A Semilabeled-Sample-Driven Bagging Technique for Ill-Posed Classification Problems

Mingmin Chi and Lorenzo Bruzzone, *Senior Member, IEEE*

***Abstract*—In this letter, a semilabeled-sample-driven bootstrap aggregating (bagging) technique based on a co-inference (inductive and transductive) framework is proposed for addressing ill-posed classification problems. The novelties of the proposed technique lie in: 1) the definition of a general classification strategy for ill-posed problems by the joint use of training and semilabeled samples (i.e., original unlabeled samples labeled by the classification process); and 2) the design of an effective bagging method (driven by semilabeled samples) for a proper exploitation of different classifiers based on bootstrapped hybrid training sets. Although the proposed technique is general and can be applied to any classification algorithm, in this letter multilayer perceptron neural networks (MLPs) are used to develop the basic classifier of the proposed architecture. In this context, a novel cost function for the training of MLPs is defined, which properly considers the contribution of semilabeled samples in the learning of each member of the ensemble. The experimental results, which are obtained on different ill-posed classification problems, confirm the effectiveness of the proposed technique.**

***Index Terms*—Bagging, ill-posed classification problems, multiple classifier systems, remote sensing images, semilabeled samples, supervised classification.**

I. Introduction

ONE OF THE most challenging issues in the supervised classification of remote sensing images lies in the solution of "ill-posed" problems [1], [2]. These problems are characterized by the availability of small-size training sets with respect to the high-dimensional input feature space and/or the large number of parameters in the classifier model. Since the collection of ground-reference data in real-world applications is an expensive and time-consuming task, in many cases the number of training samples is not sufficient for a proper learning of the classification system. This is particularly critical when considering multisensor and multisource datasets or hyperspectral images, because due to the intrinsic large dimension of the feature space, it is not possible to meet the requirements on the necessary number of training samples.

According to the remote sensing literature, one possible way of addressing ill-posed classification problems is to include semilabeled samples (i.e., samples that were originally unlabeled and were later labeled by the classification process) in the training set by using specific iterative procedures [3]–[5]. Nevertheless, the use of semilabeled samples does not guarantee good performance, since the resulting accuracy is strongly affected by the accuracy of selected semilabeled samples. In this context, in order to increase the reliability of the transductive process, we propose to use a proper ensemble method based on bagging (short for bootstrap aggregating, proposed by Breiman [6]). In the standard bagging algorithm, many subsets made up of M samples are resampled from the original training set to generate an ensemble of base classifiers to be included in the multiple classifier system. However, in ill-posed problems, standard bagging cannot be used directly because it is difficult to resample the available small-size original training set to derive classifiers to be included in the ensemble.

In this letter, we present a novel semilabeled-sample-driven bagging technique developed in the context of a co-inference process (induction and transduction) for solving ill-posed classification problems. The novelties of the proposed technique lie in the following:

- The definition of a general ensemble-based architecture for ill-posed problems by the joint use of training and semilabeled samples.
- The design of an effective bagging method (driven by semilabeled samples) for a proper definition of different classifiers based on bootstrapped hybrid sets.

Although the proposed approach is general (i.e., classifier independent), it is developed in this letter using multilayer perceptron neural networks (MLPs) [7], [8]. In this context, a cost function for the training of MLPs is defined, which properly considers the contribution of semilabeled samples (associated with the transductive process) in the learning of each member of the ensemble.

In order to assess the effectiveness of the proposed technique, many simulated ill-posed classification problems have been defined using multispectral Landsat Thematic Mapper images acquired on the Trentino area (Italy). Experimental results confirm the capabilities of the presented method to increase both the accuracy and the robustness of the classification in small-size classification problems.

This letter is organized in four sections. In Section II, the proposed semilabeled-sample-driven bagging approach is presented. Section III describes the dataset used in the experiments and the results obtained with the proposed approach. Finally, Section IV draws the conclusions of this work and discusses future developments.

Manuscript received September 29, 2004; revised November 23, 2004. This work was supported by the Italian Ministry of Education, University and Research.

The authors are with the Department of Information and Communication Technology, University of Trento, 38050 Trento, Italy (e-mail: lorenzo.bruzzone@ing.unitn.it).

Digital Object Identifier 10.1109/LGRS.2004.841478

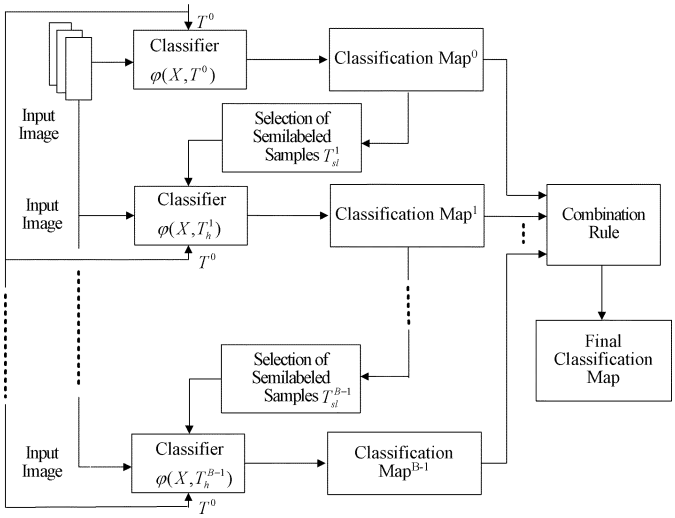

Fig. 1. Block scheme of the proposed bagging approach driven by semilabeled samples.

II. Proposed Semilabeled-Sample-Driven Bagging Approach to Ill-Posed Problems

A. Notation and Problem Definition

Let X be a d-dimensional feature vector, and $\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$ be the set of c land-cover classes that characterize the considered problem. Let T be a training set made up of N labeled samples. Given the pattern X, we can define a procedure for using the training set T to form a classifier $\varphi(X,T)$. As in the bagging algorithm, B subsets $T^b (b = 0, \ldots, B-1)$ of M bootstrapped samples are generated, and a classifier $\varphi(X,T^b)$ is built from each subset T^b. A final classification is obtained by an ensemble rule to achieve a better (more accurate and reliable) classification result than the single classifier $\varphi(X,T)$, i.e.,

$$X \in \omega_m, \omega_m \in \Omega,$$
$$\text{if and only if } \omega_m = \arg\max_{\omega_i \in \Omega} \left\{ \text{the number of } \left[\varphi(X,T^b) = \omega_i\right], \right.$$
$$\left. b = 0, \ldots, B-1 \right\}. \quad (1)$$

However, in small-size training set classification, there are no enough training samples to be bootstrapped; hence, the standard bagging algorithm cannot be directly exploited to solve ill-posed classification problems. Nevertheless, semilabeled samples can be considered similar to labeled samples to some extent and be included in the subsets of bootstrapped samples to derive corresponding base classifiers.

B. Semilabeled-Sample-Driven Bagging Technique

In the proposed approach, training sets exploited in individual base classifiers (excluding the initial classifier) are bootstrapped by the use both of all the training samples and of a subset of semilabeled patterns (in order to mitigate the problem of the small-size training set). For this reason, we define these sets as "hybrid training sets." The architecture of the proposed system is shown in Fig. 1, where $T_{\text{sl}}^b (b = 1, \ldots, B-1)$ refers to the subset of semilabeled samples injected in the hybrid training set T_h^b (which also includes the small-size original training set T^0) used for the bth classifier $\varphi(X,T_h^b)$.

TABLE I
Distribution of Training and Test Patterns in the Six Simulated Ill-Posed Classification Problems Considered

Land-cover classes	Number of test pixels	Number of training pixels					
		Set 1	*Set 2*	*Set 3*	*Set4*	*Set5*	*Set 6*
Conifers	1155	3	7	10	15	18	38
Trees	681	2	5	8	10	12	25
Grass	336	2	5	6	8	10	19
Water	84	1	1	1	1	2	3
Urban	104	1	1	2	2	3	5
Rocks	113	1	1	3	4	5	10
Overall	2473	10	20	30	40	50	100

The proposed architecture includes an initial classifier $\varphi(X,T^0)$, which only exploits the training set T^0 to generate the initial classification map (according to a standard inductive process). In this way, a "pseudolabel" is assigned to each unlabeled sample according to a parametric (or nonparametric) algorithm. Thus unlabeled samples become semilabeled, since the class label information is partially obtained. Then, the generic bth classifier $\varphi(X,T_h^b)$ of the architecture is defined by selecting a subset of semilabeled samples from the classification map of the previous classifier (this represents a transductive process), i.e., the bth hybrid training set is defined as

$$T_h^b = T^0 \cup T_{\text{sl}}^b. \quad (2)$$

This transductive process is iterated until the desired number of classifiers included in the ensemble is obtained. Finally, like in standard bagging, all the classification maps $(\text{Map}^0, \text{Map}^1, \ldots, \text{Map}^{B-1})$ are integrated by an ensemble rule to obtain the final classification map. A key issue in the proposed approach is the strategy adopted to generate the subset T_{sl}^b of semilabeled samples. In order to obtain samples representing the true class distribution in the whole image, a random selection strategy (based on a uniform spatial distribution) is applied to select semilabeled samples included in T_{sl}^b according to the classification results obtained from the classifier $\varphi(X,T_h^{b-1})$.

One of the main problems in the use of semilabeled samples is the risk of defining a negative iterative mechanism, in which unlabeled samples (transductive process) degrade the inductive learning of the classifiers. The negative mechanism may happen if the accuracy of the supervised classifier at the first iteration is below a given threshold. Nonetheless, if the accuracy is sufficiently high, the probability that this event should occur is small. This probability is decreased in the case of an ensemble of classifiers, since the ensemble proves to be robust to the presence of weak classifiers.[1] The following additional strategies can be used to mitigate the above drawback further:

1) exploiting the "confidence" associated with the classification label of each pixel to generate the semilabeled sample pool depending on the reliability of classification results (a proper threshold value on the confidence can be set);
2) using a weighing scheme in the classification algorithm, where the confidence of semilabeled samples is considered explicitly in the transductive learning process of the classifiers.

[1]It is worth noting that we expect classifiers based on different semilabeled samples to result both in different overall accuracies and in quite uncorrelated classification errors. This latter property increases the reliability of the multiple classifier system.

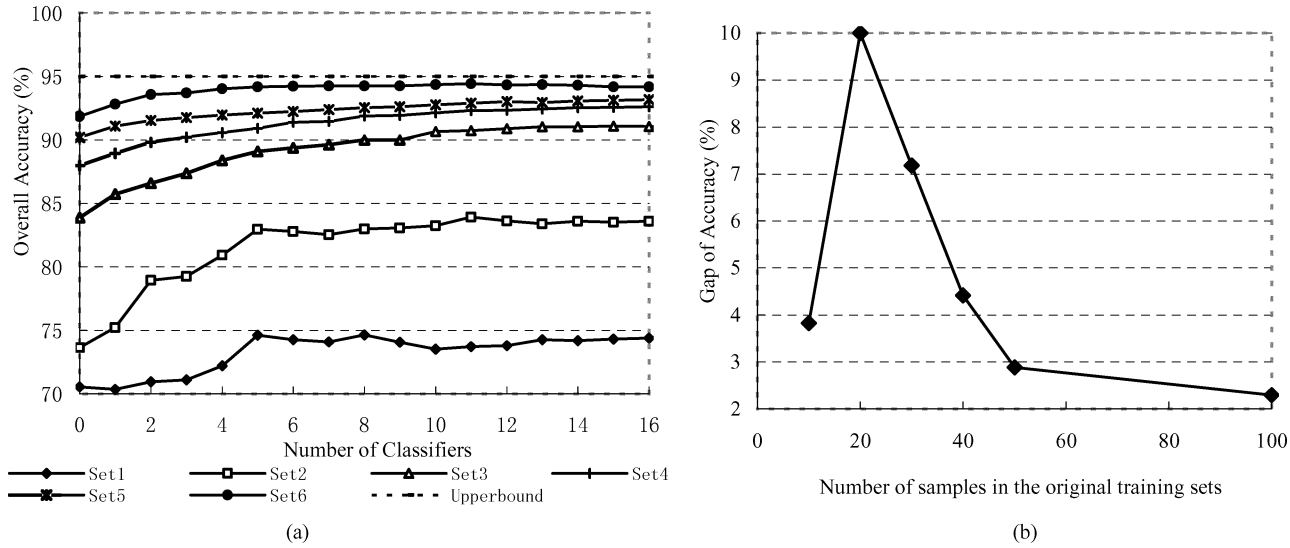

Fig. 2. (a) Overall classification accuracy (average on three trials) versus the number of classifiers included in the ensemble for the different simulated datasets considered (upper bound refers to overall accuracy with the original training set made up of 4549 samples). (b) Gap of accuracy between the standard supervised classifier (obtained by standard MLP_0 on the initial training set) and the proposed approach (applied using 17 classifiers) versus the size of original training sets.

Once classifier members included in the ensemble are defined, as in standard bagging the final classification map can be achieved following any standard combination rule (e.g., the majority voting scheme [6]).

C. Semilabeled-Sample-Driven Bagging With MLPs

In this study, we used a distribution-free technique based on multilayer-perceptron neural networks as the base classification algorithm to define the multiple classifier system (for greater details on MLPs, the reader can refer to [7]). This choice is motivated by the ability of MLPs to deal with any kind of remote sensing datasets (multisource, multisensor, etc.) without any constraint on the model of data distribution. However, it is worth noting that any classification technique can be used in the proposed architecture.

MLPs usually are trained using the error backpropagation (EBP) learning algorithm [7] applied to a proper cost function. The most widely used cost function in the standard inductive process is the MSE, which is given by

$$\text{MSE}(W) = \frac{1}{2}\sum_{j=1}^{N_{tr}}\sum_{i=1}^{c}(t_{ji} - o_{ji})^2 \tag{3}$$

where N_{tr} is the total number of training patterns, c is the number of classes, t_{ji} and o_{ji} represent the target and the network output for the jth training sample of the ith class, respectively, and W is the set of all the weights of the network. This procedure is used at the first iteration of the proposed method. After the initial classification, semilabeled samples are introduced in the training phases of all the other members of the ensemble. To this purpose, the following modified cost function is defined:

$$\text{MSE}^b(W) = \frac{1}{2}\sum_{j=1}^{N_{\text{tr}}}\sum_{i=1}^{c}(t_{ji} - o_{ji})^2 + \frac{1}{2}\sum_{j=1}^{N_{\text{sl}}}\sum_{i=1}^{c}\left(t_{ji}^b - o_{ji}^b\right)^2 q_j^b, \quad b = 1, \ldots, B-1. \tag{4}$$

where N_{sl} is the total number of randomly selected semilabeled patterns, t_{ji}^b and o_{ji}^b represent the target and the network output of the considered semilabeled pattern in the bth MLP classifier, respectively. q_j^b is a weight value for the given semilabeled pattern (which represents the reliability of the label of the jth semilabeled pattern in the bth classifier). Its value can be derived directly from the output of the $(b-1)$th MLP neural network and can be considered an estimation (optimized according to a minimum square error criterion) of the conditional posterior probability $p(\omega_i|X)$ to have the label ω_i given the jth pattern [8].

III. Experimental Results

This section reports the experimental results obtained by the proposed semilabeled-sample-driven bagging technique. The dataset considered is made up of a Landsat-5 Thematic Mapper image acquired on the Trentino area (northern Italy). For this dataset, a large number of labeled samples were collected from ground-reference data. The labeled samples were divided into a training set and a test set. In order to simulate ill-posed classification problems, subsampling (with different rates) was applied to the training set. In greater detail, from 4549 original training patterns, 10, 20, 30, 40, 50, and 100 training samples were randomly selected (see Table I) by maintaining as far as possible the prior probabilities of classes of the whole training set (with the constraint of having at least one sample for each class). In all datasets, seven features and six land-cover classes were considered in the analysis. Hence, we are clearly in the presence of an ill-posed complex classification problem (e.g., when the size of the training set is 10, 20, 30, and 40, the minority class has only one training pattern).

In order to assess the effectiveness of the proposed semilabeled-sample-driven bagging approach, we considered the 2473 test samples as unlabeled patterns, so that after classification semilabeled samples are randomly extracted from them. Three trials were carried out for each training set (with different sizes), and then the average overall accuracy was computed.

TABLE II
OVERALL ACCURACIES OBTAINED BY A STANDARD CLASSIFIER AND BY THE PROPOSED APPROACH WITH THE DIFFERENT SIZES OF ORIGINAL TRAINING SETS

Technique	Overall Accuracy (%)					
	Set1	*Set2*	*Set3*	*Set4*	*Set5*	*Set6*
Standard	70.56	73.64	83.87	87.99	90.27	91.87
Proposed	74.39	83.64	91.05	92.61	93.15	94.16

In all experiments, a three-layer neural network architecture with seven input units and six output units was used as the base classifier for the proposed bagging approach. Considering the small-size training set, few hidden units (i.e., 2, 3, and 4 for different-size training sets) were used in the proposed architecture in order to avoid overfitting. Although this leads to weak classifiers, we expect an ensemble made up of many weak classifiers (which, being based on different semilabeled samples, are expected to incur in uncorrelated errors) to provide sufficiently high classification accuracies. In all trials, twice as many semilabeled samples (whose estimated conditional posterior probabilities were greater than 0.85) as training samples were randomly selected for defining the hybrid training sets. The weights q_j^b in (4) were set to the values of the conditional posterior probabilities estimated from the MLP classifier.

Fig. 2(a) shows the behavior of overall classification accuracy versus the number of classifiers included in the ensemble for the different simulated datasets. For comparison purposes, the upper bound of the accuracy on the test set (obtained using all 4549 training patterns originally available) is also given. On analyzing the diagram, it can be observed that in general the proposed semilabeled-sample-driven bagging significantly increased overall classification accuracy. If we consider the best case (i.e., *set* 2 defined by 20 training patterns), overall accuracy obtained on the test set using only initial labeled samples was 73.64%, while the accuracy provided by the ensemble with 17 MLP classifiers was 83.64% (the accuracy sharply increased by 10%) (see Table II). More in general, on analyzing Fig. 2(b), which reports the gap of the accuracy (between overall accuracy of the standard MLP_0 classifier and that of the proposed approach) versus the size of original training sets, we can observe a different behavior of the proposed method for training sets with different sizes. In greater detail, the presented method increased the classification accuracy in all simulated datasets. However, the improvement on overall accuracy is higher when few training samples are considered (except for the case with only ten samples). This is as expected, i.e., the proposed approach is more effective when the Hughes phenomenon [9] becomes more critical. As regards *set* 1 (which resulted in an increase in accuracy of 3.83%), the small gain depends on the fact that original training samples were too few. This cannot allow the ensemble to capture and model the complexity of the classification problem. It is worth noting that in all cases, few classifiers were sufficient to obtain the convergence in the classification accuracy.

In order to analyze the influence of semilabeled samples on classification results further, Fig. 3 shows class-by-class user and producer accuracies obtained in *set* 2 (made up of 20 training samples). As can be seen, a significant improvement

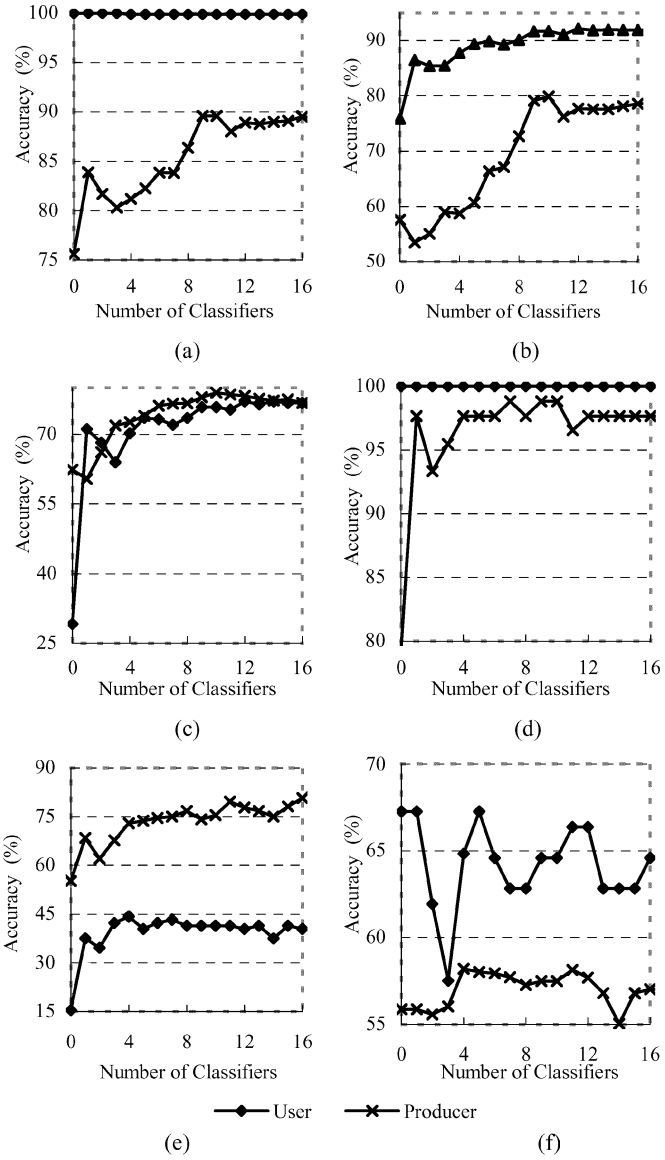

Fig. 3. Class-by-class user and producer accuracies versus the number of classifiers included in the ensemble when the size of the original training set is 20. (a) Conifer class. (b) Tree class. (c) Grass class. (d) Water Class. (e) Urban class. (f) Rock class.

(in both user and producer accuracies) was obtained for all classes except rock. The class with the highest improvement on user accuracies is grass, where a sharp increase of 46.72% can be seen. In all the other classes (with the exception of rock), user and producer accuracies increase significantly in a range of between 15% and 22%. This confirms the effectiveness of the proposed technique.

IV. DISCUSSION AND CONCLUSION

A semilabeled-sample-driven bagging technique for the classification of remote sensing data in small-size training set problems has been proposed. In particular, an architecture made up of an ensemble of classifiers has been presented, whose members are defined according to different bootstrapped hybrid training sets, made up of a balanced number of labeled and semilabeled patterns. Unlike other methods, the proposed

semilabeled-sample-driven methodology is based on a general architecture where classifiers included in the ensemble can be implemented with any kind of parametric or nonparametric classification technique, without requiring any specific constraint on the model of training data distributions. Experimental results obtained on multispectral remote sensing data (in the context of simulated ill-posed classification problems) confirmed the effectiveness of the proposed approach. In particular, this approach sharply increased both the accuracy and the stability of the obtained classification results.

As a final remark, it is worth pointing out that the use of a very few number of unrepresentative training patterns may also lead to a negative mechanism, which may degrade the accuracy of the proposed classification technique. However, we expect this situation to correspond to very poor initial training sets, which cannot be used on any reliable supervised or semisupervised classification procedure (in other words, in these extreme cases, it seems that the classification problem cannot be solved with any supervised or semisupervised nonparametric approach). In order to investigate this problem further, as a future developments of this work we plan to carry out an intensive experimental analysis on different datasets by simulating different initial training conditions with different numbers of training samples in order to estimate the probability of success of the proposed approach statistically. In addition, special attention will be devoted to the use of other classifiers to test the proposed method, by considering more stable techniques intrinsically robust to ill-posed problems (e.g., support vector machine classifiers [10]).

REFERENCES

[1] Q. Jackson and D. A. Landgrebe, "Adaptive Bayesian contextual classification based on Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 11, pp. 2454–2463, Nov. 2002.

[2] M. Chi and L. Bruzzone, "An ensemble-driven k-NN approach to ill-posed classification problems," in *Proc. 3rd Int. Workshop Pattern Recognition in Remote Sensing (PRRS'04)*, Kingston upon Thames, U.K., Aug. 2004.

[3] B. M. Shahshahani and D. A. Landgrede, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 5, pp. 1087–1092, Sep. 1994.

[4] M. T. Fardanesh and O. Ersoy, "Classification accuracy improvement of neural network classifiers by using unlabeled data," *IEEE Trans. Geosci. Remote Sens.*, vol. 36, no. 3, pp. 1020–1025, May 1998.

[5] S. Tadjudin and D. A. Landgrebe, "A decision tree classifier design for high-dimensional data with limited training samples," in *Proc. IGARSS*, vol. 1, May 1996, pp. 790–792.

[6] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 26, no. 2, pp. 123–140, 1996.

[7] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Clarendon, 1995.

[8] L. Bruzzone and S. B. Serpico, "Classification of imbalanced remote-sensing data by neural networks," *Pattern Recognit. Lett.*, vol. 18, pp. 1323–1328, 1997.

[9] G. F. Hughes, "On the mean accuracy of statistical pattern recognition," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 55–63, Jan. 1968.

[10] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, 1998.