remote sensing

MDPI

Article

A Novel Tri-Training Technique for Semi-Supervised Classification of Hyperspectral Images Based on Diversity Measurement

Kun Tan [1,†], Jishuai Zhu [1,†], Qian Du [2,*], Lixin Wu [1] and Peijun Du [3,*]

[1] Jiangsu Key laboratory of Resources and Environment Information Engineering, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China; tankun@cumt.edu.cn (K.T.); zhujishuai2012@126.com (J.Z.); awulixin@263.net (L.W.)

[2] Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS 39762, USA

[3] Key Laboratory for Satellite Mapping Technology and Applications of State Administration of Surveying, Mapping and Geoinformation of China, Nanjing University, Nanjing 210023, China

* Correspondence: du@ece.msstate.edu (Q.D.); dupjrs@126.com (P.D.); Tel./Fax: +86-0516-8359-1320 (P.D.)

† These authors contributed equally to this work.

Academic Editors: András Jung, Lenio Soares Galvao and Prasad S. Thenkabail
Received: 27 June 2016; Accepted: 4 September 2016; Published: 12 September 2016

Abstract: This paper introduces a novel semi-supervised tri-training classification algorithm based on diversity measurement for hyperspectral imagery. In this algorithm, three measures of diversity, i.e., double-fault measure, disagreement metric and correlation coefficient, are applied to select the optimal classifier combination from different classifiers, e.g., support vector machine (SVM), multinomial logistic regression (MLR), extreme learning machine (ELM) and k-nearest neighbor (KNN). Then, unlabeled samples are selected using an active learning (AL) method, and consistent results of any other two classifiers combined with a spatial neighborhood information extraction strategy are employed to predict their labels. Moreover, a multi-scale homogeneity (MSH) method is utilized to refine the classification result with the highest accuracy in the classifier combination, generating the final classification result. Experiments on three real hyperspectral data indicate that the proposed approach can effectively improve classification performance.

Keywords: classifier diversity; active learning; multi-scale homogeneity (MSH); hyperspectral imagery

1. Introduction

Conventional supervised classification algorithms (e.g., decision tree (DT) [1], naive Bayesian (NB) [2] and back propagation neural network (BPNN) [3]) can provide satisfying classification performance and have been widely used in traditional data classification, such as web page classification [4,5], medical image classification [6,7] and face recognition [8]. However, performance strongly depends on the quantity and quality of training samples. Labeled samples are often difficult, costly or time consuming to obtain, and they may not perform well on hyperspectral imagery due to the Hughes phenomenon when the number of training samples is limited [9,10]. Therefore, semi-supervised learning attempts to use unlabeled samples to improve classification [11–13]. Common semi-supervised learning algorithms include multi-view learning [14,15], self-learning [16,17], co-training [18,19], graph-based approaches [20,21], transductive support vector machines (TSVM) [22,23], etc.

Semi-supervised learning has been of great interest to hyperspectral remote sensing image analysis. In [24], semi-supervised probabilistic principal component analysis, semi-supervised local fisher discriminant analysis and semi-supervised dimensionality reduction with pairwise constraints were extended to extract features in a hyperspectral image. In [25], a new classification methodology based on spatial-spectral label propagation was proposed. Dopido and Li developed a new

framework for semi-supervised learning, which exploits active learning (AL) for unlabeled samples' selection [26]. In [27], a new semi-supervised algorithm combined spatial neighborhood information in determining class labels of selected unlabeled samples. Tan proposed a semi-supervised SVM with a segmentation-based ensemble algorithm to use spatial information extracted by a segmentation algorithm for unlabeled samples' selection in [28].

Meanwhile, Blum and Mitchell proposed a prominent approach called co-training, which has become popular in semi-supervised learning [19]. This algorithm requires two sufficient and redundant views, but this requirement cannot be met for hyperspectral imagery. Then, Gold and Zhou proposed a new co-training method called statistical co-training [29], which employed two different learning algorithms based on a single view. In [30], another new co-training method called democratic co-training was proposed. However, the aforementioned algorithms employ a time-consuming cross-validation technique to determine how to label the selected unlabeled samples and how to produce the final hypothesis. Therefore, Zhou and Li developed tri-training in [31]. It neither requires the instance space to be described with sufficient and redundant views nor imposes any constraints on supervised learning algorithms, and its applicability is broader than previous co-training style algorithms. However, tri-training has some drawbacks in three aspects: (1) selecting a complementary classifier may be difficult; (2) unlabeled samples may have error labels that are added to the training set during semi-supervised learning; (3) the final classification map may be contaminated by salt and pepper noise. In this paper, a novel tri-training algorithm is proposed. We use three measures of diversity, i.e., the double-fault measure, the disagreement metric and the correlation coefficient, to determine the optimal classifier combination, then unlabeled samples are selected using an active learning (AL) method and consistent results of any two classifiers combined with a spatial neighborhood information extraction strategy to predict the labels of unlabeled samples. Moreover, a multi-scale homogeneity (MSH) method is utilized to refine the classification result.

The remainder of this paper is organized as follows. Section 2 briefly introduces the standard tri-training algorithm, then describes the proposed approach. Section 3 presents experiments on three real hyperspectral datasets with a comparative study. Finally, Section 4 concludes the paper.

2. Methodology

2.1. Tri-Training

In the standard tri-training algorithm, three classifiers are initially trained by a dataset generated via bootstrap sampling from the original labeled data. Then, for any classifier, an unlabeled sample can be labeled as long as another two classifiers agree on the labeling of this sample. This training process will stop when the results of the three classifiers reach consistency. The final predication is produced with a variant of majority voting among all of the classifiers.

2.2. The Proposed Approach

2.2.1. Classifier Selection

The principle of classifier selection is that classifiers should be different from each other and their performance should be complementary; otherwise, the overall decision will not be better than each individual decision. Three measures of diversity are implemented to select three classifiers from SVM [32–34], multinomial logistic regression (MLR) [35,36], KNN [27,37] and extreme learning machine (ELM) [38,39]. The three measures of diversity are the double-fault measure, the disagreement metric and the correlation coefficient [40], which are described as below.

(1) The correlation coefficient (ρ):

Let $Z = [z_1, \ldots, z_n]$ be a labeled dataset, K be the number of classifiers, D_i, {i = 1...K} be the classifier and $y_i = [y_{1i}, \ldots, y_{ni}]$ be the output of D_i. If D_i recognizes correctly z_o, $y_{oi} = 1$, otherwise, $y_{oi} = 0$.

$$\rho = \frac{2}{K \times (K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \frac{N_{ij}^{11} \times N_{ij}^{00} - N_{ij}^{01} \times N_{ij}^{10}}{\sqrt{(N_{ij}^{11} + N_{ij}^{10}) \times (N_{ij}^{01} + N_{ij}^{00}) \times (N_{ij}^{11} + N_{ij}^{01}) \times (N_{ij}^{10} + N_{ij}^{00})}} \quad (1)$$

where N_{ij}^{ab} is the number of samples z_o of Z for which $y_{oi} = a$ and $y_{oj} = b$ (see Table 1). With the increase of ρ, the diversity of classifiers becomes smaller.

Table 1. The relationship between a pair of classifiers.

	D_j Correct (1)	D_j Wrong (0)
D_i correct (1)	N_{ij}^{11}	N_{ij}^{10}
D_i wrong (0)	N_{ij}^{01}	N_{ij}^{00}

(2) Disagreement metric (D):

The disagreement between classifier outputs (correct/wrong) can be measured as:

$$D = \frac{2}{K \times (K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \frac{N_{ij}^{01} + N_{ij}^{10}}{N_{ij}^{11} \times N_{ij}^{00} + N_{ij}^{01} \times N_{ij}^{10}} \quad (2)$$

where N_{ij}^{ab} is the number of samples z_o of Z for which $y_{oi} = a$ and $y_{oj} = b$ (see Table 1). With the increase of D, the diversity of classifiers becomes larger.

(3) Double-fault measure (DF):

The double-fault between classifier outputs (correct/wrong) can be measured as:

$$\mathrm{DF} = \frac{2}{K \times (K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \frac{N_{ij}^{00}}{N_{ij}^{11} + N_{ij}^{00} + N_{ij}^{01} + N_{ij}^{10}} \quad (3)$$

where N_{ij}^{ab} is the number of samples z_o of Z for which $y_{oi} = a$ and $y_{oj} = b$ (see Table 1). With the increase of DF, the diversity of classifiers becomes larger.

2.2.2. Unlabeled Sample Selection

In the standard tri-training algorithm, for any classifier, an unlabeled sample can be labeled when another two classifiers agree on the labeling of this sample. However, the training set may be small; the label of unlabeled samples that two classifiers agree on may be wrong. Therefore, for any classifier, we use a spatial neighborhood information extraction strategy with an AL algorithm to select the most useful spatial neighbors as the new training set on the condition that two classifiers agree on the labeling of these samples.

Figure 1 illustrates how to select unlabeled samples, and the selection process includes two key steps, i.e., the construction of the candidate set and active learning.

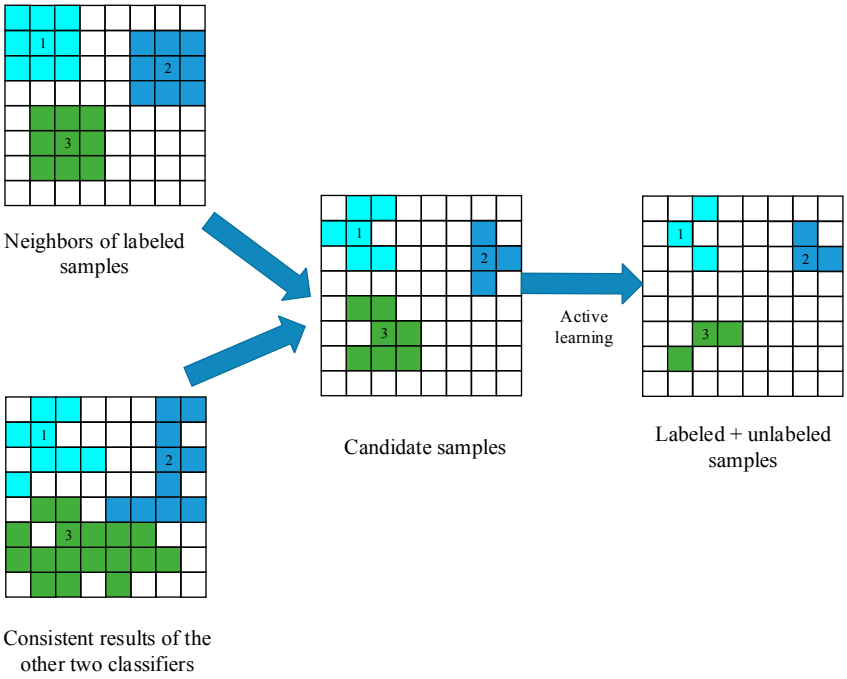

Figure 1. The process of selecting unlabeled samples.

(1) The construction of the candidate set:

For any classifier, we consider spatial neighborhood information with the consistent results of two classifiers to build the candidate set. Firstly, unlabeled samples are selected based on the consistency of two classifiers' outputs, and those samples are considered reliable according to the standard tri-training algorithm. With a local similarity assumption, the neighbors of labeled training samples are identified using a second-order spatial connectivity, and the candidate set is built by analyzing the spectral similarity of these spatial neighbors. Since the output of a classifier is based on spectral information, the candidate set is obtained based on spectral and spatial information. Thus, these samples are more reliable.

(2) Active learning:

In semi-supervised learning, the main objective is to select the most useful and informative samples from the candidate set. However, some of the samples in the candidate set may not be useful for training the third classifier, because they may be too similar to the labeled samples. To prevent the introduction of such redundant information, the breaking ties (BT) [17] algorithm is adopted to select the most informative samples.

The decision criterion of BT is:

$$x'^{BT}_m = \arg\min\{\max_{k\in K} p(y_i = k|x'_m) - \max_{k\in K\setminus\{k^+\}} p(y_m = k|x'_m)\} \quad (4)$$

where $k^+ = \arg\max_{k\in K} p(y_m = k|x'_m)$ is the most probable class for sample x'_m, $p(y_m = k|x'_m)$ is the probability when the label of sample x'_m is k and K is the number of classes.

2.2.3. Multi-Scale Homogeneity Method

Some of the existing hyperspectral image classification algorithms produce classification results with salt and pepper noise. To solve this problem, we use the multi-scale homogeneity method. Let S be the initial classification result, $\alpha, \beta, \gamma (\alpha < \beta < \gamma)$ be the scale of a homogeneous region, $\theta_i (i = 1, 2, 3)$

be the threshold of those homogeneous regions and ρ be the number of the samples that have the same label in a homogeneous region.

(1) An $\alpha \times \alpha$ homogeneous region is built in the initial classification result. If $\rho \geq \theta_1$, the samples in this region will have the same label; otherwise, the label of the samples does not change. Let this new result be the second classification result.
(2) A $\beta \times \beta$ homogeneous region is built in the second classification result. If $\rho \geq \theta_2$, the samples in this region will have the same label; otherwise, the label of the samples does not change. Let this new result be the third classification result.
(3) A $\gamma \times \gamma$ homogeneous region is built in the third classification result. If $\rho \geq \theta_3$, the samples in the homogeneity region will have the same label; otherwise, the label of the samples does not change. This new result will be the final classification result.

2.3. Semi-Supervised Classification Framework

Let $L = [(y_m, x_m), x_m \in R^d, m = 1, 2, \cdots, n]$ be the initial training set, $U = [x_1', x_2', \cdots, x_u']$ be the unlabeled set, $D_i (i = 1, 2, 3)$ be the classifiers and $S_i (i = 1, 2, 3)$ be the classification results.

The procedure of the proposed method is summarized as follows.

(1) Train the classifier D_i with L and obtain the predicted classification result S_i;
(2) For the classifier D_i, select another two classifiers agreeing on the labeling of these samples to build the first candidate set;
(3) For $x_m \in L$, the neighbors of x_m (using second-order spatial connectivity) will be labeled based on Tobler's first law, and build the second candidate set;
(4) Conduct comparative analysis of the first and the second candidate set, and select these samples that have the same label to build the third candidate set;
(5) Use the BT method to select the most useful and information samples L' from the third candidate set, $L = L \cup L'$, $U = U - L'$;
(6) Train the classifier D_i with the new L and obtain the predicted classification result S_i;
(7) Terminate if the final condition is met; otherwise, go to Step (2);
(8) Obtain S_i that has the highest classification accuracy in these three classifiers and use the multi-scale homogeneity method to process S_i to obtain the final classification result.

3. Experiments

3.1. Data Used in the Experiments

In this study, three real hyperspectral images are used to evaluate the proposed approach.

(1) The first hyperspectral image was collected by the AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) sensor over the Indian Pines region in Northwestern Indiana in 1992. This datum has a spatial size of 145 × 145 pixels. It comprises 224 spectral channels in the wave-length range from 0.4 to 2.5 um at 10-nm intervals with a spatial resolution of 20 m, and 202 channels were used in the experiment after noise and water absorption bands were removed. For illustrative purposes, the image scene in pseudocolor is shown in Figure 2a. The ground truth map available for the scene with 16 mutually-exclusive ground-truth classes is shown in Figure 2b.
(2) The second hyperspectral image was collected by the ROSIS (Reflective Optics System Imaging Spectrometer) sensor over the urban area of the University of Pavia, Italy. This datum has a spatial size of 610 × 340 pixels. It comprises 115 spectral channels in the wave-length range from 0.43 to 0.68 um with a spatial resolution of 1.3 m, and 103 channels were used in the experiment after noise and water absorption bands were removed. For illustrative purposes, the image scene in pseudocolor is shown in Figure 3a. The ground truth map available for the scene with 9 mutually-exclusive ground-truth classes is showed in Figure 3b [41].

(3) The third hyperspectral image was collected by the AVIRIS sensor over Salinas Valley, Southern California, in 1998. This datum has a spatial size of 512 × 217 pixels. It comprises 224 spectral channels in the wave-length range from 0.4 to 2.5 um with a spatial resolution of 3.7 m, and 204 channels were used in the experiment after noisy and water absorption bands were removed. For illustrative purposes, the image scene in pseudocolor is shown in Figure 4a. The ground truth map available for the scene with 16 mutually-exclusive ground-truth classes is shown in Figure 4b.

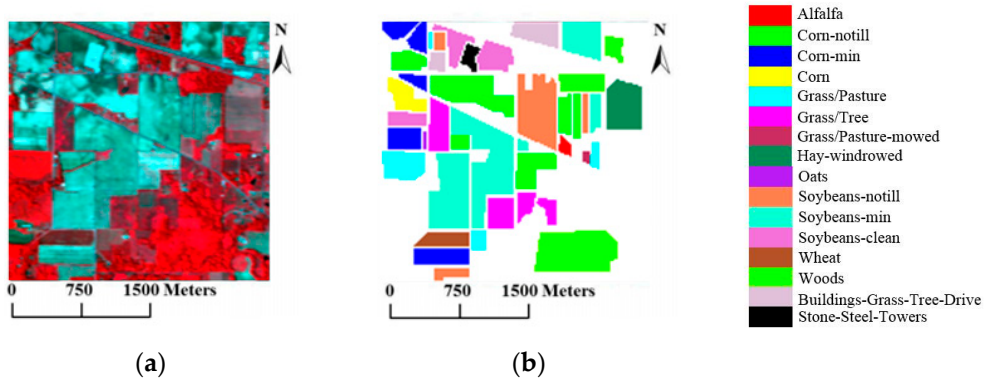

Figure 2. (**a**) Pseudocolor color composite of the AVIRIS Indian Pines data set; (**b**) the map with 16 mutually-exclusive ground-truth classes.

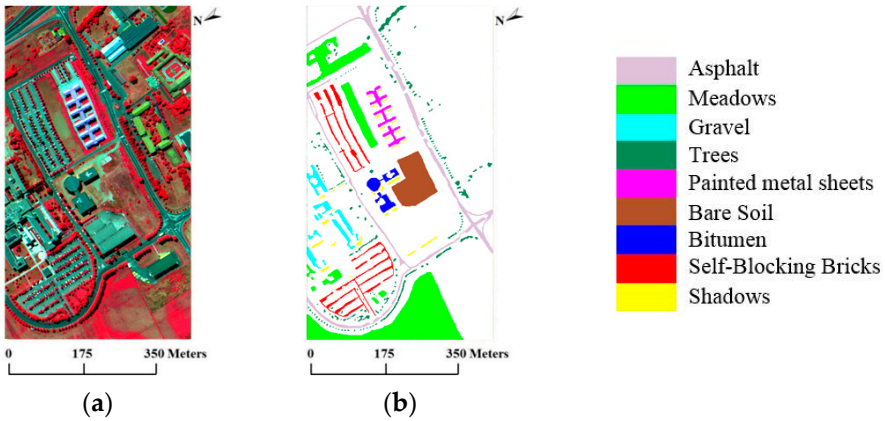

Figure 3. (**a**) Pseudocolor color composite of the Reflective Optics System Imaging Spectrometer (ROSIS) Pavia University scene; (**b**) the test area with 9 mutually-exclusive ground-truth classes.

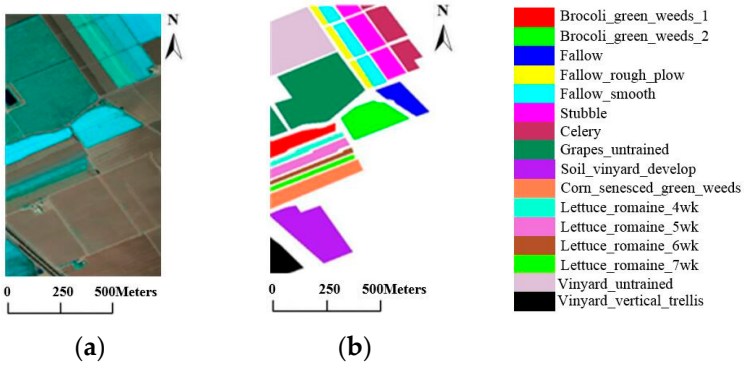

Figure 4. (**a**) Pseudocolor color composite of the AVIRIS Salinas Valley scene; (**b**) the test area with16 mutually-exclusive ground-truth classes.

3.2. Parameter Setting

In our experiments, the involved parameters were set as follows.

(1) Classifier parameter: k = 3 for KNN; the number of hidden neurons is 50; and the activation function is 'sigmoid' in the ELM; the parameter of MLR uses the default value.
(2) Multi-scale homogeneity: $\alpha = 2$, $\beta = 3$, $\gamma = 4$, $\theta_1 = 3$, $\theta_2 = 5$, $\theta_3 = 9$. The parameter α is set to follow Tobler's first law, and the parameter β is set through many experiments to ascertain the optimum value.
(3) Training sets: L = 5, 10, 15. We select 5, 10 and 15 samples per class as the initial labeled training sets.
(4) Other sets: The number of the most useful and informative samples L' in one iteration is 100. All experiments are carried out 10 times, and the averaged results are reported.

It is noteworthy that TT_AL_MSH denotes the proposed approach, and TT is the standard tri-training methods. Additionally, the performance of those approaches is objectively evaluated in terms of global accuracy (GA), which includes the overall accuracy (OA), average accuracy (AA) and the kappa coefficient (kappa). SVM and MLR have been widely used for hyperspectral image classification. ELM is a recently-developed simple and fast neural network classifier, and KNN is the traditional classifier whose kernel algorithm is the distance operation. The formation mechanisms of those classifiers are different. Therefore, we choose four base classifiers from a classifier pool, which are SVM(1), MLR(2), KNN(3) and ELM(4). In addition, three measures are used to compute their diversity (as shown in Table 2) by using the AVIRIS Indian Pines dataset. From Table 2, the same combination is selected by the D and ρ diversity measures, which contain MLR, KNN and ELM. The combination of SVM, KNN and ELM is selected by DF. In order to select the optimal combination, we selected the TT algorithm to test the performance of different classifier combination. As shown in Table 3, the combination of MLR, KNN and ELM is the optimal one.

Table 2. The diversity value (in terms of D, DF and ρ). The greatest diversity is marked in bold italics.

Classifiers Combination	*D*	DF	ρ
1,2,3	0.1745	0.1133	0.4729
1,2,4	0.1873	0.1296	0.4999
1,3,4	0.2160	**0.1495**	0.4548
2,3,4	**0.2275**	0.1311	**0.4170**

Table 3. The optimal combination selected by the diversity measures and tri-training (TT) (overall accuracy).

Classifiers Combination	AVIRIS Indian Pines			ROSIS Pavia University		
	5	10	15	5	10	15
1,3,4	58.34%	**65.29%**	69.76%	66.47%	71.32%	75.89%
2,3,4	**60.46%**	64.89%	**71.42%**	**66.86%**	**75.77%**	**78.82%**

For two methods to be compared, let f_{11} denote the number of samples that both methods can correctly classify, f_{22} the number of samples that both cannot, f_{12} the number of samples misclassified by Method 1, but not Method 2, and f_{21} the number of samples misclassified by Method 2, but not Method 1 [42]. Then, the decision criterion of McNemar's test statistic is:

$$Z = \frac{f_{12} - f_{21}}{\sqrt{f_{12} + f_{21}}} \tag{5}$$

For a 5% level of significance, the corresponding |z| value is 1.96; a |z| value greater than this quantity means that two methods have significant performance discrepancy.

Table 4 shows that the significance level of TT_MKE (i.e., MKE is the combination of MLR, KNN and ELM) compares against TT_AL_MSH_MKE, with 5, 10 and 15 initial training samples per class. Obviously, the performance of the proposed TT_AL_MSH_MKE is statistically different from TT_MKE.

Table 4. The value of Z-test in the different dataset. AL, active learning.

TT_MKE		TT_AL_MSH_MKE	
		|Z|	**Significant?**
Salinas	5 samples	36.28	Yes
	10 samples	31.28	Yes
	15 samples	44.39	Yes
Indian Pine	5 samples	26.27	Yes
	10 samples	36.82	Yes
	15 samples	32.45	Yes
Pavia university	5 samples	64.55	Yes
	10 samples	57.05	Yes
	15 samples	53.51	Yes

3.3. Experiment on the Indian Pine Dataset

Table 5 shows the OA statistical results of TT_AL_MSH_MKE, TT_AL_MSH_SKE (i.e., SKE is the combination of SVM, KNN and ELM), TT_MKE and TT_SKE. It can be obviously seen that the proposed TT_AL_MSH_MKE produces higher classification accuracy than the standard TT_MKE. With 5, 10 and 15 initial training samples per class, the OA of TT_AL_MSH_MKE increases by 17.09%, 20.14% and 17.09%, respectively, compared with TT_MKE. Figure 5 shows that the OA greatly increases with the number of unlabeled samples. When the number of unlabeled samples reaches 700, the OA becomes stable. For illustrative purposes, the classification maps of AVRIS data are provided in Figure 6. Observed from these maps, the proposed methods can effectively reduce the salt and pepper noise.

Table 5. Overall accuracy using two different techniques for the AVIRIS Indian Pines data, with 5, 10 and 15 initial training samples per class. The best OA results of each table are marked in bold italics.

Iteration Method			**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
TT_AL_MSH_MKE	$L = 5$	OA (%)	51.66	62.93	68.91	71.42	73.42	74.95	75.81	76.65	77.19	***77.55***
		Kappa (%)	47.57	59.20	65.51	68.23	70.35	72.00	72.93	73.83	74.42	74.79
		AA (%)	65.79	74.59	78.48	80.81	82.35	83.52	84.24	84.78	85.09	85.16
	$L = 10$	OA (%)	62.26	71.51	77.32	80.16	81.50	82.76	83.81	84.42	84.86	***85.03***
		Kappa (%)	58.74	68.51	74.72	77.83	79.30	80.68	81.81	82.48	82.96	83.14
		AA (%)	74.36	80.85	84.78	86.93	88.29	89.18	89.70	89.92	90.08	90.39
	$L = 15$	OA (%)	70.08	77.82	81.61	83.36	84.48	85.98	86.77	87.68	88.06	***88.51***
		Kappa (%)	67.05	75.29	79.41	81.32	82.56	84.20	85.08	86.09	86.52	87.02
		AA (%)	80.58	85.72	88.25	89.65	90.11	90.90	91.57	92.03	92.38	92.58
TT_MKE	$L = 5$	OA (%)	50.29	50.47	55.69	57.66	58.75	59.61	60.08	60.19	60.25	60.46
		Kappa (%)	45.79	45.98	51.48	53.55	54.68	55.57	56.05	56.18	56.24	56.48
		AA (%)	62.68	62.79	67.38	68.67	69.69	70.41	70.74	71.02	71.12	71.36
	$L = 10$	OA (%)	56.00	57.90	59.13	60.74	61.19	62.51	63.67	64.44	64.78	64.89
		Kappa (%)	51.54	53.62	55.04	56.39	56.48	58.06	59.48	59.76	60.18	60.53
		AA (%)	67.35	68.84	69.79	70.69	70.99	72.37	73.23	73.91	74.12	74.21
	$L = 15$	OA (%)	62.98	63.63	64.38	65.17	66.65	68.00	69.74	70.05	71.20	71.42
		Kappa (%)	59.15	59.79	60.61	61.48	62.93	64.36	66.30	66.63	67.85	68.09
		AA (%)	74.62	75.07	75.60	75.91	76.51	77.26	78.92	78.87	79.48	80.25

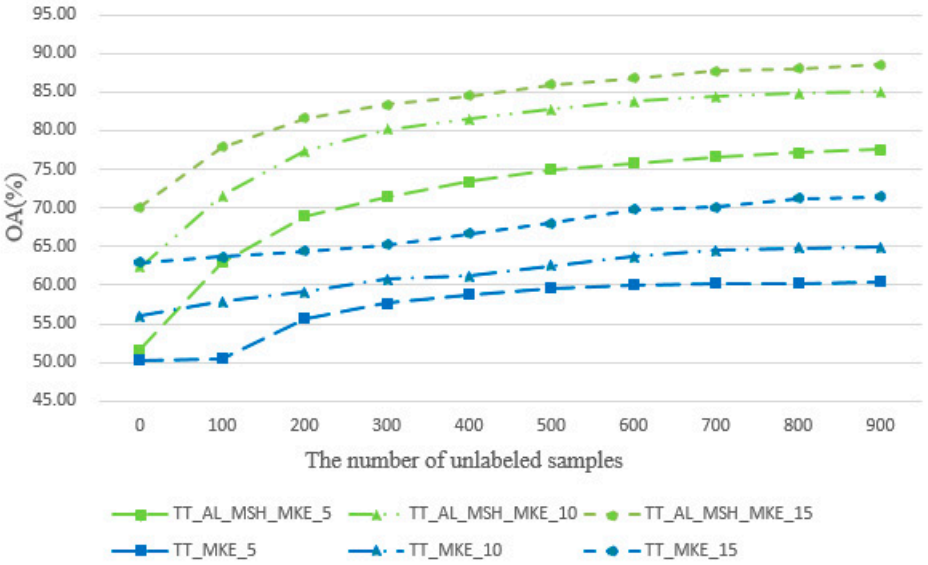

Figure 5. Overall classification accuracies obtained for the AVIRIS Indian Pines dataset using two different techniques by using 5, 10 and 15 labeled samples per class (estimated labels of unlabeled samples were used in all of the experiments).

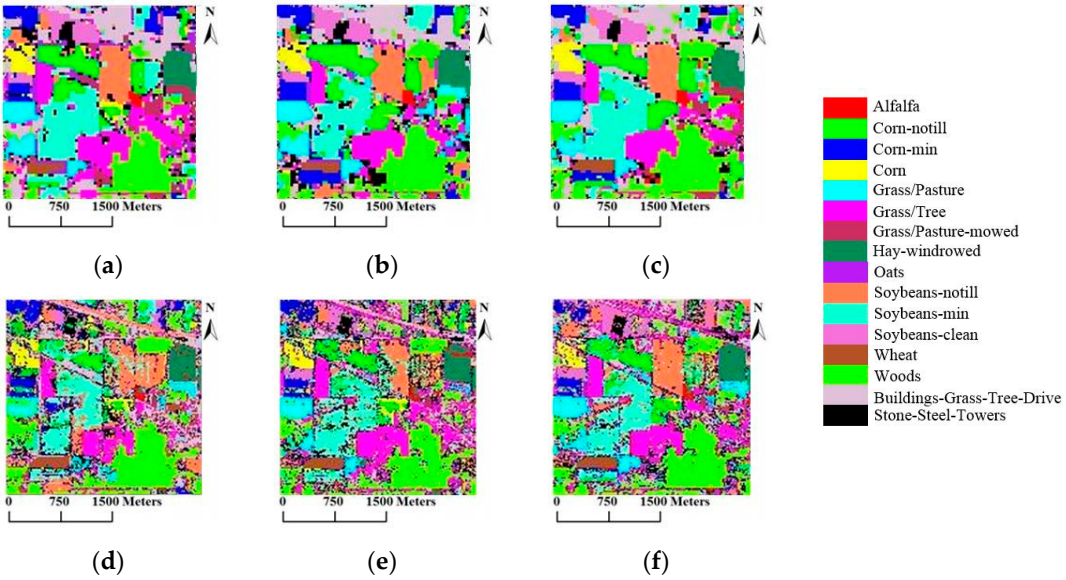

Figure 6. Classification maps for all of the methods with the AVIRIS Indian Pines dataset using 5, 10 and 15 labeled samples per class. (**a**) TT_AL_MSH_MKE (L = 5); (**b**) TT_AL_MSH_MKE (L = 10); (**c**) TT_AL_MSH_MKE (L = 15); (**d**) TT_MKE (L = 5); (**e**) TT_MKE (L = 10); (**f**) TT_MKE (L = 15).

3.4. Experiment on the University of Pavia Dataset

Table 6 shows the OA of TT_AL_MSH_MKE and TT_MKE. The proposed TT_AL_MSH_MKE can produce higher accuracy than the standard TT_MKE. With 5, 10 and 15 initial training samples per class, the OA of TT_AL_MSH_MKE increases by 15.69%, 12.84% and 13.31%, respectively, compared with TT_MKE. Figure 7 shows that the OA greatly increases with the number of unlabeled samples, and the performance of TT_AL_MSH_MKE is obviously superior to the performance of TT_MKE. However, the performance of TT_MKE is not stable, which is because unlabeled samples that are mislabeled are introduced into the training process. The classification maps of ROSIS Pavia University data are shown in Figure 8, where the proposed methods can produce smoother maps.

Table 6. Overall accuracy using two different techniques for ROSIS Pavia University data, with 5, 10 and 15 initial training samples per class. The best OA results of each table are marked in bold italics.

Iteration Method			1	2	3	4	5	6	7	8	9	10
TT_AL_MSH_MKE	$L = 5$	OA (%)	71.02	73.99	76.74	78.56	80.27	81.33	81.49	82.00	82.42	***82.55***
		Kappa (%)	63.31	67.45	71.03	73.28	75.26	76.52	76.74	77.36	77.85	78.01
		AA (%)	77.05	81.56	85.07	86.58	87.40	88.14	88.26	88.54	88.57	88.69
	$L = 10$	OA (%)	78.79	82.83	85.44	86.32	86.69	87.66	87.65	87.83	88.18	***88.61***
		Kappa (%)	72.85	78.05	81.36	82.46	82.92	84.12	84.15	84.39	84.84	85.35
		AA (%)	83.01	87.30	89.49	90.34	90.51	91.03	91.28	91.41	91.72	91.89
	$L = 15$	OA (%)	84.26	87.69	89.47	90.40	90.85	91.33	91.85	91.95	***92.11***	92.04
		Kappa (%)	79.70	84.01	86.27	87.47	88.07	88.70	89.36	89.49	89.70	89.61
		AA (%)	87.56	90.35	91.64	92.28	92.66	93.11	93.43	93.45	93.55	93.56
TT_MKE	$L = 5$	OA (%)	63.86	62.67	64.81	65.26	66.38	66.78	66.01	65.80	65.76	66.86
		Kappa (%)	54.76	53.52	55.84	56.24	57.41	57.86	57.13	57.07	56.97	58.32
		AA (%)	70.94	70.77	71.95	71.92	72.54	73.09	73.24	73.79	73.77	74.84
	$L = 10$	OA (%)	72.12	73.10	73.56	74.23	74.45	74.26	75.50	75.11	75.77	75.21
		Kappa (%)	64.55	65.72	66.22	66.90	67.22	67.07	68.53	68.07	68.82	68.27
		AA (%)	78.37	79.03	79.21	79.33	79.86	79.97	80.57	80.45	80.56	80.67
	$L = 15$	OA (%)	76.98	77.85	76.56	78.09	77.08	78.22	77.89	78.01	78.16	78.82
		Kappa (%)	70.33	71.21	69.93	71.73	70.57	71.77	71.50	71.58	71.73	72.51
		AA (%)	81.05	80.93	81.59	82.14	81.92	82.28	82.60	82.32	82.38	82.41

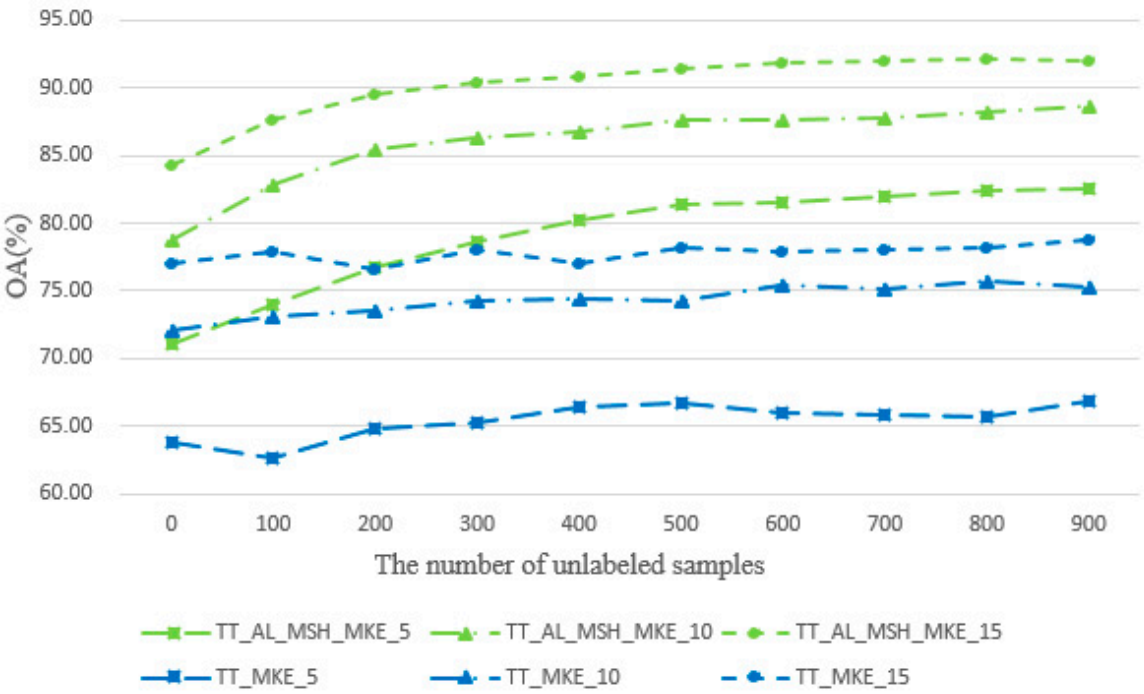

Figure 7. Overall classification accuracies obtained for the ROSIS Pavia University dataset using two different techniques by using 5, 10 and 15 labeled samples per class (estimated labels of unlabeled samples were used in all of the experiments).

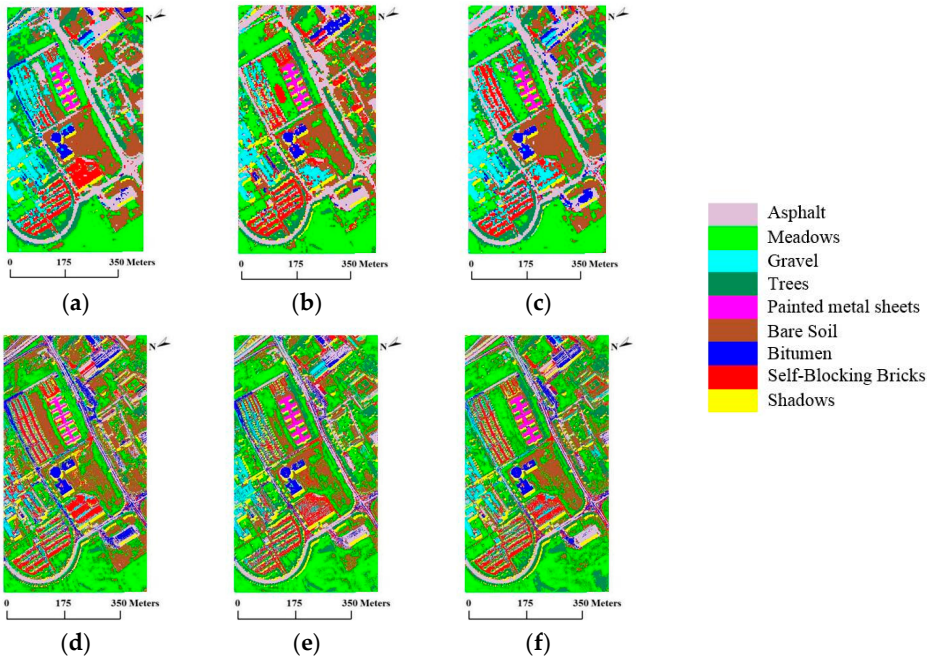

Figure 8. Classification maps for all of the methods with the ROSIS Pavia University dataset using 5, 10 and 15 labeled samples per class. (**a**) TT_AL_MSH_MKE (L = 5); (**b**) TT_AL_MSH_MKE (L = 10); (**c**) TT_AL_MSH_MKE (L = 15); (**d**) TT_MKE (L = 5); (**e**) TT_MKE (L = 10); (**f**) TT_MKE (L = 15).

3.5. Experiment on the Salinas Valley Dataset

Table 7 shows the OA of TT_AL_MSH_MKE and TT_MKE. The proposed TT_AL_MSH_MKE can produce higher accuracy than the standard TT_MKE. With 5, 10 and 15 initial training samples per class, the OA of TT_AL_MSH_MKE increases by 7.24%, 6.04% and 6.68%, respectively, compared with TT_MKE. Figure 9 shows that the OA greatly increases with the number of unlabeled samples, and the performance of TT_AL_MSH_MKE is obviously superior to the performance of TT_MKE. However, the performance of TT_MKE is not stable when the initial training samples per class is 5, which is because unlabeled samples that are mislabeled are introduced into the training process. The classification maps of Salinas data are shown in Figure 10, where the proposed methods can produce smoother maps.

Table 7. Overall accuracy using two different techniques for AVIRIS Salinas Valley data, with 5, 10 and 15 initial training samples per class. The best OA results of each table are marked in bold italics.

Iteration Method			1	2	3	4	5	6	7	8	9	10
TT_AL_MSH_MKE	L = 5	OA (%)	83.87	88.49	89.45	89.79	89.76	90.13	90.25	90.32	90.49	***90.68***
		Kappa (%)	82.09	87.22	88.29	88.66	88.63	89.04	89.17	89.25	89.43	89.65
		AA (%)	90.87	93.54	93.79	94.15	94.16	94.37	94.39	94.41	94.50	94.63
	L = 10	OA (%)	87.10	89.81	90.90	90.91	91.12	91.24	91.35	91.50	91.59	***91.64***
		Kappa (%)	85.66	88.68	89.88	89.89	90.13	90.26	90.39	90.56	90.65	90.71
		AA (%)	92.77	94.27	94.92	94.90	95.04	95.06	95.18	95.24	95.26	95.27
	L = 15	OA (%)	90.03	91.21	92.13	92.42	92.59	92.82	92.93	92.99	93.10	***93.17***
		Kappa (%)	88.92	90.23	91.25	91.57	91.75	92.01	92.13	92.21	92.33	92.40
		AA (%)	94.59	95.06	95.52	95.68	95.80	95.94	95.96	95.97	96.04	96.09
TT_MKE	L = 5	OA (%)	81.60	81.50	82.28	82.51	82.01	83.11	82.87	82.76	83.14	***83.44***
		Kappa (%)	79.57	79.45	80.32	80.58	80.05	81.26	80.99	80.88	81.28	81.61
		AA (%)	88.32	88.15	88.79	89.06	89.19	89.70	89.58	89.70	89.77	89.94
	L = 10	OA (%)	83.77	83.79	84.30	84.50	85.34	85.54	85.58	***85.64***	85.63	85.60
		Kappa (%)	82.01	82.03	82.61	82.83	83.74	83.96	84.01	84.07	84.07	84.04
		AA (%)	91.02	90.72	91.30	91.43	91.62	91.98	92.14	92.04	92.11	92.30
	L = 15	OA (%)	84.64	85.03	85.82	86.11	86.05	86.07	86.15	86.02	86.09	***86.49***
		Kappa (%)	83.00	83.42	84.30	84.60	84.55	84.57	84.67	84.52	84.61	85.04
		AA (%)	92.30	92.34	92.90	92.86	93.01	93.00	93.13	93.14	93.19	93.33

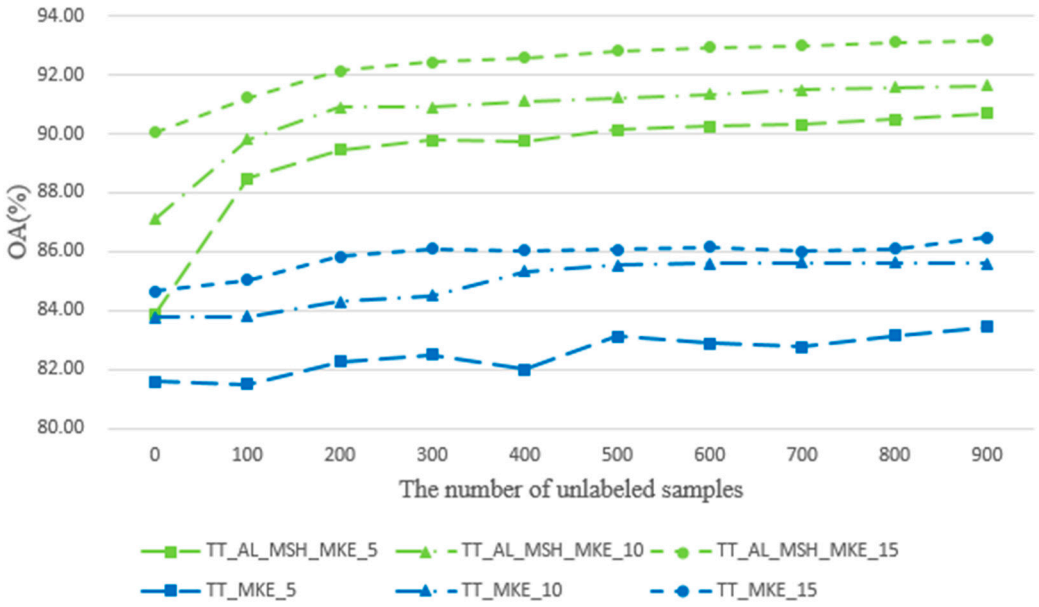

Figure 9. Overall classification accuracies obtained for the AVIRIS Salinas Valley dataset using two different techniques by using 5, 10 and 15 labeled samples per class (estimated labels of unlabeled samples were used in all the experiments).

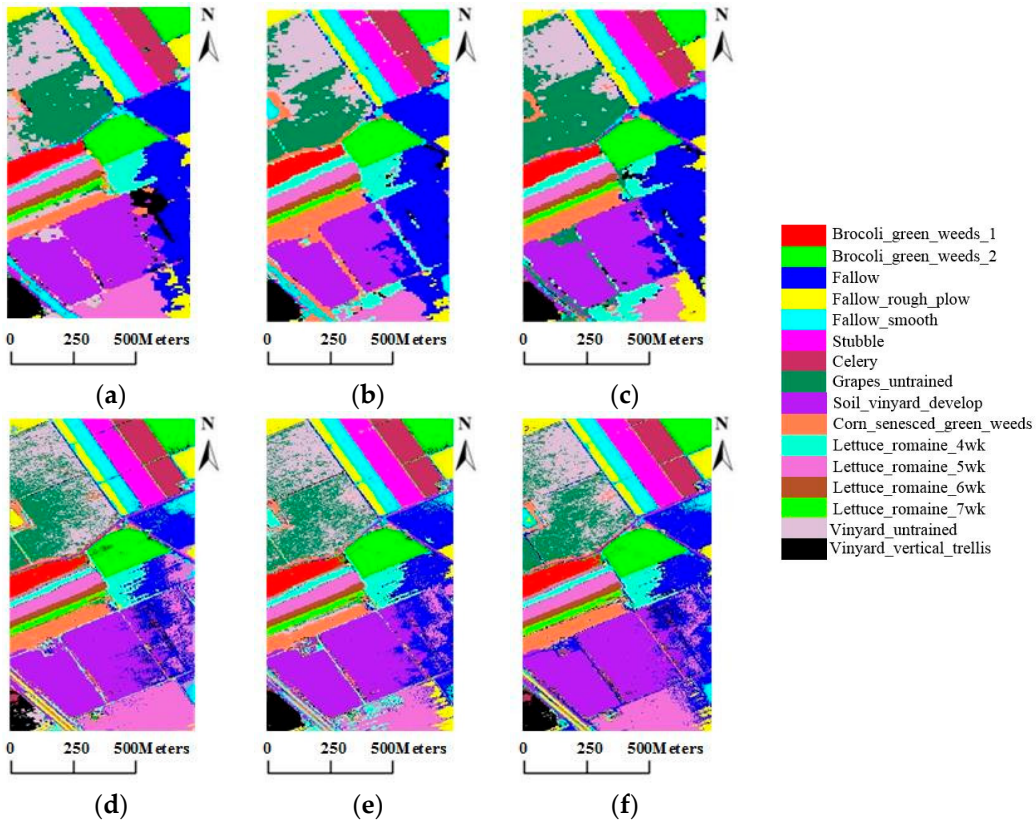

Figure 10. Classification maps for all of the methods with the AVIRIS Salinas Valley dataset using 5, 10 and 15 labeled samples per class. (**a**) TT_AL_MSH_MKE (L = 5); (**b**) TT_AL_MSH_MKE (L = 10); (**c**) TT_AL_MSH_MKE (L = 15); (**d**) TT_MKE (L = 5); (**e**) TT_MKE (L = 10); (**f**) TT_MKE (L = 15).

4. Discussion

In order to further assess the performance of the proposed method, we select some methods that combine semi-supervised spectral-spatial classification with active learning for comparison in this section. Reference results were provided in [25] for the spatial-spectral label propagation based on the support vector machine (SS-LPSVM), the transductive SVM, MLR + AL proposed in [25]. Additionally, the best reported accuracy from [27] for the MLR + KNN + SNI (i.e., SNI is the spatial neighbor information) method and from [43] for the semi-supervised classification algorithm based on spatial-spectral cluster (C-S2C) and the semi-supervised classification algorithm based on spectral cluster (SC-SC) is shown. Tables 8 and 9 illustrate the classification overall accuracy of TT_AL_MSH_MKE in comparison with the above methods for the AVIRIS Indian Pines dataset and ROSIS Pavia University dataset. With the number of initial labeled samples increasing, the OA values of all methods are increased. When $L = 5$, the best OA is obtained by TT_AL_MSH_MKE. For the AVIRIS Indian Pines dataset, the OA of TT_AL_MSH_MKE is 6.56% higher than MLR + KNN + SNI, respectively. For the ROSIS Pavia University dataset, the OA of TT_AL_MSH_MKE is 6.09%, 3.14% and 4.03% higher than MLR + KNN + SNI, respectively. The reason is that we select classifiers that are different from each other; their performances are complementary; and the classification performances are improved greatly, in particular for the small training datasets with 10 initial samples/class or less.

Table 8. Comparison of the methods, denoted as TT_AL_MSH_MKE, with the results reported in (1) [43], (2) [16], (3) [25] and (4) [27], for the AVIRIS Indian Pines dataset. The best OA results of each table are marked in bold italics. SC-SC, semi-supervised classification algorithm based on spectral cluster; TSVM, transductive support vector machine; SS-LPSVM, spatial-spectral label propagation based on the support vector machine.

Method	Training Samples		
	5	10	15
(1) SC-SC	68.79	72.84	73.11
(1) SC-S2C	68.32	75.43	77.63
(2) MLR + AL	75.00 ± 1.28	80.04 ± 1.28	81.00 ± 1.28
(3) TSVM	62.57 ± 0.23	63.45 ± 0.17	65.42 ± 0.02
(3) SS-LPSVM	69.60 ± 2.30	75.88 ± 0.22	80.67 ± 1.21
(4) MLR + KNN + SNI	70.99	**86.01**	**90.44**
TT_AL_MSH_MKE	**77.55**	85.03	88.51

Table 9. Comparison of methods, denoted as TT_AL_MSH_MKE, with results reported in (1) [43], (2) [16], (3) [25] and (4) [27], for the ROSIS Pavia University dataset. The best OA results of each table are marked in bold italics.

Method	Training Samples		
	5	10	15
(1) SC-SC	72.02	72.90	75.21
(1) SC-S2C	71.09	72.00	79.48
(2) MLR + AL	63.00 ± 1.86	83.73 ± 1.86	85.63 ± 1.86
(3) TSVM	63.43 ± 1.22	63.73 ± 0.45	68.45 ± 1.07
(3) SS-LPSVM	56.95 ± 0.95	64.74 ± 0.39	78.76 ± 0.04
(4) MLR + KNN + SNI	76.46	85.47	88.08
TT_AL_MSH_MKE	**82.55**	**88.61**	**92.11**

5. Conclusions

In this paper, a novel semi-supervised tri-training algorithm for hyperspectral image classification was proposed. In the proposed algorithm, three measures of diversity, i.e., double-fault measure, disagreement metric and correlation coefficient, are used to select the optimal classifier combination.

Then, unlabeled samples were selected using the AL method and the consistent results of another two classifiers combined with spatial neighborhood information to predict the labels of unlabeled samples. Moreover, we utilize the multi-scale homogeneity method to refine the final classification result. To confirm the effectiveness of the proposed TT_AL_MSH_MKE, experiments were conducted on three real hyperspectral data, in comparison with the standard TT_MKE. Moreover, some methods that combine semi-supervised spectral-spatial classification with active learning are selected to validate the performance of the proposed method. Experiment results demonstrate that the OA of the proposed approaches is improved more than 10% compared with TT_MKE, and the proposed method outperforms other methods in particular for the small training datasets with 10 initial samples/class or less. Meanwhile, the proposed method can effectively reduce the salt and pepper noise in the classification maps.

Acknowledgments: This research is supported in part by the Natural Science Foundation of China (No. 41471356), the Fundamental Research Funds for the Central Universities (2014ZDPY14) and the Priority Academic Program Development of Jiangsu Higher Education Institutions. The authors would also like to thank Paolo Gamba at Pavia University for providing the ROSIS dataset and Dr. David Landgrede at Purdue University for providing the AVIRIS dataset.

Author Contributions: Kun Tan. and Jishuai Zhu conceived the idea. Kun Tan, Jishuai Zhu, Qian Du, Lixin Wu and Peijun Du designed the experiments and analyzed the data. Kun Tan, Qian Du and Jishuai Zhu wrote the main manuscript text. All authors reviewed the manuscript.

Conflicts of Interest: The authors declare no competing financial interests.

References

1. Safavian, S.R.; Landgrebe, D. A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybern.* **1991**, *21*, 660–674. [CrossRef]
2. Leung, K.M. Naive Bayesian Classifier. Available online: http://cis.poly.edu/~mleung/FRE7851/f07/naiveBayesianClassifier.pdf (accessed on 28 November 2007).
3. Cun, Y.L.; Boser, B.; Denker, J.S.; Howard, R.E.; Habbard, W.; Jackel, L.D.; Henderson, D. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems 2*; David, S.T., Ed.; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1990; pp. 396–404.
4. Qi, X.; Davison, B.D. Web page classification: Features and algorithms. *ACM Comput. Surv.* **2009**. [CrossRef]
5. Chen, R.C.; Hsieh, C.H. Web page classification based on a support vector machine using a weighted vote schema. *Expert Syst. Appl.* **2006**, *31*, 427–435. [CrossRef]
6. Zhang, Y.; Dong, Z.; Wu, L.; Wang, S. A hybrid method for MRI brain image classification. *Expert Syst. Appl.* **2011**, *38*, 10049–10053. [CrossRef]
7. Hosseini, M.S.; Zekri, M. Review of medical image classification using the adaptive neuro-fuzzy inference system. *J. Med. Signals Sens.* **2012**, *2*, 49–60. [PubMed]
8. Lyons, M.J.; Budynek, J.; Akamatsu, S. Automatic classification of single facial images. *IEEE Trans. Pattern Anal. Mach. Intell.* **1999**, *21*, 1357–1362. [CrossRef]
9. Shahshahani, B.M.; Landgrebe, D. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. *IEEE Trans. Geosci. Remote Sens.* **1994**, *32*, 1087–1095. [CrossRef]
10. Hughes, G.P. On the mean accuracy of statistical pattern recognizers. *IEEE Trans. Inf. Theory* **1968**, *14*, 55–63. [CrossRef]
11. Chapelle, O.; Schölkopf, B.; Zien, A. *Semi-Supervised Learning*; MIT Press: Cambridge, MA, USA, 2006.
12. Shi, Q.; Zhang, L.; Du, B. Semisupervised discriminative locally enhanced alignment for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 4800–4815. [CrossRef]
13. Shi, Q.; Du, B.; Zhang, L.P. Spatial coherence-based batch-mode active learning for remote sensing image classification. *IEEE Trans. Image Process.* **2015**, *24*, 2037–2050. [PubMed]
14. Yu, J.; Wang, M.; Tao, D. Semisupervised multiview distance metric learning for cartoon synthesis. *IEEE Trans. Image Process.* **2012**, *21*, 4636–4648. [PubMed]
15. Culp, M.; Michailidis, G.; Johnson, K. On multi-view learning with additive models. *Ann. Appl. Stat.* **2009**, *3*, 292–318. [CrossRef]

16. Dópido, I.; Li, J.; Marpu, P.R.; Plaza, A.; Bioucas Dias, J.M.; Benediktsson, J.A. Semisupervised self-learning for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 4032–4044. [CrossRef]
17. Tuia, D.; Ratle, F.; Pacifici, F.; Kanevski, M.F.; Emery, W.J. Active learning methods for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 2218–2232. [CrossRef]
18. Samiappan, S.; Moorhead, R.J. *Semi-Supervised Co-Training and Active Learning Framework for Hyperspectral Image Classification*; IEEE: New York, NY, USA, 2015; pp. 401–404.
19. Blum, A.; Mitchell, T. Combining labeled and unlabeled data with co-training. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, Madison, WI, USA, 24–26 July 1998; ACM: New York, NY, USA, 1998.
20. Ly, N.H.; Du, Q.; Fowler, J.E. Sparse graph-based discriminant analysis for hyperspectral imagery. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 3872–3884.
21. Bai, J.; Xiang, S.; Pan, C. A graph-based classification method for hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 803–817. [CrossRef]
22. Joachims, T. Transductive support vector machines. In *Semi-Supervised Learning*; Chapelle, O., Schölkopf, B., Zien, A., Eds.; MIT Press: Cambridge, MA, UK, 2006; pp. 104–117.
23. Chen, Y.; Wang, G.; Dong, S. Learning with progressive transductive support vector machine. *Pattern Recognit. Lett.* **2003**, *24*, 1845–1855. [CrossRef]
24. Xia, J.; Chanussot, J.; Du, P.; He, X. Semi-supervised dimensionality reduction for hyperspectral remote sensing image classification. In Proceedings of the 2012 4th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Shanghai, China, 4–7 June 2012.
25. Wang, L.; Hao, S.; Wang, Q.; Wang, Y. Semi-supervised classification for hyperspectral imagery based on spatial-spectral Label Propagation. *ISPRS J. Photogramm. Remote Sens.* **2014**, *97*, 123–137. [CrossRef]
26. Dopido, I.; Li, J.; Plaza, A.; Bioucas-Dias, J.M. A new semi-supervised approach for hyperspectral image classification with different active learning strategies. In Proceedings of the 2012 4th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Shanghai, China, 4–7 June 2012.
27. Tan, K.; Hu, J.; Li, J.; Du, P. A novel semi-supervised hyperspectral image classification approach based on spatial neighborhood information and classifier combination. *ISPRS J. Photogramm. Remote Sens.* **2015**, *105*, 19–29. [CrossRef]
28. Tan, K.; Li, E.; Du, Q.; Du, P. An efficient semi-supervised classification approach for hyperspectral imagery. *ISPRS J. Photogramm. Remote Sens.* **2014**, *97*, 36–45. [CrossRef]
29. Goldman, S.; Zhou, Y. Enhancing supervised learning with unlabeled data. In Proceedings of the Seventeenth International Conference on Machine Learning, San Francisco, CA, USA, 29 June–2 July 2000.
30. Zhou, Y.; Goldman, S. Democratic co-learning, ICTAI 2004. In Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence, Boca Raton, FL, USA, 15–17 November 2004.
31. Zhou, Z.-H.; Li, M. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 1529–1541. [CrossRef]
32. Tan, K.; Du, P.-J. Hyperspectral remote sensing image classification based on support vector machine. *J. Infrared Millim. Waves* **2008**, *27*, 123–128. [CrossRef]
33. Plaza, A.; Benediktsson, J.A.; Boardman, J.W.; Brazile, J.; Bruzzone, L.; Camps-Valls, G.; Chanussot, J.; Fauvel, M.; Gamba, P.; Gualtieri, A. Recent advances in techniques for hyperspectral image processing. *Remote Sens. Environ.* **2009**, *113*, S110–S122. [CrossRef]
34. Tan, K.; Zhang, J.; Du, Q.; Wang, X. GPU Parallel implementation of support vector machines for hyperspectral image classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 4647–4656. [CrossRef]
35. Li, J.; Bioucas-Dias, J.M.; Plaza, A. Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 4085–4098. [CrossRef]
36. Li, J.; Bioucas-Dias, J.M.; Plaza, A. Semisupervised hyperspectral image classification using soft sparse multinomial logistic regression. *IEEE Geosci. Remote Sens. Lett.* **2013**, *10*, 318–322.
37. Yang, J.M.; Yu, P.T.; Kuo, B.C. A Nonparametric feature extraction and its application to nearest neighbor classification for hyperspectral image data. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 1279–1293. [CrossRef]
38. Huang, G.-B.; Zhu, Q.-Y.; Siew, C.-K. Extreme learning machine: theory and applications. *Neurocomputing* **2006**, *70*, 489–501. [CrossRef]

39. Huang, G.-B.; Zhou, H.; Ding, X.; Zhang, R. Extreme learning machine for regression and multiclass classification. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2012**, *42*, 513–529. [CrossRef] [PubMed]
40. Sirlantzis, K.; Hoque, S.; Fairhurst, M.C. Diversity in multiple classifier ensembles based on binary feature quantisation with application to face recognition. *Appl. Soft Comput.* **2008**, *8*, 437–445. [CrossRef]
41. Yu, H.Y.; Gao, L.R.; Li, J.; Li, S.S.; Zhang, B.; Benediktsson, J.A. Spectral-spatial hyperspectral image classification using subspace-based support vector machines and adaptive markov random fields. *Remote Sens.* **2016**, *8*, 1–21. [CrossRef]
42. Su, H.; Yang, H.; Du, Q.; Sheng, Y. Semisupervised band clustering for dimensionality reduction of hyperspectral imagery. *IEEE Geosci. Remote Sens. Lett.* **2011**, *8*, 1135–1139. [CrossRef]
43. Wang, L.G.; Yang, Y.S.; Liu, D.F. Semisupervised classification for hyperspectral image based on spatial-spectral clustering. *J. Appl. Remote Sens.* **2015**, *9*, 096037. [CrossRef]