Learning One-Class Hyperspectral Classifier from Positive and Unlabeled Data for Low Proportion Target

Hengwei Zhao, {Student Member, IEEE}, Yanfei Zhong, {Senior Member, IEEE}, Xin He, Xinyu Wang, {Member, IEEE}, Hong Shu

Abstract—Hyperspectral imagery (HSI) one-class classification is aimed at identifying a single target class from the HSI by using only positive labels, which can significantly reduce the requirements for annotation. However, HSI one-class classification is far more challenging than HSI multi-class classification, due to the lack of negative labels and the low target proportion, which are issues that have rarely been considered in the previous HSI classification studies. In this paper, a weakly supervised HSI one-class classifier, namely HOneCls, is proposed to solve the problem of under-fitting of the positive class occurs in the HSI data with low target proportion, where a risk estimator—the One-Class Risk Estimator—is particularly introduced to make the full convolutional neural network (FCN) with the ability of one class classification. The experimental results obtained on challenging hyperspectral classification datasets, which includes 20 kinds of ground objects with very similar spectra, demonstrate the efficiency and feasibility of the proposed One-Class Risk Estimator. Compared with the state-of-the-art one-class classifiers, the F1-score is improved significantly in the HSI data with low target proportion.

Index Terms—Hyperspectral imagery, one-class classification, low proportion target.

I. INTRODUCTION

ImagIng spectroscopy have the ability to capture hundreds of continuous wavebands from the visible spectrum to the far-infrared [1], which means HSI has ground object-rich properties, so that many kinds of objects can be identified in hyperspectral imagery (HSI) which are difficult to identify in traditional RGB remote sensing imagery. As one of the key techniques in HSI processing, HSI classification is aimed at assigning a unique label to each HSI pixel [2], and has been widely applied in agriculture [3], forestry [4], geology [5], and coastal/ocean studies [6].

In real-world application scenarios, we often meet the problem of only a single target class needing to be identified from the HSI, i.e., one-class classification [7]. However, most of the current HSI classification methods focus on the multi-class classification task, where all the classes that exist in the HSI need to be annotated, which is labor-intensive if a multi-class classification method is used for the one-class classification problem. For example, it would be preferable to only annotate the invasive tree species rather than all the species in the task of invasive tree species detection in tropical mountain areas, because labeling all the species is labor-intensive and even impractical, due to the high species richness [4]. In order to solve the above problem, this paper focuses on the HSI one-class classification problem. Compared with HSI multi-class classification, the major advantages of one-class classification can be summarized as follows:

- No predefined class system: It is difficult to establish a complete class system in the real world, and a feasible way is to define “all classes except the class of interest as negative” [8].
- Only positive annotation: One-class classification can significantly reduce the requirement for annotation for model training because there is no need to annotate the negative classes.

One-class classification is a challenging weakly supervised problem, due to the lack of negative data. According to the difference of the input data, there are two kinds of one-class classifiers that have been proposed to solve the problem of negative data deletion for remote sensing imagery: positive (P) classifiers and positive and unlabeled (PU) classifiers [7].

P classifiers, which only use positive data during the training stage, are designed to “describe” positive data by information theory [9], density estimation [10], or geometry [11]–[13]. However, it is difficult to find the decision boundary in the case of only positive data, which is the critical limitation of the P classifiers if the target of the task is “recognition” or “classification” [14]. In addition, some empirical hyperparameters are needed to balance the degree of overfit and underfit for the positive data [15].

Recent studies have demonstrated that better one-class classification results can be obtained by using extra unlabeled data [7], which is selected from the imagery randomly and contains both positive and negative data. The two-step PU classification strategy is used in the heuristic methods [16], which first obtain reliable negative samples from the unlabeled data, and then train a binary classifier by these positive and selected negative samples. However, the reliability of the
selected negative samples seriously affects the final results. Therefore, the positive data and all the unlabeled data are used simultaneously to train the model in a one-step strategy, and a biased classifier with threshold selection [4] or post-threshold calibration [15], [17], [18] is used to alleviate the negative effects of the positive samples in the unlabeled data. However, the threshold adjustment process means that the classifiers cannot be trained in an end-to-end manner, which leads to inaccurate decision boundaries and cumulative errors. Benefiting from the fact that a one-class classifier can be trained in an end-to-end manner, unbiased risk estimation based methods have attracted much attention in recent works [19], and are aimed at making the risk of the PU classifiers equal to that of the Bayesian binary classifiers. However, most of the above methods are based on RGB or multispectral remote sensing imagery, and the characteristics of HSI need to be considered.

An illustration in Fig. [1] two properties of HSI are demonstrated: inter-class similarity with intra-class variation, and low target proportion. As shown in Fig. [1b] HSI are characterized by both inter-class similarity and intra-class variation. Fig. [1b] shows that the spectral curves of several class in HongHu dataset are very similar, and taking the rape class for example, a high degree of variability in the spectral curves can be seen in the Fig. [1b]. Another characteristic of HSI is the abundance of ground objects, which will lead to low target proportion, so that the probability of the positive class, i.e., \( P(y = +1) \), being much smaller than the probability of the negative class, i.e., \( P(y = -1) \). Take the HongHu dataset (Fig. [1a]) as an example, there are more than 15 types of ground objects, and as the number of ground object categories increases, the percentage of single-class objects will be a smaller value. The classes of objects and their class probability are shown in Fig. [1c](1), the percentage of more than 11 kinds of objects is less than 0.05.

Further insights into the learning process of deep neural networks learning binary classifiers from low target proportion HSI PU data with inter-class similarity and intra-class variation are provided in this paper by studying the dynamic changes of the estimated risks (Fig. [1c]). During the training stage, the estimated risk of the positive data (the expectation of positive sample loss) decreases as the overall data risk decreases in the normal target proportion dataset, and the F1-score is 99.19, as shown in the Fig [1c](2). However, the risk of the positive data shows an opposite trend to that of the overall data when low target proportion occurs, and the F1-score is 0.00, as shown in the Fig. [1c](3). According to the above observation, the problem of low target proportion leads to under-fitting of the positive training data, because the risk of the positive data increases until it reaches the maximum risk.

One-Class Risk Estimator is proposed in this paper for HSI one-class classification, which can identify a single target class with the existing spectral-spatial feature extractor by using only the positive labels in the case of HSI data with the properties of inter-class similarity with intra-class variation and low target proportion. In summary, the main contributions of this paper are summarized as follows:

- A novel insight into low target proportion HSI is provided in the risk estimation based one-class classification, and we shown that under-fitting of the positive class is the...
key bottleneck for learning with low target proportion HSI PU data.

- A weakly supervised classifier—HOneCls is proposed to solve the problem of HSI one-class classification, where a risk estimator—the One-Class Risk Estimator—is particularly introduced to make the full convolutional neural network (FCN) with the ability of one class classification.

- Extensive experiments were conducted to demonstrate the superiority of the proposed risk estimator with the global spectral-spatial feature extractor—FreeOCNet—in both normal and low target proportion HSI datasets.

II. RELATED WORKS

A. Deep Learning Based Hyperspectral Image Classification

The goal of HSI classification is to assign a semantic label to each HSI pixel. Deep neural networks have shown remarkable performances in the HSI classification task. According to the different learning modes, deep learning based HSI classification methods can be divided into local spatial information learning frameworks and global spectral-spatial feature learning frameworks [20].

Most of the related research has been based on local spatial information learning frameworks with the learning target \( f_{local} : R^S \times S \to R \). The local spatial information learning frameworks first generate \( S \times S \) HSI patches, \( S \) is the fixed spatial size, and then the deep neural networks trained by these patches. These neural networks include convolutional neural networks (CNNs), generative adversarial networks, stacked autoencoders, recurrent neural networks and so on [21]–[25], which aim to model the mapping of \( f_{local} \).

Recently, some global spectral-spatial feature learning frameworks have been proposed for HSI classification, which consider the HSI classification task as a kind of semantic segmentation task, and aim to model the mapping function \( f_{global} : R^H \times W \to R^H \times W \) with a fully convolutional neural network (FCN). A pretrained FCN is utilized in the DMS3FE classifier [20] to extract spatial-spectral features. A two-branch FCN uses a dense conditional random field (CRF) model to balance the local and global information, and a mask matrix is proposed to assist model training in the case of sparse labels for HSI classification [27]. As a unified patch-free HSI classification framework, the fast patch-free global learning (FPGA) framework [20] utilizes an encoder-decoder based CNN to capture the global spatial information in the HSI. In addition, a global stochastic stratified sampling strategy is utilized to guarantee the convergence of the network, which means that the FCN can be optimized in an end-to-end manner.

Differing from the above supervised HSI classification methods, the classifier proposed in this paper focuses on the weakly supervised one-class classification task. Furthermore, the proposed classifier can be trained in the absence of negative data.

B. Learning From PU Data

The target of learning from PU data is to learn a binary classifier from the PU data. The unlabeled data are sampled from the input space, and include both positive and negative data. The formulation can described as follows. The variables of the input space and output space are \( X \) and \( Y \in \{+1, -1\} \), respectively. The marginal distribution of the positive and negative classes is recorded as \( P_p(x) = P(x|Y = +1) \) and \( P_n(x) = P(x|Y = -1) \), respectively, and the marginal distribution of the unlabeled data is recorded as \( P(x) \). We let \( \pi_p = P(Y = +1) \) be the class prior probability, which is assumed to be known in most cases, and can be estimated from the PU data [23], [29]. The objective is to learn a binary classifier \( f \) from \( P_p(x) \) and \( P(x) \).

The two-step strategy is used in the heuristic methods, which first obtain reliable negative samples from the unlabeled data, and then train a binary classifier by these positive and selected negative samples, such as S-EM [30]. However, the reliability of the selected negative samples seriously affects the final results.

In the one-step strategy, the positive data and all the unlabeled data are used simultaneously to train the model. Learning from PU data can be converted to a cost-sensitive problem [31], [32], a label disambiguation problem [22], or a variational problem [33]. Some of the most theoretically and practically effective methods are the unbiased risk estimation based methods [34], which aim to make the risk of the classifiers learned from the PU data equal to that of a Bayesian binary classifier. The risk of a Bayesian binary classifier can be formulated as:

\[
R_{\text{pu}}(f) = \mathbb{E}_{(X, Y) \sim P(x, y)}[l(f(X), Y)],
\]

where \( f \) is the classifier and the \( l \) is an arbitrary loss function. Many methods have been proposed based on unbiased risk estimation [35]–[39], and these methods represent the state of the art for the PU classification task.

Most risk estimation based methods obtain excellent results on the normal target proportion PU problem; however, low target proportion PU problems are widespread in the real world, but are rarely considered. For example, in addition to the HSI remote sensing community, the probability of fraud is low in financial fraud detection, and there are fewer patient samples than non-patient samples in disease detection.

III. HOneCls: One-Class Hyperspectral Imagery Classifier

In this section, the proposed one-class classifier—HOneCls—is introduced (Fig. 2), which includes the the module of one-class representation learning and the module of global spectral-spatial features extractor. One-Class Risk Estimator is responsible for learning a binary classifier from the PU data. Global spectral-spatial features can be extracted by any FCN, and a plain FCN—FreeOCNet is responsible for extracting global spectral-spatial features in HOneCls.

A. One-Class Risk Estimation

Learning a binary classifier from HSI PU data is met with these problems: the risk of the negative class \( R_n(f) \) (in (2)) cannot be calculated, due to the absence of negative data, inter-class similarity with intra-class variation, and the under-fitting of positive data caused by the low target proportion. In this
section, the problems of HSI one-class classification are first analyzed, which establishes a bridge between the problem of low target proportion and the under-fitting of the positive class in risk estimation based HSI one-class classification. After this, the conformance-based negative risk estimator is introduced to calculate a high quality risk for the negative class from the PU data, and then the positive representation enhancement strategy is introduced to solve the under-fitting problem of the positive class. In the section of experimental, the One-Class Risk Estimator proposed in this paper is empirically proved to be an effective way to solve the problem of inter-class similarity with intra-class variation in HSI one-class classification.

1) Problem Analysis for HSI One-Class Classification: From the risk of the Bayesian classifier, the binary classifier $f^*(x)$ can be obtained by $f^* = \text{argmin}_f R_{pn}(f)$, where $R_{pn}(f)$ is the expected misclassification rate of all the data distributed according to $P(x)$. Equation (1) can be simplified as follows:

$$R_{pn}(f) = \pi_p R_p^+(f) + (1 - \pi_p) R_n^-(f), \quad (2)$$

where the risk of the positive data is $R_p^+(f) = \mathbb{E}_p[(f(x^p), +1)]$ and the risk of the negative data is $R_n^-(f) = \mathbb{E}_n[(f(x^n), -1)]$. However, $R_{pn}(f)$ cannot be directly calculated in the case of a lack of negative data. What is more, the ratio of the weight of the positive risk and the weight of the negative risk is limited to $\pi_p/(1 - \pi_p) = P(Y = +1)/P(Y = -1)$.

The assumption of Rademacher complexity [40], which makes the classifier learned from PU data have tighter estimation error bounds, is not satisfied when a flexible deep neural network, i.e., a CNN, is used. In order to make the risk estimated from the flexible deep neural network be of high quality, the assumption that there is $C_l$ such that $\max_y|f(x, y)| \leq C_l$ is made to ensure that the risk estimated from PU data is consistent with $R_{pn}(f)$ [50]. In addition, so that the gradient can be computed everywhere, sigmoid loss $l_{\text{sig}}(f(x), y) = 1/(1 + e^{yf(x)})$ is usually chosen as the loss function. Compared with the cross-entropy loss function, the sigmoid loss is a two-end saturation function, where we let $t = yf(x)$, which means that $t \rightarrow -\infty$ will lead to $l_{\text{sig}}(t) \rightarrow 0$, and $t \rightarrow +\infty$ will lead to $l_{\text{sig}}(t) \rightarrow 0$.

When low target proportion meets a two-end saturated loss function, the problem of positive class under-fitting may arise. Based on the observation shown in Fig. [13], we found that, although the overall risk is decreasing, the low target proportion causes the risk of the positive class to first rise at the beginning of the training. Due to the utilization of a two-end saturated loss function, when the risk of the positive class reaches its maximum value, the gradient brought by the positive data is a small value, and the gradient from the positive data plays only a small role in the overall gradient, so that the positive data are not fully fitted.

According to the above analysis, the conformance-based negative risk estimation and a positive representation enhancement strategy are proposed to allow a deep neural network to be used for HSI one-class classification.

2) Conformance Based Negative Risk Estimation: According to (2), the risk estimation include two portion, and it is unrealistic to estimate the risk of negative class directly through negative samples due to the lack of negative data. One of the insights of this paper is how to estimate reliable negative risk in HSI one-class classification, and the proposed negative risk $\tilde{R}_{nc-n}(f) \rightarrow R_n^-(f)$ as the number of positive and unlabeled samples $n_p, n_u \rightarrow \infty$ for a fixed $f$.

As the marginal distribution of the unlabeled samples is equal to the weighted sum of the marginal distributions of the positive and negative classes, i.e., $P(x) = \pi_p P_p(x) + (1 - \pi_p) P_n(x)$, the negative risk can be calculated in another way:

$$R_n^-(f) = (R_u^- - \pi_p R_p^-(f))/(1 - \pi_p), \quad (3)$$

where $R_u^-(f) = \mathbb{E}_u[(f(x^u), -1)]$ and $R_p^-(f) = \mathbb{E}_p[(f(x^p), -1)]$. The negative risk can be estimated as follows:

$$\hat{R}_{nc-n}^-(f) = (\hat{R}_u^- - \pi_p \hat{R}_p^-)/(1 - \pi_p), \quad (4)$$

where $\hat{R}_u^-(f) = 1/n_u \sum_{i=1}^{n_u} l(f(x^u_i), -1)$ and $\hat{R}_p^-(f) = 1/n_p \sum_{i=1}^{n_p} l(f(x^p_i), -1)$, and $n_p$ and $n_u$ are the number of PU samples, respectively. However, the negative risk estimated in this way is low quality. Since there is no constraint, $\hat{R}_{nc-n}^-(f)$
will cause the overall risk to become negative, which will result in serious overfitting \cite{36}. Formally, the negative risk in the One-Class Risk Estimator can be estimated by:

\[ \widehat{R}_{oc-n}(f) = |\widehat{R}_u(f) - \pi_p \widehat{R}_p(f)| / (1 - \pi_p). \tag{5} \]

When \( \widehat{R}_{oc-n}(f) \) is estimated to be negative, \( \widehat{R}_{oc-n}(f) \) can perform gradient ascent automatically, with the help of the existing deep learning framework, to alleviate the overfitting problem. \( \pi_p \widehat{R}_p(f) \) can also be regarded as an adaptive “flood level” \cite{41}, to avoid the overfitting of the positive data in the unlabeled samples. The consistency and MSE reduction of \( \widehat{R}_{oc-n}(f) \) are proved as follows, and detailed proof is provided in the Appendix A.

For a fixed \( f \), \( \widehat{R}_{oc-n}(f) \geq \widehat{R}_n(f) \), this implies the fact that \( \widehat{R}_{oc-n}(f) \) is biased, but the consistency of \( \widehat{R}_{oc-n}(f) \) can still be proved. That is, \( \widehat{R}_{oc-n}(f) \to R_n(f) \) as \( n_p, n_u \to \infty \) for a fixed \( f \).

**Lemma 1.** The data \( (X_p, X_u) \) can be divided into two sets, \( S^+(f) = \{(X_p, X_u)\} | \widehat{R}_u(f) - \pi_p \widehat{R}_p(f) \geq 0 \) \( \) and \( S^-(f) = \{(X_p, X_u)\} | \widehat{R}_u(f) - \pi_p \widehat{R}_p(f) < 0 \) \( \). Assuming there are \( C_l \) and \( \alpha > 0 \) such that \( 0 \leq l(f(x), \pm 1) \leq C_l \) and \( (1 - \pi_p)R_n(f) \geq \alpha \), the probability measure of \( S^-(f) \) can be bounded by:

\[ P(S^-(f)) \leq \exp(-2(\alpha/C_l)^2/(\pi_p^n + 1/n_u)). \tag{6} \]

The exponential decay of the bias and the consistency of the statistics can be proved based on Lemma 1. The right-hand side of \( [9] \) is denoted as \( \Delta_f \) in the following.

**Theorem 1** (bias and consistency). As the number of positive and unlabeled samples increases, \( n_p, n_u \to \infty \), the exponential decay shows up in the bias of \( \widehat{R}_{oc-n}(f) \):

\[ 0 \leq \mathbb{E}_{X_p, X_u}[\widehat{R}_{oc-n}(f) - R_n(f)] \leq 2(1 - \pi_p)\pi_p C_l \Delta_f. \tag{7} \]

Furthermore, we let \( X_{n_p, n_u} = \pi_p / \sqrt{\pi_p} + 1 / \sqrt{n_u} C_a = C_l \sqrt{\ln(2/\sigma)/2}(1 - \pi_p) \) and for any \( \sigma > 0 \), with the probability at least \( 1 - \sigma - \Delta_f \),

\[ |\widehat{R}_{oc-n}(f) - R_n(f)| \leq C_\sigma \cdot X_{n_p, n_u}. \tag{8} \]

Equation \( [8] \) implies the fact that \( \widehat{R}_{oc-n}(f) \to R_n(f) \) as \( n_p, n_u \to \infty \) for a fixed \( f \). Moreover, the characteristic of \( \widehat{R}_{oc-n}(f) \) in the reduction of MSE can also be described in the case of \( R_n(f) \) be overestimated.

**Theorem 2** (MSE reduction). The MSE of \( \widehat{R}_{oc-n}(f) \) is less than that of \( R_n(f) \):

\[ \mathbb{E}[(\widehat{R}_{oc-n}(f) - R_n(f))^2] \leq \mathbb{E}[(R_n(f) - R_n(f))^2]. \tag{9} \]

Based on these properties, the \( R_{oc-n}(f) \) can be used for the negative risk estimation.

3) **Positive Representation Enhancement Strategy:** The problem of low target proportion leads to under-fitting of the positive class in risk estimation based one-class classification when two-end saturated loss functions are used. One solution is to enhance the representation of the positive class and accelerate the reduction of positive risk to avoid under-fitting.

The first step in positive representation enhancement is to rebalance the distribution. Based on \( [2] \), the ratio of the weight of the positive risk and the weight of the negative risk is limited to \( P(Y = +1)/P(Y = -1) \), and this ratio does not change with the number of training samples. Beyond this ratio, the balancing factor \( \alpha_p \in [0, 1] \) is introduced in the estimator to rebalance the risk between the positive and negative classes, which aims to balance the distribution of the different classes without changing the positive data distribution and negative data distribution. Formally, the One-Class Risk Estimator is defined as:

\[ \widehat{R}_{oc}(f) = \alpha_p \widehat{R}_{oc-p}(f) + (1 - \alpha_p) \widehat{R}_{oc-n}(f). \tag{10} \]

A naive way is to reweight each positive sample by \( \alpha_p \), however, positive data participate in the process of negative distribution estimation, i.e., \( \widehat{R}_p(f) \), so this naive way will change the distribution of the negative data in the input space. The distribution-rebalancing strategy operates on the class level, and does not change the distribution of the positive and negative classes, but enhances the representation of the positive class by adjusting the probability ratio of the positive and negative classes \( P(Y = +1)/P(Y = -1) \) in the input space.

The second step is hard sample mining based positive risk estimation. Hard sample mining has been widely used in supervised learning; however, due to the lack of negative samples and the positive samples participating in the calculation of negative risk, how to mine hard samples in PU data is not intuitive. In the setting of learning from PU data, the positive dataset is considered to be a clean dataset, which implies that the positive dataset has no negative samples. Inspired by Focal Loss \cite{42}, the dynamic modulating factor is only added for each positive sample in the risk estimation of the positive class. Formally, the risk of the positive class in the One-Class Risk Estimator is defined as:

\[ R_{oc-p}(f) = \mathbb{E}_p[(1 - \pi_t)^\gamma l(f(x_p^i), +1)], \tag{11} \]

which can be estimated as follows:

\[ \widehat{R}_{oc-p}(f) = (1/n_p) \sum_{i=1}^{n_p} (1 - \pi_t)^\gamma l(f(x_p^i), +1), \tag{12} \]

where \( \gamma > 0 \) is the tunable focusing parameter, and \( \pi_t \in [0, 1] \) implies the confidence of sample \( x_p^i \) being judged to be a positive sample. Two properties are noted in the positive risk estimation. 1) When sample \( x_p^i \) is misclassified, the dynamic modulating factor will be near to 1, and the risk of this sample will be unaffected. When this sample is well classified, the dynamic modulating factor will be near to 0. 2) The focusing parameter controls the weight decrease rate of the easy samples. When \( \gamma = 0 \), the positive risk estimation is equivalent to the traditional positive risk estimation.

Finally, the proposed One-Class Risk Estimator for HSI one-class classification can be formulated as:

\[ \widehat{R}_{oc}(f) = \alpha_p \widehat{R}_{oc-p}(f) + (1 - \alpha_p) \widehat{R}_{oc-n}(f), \]

where the positive risk \( \widehat{R}_{oc-p}(f) \) can be calculated as shown in \( [12] \), and the negative risk \( \widehat{R}_{oc-n}(f) \) can be calculated as shown in \( [5] \).
3) Metrics: In this paper, the methods are evaluated using the F1-score ($\times 100$). Precision represents how many of the

$$\text{Precision} = \frac{TP}{TP + FP}$$

and

$$\text{Recall} = \frac{TP}{TP + FN}$$

were the default parameters and obtained better initial model parameters, and then the sigmoid loss function is used to train the network. In this study, the network was warmed up for 20 epochs.

Another potential source of instability comes from the dynamic modulating factor, which can become a very small number as $p_i$ approaches 1, leading to NaN in the gradient calculations, so probability clamping is used to limit the maximum of $p_i$ to 0.999.

B. FreeOCNet in HOneCls

FreeOCNet is designed to extract global spectral-spatial features to verify the proposed risk estimator, which was inspired by recent work [20]. This unified FCN is depicted in Fig. 2.

The basic module of the encoder is a spectral-spatial attention (SSA)-convolutional layer (Conv $3 \times 3$) with a group normalization (GN)-rectified linear unit (ReLU). The SSA refines the features by weighting each pixel adaptively, which differs from spectral attention [20], where weighting is conducted only in the spectral dimension. The implementation of SSA is similar to that of a convolutional block attention module (CBAM) [43]. The combination of a Conv $3 \times 3$-ReLU with stride 2 replaces the $3 \times 3$ pooling layer to reduce the spatial size of the feature maps, to align the center of the receptive field and its projected location. A stem is placed at the front of the network to compress the data with different spectral channels into uniform channels.

A lightweight decoder with a fixed number of channels (128) is used in this FCN. The decoder consists of an alternately stacked $3 \times 3$ convolutional layer and an upsampling layer with scale 2.

We use the fusion of high-level features and low-level features to obtain better semantic information and maintain better spatial details. Differing from the fusion approach of UNet, the approach of point-wise addition between features with the same spatial size is adopted to fuse the high-level and low-level features, which makes the optimization easier by residual connection [44]. The $1 \times 1$ convolution is adopted to adjust the number of channels in the low-level feature maps to match the channel requirement of the lightweight decoder.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental Settings

1) Dataset: Three benchmark aerial hyperspectral images were used in the experiments, i.e., the WHU-Hi-HongHu dataset (Fig. 3a), the WHU-Hi-LongKou dataset (Fig. 3b), and the WHU-Hi-HanChuan dataset (Fig. 3c) [45]. The classes in these hyperspectral images are similar in texture and spectra. Therefore, one-class classification on these three datasets is a huge challenge.

One hundred positive samples and unlabeled samples, which were 40 times the number of positive samples, were randomly selected from the images to participate in the training. In addition to some inaccurate annotations, 20 kinds of ground objects were selected for the one-class classification. More detailed information about the HSI datasets is provided in Table I. In order to avoid the impact of inaccurate estimation of the class prior on the results of the PU learning, we adopted the approximate real class prior estimated from the ground truth.

2) Training Details: In order to demonstrate the robustness of the proposed method, all the ground objects were classified with the same hyperparameters in the proposed HOneCls. To make a fair comparison in the deep learning-based methods, we also aligned the hyperparameters of all the deep learning-based methods. All deep learning-based methods were trained for 1000 epochs using a stochastic gradient descent optimizer. In addition, the weight decay was set to 0.0001 and the momentum was set to 0.9 for all the experiments, with no modification. $\alpha_p = 0.3$ and $\gamma = 0.1$ were the default settings, and experiments were conducted to determine these two parameters, according to [42]. The experiments were conducted using an NVIDIA Tesla P100 GPU.

3) Metrics: In this paper, the methods are evaluated using the F1-score ($\times 100$). Precision represents how many of the

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental Settings

1) Dataset: Three benchmark aerial hyperspectral images were used in the experiments, i.e., the WHU-Hi-HongHu dataset (Fig. 3a), the WHU-Hi-LongKou dataset (Fig. 3b), and the WHU-Hi-HanChuan dataset (Fig. 3c) [45]. The classes in these hyperspectral images are similar in texture and spectra. Therefore, one-class classification on these three datasets is a huge challenge.

One hundred positive samples and unlabeled samples, which were 40 times the number of positive samples, were randomly selected from the images to participate in the training. In addition to some inaccurate annotations, 20 kinds of ground objects were selected for the one-class classification. More detailed information about the HSI datasets is provided in Table I. In order to avoid the impact of inaccurate estimation of the class prior on the results of the PU learning, we adopted the approximate real class prior estimated from the ground truth.

2) Training Details: In order to demonstrate the robustness of the proposed method, all the ground objects were classified with the same hyperparameters in the proposed HOneCls. To make a fair comparison in the deep learning-based methods, we also aligned the hyperparameters of all the deep learning-based methods. All deep learning-based methods were trained for 1000 epochs using a stochastic gradient descent optimizer. In addition, the weight decay was set to 0.0001 and the momentum was set to 0.9 for all the experiments, with no modification. $\alpha_p = 0.3$ and $\gamma = 0.1$ were the default settings, and experiments were conducted to determine these two parameters, according to [42]. The experiments were conducted using an NVIDIA Tesla P100 GPU.

3) Metrics: In this paper, the methods are evaluated using the F1-score ($\times 100$). Precision represents how many of the
positive samples predicted by the classifier are truly positive samples, and recall is how many truly positive samples are correctly predicted by the classifier. As the harmonic average of the precision and recall, the F1 score is suitable for the scenarios of one-class classification.

\[ F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \]

The focus of this paper is on one-class classification, so attention is paid to the F1-score of each ground object. Experiments are required for each class of ground object, and there is no correlation between one-class classification experiments for different ground objects in the same dataset. What’s more, the average F1-score of different ground objects was calculated to test the robustness of the one-class classification algorithm on different ground objects. Since different datasets were shot under different conditions, the average F1-score was calculated for different ground objects on the same dataset.

All the experiments were repeated five times. The average F1-score of five experiments and the distribution map with the highest F1-score are reported.

4) Methods: As a control experiment, unlabeled samples were treated as negative data trained by binary cross-entropy (BCE) loss, to show that it is inappropriate to simply treat unlabeled data as negative data. We compared the proposed method with the other main one-class classification methods in the remote sensing community: one-class support vector machine (OCSVM) [46], the positive and unlabeled learning algorithm (PUL) [18], the positive and background learning algorithm (PBL) [15], and biased support vector machine (BSVM) [4]. The \textit{HOneCls} proposed in this paper can demonstrate the superiority of the risk estimation in the classification of HSI, which has inter-class similarity with intra-class variation, in the absence of negative data compared with the above methods. The proposed method was compared with uPU [35] and nnPU [36], to demonstrate the importance of considering the problem of low target proportion in HSI. \textit{FreeOCNet} was used as the global spectral-spatial feature extractor for BCE, uPU, nnPU, and \textit{HOneCls}.

B. Experiment 1: HongHu Dataset

The HongHu HSI has 270 channels from 400 to 1000 nm, and the spatial resolution of this imagery is about 0.043 m. The HongHu data used in this study were selected from the original HongHu imagery to remove most of the pixels without the labels. The selected hyperspectral image was 678 \times 465 pixels in size, and eight unambiguous land-cover types were selected for the one-class classification. The distribution maps are shown in Fig. 4 and the F1-scores are listed in Table II. From the distribution maps, \textit{HOneCls} accurately identifies all the land-cover types. The \textit{HOneCls} framework obtains the best F1-score in all the selected target objects, and the average F1-score (95.61) of \textit{HOneCls} is significantly higher than that of the other classical one-class classification methods, which indicates that \textit{HOneCls} shows good robustness on different objects. It can be seen that the risk estimation based methods (uPU, nnPU, \textit{HOneCls}) obtain better results than other one-class classification methods, which usually used for RGB or multi-spectral remote sensing images, in the remote sensing community. The risk estimation based \textit{HOneCls} has better performance on HSI with inter-class similarity and intra-class variation. However, uPU and nnPU do not recognize the positive class for the low target proportion

| Dataset         | Classes selected for one-class classification | Labeled samples for each class | Unlabeled samples for each class | Validation samples for each class |
|-----------------|-----------------------------------------------|-------------------------------|----------------------------------|-----------------------------------|
| HongHu          | cotton, rape, chinese cabbage, cabbage, tuber mustard, brassica parachinensis, carrot, white radish (8 Classes) | 100                           | 4000                             | 290878                            |
| LongKou         | corn, sesame, broad-leaf soybean, rice, water (5 Classes)                           | 100                           | 4000                             | 203642                            |
| HanChuan        | strawberry, cowpea, soybean, water spinach, watermelon, road, water (7 Classes)     | 100                           | 4000                             | 255930                            |

| Class            | Class prior | OCSVM | BSVM | PUL | PBL | BCE | uPU | nnPU | \textit{HOneCls} |
|------------------|-------------|-------|------|-----|-----|-----|-----|------|------------------|
| cotton           | 0.3769      | 60.34 | 87.49| 92.50| 92.49| 10.02| 58.69| 99.19| 99.33            |
| rape             | 0.1317      | 59.96 | 83.83| 65.84| 66.05| 55.83| 94.25| 94.70| 98.59            |
| chinese cabbage  | 0.0544      | 55.89 | 64.94| 57.25| 57.00| 79.77| 34.44| 34.11| 90.58            |
| tuber mustard    | 0.0367      | 17.40 | 36.40| 38.09| 38.77| 58.88| 0    | 0    | 95.52            |
| cabbage          | 0.0319      | 62.71 | 92.59| 66.78| 65.20| 79.42| 0    | 0    | 98.39            |
| brassica parachinensis | 0.0194 | 29.16 | 35.77| 53.57| 52.40| 87.06| 0    | 0    | 94.56            |
| white radish     | 0.0119      | 23.07 | 52.76| 40.00| 38.74| 86.52| 0    | 0    | 94.30            |
| carrot           | 0.0102      | 32.60 | 44.59| 48.88| 48.63| 81.21| 0    | 0    | 93.58            |
| Average F1-score | 42.64       | 62.30 | 57.86| 57.41| 67.34| 23.05| 28.50|      | 95.61            |

TABLE I
Details of the HSI Datasets (20 Classes in Total)

TABLE II
The F1 scores for the HongHu Dataset
ground objects, i.e., under-fitting of the positive class occurs, and the proposed $HOneCls$ has the ability to extract ground objects accurately in the case of low target probability. In other words, objects with a small class probability that cannot be identified without considering the low target proportion are identified by $HOneCls$.

Another interesting finding is that BCE outperforms existing one-class classification methods in the remote sensing...
community for some low probability targets, such as tuber mustard, brassica parachinensis, white radish and carrot. The low target probability means that there are fewer positive samples in the unlabeled data, when the BCE will be less affected by mislabeled negative training samples, however, the BCE performs worse in one-class classification of ground objects with a large proportion of classes, such as cotton and rape. The F1-score of the method proposed in this paper over 90 for all ground objects and significantly improves the average F1-score of HOneCls by 28.27 compared to the second best method.

C. Experiment 2: LongKou Dataset

The LongKou HSI has 270 bands from 400 to 1000 nm, and the spatial resolution is about 0.463 m. The image is of 550×400 pixels in size. To avoid the imprecise labels, five of the nine ground objects are selected for the one-class classification.

The detection maps for the LongKou dataset are shown in Fig. 5 and the F1-scores are listed in Table III. Similar to the results for the HongHu dataset, HOneCls shows good robustness to the ground objects in the LongKou dataset. From the distribution maps, it can be seen that the risk estimation based methods obtain better one-class classification results, but uPU and nnPU do not recognize the positive class for the low target proportion ground objects, i.e., the under-fitting of the positive class occurs. The HOneCls framework obtains good F1-scores in all the land-cover types and HOneCls accurately identifies all the land-cover types. In particular, objects with a small class prior that cannot be identified without considering the low target proportion are identified by HOneCls. The One-Class Risk Estimator performs slightly worse than nnPU in the corn class, but the proposed risk estimator shows a significant improvement in average F1-score. The average F1-score of HOneCls is 98.07, and the second best average F1-score is BSVM with 84.87. The results of BCE show that this classical binary semantic segmentation method is not suitable for the PU data based one-class classification task, and the results show obvious overfitting.

D. Experiment 3: HanChuan Dataset

The HanChuan imagery has 274 channels from 400 to 1000 nm, and the image is of 1217×303 pixels in size. The spatial resolution of the HanChuan imagery is 0.109 m. In particular, the HanChuan dataset is overlaid with distinct shadows, which significantly increases the spectral variability. Seven kinds of ground objects without obvious imprecise labeling were selected for detection.

The distribution maps for the HanChuan dataset (Fig. 6) also show the improvement of the robustness of the One-Class Risk Estimator for the ground targets in the HanChuan dataset. HOneCls obtains excellent one-class classification results, for both the normal and low target proportion data, for all the classes. The F1-scores for the HanChuan dataset are listed in Table IV. The HOneCls framework obtain a good F1-score in all the land-cover types. The One-Class Risk Estimator performs slightly worse than nnPU in the water class, but the proposed risk estimator shows a significant improvement in average F1-score. The average F1-score of HOneCls is 93.63, which is significantly higher than that of the second best method with an improved F1-score of 21.04.

| Class          | Class prior | OC-SVM  | BSVM  | PUL  | PBL  | BCE  | uPU  | nnPU | HOneCls |
|----------------|-------------|---------|-------|------|------|------|------|------|---------|
| water          | 0.3048      | 68.80   | 88.35 | 95.73| 95.35| 63.64| 98.06| 98.84| 98.89   |
| broad-leaf soybean | 0.2873     | 63.58   | 78.12 | 84.69| 84.74| 14.53| 48.68| 90.67| 92.60   |
| corn           | 0.1569      | 67.99   | 93.97 | 89.24| 89.41| 26.05| 93.37| 99.65| 99.58   |
| rice           | 0.0539      | 65.24   | 94.96 | 87.09| 85.92| 55.05| 57.72| 73.23| 99.92   |
| sesame         | 0.0138      | 24.34   | 68.95 | 67.13| 66.87| 95.01| 0    | 0    | 99.39   |
| Average F1-score |           | 57.99   | 84.87 | 84.78| 84.46| 50.85| 59.57| 72.48| 98.07   |

| Class          | Class prior | OC-SVM  | BSVM  | PUL  | PBL  | BCE  | uPU  | nnPU | HOneCls |
|----------------|-------------|---------|-------|------|------|------|------|------|---------|
| water          | 0.2045      | 60.81   | 94.38 | 90.74| 90.69| 81.47| 92.52| 97.91| 97.82   |
| strawberry     | 0.1213      | 67.86   | 80.43 | 35.70| 35.89| 72.40| 51.20| 87.64| 93.68   |
| cowpea         | 0.0617      | 34.46   | 45.38 | 34.27| 37.26| 37.49| 0    | 0    | 85.66   |
| road           | 0.0503      | 37.46   | 65.81 | 13.60| 13.62| 70.69| 0    | 0    | 91.56   |
| soybean        | 0.0279      | 49.85   | 46.27 | 52.39| 52.23| 67.05| 0    | 0    | 96.92   |
| watermelon     | 0.0123      | 13.99   | 17.80 | 20.46| 21.18| 82.42| 0    | 0    | 91.17   |
| water spinach  | 0.0033      | 21.65   | 19.47 | 24.71| 26.19| 96.67| 0    | 0    | 98.63   |
| Average F1-score |           | 40.87   | 52.79 | 38.84| 39.58| 72.59| 20.53| 26.51| 93.63   |
E. Analysis of $\alpha_p$ and $\gamma$

Ablation experiments were carried out to analyze $\alpha_p$ and $\gamma$. The purpose of this paper is to propose a one-class classification algorithm that can be applied to most of the ground objects, we therefore chose to select a set of parameters that performed well in most of the ground objects by reporting the average F1-score of all class of objects on the HongHu dataset (Table V). The experiment was repeated five times, and the average F1-score and the average standard deviation are reported.

Two key conclusions can be drawn from Table V: 1) The distribution-rebalancing strategy is an effective way to solve the problem of low target proportion. The weight of the class risk $\alpha_p$ is analyzed when $\gamma$ is fixed at 0 in Table V(a), and reliable one-class classification results can be obtained when $\alpha \geq 0.2$. These results show that the ratio of the weight of the risk between the positive and negative classes may not have to be limited to $\pi_p/(1 - \pi_p)$, and a reliable HSI one-class classifier can also be obtained from the conformance-based negative risk. We believe that the distribution rebalancing acts as a regularization in one-class classification because we force the model to assign approximate weights to the classes with different distribution frequencies. We note that the average F1-score decreases when $\alpha = 0.4$. What is more, we note that this is due to the larger standard deviation of the results when $\alpha = 0.4$. When $\gamma$ is added, this shortcoming is compensated, and the standard deviation of the results decreases under certain combinations of $\alpha$ and $\gamma$. For example, the results have a lower standard deviation when $\alpha = 0.2$ and $\gamma = 0.5$ or $\alpha = 0.3$ and $\gamma = 0.1$. 2) The second conclusion is that the dynamic weighting mechanism for the positive samples can further improve the performance of the One-Class Risk Estimator, thanks to the attention paid to the hard positive samples, and the dynamic weighting mechanism has the ability to reduce the standard deviation. From Table V(b) to

| $\alpha$ | Average F1 | $\alpha$ | $\gamma$ | Average F1 | $\alpha$ | $\gamma$ | Average F1 | $\alpha$ | $\gamma$ | Average F1 | $\alpha$ | $\gamma$ | Average F1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.1 | 74.09 (21.12) | 0.2 | 0.0 | 93.18 (6.12) | 0.3 | 0.0 | 93.64 (5.13) | 0.4 | 0.0 | 89.93 (12.38) | 0.5 | 0.0 | 93.09 (5.63) |
| 0.2 | 93.18 (6.12) | 0.2 | 0.1 | 90.95 (7.06) | 0.3 | 0.1 | 95.61 (0.62) | 0.4 | 0.1 | 94.00 (3.04) | 0.5 | 0.1 | 92.12 (6.19) |
| 0.3 | 93.64 (5.13) | 0.2 | 0.3 | 94.35 (3.86) | 0.3 | 0.3 | 95.53 (0.95) | 0.4 | 0.3 | 94.40 (3.72) | 0.5 | 0.3 | 93.82 (3.69) |
| 0.4 | 89.93 (12.38) | 0.2 | 0.5 | 95.53 (0.66) | 0.3 | 0.5 | 94.21 (3.78) | 0.4 | 0.5 | 95.49 (0.81) | 0.5 | 0.5 | 93.58 (5.19) |
| 0.5 | 93.09 (5.63) | 0.2 | 1.0 | 95.16 (0.79) | 0.3 | 1.0 | 94.89 (1.12) | 0.4 | 1.0 | 95.26 (0.98) | 0.5 | 1.0 | 94.42 (3.04) |
| 0.6 | 91.85 (7.02) | 0.2 | 3.0 | 94.45 (0.94) | 0.3 | 3.0 | 92.99 (4.28) | 0.4 | 3.0 | 93.22 (3.80) | 0.5 | 3.0 | 92.14 (4.69) |

(a) Varying $\alpha$ ($\gamma = 0$)  (b) Varying $\gamma$ ($\alpha = 0.2$)  (c) Varying $\gamma$ ($\alpha = 0.3$)  (d) Varying $\gamma$ ($\alpha = 0.4$)  (e) Varying $\gamma$ ($\alpha = 0.5$)
Fig. 7. Risk in the training stage. First row: When the class prior is small, the risk of the positive class does not decrease without considering the low target proportion. The class prior of the cotton class is larger, and the risk is normally reduced. Second row: After considering the low target proportion, the risk of the positive class decreases normally.

Table VI lists the average F1-scores of the different FCNs, where the best results are obtained by the proposed FreeOCNet extractor. The FCN proposed in this paper, i.e., FreeOCNet, avoids the boundary distortion caused by the sparse labels. The feature maps with the same size in the encoder and decoder are fused to maintain better spatial details, and the approach of point-wise addition obtains better results than the approach of concatenation.

F. Analysis of Different FCNs

Differing from the labels in the traditional semantic segmentation task, HSI labels are sparse. We used the mainstream FCNs (PSPNet, UNet, DeepLab-v3, DeepLab-v3+, and HRNet) as global spectral-spatial feature extractors with the proposed One-Class Risk Estimator for the one-class classification task. Due to the lack of boundary information in the sparse labels, the mainstream FCNs show the phenomenon of boundary distortion with HOneCls (Fig. 8). The average F1-scores of the different FCNs are listed in Table VI, where the best results are obtained by the proposed FreeOCNet extractor. The FCN proposed in this paper, i.e., FreeOCNet, avoids the boundary distortion caused by the sparse labels. The feature maps with the same size in the encoder and decoder are fused to maintain better spatial details, and the approach of point-wise addition obtains better results than the approach of concatenation.
G. Analysis of the Role of Class Prior in the Adaptive Flood Level

In this part, the role of class prior for HSI one-class classification is analyzed. The influence of the class prior on the detection results was analyzed at an interval of 0.05. The results are shown in Fig. 9.

The dots in Fig. 9 are the results of the approximate real class prior and its F1 score. We found that the class prior controls the precision and recall of the model. When the input class prior is smaller than the real class prior, the model obtains a high precision, and the recall increases with the increase of the class prior. When the input class prior is larger than the real class prior, the model obtains a high recall, but with the increase of the class prior, the model precision gradually declines. What is more, in the case of inputting a real class prior, the detection result will generally be better, but the input of a real class prior may not obtain the best result, as with the broad-leaf soybean class in the LongKou dataset. It can be seen from Fig. 9 that reliable results can still be obtained even when there is a 0.05 error in the estimated class prior. In conclusion, the class prior in adaptive “flood level” \((\pi_p^p, \tilde{R}_p^p(f))\) controls the fitting degree of the model to the positive class.

V. Conclusion

In this paper, we have proposed a novel framework for the HSI one-class classification task—HOneCls—which pushes the HSI classification methods from multi-class classification to one-class classification and reduces the annotation work in binary classification from the perspective of the class system. In the HOneCls framework, the One-Class Risk Estimator has also been proposed in this paper, which suitable for the characteristics of low target probability of HSI with inter-class similarity and intra-class variation and make it possible for the existing HSI multi-class feature extractors to extract one-class features. The conformance-based negative risk estimator is introduced to One-Class Risk Estimator to calculate a high quality risk for the negative class without negative data, and then the positive representation enhancement strategy is introduced to solve the under-fitting problem of the positive class. The unified FreeOCNet extractor has also been proposed, which can obtain better segmentation results under the case of sparse HSI labeling, compared with the current mainstream FCNs. The detection results were verified on three challenging aerial hyperspectral semantic segmentation datasets, where the proposed HOneCls framework showed a significant improvement over the other methods.

In the future, we will work on developing HSI one-class classification algorithms for the few-shot learning, and applying HSI one-class classification algorithms to specific species identification in forestry and agriculture.

REFERENCES

[1] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, “Hyperspectral remote sensing data analysis and future challenges,” IEEE Geoscience and Remote Sensing Magazine, vol. 1, no. 2, pp. 6–36, Jun 2013.

[2] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, “Advanced spectral classifiers for hyperspectral images: A review,” IEEE Geoscience and Remote Sensing Magazine, vol. 5, no. 1, pp. 8–32, Mar 2017.
International Conference on Machine Learning}, PMLR, Jan 2019, pp. 2820–2829.

[39] X. Chen, W. Chen, T. Chen, Y. Yuan, C. Gong, K. Chen, and Z. Wang, “Self-PU: Self boosted and calibrated positive-unlabeled training,” in \textit{Proceedings of the 37th International Conference on Machine Learning}, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, Jul 2020, pp. 1510–1519.

[40] G. Niu, M. C. du Plessis, T. Sakai, Y. Ma, and M. Sugiyama, “Theoretical comparisons of positive-unlabeled learning against positive-negative learning,” in \textit{Advances in Neural Information Processing Systems}, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., Jan 2016.

[41] T. Ishida, I. Yamane, T. Sakai, G. Niu, and M. Sugiyama, “Do we need zero training loss after achieving zero training error?” in \textit{Proceedings of the 37th International Conference on Machine Learning}, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, Jul 2020, pp. 4604–4614.

[42] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” \textit{IEEE Transactions on Pattern Analysis and Machine Intelligence}, vol. 42, no. 2, pp. 318–327, Feb 2020.

[43] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cham: Convolutional block attention module,” in \textit{Computer Vision – ECCV 2018}, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, Oct 2018, pp. 3–19.

[44] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in \textit{2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)}, Jun 2016, pp. 770–778.

[45] Y. Zhong, X. Hu, C. Luo, X. Wang, J. Zhao, and L. Zhang, “Whu-hi: Uav-borne hyperspectral with high spatial resolution (h2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with crf,” \textit{Remote Sensing of Environment}, vol. 250, p. 112012, Dec 2020.

[46] B. Schölkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, “Support vector method for novelty detection,” in \textit{Advances in Neural Information Processing Systems}, S. Solla, T. Leen, and K. Müller, Eds., vol. 12. MIT Press, Nov 2000.

\textbf{APPENDIX A. PROOF OF THE BIAS, CONSISTENCY AND MEAN SQUARED ERROR}

\textbf{1. Proof of the Bias and Consistency}

The probability density functions of $X_p$ and $X_u$ can be denoted as $p_p(X_p) = p_p(x^p) \cdots p_p(x^p_{n_p})$ and $p_u(X_u) = p(x^u_1) \cdots p(x^u_{n_u})$, respectively. The joint cumulative distribution function of $(X_p, X_u)$ can then be denoted as

$$F(X_p, X_u) = F_p(X_p) \cdot F_u(X_u),$$

where $F_p(X_p)$ and $F_u(X_u)$ are, respectively, the cumulative distribution functions of the PU data.

The probability measure of $S^{-}(f)$ can be defined as

$$P(S^{-}(f)) = \int_{(x_p, x_u) \in S^{-}(f)} dF(X_p, X_u).$$

For $P(S^{-}(f))$, we have the following mathematical transformation based on the assumptions:

$$P(S^{-}(f)) = P(\hat{R}_{-}^-(f) - \pi_p \hat{R}_{p}^-(g) < 0) \leq P(\hat{R}_{-}^-(f) - \pi_p \hat{R}_{p}^-(g) \leq (1 - \pi_p) R_{-}^- (f) - \alpha) \leq P((1 - \pi_p) R_{-}^- (f) - (\hat{R}_{-}^-(f) - \pi_p \hat{R}_{p}^-(g)) \geq \alpha).$$

To further prove Lemma 1, we apply McDiarmid’s inequality to the probability expression. Given the assumption that $0 \leq l(f(x), \pm 1) \leq C_l$, there will be no more than $C_l/n_p$ in the change of $\hat{R}_{p}^-(f)$ when a sample of $X_p$ is replaced, and no more than $C_l/n_u$ in the change of $\hat{R}_{u}^-(f)$ when a sample of $X_u$ is replaced. According to McDiarmid’s inequality, for $\alpha$ we have

$$P((1 - \pi_p) R_{-}^- (f) - (\hat{R}_{-}^-(f) - \pi_p \hat{R}_{p}^-(g)) \geq \alpha) \leq P(E[\hat{R}_{-}^-(f) - \pi_p \hat{R}_{p}^-(g)] - (\hat{R}_{-}^-(f) - \pi_p \hat{R}_{p}^-(g)) \geq \alpha) \leq \exp(-2\alpha^2/n_p (C_l \pi_p/n_p)^2 + n_u (C_l/n_u)^2) \leq \exp(-2\alpha^2/C_l^2).$$

Thus, Lemma 1 can be proved:

$$P(S^{-}(f)) \leq \exp(-2(\alpha/C_l)^2/(n_p^2/n_p + 1/n_u)).$$

Then, to prove Theorem 1, we have two steps. STEP 1 prove (7) with the definition and Lemma 1; STEP 2: prove the further conclusion (8) with (9), the absolute value inequality and McDiarmid’s inequality.

**STEP 1:**

Based on $\tilde{R}_{oc}^-(f) - \tilde{R}_{-}^-(f) = 0$ on $S^+(f)$, we have

$$E[\tilde{R}_{oc}^-(f) - \tilde{R}_{-}^-(f)] = E[\tilde{R}_{oc}^- - \tilde{R}_{-}^-].$$

**STEP 2:**

To gain the further conclusion, consider the deviation bound of $\tilde{R}_{oc}^-(f)$:

$$|\tilde{R}_{oc}^- - \tilde{R}_{-}^-| \leq |\tilde{R}_{oc}^- - \tilde{R}_{-}^-| + |\tilde{R}_{-}^- - \tilde{R}_{-}^-| \leq |\tilde{R}_{oc}^- - \tilde{R}_{-}^-| + |\tilde{R}_{-}^- - \tilde{R}_{-}^-|.$$

Still, we apply McDiarmid’s inequality to the part $|\tilde{R}_{-}^- - \tilde{R}_{-}^-|$. There will be no more than $n_p C_l/(n_p - \pi_p n_p)$ in the change of $\tilde{R}_{-}^-(f)$ when a sample of $X_p$ is replaced, and no more than $C_l/(n_u - \pi_p n_u)$ in the change of $\tilde{R}_{-}^- (f)$ when a sample of $X_u$ is replaced. For any $\epsilon > 0$:

$$P(\tilde{R}_{-}^-(f) - \tilde{R}_{-}^-) \geq \epsilon) \leq 2 \exp\left(-\frac{2\epsilon^2}{n_p C_l/(n_p - \pi_p n_p)^2 + n_u (C_l/n_u - \pi_p n_u)^2}\right) \leq 2 \exp\left(-\frac{2\epsilon^2}{C_l^2 (\pi_p n_p + 1/n_u)}\right).$$
Let \( \sigma \) equal the right side of the above formula, and after transformation we have
\[
\epsilon = \frac{1}{1 - \pi_p} \sqrt{\frac{\ln(2/\sigma)C^2(\pi_p^2/n_{p} + 1/n_{u})}{2}} 
\leq \sigma \left( \frac{\pi_p}{\sqrt{n_{p}}} + \frac{1}{\sqrt{n_{u}}} \right).
\]
Thus, the result is that, with the probability of at least \( 1 - \sigma \):
\[
|\hat{R}_n^-(f) - \mathbb{E}[\hat{R}_n^-(f)]| \leq \epsilon 
\leq C_\sigma \left( \frac{\pi_p}{\sqrt{n_{p}}} + \frac{1}{\sqrt{n_{u}}} \right) 
= C_\sigma \cdot \mathcal{X}_{n_p,n_u}.
\]
Given that \( |\hat{R}_{oc-n}^-(f) - \hat{R}_n^-(f)| > 0 \) with the probability of at most \( \Delta_f \), we can prove (8).
For any \( \sigma > 0 \) with the probability of at least \( 1 - \sigma - \Delta_f \)
\[
|\hat{R}_{oc-n}^-(f) - R_n^-(f)| \leq 0 + |\hat{R}_n^-(f) - \mathbb{E}[\hat{R}_n^-(f)]| 
\leq C_\sigma \cdot \mathcal{X}_{n_p,n_u}
\]
The proof of Theorem 1 is then finished.

2. Proof of the Mean Squared Error

Based on Theorem 1, the following formula also holds:
\[
\mathbb{E}[\hat{R}_{oc-n}^-(f)] - R_n^-(f) \leq \frac{2}{1 - \pi_p} \pi_p C_1 \Delta_f. \quad (16)
\]
From [16], when \( n_p, n_u \to \infty \) we can get \( R_n^-(f) = \mathbb{E}[\hat{R}_{oc-n}^-(f)] \).
Then:
\[
\mathbb{E}[(\hat{R}_n^-(f) - R_n^-(f))^2] 
= \mathbb{E}[(\hat{R}_n^-(f))^2] - 2R_n^-(f) \cdot \mathbb{E}[\hat{R}_n^-(f)] + R_n^-(f)^2,
\]
\[
\mathbb{E}[(\hat{R}_{oc-n}^-(f) - R_n^-(f))^2] 
= \mathbb{E}[(\hat{R}_{oc-n}^-(f))^2] - 2R_n^-(f) \cdot \mathbb{E}[\hat{R}_{oc-n}^-(f)] + R_n^-(f)^2.
\]
As a result:
\[
\mathbb{E}[(\hat{R}_n^-(f) - R_n^-(f))^2] - \mathbb{E}[(\hat{R}_{oc-n}^-(f) - R_n^-(f))^2] 
= 2R_n^-(f) \cdot \mathbb{E}[\hat{R}_{oc-n}^-(f) - \hat{R}_n^-(f)] 
= \int_{(X_p,X_u) \in S^-} 2R_n^-(f) \cdot [\hat{R}_{oc-n}^-(f) - \hat{R}_n^-(f)]dF(X_p,X_u) 
= \int_{(X_p,X_u) \in S^-} -4R_n^-(f) \cdot \hat{R}_n^-(f)dF(X_p,X_u)
\]
Given that \( R_n^-(f) > 0 \) and \( \hat{R}_n^-(f) < 0 \) on \( S^- \), we get
\[-4R_n^-(f) \cdot \hat{R}_n^-(f) > 0 \] on \( S^- \). Theorem 2 can then be proved:
\[
\mathbb{E}[(\hat{R}_n^-(f) - R_n^-(f))^2] - \mathbb{E}[(\hat{R}_{oc-n}^-(f) - R_n^-(f))^2] \geq 0
\iff E[(\hat{R}_{oc-n}^-(f) - \mathbb{E}[\hat{R}_{oc-n}^-(f)])^2] \leq \mathbb{E}[(\hat{R}_n^-(f) - \mathbb{E}[\hat{R}_n^-(f)])^2]