Semisupervised Learning for Image Mapping: Combining Clustering with Support Vector Machines

Debasis Chakraborty
Dept. of CSE(IoT,CS,BCT)
Asansol Engineering College
Asansol, India

Anup Kumar Mukhopadhyay
Dept. of CA
Asansol Engineering College
Asansol, India

Sayantani Chakraborty
MCA Student, Dept. of CA
Asansol Engineering College
Asansol, India

Shemim Begum
Dept. of CSE
Govt College of Engg. & Textile Technology
Berhampore, India

***Abstract*—This study presents a semisupervised learning approach using SVMs combined with fuzzy clustering for pixel classification in RS images. The method iteratively updates the training set by selecting informative unlabeled points through three fuzzy clustering strategies—center-based, random, and border-based selection. Experimental results on two remote sensing datasets show that center-based selection significantly enhances classifier performance compared to other strategies. Applied to Kolkata SPOT and Mumbai IRS images, the proposed approach improves classification accuracy, reduces redundancy in unlabeled data, and achieves higher kappa values and cluster quality indices than conventional SVM methods.**

***Index Terms*—Support vector machine, Remote sensing satellite images, Quadratic programming, Cluster based support vector machine, kernel function.**

I. Introduction

Image classification has long been central to remote sensing, forming the basis for numerous environmental and socioeconomic analyses. The success of classification depends on a suitable system design and sufficient training samples [1]. Traditional classifiers such as Naïve Bayes, k-NN, and SVM have been widely applied, while advanced algorithms like decision trees [2], ANNs [3], and evolutionary techniques [4] have demonstrated improved performance [5]. Despite progress, enhancing accuracy in land cover extraction remains a challenge, encouraging exploration of new learning approaches. A major limitation in supervised classification lies in the scarcity and quality of labeled samples, leading to ill-posed problems. To address this, two main strategies have been explored: semisupervised learning using both labeled and unlabeled data [6] and SVMs. Semisupervised approaches have proven effective for ill-posed classification problems by leveraging additional unlabeled data to refine decision boundaries [7]. Building on these ideas, the proposed CBSVM algorithm iteratively selects informative unlabeled points through clustering, labels them using SVM, and augments the training set to improve classification accuracy.

A useful unsupervised method in learning is clustering that divides the space of input into C clusters based on some similarity measure in which the value of C may not known beforehand. A clustering method partition the data set, which is represented as: $V(X) = C \times n$ partition matrix of the patterns in the set of n patterns. The partition matrix can be denoted by $V = [v_{cj}]$, c = 1,2,... and j = 1, 2,..., n, where v_{cj} is the membership of x_j to the cth cluster. Now v_{cj} is used to define the *crisp* and *fuzzy* clustering. To achieve crispness, v_{kj} is 0 or 1. In this case, any pattern can only be in one and no more than one class. Where as in case of fuzzy theory, the membership is :$0 \leq v_{cj} \leq 1$.

The proposed approach integrates Fuzzy C-Means (FCM) clustering with SVMs to improve accuracy. FCM estimates membership values for unlabeled points based on Euclidean distance, and points with high membership degrees are iteratively labeled using the trained SVM. These newly labeled points are appended to the training set and removed from the unlabeled pool until all samples are classified. The key idea is to selectively incorporate informative samples while avoiding redundant ones that may reduce accuracy. The proposed CB-SVM was evaluated on two numerical remote sensing datasets and two satellite images, showing improved generalization and higher kappa, accuracy, and cluster quality indices compared to the standard SVM.

II. Fuzzy clustering and validity assessment

FCM [8] is employed in associating the pattern x_j to values of membership to various clusters and certain validity indices such as Xie-Beni (XB) [9] and $\mathcal{I}$ [10] indices.

A. FCM

FCM is applied to generate a matrix $V(X)$ and minimizes the measure.

$$J_m = \sum_{j=1}^{n} \sum_{c=1}^{C} v_{cj}^{m} d^2(z_c, x_j), \quad 1 \leq m \leq \infty \tag{1}$$

where n is the samples numbers,C is the number of clusters, $u_{cj} \in \{0, 1\}$ is membership of the jth point in the c th cluster and m is the exponent. With the increase in m, there is additional fuzzification. $d(z_c, x_j)$ is the distance between point x_j and the c th centre z_c. The membership v_{ci} values of all the un-labeled points are determined with the help of the following equation:

$$v_{vi} = \frac{1}{\sum_{j=1}^{C} \left(\frac{d(z_c,x_i)}{d(z_j,x_i)}\right)^{\frac{2}{m-1}}}, \quad for \;\; 1 \leq c \leq C, \;\; 1 \leq i \leq n \tag{2}$$

The cluster centres are then updated as follows:

$$z_c = \frac{\sum_{i=1}^{n} v_{ci}^{m} \;\, x_c}{\sum_{i=1}^{n} v_{ci}^{m}} \quad 1 \leq c \leq C \tag{3}$$

The algorithm converges when the cluster centers stabilize between iterations. Upon convergence, each point is assigned to a cluster for which it offers the highest value.

B. Xie-Beni index

The details of XB measure is available in [9].The objective is to estimate minimum XB to achieve optimal cluster.

C. $\mathcal{I}$ index

The index $\mathcal{I}$, is defined in [10]. Higher $\mathcal{I}$ shows better solution.

III. Support Vector Machines

SVM operates by identifying the optimal hyperplane that maximizes the margin of separation between two classes [11]. Although originally developed for binary classification, the SVM framework has since been extended to solve multi-class problems.

SVM solves the problem as follows.

$$J(\mathbf{w}, \xi) = \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \xi_i \tag{5}$$

constrained to:

$$y_i(\phi(\mathbf{x_i}) \cdot \mathbf{w} + b) \geq 1 - \xi_i, \quad \xi_i \geq 0; \; i = 1, 2, \cdots, n. \tag{6}$$

Here, $\mathbf{w}$ and b define the linear decision boundary in the feature space that maximizes the margin. The slack variables ξ_i allow for misclassifications on the training data, and the user-specified parameter C controls classification error, as defined in equation (5).

Minimizing the first term in Eqn. (5) reduces the model complexity by controlling the VC-dimension, while minimizing the second term reduces the training misclassification error [11]. This formulation leads to a constrained quadratic programming (QP) problem.

The solution yields a decision function of the form:

$$f(\mathbf{x}) = sgn \left[\sum_{i=1}^{n} y_i \alpha_i k(\mathbf{x}, \mathbf{x_i}) + b \right] \tag{7}$$

with the kernel function $k(.,.)$ defined as.

$$k(x, x_i) \; = <\phi(x), \phi(x_i)> \tag{8}$$

The corresponding data points $\mathbf{x_i}$ associated with the nonzero α_i coefficients are called support vectors, and they entirely define the decision function. The term K(x, x) represents a nonlinear kernel function. In this work, we employ the RBF kernel, $k(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma||\mathbf{x_i} - \mathbf{x_j}||^2)$, where γ is a kernel parameter that controls the influence of each support vector.

IV. Cluster Based Support Vector Machine

Supervised algorithms require sufficient and representative training data to establish accurate decision boundaries, but acquiring such labeled data is often costly and time-consuming. To address this, clustering can be used to select informative unlabeled points, which are labeled and incorporated into the training set . However, incorrect labeling may degrade accuracy, so points with low membership confidence are excluded. In the proposed CBSVM, the Support Vector Machine (SVM) is incrementally trained using both labeled and confidently semilabeled samples. Fuzzy C-Means (FCM) clustering estimates membership values v_i providing labeling confidence and enabling selection of points near cluster centers for inclusion in the training set. This approach suppresses outliers, refines class means, and adjusts the decision boundary to minimize generalization error using both labeled and semilabeled data. Experiments also compare random, border-based, and center-based selection strategies on SPOT and IRS datasets. In each iteration, a fixed number N_0 of high-confidence unlabeled points are added to the training set until convergence, yielding improved classification accuracy and robust model generalization (Fig. 1).

V. Experimental Results

This experiment utilizes two numeric satellite datasets—a SPOT and IRS images [3]—where landcover types are represented by pixel intensity values without spatial information (Scatter plot shown in Fig.2(a) and 2(b) . Each dataset is randomized, partitioned into two equal subsets, and one subset forms an initial training set with 20% labeled and 80%

```
**Input**   Labeled points: $L = [(\mathbf{x_i}, y_i)]\,,\ i = 1, 2, \ldots, l$ and
un-labeled points: $U = [(\mathbf{x_i})]\,,\ i = l+1, \ldots, n.$
**classifier** SVM and FCM.
**Output**   Final SVM classifier with updated labeled set $W$.
1. Initialize the working set $W = L$, $N = N_0$ and $\epsilon = \epsilon_0$.
2. Train SVM with $W$.
3. Estimate the memebership values $v_i$ using Eqn. 2.
4. Obtain $N$ new input examples from $U$ with membership $v_i$
such that
    such that $v_i \geq \epsilon$ .
5. Obtain label vector of $N$ using the trained SVM.
6. Update $W$ by adding $N$ examples.
7. Remove $N$ from the unlabeled set $U$.
8.Repeat steps 2-7 until convergence is achieved.
```

Fig. 1. Cluster based support vector machine learning

unlabeled points. The proposed center-based active learning strategy is applied to this data, and the newly curated training set is used for classification, with random and border-based selection methods also employed for comparative analysis of unlabeled data incorporation. The three-dimensional SPOT dataset [3] contains 932 samples in green, red, and NIR bands across seven complex, overlapping classes: Turbid Water (TW), Pond Water (PW), Concrete, Vegetation, Habitation, Baren Land (BL), and Roads/Bridges (B/R). The IRS dataset [3] comprises 198 samples with four spectral bands (green, red, NIR, and infrared) partitioned into six classes: PW, TW1, TW2, Vegetation, BL, Habitation, and Concrete.

A. Input parameters

The parameters (C, γ) were estimated using a grid search. For instance, the optimal values found for one classifier were (C, γ) = (14.2, 3.0)as this pair yielded the smallest generalization error. The parameter ϵ was set to 0.5. Following common practice in the literature [4], [12], the value of the value of m is 2.

B. Evaluation Criterion

The algorithms are compared using four metrics-the Kappa index introduced by Cohen is available in [13], is a widely adopted measure of classification accuracy in many areas. It is computed from a confusion matrix (or contingency table), where each C_{ij} stands for the no. of instances from actual class i that were assigned to class j. The diagonal elements indicates correctly classified value. Other matrices J_m, XB and $\mathcal{I}$ are explained in section II.

TABLE I
CONFUSION MATRIX FOR SPOT DATA USING CBSVM (CENTER BASED)

		Actual						
		1	2	3	4	5	6	7
Predicted	1	60	0	0	0	0	0	0
	2	0	70	0	0	3	0	6
	3	0	5	76	0	0	0	7
	4	0	0	0	116	8	2	0
	5	0	2	0	6	30	0	0
	6	0	13	0	0	0	38	0
	7	0	13	6	0	0	0	5

TABLE II
CONFUSION MATRIX FOR NUMERIC IRS USING CBSVM (CENTER BASED)

		Actual					
		1	2	3	4	5	6
Predicted	1	8	0	0	0	0	0
	2	0	23	4	0	0	0
	3	0	0	31	1	1	0
	4	0	0	0	9	0	0
	5	0	0	0	4	10	0
	6	0	0	0	0	1	7

TABLE III
MEAN KAPPA AND CORRECT(%) FOR TWO DATASETS

Dataset	Classifier	Kappa index	%correct
SPOT	CBSVN	0.8150	84.77
	CSVM	0.7786	81.62
IRS	CBSVM	0.8569	88.89
	CSVM	0.8102	84.79

C. Comparative Results

A quantitative assessment of the two classifiers is presented in Table III, which reports the overall accuracy and Kappa index. The results demonstrate that the CBSVM classifier achieves a higher Kappa value and overall accuracy than the CSVM classifier on both datasets.

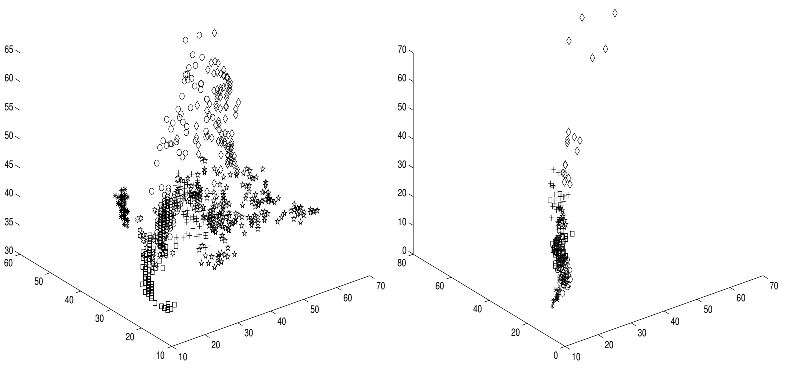
Furthermore, the confusion matrices generated by CBSVM, detailed in Tables I and II, exhibit superior classification performance compared to those of CSVM. This qualitative improvement is corroborated by quantitative cluster validity indices, reported in Table IV. For instance, on the SPOT dataset, CBSVM produces an $\mathcal{I}$ value of 168.22, outperforming the value of 158.46 produced by CSVM. The superiority of the CBSVM method is consistently reflected across the other performance metrics as well

D. Effect of cluster based Support Vector Learning

Figures 3(a) and 3(b) depict the accuracy (test) vs the number of training examples for three selection strategies: center-based, random, and border-based. The results demonstrate that center-based selection consistently yields the highest test accuracy across both datasets.In each iteration, up to $N = 10$ and $N = 5$ samples were selected for the SPOT and IRS datasets, respectively. A significant improvement in accuracy is observed with center-based selection as the training sample numbers increases, showing that points near cluster centers are more informative than those selected by the other methods. The proposed scheme effectively identifies these informative points within the first few iterations. However, accuracy declines with further iterations, likely due to the accumulation of mislabeled samples. This observation suggests that optimal performance is achieved with a limited number of iterations. Consequently, our experiments with CBSVM were conducted using 7 to 8 iterations.

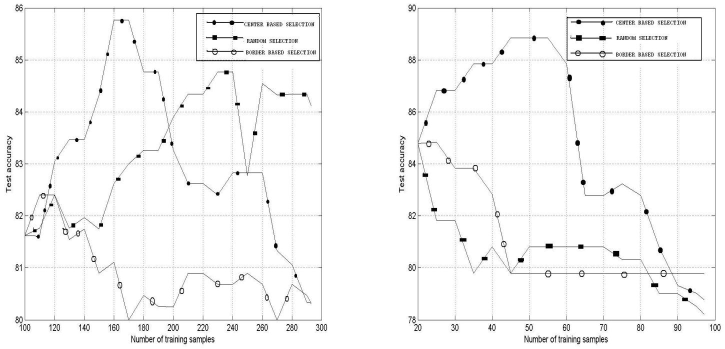
VI. IMAGE MAPPING

The image data employed here is SPOT and IRS images [3], [12]. The SPOT contains three spectral bands whereas IRS has



(a) (b)

Fig. 2. Scatter plot of (a) SPOT and (b) IRS



(a) (b)

Fig. 3. Plot of test accuracy vs number of training data using three selection techniques (a) numeric SPOT data with $L = 100$ and $N = 10$ and (b) numeric IRS data with $L = 20$ and $N = 5$

TABLE IV
VALIDITY MEASURES FOR TWO DATASETS

dataset	Classifier	J_m	XB	$\mathcal{I}$
SPOT	CBSVM	8.2232E+3	0.6917	168.22
	CSVM	8.4198E+3	0.7933	158.46
IRS	CBSVM	3.7001E+3	0.4300	208.26
	CSVM	3.8352E+3	0.4942	179.88

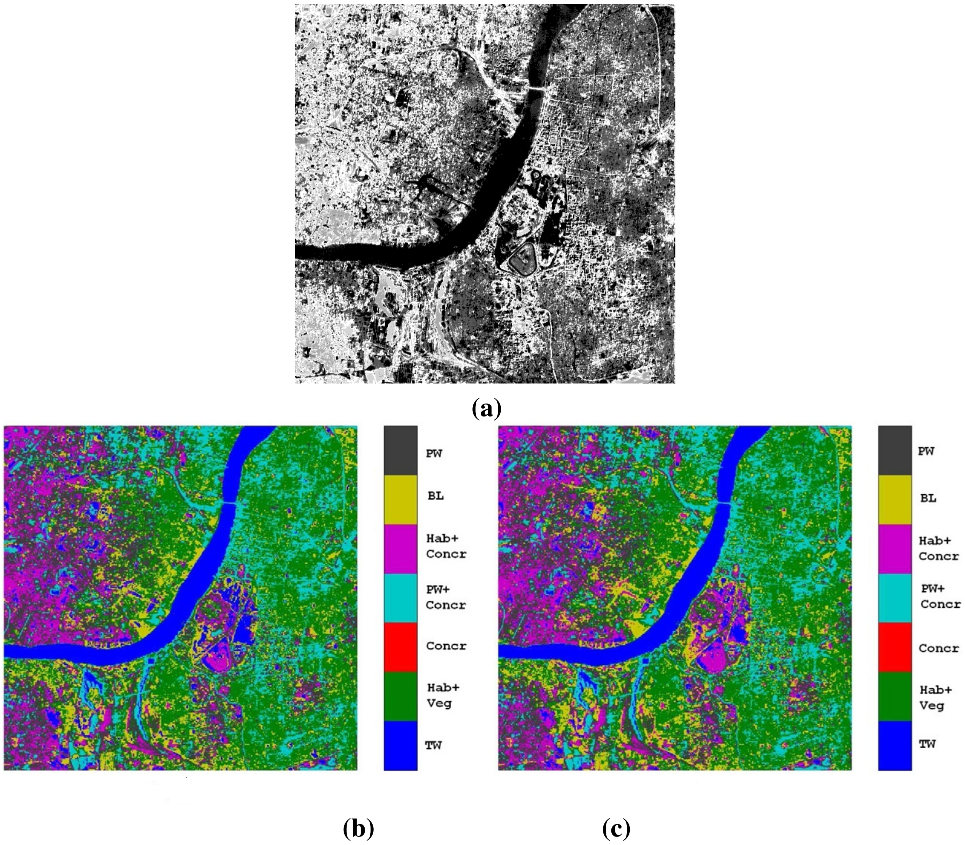
four bands. Each image has 262144 unclassified pixels in both the datasets.

In this one, the problem is to label the image with various landcover units, with the assistance of the given features (spectral bands). SPOT and IRS have been taken as the starting training set. Also, a random selection of image database has been done to get a set of unlabeled points which include four times as many points as the train set. Its training sets are then updated through a series of training with the proposed technique. Image datasets will be assigned a value of N.

Tables V and VI in the SPOT and IRS image respectively tabulate the achievements of two methods in terms of three indices. Classified images will also be verifiable on the efficiency of the learners.

A. SPOT image

Fig. 4(a) shows the green band in grayscale. Classification results for this image, which contains seven land cover classes [3] are presented in Fig. 4(b) for CBSVM and Figure 4(c) for CSVM. An analysis of the CBSVM result in Fig. 4(b) reveals several key features: The Hooghly river is correctly classified as TW. Two distinct water bodies south of the river—the Kidderpore Dockyard (right) and Lake Garden Reach (left)—are accurately identified as a mix of PW and TW. In contrast, the CSVM result (Fig. 4(c)) misclassifies these entirely as PW.The Talis Nala, a thin channel extending from the river, is correctly classified as PW.The Race Course, visible as a triangular patch, is more vividly delineated by CBSVM than by CSVM.The Beleghata Canal, extending from the top-right corner, is correctly identified as PW.Urban areas in the top-right and central sections are well-captured. Rabindra Setu bridge, which crosses the river Hooghly, is captured as a mixture of Concrete and B/R.The remaining areas in the CSVM result (Figure 4(c)) are broadly comparable to the CBSVM classification but lack the same level of detail and accuracy in the specific regions noted above.Table V presents the index values for SPOT classified by the evaluating tech-



(a)

(b) (c)

Fig. 4. (a) SPOT image and classified image using (b) CBSVM and (c) CSVM

niques. The results demonstrate that CBSVM technique yields superior classification quality. This superiority is contingent on obtaining a correct model during incremental learning. The improvement is due to the additional semi-labeled examples incorporated by CBSVM contain the informational content necessary to enhance classification accuracy.

TABLE V
RESULTS FOR THE KOLKATA SPOT DATA

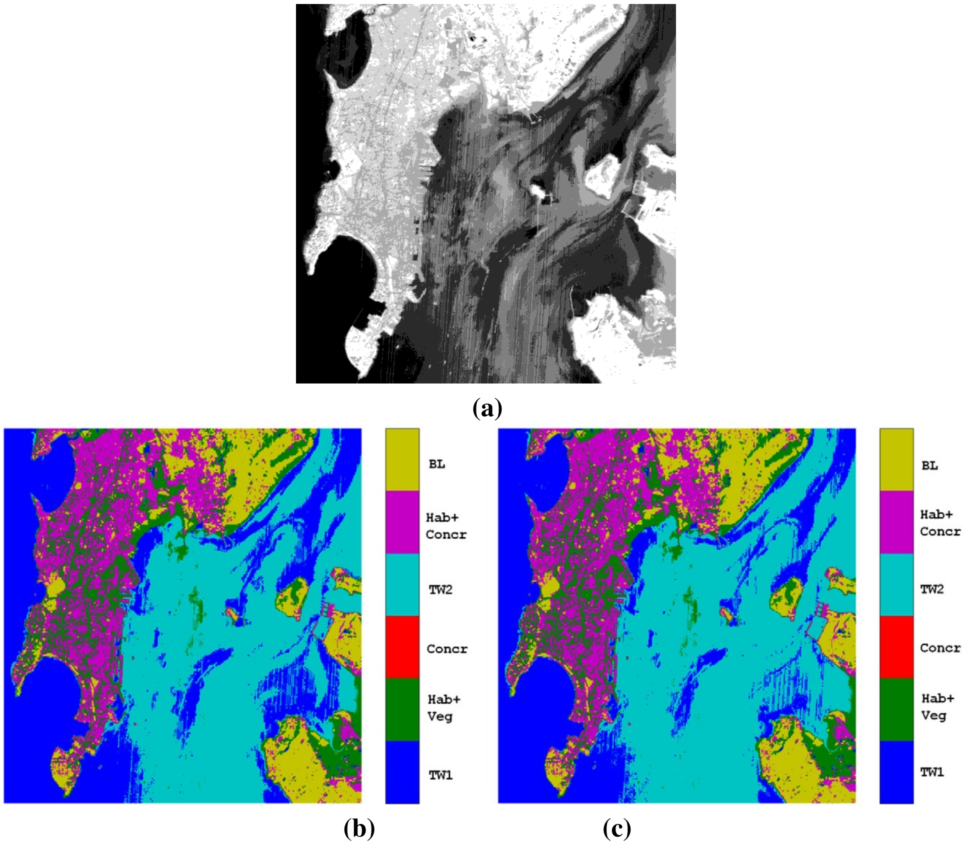
Classifier	J_m	XB	$\mathcal{I}$
CBSVM	1.9603E+6	0.2281	40.21
CSVM	1.9561E+6	0.2367	37.46

B. IRS image

Figure 5(a) displays the Mumbai IRS image in grayscale for the infrared band (Band 4). The corresponding classification results obtained using the CBSVM and CSVM shown in in Fig 5(b) and 5(c), respectively.Based on available ground truth, the image features six primary land cover classes: turbid water (subdivided into TW1 and TW2), concrete, habitation, vegetation, and barren land.An analysis of the CBSVM result in Fig. 5(b) shows that the the city is bounded by the Arabian Sea on three sides.. The classifier successfully distinguishes two distinct spectral properties of the seawater, categorizing it as TW1 and TW2, which aligns with the variations visible in the original image (Figure 5(a)).Several islands, including Elephanta Islands, are visible in the bottom-right portion of image and are largely classified. The dockyard, located on the southeastern coast and characterized by a distinctive three-fingered structure, is also accurately identified. As expected, BL within the islands are correctly classified. The heavily industrialized and urbanized mainland is predominantly identified as mixture of habitation and concrete.Fig. 5(c) presents the classification result from the CSVM method. While the CSVM classifier also distinguishes two spectral regions in the Arabian Sea, the result is less precise than the segmentation achieved by CBSVM in Figure 5(b). The classification of other land cover types is broadly consistent with the results shown previously. Table VI presents the J_m, XB and $\mathcal{I}$ index values for the image of IRS using two algorithms. The results clearly demonstrate the superior performance CBSVM method over traditional CSVM.

TABLE VI
RESULTS FOR THE MUMBAI IRS DATA

Classifier	J_m	XB	$\mathcal{I}$
CBSVM	2.0970E+6	0.1941	97.49
CSVM	2.0985E+6	0.2170	87.34



(a)

(b) **(c)**

Fig. 5. (a) IRS image and classified image using (b) CBSVM and (c) CSVM

VII. Conclusion

This article presents a semi-supervised pixel classification method for remote sensing imagery, which integrates fuzzy clustering to enhance SVM classifier. The objective is to iteratively construct an accurate model by selectively incorporating the most informative points from the unlabeled data. The proposed algorithm, CBSVM, begins with a small initial training set and incrementally expands it with high-value samples. The strength of CBSVM lies in this targeted, incremental growth of the labeled dataset. Experimental outcomes on both numerical benchmarks and real datasets indicates that this technique achieves superiority over CSVM approach. Consequently, this method is a promising candidate for deployment in other domains where labeled data is scarce or costly to obtain.

References

[1] A. M. Cingolani, D. Renison and M. R. Cabido, "Mapping vegetation in a heterogeneous mountain rangeland using lanasat data: an alternative method to define and classify land-cover units," *Remote Sensing of Environment*, vol. 92, pp. 84–97, 2004.

[2] M. A. Friedl, C. E. Brodley and A. H. Strahler, "Maximizing land cover classification accuracies produced by decision trees at continental to global scales," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, pp. 969–979, 1999.

[3] S. Bandyopadhyay and S. K. Pal, "Pixel classification using variable string genetic algorithms with chromosome differentiation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 2, pp. 303–308, 2001.

[4] S. Bandyopadhyay and U. Maulik, "Genetic clustering for automatic evalution of clusters and application to image classification," *Patter Recognition*, vol. 35, no. 2, pp. 1197–1208, 2002.

[5] T. S. Li, C. Y. Chen and C. T. Su, "Comparison of neural and statistical algorithms for supervised classification of multidimensional Data," *International Journal of Industrial Engineering - Theory, Application and Practice*, vol. 10, pp. 73–81, 2003.

[6] M. Chi and L. Bruzzone, " A semilabeled-sample driven bagging technique for ill-posed classification problem," *IEEE transactions on Geoscience and Remote Sensing Letterrs*, vol. 2, no. 1, pp. 69–73, 2005.

[7] Q. Jackson and D. A. Landgrebe, "A adaptive method for combined covariance estimation and classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 5, pp. 1082–1087, 2002.

[8] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, New York: Plenum, 1981.

[9] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 841–847, 1991.

[10] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1650–1654, 2002.

[11] C. J. C. Burges, "A tutorial on support vector machines for pattern cecognition," *Knowledge Discovery and Data Mining*, vol. 2, pp. 121–167, 1998.

[12] S. Bandyopadhyay, U. Maulik, and A. Mukhopadhyay, "Multiobjective genetic clustering for pixel classification in remote sensing imagery," *IEEE transactions on Geoscience and Remote Sensing*, vol. 45, no. 5, pp. 1506–1511, 2007.

[13] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurements*, vol. 20, no. 12, pp. 37–46, 1960. Algorithms for clustering data, Prentice Hall, Englewood Cliffs, NJ, 1988.