Combining clustering and classification for remote-sensing images using unlabeled data

Xiaoyong Bian (边小勇)[1,2*], Tianxu Zhang (张天序)[1], and Xiaolong Zhang (张晓龙)[2]

[1] *Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan 430074, China*

[2] *School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430081, China*

[*] *Corresponding author: xyjconf100@163.com*

Received August 12, 2010; accepted October 28, 2010; posted online January 1, 2011

A joint clustering and classification approach is proposed. This approach exploits unlabeled data for efficient clustering, which is applied in the classification with support vector machine (SVM) in the case of small-size training samples. The proposed method requires no prior information on data labels, and yields better cluster structures. Through cluster assumption and the notions of support vectors, the most confident k cluster centers and data points near the cluster boundaries are labeled and used to train a reliable SVM classifier. Our method gains better estimation of data distributions and mitigates the unrepresentative problem of small-size training samples. The data set collected from Landsat Thematic Mapper (Landsat TM-5) validates the effectiveness of the proposed approach.

OCIS codes: 100.3008, 100.5010.

doi: 10.3788/COL201109.011002.

Clustering and classification techniques have been widely studied and applied in the fields of information processing, data analysis, and computer vision, among others[1,2]. Clustering in the form of unsupervised methods involves partitioning the unlabeled data points into disjoint subsets (clusters) based on the underlying structure of the data. One of the most widely used clustering algorithms is k-means and its variation[3]. For the supervised classification methods, most existing algorithms require sufficient training samples to train a reliable classifier and aim to generalize well on new data points. However, the general supervision information provided by pairwise constraints or label information is often unavailable in certain application domains[4,5].

To the best of our knowledge, abundant unlabeled data available in remote-sensing images have hardly been fully used in the classification process over the past decade, even if they should be more representative and can be exploited to enhance classification tasks. This characteristic has recently motivated an increasing number of research interests in the semi-supervised learning (SSL) paradigm, which aims to improve learning performance by incorporating unlabeled data into the learning process. Therefore, combining clustering and classification techniques to the analysis of remote-sensing images has attracted great attention. It has been shown in Ref. [4] that semi-supervised clustering is guided by pairwise constraints provided by the user, thus this method is not always accurate. Chi *et al.* carried out directly in the primal representation the optimization problems on support vector machine (SVM) for the classification of hyperspectral remote-sensing data, with the computation complexity of $O(nd^2 + d^3)$ when $n \ll d$, where n is the labeled samples and d is the features[5]. As in all supervised learning methods, this alternative implementation technique requires the manual selection of training data. In Ref. [6], binary transductive SVM (TSVM) was proposed to classify multi-class hyperspectral remote-sensing images using transductive samples, which may result in a nonconvex optimization problem. In addition, a recent effort in Ref. [7] employed k Gaussian mixture models constrained by labeled samples to estimate the data distribution, and the reported classification accuracy in each training data set was generally lower than 91% even when there were hundreds of training samples per class. All these methods explicitly use pairwise constraints or label information to guide clustering and classification learning.

One of the two major issues related to remote-sensing image classification problems is that labeled data are often difficult to be obtained and very sparse in practical applications. Another critical issue is that the training samples are often collected from the same area of the scene regardless of the variation of spectral signatures of land cover classes in the spatial domain and fail to estimate the distributions for the entire data. Both problems result in the risk of overfitting the training samples and may involve poor generalization capabilities in the classifier[6,7].

In this letter, a novel approach combining clustering and classification techniques is proposed to handle the unrepresentative problem of small-size training samples for SVM classification. We propose to import appropriately unlabeled data through the clustering method to the classification of remote-sensing images; bipartition-based k-means (BKM) is utilized to better estimate k initial centers, and the confidently clustered data can then be generated. Through cluster assumption[8] and the notions of convex hulls and reduced ellipse-like structures of the cluster regions[9], the confident data points from different cluster regions can be further evaluated, extracted, and labeled as training samples; subsequently, SVM can be trained with the labeled data. The use of clustering prior to classification approach is a natural and practical choice because the labeled data may be unavailable. In contrast to Ref. [3], our version of clustering algorithm works in a completely unsupervised manner. The final classification results of the proposed approach are reported on real remote-sensing image, and the results are appealing when compared to the state-of-the-art SVM.

Traditional k-means is sensitive to the initial cluster center, and the clustering result varies with different initial centers. It has been shown that k-means is prone to local minima and uncertainty regarding the number of clusters in the given data set. Several solutions similar to the ones used for the traditional algorithm, such as split and merge techniques, may be adopted. In this study, we propose a BKM approach to partition the data set, wherein k is set as the number of classes according to prior class knowledge. The large number of unlabeled data is utilized by the BKM procedure to estimate the data distribution. Each sequence of the investigated image is divided into two disjoint parts based on gray mean, and multi-binary images are obtained. The key issue is to define the "density" function in order to find the k densest regions and to approximate the underlying distribution. For convenience, the densest region should contain the maximum data points in the region, and the jth $(j = 2, \cdots, k)$ densest region should have the jth maximum data points in the region; simultaneously, the farthest distance to all first $j-1$ cluster regions can be derived according to the "density" definition. Starting from $j = 2$, they can be yielded in terms of

$$\max(\min(d(C_i, Z_1), d(C_i, Z_2), \cdots, d(C_i, Z_{j-1})),$$

$$(i = 1, \cdots, 2k;\ j = 2, \cdots, k) \tag{1}$$

where C_i is ith region center and Z_j is the jth region that satisfies the C_i in expression (1). The distance measure $d(x_i, x_j)$ can be calculated and the object function J in our clustering approach adopts the default as that in k-means; a detailed description can be seen in Ref. [3].

Assuming that the data is $i.i.d$ data, where n is data points and b is the size of the 3rd dimension of the data set, a BKM algorithm can be illustrated as follows.

Algorithm: BKM

1) Input: Data set $X = (x_{i,j}), x \in R^d, i = 1, \cdots, n; j = 1, \cdots, b$; cluster number k.

2) Partition iteratively on X into $2k$ binary regions based on the gray mean per band.

3) Compute the "density" value for all regions and obtain the 1st region, Z_1.

4) Get the remainder $k-1$ cluster regions, $Z_2, \cdots, Z_k$, according to expression (1).

5) Run k-means with the selected k cluster centers.

6) Output: $\{Z_i\ (i = 1, \cdots, k)\}$.

As can be seen, our clustering method does not require prior information other than the cluster number k (equals the number of classes). We reasonably presume that k cluster centers can be with the most confident class label under the clustering model and assigned a deterministic label, respectively; more confident data points are studied as follows.

As described above, k cluster centers are confidently labeled and then propagated to a part of unlabeled data from different cluster regions. However, the question is: how can we extend the labeled training samples? As standard SVM could optimize maximum margin hyperplane by introducing support vectors, we attempted to locate such data points nearby cluster boundaries aside from the centroids. Figure 1 illustrates this motivation.

The benefits of taking much more confident unlabeled data into account can be seen in Fig. 1. Figure 1(a) shows that our proposed approach (solid line) can outperform the boundary given by the direct SVM, with one labeled data for each class (class boundary is denoted by the dashed line). The boundary of the proposed approach is clearly more reasonable. Figure 1(b) illustrates the extended labeled data based on the cluster centers with the signs + and −. The two scaled ellipses enclosed by the dashed line expanded the labeled data by applying reduced ellipse-like region structure and convex hull on the basis of the clustering result; some originally unlabeled data can be viewed as labeled data with high confidence. As more confident training samples are added to train a SVM model, the classifier with higher accuracy can be obtained. As depicted in Refs. [8, 10], cluster assumption favors decision boundaries for classification passing through low-density regions in the image space; for instance, the most confident data points can be the centroids or the ones near the boundaries, which can be confidently classified. Some studies have indicated that the training samples on the cluster boundaries better represent the distribution of real data. Therefore, we also exploited such data points.

The pseudo code of the proposed cluster labeling and evaluation algorithm is illustrated as follows.

Algorithm: cluster labeling and evaluation.

1) Input: data set $X = (x_{i,j}), x \in R^d$; cluster result $Z_i\ (i = 1, \cdots, k)$; region properties "regionprops".

2) Explore the key region properties including "centroid", "area", "convex hull", and reduced "ellipse"; a part of the confident data points from each cluster region have been surveyed.

3) Label c% data points and generate a group of training samples, and run k-fold cross-validation (CV) process.

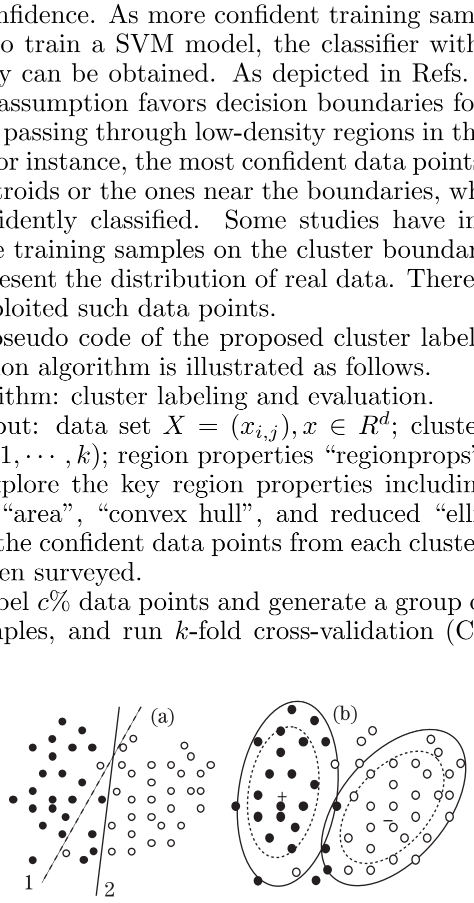

Fig. 1. An illustrative example of combining clustering and classification; + and − in the centers of the two cluster are newly labeled by clustering; other data points (black and white) are all unlabeled data. (a) Decision planes of the supervised learning methods (1) and our approach (2) given by the SVMs; (b) expanded labeled data in dashed ellipses.

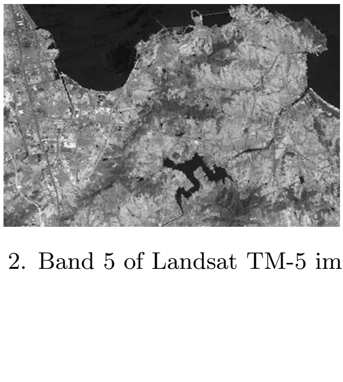

Fig. 2. Band 5 of Landsat TM-5 image.

4) Train a SVM classifier with optimal radial basis function (RBF) kernel parameters c and g.

5) Evaluate the unrepresentative problem of the training samples in terms of CV accuracy and the ground truth.

6) Output: SVM model and classification map of the investigated image.

As can be seen from the above algorithm, we chose the most confident $c\%$ data points and labeled them. If we choose c=100, all data points from the same cluster are labeled and used as training samples. This is impractical because the clustering procedure might introduce incorrectly labeled data points into the training samples. As a result, the classification performance will decrease because of error-labeling, although we can avoid this case to a certain extent. Finally, the optimal classification model can be used to classify all unlabeled data.

The proposed BKM-SVM approach is the integration of the clustering and SVM techniques previously discussed. The benefits of combining clustering and classification to the analysis of remote-sensing image are validated in the experiments.

We evaluate our approach on a real remote-sensing image. Some other methods are implemented for comparison, i.e., k-means-based SVM (KM-SVM) and standard SVM. The performance study in real remote-sensing image classification is conducted.

The data used in the experiments were collected using the multispectral Landsat-5 Thematic Mapper (Landsat TM-5) on the Jiaodong Biland in Shandong Province, China. There were originally seven bands of TM-5 sensor with median spectral and spatial resolutions, with image size of 332 × 464 pixels. Channel 6 was removed due to panchromatic, thus the data set contained six features in four land-cover types. Figure 2 shows channel 5 of the investigated image.

Unlike purely supervised methods through manual selection of regions of interest (ROIs), the training samples in our proposed approach were automatically generated by our clustering method in spatially disjoint patches scattered throughout the scenario. The total number of training samples varied from 20 to 400 to estimate the impact of different training sample sizes, and the whole image was used as test sample. The ground truth of the investigated image was created by human labeling, which indicated that the training and test samples were generally balanced.

To demonstrate the effectiveness of our proposed algorithm, we compared the following three algorithms.

KM-SVM: This classification algorithm simply employs k-means to cluster the original data set in advance, and then every cluster center and the ones closest to the centroids are labeled and extracted to train a SVM classifier.

BKM-SVM: Unlike KM-SVM, BKM clustering can capture the intrinsic geometrical structure of unlabeled data, and the most confident data points are generated with the aid of convex hull and reduced ellipse-like region structures; it used to learn a Gaussian RBF kernel.

SVM: The standard SVM approach demands a number of labeled data, which are hard to obtain even if necessary prior knowledge is provided.

In our experiments, the training and test samples were normalized in the range between 0 and 1. A one-against-one multi-class scheme was adopted with LIBSVM data, and the optimal RBF kernel parameters c_{opt} and g_{opt} can be achieved by 10-cross validation, where $c = (2^{-5}, \cdots, 2^{5})$ and $g = (2^{-5}, \cdots, 2^{5})$. The evaluation metric, classification accuracy (CA) was used, which is defined as

$$\mathrm{CA} = \sum\nolimits_{i=1}^{n} \frac{C(l_i, g_i)}{n}, \qquad (2)$$

where n denotes the total number of data points, $C(l_i, g_i)$ is the counter function that adds 1 to itself if and only if $l_i = g_i$, and the predicted label l_i is permuted to match the label given by ground truth g_i. The classification performance was evaluated by comparing the predicted labels of the given algorithm with that given by ground truth. All our experiments have been performed on a P4 1.6-GHz Windows XP computer with 512-MB memory.

In order to survey nonconvex cluster structure problem, we first obtain the ellipse-like structure of each cluster region and scale it with a reduced proportion. After assuring that the data points contained in the reduced ellipse are confidently coincident with the same cluster, we label them. Figure 3 illustrates a reduced ellipse-like

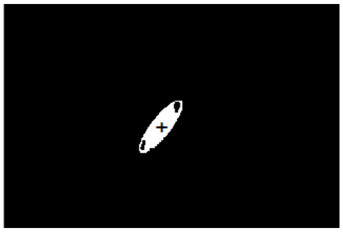

Fig. 3. Reduced ellipse-like region structure extracted from cluster result and its center signed with +.

Table 1. Obtained Training Samples by Different Cluster Structures

Chasses	Training Data						
	Ground Truth	Size 20	Size 50	Size 100	Size 200	Size 300	Size 400
Urban Area	29107	5	12	27	56	85	115
Forest	33018	3	8	20	45	68	91
Farm	31503	3	10	22	40	75	86
Water	60420	9	20	31	59	72	108
Overall	154048	20	50	100	200	300	400

Table 2. Classification Accuracy Comparison on the Investigated Image

Training Data	Accuracy (%)		
	KM-SVM	SVM	BKM-SVM
Size 20	88.64	87.44	92.36
Size 50	89.27	88.31	93.52
Size 100	90.16	93.05	93.74
Size 200	91.21	93.82	94.28
Size 300	91.86	93.95	94.34
Size 400	92.2	94.21	95.15
Average	90.56	91.8	93.9

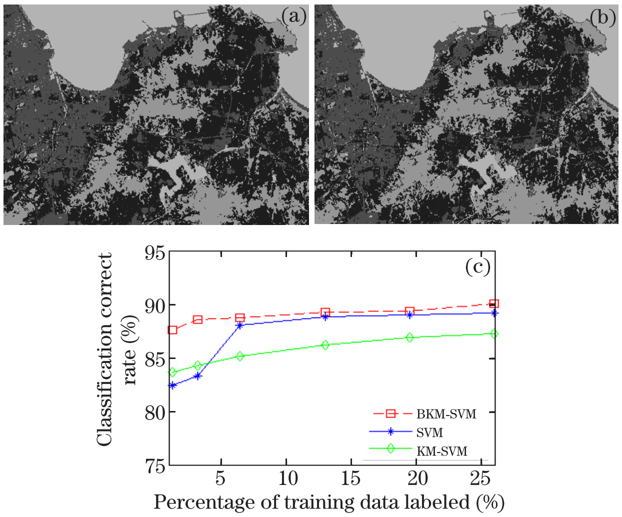

Fig. 4. Comparison among the classification results provided by SVM and our approach. (a) SVM obtained manually; (b) our approach obtained automatically; (c) classification correct rate versus percentage of labeled data on real TM-5 image.

structure of the investigated region, which can better capture the data distribution. This structure may introduce noise, i.e., the black inner holes, and they must be removed and never labeled.

The proposed algorithm was applied to classify the remote-sensing image in the case of the small-size training samples acquired. Table 1 shows the different numbers of training samples in different splits. Experiments with 20, 50, 100, 200, 300, and 400 training samples were designed. However, our training samples were obtained through automatic sub-sampling strategy instead of random selection. Table 2 reports the classification accuracies provided by the proposed algorithm in comparison to the standard SVM and KM-SVM with different training sample sizes. The classification accuracies provided by our approach in all the data sets consistently maintains a stable increase, with the maximum increase of up to 5.21% compared with the standard SVM; other significant case is the data set of size 20, where the increase obtained by the proposed approach with respect to the SVM was 4.92%. On average, our approach outperforms SVM by 2.1% and KM-SVM by 3.34%. Direct k-means procedure may clearly introduce incorrect labels. This confirms that the proposed approach can increase not only classification accuracy, but also stability. Furthermore, the computation complexity of our approach is no more than the standard SVM. Due to the use of the structure of unlabeled data, each class is relatively well separated from other classes in Fig. 4. Therefore, the data distribution can be better estimated and the unrepresentative problem of small-size training samples is mitigated.

In conclusion, we study a method to generate the most confident training samples in an unsupervised manner, employing the clustering technique to obtain small-size training samples by introducing much more unlabeled data and thus enhancing the classification of remote-sensing images. We validate the effectiveness of the proposed approach in comparison to the state-of-the-art SVM. To the best of our knowledge, there is no study that focuses on clustering unlabeled data to yield small-size labeled data for classification. The most related study on clustering aided by labeled data to guide classification could be seen in Refs. [7, 10]. These studies require class label prior knowledge, whereas in our case, small-size labeled data can be automatically obtained and evaluated. Future research should study the application of the proposed approach to classify hyperspectral data as well as the use of transductive SVM and Gaussian mixture model (GMM) methods.

We thank Dr. C.-J. Lin for providing the LIBSVM. This work was supported by the National Natural Science Foundation of China (Nos. 60736010 and 60975031), in part by the Science Foundation of Wuhan University of Science and Technology (No. 2008TD04), the Open Foundation of State Key Laboratory of Bioelectronics, Southeast University, and Hubei Provincial Natural Science Foundation (No. 2009CAD034)

References

1. H. Deng, J. Liu, and Z. Chen, Chin. Opt. Lett. **8,** 24 (2010).
2. H. Su and Y. Sheng, Chin. Opt. Lett. **8,** 811 (2010).
3. G. Hu, S. Zhou, J. Guan, and X. Hu, Information Processing and Management **44,** 1397 (2008).
4. W. Tang, H. Xiong, S. Zhong, and J. Wu, in *Proceedings of Knowledge Discovery and Data Mining* (*KDD'07*) 707 (2007).
5. M. Chi, R. Feng, and L. Bruzzone, Advances in Space Research **41,** 1793 (2008).
6. L. Bruzzone, M. Chi, and M. Marconcini, IEEE Trans. Geosci. Remote Sens. **44,** 3363 (2006).
7. M. Chi, Q. Qian, and J. A. Benediktsson, in *Proceedings of IEEE International Geoscience and Remote Sensing Symposium* (*IGARSS'08*) I-209 (2008).
8. H.-J. Zeng, X.-H. Wang, Z. Chen, H. Lu, and W.-Y. Ma, in *Proceedings of the 3rd IEEE International Conference on Data Mining* (*ICDM'03*) (2003).
9. Z. Liu, J. Liu, C. Pan, and G. Wang, IEEE Trans. Neural Networks **20,** 1215 (2009).
10. O. Chapelle and A. Zien, in *Proceedings of the International Workshop on Artificial Intelligence and Statistics* (*AISTATS'05*) (2005).