0.8)</td>
<td>12.41 (0.35)</td>
<td>12.49 (0.98)</td>
<td>11.65 (0.44)</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>1000</td>
<td>12.16 (0.23)</td>
<td>12.19 (0.75)</td>
<td>12.88 (0.37)</td>
<td>13.66 (0.54)</td>
<td>12.09 (0.42)</td>
</tr>
</tbody>
</table>

Figure 3. Box plot of the classification errors for the proposed U\textsuperscript{m} -SSC method tested on learning from 50 U sets with inaccurate class priors (\( \epsilon = 0 \) means true; larger \( \epsilon \), larger noise).

- MMC-U\textsuperscript{2-b} \cite{Scott2020}: the classification risk based method in the multiple mutual contamination (MMC) framework, i.e., (8);
- MMC-U\textsuperscript{2}: the method that improves MMC-U\textsuperscript{2-b} with unbiased risk estimators \( \hat{R}_{U^2}(f) \);
- MMC-U\textsuperscript{2-c}: the method that improves MMC-U\textsuperscript{2} with non-negative risk correction \( \hat{R}_{U^2,c}(f) \);
- LLP-VAT \cite{Tsai2020}: the empirical proportion risk based method, i.e., (6).

More details about the implementation of baselines can be found in Appendix B.3.

4.1. Comparison with State-of-the-art Methods

We first compare our proposed method with state-of-the-art methods for the U\textsuperscript{m} classification problem. The experimental results of learning from 10 and 50 U sets are reported in Figure 2 and a table of the final errors is in Appendix C.1.

We can see that the classification risk based methods, i.e., MMC-U\textsuperscript{2-b}, MMC-U\textsuperscript{2}, MMC-U\textsuperscript{2-c}, and our proposed U\textsuperscript{m}-SSC method generally outperform the empirical proportion risk based method, i.e., LLP-VAT, with lower classification error and more stability, which demonstrates the superiority of the consistent methods.

Within the classification risk based methods, our observations are as follows. First, the proposed U\textsuperscript{m} -SSC method outperforms others in most cases. We believe that the advantage comes from the surrogate set classification mechanism in U\textsuperscript{m}-SSC, which implies the classifier-consistent methods perform better than the risk-consistent methods. Second, compared to MMC-U\textsuperscript{2-b}, we can see our advantage becomes bigger when \( \pi_D \approx \frac{1}{2} \) is not the case, e.g., Fashion-MNIST and CIFAR-10. Moreover, the performance of the improved MMC-U\textsuperscript{2} method (combination of unbiased risk estimators) is better than MMC-U\textsuperscript{2-b} (combination of balanced risk estimators) in all cases. These empirical findings corroborate our analysis that the balanced classification risk (7) can be biased in such cases. Third, we confirm that the training risks of MMC-U\textsuperscript{2-b} and MMC-U\textsuperscript{2} go negative as training proceeds, which incurs overfitting. Other methods do not have this negative empirical training risk issue. And we can see that the improved MMC-U\textsuperscript{2-c} method effectively mitigates this overfitting but its performance is still inferior to our proposed method. These results are consistent with the observations in Lu et al. (2020). We also notice that the empirical training risks of the proposed U\textsuperscript{m}-SSC method are obviously higher than other baseline methods. This is due to the fact that the added transition layer rescales the range of model output. We provide a detailed analysis on this point in Appendix C.1. A notable effect is that a relatively small learning rate is more suitable for our method.

4.2. On the Variation of Set Size

In practice, the size of the U sets may vary from a large range depends on different tasks. However, as the set size varies, given the data generation process in (3), the marginal density of our training data \( p_{tr}(x) \) shifts from that of the test one,
which may cause severe covariate shift (Shimodaira, 2000; Zhang et al., 2020). To verify the robustness of our proposed method against covariate shift, we conducted experiments on the variation of set size. Recall that in other experiments, we use uniform set size, i.e., all sets contain \( n_{tr}/m \) U data. In this subsection, we investigate two kinds of set size shift:

- Randomly select \( \lceil m/2 \rceil \) U sets and change their set sizes to \( \tau \cdot n_{tr}/m \) where \( \tau \in [0, 1] \);
- Randomly sample each set size \( n_j \) from range \( [0, n_{tr}] \) such that \( \sum_{j=1}^{m} n_j = n_{tr} \).

As shown in Table 3, the proposed method is reasonably robust as \( \tau \) moves towards 0 in the first shift setting. The slight performance degradation may come from the decreased total number of training samples \( n_{tr} \) as \( \tau \) decreases. We also find that our method reaches the best performance in 3 out of 4 benchmark datasets in the second shift setting. Since it is a more natural way for generating set sizes, the robustness of the proposed method on varied set sizes can be verified.

### 4.3. On the Variation of Set Numbers

Another main factor that may affect the performance is the number of available U sets. As the U data can be easily collected from multiple sources, the learning algorithm is expected to be able to handle the variation of set numbers well. The experimental results of learning from 10 and 50 U sets have been shown in Section 4.1. In this subsection, we test the proposed U\(^m\)-SSC method on extremely small set numbers e.g., \( m = 2 \), and large set numbers, e.g., \( m=1000 \). The experimental results of learning form 2, 100, 500, and 1000 U sets are reported in Table 4.

From the results, we can see that the performance of the proposed method is reasonably well on different set numbers. In particular, a lower classification error can be observed for \( m = 2 \) across all 4 benchmark datasets. The better performance may come from the larger number of U data contained in a single set, i.e., \( n_{tr}/2 \) in this case. Since our method uses class priors as the only weak supervision, an increasing number of the sampled data within each U set guarantees a better approximation of them. These experimental results demonstrate the effectiveness of the proposed method on the variation of set numbers. We note that it is a clear advantage over the LLP methods, whose performance drops significantly when the set number becomes small and set size becomes large, because label proportions converge to the same class prior in their setup, making the LLP problem computationally intractable.

### 4.4. Robustness against Inaccurate Class Priors

Hitherto, we have assumed that the values of class priors \( \{ \pi_j \}_{j=1}^m \) are accessible and accurately used in the construction of our method, which may not be true in practice. In order to simulate U\(^m\) classification in the wild, where we may suffer some errors from estimating the class priors, we design experiments that add noise to the true class priors. More specifically, we test the U\(^m\)-SSC method by replacing \( \pi_j \) with the noisy \( \pi'_j = \pi_j + \gamma \cdot \epsilon \), where \( \gamma \) uniform randomly take values in \( \{+1, -1\} \) and \( \epsilon \in \{0, 0.05, 0.1, 0.15, 0.2\} \), so that the method would treat noisy \( \pi'_j \) as the true \( \pi_j \) during the whole learning process. The experimental setup is exactly same as before except the replacement of \( \pi_j \). Note that we tailor the noisy \( \pi_j \) to \([0, 1] \) if it surpasses the range.

The results on learning from 50 U sets with inaccurate class priors are reported in Figure 3 and a table of the final test errors is in Appendix C.2, where \( \epsilon = 0 \) means true class priors. We can see that our method works reasonably well using noisy \( \pi'_j \) though the classification error slightly increases for higher noise level \( \epsilon \) which is as expected.

### 5. Conclusions

In this work, we focused on learning from multiple sets of U data and proposed a new method based on a surrogate set classification task. We bridged the original and surrogate class-posterior probabilities via a linear-fractional transformation, and then studied its properties. Based on them, we proposed the U\(^m\)-SSC algorithm and implemented it by adding a transition layer to the neural network. We also proved that the U\(^m\)-SSC method is classifier-consistent and established an estimation error bound for it. Extensive experiments demonstrated that the proposed method could successfully train binary classifiers from multiple U sets, and it compared favorably with state-of-the-art methods.

### Acknowledgments

The authors would like to thank Tianyi Zhang, Wenkai Xu, Nontawat Charoenphakdee, and Yuting Tang for helpful discussions. NL was supported by the MEXT scholarship No. 171536 and MSRA D-CORE Program. GN and MS were supported by JST AIP Acceleration Research Grant Number JPMJCR20U3, Japan. MS was also supported by the Institute for AI and Beyond, UTokyo.
References


