Available online at www.sciencedirect.com

SCIENCE DIRECT®

Remote Sensing of Environment 93 (2004) 107–117

Remote Sensing of Environment

www.elsevier.com/locate/rse

Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification

Giles M. Foody*, Ajay Mathur

School of Geography, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom

Received 10 May 2004; received in revised form 18 June 2004; accepted 26 June 2004

Abstract

Conventional approaches to training a supervised image classification aim to fully describe all of the classes spectrally. To achieve a complete description of each class in feature space, a large training set is typically required. It is not, however, always necessary to have training statistics that provide a complete and representative description of the classes, especially if using nonparametric classifiers. For classification by a support vector machine, only the training samples that are support vectors, which lie on part of the edge of the class distribution in feature space, are required; all other training samples provide no contribution to the classification analysis. If regions likely to furnish support vectors can be identified in advance of the classification, it may be possible to intelligently select useful training samples. The ability to target useful training samples may allow accurate classification from small training sets. This potential for intelligent training sample collection was explored for the classification of agricultural crops from multispectral satellite sensor data. With a conventional approach to training, only a quarter of the training samples acquired actually made a positive contribution to the analysis and allowed the crops to be classified to a high accuracy (92.5%). The majority of the training set, therefore, was unnecessary as it made no contribution to the analysis. Using ancillary information on soil type, however, it would be possible to constrain the training sample acquisition process. By limiting training sample acquisition only to regions with a specific soil type, it was possible to use a small training set to classify the data without loss of accuracy. Thus, a small number of intelligently selected training samples may be used to classify a data set as accurately as a larger training set derived in a conventional manner. The results illustrate the potential to direct training data acquisition strategies to target the most useful training samples to allow efficient and accurate image classification.

Keywords: Intelligent training; SVM classification; Support vectors

1. Introduction

Supervised classification is one of the most commonly undertaken analyses of remotely sensed data. The output of a supervised classification is effectively a thematic map that provides a snapshot representation of the spatial distribution of a particular theme of interest such as land cover. Thematic maps are crucial to many applications, as well as, for example, the basis for techniques to study land cover changes or the information required to parameterize many environmental models.

Although remote sensing has considerable potential as a source of thematic maps many problems are encountered and the accuracy of the maps derived is sometimes viewed as insufficient by users (Foody, 2002). Many factors may be responsible for this situation. These range from the nature of the classes being studied through the properties of the sensing system used to acquire the imagery to the techniques used to extract thematic information from the imagery. Here, attention is focused on one key aspect of the use of supervised image classification techniques for the accurate derivation of thematic information from remotely sensed imagery, the training data used to direct the classification.

Supervised classifications may be considered to comprise three distinct stages: training, allocation and testing.

* Corresponding author. Fax: +44 2380 593 295.
E-mail address: g.m.foody@soton.ac.uk (G.M. Foody).

0034-4257/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.rse.2004.06.017

Quantitative descriptions of each class to be mapped are derived in the training stage. For this, areas of known class membership, training sites, are identified in the image and their remotely sensed response characterised from the sample of pixels they contain. The quantitative descriptions, training statistics, derived from the training stage are used in association with the selected classification decision rule to allocate each pixel in the image to the class with which it has greatest similarity in the class allocation stage of the analysis. In the final stage of a supervised classification, the accuracy of the thematic classification derived is assessed. To avoid optimistic bias (Hammond & Verbyla, 1996), accuracy assessment is typically based on a sample of pixels not used in training the classifier.

Many factors may act to limit classification accuracy below that possible. Consequently, researchers have sought means to increase the accuracy of thematic information extraction from remotely sensed imagery. Much research has, for example, focused on the classification technique used. Crucially, however, classification accuracy is dependent on a range of pre-class allocation issues. Of particular importance are factors connected with the training stage of the classification. Indeed the nature of the training stage can have a larger impact on classification accuracy than the classification technique used (Campbell, 2003; Hixon et al., 1980).

A range of issues must be considered in the design and execution of the training stage of a supervised image classification. It is generally perceived that the aim of the training stage is, essentially, to define an accurate model of the classes (Campbell, 2003; Kuo & Landgrebe, 2002). For each class, the training sample should, therefore, provide a representative description of the overall population, one that fully describes the classes spectrally (Chen & Stow, 2002; Eastman & Laney, 2002; Lillesand et al., 2004; Muchoney & Strahler, 2002). Thus, the literature often stresses the need to select training sites that are typical and representative of the classes in order to derive appropriate training statistics on which to base the classification (e.g., Campbell, 2003; McCaffrey & Franklin, 1993; Tso & Mather, 2001). This requires training sites to be sufficiently numerous and spread over the entire region to be mapped to accommodate for the effects of natural variability in the spectral response (Campbell, 2003; Mather, 2004).

Although classifiers differ in their training requirements (Mather, 2004), much of the literature focused on the specification of the training stage appears to be based on a classical statistical view of the classification process and often with a conventional probabilistic algorithm assumed as the classification technique to be employed. A variety of issues have been addressed in research, including the size and location of training sites and the composition of training sets but most attention has been directed at the issue of training set size, the number of training samples used. Thus, for example, with a maximum likelihood classification a key concern is to form a training set comprising of at least 10–30 independent training cases per class per discriminatory variable (e.g., waveband) to allow the formation of a representative description of the class so that its mean and variance can be reasonably estimated (Mather, 2004; Piper, 1992).

Many studies have investigated the effect of training set size on the accuracy of a variety of classification methods. In general, these studies have found that classification accuracy tends to be positively related to training set size (Arora & Foody, 1997; Foody & Mathur, 2004; Foody et al., 1995; Pal & Mather, 2003; Zhuang et al., 1994). The recommended size of a training set is often linked to important characteristics of the classifier and data set to be classified. For example, the dimensionality of the data set to be classified and complexity of the classifier have marked impacts on the size of the training set required for accurate classification (Kavzoglu & Mather, 2003; Tso & Mather, 2001). However, it is costly, notably in terms of time and finance, to acquire large training sets (Buchheim & Lillesand, 1989; Jackson & Landgrebe, 2001). Numerous classification analyses have, therefore, been undertaken with training sets that may be smaller than that which might be expected to be required for an accurate classification, although the actual size required is a function of factors such as the degree of interclass separability.

Sometimes the smallness of the training set represents a major problem to classification analyses and may result in classifications of low accuracy being derived (Bishop, 1995; Jackson & Landgrebe, 2002; Tadjudin & Landgrebe, 1999; 2000); this is especially apparent for analyses using hyperspectral sensor data. Many have, therefore, tried to reduce the need for large training sets. Studies have investigated a range of issues including the accommodation of spatial dependence in the data to define an efficient sampling design for training sample acquisition (Atkinson, 1991; Campbell, 1981; Webster et al., 1989), use of feature selection and feature extraction methods to reduce the dimensionality of the data set to be classified (Kuo & Landgrebe, 2002), use of unlabelled and semi-labelled samples in training as well as unsupervised classifications to help guide the analysis (Huang, 2002; Jackson & Landgrebe, 2002; Tadjudin & Landgrebe, 2000) and means to update and transfer training statistics derived from earlier investigations or other locations (Bruzzone & Pieto, 2001, 2002; Van Collie et al., in press; Wilkinson, 1997). An alternative approach to addressing the problems associated with small training sets is to recognise that the individual training samples vary in value and, from the outset, to seek the use of the most informative training samples only. By focusing on only those training samples that are helpful in fitting the decision boundary or hyperplane(s) that can separate the classes accurately, the total size of the training set needs not be large. This approach to training set selection is based on the perspective that the basic approach to training a supervised classifier outlined above is generally founded on the desire to fully describe each class in feature

space which may be an unnecessary and, in some instances, an unhelpful target.

Recently, many studies have shown that techniques such as evidential reasoning, neural networks, decision trees and support vector machines (SVM) may often be able to classify a data set to a higher accuracy than conventional statistical classifiers (Arora & Foody, 1997; Benediktsson et al., 1990; Foody & Mathur, 2004; Huang et al., 2002; Peddle et al., 1994). With these techniques, which are not based on an assumed parametric model, the standard requirement for a full and representative description of the spectral response of each class may no longer be necessary or appropriate in training the classification analysis. Thus, for some classifiers, an ideal training stage may differ markedly from that required for a conventional statistical classification. Given that these classifiers may be more accurate than the conventional and widely used probabilistic classifiers, they should be adopted increasingly and, critically, the design of their training stage constructed around their nature.

Good training statistics are those that may be used to derive an accurate classification (Bolstad & Lillesand, 1991). These statistics need not be derived from typical sites or provide a representative description of the classes. The key concern is that the training statistics should enable a classifier to accurately discriminate between cases of the different classes. There is, therefore, no requirement to fully describe the classes in feature space. In reality, the concern is sometimes to identify and characterise the remotely sensed response of sites that lie near to the locations of the classification hyperplanes or decision boundaries that partition feature space (Fig. 1). This focus on the vicinity of the hyperplanes, similar to aspects of decision boundary feature extraction (Lee & Landgrebe, 1993), may facilitate accurate classification. For example, it has been established that the most useful training samples for classification with a multilayer perceptron neural network are those that lie on the edge of the class distributions in feature space and between the distributions of two or more classes (Foody, 1999; Tambouratzis, 2000). These samples are termed here as border training samples and may be relatively atypical of their own class but are considerably more useful for an accurate classification than training samples that lie in the core area of the class distribution in feature space. Moreover, reducing the size of the training set by decreasing the number of training samples from the central core of the class in feature space, it is possible to speed up network training and facilitate the adoption of a simple network architecture without degrading classification accuracy (Tambouratzis, 2000). Such border training samples, however, provide a biased sample which, for example, would be inappropriate for training classical classification techniques such as the minimum distance to means and maximum likelihood classifiers. Thus, training samples are not equally informative and may vary in usefulness between classifiers.

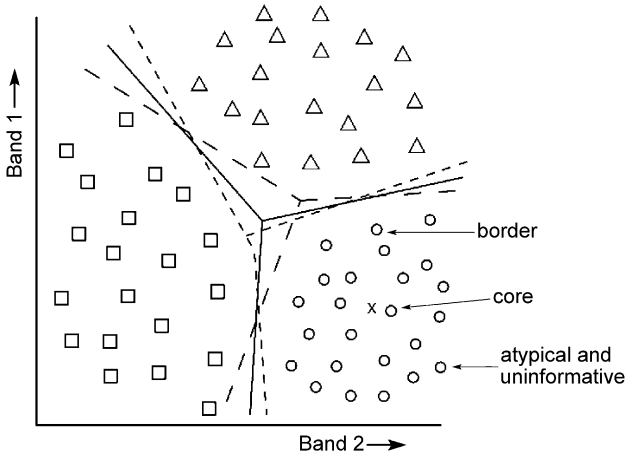

Fig. 1. Hyperplanes to separate the training samples of three classes in feature space (Note that, for brevity, the term hyperplane will be used throughout this article to indicate a decision boundary but applies strictly only when the decision surface has three or more dimensions). The three sets of lines plotted represent candidate hyperplanes that are each able to completely separate the three classes. Note, however, that the solid line represents the hyperplane with the greatest level of generalizability; with the other two hyperplanes, there is more chance of an unseen case lying on the wrong side of a hyperplane. For one of the classes, a training sample that may be considered to be a border training sample has been highlighted. Border training samples are the most useful for the SVM classification even though they are relatively atypical of the class. Also shown are a training sample from the class core (close to the class's centroid, x) and an atypical sample, each of which is uninformative in the SVM classification.

The unequalness of training samples in terms of their value for fitting an appropriate set of decision boundaries should be considered in the design of training data acquisition and refinement strategies for a classification analysis. Sometimes the methods used in training data collection or post-acquisition refinement act to emphasise the typical (average) response of a class. For example, various automated means of training site selection and growth from seeds utilise criteria that act to exclude relatively atypical sites (McCaffrey & Franklin, 1993). Similarly, the post-acquisition refinement of training statistics often focuses on the removal or down-weighting of training samples that appear to be atypical (Arai, 1992; Rosin & Fierens, 1998; Wang, 1990). Such approaches focus on the class centroid which may be useful with some classifiers but may also be undesirable for others. Training set acquisition and refinement operations that focus on the core of a class may, for example, actually act to reduce the number of very informative training samples for classifiers such as a SVM. More critically, with some classifiers it may be possible to intelligently select training samples and/or to intelligently revise training statistics prior to the class allocation stage of the analysis. Intelligent selection of training samples may allow small training sets to be used without loss of discriminatory power, substantially reducing the costs of training a classifier. Additionally, training samples once acquired may be intelligently refined, with those providing no or little useful contribution to the

placement of an accurate hyperplane in feature space deleted without loss of discriminatory ability but with possible gains in computational efficiency.

The potential for intelligent training is particularly evident for SVM-based classification as the process is based on the notion that only the training samples that lie on the class boundaries are necessary for discrimination. Consequently, only a small amount of the training samples typically acquired, perhaps a few percent, are actually used in fitting the hyperplanes (Brown et al., 1999) leading to the potential to define small training sets (Huang et al., 2002). The realization of this potential requires, however, a means to identify the useful training sites in advance of the class allocation stage of the classification (Foody & Mathur, 2004).

This paper aims to evaluate the potential to target training data collection to regions that may contain useful training samples at the expense of those that will contribute insignificantly to classification by a SVM. The approach is based on the use of knowledge on the variables that influence the spectral response of the classes.

A class typically occupies a relatively discrete area of feature space. The size and location of this area in feature space is determined by the spectral variation of the class which is a function of its intrinsic properties and the environmental context. For example, the spectral response of an agricultural crop class in an image might vary as a function of variables such as the crop's growth stage, topographic position, density of cover, health, impact of management activities, substrate conditions and sensor view angle. In some instances, cases of the crop with a particular variable set (e.g., crops growing on a particular soil) may be expected to lie generally within a limited part of the area occupied by the class as a whole in feature space. It is hypothesised that if cases of a class with a particular variable set are clustered in feature space and lie near the edge of the class distribution as well as close to that of another class, then these cases are the most useful for the specification of a hyperplane to separate the classes. Alternatively, those cases that are clustered away from the location of the hyperplane could be ignored, masked out of the training selection process without any significant loss of accuracy. Thus, if there is prior knowledge or some ancillary data set that can be used to help target the location of training sites to regions from which the most informative training samples can be derived, it may be possible to acquire a small, intelligently selected, training set that can be used to accurately classify imagery. This is evaluated here with a SVM classification of agricultural crops.

2. SVM classification

Recent research has indicated the considerable potential of SVM-based approaches for the supervised classification of remotely sensed data. Comparative studies have shown that classification by a SVM can be more accurate than popular contemporary techniques such as neural networks and decision trees as well as conventional probabilistic classifiers such as the maximum likelihood classification (Foody & Mathur, 2004; Huang et al., 2002). SVMs were deigned for binary classification but various methods exist to extend the binary approach to multiclass classification (Hsu & Lin, 2002; Huang et al., 2002). Here, for simplicity only the basic binary approach is considered to illustrate the potential for accurate classification based on small training sets through intelligent training sample selection. Through undertaking a set of binary analysis, however, a multiclass classification may be derived if desired.

The basis of SVM classification is illustrated in Fig. 2. In essence, the SVM classification is based on the notion of fitting an optimal separating hyperplane between classes by focusing on the training samples that lie at the edge of the class distributions, the support vectors. All of the other training samples are effectively discarded as they do not contribute to the estimation of hyperplane location (Belousov et al., 2002; Brown et al., 2000). In this way not only is an optimal hyperplane fitted, in the sense that it is expected to have a large degree of generalizability, but also a high accuracy may be obtained with the use of a small training set. Given the costs of training data acquisition, the potential to limit training set size may be an advantageous feature if it is achieved without any significant loss of discriminatory power in the final classification.

The simplest means of outlining the nature of classification with a SVM is for the situation in which the two

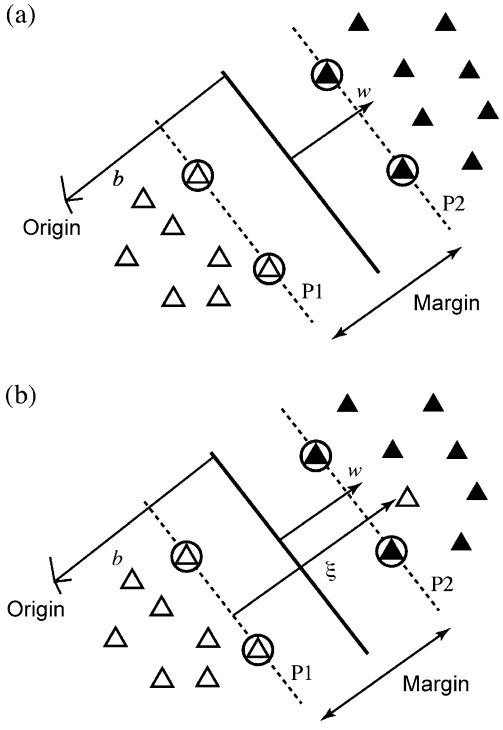

Fig. 2. Basis of SVM classification. (a) The linearly separable situation and (b) accommodation of outliers. Circled cases represent support vectors.

classes are linearly separable in q dimensional space. Using the training data represented by $\{x_i,y_i\}$, $i=1,\ldots,r$, $y_i\in\{1,-1\}$, the aim is to develop a classifier that generalizes accurately. For this, decision boundaries or hyperplanes that separate the classes in features space are required. Many candidate hyperplanes could be fitted to separate the classes but there is only one optimal separating hyperplane, which is expected to generalize well in comparison to other possible hyperplanes (Figs. 1 and 2). This optimal hyperplane should lie between the two classes in feature space and be positioned such that all of the samples of a class are located to one side of it and it is located such that the distance to the closest training data samples in both of the classes is as large as possible.

A hyperplane in feature space is defined by the equation $wx+b=0$, where x is a point lying on the hyperplane, w is normal to the hyperplane and b is the bias (Fig. 2). A separating hyperplane can be defined for the two classes as: $wx_i+b\geq+1$ (for the class $y_i=+1$) and $wx_i+b\leq-1$ (for the class $y_i=-1$). These two equations may be combined to give:

$$y_i(wx_i+b)-1\geq 0 \tag{1}$$

The training samples on these two hyperplanes are termed the support vectors and are central to the establishment of the optimal separating hyperplane that is used in SVM classification.

The support vectors of the two classes lie on two hyperplanes, which themselves are parallel to the optimal hyperplane and are defined by $wx_i+b=\pm 1$. The margin between these planes is $2/||w||$ and the analysis aims to maximise this margin through the constrained optimization problem,

$$\min\left\{\frac{1}{2}||w||^2\right\} \tag{2}$$

under the inequality constraints of Eq. (1).

If the classes are not linearly separable, slack variables, $\{\xi_i\}_{i=1}^{r}$, that indicate the distance the sample is from the hyperplane P1 or P2 passing through the support vectors of the class to which the sample belongs (Fig. 2), and so the amount of violation of the constraints allowed, is introduced. Eq. (1) may then be rewritten as,

$$y(wx_i+b)>1-\xi_i \tag{3}$$

If outliers exist in the data set, Eq. (3) can always be satisfied by making ξ_i very large and, so, a penalty term, $C\sum_{i=1}^{r}\xi_i$ is added to penalize solutions for which ξ_i are very large. The constant C controls the magnitude of the penalty associated with training samples that lie on the wrong side of the hyperplane. If C is very low an inappropriately large a fraction of support vectors may be derived. Alternatively, if C is very large the analysis may over-fit to the training data which may yield a low level of generalization ability. Although the parameter C must, therefore, be selected with care, there is a large degree of robustness of SVM-based classification to variation in its parameters (Belousov et al., 2002). With the addition of the penalty term to the analysis, the optimization problem becomes,

$$\min\left[\frac{||w||^2}{2}+C\sum_{i=1}^{r}\xi_i\right] \tag{4}$$

under the constraints of Eq. (3). The first part of Eq. (4) seeks to maximize the margin between the classes while the second part aims to penalize the samples located on the incorrect side of the hyperplane with C controlling the relative balance of these two competing objectives. If the classes overlap considerably in feature space, then $C\sum_{i=1}^{r}\xi_i$ can be very large and the optimal hyperplane may be expected to have limited generalization ability.

The basic approach to SVM classification may be extended to allow for nonlinear decision surfaces. For this, the input data are mapped into a high dimensional space through some nonlinear mapping which has the effect of spreading the distribution of the data points in a way that facilitates the fitting of a linear hyperplane. With this, the classification decision function becomes

$$f(x)=sgn\left(\sum_{i=1}^{r}\alpha_i y_i k(x,x_i)+b\right) \tag{5}$$

where α_i, $i=1,\ldots,r$ are Lagrange multipliers and $k(x,x_i)$ is a kernel function. The magnitude of α_i is determined by the parameter C and lies on a scale of 0–C (Belousov et al., 2002). The kernel used must meet Mercer's condition (Vapnik, 1995) and one such kernel that is widely used is the radial basis function,

$$k(x,x_i)=e^{-\gamma||(x-x_i)||^2} \tag{6}$$

where γ is the parameter controlling the width of the Gaussian kernel.

The accuracy with which a SVM may classify a data set is dependent on the magnitude of the parameters C and γ. With a large value of γ and/or C, there is a tendency for the SVM to over-fit to the training data, yielding a classifier that may generalize poorly. In this situation, it may be possible to classify the training data set accurately but the accuracy with which an independent testing set is classified may be low. Since the interest in remote sensing is typically to derive a classifier from a limited training set which may be applied usefully to other samples, the generalization ability of the classifier is of fundamental importance and hence over-fitting is undesirable. Consequently, the magnitude of C and γ must be determined carefully for the task in-hand. In practice, a large generalization ability may be obtained by, for example, setting γ appropriately for a given value of C. To help reduce subjectivity in setting the SVM parameters it is common for a cross-validation approach or a validation set, distinct from both the training and testing

sets, to be used to help select appropriate values for the parameters (Belousov et al., 2002).

3. Test site and data

Thematic classifications of imagery acquired for the region around the village of Feltwell, UK (Fig. 3) were undertaken. This test site was chosen for a variety of reasons. In particular, the test site is flat, removing undesired topographic influences in the remotely sensed response, and single crops are grown in large fields. Moreover, within a relatively small area there are marked differences in soil type that might impact considerably on the spectral response of the classes (Fig. 4). Here, attention focussed on the region comprising two main soil types. In the western part of the test site, the soils were predominantly earthy eutro-amorphous peats (hereafter referred to as peat soils) with the remainder of the sites covered mainly with brown rendzina soil with a small proportion of typical brown sands (hereafter referred to as brown soils). The peat and brown soils differed considerably in composition, notably in their organic matter content and texture which resulted in marked differences in appearance as well as spectral response (Fig. 4). Since the soil background to a crop will, depending on the extent of canopy cover and view angle effects, contribute to the overall spectral response observed from an area, it may have a substantial impact on the distribution of a class in feature space. The soil background may also affect the spectral response of a crop when the extent of ground cover is large or even complete through impacts on growth and condition. Differences in soil type may, therefore, allow stratification of the site for training data collection, allowing the potential to direct this activity to regions that may be expected to contain border training samples and away from regions that may provide relatively uninformative training samples. Thus knowledge of the soil type may open the way for accurate classification with small, intelligently defined, training sets.

The remotely sensed data used were extracted from a multispectral, three-waveband, SPOT HRV image of the site acquired on 16 June 1986 (Fig. 3). Near the time of image acquisition, a map depicting the crop type planted in each of the large fields was derived from traditional methods. This crop map was used as the ground data in support of the analyses undertaken.

The focus was on two classes, winter wheat and spring barley, that occupied much of the land at the test site. Winter wheat was grown across the entire test site while spring barley was restricted geographically, with fields planted to this crop typically constrained to the region with brown soil. At the time of image acquisition, these crops would have been at a mature growth stage with typically >95% ground cover. For each class, 75 pixels sampled from across the entire region were selected to form the training set for a SVM classification. This sample size lies within the range

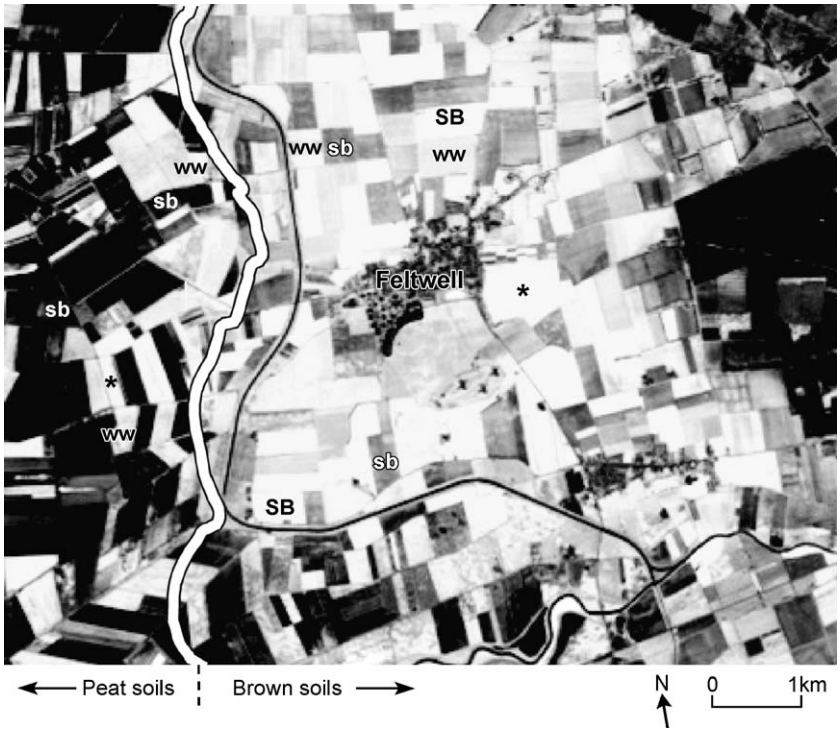

Fig. 3. Infrared image of the test site with some fields of winter wheat (WW) and spring barley (SB) highlighted as examples and the main division by soil type indicated. The variation in soil background had a marked effect on the observed spectral response. This was most apparent for fields that had been planted to sugar beet (sb) which, at the time of image acquisition, had very low ground cover (<20%) and, consequently, the soil background contributed markedly to the observed spectral response.

typically suggested for the classification of multispectral data, representing 25 times the number of spectral wavebands used in the analysis. In this training set, all the samples for the spring barley class were extracted from the region with brown soils. For the winter wheat class, however, 40 training samples were drawn from the region with brown soils and the remaining 35 from that with the dark toned peat soils.

Initial classifications focused on the accuracy with which the two crops could be discriminated. The analyses were then repeated with the training samples for the winter wheat class on the peat soil, which were hypothesised from the class distributions in feature space (Fig. 5) to be relatively uninformative, deleted from the training set. Since imbalanced training sets can sometimes degrade classification accuracy, a further set of analyses were undertaken with balanced training sets. For this, the 35 training samples of winter wheat on peat soil removed from the analysis were replaced with an additional set of 35 samples of winter wheat drawn from the region of brown soil.

For comparative purposes, a further classification was undertaken in which the winter wheat training samples drawn from the region of brown soil were deleted. Thus, in this classification only the training samples for winter wheat drawn from the region with peat soil were used to train a SVM classification to classify the winter wheat and spring barley classes.

Fig. 4. Ground photographs of the test site to show different soil types acquired in the year after image acquisition: (a) brown soil; (b) peat soil. The location of the fields shown is indicated in Fig. 3 with the * symbols.

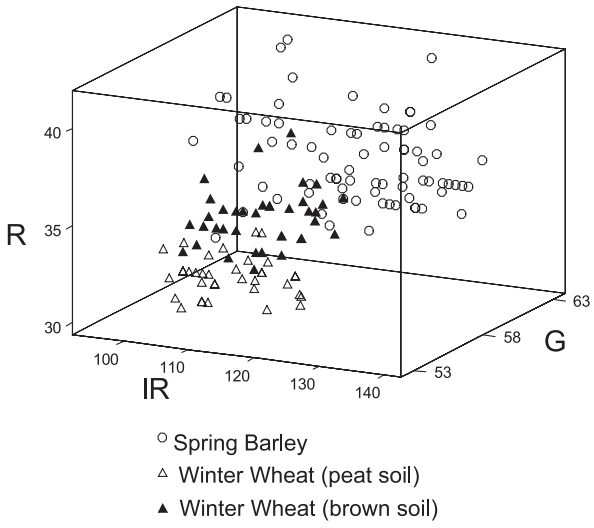

Fig. 5. Distribution of the training samples in feature space (IR—infrared; R—red; G—green).

The training sets were used to drive a series of SVM classifications. The values of the SVM parameters, C and γ, were determined using a fivefold cross-validation approach. From this analysis, the values of C and γ were set at 16 and 0.0039, respectively. Note that because of this parameter set, the magnitude of α_i lies on a scale 0–16. The magnitude of α_i determined for each training sample used in the analysis is positively related to the contribution made by that training sample to the fitting of the hyperplane and so to the classification derived. This is evident from Eq. (5), with training samples for which α_i=0 making no contribution to the fitting of the hyperplane. Such training samples carry no useful information, unlike those that lie in the border region where α_i tends to its maximum value.

The accuracy of each classification was assessed using the same testing set. This testing set comprised 40 pixels of each class. Of these, 20 of the pixels representing the winter wheat class were drawn from the region with peat soils, all of the remaining testing pixels were drawn from the region with brown soil. Classification accuracy was expressed in terms of the percentage of testing cases correctly allocated by the classifications. The statistical significance of differences in classification accuracy was assessed with an approach based on McNemar's test. This approach was adopted since it recognises the use of the same testing sample in the accuracy assessment of each classification compared (Foody, 2004).

4. Results and discussion

The feature space plot of the data extracted showed that soil type had a marked impact on the spectral response of the classes. For the winter wheat class, which was found

on both soil types, it was evident that the area it covered in feature space was comprised of effectively two clusters (Fig. 5). Moreover, one of these clusters of winter wheat samples, containing the samples drawn from the region with brown soil, lay between the other and the distribution of the samples of the spring barley class. Since there is no need to fully describe classes in feature space but only focus on the border region between the classes with a SVM, it may be anticipated that the training samples for winter wheat drawn from the region with peat soils contributed little or nothing to the fitting of the hyperplane between the two classes. Evidence for this view was apparent in the results of the SVM classifications.

The initial SVM classification was accurate, with 92.5% of cases in the testing set correctly allocated. Although this classification was based on a training set comprising 150 pixels (75 of each class) only 37 of these training samples were identified as support vectors in the analysis. Thus, the fitting of the hyperplane was based on just one quarter of the training set, the remaining training samples made no contribution to the analysis. Furthermore, it was apparent that only one of these support vectors corresponded to a training sample that had been drawn from the region with peat soil (Table 1). Moreover, this single support vector drawn from the region of peat soils had the smallest α_i, 4.3247, and so made the least contribution of all the support vectors to the fitting of the hyperplane between the classes.

Repeating the SVM classification but with the training samples for the winter wheat class drawn from the region of peat soils excluded resulted in the same set of class allocations. Thus, classification accuracy was maintained despite the reduction in the training set size. Similarly, replacing the training samples excluded with additional training samples for the winter wheat class drawn from the region of brown soils had no significant effect on classification accuracy, with the exact same set of class allocations made for the testing cases as before.

The importance of selecting useful training sites was particularly evident in the results of the classification in which the winter wheat class was trained upon only samples drawn from the region with peat soil. In this instance, the two classes, spring barley and winter wheat (on peat), were very separable (Fig. 5). It was, therefore, unsurprising that a hyperplane could be fitted to the training data set that allowed the training samples to be classed to a very high accuracy (99.1%). However, relative to the hyperplanes fitted in earlier classifications, the hyperplane was shifted away from the spring barley class. The hyperplane now cut through the region of feature space occupied by the winter wheat on brown soil. Thus, when this hyperplane was used to classify the testing set, which contained cases of winter wheat from the region with brown soil, misclassification was common. The accuracy with which the testing set was classified, 73.8%, was significantly less ($p<0.05$) than when training samples of winter wheat on brown soil were included in the analysis. This result emphasised the need to ensure that the training samples that lie at the informative edge of the class distribution in feature space are included in the analysis. Their absence can lead to the hyperplane being fitted inappropriately resulting in a classification of low accuracy when the classes may actually display a large degree of separability.

Table 1
Summary of the 37 support vectors identified in the analysis, ordered from top to bottom by decreasing importance

Class	Soil Type	α_i
Spring barley	Brown	16.0000
Spring barley	Brown	16.0000
Spring barley	Brown	16.0000
Spring barley	Brown	16.0000
Spring barley	Brown	16.0000
Spring barley	Brown	16.0000
Spring barley	Brown	16.0000
Spring barley	Brown	16.0000
Spring barley	Brown	16.0000
Spring barley	Brown	16.0000
Spring barley	Brown	16.0000
Spring barley	Brown	16.0000
Spring barley	Brown	16.0000
Winter wheat	Brown	16.0000
Winter wheat	Brown	16.0000
Winter wheat	Brown	16.0000
Winter wheat	Brown	16.0000
Winter wheat	Brown	16.0000
Winter wheat	Brown	16.0000
Winter wheat	Brown	16.0000
Winter wheat	Brown	16.0000
Winter wheat	Brown	16.0000
Winter wheat	Brown	16.0000
Winter wheat	Brown	16.0000
Winter wheat	Brown	16.0000
Winter wheat	Brown	16.0000
Winter wheat	Brown	16.0000
Winter wheat	Brown	16.0000
Spring barley	Brown	11.8711
Spring barley	Brown	9.7442
Spring barley	Brown	8.0480
Spring barley	Brown	7.5299
Spring barley	Brown	6.5543
Winter wheat	Brown	6.5322
Spring barley	Brown	4.4947
Winter wheat	Brown	4.4371
Winter wheat	Peat	4.3247

Note that the only support vector relating to a sample drawn from the region on peat soil was the least important.

The results show that the use of ancillary data can allow the analyst to focus training site selection activities on regions likely to furnish useful support vectors. This, therefore, allows the definition of a small training set that may be used to derive an accurate classification.

5. Summary and conclusions

The standard desire in training a supervised classifier has been to derive an accurate and complete description of the spectral properties of each class. For brevity, assuming

normally distributed data, a class may then be described by its mean and variance. To derive an accurate description of a class's spectral response, it has been suggested that a large number of training samples be acquired from sites distributed across the entire region, to capture geographical variation. Such approaches will help ensure accurate estimates of the mean and class variance are derived, which is important if using a conventional statistical classifier such as the maximum likelihood classification. Moreover, some training data acquisition approaches and post-acquisition training statistic refinement operations often seek to down-weight the contribution of outlying training samples in favour of those that lie closer to the class mean or centroid. While such an approach may be appropriate with some classification approaches, it need not always be appropriate or necessary. With some classifiers, for example, there is no requirement to obtain a complete description of each class's spectral response. For example, with a basic minimum distance to means classifier, the focus is on the class centroid while with a parallepiped classifier the focus is on the extreme values. With classifiers such as a SVM, only the training samples that lie close to the hyperplanes that separate the classes in feature space are necessary. Thus, individual training samples may vary greatly in importance and value. Furthermore, this relative importance of the training sites is also a function of the specific classifier to be trained. If the most informative training samples can be identified, pre- or post-acquisition, this can be beneficial to a classification analysis. Critically, if the regions expected to contain informative training sites can be identified in advance of the training data collection programme, then it may be possible to reduce the number of training sites and samples acquired without loss of accuracy. This is particularly important as training data are often in short supply and so targeting the most informative training sites may be beneficial.

The potential to identify the most informative and useful training samples was illustrated with SVM classifications of agricultural crops. Key points of note from the analyses were:

(1) Soil type had a marked effect on the spectral response of the winter wheat class, such that in feature space the samples drawn from a region of brown soil lay close to the spring barley class while samples drawn from the peat soil were distant from the spring barley class. Given the nature of SVM classification, it would, therefore, be expected that the samples of winter wheat drawn from the region of brown soil would furnish useful support vectors while samples drawn from the region of peat soil would be relatively uninformative. This information can be used to help direct training data acquisition activities to sites where informative training samples may be most likely to lie.
(2) A SVM trained using samples drawn from both soil types classified the testing set with an accuracy of 92.5%. In this classification, only 37 of the 150 training samples were support vectors and made a contribution to the fitting of the hyperplane to separate the winter wheat and spring barley classes. Thus, approximately three quarters of the training set made no contribution to the classification; accurate classification can be achieved with a small training set.
(3) Removing the training samples of winter wheat drawn from the region with peat soil from the training set had no significant effect on classification accuracy.
(4) On the basis of the distribution of winter wheat samples drawn from the two soil types in feature space, it was, therefore, possible to intelligently direct training data acquisition activities to regions that may furnish useful training samples (i.e., to the brown soil region) and away from sites that may provide relatively uninformative training samples (i.e., the region of peat soil). Thus, if returning to this test site in the future to map the same classes, the analyst would be able to restrict the training data acquisition activities for the winter wheat class to just the region with brown soil without compromising on classification accuracy. If classifying another site, the analyst could attempt to deduce the effect of site variables on the spectral response of the classes present and use this information to target the training sample acquisition activities. Alternatively, a small pre-training survey could be undertaken to evaluate the distribution of the classes in feature space and direct the training data acquisition strategy.

The key conclusion of the analyses is that a complete description of each class in feature space is not required for an accurate classification. With a SVM, only training samples located near the hyperplane position are required with other samples not contributing the analysis. The acquisition of training samples from beyond the border region is, therefore, unnecessary and wasted effort. Thus in situations such as in the example discussed above, when knowledge of a major difference in a variable such as soil type that may impact on the spectral response of a class is available, this could be used to direct training activities. In particular, this knowledge may allow small, intelligently selected, training samples to be acquired that may be used to discriminate between the classes as accurately as a much larger, unintelligently selected, sample. Knowledge may, therefore, be used to reduce training set size without loss of classification accuracy by directing the training site acquisition process. Having shown an initial step toward intelligent training, that some training sites are more useful than others and may be identified in advance of the class allocation stage of the classification analysis, current work aims to extend and test the approach to a multiclass classification undertaken for an operational application.

Acknowledgements

The SPOT HRV data were supplied by the NERC EO data centre and the crop map acquired through involvement in the European AgriSAR campaign. The information on soils was derived from a map produced by the Soil Survey of England and Wales. We are grateful to the Commonwealth Scholarship and Fellowship plan administered by the British Council for provision of a scholarship to support AM's work for a PhD degree at the University of Southampton while on leave from the Punjab Remote Sensing Centre, India.

References

Arai, K. (1992). A supervised Thematic Mapper classification with a purification of training samples. *International Journal of Remote Sensing*, *13*, 2039–2049.

Arora, M. K., & Foody, G. M. (1997). Log-linear modelling for the evaluation of the variables affecting the accuracy of probabilistic, fuzzy and neural network classifications. *International Journal of Remote Sensing*, *18*, 785–798.

Atkinson, P. M. (1991). Optimal ground-based sampling for remote-sensing investigations—estimating the regional mean. *International Journal of Remote Sensing*, *12*, 559–567.

Belousov, A. I., Verzakov, S. A., & von Frese, J. (2002). A flexible classification approach with optimal generalisation performance: Support vector machines. *Chemometrics and Intelligent Laboratory Systems*, *64*, 15–25.

Benediktsson, J. A., Swain, P. H., & Ersoy, O. K. (1990). Neural network approaches versus statistical-methods in classification of multisource remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, *28*, 540–552.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Oxford Univ. Press.

Bolstad, P. V., & Lillesand, T. M. (1991). Semi-automated training approaches for spectral class definition. *International Journal of Remote Sensing*, *13*, 3157–3166.

Buchheim, M. P., & Lillesand, T. M. (1989). Semi-automated training field extraction and analysis for efficient digital image classification. *Photogrammetric Engineering and Remote Sensing*, *55*, 1347–1355.

Brown, M., Gunn, S. R., & Lewis, H. G. (1999). Support vector machines for optimal classification and spectral unmixing. *Ecological Modelling*, *120*, 167–179.

Brown, M., Lewis, H. G., & Gunn, S. R. (2000). Linear spectral mixture models and support vector machines for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, *38*, 2346–2360.

Bruzzone, L., & Pieto, D. F. (2001). Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, *39*, 456–460.

Bruzzone, L., & Pieto, D. F. (2002). A partially unsupervised cascade classifier for the analysis of multitemporal remote-sensing images. *Pattern Recognition Letters*, *23*, 1063–1071.

Campbell, J. B. (1981). Spatial correlation-effects upon accuracy of supervised classification of land cover. *Photogrammetric Engineering and Remote Sensing*, *47*, 355–363.

Campbell, J. B. (2003). *Introduction to remote sensing* (3rd ed.). London: Taylor and Francis.

Chen, D. M., & Stow, D. (2002). The effect of training strategies on supervised classification at different spatial resolutions. *Photogrammetric Engineering and Remote Sensing*, *68*, 1155–1161.

Eastman, J. R., & Laney, R. M. (2002). Bayesian soft classification for sub-pixel analysis: A critical evaluation. *Photogrammetric Engineering and Remote Sensing*, *68*, 1149–1154.

Foody, G. M. (1999). The significance of border training patterns in classification by a feedforward neural network using backpropagation learning. *International Journal of Remote Sensing*, *20*, 3549–3562.

Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, *80*, 185–201.

Foody, G. M. (2004). Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy. *Photogrammetric Engineering and Remote Sensing*, *70*, 627–633.

Foody, G. M., & Mathur, A. (2004). A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, *42*, 1335–1343.

Foody, G. M., McCulloch, M. B., & Yates, W. B. (1995). The effect of training set size and composition on artificial neural network classification. *International Journal of Remote Sensing*, *16*, 1707–1723.

Hammond, T. O., & Verbyla, D. L. (1996). Optimistic bias in classification accuracy assessment. *International Journal of Remote Sensing*, *17*, 1261–1266.

Hixon, M., Scholz, D., & Fuhs, N. (1980). Evaluation of several schemes for classification of remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, *46*, 1547–1553.

Hsu, C. -W., & Lin, C. -J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, *13*, 415–425.

Huang, C., Davis, L. S., & Townshend, J. R. G. (2002). An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, *23*, 725–749.

Huang, K. Y. (2002). The use of a newly developed algorithm of divisive hierarchical clustering for remote sensing image analysis. *International Journal of Remote Sensing*, *23*, 3149–3168.

Jackson, Q., & Landgrebe, D. (2001). An adaptive classifier design for high-dimensional data analysis with a limited training data set. *IEEE Transactions on Geoscience and Remote Sensing*, *39*, 2664–2679.

Jackson, Q., & Landgrebe, D. (2002). An adaptive method for combined covariance estimation and classification. *IEEE Transactions on Geoscience and Remote Sensing*, *40*, 1082–1087.

Kavzoglu, T., & Mather, P. M. (2003). The use of backpropagating artificial neural networks in land cover classification. *International Journal of Remote Sensing*, *24*, 4907–4938.

Kuo, B. C., & Landgrebe, D. A. (2002). A covariance estimator for small sample size classification problems and its application to feature extraction. *IEEE Transactions on Geoscience and Remote Sensing*, *40*, 814–819.

Lee, C., & Landgrebe, D. A. (1993). Decision boundary feature extraction for nonparametric classification. *IEEE Transactions on Systems, Man and Cybernetics*, *23*, 433–444.

Lillesand, T. M., Kiefer, R. W., & Chipman, J. W. (2004). *Remote sensing and image interpretation* (5th ed.). New York: Wiley.

Mather, P. M. (2004). *Computer processing of remotely sensed images* (3rd ed.). Chichester: Wiley.

McCaffrey, T. M., & Franklin, S. E. (1993). Automated training site selection for large-area remote sensing image analysis. *Computers & Geosciences*, *19*, 1413–1428.

Muchoney, D. M., & Strahler, A. H. (2002). Pixel- and site-based calibration and validation methods for evaluating supervised classification of remotely sensed data. *Remote Sensing of Environment*, *81*, 290–299.

Pal, M., & Mather, P. M. (2003). An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sensing of Environment*, *86*, 554–565.

Peddle, D. R., Foody, G. M., Zhang, A., Franklin, S. E., & LeDrew, E. F. (1994). Multisource image classification: II. An empirical comparison of evidential reasoning and neural network approaches. *Canadian Journal of Remote Sensing*, *20*, 396–407.

Piper, J. (1992). Variability and bias in experimentally measured classifier error rates. *Pattern Recognition Letters*, *13*, 685–692.

Rosin, P. L., & Fierens, F. (1998). The effects of data filtering on neural network learning. *Neurocomputing*, *20*, 155–162.

Tadjudin, S., & Landgrebe, D. A. (1999). Covariance estimation with limited training samples. *IEEE Transactions on Geoscience and Remote Sensing*, *37*, 2113–2118.

Tadjudin, S., & Landgrebe, D. A. (2000). Robust parameter estimation for mixture model. *IEEE Transactions on Geoscience and Remote Sensing*, *38*, 439–445.

Tambouratzis, T. (2000). Counter-clustering for training pattern selection. *Computer Journal*, *43*, 177–190.

Tso, B., & Mather, P. M. (2001). *Classification methods for remotely sensed data*. London: Taylor and Francis.

Van Collie, F. M. B., Verbeke, L. P. C., & De Wulf, R. R. (2004). Previously trained neural networks as ensemble members: Knowledge extraction and transfer. *International Journal of Remote Sensing*, (in press).

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Wang, F. (1990). Fuzzy supervised classification of remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, *28*, 194–201.

Webster, R., Curran, P. J., & Munden, J. W. (1989). Spatial correlation in reflected radiation from the ground and its implications for sampling and mapping by ground-based radiometry. *Remote Sensing of Environment*, *29*, 67–78.

Wilkinson, G. G. (1997). Open questions in neurocomputing for Earth observation. In I. Kanellopoulos, G. G. Wilkinson, F. Roli, & J. Austin (Eds.), *Neurocomputation in remote sensing data analysis* (pp. 3–13). Berlin: Springer.

Zhuang, X., Engel, B. A., Lozanogarcia, D. F., Fernandez, R. N., & Johannsen, C. J. (1994). Optimization of training data required for neuro-classification. *International Journal of Remote Sensing*, *15*, 3271–3277.