A Regularization-Based Positive and Unlabeled Learning Algorithm for One-Class Classification of Remote Sensing Data

Yan Zhang, Tianjiao Yang(✉), and Chenguang Zhang

College of Science, Hainan University, Haikou 570228, China

Abstract. Given some pixels with user-defined land cover types as labeled positive and negative samples, traditional remote sensing classification methods are sufficient to obtain optimal classification results. However, in many cases, only the positive pixels that users are interested in are labeled, and the negative samples are too diverse to be labeled. Such classification problems are referred to as one-class classification. Traditional learning methods are not suitable for one-class classification problems because labeled negative samples are required for these methods. In this paper, we propose a regularization-based positive and unlabeled learning method called RPUL for one-class classification of high-spatial-resolution aerial photographs. RPUL uses the implicit mixture model of restricted Boltzmann machines (IRBM) as the base framework of the classifier. With the help of a regularization term embedded into the loss function, an additional restriction is imposed on the negative class conditional PDF to ensure that it is as far from the positive class conditional PDF as possible. Thus, although no labeled negative training samples are available, the negative class conditional PDF can be estimated directly to obtain a binary classifier for the detection of the class of interest. The experimental results indicate that the new method provides high classification accuracy and outperforms state-of-the-art methods, including the cost-sensitive positive and unlabeled learning (CSPUL) and Gaussian domain descriptor methods.

Keywords: Remote-sensing · Positive and unlabeled data · Regularization term · Restricted boltzmann machines (RBM)

1 Introduction

Remote sensing technology has been widely used in various urban and environmental applications, such as land use change monitoring, water quality measurement and vegetation mapping. In general, remote sensing technologies rely on the classification and detection of targets in remote sensing images. Target detection refers to the technical process of distinguishing target and nontarget areas in an image and can essentially be seen as a process of machine learning: learn and construct a statistical classification model on the set of positive and negative labeled data and use this model to obtain the class label of other unlabeled pixels.

J. Cheng et al. (Eds.): CSS 2020, LNCS 12653, pp. 172–183, 2021.
https://doi.org/10.1007/978-3-030-73671-2_16

In recent years, with the development of machine learning and image processing technologies, remote sensing object detection methods have provided relatively good detection results. However, in some applications, we may be interested in only specific target areas and not other areas, which may incur the absence of negative labelled data [1–5]. For example, if the goal of a project is to detect roads from remote sensing data and update the information of an existing transport system, we may be reluctant to label forests and agricultural areas in the images as labeled negative training data. Moreover, even if we can afford the time and labor cost, it is still difficult to obtain a proper negative training dataset due to the high diversity of negative classes, particularly when high-spatial-resolution images are used. The classification problem in which the training data include only labeled positive training samples (target region) and not negative labeled training samples (non-target region) is called the one-class learning problem in machine learning [6, 7]. For this type of problem, traditional supervised classification methods are usually inefficient because traditional supervised classifiers require the classes in the remote sensing image to all have labeled training pixels. Thus, it is necessary to develop a stable and efficient remote sensing image target region detection method for cases where the training set contains only positive labeled samples.

At present, two strategies are used to address the one-class classification problem in the literature. The first strategy completely ignores unlabeled data and trains a classifier on only positive labeled data. Typical approaches of this type include the Gaussian model (GM) [7], one-class support vector machine (OCSVM) [8, 9] and support vector data description (SVDD) [10]. The GM assumes that the positive data are sampled from a Gaussian distribution. After density estimation of the positive labeled data, GM discriminates the positive class from the other classes by specifying an appropriate threshold. The disadvantage of GM is its inability to determine a suitable threshold. Moreover, when the data feature dimensionality is high, density estimation is usually very difficult. SVDD and OCSVM regard the original point as the only negative training case and find a hyperellipsoid that can exactly accommodate all positive examples or a hyperplane to separate the positive labeled data from the original point with the maximum margin. The disadvantage of these two methods is that their classification results are sensitive to the parameter values, so careful parameter tuning is required. The second strategy is semi-supervised learning, where unlabeled data are added to the learning process to compensate for missing negative labeled data. Representative works include semi-supervised one-class SVM (S^2OC-SVM) [11], 1-SVMs [13], positive and unlabeled learning method (PUL) [3] and cost-sensitive positive and unlabeled learning method (CSPUL) [13]. S^2OC-SVM and 1-SVMs improve the classifier by introducing manifold regular terms into the learning goal to make the labels smoother. However, the classification outcome is still sensitive to the parameter values. PUL and CSPUL are state-of-the-art methods for one-class classification. They use the estimated class prior to learning a classifier on positive and unlabeled data directly, where the unlabeled data play a similar role as the negative labeled data. However, the two-step strategy makes the classification precision strongly dependent on the class prior estimated in their first step.

In this paper, we propose a PUL method based on regularization, which is formalized as the Bhattacharyya coefficient (BC). The BC is a measure of the amount of overlap between two statistical samples or populations and is widely used in research on feature extraction and selection, image processing, speaker recognition, and phone clustering. We use the BC to impose an additional restriction on the unknown negative class conditional PDF to ensure that it is as far from the positive class conditional PDF as possible. Since the positive class conditional PDF and the mixture PDF of both the positive class and negative class can be estimated from the positive data and the unlabeled data, respectively, such a learning strategy makes it possible to obtain an estimate of the negative class conditional PDF. Moreover, we adopt an implicit mixture model of restricted Boltzmann machines (IRBM) to depict the data distribution to avoid the problem of simultaneously estimating the value of unknown class priors and unknown density functions. Thus, RPUL is established by embedding the BC between two class conditional densities into the risk function, i.e., the KL divergence between samples and the IRBM model, as a regularization item.

In contrast to other one-class methods, RPUL makes no assumptions about the data generation mechanism and requires no processing steps to estimate the threshold or class prior. We apply RPUL to classify data extracted from three scenes of a high-spatial-resolution image under the assumption that only positive data and unlabeled data are available for training. The experimental results illustrate the superiority of the proposed method compared with other state-of-the-art strategies.

2 The Proposed Approach

2.1 Preliminaries

Bhattacharyya Coefficient. The BC between two probability densities $p_1(\mathbf{v})$ and $p_2(\mathbf{v})$, with $\mathbf{v} \in \mathrm{R}^d$, is defined as

$$B = \int_{R^d} \sqrt{p_1(\mathbf{v})p_2(\mathbf{v})}d\mathbf{v}. \tag{1}$$

Clearly, the value of B is always confined within the interval [0, 1].

Implicit mixture model of RBMs (IRBM) [14]. The IRBM is a mixture model of RBMs with the mixed weights implicitly parameterized.

Let $\mathbf{v} \in \mathrm{R}^d$ be a vector of visible (observed) variables and $\mathbf{h}$ be a vector of hidden variables. Let K be the number of components (classes): K is two in this paper since we discuss only situations with two classes. Let $\mathbf{q}$ be a K-dimensional binary vector with only one element being one. Further, if $q_1 = 1$ and $q_2 = 0$, then the current $\mathbf{v}$ is a case of the positive class; otherwise, it is a case of the negative class. The energy function for IRBM is

$$E(\mathbf{v}, \mathbf{h}, \mathbf{q}) = \frac{1}{2}\sum_i (v_i - c_i)^2 - \sum_j h_j d_j - \sum_k q_k \sum_{i,j} W_{ijk} v_i h_j \tag{2}$$

$$c_i = \sum_k q_k C_{ik},\, d_j = \sum_k q_k D_{jk} \tag{3}$$

where W, C and D are the weight parameters, the visible unit biases and the hi-den unit biases, respectively, and k represents the component index. The joint distribution for the mixture model is

$$p_{\text{model}}(\mathbf{v}^s, \mathbf{h}^s, \mathbf{q}^s) = exp\big(-E(\mathbf{v}^s, \mathbf{h}^s, \mathbf{q}^s)\big)/Z \tag{4}$$

where

$$Z = \sum_{\mathbf{v},\mathbf{h},\mathbf{q}} exp(-E(\mathbf{v}, \mathbf{h}, \mathbf{q})) \tag{5}$$

is the partition function of the implicit mixture model. The components of IRBM are standard RBMs. The energy function of the k^{th} component derived from (2) is

$$E_k(\mathbf{v}, \mathbf{h}) = E(\mathbf{v}, \mathbf{h},\ q_k = 1) \tag{6}$$

The corresponding distribution function of the k^{th} component is

$$\begin{aligned} &p_{\text{model}}(\mathbf{v}^s, \mathbf{h}^s|q_k = 1) = exp\big(-E_k(\mathbf{v}^s, \mathbf{h}^s)\big)/Z_k \\ &Z_k = \sum_{\mathbf{v},\mathbf{h}} exp(E_k(\mathbf{v}, \mathbf{h})) \end{aligned} \tag{7}$$

Let $\theta = \{W, C, D\}$ be the set of model parameters. Given a set of N training cases $\{\mathbf{v}^1, ..., \mathbf{v}^N\}$, the learning process of IRBM is to maximize the log likelihood of $L = \sum_{n=1}^{N} \log p_{\text{model}}(\mathbf{v}^n; \theta)$ or to minimize the Kullback–Leibler (KL) distance between the empirical data distribution and the model distribution $KL(p_{data}(\mathbf{v})||p_{\text{model}}(\mathbf{v}\ ; \theta))$, where $p_{data}(\mathbf{v}) = \frac{1}{N}\sum_{i=1}^{n} \delta(\mathbf{v} - \mathbf{v}^n)$ and $\delta(\mathbf{v}-\mathbf{v}^n)$ is 1 only when $\mathbf{v} = \mathbf{v}^n$; otherwise, it is 0. IRBM can be trained by a contrastive divergence-like algorithm by sampling the conditional distributions $p(\mathbf{h}, \mathbf{q}|\mathbf{v})$ and $p(\mathbf{v}|\mathbf{h}, \mathbf{q})$. Sampling $p(\mathbf{h}, \mathbf{q}|\mathbf{v})$ is not straightforward and performed in two steps. First, the K-way discrete distribution $p(\mathbf{q}|\mathbf{v})$ is computed (see below) and sampled. Then, given $q_k = 1$, the k^{th} component RBM is selected and its conditional distribution $p(\mathbf{h}|\mathbf{v})$ is sampled. $p(\mathbf{q}|\mathbf{v})$ is given by

$$p(q_k = 1|\mathbf{v}) = \frac{\exp(-F(\mathbf{v}, q_k = 1))}{\sum_m \exp(-F(\mathbf{v}, q_m = 1))} \tag{8}$$

where

$$F(\mathbf{v}, q_k = 1) = \frac{1}{2}\sum_i (v_i - c_i)^2 - \sum_j \log\left(1 + \exp\left(\sum_i W_{ijk} v_i\right)\right) \tag{9}$$

2.2 Learning Framework

Notation

Let $\mathcal{Y} = \{+1, -1\}$ be the set of possible labels. Without loss of generality, we suppose only the first l cases in $\left\{\mathbf{v}^1, ..., \mathbf{v}^N\right\}$ are labeled with positive label $+1$ and the rest are unlabeled. Let $P = \left\{\mathbf{v}^1, ..., \mathbf{v}^l\right\}$ be the set of positive samples, and let $U = \left\{\mathbf{v}^{l+1}, ..., \mathbf{v}^N\right\}$ be the set of unlabeled samples.

Method

The goal of our method is to learn the posterior probability function $p(q_1 = 1|\mathbf{v})$. According to Bayes' rule,

$$p(q_1 = 1|\mathbf{v}) = \frac{p(\mathbf{v}|q_1 = 1)p(q_1 = 1)}{p(\mathbf{v})}. \tag{10}$$

Then, the positive conditional density function $p(\mathbf{v}|q_1 = 1)$,the mixture density $p(\mathbf{v})$ and the class prior $p(q_1 = 1)$ must be estimated. As the IRBM was adopted as the data description model, estimation of the class prior is replaced by estimation of the negative class conditional density function. However, because of the lack of labeled negative data, estimation of the negative class conditional density is not straightforward. To address this problem, we introduce the BC to obtain supernumerary information about the negative class conditional density to compensate for the absence of negative labeled data. This approach is reasonable. In fact, minimizing the BC between the conditional densities of two class, i.e., the amount of overlap, would lead to a negative class conditional density that is far from the positive class conditional density. Then, in the area far from the negative data, it holds that $p(\mathbf{v}|q_1 = 1)p(q_1 = 1)$ is approximately equal to $p(\mathbf{v})$. Notably, approximating $p(\mathbf{v}|q_1 = 1)p(q_1 = 1)$ as $p(\mathbf{v})$ is the starting point of the state-of-the-art one-class method [13] for estimating the class prior.

Finally, the proposed framework of RPUL is formulated to minimize

$$\begin{aligned} Z(\theta) = {} & KL(p_{\text{data}}(\mathbf{v}|q_1 = 1), p(\mathbf{v}|q_1 = 1; \theta_1)) \\ & + KL(p_{\text{data}}(\mathbf{v}), p(\mathbf{v}; \theta)) + \mu B(p(\mathbf{v}|q_1 = 1; \theta_1), p(\mathbf{v}|q_2 = 1; \theta_2)) \end{aligned} \tag{11}$$

where $\theta = \{\theta_1, \theta_2\}$ is the set of model parameters and θ_k is the set of parameters of the k^{th} component of IRBM. $KL(\bullet)$ is the Kullback–Leibler divergence. The first two items on the right side of the equal sign measure the degree of fit between the positive data and the first positive component of IRBM and the degree of fit between the unlabeled data and the complete IRBM, respectively. The final item is the BC regularization item, which ensures that the second component of IRBM captures the negative class conditional density precisely, as mentioned in the previous analysis. The trade-off between the data fit items and the regularization item is positive parameter μ, which is fixed at 0.1 in this paper.

Solution

As in the training process of IRBM, gradient descent is employed to solve optimization problem (11). To make the notation concise, the three terms on the right side of the equal sign of (11) are denoted by $KL_1(\theta_1)$, $KL_2(\theta)$, and $B(\theta)$. Given the samples $\mathbf{v}^s \in U$, the estimate of $B(\theta)$ is computed by

$$B(\theta) = \sum\nolimits_{\mathbf{v}^s} \sqrt{f(\mathbf{v}^s;\theta_1)g(\mathbf{v}^s;\theta_2)}. \tag{12}$$

Then, the derivative of $B(\theta)$ with respect to θ_k is

$$\frac{\partial B}{\partial \theta_k} = \frac{-1}{2\sqrt{p(q_1=1)p(q_2=1)}} \left[\begin{array}{l} \sum\nolimits_{\mathbf{v}^s} p(\mathbf{v}^s)\sqrt{p(q_1=1|\mathbf{v}^s)p(q_2=1|\mathbf{v}^s)}\dfrac{\partial F_k(\mathbf{v}^s)}{\partial \theta_k} \\ -\left(\dfrac{\sum_{\mathbf{v}^s} p(\mathbf{v}^s)\sqrt{p(q_1=1|\mathbf{v}^s)p(q_2=1|\mathbf{v}^s)}}{\sum_{\mathbf{v}} p(\mathbf{v})p(q_k=1|\mathbf{v})}\right) \\ \left(\sum\nolimits_{\mathbf{v}} p(\mathbf{v})p(q_k=1|\mathbf{v})\dfrac{\partial F_k(\mathbf{v})}{\partial \theta_k}\right) \end{array}\right] \tag{13}$$

where θ is omitted for brevity and $F_k(\mathbf{v}) = F(\mathbf{v}, q_k = 1)$. To compute the terms associated with the variable $\mathbf{v}$ of (13) exactly, we would need to sum over the joint space of all possible visible variables, which is an intractable task. Fortunately, we can address this problem using the CD learning algorithm, which has been found to be effective for training a variety of energy-based models. Based on the CD algorithm, we sample the mixed probability density $p(\mathbf{v})$ to compute the corresponding expectation terms and then obtain the approximation to the derivative of $B(\theta)$:

$$\frac{\partial B(\theta)}{\partial \theta_k^n} \approx \frac{-1}{2\sqrt{p(q_1=1)p(q_2=1)}} \left[\begin{array}{l} \sum\limits_{s=l}^{l+u} p_{data}(\mathbf{v}^s)\sqrt{p(q_1=1|\mathbf{v}^s)p(q_2=1|\mathbf{v}^s)}\dfrac{\partial F_k(\mathbf{v}^s)}{\partial \theta_k^n} \\ -\left(\dfrac{\sum\limits_{s=l}^{l+u} p_{data}(\mathbf{v}^s)\sqrt{p(q_1=1|\mathbf{v}^s)p(q_2=1|\mathbf{v}^s)}}{\sum\limits_{s=l}^{l+u} p\big((\mathbf{v}^s)^-\big)p(q_k=1|(\mathbf{v}^s)^-)}\right) \\ \left(\sum\limits_{s=l}^{l+u} p\big((\mathbf{v}^s)^-\big)p(q_k=1|(\mathbf{v}^s)^-)\dfrac{\partial F_k((\mathbf{v}^s)^-)}{\partial \theta_k^n}\right) \end{array}\right] \tag{14}$$

where $(\mathbf{v}^s)^-$ is obtained by the negative phase, which are the values of the visible variables after M steps of alternating sampling and $p(\mathbf{h}, \mathbf{q}|\mathbf{v})$ and $p(\mathbf{v}|\mathbf{h}, \mathbf{q})$. Otherwise, given $\mathbf{v}^s$, if the sampled $q_1 = 1$, let $s_k = 1$; else, let $s_k = 2$. Similarly, given $(\mathbf{v}^s)^-$, we can obtain the value of $(s_k)^-$. The derivative of F_k in (14) can be computed approximately as follows:

$$\frac{\partial F_k(\mathbf{v}^s)}{\partial W_{ijk}} = -p\big(h_j|\mathbf{v}^s, q_k=1\big)v_i^s \approx \begin{cases} -h_j^s v_i^s, & k = s_k \\ 0, & k \neq s_k \end{cases}, \tag{15}$$

$$\frac{\partial F_k((\mathbf{v}^s)^-)}{\partial W_{ijk}} = -p\big(h_j|(\mathbf{v}^s)^-, q_k=1\big)(v_i^s)^- \approx \begin{cases} -(h_j^s)^-(v_i^s)^-, & k = s_k^- \\ 0, & k \neq s_k^- \end{cases}, \tag{16}$$

$$\frac{\partial F_k(\mathbf{v}^s)}{\partial C_{ik}} = \begin{cases} -v_i^s + c_i, & k = s_k \\ 0, & k \neq s_k \end{cases}, \tag{17}$$

$$\frac{\partial F_k((\mathbf{v}^s)^-)}{\partial C_{ik}} = \begin{cases} -(v_i^s)^- + c_i, & k = s_k^- \\ 0, & k \neq s_k^- \end{cases} \tag{18}$$

$$\frac{\partial F_k(\mathbf{v}^s)}{\partial D_{jk}} = -p\big(h_j|\mathbf{v}^s, q_k = 1\big) \approx \begin{cases} -h_j^s, & k = s_k \\ 0, & k \neq s_k \end{cases}, \tag{19}$$

$$\frac{\partial F_k((\mathbf{v}^s)^-)}{\partial D_{jk}} = -p\big(h_j|(\mathbf{v}^s)^-, q_k = 1\big) \approx \begin{cases} -(h_j^s)^-, & k = s_k^- \\ 0, & k \neq s_k^- \end{cases} \tag{20}$$

The derivative of $KL_1(\theta_1)$ with respect to θ_1 and the derivative of $KL_2(\theta)$ with respect to θ can be computed by CD algorithm, as done in the preliminaries. After the derivatives are computed, the parameters of our model are iteratively updated as follows:

$$\theta_{new} = \theta_{old} - \eta\Delta\theta, \tag{21}$$

where η is the learning rate and

$$\Delta\theta = \frac{\partial(\mu B + KL_1 + KL_2)}{\partial\theta}. \tag{22}$$

Finally, for any given sample $\mathbf{v}$, following Bayes' decision theory, if $p(q_1 = 1|\mathbf{v}) > p(q_2 = 1|\mathbf{v})$, the label is positive. Otherwise, the label is negative, and $p(q_1 = 1|\mathbf{v})$ can be computed via formula (8).

Note that the computation of (22) simply involved applying the CD algorithm to the P set and U set, and the time complexity of the proposed method is the same as that of IRBM.

3 Experiments

In this section, we investigate the performance of the proposed RPUL for one-class classification of remote sensing data. The cost-sensitive LPU method (called CSLPU below) proposed in [13] is a state-of-the-art alternative learning method for the same positive/unlabeled scenario, and the Gaussian domain descriptor (GDD) methods are commonly used one-class classifiers. Hence, these methods are also compared with the proposed RPUL in our experiments.

3.1 Dataset Description

The initial dataset used in this paper was RIT-18 [15, 16], which is composed of very-high-resolution aerial photographs (4.7 cm GSD) acquired by an unmanned aircraft system (see Fig. 1). The dataset includes 6 VNIR spectral bands and 18 labeled object classes. The 2nd, 14th, 15th and 16th classes were chosen as positive classes in this paper because they are the first four classes that have a sufficient number of pixels (at least 1%

of the total pixels). The size of the photographs is 9393 × 5642, with a total of 52995306 pixels. We slid a 5 × 5 pixel template over the image and extracted 88 features for each pixel, including the mean, variance, homogeneity, contrast, and second moment of the six bands. All features were rescaled to the range [0, 1].

RPUL and CSLPU require positive and unlabeled data for training, whereas GDD and SVDD require only positive data. In general, more labeled training data results in higher accuracy but also increases the required labeling effort. In our experiments, for each class extraction, we randomly selected only 50 pixels of a class as labeled positive training samples: the labeled pixels were less than 9e−5% of the entire image. Additionally, for RPUL and CSLPU, we randomly selected an additional 1000 pixels from the entire image as the unlabeled dataset. As mentioned in the introduction, the classification results of GDD strongly depend on the tuned model parameters: high classification accuracy on the testing dataset is difficult to guarantee if these parameters are tuned with only positive data. To investigate the optimal performance, we used 1000 randomly selected background pixels of other classes in addition to the previously prepared positive labeled samples to tune the parameters. Finally, the remaining pixels of the photographs formed the test dataset. Moreover, to obtain statistically reliable results, ten different random realizations of the training data were considered for each classification, and the classification results were evaluated in terms of the overall accuracy (OA), F-measure (F), recall (R) and kappa coefficient (K) [17].

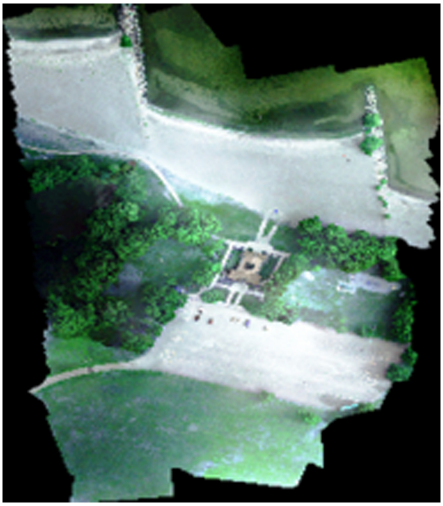

Fig. 1. RGB visualization of the RIT-18 dataset. This dataset has six spectral bands.

3.2 Model Development

RPUL. The RPUL model was developed in MATLAB. Typically, we used models with 200 latent variables. The value of the parameter μ in (22) was fixed at 0.1; the learning rate in (21) was set to 1e−3; and the weight decay was set to 1e−2. A momentum term was also used: 0.9 of the previous accumulated gradient was added to the current gradient. A temperature parameter was introduced to scale the free energies, similar to the training process of IRBM: the parameter was set to 100. We trained the model using the entire sample in both the *P* set and the *U* set until the class labels of the data did not change or the number of iterations reached 2000.

CSLPU. The CSLPU model was implemented in MATLAB. We used a Gaussian radial basis function (RBF) kernel and followed the empirical approach in [6] to tune the parameters. The number of basis functions was set to 300. The regularization parameter was tuned in the range [−3, 10] on a log scale with a step size of 1. The kernel width was tuned in the range that was computed by first estimating the median value of the distances from all samples to the randomly selected centroids and multiplying the median value by the numbers in the interval [−2, 10] on a log scale with a step size of 1. Moreover, CSLPU needs the class prior to be known first. We used the method in [13] to estimate the class prior, with the parameters tuned under the same settings as those used for CSLPU.

GDD. The GDD model was implemented via dd_tools. We used the simple Gaussian target distribution and tuned two parameters: the error on the target class in the range [0.1, 1] with a step size of 0. 1 and the regularization parameter in the range [0.1, 1] with a step size of 0.1. As for SVDD, only the samples in the *P* set were used to train the classifier using the tuned parameters.

3.3 Experimental Results

Every experiment was repeated ten times with randomly selected positive and unlabeled samples. Figure 2 shows the classification maps of one of the experiments of Fig. 1 for each land type, where (a) is the benchmarks, i.e., the true pixel labels, and (b), (c) and (d) are the classification results of RPUL, CSLPU and GDD, respectively. In general, RPUL provides the best classification results in the extraction of a single land type from the aerial photograph. Note that such good classification results are obtained in the situation with only 50 positive labeled pixels and no negative labeled pixels for training. Therefore, with the help of the regularization item, RPUL can learn additional information about the unknown negative training samples from the positive and unlabeled samples to construct a proper classifier even without the labeled negative training samples. CSLPU also provides relatively good results, particularly for water areas, but GDD produces poor results. Both RPUL and CSLPU used unlabeled samples to build the classifier, which may be the reason that they have better classification results than GDD. Moreover, CSLPU is slightly inferior to RPUL. The main reason is likely that the distribution of the positive class in the training set is not identical to the distribution in the unlabeled set since we selected only 50 positive samples as labeled samples; therefore, CSLPU might not be able to obtain an optimal estimate of the class prior to train the classifier. Table 1 compares the accuracy, F-measure, recall and kappa coefficient of the three methods for different land types. The results in Table 1 show that RPUL and CSLPU had similar best evaluation values and GDD provided the worst classification results, even with the parameters tuned on the set of additional negative labeled samples and positive labeled samples.

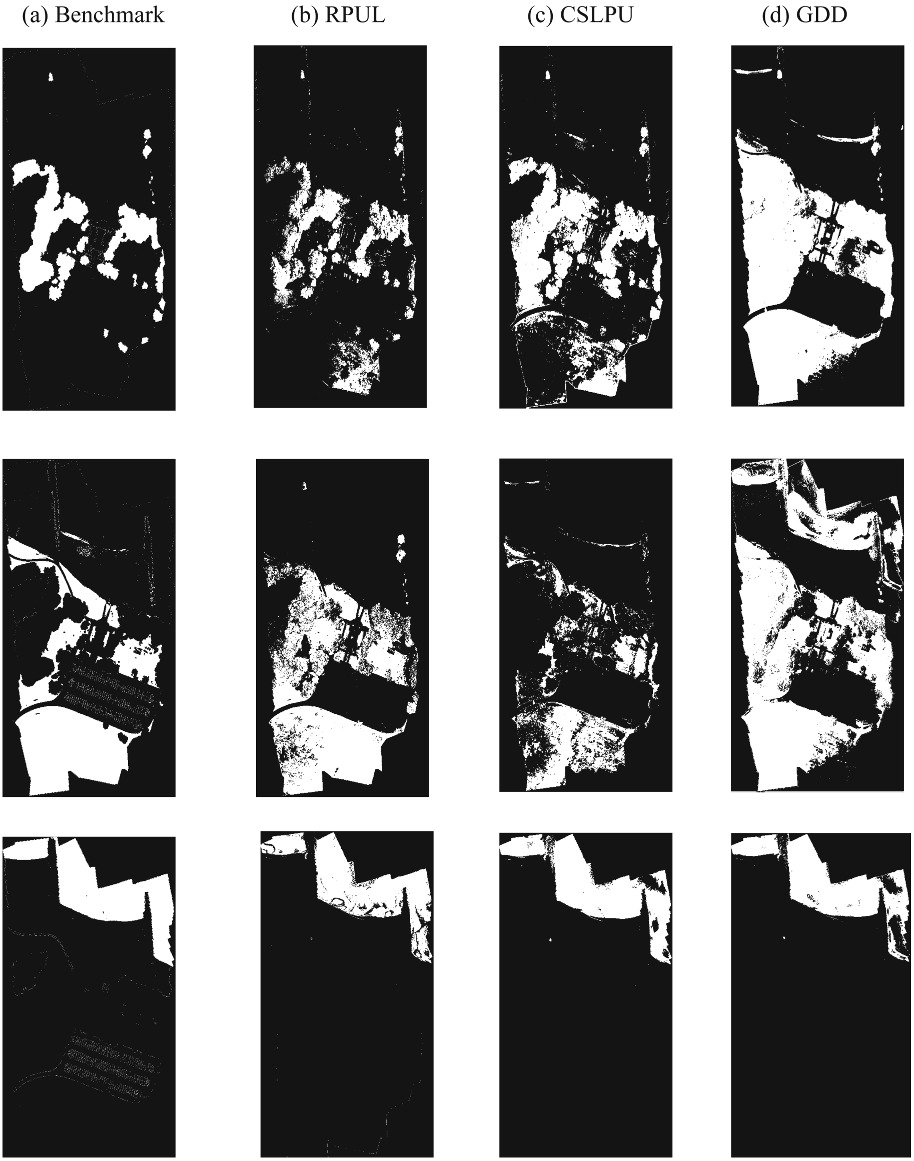

Fig. 2. Prediction maps of each land type. From the first row to the last row, prediction maps of tree, grass and water. White: positive; black: negative.

Table 1. The accuracy (OA), F-measure (F), recall (R) and kappa coefficient (K) of RPUL, CSLPU and GDD for all land types

Land Types	RPUL				CSLPU				GDD			
	OA	F	R	K	OA	F	R	K	OA	F	R	K
Tree	0.93	0.96	0.96	0.67	0.90	0.94	0.90	0.64	0.77	0.85	0.75	0.36
Grass	0.86	0.90	0.89	0.63	0.89	0.93	0.98	0.61	0.74	0.82	0.72	0.43
water	0.98	0.98	0.99	0.91	0.98	0.99	0.99	0.90	0.97	0.98	0.99	0.88

4 Conclusion

In this paper, we addressed the problem of one-class classification of remote sensing data by proposing a new BC-based positive and unlabeled learning algorithm. In contrast to other one-class methods, the proposed method makes no assumptions about the data generation mechanism and does not need a processing step to estimate the threshold or the class prior. Moreover, the proposed method is a semi-supervised learning method that requires only a small set of labeled positive data for classifier training. The experimental results indicated that the new algorithm achieves high classification accuracy, outperforming the CSPUL, SVDD, and GDD methods. In future work, we will apply the learning strategy to a generative adversarial network to further improve the performance of LPU methods.

Acknowledgment. This work is supported by Hainan Provincial Natural Science Foundation of China (Project No. 119MS004), the National Natural Science Foundation of China (61650303) and the Undergraduate Bilingual Course Construction Project of Hainan University "Advanced Mathematics" (Project No. hdsyk201903).

References

1. Rao, T., Rajinikanth, T.: Supervised classification of remote sensed data using support vector machine. Global J. Comput. Sci. Technol. **14** (2014)
2. Wenkai, L., Qinghua, G.: A new accuracy assessment method for one-class remote sensing classification. IEEE Trans. Geosci. Remote Sens. **52**, 4621–4632 (2014)
3. Wenkai, L., Qinghua, G., Elkan, C.: A positive and unlabeled learning algorithm for one-class classification of remote-sensing data. IEEE Trans. Geosci. Remote Sens. **49**, 717–725 (2011)
4. Xueqing, D., Wenkai, L., Xiaoping, L.: One-class remote sensing classification: one-class vs. binary classifiers. Int. J. Remote Sens. **39**, 1890–1910 (2018)
5. Kristen, J., Andreas, S.: Positive and unlabeled learning algorithms and applications: a survey. In: International Conference on Information, Intelligence, Systems and Applications (IISA) (2019)
6. du Plessis, M.C., Niu, G., Sugiyama, M.: Analysis of learning from positive and unlabeled data. In: Advances in Neural Information Processing Systems, MIT Press, Montreal, Quebec, Canada, pp. 703–711 (2014)
7. David, M.: Tax, One-class classification, Concept-learning in the absence of counter-examples. In: ASCI dissertation series, Delft Univ.Technol., Delft,The Netherlands (2001)

8. Schölkopf, B., Williamson, R.C., Smola, A.J., Shawe-Taylor, J., Platt, J.C.: Support vector method for novelty detection. In: NIPS, MIT Press, Colorado, USA, pp. 582–588 (1999)
9. Guerbai, Y., Chibani, Y., Hadjadji, B.: The effective use of the one-class SVM classifier for handwritten signature verification based on writer-independent parameters. Pattern Recogn. **48**, 103–113 (2015)
10. Tax, D.M., Duin, R.P.: Support vector domain description. Pattern Recogn. Lett. **20**, 1191–1199 (1999)
11. Goh, K.-S., Chang, E.Y., Li, B.: Using one-class and two-class SVMs for multiclass image annotation. IEEE Trans. Knowl. Data Eng. **17**, 1333–1346 (2005)
12. Mũnoz-Marí, J., Bovolo, F., Gomez-Chova, L., Bruzzone, L., Camp-Valls, G.: Semisupervised one-class support vector machines for classification of remote sensing data. IEEE Trans. Geosci. Remote Sens. **48**, 3188–3197 (2010)
13. du Plessis, M.C., Sugiyama, M.: Class prior estimation from positive and unlabeled data. IEICE Trans. Inf. Syst. **97**, 1358–1362 (2014)
14. V. Nair, G.E. Hinton, Implicit mixtures of restricted Boltzmann machines. In: Advances in Neural Information Processing Systems, pp. 1145–1152 (2009)
15. Kemker, R., Salvaggio, C., Kanan, C.: Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning, arXiv preprint arXiv:1703.06452 (2017)
16. Guoping, Y., Yingli, Z.: Target detection method of remote sensing image based on deep learning. In: Cross Strait Radio Science & Wireless Technology Conference (CSRSWTC) (2020)
17. Stehman, S.: Estimating the kappa coefficient and its variance under stratified random sampling. Photogram. Eng. Remote Sens. **62**, 401–407 (1996)