SVM-Based Data Editing for Enhanced One-Class Classification of Remotely Sensed Imagery

Xiaomu Song, *Member, IEEE*, Guoliang Fan, *Senior Member, IEEE*, and Mahesh Rao

***Abstract*—This paper studies a specific one-class classification problem where the training data are corrupted by significant outliers. Specifically, we are interested in the one-class support vector machine (OCSVM) approach that normally requires good training data. However, perfect training data are usually hard to obtain in most real-world applications due to the inherent data variability and uncertainty. To address this issue, we propose an OCSVM-based data editing and classification method that can iteratively purify the training data and learn an appropriate classifier from the trimmed training set. The proposed method is compared with a general OCSVM approach trained from two types of bootstrap samples, and applied to the mapping and compliance monitoring tasks for the U.S. Department of Agriculture's Conservation Reserve Program using remotely sensed imagery. Experimental results show that the proposed method outperforms the general OCSVM using bootstrap samples at a lower computational load.**

***Index Terms*—Bootstrap techniques, compliance monitoring, Conservation Reserve Program (CRP), data editing, mapping, one-class classification, support vector machines (SVMs).**

I. Introduction

ONE-CLASS classification is a special case and an extension of the two-class classification problem, where the two classes are called the target and the outlier class, respectively. Detecting a target class of interest remains challenging in remote sensing data analysis due to the spectral and spatial variability of different classes in a feature space. The difficulty lies in determining an appropriate classification boundary that embraces the data of the target class in the feature space.

Support vector machines (SVMs) have been widely used in remote sensing applications [1], [2] and show advantages over conventional classifiers, such as maximum likelihood and artificial neural networks [3]. The SVM separates the data of two classes by maximizing the margin between them, and may have a good generalization performance if the SVM capacity is well controlled [4]. The margin-based criterion also makes the SVM not suffer from the curse of dimensionality when the data size does not significantly outnumber the feature dimension. The SVM is often referred to as two-class SVM (TCSVM) [4], and is a supervised learning algorithm that requires the training data for each class. It was extended to one-class learning, and led to the one-class SVM (OCSVM) [5], [6], which can be applied to either supervised or unsupervised classifications.

In SVM learning, the presence of outliers in the training data will degrade the learning and classification performance. Bootstrap techniques that create pseudoreplicated data samples via resampling are proposed to reduce the effect of outliers [7]. Additionally, there are other techniques such as "condensing" and "editing" that suppress the effect of outliers by trimming the training data [8], [9]. The bootstrap techniques have been applied to SVM learning and show improved classification results [10]. However, if the outlier proportion is significant, the bootstrap techniques are not effective in improving the classification performance. Moreover, it is usually time consuming to compute bootstrap samples.

In this paper, we propose an OCSVM-based data editing and classification method that first iteratively removes outliers in the training data, and then trains an OCSVM with the purified training data as the classifier. There are two assumptions for the training data: 1) the outlier percentage is less than half, and 2) target and outlier samples do not significantly overlap in the feature space. Particularly, we apply this method to the mapping and compliance monitoring tasks of the U.S. Department of Agriculture's Conservation Reserve Program (CRP) using Landsat Thematic Mapper (TM) imagery. The CRP is a voluntary program that encourages farmers to plant long-term native vegetative covers to improve soil, water, and wildlife resources.[1] Field boundaries of CRP tracts (used as reference data in this paper) were manually digitized by the Natural Resources Conservation Service (NRCS) from archived aerial photographs. The possible dislocation and spatial misalignment of CRP tracts will deteriorate the reliability of reference data for training purposes. Moreover, CRP enrollments change with sign-ups and some CRP tracts are not compliant with contract stipulations. Therefore, most CRP reference data are not accurate and/or up-to-date. In our previous work, we have separately addressed CRP mapping [11] and compliance monitoring [12], [13] by developing SVM-based approaches. The proposed method is the first attempt to jointly implement CRP mapping and compliance monitoring into one flow.

[1] http://www.fsa.usda.gov/dafp/cepd/crp.htm

Manuscript received January 5, 2007; revised July 10, 2007. This work was supported in part by the Oklahoma NASA EPSCoR Research Initiation Grant (2003, 2005) and in part by the Water Research Center Grant (2003, 2004), Oklahoma State University.

X. Song was with the School of Electrical and Computer Engineering, Oklahoma State University, Stillwater, OK 74078 USA. He is now with the Evanston Northwestern Healthcare Research Institute and Department of Radiology, Northwestern University, Evanston, IL 60208 USA.

G. Fan is with the School of Electrical and Computer Engineering, Oklahoma State University, Stillwater, OK 74078 USA.

M. Rao is with the Department of Geography, Oklahoma State University, Stillwater, OK 74078 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2008.916832

II. Background

A. SVMs

The TCSVM learning constructs a classification hyperplane to maximally separate the two classes in the feature space. Since the TCSVM is a supervised learning algorithm, it cannot be directly used in the situations where no training prototypes are provided. The OCSVM is an extension of the TCSVM, and is applicable to both supervised and unsupervised learning [5], [6]. It seeks an approximation function that embraces a majority of data. There are two OCSVM implementations. One finds a hypersphere with minimum volume to contain a majority of feature points [5]. The other separates a majority of feature points from the origin using a hyperplane with a maximum margin [6]. Given n training samples $\mathbf{x}_i, i = 1, \ldots, n$ with their class labels y_i, the hyperplane is constructed by solving

$$\min_{\mathbf{w}\in F, \xi\in R^n, \rho\in R} \frac{1}{2}\|\mathbf{w}\|^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i \tag{1}$$

subject to: $y_i(\mathbf{w}\cdot\mathbf{x}_i) \geq \rho - \xi_i$, where F indicates the feature space, and the margin size is $\rho/\|\mathbf{w}\|$. The parameter $\nu \in (0, 1]$ that controls the upper bound percentage of outliers usually has to be given or estimated. The slack variable ξ_i allows training errors for nonseparable cases. Kernel methods are often used in OCSVM and TCSVM to project the original input data into a high dimensional feature space, where a linear classification is equivalent to a nonlinear classification in the input space. Radial basis function (RBF) is one of the commonly used kernels, and it is defined as $k(x, x_i) = e^{-\gamma\|x-x_i\|^2}$, where x_i is the kernel center, and γ is the width parameter. A large γ corresponds to a small kernel width. The second OCSVM implementation with the RBF kernel is used in this paper.

B. Training Data Refinement Techniques

The purity of training data is very important to the classification performance. Some preprocessing techniques are necessary to remove possible outliers in the training data, such as bootstrap, condensing, and editing. Four bootstrap techniques were discussed in [7] where each original training sample is replaced by a bootstrap sample generated by the weighted average or local mean of its nearest neighbors. They reduce the effect of outliers during training and also bring the training samples closer to the mean of each class. The bootstrap techniques have been combined with the SVM to improve the classification performance [10], but they may not be applicable in the CRP-related applications because there could be significant outliers in training data when CRP tracts are not fully compliant or have expired.

The condensing technique proposed in [8] aims in finding a minimal subset of the original training data that is able to classify the data without sacrificing the performance. The editing technique tries to remove mislabeled patterns from initial training data to improve the classification accuracy [9]. Data editing can reduce training data size and improve classification performance. Here we propose an OCSVM-based data editing method for the one-class classification where training data contains significant outliers.

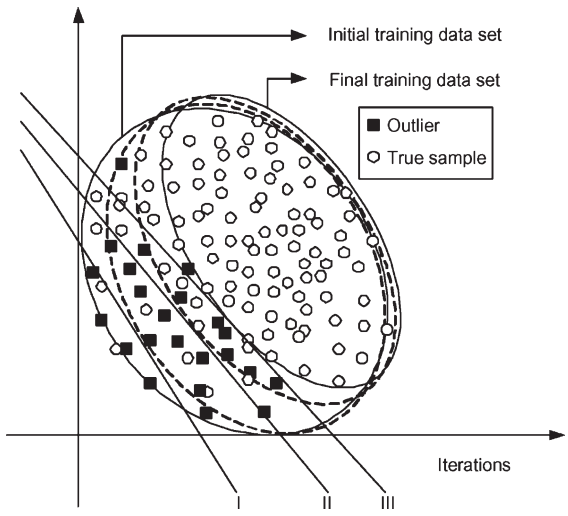

Fig. 1. Training data refinement in three iterations.

III. SVM-Based Editing and One-Class Classification

A. Proposed Method

We propose a new OCSVM-based editing and classification method that trims the training data by recursively removing the samples likely to be outliers. Let $\mathcal{R}$ be the initial training data and parameter ν of the OCSVM is assigned with ν_0 of a small value that determines the trimming step. The proposed algorithm is described as follows.

1) $\mathcal{S}_0 = \mathcal{R}$ is the initial training data set.
2) For $i = 1$ to M.
   a) The ith OCSVM is learned from $\mathcal{S}_{i-1}$ with $\nu = \nu_0$.
   b) The samples classified as outliers, denoted by $\mathcal{U}_i$, are removed to construct $\mathcal{S}_i = \mathcal{S}_{i-1} - \mathcal{U}_i$.
   c) The percentage of removed outliers is computed as $\eta_i = \sharp(\mathcal{U}_i)/\sharp(\mathcal{S}_{i-1})$ ($\sharp$ is the number of samples).

   End
3) Find $k = \arg\min_i \eta_i$, and use the kth OCSVM learned from $\mathcal{S}_{k-1}$ as the classifier if the test data is $\mathcal{R}$, or apply S_k to train a new OCSVM with an appropriate RBF kernel width if the test data contains many new classes.

The proposed algorithm can be visualized in Fig. 1 where the given data set contains significant outliers. If the target and outlier classes do not significantly overlap in the feature space, when applying the proposed method iteratively with a sufficiently small ν_0, we may obtain a minimum outlier percentage when the hyperplane gradually moves and gets close to the optimal separation boundary in terms of minimum classification error. Therefore, we set M iterations in the proposed method where M is upper bounded by $0.5/\nu_0$ (the amount of outliers is assumed to be less than half), and find the optimal iteration k with the minimum η_i. By trimming the outliers iteratively, the original training data are greatly purified, and the true outlier percentage could be estimated by aggregating removed outliers. It is worth noting that ν_0

should be much smaller than the outlier proportion to gradually move the learned hyperplane toward the optimal one. Too large values may run the risk of mislocating the cluster boundary and overtrimming the training data. Since the outlier proportion is usually unknown, we conservatively set the ν_0 value between 1%–5% in this paper.

The RBF kernel defines the structure of feature space where the classification hyperplane is constructed. When the kernel width is too small, every training point becomes a support vector, and the SVM overfits the training data. On the other hand, if the kernel width is too large, not enough training points are selected as support vectors, and the SVM underfits the training data. Typically, if the distributions of training data and testing data are similar, we can use the same kernel width for both data editing and classifier training. However, if the testing data includes new classes that do not exist in the training data, the basic assumption of SVM learning that the training and testing data follow a same unknown distribution might not hold [4], and the kernel width used for editing may not be appropriate for classification. This practical issue has been addressed in different contexts and applications [14], [15]. In this project, the data editing is performed on CRP tracts that contains relatively simple classes, while the testing is performed on the whole study area that includes many new classes. To avoid significant misdetections, we can implicitly assign a higher cost to false positive by decreasing the RBF kernel width (i.e., increase γ) to increase the system fitness to the target class when training the OCSVM based on the edited training data. It is expected that the proposed method can remove most outliers in the training data and improve the classification performance.

B. Simulation Results on Random Data

A 2000-sample random data set was generated where each sample is represented by a 30-D feature vector and each dimension is composed of a normally distributed random number. The target class samples have a mean value of 2.56 whereas outliers (25.2%) have a zero mean. Three criteria were used for objective evaluation: 1) accuracy (P_{a}) is the percentage of samples that are correctly classified in terms of target and outlier classes; 2) precision (P_{b}) indicates the percentage of detected target samples that are true ones; and 3) recall (P_{c}) is the percentage of true target samples that can be detected. The general OCSVM using bootstrap samples was used as the benchmark. Two bootstrap methods were used here. Bootstrapping II generates bootstrap samples by computing a weighted average of nearest neighbors and Bootstrapping IV uses the mean of nearest neighbors to get a bootstrap sample [7]. The bootstrap samples were classified by OCSVM where the outlier percentage was assumed to be known $(\nu = 0.25)$. The LIBSVM package [16] was used to implement OCSVM.

In the proposed method, the outlier percentage is assumed to be unknown. The RBF kernel width was set to be $\gamma = 0.03$. $\nu_0 = 0.05$ was used in each iteration. The minimum percentage of detected outliers occurs in the sixth iteration, with approximately 26.4% totally removed outliers. The trained OCSVM in this iteration was applied to classify "targets" from "outliers" in the original data set. As shown in Table I, although the proposed method was implemented in an unfavorable condition (without knowing the true outlier proportion), it still achieves higher P_{a} and P_{b} than the general OCSVM using bootstrap samples. Additionally, using a computer with a Pentium IV CPU and 384 MB of memory, the proposed method only needs several seconds to get final results, while the bootstrap methods spend about 2 min to generate 2000 bootstrap samples.

TABLE I
COMPARISON ON THE RANDOM DATA

	Accuracy (P_a)	Precision (P_b)	Recall (P_c)
OCSVM + Bootstrapping II	74.65	83.03	83.09
OCSVM + Bootstrapping IV	74.6	82.93	83.16
Edited OCSVM (ν_0 = 0.05)	75.1	83.95	82.49

C. Simulation Results on Synthetic Mosaics

The two methods were also tested on two synthetic mosaics shown in Fig. 2(a). One has 25% outliers, and the other has 8.41% outliers, shown as the ground truth images in Fig. 2(b). The pixelwise 25-D feature vectors were extracted using the method proposed in [17]. The RBF kernel width was set to be $\gamma = 0.1$. Fig. 2(c) and (d) represent the classification results of the general OCSVM (given the true outlier percentages) using samples generated by Bootstrapping II and IV, and Fig. 2(e) shows the classification result of the proposed method with unknown outlier percentage. For the first mosaic, $\nu_0 = 0.05$ $(\nu_0 \ll 0.25)$ was used in each iteration of the proposed method. The minimum percentage of outlier detection occurs in the sixth iteration as shown in Fig. 3(a). The correspondingly trained OCSVM was applied to classify the mosaic. The comparison of numerical results in Table II indicates that the proposed method outperforms the general OCSVM using bootstrap samples. For the second mosaic with 8.41% outliers, $\nu_0 = 0.02$ $(\nu_0 \ll 0.08)$ was used in each iteration of the proposed method, and the minimum percentage of outlier detection was found in the sixth iteration as shown in Fig. 3(b). The comparisons is listed in Table III, showing that the proposed method gets higher P_{a} and P_{b} with a little bit lower P_{c}.

The proposed method has two major advantages. First, although the outlier percentage is unknown, it can improve the classification performance by iteratively trimming training data, as long as ν_0 is sufficiently smaller compared with the true outlier percentage. Second, it is computationally more efficient compared with the conventional bootstrap techniques. Next, we will study a specific one-class remote sensing data analysis application (CRP mapping and compliance monitoring) where no perfect reference (training) data are available.

IV. EXPERIMENTS AND DISCUSSION

A. Study Area and Experimental Setup

The multiseasonal Landsat TM data (30 × 30 m) for Texas County, OK, obtained in February 2000 and June 2000 were used in the experiment. Fig. 4(a) shows the Landsat image of the study area in June 2000, superimposed by the CRP reference data acquired in the 1990s by the NRCS. Since the reference data were manually delineated from aerial

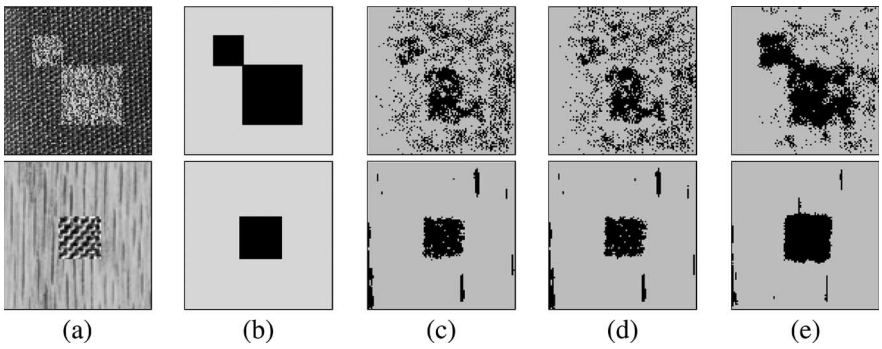

Fig. 2. (a) Two mosaics. (b) Ground truth with (top) 25% and (bottom) 8.41% outliers. (c) Bootstrapping II. (d) Bootstrapping IV. (e) Proposed method.

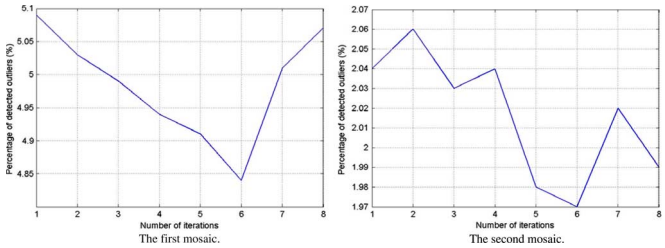

Fig. 3. Percentages of detected outliers versus iterations for two mosaics.

TABLE II
COMPARATIVE STUDIES ON THE FIRST MOSAICS

	Accuracy (P_a)	Precision (P_b)	Recall (P_c)
OCSVM + Bootstrapping II	74.46	85.51	81.38
OCSVM + Bootstrapping IV	73.90	85.10	81.06
Edited OCSVM (ν_0 = 0.05)	88.31	95.51	89.36

TABLE III
COMPARATIVE STUDIES ON THE SECOND MOSAICS

	Accuracy (P_a)	Precision (P_b)	Recall (P_c)
OCSVM + Bootstrapping II	95.41	97.49	97.48
OCSVM + Bootstrapping IV	95.35	97.45	97.47
Edited OCSVM (ν_0 = 0.02)	97.13	100.0	96.87

photographs and is not current, four types of errors could exist: 1) registration errors between the reference data and Landsat imagery; 2) unrecorded new CRP enrollments; 3) expired CRP tracts that returned to cultivation; 4) noncompliance issues in CRP tracts Error; 1), 2), and 3) are related to CRP mapping, and error 4) concerns compliance monitoring. The reference data provide initial training prototypes and a benchmark for algorithm evaluation.

A 35-D pixelwise feature vector was constructed to represent each pixel in the study area. The first ten dimensions consist of Landsat TM bands 2, 3, 4, 5, and 7. The next 20 dimensions are texture features including the local mean and local variance within a 3×3 window in each band of each season. The last five dimensions consist of several derived features, including normalized difference vegetation index, band ratios, and band differences [13]. There are 14 different CRP grass types in the study area, and they were regrouped into seven subclasses according to the spectral similarity. For each subclass, the initial training data are collected from the 35-D feature set according to the reference data.

The proposed method was compared with the general OCSVM using Bootstrapping II and IV to generate training data. When implementing the proposed method, we chose $\nu_0 = 0.02$ and $\gamma = 0.03$ for the SVM-based data editing, and $\gamma = 1.0$ was used for the training of the classifier. This process was applied to all seven subclasses, and the classification results were combined to obtain the final CRP map. The training data purification performs CRP compliance monitoring where all CRP tracts given in the reference data are refined recursively, and the final classification provides both the CRP compliance monitoring and mapping results. Therefore, two CRP-related tasks are jointly implemented in one flow. When applying the general OCSVM using bootstrap samples, ν was fixed to be 0.1 (i.e., assume each subclass has about 10% outliers) for OCSVM, and $\gamma = 1.0$ for the RBF kernel. Length filtering and morphological closing/opening operation were applied to the classification results of all approaches to produce clean segmentation maps.

B. Experimental Results

Fig. 4(b)–(d) shows the results obtained from the proposed method and the general OCSVM using bootstrap samples. Fig. 4(b) shows that some CRP tracts presented in the reference data were not detected, implying possible retired and/or noncompliant CRP tracts. Fig. 5 shows the mapping results of several clips where the top row is the June Landsat images overlaid by the original reference data, and the bottom row is the mapping and compliance monitoring results. Fig. 5(a)–(i) shows different CRP tracts that are correctly classified. Particularly, the proposed method can remove relatively small man-made structures in the CRP map, as shown in (b), (d), (g), (h), and (i). If our assumption that the dominant class (CRP) in the training data is contract compliant does not hold, then the proposed method could fail. For example, Fig. 5(j) represents a CRP tract where the non-CRP is the dominant class (the red color represents a non-CRP cultivation area) and the final CRP classification result is not valid. It is also observed that the general OCSVM significantly underdetected the CRP tracts when using bootstrap samples generated by Bootstrap-II and -IV as shown in Fig. 4(c) and (d), indicating that the proposed method is more effective for CRP mapping and compliance monitoring.

V. CONCLUSION

We have proposed an SVM-based data editing and one-class classification method that can be used to find the class of

Fig. 4. (a) CRP reference data superimposed on Landsat TM image (552 × 523). (b) CRP map generated by the proposed method. (c), (d) CRP maps obtained by the general OCSVM using bootstrap samples generated by Bootstrapping II, and IV. Significant underdetections are observed.

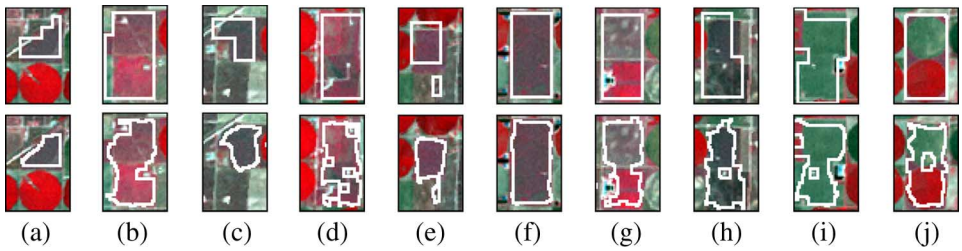

Fig. 5. Top row shows the June images overlaid by the original reference data, and the bottom row shows the CRP compliance monitoring and mapping results of the proposed method. (a)–(i) Classification results of different CRP tracts. (j) Misclassification of a non-CRP area.

interest in remotely sensed imagery. The major advantage of the proposed method is its ability to handle the unknown amount of outliers in the training data by the recursive sample refinement. Specifically, the proposed technique was applied to the mapping and compliance monitoring of CRP tracts where these two CRP-related tasks are integrated into one flow. The experimental results show that this method can automatically update CRP maps and identify noncompliant CRP tracts. Compared with manual delineation, it considerably reduces the time and labor cost, thereby facilitating CRP management tasks. The major limitation is the assumption that the class of interest is the majority in the training data may not always be true. This issue could be addressed by examining the validity of the training data for each CRP species prior to the SVM training or by preselecting reliable training samples for each class. Nevertheless, this approach provides an integrated tool for one-class remote sensing data analysis where the training data may be corrupted by a significant amount of outliers.

ACKNOWLEDGMENT

The authors would like to thank the anonymous referees and the associate editor whose comments and suggestions have significantly improved this paper.

REFERENCES

[1] F. Roli and G. Fumera, "Support vector machines for remote-sensing image classification," *Proc. SPIE*, vol. 4170, pp. 160–166, 2001.
[2] M. Brown, H. G. Lewis, and S. R. Gunn, "Linear spectral mixture models and support vector machines for remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 5, pp. 2346–2360, Sep. 2000.
[3] M. Pal and P. M. Mather, "Support vector machines for classification in remote sensing," *Int. J. Remote Sens.*, vol. 26, no. 5, pp. 1007–1011, 2005.
[4] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.
[5] D. M. J. Tax, "One-class classification," Ph.D. dissertation, Tech. Univ. Delft, Delft, The Netherlands, 2001.
[6] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, 2001.
[7] Y. Hamamoto, S. Uchimura, and S. Tomita, "A bootstrap technique for nearest neighbor classifier design," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 1, pp. 73–79, Jan. 1997.
[8] P. E. Hart, "The condensed nearest neighbor rule," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 5, pp. 515–516, May 1968.
[9] P. A. Devijver and J. Kittler, "On the edited nearest neighbor rule," in *Proc. 5th Int. Conf. Pattern Recog.*, 1980, pp. 72–80.
[10] X. Li, Y. Zhu, and E. Sung, "Sequential bootstrapped support vector machines—A SVM accelerator," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2005, vol. 3, pp. 1437–1442.
[11] X. Song, G. Fan, and M. Rao, "Automatic CRP mapping using nonparametric machine learning approaches," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 888–897, Apr. 2005.
[12] G. Cherian, X. Song, G. Fan, and M. Rao, "Application of support vector machines for automatic compliance monitoring of the conservation reserve program (CRP) tracts," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Sep. 2004, pp. 242–245.
[13] X. Song, G. Cherian, and G. Fan, "A ν-insensitive SVM approach for automatic compliance monitoring of the conservation reserve program (CRP) tracts," *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 2, pp. 99–103, Apr. 2005.
[14] J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Schölkopf, *Correcting Sample Selection Bias by Unlabeled Data*, vol. 19. Cambridge, MA: MIT Press, 2007, pp. 601–608.
[15] G. Wu and E. Chang, "KBA: Kernel boundary alignment considering imbalanced data distribution," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 786–795, Jun. 2005.
[16] C. C. Chang and C. J. Lin, *LIBSVM: A Library for Support Vector Machines*, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/cjlin/libsvm
[17] K. I. Kim, K. Sung, S. H. Park, and H. J. Kim, "Support vector machines for texture classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 11, pp. 1542–1550, Nov. 2002.