A novel semi-supervised hyperspectral image classification approach based on spatial neighborhood information and classifier combination

Kun Tan a, Jun Hu b, Jun Li b, Peijun Du c,⇑

a Jiangsu Key Laboratory of Resources and Environment Information Engineering, China University of Mining and Technology, Xuzhou, China
b Guangdong Key Laboratory for Urbanization and Geo-Simulation, School of Geography and Planning, Sun Yat-Sen University, Guangzhou, China
c Key Laboratory for Satellite Mapping Technology and Applications of State Administration of Surveying, Mapping and Geoinformation of China, Nanjing University, Nanjing, China

to tackle the aforementioned difficulties for remote sensing image processing tasks. However, it is often difficult for a traditional classifier to offer satisfactory performance in hyperspectral image classification especially with limited small training set. This observation has fostered the idea of semi-supervised learning, which adds unlabeled samples to the training set, without significant cost, to improve the capability of the classifier (Shahshahani and Landgrebe, 1994). In general, semi-supervised learning consists of five different models: generative models (Chapelle et al., 2006; Jin et al., 2013), graph-based methods (Zha et al., 2009; Zhu and Lafferty, 2005), transductive support vector machines (TSVMs) (Joachims, 1999; Tong and Koller, 2002), self-learning approaches (Grandvalet and Bengio, 2005; Rosenberg et al., 2005; Tuia et al., 2009), and multiview learning (Di and Crawford, 2010, 2011, 2012).

The main problems of semi-supervised learning approach are how to select the most helpful unlabeled samples and how to determine the class label of these new selected samples. In this paper breaking ties (BT) (Luo et al., 2004) method is applied to select the most useful unlabeled samples. It can greatly reduce the amount of computing time and improve the efficiency of the algorithm. At first, the label is estimated just through the initial classification map (Dópido et al., 2013). However, a small number
of initial labeled samples, likely with poor generalization, makes the classification problem very difficult. A modified BT method, called MBT, was integrated in classification (LORSAL) and segmentation (LORSAL-MLL), resulting in two new methods with active learning, called LORSAL-AL and LORSAL-MLL-AL (Li et al., 2011). Unlabeled samples are applied to improve the estimation of the class distributions, and the obtained classification is refined by using a spatial multi-level logistic prior (Li et al., 2010). Begüm proposed a strategy that can be embedded in any AL method to identify the most informative samples and to reduce the overall cost (Demir et al., 2014). With the combination of spectral and spatial information, it is widely used in remote sensing image classification, leading to obvious improvements in classification accuracy (Bioucas-Dias et al., 2013; Li et al., 2013). As a consequence, discriminative approaches can mitigate the curse of dimensionality because they just require smaller training sets (Di and Crawford, 2012; Tan et al., 2014).

In this paper, a novel approach is proposed to confirm the labels of unlabeled samples. This method is to reduce the difficulty of samples selection in the semi-supervised classification based on the spatial neighborhood information and classifier combination. “Spatial Neighborhood Information of Labeled samples” (SNI-L) based on 4-neighborhood or 8-neighborhood is usually applied in the semi-supervised learning process (Dópido et al., 2013; Wang et al., 2014). However, the spatial neighborhood information of the selected unlabeled samples (SNI-unL) is rarely used in the determination process of samples label. When the class of each pixel is known, all the pixels can be regarded as training samples and the label of unlabeled samples must be same with one of the 8-neighborhood pixels. When the amount of initial training samples is small, the label of unlabeled samples also should be same as one of the training samples’ appearing in the neighborhood, but the nearest training sample may not be the right label. So the final label should be judged by classifier and the information of near training samples is helpful for the classifier. If the label assigned by a classifier is same as that of the training samples’ appearing in the neighborhood, it could be chosen as right training samples.

At present, support vector machines (SVM) (Schölkopf and Smola, 2002), multinomial logistic regression (MLR) (Böhning, 1992), and ensemble classifiers (EC) (Du et al., 2012) are widely used. There is still a need to discuss the final label given only by one classifier. So, in order to improve the accuracy of sample selection, two classifiers multinomial logistic regression (MLR) and k-nearest neighbor (KNN) (Li et al., 2005) are combined together.

The remainder of this paper is organized as follows. Section II describes the proposed approach for semi-supervised self-learning. Section III gives the classification results of two real hyperspectral images collected by the Airborne Visible-Infrared Imaging Spectrometer (AVIRIS) (Green et al., 1998) and the Reflective Optics Spectrographic Imaging System (ROSIS) (Benediktsson...
et al., 2005). Finally, Section IV concludes the paper and outlines some of the plausible future research lines.

2. Proposed classification approach

First, we define the notations used in this paper. Let \( K = [1, \ldots, k] \) denote a vector of \( K \) class labels, \( S = \{1, \ldots, n\} \) is a set of integers indexing the \( n \) pixels of an image, \( X = [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n} \) denotes an image of \( d \)-dimensional feature vectors, \( y = (y_1, \ldots, y_n) \) is an image of labels. \( D \) is the search radius and \( L \) denotes the number of initial training samples per class.

2.1. Selecting the most informative unlabeled samples based on BT

In this paper, BT (Luo et al., 2004) algorithm is used to obtain the most informative unlabeled samples and construct a data set of unlabeled samples by the posterior density \( p(y|x) \). This greatly improves the operation rate and reduces the amounts of added unlabeled samples. The BT technique is focused on the diversity of the unlabeled samples. It obtains the minimum difference between the two highest probabilities from the class posterior density \( p(y|x) \). Many experiments have proved that BT can complete this task well (Li et al., 2011; Wang et al., 2014).

2.2. Labeling unlabeled samples based on Spatial Neighborhood Information (SNI)

If the class label of the unlabeled samples is only determined by the primary classification map, which can be used as a priori knowledge, it is very difficult to guarantee satisfied accuracy. “Tobler’s First Law of Geography” reveals that all attribute values on a geographic surface are related to each other, but closer values are more strongly related than are more distant ones. This law also gives us a revelation that the same kind of objects should be in the same region and the most likely label of the unlabeled samples should be the same as one of the existing categories of the training samples in the spatial neighborhood area. In this paper, SNI-unL is constructed to help the classifier label the selected unlabeled samples. This is the first innovation point. The detail steps are shown as following.

Step 1: Circle-neighborhood (CN): The amount of the initial samples per class is too small. We empirically find out that, the 4- or 8-neighborhood are too small for the searching of the useful samples, a larger searching area is needed for providing a better \( D \). In this paper, we select a circle neighborhood which takes the selected unlabeled samples as the center in the training set. It is convenient for circular neighborhood to adjust the size of search radius \( D \) to find the optimal one and can avoid the influence of different spatial distance in 4-neighborhood or 8-neighborhood.

Step 2: Spatial Neighborhood Information (SNI). The closer training samples which appeared in the circle-neighborhood of unlabeled samples were selected through two-dimensional (2-D) Euclidean distance. Then, it was made a statistic of the selected training samples in the circular neighborhood and structured a Possible Label Set (PLS). When the label of unlabeled samples given by classifier can be found in PLS,
it is considered that the label given by the classifier is correct. If the condition is not satisfied, this unlabeled sample is removed. Different sizes of images should be a correspondence with different sizes of neighborhood. Fig. 1 shows how the circle neighborhood is established and the spatial neighborhood information is adopted.

### 2.3. Classifier combination based on MLR and KNN

In each cycle process of semi-supervised learning, MLR, a probabilistic classifier developed by Böhning (1992), is used to model the class posterior density $p(y|x)$ (the probability of each pixel belonging to each class) which is used to select the most informative unlabeled samples through BT method and obtain the final classification map.

MLR classifier is shown as follows.

$$p(y_i = k|x_i, \omega) = \frac{\exp(\omega_h^T h(x_i))}{\sum_{k=1}^K \exp(\omega_k^T h(x_i))} \tag{1}$$

where $h(x) = [h_1(x), \ldots, h_K(x)]^T$ is a vector of $l$ fixed functions of the input, which are often termed features. It should be noted that function $h$ may be linear or nonlinear. To obtain better results, kernels have been widely used to improve the data divisibility in the transformed space. In this paper, we take a Gaussian radial basis function (RBF) kernel, which is widely used with hyperspectral images (Camps-Valls and Bruzzone, 2005). $\omega = [\omega_1^T, \ldots, \omega^K]^T$ is the logistic regressor, and $\omega_k = [\omega_k^T, \ldots, \omega_k^T]$ denotes the regression parameters of $l + 1$ characteristic values per class. We use LORSAL (Li et al., 2011) to get $\omega^K$.

Due to a small number of initial training samples per class, it is difficult to test whether the label of a certain unlabeled sample is correct or not through the classification map provided by the MLR classifier. The result derived from only one classifier sometimes make an error. MLR is a fitting classifier and cannot obtain good model parameters just with a small number of initial samples. In this paper, MLR classifier is combined with the classifier KNN (Li et al., 2005) to improve the performance of “classifier”.

SVM classifier usually cannot get well performance in the boundary position. It was pointed out that the combination of SVM and KNN can overcome this shortcomings and improve the performance of multi-class classification (Baassou et al., 2013; Yuan et al., 2008). It indicates that KNN can effectively classify with samples located at the category Borders. Hence, MLR has a requirement for boundary information and KNN can meet this demand. So we make a combination with MLR and KNN, and the experiments also achieve a well performance.

Fig. 2 shows the detailed fusion way of MLR and KNN and the deterministic process of final label for unlabeled samples to be determined. When two classifiers’ label from classification map appear in the possible label set (PLS) and they have the same labels or When only one classifier’s label appear in the possible set, we label this unlabeled sample through the way of part B. If the labels MLR and KNN given appear in the possible label set but the labels MLR and KNN given are different or two classifiers’ label cannot appear in the possible set, these conditions indicate that the combination classifier does not has the ability to determine the labeled of this unlabeled sample at present. Some uncertain unlabeled samples which are difficult to confirm were released back to keep the selected unlabeled samples sets high accuracy. In the next cycle, the unselected unlabeled labeled may be determined.

$$y = \begin{cases} y_{MLR}, & \text{if } (C_{MLR} = 1, C_{KNN} = 0) \\ y_{KNN}, & \text{if } (C_{MLR} = 0, C_{KNN} = 1) \\ \text{if } (C_{MLR} = 1, C_{KNN} = 1, y_{MLR} = y_{KNN}) \end{cases}, \tag{2}$$

where

- $C_{MLR} = 1$, if (Label$_{MLR}$ ∈ CN)
- $C_{MLR} = 0$, if (Label$_{MLR}$ # CN)
- $C_{KNN} = 1$, if (Label$_{KNN}$ ∈ CN)
- $C_{KNN} = 0$, if (Label$_{KNN}$ # CN)

Finally, in order to tackle the whole experiment process, Fig. 3 shows a detailed description of the whole proposed semi-supervised learning approach.

### 2. Experimental results

#### 2.1. Data sets

It is well known that a good algorithm should be suitable for different conditions. These two data sets can satisfy this requirement and have been widely used for testing the accuracy of hyperspectral data classification algorithms. The use of the AVIRIS data set is often faced with classification problems that arise from the existence of mixed pixels, the unbalanced number of available labeled pixels per class, and the inseparability just based on the spectrum. The use of the ROSIS data set is often confronted with some challenges in that the objects are unsystematically and
irregularly distributed in the map, and it is not easy to distinguish between the bare soil and meadow classes. The initial training samples were randomly selected from test data set per class. The final results were obtained by averaging the result after conducting 10 independent Monte Carlo (MC).

(1) The AVIRIS Indian Pines data set was collected by the AVIRIS sensor over the Indian Pines region in Northwestern Indiana in 1992. The size of this data set is 145 lines by 145 samples, with 224 spectral channels in the wavelength range of 0.4–2.5 \( \mu \text{m} \). Due to noise and water absorption, 22 radiance channels were removed from the data. Fig. 4(a) shows a false color image of the AVIRIS data set. Fig. 4(b) shows the ground-truth map with 16 mutually exclusive ground-truth classes, in a total of 10,366 samples.

(2) The ROSIS Pavia University hyperspectral data set was collected by the ROSIS optical sensor over the urban area of the University of Pavia, Italy. The size of this image is 610 lines by 340 samples, with 103 spectral channels in the wavelength range of 0.43–0.86 \( \mu \text{m} \). Fig. 5(a) shows a false color composite of the ROSIS scene, and Fig. 5(b) shows the test area with 9 mutually exclusive ground-truth classes, in a total of 42,776 samples.

2.2. Influence of the suitable search radius \( D \)

For the proposed algorithm, one of the key questions is how to confirm the suitable search radius \( D \), which influences the accuracy and the numbers of selected unlabeled samples. In this part, we...
carried out experiments to find the optimal parameter $D$ with AVIRIS Indian Pines and ROSIS Pavia University set data.

Fig. 6(a)–(c) shows the result of AVIRIS data. When the number of initial training samples per class is 5, 10, 15 and the search radius $d$ ranges from 1 to 6, the results corresponding $d = 4$ or $d = 5$ are acceptable; Fig. 6(d)–(f) shows the result of ROSIS data. When the number of initial training samples per class is 5, 10, 15 and the search radius $d$ ranges from 5 to 25, the results corresponding $d = 10$ is stable. It is proved that the influence of the suitable parameter $D$ mainly focus on the size of image data. When the size of image is known, the value of $D$ is stable.

2.3. Comparison experiments

In this part, three groups of comparison experiments are listed to explain various problems.

(1) Experiment-1: the influence of spatial neighborhood information. As shown in Fig. 7 that when the selected unlabeled samples were labeled just through the classification map of MLR, the accuracy decrease with the increase of iteration. The main reason was that a large number of the labels of unlabeled samples were wrong. When the SNI-unL is used for MLR, the performance improves. This experiment proves that the proposed SNI-unL is helpful for the classifier.

(2) Experiment-2: the influence of classifier combination. Table 1 shows the overall accuracy (OA) and Kappa coefficient statistics of the AVIRIS data set with the increase of search radius $D$ in step 1. The number of initial training samples per class $L$ are 5, 10 and 15. Three different cases are listed for comparison: the MLR, the KNN and the fusion of these two classifier. Some important rules are showed in Table 1: (1) the overall accuracy (OA) and Kappa are first increased and then decreased with the increase of the search radius $D$ in step 1, when the initial number of training samples per class $L$ is fixed; (2) the overall accuracy (OA) and Kappa toward higher with the increase of the initial number of training samples per class $L$, when the search radius $D$ is fixed; (3) the best classification result is when $D$ is 3 or 4 in all cases; (4) when MLR and KNN are fused together by
Table 2
Overall accuracy and kappa coefficient statistics using MLR and KNN applied to the ROSIS data set, with 5, 10 and 15 initial training samples per class. The best OA and Kappa results of each table were marked bold italics.

<table>
<thead>
<tr>
<th>ROSIS</th>
<th>OA (%)</th>
<th>Kappa (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MLR + SNI</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L = 5$</td>
<td>62.78</td>
<td>53.47</td>
</tr>
<tr>
<td>$L = 10$</td>
<td>64.24</td>
<td>56.58</td>
</tr>
<tr>
<td>$L = 15$</td>
<td>81.00</td>
<td>75.68</td>
</tr>
<tr>
<td><strong>KNN + SNI</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L = 5$</td>
<td>61.87</td>
<td>53.56</td>
</tr>
<tr>
<td>$L = 10$</td>
<td>78.83</td>
<td>72.26</td>
</tr>
<tr>
<td>$L = 15$</td>
<td>78.70</td>
<td>72.75</td>
</tr>
<tr>
<td><strong>MLR + KNN + SNI</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$L = 5$</td>
<td>66.01</td>
<td>56.71</td>
</tr>
<tr>
<td>$L = 10$</td>
<td>79.34</td>
<td>73.53</td>
</tr>
<tr>
<td>$L = 15$</td>
<td>79.12</td>
<td>73.44</td>
</tr>
</tbody>
</table>

Fig. 8. The classification map with three different ways using AVIRIS $D = 4$, $L$ is 5, 10 and 15.
<table>
<thead>
<tr>
<th>Method</th>
<th>Parameters</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLR+SNI</td>
<td>L=5, D=10</td>
<td>63.65%</td>
</tr>
<tr>
<td>KNN+SNI</td>
<td>L=5, D=10</td>
<td>69.69%</td>
</tr>
<tr>
<td>MLR+KNN+SNI</td>
<td>L=5, D=10</td>
<td>65.44%</td>
</tr>
<tr>
<td>MLR+SNI</td>
<td>L=10, D=10</td>
<td>82.89%</td>
</tr>
<tr>
<td>KNN+SNI</td>
<td>L=10, D=10</td>
<td>79.86%</td>
</tr>
<tr>
<td>MLR+KNN+SNI</td>
<td>L=10, D=10</td>
<td>84.91%</td>
</tr>
<tr>
<td>MLR+SNI</td>
<td>L=15, D=10</td>
<td>84.94%</td>
</tr>
<tr>
<td>KNN+SNI</td>
<td>L=15, D=10</td>
<td>81.73%</td>
</tr>
<tr>
<td>MLR+KNN+SNI</td>
<td>L=15, D=10</td>
<td>88.08%</td>
</tr>
</tbody>
</table>

**Fig. 9.** The classification map with three different ways using ROSIS (D = 10, L is 5, 10 and 15).
the proposed way, it can reach the best performance; (5) when MLR or KNN is connected with local location information through circle search area (CN), some significant results can also be achieved.

In order to express the differences among these three methods intuitively, Fig. 8 shows the best classification map in the same $D$ and different $L$ after 30 iteration times. $D = 4$ and $L$ is 5, 10 and 15.

Table 2 shows the overall accuracy (OA) and Kappa coefficient statistics of the ROSIS data set with the increase of search radius $D$ in step 5. The number of initial training samples per class $L$ are 5, 10 and 15. Three different cases are listed for comparison. Some important rules are showed in Table 2: (1) the overall accuracy (OA) and Kappa are first increased and then decreased with the increase of the search radius $D$ in step 5, when the initial number of training samples per class $L$ is fixed; (2) the overall accuracy (OA) and Kappa toward higher with the increase of the initial number of training samples per class $L$, when the search radius $D$ is fixed; (3) when MLR and KNN are fused together by the proposed way, it can get the best performance; (4) when MLR or KNN is connected with local location information through circle search area (CN), some significant results can also be achieved. These four rules are same with the AVIRIS data; (5) the best classification result is when $D = 15$ or 20 nearly.

In order to express the differences among these three methods intuitively, Fig. 9 shows the best classification map in the same $D$ and different $L$ after 20 iteration times. $D = 10$ and $L$ is 5, 10 and 15.

Fig. 10. The comparison results using SNI-L and SNI-unL.
When the time of iteration is toward 30, the OAs are 75%, 87% and 90%; Figs. 10(d)–(f) and 11(b) tell us that when the number of initial training samples per class are 5, 10, 15, the proposed SNI-unl can get better performance than SNI-L. The same conclusions can be got through these two data sets.

Fig. 11 shows the whole training sample numbers (labeled and unlabeled) of “MLR + KNN + SNI” way, which were obtained by averaging the result after conducting 10 independent Monte Carlo (MC) runs. We randomly selected the training samples from the whole image in each time. Figs. 10(a)–(c) and 11(a) tell us that when the number of initial training samples per class are 5, 10 and 15 for AVIRIS, the OA are toward 70%, 80% and 85% after 10 iterations. However, the number of whole samples are less than 1000. When the time of iteration is toward 30, the OAs are 75%, 87% and 90%; Figs. 10(d)–(f) and 11(b) tell us that when the number of initial training samples per class are 5, 10 and 15, the OA are toward 73%, 82% and 85% after ten iterations. However, the number of whole samples are less than 400. When the time of iteration is toward 30, the OAs are 78%, 87% and 89%.

3. Conclusion and further work

In this paper, we proposed a new approach based on spatial neighborhood information (SNI) and classifier combination for the semi-supervised classification of hyperspectral images. In the semi-supervised process, the improvement in the classifier function, the use of the BT algorithm, and the decision method for the label of unlabeled samples are all key points in this note. After selecting the most informative unlabeled samples by the use of breaking ties (BT) technology, our main task is the determination of the labels of new unlabeled samples. On one hand, we use the spatial neighborhood through circular search area to reduce the difficulty of judgments. On the other hand, multinomial logistic regression (MLR) and k-nearest neighbor (KNN) are combined together to improve the accuracy of judgment. Experimental results with two classical hyperspectral data sets indicate that the proposed approach can obtain well performance. We also explored the optimization of parameter D by searching experiments. It has been proved that the parameter D mainly depend on the size of image data. When the size of image is known, the value of D is stable. We can set a proper range of parameter D before running the searching steps. For the AVIRIS data, the better parameter D is about 5. For ROSIS data, the better parameter D is about 15. It should set the initial D as the size of the image. In our future work, the case that when the all labels two classifiers give appear in the search area but are inconsistent.

Acknowledgments

The authors would like to thank Professors David Landgrebe and Paolo Gamba for providing the data used in the experiments. This research is supported in part by Natural Science Foundation of China (No. 41471356), Fundamental Research Funds for the Central Universities (2014QNAA33, 2014ZDPY14), Priority Academic Program Development of Jiangsu Higher Education Institutions.

References
