Feature Selection for Classification of Hyperspectral Data by SVM

Mahesh Pal and Giles M. Foody, *Member, IEEE*

***Abstract*—Support vector machines (SVM) are attractive for the classification of remotely sensed data with some claims that the method is insensitive to the dimensionality of the data and, therefore, does not require a dimensionality-reduction analysis in preprocessing. Here, a series of classification analyses with two hyperspectral sensor data sets reveals that the accuracy of a classification by an SVM does vary as a function of the number of features used. Critically, it is shown that the accuracy of a classification may decline significantly (at 0.05 level of statistical significance) with the addition of features, particularly if a small training sample is used. This highlights a dependence of the accuracy of classification by an SVM on the dimensionality of the data and, therefore, the potential value of undertaking a feature-selection analysis prior to classification. Additionally, it is demonstrated that, even when a large training sample is available, feature selection may still be useful. For example, the accuracy derived from the use of a small number of features may be noninferior (at 0.05 level of significance) to that derived from the use of a larger feature set providing potential advantages in relation to issues such as data storage and computational processing costs. Feature selection may, therefore, be a valuable analysis to include in preprocessing operations for classification by an SVM.**

***Index Terms*—Classification accuracy, feature selection, Hughes phenomenon, hyperspectral data, support vector machines (SVM).**

I. Introduction

PROGRESS in hyperspectral sensor technology allows the measurement of radiation in the visible to infrared spectral region in many finely spaced spectral features or wavebands. Images acquired by these hyperspectral sensors provide greater detail on the spectral variation of targets than those acquired by conventional multispectral systems, providing the potential to derive more information about different objects in the area imaged [1]. Analysis and interpretation of data from these sensors present new possibilities for applications such as land-cover classification [2]. However, the availability of large amounts of data also represents a challenge to classification analyses. For example, the use of many features may require the estimation of a considerable number of parameters during the classification process [3]. Ideally, each feature (e.g., spectral waveband) used in the classification process should add an independent set of information. Often, however, features are highly correlated, and this can suggest a degree of redundancy in the available information which may have a negative impact on classification accuracy [4].

One problem often noted in the classification of hyperspectral data is the Hughes effect or phenomenon. The latter can have a major negative impact on the accuracy of a classification. The key characteristics of the phenomenon, assuming a fixed training set, may be illustrated for a typical scenario in which features are incrementally added to a classification analysis. Initially, classification accuracy increases with the addition of new features. The rate of increase in accuracy, however, declines, and eventually, accuracy will begin to decrease as more features are included. Although it may at first seem counterintuitive for the provision of additional discriminatory information to result in a loss of accuracy, the problem is often encountered [5]–[7] and arises as a consequence of the analysis requiring the estimation of more parameters from the (fixed) training sample. Thus, the addition of features may lead to a reduction in classification accuracy [8].

The Hughes phenomenon has been observed in many remote sensing studies based upon a range of classifiers [3], [5], [9], [10]. For example, a parametric technique, such as the maximum likelihood classifier, may not be able to classify a data set accurately if the ratio of sample size to number of features is small, as it will not be able to correctly estimate the first- and second-order statistics (i.e., mean and covariance) that are fundamental to the analysis [6]. Note that, with a fixed training set size, this ratio declines as the number of features is increased. Thus, two key attributes of the training set are its size and fixed nature. If, for example, the training set was not fixed but was instead increased appropriately with the addition of new features, the phenomenon may not occur. Similarly, if the fixed training set size was very large so that even when all features of a hyperspectral sensor were used, the Hughes effect may not be observed as all parameters may be estimated adequately. Unfortunately, however, the size of the training set required for accurate parameter estimation may exceed that available to the analyst. Given that training data acquisition may be difficult and costly [11]–[13], some means to accommodate the negative issues associated with high-dimensional data sets are required.

Various approaches could be adopted for the appropriate classification of high-dimensional data. These span a spectrum from the adoption of a classifier that is relatively insensitive to the Hughes effect [14] through the use of methods to effectively increase training set size [5], [11] by the application of some form of dimensionality-reduction procedure prior to the

Manuscript received May 12, 2009; revised September 9, 2009. First published February 22, 2010; current version published April 21, 2010. The work of Dr. Pal was supported by the Association of Commonwealth Universities with a fellowship at the University of Nottingham carried out during the period October 2008–March 2009.

M. Pal is with the National Institute of Technology, Kurukshetra 136 119, India (e-mail: mpce_pal@yahoo.co.uk).

G. M. Foody is with the School of Geography, University of Nottingham, NG7 2RD Nottingham, U.K. (e-mail: giles.foody@nottingham.ac.uk).

Digital Object Identifier 10.1109/TGRS.2009.2039484

classification analysis. It may also sometimes be appropriate to use a combination of approaches to reduce the possibility of the Hughes effect being observed. The precise approach adopted may vary with study objectives, data sets, and classification approach. One classification method that has been claimed to be independent of the Hughes effect and so promoted for use with hyperspectral data sets is support vector machines (SVM) [15], although, as will be discussed later, there is some uncertainty relating to the role of feature reduction with this method.

The SVM has become a popular method for image classification. It is based on structural risk minimization and exploits a margin-based criterion that is attractive for many classification applications [16]. In comparison with approaches based on empirical risk, which minimize the misclassification error on the training set, structural risk minimization seeks the smallest probability of misclassifying a previously unseen data point drawn randomly from a fixed but unknown probability distribution. Furthermore, an SVM tries to find an optimal hyperplane that maximizes the margin between classes by using a small number of training cases, the support vectors. The complexity of SVM depends only on these support vectors, and it is argued that the dimensionality of the input space has no importance [15], [17], [18]. This hypothesis has been supported by a range of studies with SVM, such as those employing the popular radial basis function (RBF) kernel for land-cover classification applications [19]–[21].

The basis of the SVM and the results of some studies, therefore, suggest that SVM classification may be unaffected by the dimensionality of the data set and, therefore, the number of features used. However, other studies have shown that the accuracy of SVM classification could still be increased by reducing the dimensionality of the data set [22], [23]; hence, there is a degree of uncertainty over the role of feature reduction in SVM-based classification. Feature reduction, however, impacts on more than just the accuracy of a classification. A feature-reduction analysis may be undertaken for a variety of reasons. For example, it may speed up the classification process by reducing data-set size and may increase the predictive accuracy as well as ability to understand the classification rules [24]. It may also simply provide advantages in terms of reducing data-storage requirements. Feature reduction may, therefore, still be a useful analysis even if it has no positive effect on classification accuracy.

Two broad categories of feature-reduction techniques are commonly encountered in remote sensing: feature extraction and feature selection [25], [26]. With feature extraction, the original remotely sensed data set is typically transformed in some way that allows the definition of a small set of new features which contain the vast majority of the original data set's information. More popular, and the focus of this paper, are feature-selection methods. The latter aim to define a subset of the original features which allows the classes to be discriminated accurately. That is, feature selection typically aims to identify a subset of the original features that maintains the useful information to separate the classes with highly correlated and redundant features excluded from the classification analysis [25].

Feature-selection procedures are dependent on the properties of the input data as well as on the classifier used [27], [28]. These procedures require that a criterion be defined by which it is possible to judge the quality of each feature in terms of its discriminating power [29]. A computational procedure is then required to search through the range of potential subsets of features and select the "best" subset of features based upon some predefined criterion. The search procedure could simply consist of an exhaustive search over all possible subsets of features since this is guaranteed to find the optimal subset. In a practical application, however, the computational requirements of this approach are unreasonably large, and a nonexhaustive search procedure is usually used [30]. A wide variety of feature-selection methods have been applied to remotely sensed data [30]–[33]. Based on whether they use classification algorithms to evaluate subsets, the different methods can be grouped into three categories: filters, wrappers, and embedded approaches. These approaches may select different subsets, and these, in turn, may vary in suitability for use as a preprocessing algorithm for different classifiers. Because of these differences and the range of reasons for undertaking a feature selection, as well as the numerous issues that influence outputs and impact on later analyses, feature selection remains a topic for research [34].

Although the literature includes claims that classification by SVM is insensitive to the Hughes effect [19]–[21], [35], it also includes case studies using simulated data [36], [37] and theoretical arguments that indicate a positive role for feature selection in SVM classification [38], [39]. Both Bengio *et al.* [38] and Francois *et al.* [39] based their arguments on the use of local kernels, such as the popular RBF, with kernel-based classifiers in which the cases lying in the neighborhood of the case being used to calculate the kernel value have a large influence [40]. In their argument, Bengio *et al.* [38] used the bias-variance dilemma [41] to suggest that the classifiers with local kernel would require exponentially large training data set to have the same level of classification error in high-dimensional space as that in a lower space, suggesting the sensitivity of SVM classifier to the curse of dimensionality. On the other hand, Francois *et al.* [39] suggested that the locality of a kernel is an important property that makes the generated model more interpretable and used an algorithm more stable than the algorithms using global kernels. They argued that an RBF kernel loses the properties of a local kernel with increasing feature space, a reason why they may be unsuitable in high-dimensional space. With the latter, for example, it has been argued that classifiers using local kernels are sensitive to the curse of dimensionality as the properties of learned function at a case depends on its neighbors, which fails to work in high-dimensional space. There is, therefore, uncertainty in the literature over the sensitivity of classification by an SVM to the dimensionality of the data set and, therefore, of the value of feature selection within such an analysis. This paper aims to address key aspects of this uncertainty associated with the role of feature selection in the classification of hyperspectral data sets. Specifically, this paper aims to explore the relationship between the accuracy of classification by an SVM and the dimensionality of the input data. The latter will also be controlled through application of a series

of feature-selection methods and, therefore, also highlight the impact, if any, of different feature-selection techniques on the accuracy of SVM-based classification. Variation in the accuracy of classifications derived using feature sets of differing size will be evaluated using statistical tests of difference and noninferiority [42], [43] in order to evaluate the potential role of feature selection in SVM-based classification. This paper is, to our knowledge, the first rigorous assessment of the Hughes effect on SVM with hyperspectral data set. Other studies (e.g., [19]–[21]) have commented on the Hughes effect in relation to the SVM-based classification of remotely sensed data, but this paper differs in that the experimental design adopted gives an opportunity for the effect to occur (e.g., by including analyses based on small training sets), and the statistical significance of differences in accuracy is evaluated rigorously (e.g., including formal tests for the difference and noninferiority of accuracy). To set the context of this paper, Section II briefly outlines the classification by an SVM. Section III provides a summary of the main methods and data sets used. Section IV presents the results, and Section V details the conclusions of the research undertaken.

II. SVM

The SVM is based on a statistical learning theory [14] and seeks to find an optimal hyperplane as a decision function in high-dimensional space [44], [45]. In the case of a two-class pattern-recognition problem in which the classes are linearly separable, the SVM selects from among the infinite number of linear decision boundaries the one that minimizes the generalization error. Thus, the selected decision boundary (represented by a hyperplane in feature space) will be one that leaves the greatest margin between the two classes, where margin is defined as the sum of the distances to the hyperplane from the closest cases of the two classes [14]. The problem of maximizing the margin can be solved using standard quadratic programming optimization techniques.

The simplest scenario for classification by an SVM is when the classes are linearly separable. This scenario may be illustrated with the training data set comprising k cases and be represented by $\{\mathbf{x}_i, y_i\}, i = 1, \ldots, k$, where $x \in \mathbf{R}^N$ is an N-dimensional space and $y \in \{-1, +1\}$ is the class label. These training patterns are linearly separable if there exists a vector $\mathbf{w}$ (determining the orientation of a discriminating plane) and a scalar b (determining the offset of the discriminating plane from the origin) such that

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0. \tag{1}$$

The hypothesis space can be defined by the set of functions given by

$$f_{\mathbf{w},b} = \operatorname{sign}(\mathbf{w} \cdot \mathbf{x} + b). \tag{2}$$

The SVM finds the separating hyperplanes for which the distance between the classes, measured along a line perpendicular to the hyperplane, is maximized. This can be achieved by solving the following constrained optimization problem:

$$\min_{w,b} \frac{1}{2}\|\mathbf{w}\|^2. \tag{3}$$

For linearly nonseparable classes, the restriction that all training cases of a given class lie on the same side of the optimal hyperplane can be relaxed by the introduction of a "slack variable" $\xi_i \geq 0$. In this case, the SVM searches for the hyperplane that maximizes the margin and that, at the same time, minimizes a quantity proportional to the number of misclassification errors. This tradeoff between margin and misclassification error is controlled by a positive constant C such that $\infty > C > 0$. Thus, for nonseparable data, (3) can be written as

$$\min_{w,b,\xi_1,\ldots\xi_k} \left[\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{k}\xi_i\right]. \tag{4}$$

For nonlinear decision surfaces, a feature vector $x \in \mathbf{R}^N$ is mapped into a higher dimensional Euclidean space (feature space) F via a nonlinear vector function $\Phi : \mathbf{R}^N \mapsto F$[44]. The optimal margin problem in F can be written by replacing $\mathbf{x}_i \cdot \mathbf{x}_j$ with $\mathbf{\Phi}(\mathbf{x}_i) \cdot \mathbf{\Phi}(\mathbf{x}_j)$ which is computationally expensive. To address this problem, Vapnik [14] introduced the concept of using a kernel function K in the design of nonlinear SVM. A kernel function is defined as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{\Phi}(\mathbf{x}_i) \cdot \mathbf{\Phi}(\mathbf{x}_j) \tag{5}$$

and with the use of a kernel function, (2) becomes

$$f(x) = \operatorname{sign}\left(\sum_i \lambda_i y_i K(\mathbf{x}_\mathrm{i}, \mathbf{x}_j) + b\right) \tag{6}$$

where λ_i is a Lagrange multiplier. A detailed discussion of the computational aspects of SVM can be found in [14] and [45], with many examples also in the remote sensing literature [19], [21], [46], [47].

III. Data and Methods

A. Test Areas

Data sets for two study areas were used. The first study area, La Mancha Alta, lies to the south of Madrid, Spain. It is an area of Mediterranean semiarid wetland, which supports rain-fed cultivation of crops such as wheat, barley, vines, and olives. A hyperspectral image data set was acquired for the test site by the Digital Airborne Imaging Spectrometer (DAIS) 7915 sensor on June 29, 2000. The sensor was a 79-channel imaging spectrometer developed and operated by the German Space Agency [48]. This instrument operated at a spatial resolution of 5 m and acquired data in the wavelength range of 0.502–12.278 μm. Attention here focused on the data acquired in only the visible and near-infrared spectra. Thus, the data acquired in the seven features located in the mid- and thermal-infrared regions were removed. Of the remaining 72 features covering spectral region 0.502–2.395 μm, further seven features were

removed because of striping noise distortions in the data. The features removed were bands 41 (1.948 μm), 42 (1.964 μm), and 68–72 (2.343–2.395 μm). After these preprocessing operations, an area of 512 pixels by 512 pixels from the remaining 65 features covering the test site was extracted for further analysis.

The second study area was a region of agricultural land in Indiana, U.S. For this site, a hyperspectral data set acquired by Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) was used. This data set is available online from [49]. The data set consists of a scene of size 145 pixels $\times$ 145 columns. Of the 220 spectral bands acquired by the AVIRIS sensor, 35 were removed as they were affected by noise. For ease of presentation, the bands used were renumbered 1–65 and 1–185 in order of increasing wavelength for the DAIS and AVIRIS data sets, respectively.

B. Training and Testing Data Sets

For the DAIS data set, field observations of the test site were undertaken in late June 2001, exactly one year after the image data were acquired, to generate a ground-reference data set. Visual examination of the DAIS imagery combined with field experience showed that the region comprised mainly eight land-cover types: wheat, water, salt lake, hydrophytic vegetation, vineyards, bare soil, pasture, and built-up land. A ground-reference image was generated from the field information. With the AVIRIS data set, a ground-reference image available on [49] was used to collect the training and test pixels for a total of nine land-cover classes (corn-no till, corn-min till, grass/pasture, grass/trees, hay-windrowed, soybeans-no till, soybeans-min till, soybean-clean, and woods). Stratified random sampling, by class, was undertaken in order to collect independent data sets for training (up to 100 pixels per class) and testing the SVM classifications of the DAIS and AVIRIS data sets.

To evaluate the sensitivity of the SVM to the Hughes effect, a series of training sets of differing sample size was acquired. These data sets were formed by selecting cases randomly from the total available for training each class. A total of six training set sizes, comprising 8, 15, 25, 50, 75, and 100 pixels per class, were used. These training samples are typical of the sizes used in remote sensing studies (e.g., [26], [46], and [50]–[53]) but critically also include small sizes at which the Hughes effect would be expected to manifest itself, if at all. For each size of training set, except that using all 100 pixels available for each class, five independent samples were derived from the available training data. Each of the five training sets of a given size was used to train a classification, and to avoid extreme results, the main focus here is on the classification with the median accuracy.

SVM classifications using training sets of differing sizes were undertaken in which the dimensionality of the input data set, indicated by the number of features used, was varied. Since the main concern was to determine if the Hughes effect would be observed and not the design of an optimal classification, most attention focused on the scenario in which the features were entered in a single fashion for comparative purposes. With this, features were added incrementally in groups of five in order of wavelength. Thus, the first analysis used features 1–5, the second features 1–10, and so on until all the 13th and 37th analyses with DAIS and AVIRIS data, respectively. A number of additional analyses were undertaken with DAIS data in which features were added individually in order of decreasing discriminatory power (i.e., the feature estimated to provide most discriminatory information was entered first, and that which provided the least discriminatory information was added last). Irrespective of the method of incrementing features, the accuracy with which an independent testing set was classified was calculated at each incremental step.

Classification accuracy was estimated using a testing set that comprised a sample of 3800 pixels (500 pixels for seven classes and 300 pixels for the relatively scarce pasture class) with the DAIS data and 3150 pixels (350 pixels per class) with the AVIRIS data sets. In all cases, accuracy was expressed as the percentage of correctly allocated cases. The statistical significance of differences in accuracy was assessed using the McNemar test and confidence intervals [43], [54], [55]. Two types of test were undertaken to elucidate the effect of feature selection on SVM classification accuracy. First, the statistical significance of differences in accuracy was evaluated. This testing was undertaken because one characteristic feature of an analysis that is sensitive to the Hughes effect is a decrease in accuracy following the inclusion of additional features. Thus, the detection of a statistically significant decrease in classification accuracy following the addition of features to the analysis would be an indication of sensitivity to the Hughes effect. A standard one-sided (as the focus is on a directional alternative hypothesis) test of the difference in accuracy values was derived using the McNemar test [55]. However, as feature selection has positive impacts beyond those associated with classification accuracy (e.g., reduced data-processing time and storage requirements), a positive role would also occur if a small feature set could be used without any significant loss of classification accuracy. This cannot be assessed with a test for difference as a result indicating no significant difference in accuracy is not actually a proof of similarity [56]. Indeed, in this situation, the desire is not to test for a significant difference in accuracy but rather to test for the similarity in accuracy, which could be met in this situation through the application of a test for noninferiority [42], [43]. In essence, the aim is to determine if a small feature set, which provides advantages to the analyst, can be used to derive a classification as accurate as that from a large, or indeed, full feature set. The latter test for noninferiority was achieved using the confidence interval fitted to the estimated differences in classification accuracy [43]. For the purpose of this paper, it was assumed that a 1.00% decline in accuracy from the peak value was of no practical significance, and this value is taken to define the extent of the zone of indifference in the test. Critically, a positive role for feature-selection analyses would be indicated if the test for difference was significant (showing that accuracy can be degraded by the addition of new features) and/or if the test for noninferiority was significant (showing that a small feature set derives a classification as accurate as that from the use of a large feature set but providing advantages in relation to data storage and processing, etc.).

C. Feature-Selection Algorithms

From the range of feature-selection methods available, four established methods, including one from each of the main categories of methods identified earlier, were applied to the DAIS data. The salient issues of each method are briefly outlined next.

1) SVM Recursive Feature Elimination (SVM-RFE): The SVM-RFE is a wrapper-based approach utilizing the SVM as base classifier [22]. The SVM-RFE utilizes the objective function $(1/2)\|\mathbf{w}\|^2$ as a feature-ranking criterion to produce a list of features ordered by apparent discriminatory ability. At each step, the coefficients of the weight vector w are used to compute the ranking scores of all features remaining. The feature with the smallest ranking score $(w_i)^2$ is eliminated, where w_i represents the corresponding ith component of w. This approach to feature selection, therefore, uses a backward feature-elimination scheme to recursively remove insignificant features (i.e., at each step, the feature whose removal changes the objective function least is excluded) from subsets of features in order to derive a list of all features in ranked order of value.

2) Correlation-Based Feature Selection (CFS): The CFS is a filter algorithm that selects a feature subset on the basis of a correlation-based heuristic evaluation function [57]. The heuristics by which CFS measures the quality of a set of features take into account the usefulness of individual features for predicting the class and can be summarized as

$$\frac{fC_{ci}}{\sqrt{f + f(f-1)C_{ii}}} \tag{7}$$

where f is the number of features in the subset, C_{ci} is the mean feature correlation with the class, and C_{ii} is the average feature intercorrelation. Both C_{ci} and C_{ii} are calculated by using a measure based on conditional entropy [58]. The numerator provides an indication of how predictive of the class a group of features are, whereas the denominator indicates about the redundancy among the features. The evaluation criterion used in this algorithm is biased toward the feature subsets that are highly predictive of the class and not predictive of each other. This criterion acts to filter out the irrelevant features as they have low correlations with the class, and redundant features are ignored as they will be highly correlated with one or more features, thus providing a subset of best selected features. In order to reduce the computation cost, a bidirectional search (a parallel implementation of sequential forward and backward selections) may be used. This approach searches the space of feature subsets by greedy hill climbing in a way that features already selected by sequential forward selection are not removed by backward selection, and the features already removed by backward selection are not selected by forward selection.

3) Minimum-Redundancy–Maximum-Relevance (mRMR): The mRMR feature selection is a filter-based method that uses mutual information to determine the dependence between the features [59]. The mRMR uses a criterion which selects features that are different from each other and still have the largest dependence on the target class. This approach consists in selecting a feature f_i among the not selected features f_S that maximizes $(u_i - r_i)$, where u_i is the relevance of f_i to the class c alone and r_i is the mean redundancy of f_i to each of the already selected features. In terms of mutual information, u_i and r_i can be defined as

$$u_i = \frac{1}{|f|} \sum_{f_i \in f} I(f_i; c) \tag{8}$$

$$r_i = \frac{1}{|f|^2} \sum_{f_j \in f} I(f_i, f_j) \tag{9}$$

where $I(f; c)$ is the mutual information between the two random variables f and c. At each step, this method selects a feature that has the best compromised relevance redundancy and can be used to produce a ranked list of all features in terms of discriminating ability.

4) Random Forest: The random-forest-based approach is an embedded method of feature selection. The random forest consists of a collection of decision-tree classifiers [60] where each tree in the forest has been trained using a bootstrap sample of training data and a random subset of features sampled independently from the input features. A subset of the training data set is omitted from the training of each classifier [61]. These left-out data are called out-of-bag (out of the bootstrap) samples and are used for feature selection by determining the importance of different features during classification process [60], [62]. The latter is based on a Z score, which can be used to assign a significance level (importance level) to a feature, and from this, a ranked list of all features may be derived [60].

D. Methods

SVMs were initially designed for binary classification problems. A range of methods has been suggested for multiclass classification [21], [63], [64]. One of these, the "one-against-one" approach, was used here [65] with both hyperspectral data sets. Throughout, an RBF kernel was used with kernel width parameter $\gamma = 2$ and $C = 5000$, values which were used successfully with the DAIS hyperspectral data set in other studies [19], [20], [33], [66]. For analyses of the AVIRIS data set, an RBF kernel with $\gamma = 1$ and regularization parameter $C = 50$ was used [66].

With the feature selection by random forests, one-third of the total data set available for training was used to form the out-of-bag sample. The random-forest classifier also requires finding the optimal value of a number of features used to generate a tree as well as the total numbers of trees. After several trials, 13 features and 100 trees were found to be working well with the DAIS data set [33].

IV. Results

The accuracy of classification by an SVM varied as a function of the number of features used and the size of the training set using the DAIS data set (Fig. 1). In general terms, classification accuracy tended to increase with an increase in the number of features. Critically, however, when a fixed training set of small size (≤ 25 cases per class) was used, the accuracy initially rose with the addition of features to a peak, but thereafter declined with the addition of further features. Moreover, the

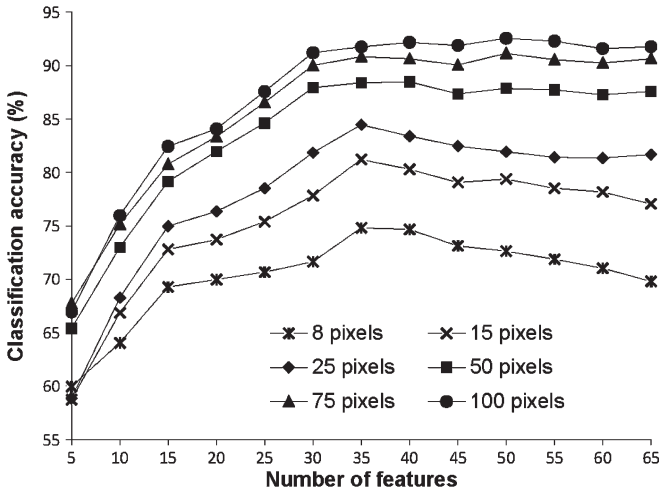


Fig. 1. Variation of classification accuracy with the number of features for analyses based on training sets of differing size using the DAIS data set.

TABLE I
DIFFERENCE BETWEEN PEAK ACCURACY AND THAT DERIVED FROM THE USE OF ALL 65 FEATURES OF DAIS DATA SET FOR THE RESULTS SUMMARIZED IN FIG. 1. THE Z VALUE STATED WAS DERIVED FROM THE MCNEMAR TEST. FOR THE ONE-SIDED TEST ADOPTED, A DIFFERENCE IS SIGNIFICANT AT THE 0.05 LEVEL IF $Z > 1.64$

	Training set size per class					
	8 pixels	15 pixels	25 pixels	50 pixels	75 pixels	100 pixels
Peak accuracy, % (number of features)	74.79 (35)	81.21 (35)	84.45 (35)	88.47 (40)	91.13 (50)	92.53 (50)
Accuracy with 65 features (%)	69.79	77.05	81.66	87.58	90.63	91.76
Difference (%)	5.00	4.16	2.79	0.89	0.50	0.77
Z value	6.04	5.35	4.02	1.69	1.48	2.22

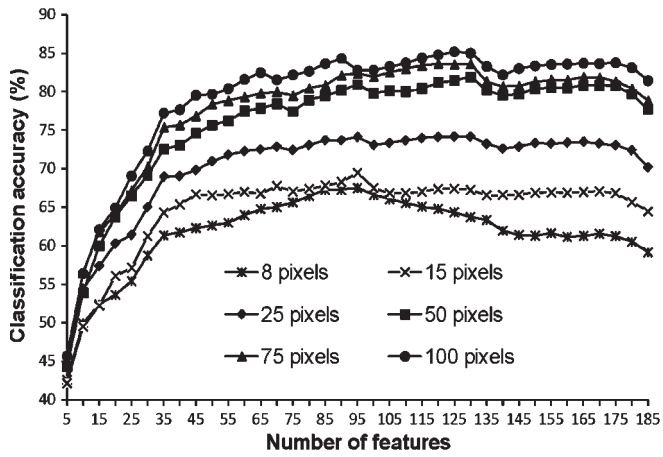


Fig. 2. Variation of classification accuracy with the number of features for analyses based on training sets of differing size using the AVIRIS data set.

TABLE II
DIFFERENCE BETWEEN PEAK ACCURACY AND THAT DERIVED FROM THE USE OF ALL 185 FEATURES OF AVIRIS DATA SET FOR THE RESULTS SUMMARIZED IN FIG. 2. THE Z VALUE STATED WAS DERIVED FROM THE MCNEMAR TEST. FOR THE ONE-SIDED TEST ADOPTED, A DIFFERENCE IS SIGNIFICANT AT THE 0.05 LEVEL IF $Z > 1.64$

	Training set size per class					
	8 pixels	15 pixels	25 pixels	50 pixels	75 pixels	100 pixels
Peak accuracy, % (number of features)	67.53 (95)	69.49 (95)	74.21 (130)	81.94 (130)	83.65 (120)	85.21 (125)
Accuracy with 65 features (%)	59.17	64.48	70.19	77.75	78.89	81.46
Difference (%)	8.36	5.01	4.02	4.19	4.76	3.75
Z value	9.44	5.92	8.77	6.92	7.18	6.10

decline in accuracy was statistically significant, even for the classification based on the largest training set size (Table I). For example, the largest difference between the peak accuracy and that obtained from the use of all 65 features was 5.00%, a difference that was significant at the 0.05 level of significance (Table I).

Similar general trends to those found with the analysis of the DAIS data were observed with the results of the analyses of the AVIRIS data set (Fig. 2). Critically, classification accuracy was observed to decline with the addition of features. Moreover, with this data set, a statistically significant (at 0.05 level) decline in accuracy with the addition of features was observed for all training set sizes (Table II). The largest difference between the peak accuracy and that obtained from the use of all 185 features was 8.36%.

Consequently, the key negative characteristic of the curse of dimensionality or Hughes effect was observed with SVM classification when a small training set was used. Although this result contradicts some statements in the literature that suggest that the SVM is independent of the dimensionality of the data set [20], [21], it should be noted that these studies used relatively large training sets and do not include a rigorous statistical test of the significance of differences in accuracy. For example, Melgani and Bruzzone [21] used over 230 training cases for each class, while Pal and Mather [20] used sample sizes of at least 100 pixels per class. The size of the training sets used in these studies may have been sufficiently large to ensure that Hughes effect was not manifest in the analyses reported. Thus, in these studies, the experimental designs adopted may not have provided an opportunity for the Hughes effect to arise and be detected. Additionally, it may be expected that the degree to which the effect is observed may vary from study to study as a function of the classes (e.g., their number and spectral separability) and data set (e.g., number and location of spectral wavebands). Note, for example, that the Hughes effect appeared to occur at each training set size studied with the AVIRIS data

TABLE III
RESULTS OF THE APPLICATION OF THE FOUR FEATURE-SELECTION METHODS USING DAIS DATA SET HIGHLIGHTING THE CHARACTERISTICS OF THE CLASSIFICATION BASED ON EACH TRAINING SET SIZE THAT WAS OF MOST COMPARABLE ACCURACY WITH THAT DERIVED WITHOUT FEATURE SELECTION

Feature selection Method	Training set size per class											
	8 pixels		15 pixels		25 pixels		50 pixels		75 pixels		100 pixels	
	Accuracy (%)	Feature size	Accuracy (%)	Feature size	Accuracy (%)	Feature size	Accuracy (%)	Feature size	Accuracy (%)	Feature size	Accuracy (%)	Feature size
None	69.29	65	74.82	65	80.58	65	87.10	65	90.71	65	91.76	65
SVM-RFE	69.84	4	75.39	10	81.68	7	87.45	15	90.87	16	91.89	13
mRMR	69.71	8	76.34	11	81.02	12	87.13	13	90.87	42	91.84	37
CFS	69.50	4	75.82	7	82.18	8	87.11	12	91.32	14	91.84	17
Random forest	71.94	6	76.39	9	81.95	9	87.11	14	90.82	25	92.08	21

TABLE IV
SELECTED FEATURES WITH DIFFERENT DATA SETS AND THE NUMBER OF COMMON FEATURES SELECTED BY VARIOUS APPROACHES USING DAIS DATASET.

Feature selection approach	Training set size per class						Number of common features
	8 pixels	15 pixels	25 pixel	50 pixel	75 pixels	100 pixels	
SVM-RFE	1,4,35,53	1,4,6,27,32,36 ,37,50,51,57	1,3,4,26,32,37 ,42	1,2,3,4,18, 26,27,31,32, 36,37,46,48,52, 56	1,2,3,4,5,26,27, 30,31,32,34,36, 37,40,52,56	1,2,3,21,26,27 ,30,34,36,37, 51,52,56	1
mRMR	10,15,16,17, 24,25,49,56	9,16,22,24, 25,26,32,48, 49,50,65	9,15,22,24, 25,26,29,31, 32,48,49,51	8,21,22,23, 24,25,26,27,28, 30,49,50,65	2,3,6,7,8,9, 10,12,13,14,15, 16,17,18,19,20, 21,22,23,24,25, 26,27,28,29,30, 31,32,36,37,38, 41,47,48,49,50, 51,52,53,63,64, 65	6,7,8,9,12, 13,14,15,16, 17,18,19,20, 21,22,23,24,2 5,26,27,28,29, 30,31,32,33,3 8,41,47,48,49, 50,51,52,53, 63,65	3
CFS	2,10,15,17	3,10,15,23, 24,29,36	2,5,10,13, 21,24,25, 29	1,2,5,10,21, 22,24,25,27, 28,30,31	1,2,5,9,20, 22,27,28,29,31, 32,37,40, 44	1,2,4,13,17,20 ,24,25,27,28, 30,31,32,36, 37,39,45	0
Random forest	14,28,29, 30,41,58	10,21,22,24, 27,30,32,40,4 1	1,2,5,12,21,28 ,29,31,32	1,2,3,4,5,24,25, 26,30,31,32,39, 42,50	1,2,4,5,6,7, 23,24,26,27,29, 30,31,32,39,41, 42,44,49,50,53, 61,63,64,65	1,2,3,5,22, 23,26,27,28, 29,30,31,32, 39,40,41,42, 50,59,63,64	0

(Fig. 2) but only when small ($\leq$ 25 cases per class) training sets were used with the DAIS data set (Fig. 1).

Having established that the accuracy of classification by an SVM is sensitive to the number of features used, the four different feature-selection methods were applied to the DAIS data in order to evaluate the sensitivity of SVM classification to different types of feature-selection method. The aim was not to define an optimal feature selection but to provide insight into the sensitivity of the SVM classification to the method used.

The classifications derived after application of the four feature-selection methods varied in accuracy. Unlike the previous analyses, features were added individually to classifications in the order suggested by the feature-selection analysis. To focus on key trends, Table III shows the accuracy derived without feature selection and the accuracy that was of closest magnitude after the application of each of the feature selection methods. Critically, the table also identifies the number of features used to derive the classification accuracy closest to that derived when no feature selection was undertaken. Irrespective of the feature-selection algorithm employed, the results suggest that a small subset of selected features ($\leq$ 12) would be sufficient to achieve comparable accuracy with the small training sets comprising 8, 15, and 25 pixels per class. In comparison, the training sets with 50, 75, and 100 pixels per class require a larger subset of selected features to achieve the comparable classification accuracy with that derived from the full data set (and the accuracy values were also of a higher magnitude).

It was evident from Table III that the feature-selection methods varied in efficiency, measured in terms of the number of features required to derive a classification of comparable accuracy

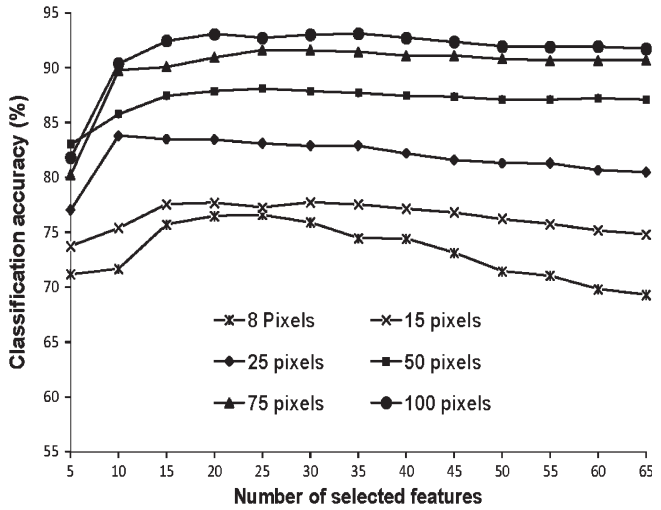


Fig. 3. Relationship between classification accuracy and the number of features selected by the SVM-RFE using the DAIS data set.

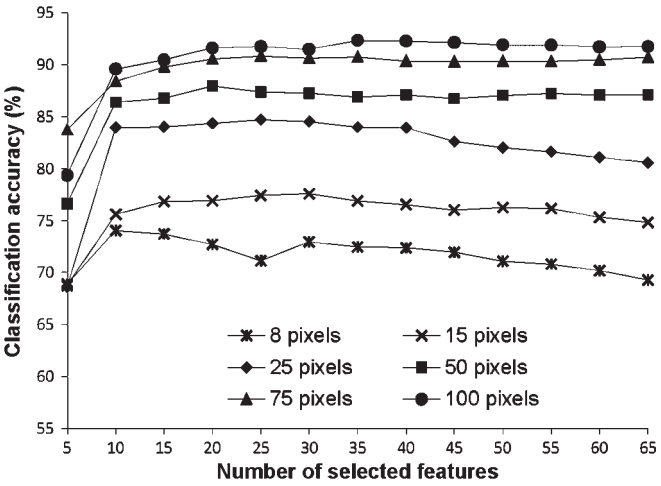


Fig. 4. Relationship between classification accuracy and the number of features selected by the random forest using the DAIS data set.

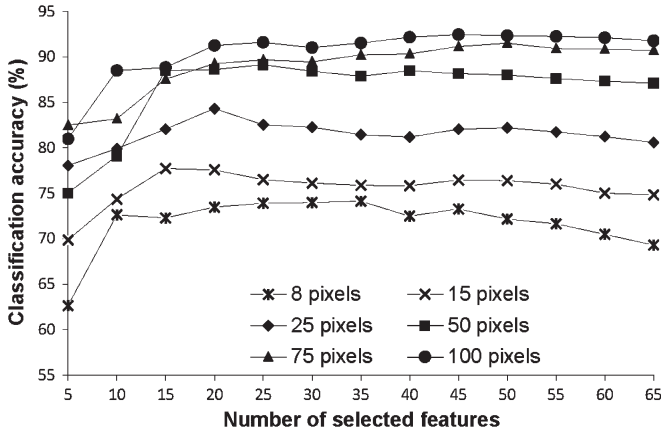


Fig. 5. Relationship between classification accuracy and the number of features selected by the mRMR using the DAIS data set.

TABLE V
SUMMARY OF THE TEST FOR THE DIFFERENCE IN ACCURACY BETWEEN THE PEAK ACCURACY AND THAT DERIVED FROM THE USE OF THE FULL FEATURE SET USING DAIS DATA SET. VALUES IN BRACKET GIVE THE NUMBER OF FEATURES PROVIDING PEAK CLASSIFICATION ACCURACY, SHOWN IN FIGS. 2–4. THE Z VALUE STATED WAS DERIVED FROM THE MCNEMAR TEST. FOR THE ONE-SIDED TEST ADOPTED, A DIFFERENCE IS SIGNIFICANT AT THE 0.05 LEVEL IF $Z > 1.64$

	Z value					
	8 pixels	15 pixels	25 pixels	50 pixels	75 pixels	100 pixels
SVM-RFE	11.54 (25)	5.19 (20)	7.10 (15)	2.33 (25)	2.35 (25)	4.84 (35)
Random forest	7.29 (10)	5.54 (30)	7.84 (25)	1.64 (20)	0.25 (25)	1.67 (35)
MRMR	8.73 (35)	4.80 (15)	7.12 (20)	4.01 (20)	2.65 (50)	2.44 (45)

to that derived without feature selection. Note, for example, that in the two filter-based feature-selection approaches, the CFS uses a smaller subset of features in comparison with mRMR. This suggests that, for this data set, at least, CFS is more suitable than the mRMR method.

It was also evident that the specific features selected by the different methods varied. Table IV identifies the selected features that provided the classification of comparable accuracy to that derived from the full (65 features) data set. It was evident that a dissimilar feature list was obtained from analyses based on training sets of differing size, with at most only three common features observed with any one feature-selection method. The outputs of the feature-selection methods were therefore a function of the training set size. Moreover, the lack of commonalities in the features selected with different training set sizes also confirms that the best set of features selected by a nonexhaustive search need not contain the best feature or a set of best features from the full feature space [67].

For comparison against the results given in Fig. 1, Figs. 3–5 show the relationship between classification accuracy and the number of selected features using three of the feature-selection methods. The CFS method was excluded from this analysis, as this approach does not provide a ranked list of the features. For the purpose of comparability with Fig. 1, the features have been added in groups of five (in order of discriminating ability). The statistical significance of the difference in accuracy between the peak accuracy value and that derived with the use of the full feature set for each classification summarized in Figs. 3–5 was evaluated with a McNemar test. The derived Z values are provided in Table V, which suggests a similar trend as achieved with earlier combination of features (Fig. 1) using the training sample size of 8, 15, and 25 pixels per class. It was evident, however, that the peak accuracy was derived with a smaller number of features as, in this case, the features were added in order of discriminating power.

The results highlight that a statistically significant negative impact of feature-set size on classification accuracy was observed when a small training sample was used, confirming the results of the McNemar test for a significant difference. Although this in itself points to a dependence of SVM classification on the dimensionality of the data set and highlights a positive role for feature-selection analysis, the latter has

TABLE VI
DIFFERENCE AND NONINFERIORITY TEST RESULTS BASED ON 95% CONFIDENCE INTERVAL ON THE ESTIMATED DIFFERENCE IN ACCURACY FROM THE PEAK VALUE FOR FEATURE SETS SELECTED WITH THE SVM-RFE USING DAIS DATA SET: BASED ON TRAINING SET OF 100 CASES PER CLASS WITH PEAK ACCURACY OF 93.13% WITH 35 FEATURES

Number of features	Accuracy (%)	Difference from peak accuracy (%)	95% confidence interval	Conclusion (at 0.05 level of significance)
5	81.82	11.31	11.298 - 11.322	Different
10	90.40	2.73	2.721 - 2.739	Different
15	92.47	0.66	0.653 - 0.667	Non-inferior
20	93.08	0.05	0.044 - 0.056	Non-inferior
25	92.74	0.39	0.384 - 0.396	Non-inferior
30	93.03	0.10	0.096 - 0.104	Non-inferior
35	93.13	0.00	0.000 – 0.000	(no change)
40	92.74	0.39	0.386 – 0.394	Non-inferior
45	92.37	0.76	0.755 – 0.765	Non-inferior
50	91.97	1.16	1.154 – 1.166	Different
55	91.92	1.21	1.204 – 1.216	Different
60	91.95	1.18	1.174 – 1.186	Different
65	91.76	1.37	1.364 – 1.376	Different

TABLE VII
DIFFERENCE AND NONINFERIORITY TEST RESULTS BASED ON 95% CONFIDENCE INTERVAL ON THE ESTIMATED DIFFERENCE IN ACCURACY FROM THE PEAK VALUE FOR FEATURE SETS SELECTED WITH THE RANDOM FOREST USING DAIS DATA SET: BASED ON TRAINING SET OF 100 CASES PER CLASS WITH PEAK ACCURACY OF 92.34% WITH 35 FEATURES

Number of features	Accuracy (%)	Difference from peak accuracy (%)	95% confidence interval	Conclusion (at 0.05 level of significance)
5	79.37	12.97	12.958 – 12.982	Different
10	89.58	2.76	2.751 – 2.769	Different
15	90.47	1.87	1.862 – 1.878	Different
20	91.61	0.73	0.724 – 0.736	Non-inferior
25	91.76	0.58	0.573 – 0.587	Non-inferior
30	91.50	0.84	0.835 – 0.845	Non-inferior
35	92.34	0.00	0.000 – 0.000	(no change)
40	92.29	0.05	0.046 – 0.054	Non-inferior
45	92.13	0.21	0.205 – 0.215	Non-inferior
50	91.92	0.42	0.414 – 0.426	Non-inferior
55	91.89	0.45	0.444 – 0.456	Non-inferior
60	91.71	0.63	0.623 – 0.637	Non-inferior
65	91.76	0.58	0.573 – 0.587	Non-inferior

TABLE VIII
DIFFERENCE AND NONINFERIORITY TEST RESULTS BASED ON 95% CONFIDENCE INTERVAL ON THE ESTIMATED DIFFERENCE IN ACCURACY FROM THE PEAK VALUE FOR FEATURE SETS SELECTED WITH THE mRMR USING DAIS DATA SET: BASED ON TRAINING SET OF 100 CASES PER CLASS WITH PEAK ACCURACY OF 92.45% WITH 45 FEATURES

Number of features	Accuracy (%)	Difference from peak accuracy (%)	95% confidence interval	Conclusion (at 0.05 level of significance)
5	80.97	11.48	11.468 – 11.492	Different
10	88.5	3.95	3.940 – 3.960	Different
15	88.82	3.63	3.620 – 3.640	Different
20	91.24	1.21	1.202 – 1.218	Different
25	91.58	0.87	0.862 – 0.878	Non-inferior
30	91.03	1.42	1.413 – 1.427	Different
35	91.53	0.92	0.914 – 0.926	Non-inferior
40	92.16	0.29	0.286 – 0.294	Non-inferior
45	92.45	0.00	0.000 – 0.000	(no change)
50	92.34	0.11	0.106 – 0.114	Non-inferior
55	92.24	0.21	0.206 – 0.214	Non-inferior
60	92.11	0.34	0.335 – 0.345	Non-inferior
65	91.76	0.69	0.685 – 0.696	Non-inferior

other advantages, and the results suggest that feature selection may be valuable even when a large training sample was available. Note, for example, that in all series of analyses (Fig. 1 and Figs. 3–5), when the largest training sample was used (100 cases per class), the accuracy was largely maintained when the number of features is reduced from the full (65 features) to a small subset; only at a very small number of features did the classification accuracy decline markedly. This similarity in accuracy values shows that the positive benefits of feature selection (e.g., reduced data storage and processing requirements) may be achieved without significant negative effect on classification accuracy. The latter is evident in the results of the noninferiority testing summarized in Tables VI–VIII. Critically, the accuracy of classifications derived with the use of relatively small training sets was not statistically inferior to the peak accuracy derived from the use of a larger feature-set size.

V. CONCLUSION

The SVM has been widely used and promoted for land-cover classification studies, including multispectral and hyperspectral data with some studies suggesting that the method is not affected by the Hughes phenomena. A major conclusion of this paper is that the accuracy of SVM classification is influenced by the number of features used and, therefore, is affected by the Hughes phenomenon with the impact most evident when a small training set is used (Figs. 1 and 2, Tables I and II). It is possible that the Hughes effect had not been observed in some other studies because the opportunity for it to become manifested in the results was limited through experimental design, notably through the use of a large training set. The results presented in this paper have shown that the accuracy of classification by an SVM can be significantly reduced by the addition of features and that the effect is most apparent with small training sets. With the AVIRIS data set, a significant reduction in accuracy with the addition of features was observed at all training set sizes evaluated. With the DAIS data set, a statistically significant decline in accuracy was also observed for small

training sets (≤ 25 cases per class). However, even with a large training sample using the DAIS data set, feature selection may have a positive role, providing a reduced data set that may be used to yield a classification of similar accuracy to that derived from use of a much larger feature set. As the accuracy of SVM classification was dependent on the dimensionality of the data set and the size of the training set, it may therefore be beneficial to undertake a feature-selection analysis prior to a classification analysis. The results, however, also highlight that the choice of the feature-selection methods may be important. For example, the results derived from analyses with the four different feature-selection methods show that the number of features selected varied greatly.

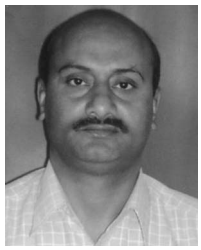
ACKNOWLEDGMENT

The authors would like to thank Prof. J. Gumuzzio of the Autonomous University of Madrid, Spain, for making available the DAIS data that were collected and processed by DLR and also the three referees for their constructive comments on the original version of this paper. M. Pal would like to thank the School of Geography, University of Nottingham, for the computing facilities.

REFERENCES

[1] C.-I Chang, *Hyperspectral Data Exploitation: Theory and Applications*. Hoboken, NJ: Wiley, 2007.
[2] J. B. Campbell, *Introduction to Remote Sensing*, 3rd ed. New York: Guilford Press, 2002.
[3] J. A. Benediktsson and J. R. Sveinsson, "Feature extraction for multisource data classification with artificial neural networks," *Int. J. Remote Sens.*, vol. 18, no. 4, pp. 727–740, Mar. 1997.
[4] P. Zhong, P. Zhang, and R. Wang, "Dynamic learning of SMLR for feature selection and classification of hyperspectral data," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 2, pp. 280–284, Apr. 2008.
[5] B. M. Shahshahani and D. A. Landgrebe, "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. 1087–1095, Sep. 1994.
[6] S. Tadjudin and D. A. Landgrebe, "Covariance estimation with limited training samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 4, pp. 2113–2118, Jul. 1999.
[7] M. Chi, R. Feng, and L. Bruzzone, "Classification of hyperspectral remote-sensing data with primal SVM for small-sized training dataset problem," *Adv. Space Res.*, vol. 41, no. 4, pp. 1793–1799, 2008.
[8] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.
[9] S. Lu, K. Oki, Y. Shimizu, and K. Omasa, "Comparison between several feature extraction/classification methods for mapping complicated agricultural land use patches using airborne hyperspectral data," *Int. J. Remote Sens.*, vol. 28, no. 5, pp. 963–984, Jan. 2007.
[10] S. Tadjudin and D. A. Landgrebe, "A decision tree classifier design for high-dimensional data with limited training samples," in *Proc. IEEE Geosci. Remote Sens. Symp.*, May 27–31, 1996, vol. 1, pp. 790–792.
[11] M. Chi and L. Bruzzone, "A semilabeled-sample-driven bagging technique for ill-posed classification problems," *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 1, pp. 69–73, Jan. 2005.
[12] P. Mantero, G. Moser, and S. B. Serpico, "Partially supervised classification of remote sensing images through SVM-based probability density estimation," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 559–570, Mar. 2005.
[13] G. M. Foody and A. Mathur, "Toward intelligent training of supervised image classifications: Directing training data acquisition for SVM classification," *Remote Sens. Environ.*, vol. 93, no. 1/2, pp. 107–117, Oct. 2004.
[14] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
[15] C. Cortes and V. N. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
[16] V. N. Vapnik, *Estimation of Dependences Based on Empirical Data*. New York: Springer-Verlag, 1982.
[17] D. M. J. Tax, D. de Ridder, and R. P. W. Duin, "Support vector classifiers: A first look," in *Proc. 3rd Annu. Conf. Adv. School Comput. Imaging*, H. E. Bal, H. Corporaal, P. P. Jonker, and J. F. M. Tonino, Eds., Heijen, The Netherlands, Jun. 2–4, 1997, pp. 253–258.
[18] J. A. Gualtieri, "The support vector machine (SVM) algorithm for supervised classification of hyperspectral remote sensing data," in *Kernel Methods for Remote Sensing Data Analysis*, G. Camps-Valls and L. Bruzzone, Eds. Chichester, U.K.: Wiley, 2009.
[19] M. Pal and P. M. Mather, "Assessment of the effectiveness of support vector machines for hyperspectral data," *Future Generation Comput. Syst.*, vol. 20, no. 7, pp. 1215–1225, Oct. 2004.
[20] M. Pal and P. M. Mather, "Some issues in classification of DAIS hyperspectral data," *Int. J. Remote Sens.*, vol. 27, no. 14, pp. 2895–2916, Jul. 2006.
[21] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.
[22] I. Guyon, J. Weston, S. Barnhill, and V. N. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 389–422, Jan. 2002.
[23] A. Gidudu and H. Ruther, "Comparison of feature selection techniques for SVM classification," in *Proc. 10th Int. Symp. Phys. Meas. Spectral Signatures Remote Sens.*, vol. XXXVI, *Intl. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, M. E. Schaepman, S. Liang, N. E. Groot, and M. Kneubühler, Eds., Davos, Switzerland, 2007, pp. 258–263.
[24] H. Liu, "Evolving feature selection," *IEEE Intell. Syst.*, vol. 20, no. 6, pp. 64–76, Nov. 2005.
[25] H. Liu and H. Motoda, *Feature Extraction, Construction and Selection: A Data Mining Perspective*. Norwell, MA: Kluwer, 1998.
[26] P. M. Mather, *Computer Processing of Remotely-Sensed Images: An Introduction,* 3rd ed. Chichester, U.K.: Wiley, 2004.
[27] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1/2, pp. 273–324, Mar. 1997.
[28] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, no. 7/8, pp. 1157–1182, Mar. 2003.
[29] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal., Int. J.*, vol. 1, no. 3, pp. 131–156, 1997.
[30] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 153–158, Feb. 1997.
[31] T. Kavzoglu and P. M. Mather, "The role of feature selection in artificial neural network applications," *Int. J. Remote Sens.*, vol. 23, no. 15, pp. 2787–2803, Aug. 2002.
[32] S. B. Serpico and L. Bruzzone, "A new search algorithm for feature selection in hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1360–1367, Jul. 2001.
[33] M. Pal, "Support vector machine-based feature selection for land cover classification: A case study with DAIS hyperspectral data," *Int. J. Remote Sens.*, vol. 27, no. 14, pp. 2877–2894, Jul. 2006.
[34] J. Loughrey and P. Cunningham, "Overfitting in wrapper-based feature subset selection: The harder you try the worse it gets," in *Research and Development in Intelligent Systems XXI*, M. Bramer, F. Coenen, and T. Allen, Eds. London, U.K.: Springer-Verlag, 2004, pp. 33–43.
[35] G. H. Halldorsson, J. A. Benediktsson, and J. R. Sveinsson, "Source-based feature extraction for support vector machines in hyperspectral classification," in *Proc. IEEE Geosci. Remote Sens. Symp.*, Sep. 20–24, 2004, vol. 1, pp. 536–539.
[36] O. Barzilay and V. L. Brailovsky, "On domain knowledge and feature selection using a support vector machine," *Pattern Recognit. Lett.*, vol. 20, no. 5, pp. 475–484, May 1999.
[37] A. Navot, R. Gilad-Bachrach, Y. Navot, and N. Tishby, "Is Feature Selection Still Necessary?" *Lecture Notes in Computer Science*. vol. 3940, Berlin, Germany: Springer-Verlag, 2006, pp. 127–138.
[38] Y. Bengio, O. Delalleau, and N. Le Roux, "The curse of highly variable functions for local kernel machines," in *Advances in Neural Information Processing Systems*, vol. 18. Cambridge, MA: MIT Press, 2006, pp. 107–114.
[39] D. Francois, V. Wertz, and M. Verleysen, "About the locality of kernels in high dimensional space," in *Proc. Int. Symp. Appl. Stochastic Models Data Anal.*, Brest, France, May 17–20, 2005, pp. 238–245.
[40] B. Scholkopf, S. Mika, C. J. C. Burges, P. Knirsch, K. R. Muller, G. Ratsch, and A. J. Smola, "Input space versus feature space in kernel-based methods," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 1000–1017, Sep. 1999.

[41] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Comput.*, vol. 4, no. 1, pp. 1–58, Jan. 1992.
[42] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical Methods for Rates & Proportions*, 3rd ed. New York: Wiley-Interscience, 2003.
[43] G. M. Foody, "Classification accuracy comparison: Hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority," *Remote Sens. Environ.*, vol. 113, no. 8, pp. 1658–1663, Aug. 2009.
[44] B. Boser, I. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. Workshop Comput. Learn. Theory*, 1992, pp. 144–152.
[45] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[46] G. M. Foody and A. Mathur, "A relative evaluation of multiclass image classification by support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 6, pp. 1335–1343, Jun. 2004.
[47] G. Camps-Valls and L. Bruzzone, *Kernel Methods for Remote Sensing Data Analysis*. Chichester, U.K.: Wiley.
[48] P. Strobl, R. Richter, F. Lehmann, A. Mueller, B. Zhukov, and D. Oertel, "Preprocessing for the airborne imaging spectrometer DAIS 7915," *Proc. SPIE*, vol. 2758, pp. 375–382, Jun. 1996.
[49] *AVIRIS NW Indiana's Indian Pines*, 1992. data set, ftp://ftp.ecn.purdue.edu/biehl/MultiSpec/92AV3C.lan (original files) and ftp://ftp.ecn.purdue.edu/biehl/PC_MultiSpec/ThyFiles.zip (ground truth).
[50] G. M. Foody and M. K. Arora, "An evaluation of some factors affecting the accuracy of classification by an artificial neural network," *Int. J. Remote Sens.*, vol. 18, no. 4, pp. 799–810, Mar. 1997.
[51] G. M. Foody, A. Mathur, C. Sanchez-Hernandez, and D. S. Boyd, "Training set size requirements for the classification of a specific class," *Remote Sens. Environ.*, vol. 104, no. 1, pp. 1–14, Sep. 2006.
[52] M. Pal and P. M. Mather, "An assessment of the effectiveness of decision tree methods for land cover classification," *Remote Sens. Environ.*, vol. 86, no. 4, pp. 554–565, Oct. 2003.
[53] T. G. Van Niel, T. R. McVicar, and B. Datt, "On the relationship between training sample size and data dimensionality of broadband multi-temporal classification," *Remote Sens. Environ.*, vol. 98, no. 4, pp. 468–480, Oct. 2005.
[54] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998.
[55] G. M. Foody, "Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 5, pp. 627–633, May 2004.
[56] D. G. Altman and J. M. Bland, "Absence of evidence is not evidence of absence," *Brit. Med. J.*, vol. 311, no. 7003, p. 485, Aug. 1995.
[57] M. A. Hall and L. A. Smith, "Feature subset selection: A correlation-based filter approach," in *Proc. Int. Conf. Neural Inf. Process. Intell. Inf. Syst.*, 1997, pp. 855–858.
[58] W. H. Press, *Numerical Recipes*. Cambridge, U.K.: Cambridge Univ. Press, 1988.
[59] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[60] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
[61] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.
[62] R. Díaz-Uriarte and S. A. de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinf.*, vol. 7, no. 1, p. 3, 2006.
[63] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multi-class support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.
[64] M. Pal, "Multiclass approaches for support vector machine based land cover classification," in *Proc. 8th Annu. Int. Conf., Map India*, 2005. [Online]. Available: http://www.gisdevelopment.net/technology/rs/mi0554.htm. [Accessed: Dec. 12, 2008].
[65] S. Knerr, L. Personnaz, and G. Dreyfus, "Single-layer learning revisited: A stepwise procedure for building and training neural network," in *Neurocomputing: Algorithms, Architectures and Applications*. Berlin, Germany: Springer-Verlag, 1990.
[66] M. Pal, "Margin-based feature selection for hyperspectral data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 11, no. 3, pp. 212–220, Jun. 2009.
[67] T. M. Cover, "The best two independent measurements are not the two best," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-4, no. 1, pp. 116–117, Jan. 1974.



Mahesh Pal received the Ph.D. degree from the University of Nottingham, Nottingham, U.K., in 2002.

He is currently an Associate Professor with the Department of Civil Engineering, National Institute of Technology, Kurukshetra, India. His major research areas are land-cover classification, feature selection, and application of artificial intelligence techniques in various civil engineering application.

Dr. Pal is in the editorial board of the recently launched journal *Remote Sensing Letters*.

Giles M. Foody (M'01) received the B.Sc. and Ph.D. degrees in geography from the University of Sheffield, Sheffield U.K., in 1983 and 1986, respectively.

He is currently a Professor of geographical information science with the University of Nottingham, Nottingham, U.K. His main research interests focus on the interface between remote sensing, ecology, and informatics.

Dr. Foody is currently the Editor-in-Chief of the *International Journal of Remote Sensing* and of the recently launched journal *Remote Sensing Letters*. He holds editorial roles with *Landscape Ecology* and *Ecological Informatics* and serves on the editorial board of several other journals. He was the recipient of the Remote Sensing and Photogrammetry Society's Award, its highest award, for services to remote sensing in 2009.