A Fast Separability-Based Feature Selection Method for High-Dimensional Remotely-Sensed Image Classification

Baofeng Guo, R. I. Damper, Steve R. Gunn and J. D. B. Nelson

Information: Signals, Systems, Images (ISIS) Research Group
School of Electronics and Computer Science
University of Southampton,
Southampton SO17 1BJ, United Kingdom
Emails: {bg|rid|srg|jn}@ecs.soton.ac.uk

Abstract

Because of the difficulty of obtaining an analytic expression for Bayes error, a wide variety of separability measures has been proposed for feature selection. In this paper, we show that there is a general framework based on the criterion of mutual information (MI) that can provide a realistic solution to the problem of feature selection for high-dimensional data. We give a theoretical argument showing that the mutual information of multi-dimensional data can be broken down into several one-dimensional components, which makes numerical evaluation much easier and more accurate. It also reveals that selection based on the simple criterion of only retaining features with high associated MI values may be problematic when the features are highly correlated. Although there is a direct way of selecting features by jointly maximising mutual information, this suffers from combinatorial explosion. Hence, we propose a fast feature selection scheme based on a 'greedy' optimisation strategy. To confirm the effectiveness of this scheme, simulations are carried out on 16 land-cover classes using the 92AV3C dataset collected from the 220-dimensional AVIRIS hyperspectral sensor. We replicate our earlier positive results (which used an essentially heuristic method for MI-based band-selection) but with much reduced computational cost and a much sounder theoretical basis.

Key words:
Feature selection, mutual information, remote sensing, hyperspectral image classification.

1 Introduction

Feature selection involves choosing a subset of features that will best represent the original data under a certain criterion. It is an important pre-processing step for classifiers, to discard irrelevant and redundant features that may affect classifier performance and efficiency [1,2]. In this paper, we investigate a computationally-efficient solution to the problem of feature selection for image classification, and test it on high-dimensional remotely-sensed data.

In remote-sensing research, hyperspectral sensors observe the earth's surface by simultaneously sampling hundreds of contiguous spectral bands with a fine resolution, e.g., 0.01 μm. For instance, the AVIRIS hyperspectral sensor [3] has 224 spectral bands ranging from 0.4 μm to 2.5 μm. Such a large number of bands implies high-dimensionality data, presenting several significant challenges to image classification. First, it is well known that the dimensionality of input space strongly affects the performance of many supervised classification methods [4]. Second, because hyperspectral data are sensed in contiguous, finely-spaced bands, there is almost certain to be redundancy between them. Third, complex atmospheric transmission and interference means some bands contain less discriminatory information than others. Finally, high-dimensional data impose requirements for storage space, computational load and communication bandwidth that tell against time-critical applications. It is therefore advantageous to remove bands that convey little or no discriminatory information. Many band-selection techniques have been proposed [5–15]. However, there are still some challenges to apply these techniques effectively, such as high computational cost, presence of local minima problems, difficulties for real-time implementation, etc.

Two main factors in feature selection are selection criterion as it affects accuracy, and searching algorithm as it affects speed. Bayes error is the generally-agreed ideal criterion to guide selection, but it is difficult to obtain an explicit and analytic expression for this. Hence, alternatives are sought such as separability measures, which usually provide bounds around the Bayes error. In this particular research, dimensionality-reduction techniques such as principal components have been deliberately avoided so as to retain the raw hyperspectral data for purposes of registration with other source images (e.g., SAR imagery), and not to lose interpretability of the original physical meaning (e.g., surface temperature).

Like many separability measures, mutual information (MI) evaluates the statistical dependence between two random variables and so can be used to measure the utility of selected features to classification. Mutual information has obvious potential for band selection [8,16,17], but this has not been fully exploited in the past. Several valuable techniques [18,19] have been developed to use mutual information for feature selection, which usually require strong assumptions and approximations to be made. A well-found theoretical framework is still absent, and existing techniques may not be as effective as they could be when the application scenario changes or the underlying assumptions do not hold.

In this paper, we propose a theoretically-consistent solution to the use of mutual information for feature selection. First, several commonly-used separability measures are examined and compared with the mutual information. It is then argued that MI is a suitable choice for multi-class problems and non-Gaussian data. Subsequently, we provide theoretical arguments showing that the evaluation of high-dimensional MI can be simplified by evaluating several one-dimensional MI terms. This paves a way for an effective solution when it is difficult to obtain enough training data to validate the estimation of high-dimensional MI. It also gives insight to explain why some 'good' individual features could be 'bad' when employed jointly. Finally, a fast feature-selection method is proposed by using greedy optimisation. In this method, features are sequentially selected not only on the basis of their associated MI values but also their 'complementary' level to the already selected ones. The proposed method avoids iterative searching and intensive matrix operations (e.g., matrix inverse and determinant), and so provides a low computational cost solution for time-critical applications. The underlying idea of this method can be extended to other separability measures as well. Besides the tested hyperspectral data, it is also applicable to other high-dimensional data.

The remainder of this paper is organised as follows. After a brief introduction to typical separability measures in Section 2, we derive a theoretical framework for the use of mutual information for feature selection in Section 3. In Section 4, we propose a greedy optimisation strategy to maximise the mutual information between the selected features and the class label. Simulations are carried out to test the performance of the proposed method based on 220-dimensional remotely-sensed data, which are presented in Section 5. Finally, Section 6 concludes.

2 Feature selection based on separability measures

Let $\mathbf{x}' = (x'_1, x'_2, \ldots, x'_N)$ be a N-dimensional data vector, with each component $x'_i \in \mathbb{R}$ standing for an observed variable. For remotely sensed data, this could be a spectral reflectance value measured at band i, $i = 1, 2, \ldots, N$. The objective of feature selection is to find a subset of the components, $\mathbf{x} = (x_1, x_2, \ldots, x_M)$, $M < N,\ x_j \in \{x'_1, x'_2, \ldots, x'_N\},\ j = 1, 2, \ldots, M$ satisfying a certain cost criterion, $J(\,)$, such as:

$$J(\mathbf{x}^0) = \min_{\mathbf{x} \in \Xi} J(\mathbf{x})$$

where Ξ is a set of any M-dimensional vectors selected from the original N-dimensional vector $\mathbf{x}'$. The size of set Ξ, i.e., the number of all possible combinations is $\binom{N}{M}$.

In image classification, the ideal criterion for feature selection should be Bayes error. But direct minimisation of the Bayes error cannot be analytically performed, so a wide range of alternative criteria that are easier to evaluate have been proposed.

Generally speaking, these practical criteria fall into three major categories: probabilistic distance, divergence and correlation-based. Because the Bayes error may be bounded by these easily-evaluated criteria, a realistic performance can be expected using these approximations to Bayes error. Typical measures of probabilistic distance are:

$$\text{Chernoff:} \quad D_{\mathrm{C}} = -\log \int_{\mathbf{x}} p^{\alpha}(\mathbf{x}|\omega_1) p^{1-\alpha}(\mathbf{x}|\omega_2)\, \mathrm{d}\mathbf{x}$$

$$\text{Bhattacharyya:} \quad D_{\mathrm{B}} = -\log \int_{\mathbf{x}} \sqrt{p(\mathbf{x}|\omega_1) p(\mathbf{x}|\omega_2)}\, \mathrm{d}\mathbf{x}$$

$$\text{Jeffreys-Matusita:} \quad D_{\mathrm{JM}} = \left\{ \int_{\mathbf{x}} \left[\sqrt{p(\mathbf{x}|\omega_1)} - \sqrt{p(\mathbf{x}|\omega_2)} \right]^2 \mathrm{d}\mathbf{x} \right\}^{1/2}$$

where $0 < \alpha < 1$; $p(\mathbf{x}|\omega_1)$ and $p(\mathbf{x}|\omega_2)$ are the conditional density functions given two classes ω_1 and ω_2, respectively. It is seen that the Bhattacharyya distance is a special case of Chernoff distance when $\alpha = 0.5$.

The most-used divergence measures are the Kullback-Leibler divergence:

$$D_{\mathrm{KL}}(p(\mathbf{x}|\omega_1) \| p(\mathbf{x}|\omega_2)) = \int_{\mathbf{x}} p(\mathbf{x}|\omega_1) \log \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)}\, \mathrm{d}\mathbf{x}$$

and the symmetric Kullback-Leibler divergence:

$$D_{\mathrm{SKL}}(p(\mathbf{x}|\omega_1) \| p(\mathbf{x}|\omega_2)) = \int_{\mathbf{x}} [p(\mathbf{x}|\omega_1) - p(\mathbf{x}|\omega_2)] \cdot \log \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)}\, \mathrm{d}\mathbf{x}$$

These existing criteria are closely related to the Bayes error, but they are often difficult to evaluate when data are non-Gaussian distributed. Moreover, most of these distances or divergences are defined based on two-class problems, i.e., they are naturally pairwise measures. Their extensions to the multi-class case, a common scenario in remotely-sensed image classification, are usually averages of all pairwise measures. This increases computational burden in feature selection enormously. A possible inconsistency also arises, in that particular features might contribute quite differently to classification performance in different pairwise measures. This will introduce a further dilemma; how should this be treated in feature selection? Hence, it is preferable to introduce a more suitable separability measure for the problem of multi-class and non-Gaussian data.

In information theory, mutual information is a quantity that measures the mutual dependence of the two variables, and is defined as:

$$I(X,Y) = \int_Y \int_X p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}\, \mathrm{d}x\, \mathrm{d}y \tag{1}$$

where $p(x, y)$ is the joint probability density function of continual random variables X and Y, and $p(x)$ and $p(y)$ are the marginal probability density functions respectively. Mutual information is related to entropy as:

$$\begin{aligned} I(X,Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{2}$$

given the Shannon entropy (discrete) defined as:

$$H(X) = -\sum_X p(x) \log p(x)$$

Mutual information is related to the Kullback-Leibler divergence, in the form of divergence of the product of the marginal distributions $p(x)$ and $p(y)$, from the joint distribution $p(x, y)$:

$$I(X,Y) = D_{\mathrm{KL}}(p(x,y) \| p(x)p(y)) \tag{3}$$

There are two strong motivations for us to consider mutual information as the selection criterion in this work, The first is its natural suitability for multi-class problems; the second is its simplicity of evaluation for non-Gaussian data. From (1) or (3) it can be seen that, contrary to the separability measures discussed previously, MI is not evaluated based on two conditional probability densities but between the joint probability distribution function and the product of marginal probability distribution functions. Thus, MI is not evaluated pairwise, and this fact makes it automatically suited to multi-class problems. Mutual information can be effectively evaluated as in (2), based on non-parametric density estimation methods such as two-dimensional histograms [18,20]. Therefore, the technique is no longer confined to Gaussian data.

In applying MI to remotely-sensed image classification, we can treat the spectral signal as a random variable X with continuous values of spectral reflectance, and its ground truth (i.e., the corresponding reference map—see later) as Y with discrete class labels $\omega_1, \omega_2, \ldots, \omega_n$. Thus, MI can be used to estimate the dependency between them as:

$$
\begin{aligned}
I(X, Y) = &- \int_X p(x) \log p(x) \, \mathrm{d}x - \sum_{\mathbf{y}} P(\mathbf{y}) \log P(\mathbf{y}) \\
&+ \sum_{\mathbf{y}} \int_X p(x, \mathbf{y}) \log p(x, \mathbf{y}) \, \mathrm{d}x
\end{aligned}
$$

Since the ground truth defines the required classification result, MI measures the capability of using the spectral signal to predict the classification objective.

Let $Y = \{\omega_1, \omega_2, \ldots, \omega_n\}$. For a multi-class problem, it can be shown from (2) that MI will be:

$$
\begin{aligned}
I(X, Y) &= H(X) - H(X|Y) \\
&= - \int_X p(x) \log p(x) \mathrm{d}x + \sum_{\omega_i} \int_X p(x|\omega_i) \log p(x|\omega_i) \mathrm{d}x
\end{aligned} \tag{4}
$$

Based on (4), we can further explore the relation between MI and other separability measures. For the two-class problem, i.e., $\mathbf{y} = \{\omega_1, \omega_2\}$,

$$
\begin{aligned}
-H(X|Y) &= \sum_{\mathbf{y}} \int_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) \log p(\mathbf{x}|\mathbf{y}) \mathrm{d}\mathbf{x} \\
&= \int_{\mathbf{x}} p(\mathbf{x}|\omega_1) \log p(\mathbf{x}|\omega_1) \mathrm{d}\mathbf{x} + \int_{\mathbf{x}} p(\mathbf{x}|\omega_2) \log p(\mathbf{x}|\omega_2) \mathrm{d}\mathbf{x} \\
&\leq \int_{\mathbf{x}} p(\mathbf{x}|\omega_1) \log p(\mathbf{x}|\omega_1) \mathrm{d}\mathbf{x} + \int_{\mathbf{x}} p(\mathbf{x}|\omega_2) \log p(\mathbf{x}|\omega_2) \mathrm{d}\mathbf{x} \\
&\qquad - \int_{\mathbf{x}} p(\mathbf{x}|\omega_1) \log p(\mathbf{x}|\omega_2) \mathrm{d}\mathbf{x} - \int_{\mathbf{x}} p(\mathbf{x}|\omega_2) \log p(\mathbf{x}|\omega_1) \mathrm{d}\mathbf{x} \\
&= \int_{\mathbf{x}} [p(\mathbf{x}|\omega_1) - p(\mathbf{x}|\omega_2)] \cdot \big[\log p(\mathbf{x}|\omega_1) - \log p(\mathbf{x}|\omega_2)\big] \, \mathrm{d}\mathbf{x} \\
&= D_{\mathrm{SKL}}(p(\mathbf{x}|\omega_1) \| p(\mathbf{x}|\omega_2))
\end{aligned}
$$

So given a fixed $H(x)$, MI is lower-bounded by the symmetric Kullback-Leibler divergence. It is known [21] that the Jeffreys-Matusita distance is also lower-bounded by the symmetric Kullback-Leibler divergence as:

$$
D_{\mathrm{JM}} \leq 2 \left[1 - \exp\left(-\frac{D_{\mathrm{SKL}}}{8} \right) \right]
$$

which implies the underlying connections between MI, probabilistic distances and Bayes error [22,23].

3 Feature selection through manipulating MI

With the general framework discussed above, feature selection based on MI can be described as follows: Given a set of original data vectors $\mathbf{x}'$ with N components or variables, and $\mathbf{y}$ the corresponding output class label (e.g., the ground truth), find a subset of variables $\mathbf{x} \subset \mathbf{x}'$ with M components $(M < N)$ that maximises MI $I(\mathbf{x}, \mathbf{y})$, i.e., $J(\mathbf{x}^0) = \max_{\mathbf{x} \subset \mathbf{x}'} I(\mathbf{x}, \mathbf{y})$.

Direct implementation of the above solution needs numerical evaluation of multi-dimensional entropies, which is difficult either using non-parametric density estimation or parametric methods. Most of these techniques require the estimation of statistics in some high-dimensional space. For example, the number of bins in histogram estimation increases exponentially with the dimensionality, as does the number of Gaussian parameters (i.e., the means and variances) in the Parzen window method. In practical applications, the amount of training data is always limited, and cannot meet the demands of the above methods with sufficient accuracy. Aiming at this problem, we devise the algorithm that will now be described.

Let $\mathbf{x} = (x_1, x_2, \ldots, x_M)$ be a selected data vector and $\mathbf{y}$ the corresponding class label. The mutual information between them is:

$$I(\mathbf{x}, \mathbf{y}) = I\left((x_1, x_2, \ldots, x_M), \mathbf{y}\right)$$

If $\mathbf{x}$ only has two components, i.e., $\mathbf{x} = (x_1, x_2)$, this becomes:

$$\begin{aligned} I(\mathbf{x}, \mathbf{y}) &= I((x_1, x_2), \mathbf{y}) \\ &= H(x_1, x_2) - H(x_1, x_2|\mathbf{y}) \end{aligned} \tag{5}$$

From (2), we have:

$$H(x_1, x_2) = H(x_1) + H(x_2) - I(x_1, x_2) \tag{6}$$

$$\text{and} \quad H(x_1, x_2|\mathbf{y}) = H(x_1|\mathbf{y}) + H(x_2|\mathbf{y}) - I(x_1, x_2|\mathbf{y}) \tag{7}$$

Substituting $H(x_1, x_2)$ and $H(x_1, x_2|\mathbf{y})$ of (5) into (6) and (7), we get:

$$
\begin{aligned}
I(\mathbf{x},\mathbf{y}) &= H(x_1,x_2) - H(x_1,x_2|\mathbf{y}) \\
&= H(x_1) + H(x_2) - I(x_1,x_2) - H(x_1|\mathbf{y}) \\
&\quad - H(x_2|\mathbf{y}) + I(x_1,x_2|\mathbf{y}) \\
&= \sum_{i=1,2} I(x_i,\mathbf{y}) - I(x_1,x_2) + I(x_1,x_2|\mathbf{y})
\end{aligned}
$$

Extending to more than than two components, we get:

$$
I(\mathbf{x},\mathbf{y}) = \sum_i I(x_i,\mathbf{y}) - \sum_i \sum_{j\neq i} I(x_i,x_j) + \sum_i \sum_{j\neq i} I(x_i,x_j|\mathbf{y}) \tag{8}
$$

Equation (8) reveals two important issues:

(1) The mutual information of a vector can be broken down into the MI of each component, which removes the obstacle of evaluating high-dimensional mutual information;
(2) To maximise the MI, we need to optimise not only the sum of the MI of all individual components, i.e., the term $\sum_i I(x_i,\mathbf{y})$, but also to minimise the redundancy (or inversely maximise the complementary level, in the terminology to be introduced shortly) within the components, i.e., the term $\sum_j I(x_i,x_j) - \sum_j I(x_i,x_j|\mathbf{y})$. These conclusions can be illustrated intuitively by Figure 1.

Figure 1 illustrates several typical scenarios in respect of the MI between the class label $\mathbf{y}$ and individual observation variables. Here, x_{i+1} represents a variable highly correlated with x_i, and x_j and x_k are much less correlated with x_i. The MI between vectors (x_i, x_{i+1}) and $\mathbf{y}$ is indicated by a shadowed area consisting of three different patterns of patches, i.e., $I\left((x_i,x_{i+1}),\mathbf{y}\right) = A + B + C$, where A, B and C are defined by different cases of overlap. In detail:

- $(A+B)$ is the mutual information between x_i and $\mathbf{y}$, i.e., $I(x_i,\mathbf{y})$;
- $(B+C)$ is the MI between x_{i+1} and $\mathbf{y}$, i.e., $I(x_{i+1},\mathbf{y})$;
- $(B+D)$ is the MI between x_i and x_{i+1}, i.e., $I(x_i,x_{i+1})$;
- D is the conditional mutual information between x_i and x_{i+1} given $\mathbf{y}$, i.e., $I(x_i,x_{i+1}|\mathbf{y})$.

To calculate $I\left((x_i,x_{i+1}),\mathbf{y}\right)$ using equation (8), the mutual information between each individual variable and the class label is first summed, i.e., $\sum_{m=\{i,i+1\}} I(x_m,\mathbf{y}) = (A+B) + (B+C)$. Any correlation between x_i and x_{i+1} means that the amount of information B was double counted in this step (cf. inclusion-exclusion principle). So the double counting is subsequently corrected by subtracting the amount of information $(B+D)$ in the second term of the right side in (8), i.e., $-I(x_i,x_{i+1}) = -(B+D)$. Finally, the third term on the right side of (8), i.e., $I(x_i,x_{i+1}|\mathbf{y})$, corrects by adding the amount of information D that

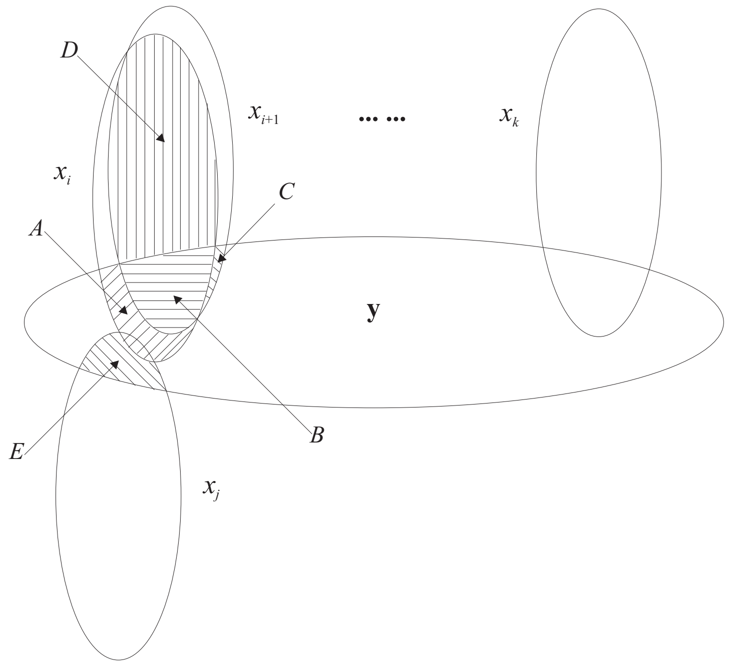

Fig. 1. Illustration of mutual information for different scenarios. Here, x_{i+1} is a variable that is highly correlated with x_i but x_j and x_k are less correlated with x_i; $\mathbf{y}$ is a class label variable representing the ground truth.

should not have been deleted in the last step, giving:

$$I\left((x_i, x_{i+1}), \mathbf{y}\right) = (A + B + B + C) - (B + D) + D = A + B + C$$

This illustration clearly shows that the features maximising the MI depend not only on their predictive information individually, e.g., $(A + B + B + C)$, but also need to take account of redundancy between them. In this example, variable x_j should have priority for selection over x_{i+1} in spite of the latter having larger individual MI with $\mathbf{y}$. This is because x_j provides more *complementary* information to variable x_i to predict $\mathbf{y}$ than does x_{i+1} (as $E > C$ in Figure 1). The above discussion justifies and provides a theoretical framework for the heuristic approximation used by previous researchers [18,19,17].

4 Maximisation of MI through progressive optimisation

Following the selection criterion and evaluation solution introduced in the last section, we now face the problem of determining a suitable search strategy. The number of ways of choosing M from the N features is $\binom{N}{M}$, which means that a huge number of MI evaluations might be needed for high-dimensional remotely-sensed data. For example, the AVIRIS hyperspectral sensor has 220 spectral bands;

if 80 bands are to be selected, the number of choices is 2.37×10^{61}, an unaffordable computing burden.

Aiming at this problem, we propose a greedy optimisation strategy to maximise (8), described as follows. According to (8), to maximise $I(\mathbf{x}, \mathbf{y})$ the first variable can be chosen as:

$$x_1^0 = \max_i I\left(x_i, \mathbf{y}\right)$$

where x_k^0 represents the result of maximisation at step k. Then, the second variable is chosen as:

$$x_2^0 = \max_{i \neq 1} \left[I\left(x_i, \mathbf{y}\right) - \sum_{i \neq 1} I\left(x_i, x_1^0\right) + \sum_{i \neq 1} I\left(x_i, x_1^0 | \mathbf{y}\right) \right]$$

The remaining variables are chosen in the same way until the pre-specified number, M, of variables is reached:

$$x_n^0 = \max_{i \neq j} \left[I\left(x_i, \mathbf{y}\right) - \sum_j \sum_{i \neq j} I\left(x_i, x_j^0\right) + \sum_j \sum_{i \neq j} I\left(x_i, x_j^0 | \mathbf{y}\right) \right] \tag{9}$$

where x_j^0, $j = 1, 2, \cdots, n-1$ are the variables already selected.

By using (8), the calculation of MI in each step of greedy optimisation is simplified. Except the conditional ones, i.e., $I(x_i, x_j|\mathbf{y})$, the other two kinds of MI in (8), (i.e., $I(x_i, \mathbf{y})$ and $I(x_i, x_{i+1})$) can be conveniently evaluated by using one-dimensional mutual information estimates.

The problem of evaluating the conditional MI can be solved by realising the following two facts:

(1) $I\left(x_i, x_j|\mathbf{y}\right) < I\left(x_i, x_j\right)$;
(2) $I(x_i, x_j|\mathbf{y}) \to 0$ as x_i and x_j become progressively less correlated.

While the first fact is apparent (the formal proof can be found in any textbook of information theory), the second can be illustrated from observing the MI between (x_i, x_j) and $\mathbf{y}$ in Figure 1, in which the conditional mutual information $I(x_i, x_j|\mathbf{y})$ is zero. In other words, the third term in (8) (i.e., the area D) will vanish as the correlation between x_i and x_j becomes less.

To indicate the likely correlation among the variables for this specific hyperspectral problem, we can set up a spectral window with a certain bandwidth W. If two variables are in spectral bands far away from each other, i.e, not within the chosen bandwidth W, the correlation between them is unlikely to be high. This is the scenario envisaged for x_i and x_j or x_k in Figure 1. Then (9) reduces to:

$$x_n^0 = \max_{i \neq j} \left[I\left(x_i, \mathbf{y}\right) - \sum_j \sum_{i \neq j} I\left(x_i, x_j^0\right) \right]$$

On the other hand, if the variables to be evaluated lie in bands closer together than W, they are assumed likely to be correlated, such as x_i and x_{i+1} in Figure 1. Based on the fact that $I\left(x_i, x_j|\mathbf{y}\right) < I\left(x_i, x_j\right)$, we can use the following approximation:

$$I\left(x_i, x_j|\mathbf{y}\right) \approx \alpha I\left(x_i, x_j\right) \tag{10}$$

where $0 < \alpha < 1$. Thus, (9) is approximated by:

$$\begin{aligned} x_n^0 &= \max_{i \neq j} \left[I\left(x_i, \mathbf{y}\right) - \sum_j \sum_{i \neq j} I\left(x_i, x_j^0\right) + \sum_j \sum_{i \neq j} \alpha I\left(x_i, x_j^0\right) \right] \\ &= \max_{i \neq j} \left[I\left(x_i, \mathbf{y}\right) - \sum_j \sum_{i \neq j} (1-\alpha) I\left(x_i, x_j^0\right) \right] \\ &= \max_{i \neq j} \left[I\left(x_i, \mathbf{y}\right) - \sum_j \sum_{i \neq j} \beta I\left(x_i, x_j^0\right) \right] \end{aligned} \tag{11}$$

where $\beta = 1 - \alpha$ and $0 < \beta < 1$.

Here, W and β can be seen as two application-related parameters, which reflect the nature of the correlation among different observation variables. In practical applications, they can be estimated using the training data in the same way as estimating other model parameters (such as means, variances, etc).

The above strategy selects features sequentially, i.e., it is greedy, and so avoids the problem of combinatorial explosion. At each step, the next feature will be selected so as to maximise $I(\mathbf{x}, \mathbf{y})$ incrementally. This is a similar idea to increasing gradually the class distance in the steepest ascent (SA) [10] or other hill-climbing algorithms, whereas the advantages of the proposed method are as follows.

(1) There is a determinate number of cost function evaluations. To maximise (9), we require $(N - M)$ evaluations of the cost function $I(\mathbf{x}, \mathbf{y})$ where N is the number of features in the original data and M is the number of selected

features. In other hill-climbing-based methods, the number of iterations depends on random initialisation and the particular termination criterion chosen.

(2) The computational requirement is low. Evaluation of $I(\mathbf{x}, \mathbf{y})$ does not need pairwise distance calculations; it is based on non-parametric MI estimation techniques [18,20]. Computation of matrix inverses and determinants is also avoided.

(3) Under a certain approximation (see (10) and (11)), the computational complexity of term $\left[I(x_i, \mathbf{y}) - \sum_j \sum_{i \neq j} I\left(x_i, x_j^0\right) + \sum_j \sum_{i \neq j} I\left(x_i, x_j^0 | \mathbf{y}\right)\right]$ is $\mathcal{O}(M(N-M))$. All the one-dimensional MI terms, i.e., $I(x_i, \mathbf{y})$ and $I\left(x_i, x_j\right)$ can be calculated beforehand and reused in each following step. This makes the algorithm much faster.

5 Simulations

To verify the performance of the proposed method, simulations are carried out on a 220-dimensional remotely-sensed hyperspectral dataset, where ‘band’ is used to indicate each individual feature measured at a specific wavelength.

5.1 AVIRIS 92AV3C dataset

The public AVIRIS 92AV3C hyperspectral dataset has been researched extensively. It is illustrative of the problem of hyperspectral image analysis to determine land use, and can be downloaded from `ftp://ftp.ecn.purdue.edu/biehl/MultiSpec/`. Although the AVIRIS sensor collects nominally 224 bands of data, 4 of these contain only zeros and so are discarded, leaving 220 bands in the 92AV3C dataset. At certain frequencies, the spectral images are known to be adversely affected by atmospheric water absorption. This affects some 20 bands. Each image is of size 145 × 145 pixels. The data were collected over a test site called Indian Pine in north-western Indiana, USA [24,25]. Only a single datacube is available.

The dataset is accompanied by a reference map, indicating partial ground truth, whereby pixels are labelled as belonging to one of 16 land-cover classes (mainly vegetation). Not all pixels are so labelled (e.g., highway, rail track, etc.), presumably because they correspond to uninteresting regions or were too difficult to label.

5.2 Evaluation of mutual information as a selection criterion

To implement the algorithm described in Section 4, the MI between each of the 220 spectral images (i.e., each band) and the corresponding ground-truth reference map accompanying the 92AV3C dataset was calculated, as shown in Figure 2(a). It is instructive to compare this MI curve to selected examples of AVIRIS images shown below the MI curve in Figure 2(a). It can be seen that the bands most similar to the reference map are those having higher values of MI. For example, the images of the spectral bands 31–34 bear more obvious resemblance to the reference map than those in bands 35–40, and their MI values are correspondingly higher. Figure 2(a) also reveals clearly the effect of atmospheric water absorption, giving the lowest MI values in bands 104–108 and 150–163 at precisely those frequencies where absorption occurs. Figure 2(b) further illustrates 200 samples (hence the 'thick' lines) of hyperspectral signals extracted from the AVIRIS 92AV3C dataset, for two land-cover classes: 'Corn-notill' and 'Grass/Pasture' respectively. We can also compare them with the MI, and confirm a certain degree of agreement for visual estimation of separability (such as the degree of overlap and scatter between two classes).

In Figure 3, we compare the MI with the commonly-used Bhattacharyya coefficients. Bhattacharyya coefficients were calculated between each band and the reference map and are shown in the solid line, together with the mutual information (dashed line). It can be seen that the overall shapes of the MI curve and the Bhattacharyya coefficients are very similar, indicating an agreement of MI with another commonly-used (but more computationally expensive) separability measure.

Figures 2 and 3 clearly show that the values of both MI and Bhattacharyya coefficients are not constant across various bands. This indicates that in hyperspectral data the discriminatory information is non-uniformly distributed across the spectrum. Some bands may contain more useful information for classification than others, and so have larger separability indices. From Figure 2(a), we see that many neighbouring bands show great similarity and are highly correlated accordingly. These two observations confirm the existence of 'weaker' and/or redundant features, and motivate the necessity to select effective spectral bands. They also verify the agreement of MI and the 'discriminatory information', through both visual inspection and comparison with the Bhattacharyya coefficients. Hence, mutual information can be used to encode the relevance of a spectral band and in band selection.

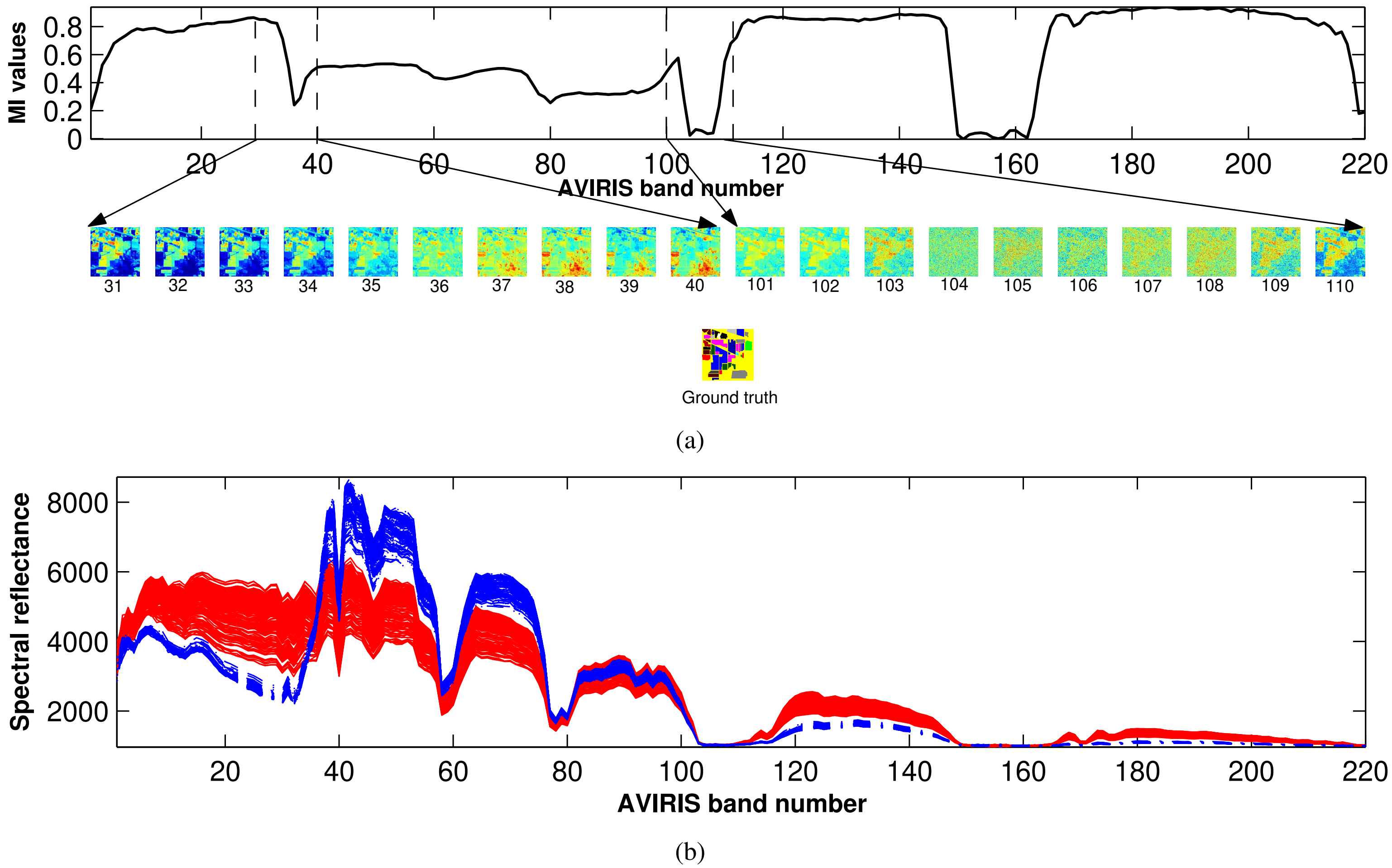

Fig. 2. (a) Mutual information of AVIRIS bands 1 to 220 with respect to the reference map; (b) 200 sampled spectral responses for two land-cover classes in AVIRIS 92AV3C, Corn-notill (light) and Grass/Pasture (dark).

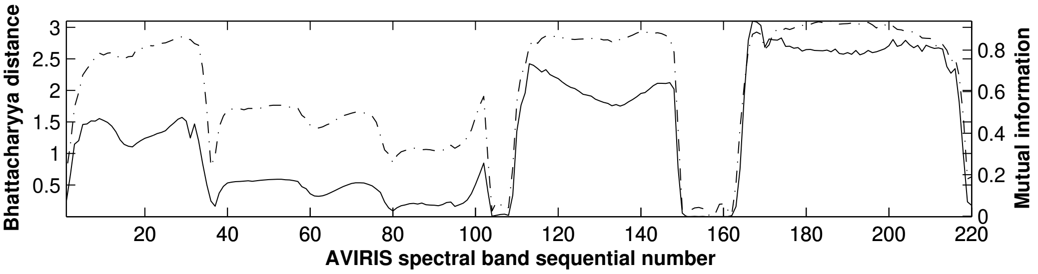

Fig. 3. Measurements of Bhattacharyya distance (solid line) and mutual information (dashed line).

5.3 Comparing results for hyperspectral band selection

Simulations of classification performance have been carried out to assess the proposed band-selection method on the AVIRIS 92AV3C dataset. Currently-popular support vector machines (SVMs) [26,27] were chosen as the classifiers in these simulations since previous studies of hyperspectral data classification have shown competitive performance with the best available algorithms [25,15]. Although SVMs are used here, the proposed method is not limited to supervised algorithms. Other classification algorithms are also relevant since the MI metric is calculated directly from the data, without feedback from the classifiers.

Half of the pixels from each class were randomly chosen for training, with the remaining 50% forming the test set on which performance was assessed (Table 1). The class labels from the reference map accompanying the dataset were utilised for supervised training. Since SVMs are inherently binary (two-class) classifiers, $\binom{16}{2}$ one-against-one classifiers were used with subsequent majority voting to give a multi-class result. The kernel function used was an inhomogeneous polynomial of order 5. The penalty parameter C was tested for values between 10^{-3} and 10^5 by a validation procedure using training data, and 10^3 was chosen as an appropriate value.

The main objective of band selection is to choose the most 'informative' or 'relevant' spectral bands, while achieving the highest possible classification accuracy for the reduced number of bands. The simulations were designed to assess the change of classification accuracy as spectral bands are progressively added. The solid line in Figure 4 shows the results for bands selected according to the algorithm described in Section 4. These results are compared to a benchmark algorithm [14,28] (dashed line) that chooses bands based on metrics (here, MI) that only measure separability between each individual band and the reference map. From our discussion in Sections 3 and 4, this strategy only optimises the first term in (8), i.e., max $\sum_i I(x_i, \mathbf{y})$. If the correlation between features is significant, the result will be inaccurate as too much preference is given to neighbouring bands that have higher MI values, but these are also highly correlated with each other

Table 1
Number of training and testing pixels in each class.

Class	Pixels in training set	Pixels in testing set
1. Alfalfa	27	27
2. Corn-notill	717	717
3. Corn-min	417	417
4. Corn	117	117
5. Grass/Pasture	249	248
6. Grass/Trees	374	373
7. Grass/pasture-mowed	13	13
8. Hay-windrowed	245	244
9. Oats	10	10
10. Soybeans-notill	484	484
11. Soybeans-min	1234	1234
12. Soybean-clean	307	307
13. Wheat	106	106
14. Woods	647	647
15. Bldg-Grass-Tree-Drives	190	190
16. Stone-steel towers	48	47

(see Figure 2(a) for the highest MI region with the neighbouring bands 180–200). Therefore, the selected bands may fail to complement each other, because of the considerable redundancy among them. This is confirmed by the simulations shown in Figure 4: It is seen that a significant improvement has been made by adopting the proposed algorithm when up to 100 bands are selected. As more and more bands are selected, the difference between the two methods tends to narrow because, at this stage, most of the essential bands are already included.

To validate the new method more fully, performance has been compared against some other representative band-selection algorithms on the 92AV3C dataset, namely steepest ascent (SA) searching [10,15] and the very well-known Pearson correlation coefficient. The comparison is carried out for 20 to 80 bands retained, since the SA algorithm is very computationally-expensive and for the 92AV3C dataset the classification accuracy shows little change beyond 80 selected bands.

Table 2 shows the results. To account for local maxima, the SA algorithm was run

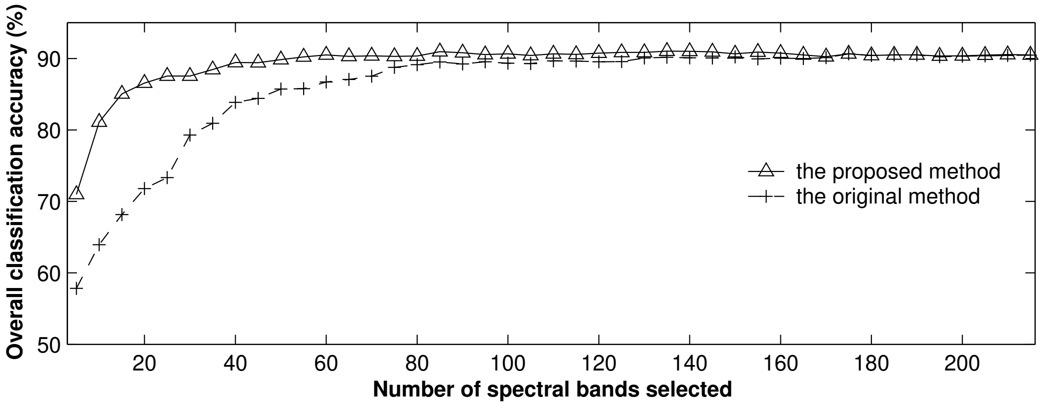

Fig. 4. Comparison of proposed algorithm with the benchmark algorithm on the basis of overall percentage accuracy.

Table 2
Comparison of proposed method with competitor algorithms on the basis of overall percentage accuracy (best results in each row shown in bold).

Bands Retained	Method		
	MI	SA	Correlation
20	86.57	**88.07**	84.64
30	87.55	**89.60**	86.43
40	89.42	**90.12**	87.51
50	89.83	**90.27**	88.32
60	**90.48**	90.35	88.75
70	**90.37**	90.33	89.29
80	**90.35**	90.33	90.12

five times with random initialisations, and the best result chosen for tabulation. The proposed method easily outperformed the method based on correlation, justifying the use of MI as a selection criterion. It is also competitive with the search-based SA algorithm, with slightly lower accuracy at band selection numbers 20 to 50. However, the SA algorithm involves computationally-expensive iterative search (and repetition to account for local maxima). Each iteration of the search requires $M(N-M)$ evaluations of Jeffreys-Matusita distance. The Jeffreys-Matusita distance is a sum of Bhattacharyya distances that has to be evaluated pairwise. For the 16-class problem, $\binom{16}{2}$ Bhattacharyya distance evaluations are needed, and at each time, determinants and inverses of the covariance matrices have to be calculated [10]. It was also noticed in implementing the SA algorithm that when the number of selected bands increased above about 40, the covariance matrices used to calculate the Bhattacharyya distance tended to become singular.

To cope with this ill-posed problem, some regularisation method has to be applied but that will introduce extra errors. Avoidance of this problem is another reason to favour the new method. Finally, we avoid the difficulty, inherent in methods based on pairwise comparisons, of having to decide how to deal with conflicting evidence for the retention/deletion of a given feature (i.e., the feature scores well in combination with one feature but poorly in combination with another).

The results presented here are comparable to those in our earlier work [17], also based on selection of ‘complementary’ bands via MI evaluation. While the research in [17] mainly focuses on estimating the reference variable for the MI calculation, its selection criterion is based on a heuristic observation of hyperspectral imagery to avoid selecting redundant neighboring bands. In a little more detail, to reduce correlation between bands two extra parameters, a rejection bandwidth and a complementary threshold, were introduced (see Algorithm 1, p. 525 of [17]) and chosen by empirical cross-validation. On the contrary, the method proposed in this research is based on two different foundations, namely:

(1) decomposition of multi-dimensional MI; and
(2) progressive maximisation of MI.

Both are obtained completely from theoretical analysis of mutual information. Therefore, comparing with the heuristic approach in [17], the generic algorithm presented here has a much sounder theoretical footing, and can be applied to any high-dimensional data besides hyperspectral images. Moreover, the theoretical consistency embedded in the new method helps it to achieve very significant computational efficiencies, such as fewer parameters to be validated.

6 Conclusion

A framework of feature selection using mutual information for image classification has been proposed, with natural applicability to multi-class problems and non-Gaussian data. Given the relationship between MI and Bayes error, maximising MI is analogous to the idea of maximising separability that many other methods employ. Theoretical analysis revealed that multi-dimensional mutual information can be efficiently evaluated by breaking down into a series of one-dimensional mutual information terms. A ‘greedy’ optimisation scheme is proposed to reduce the search space for the maximisation of mutual information. Experiments on the AVIRIS high-dimensional remotely-sensed dataset show that the proposed method outperforms or is competitive with state-of-the-art methods, but having the advantages of easy implementation and low computational cost.

Acknowledgement

This work was supported by the Data and Information Fusion (DIF) Defence Technology Centre funded by the UK Ministry of Defence and managed by General Dynamics Limited and QinetiQ.

References

[1] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, 1997.

[2] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1–2, pp. 245–271, 1997.

[3] AVIRIS, "Airborne visible/infrared imaging spectrometer." [Online]. Available: http://aviris.jpl.nasa.gov/

[4] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 55–63, 1968.

[5] J. Price, "Spectral band selection for visible-near infrared remote sensing: Spectral-spatial resolution tradeoffs," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 35, no. 5, pp. 1277–1285, 1997.

[6] M. Velez-Reyes and L. Jimenez, "Subset selection analysis for the reduction of hyperspectral imagery," in *Proceedings of IEEE International Geoscience and Remote Sensing Symposium*, vol. 3, Seattle, WA, 1998, pp. 1577–1581.

[7] G. Petrie, P. Heasler, and T. Warner, "Optimal band selection strategies for hyperspectral data sets," in *Proceedings of IEEE International Geoscience and Remote Sensing Symposium*, vol. 3, Seattle, WA, 1998, pp. 1582–1584.

[8] C.-I. Chang, D. Qian, T.-L. Sun, and M. Althouse, "A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 6, pp. 2631–2641, 1999.

[9] N. Keshava, "Best bands selection for detection in hyperspectral processing," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'01*, vol. 5, Salt Lake City, UT, 2001, pp. 3149–3152.

[10] S. Serpico and L. Bruzzone, "A new search algorithm for feature selection in hyperspectral remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 7, pp. 1360–1367, 2001.

[11] P. Groves and P. Bajcsy, "Methodology for hyperspectral band and classification model selection," in *IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data*, Greenbelt, MD, 2003, pp. 120–128.

[12] S. Kaewpijit, J. L. Moigne, and T. El-Ghazawi, "Automatic reduction of hyperspectral imagery using wavelet spectral analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 4, pp. 863–871, 2003.

[13] D. Qian, "Band selection and its impact on target detection and classification in hyperspectral image analysis," in *IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data*, Greenbelt, MD, 2003, pp. 374–377.

[14] P. Bajcsy and P. Groves, "Methodology for hyperspectral band selection," *Photogrammetric Engineering and Remote Sensing Journal*, vol. 70, no. 7, pp. 793–802, 2004.

[15] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.

[16] C. Conesea and F. Masellia, "Selection of optimum bands from TM scenes through mutual information analysis," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 48, no. 3, pp. 2–11, 1993.

[17] B. Guo, S. R. Gunn, R. I. Damper, and J. B. Nelson, "Band selection for hyperspectral image classification using mutual information," *IEEE Geoscience and Remote Sensing Letters*, vol. 3, no. 4, pp. 522–526, 2006.

[18] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, Jul 1994.

[19] N. Kwak and C.-H. Choi, "Input feature selection for classification problems," *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 143–159, 2002.

[20] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Physical Review A*, vol. 33, no. 2, pp. 1134–1140, 1986.

[21] P. Swain and R. King, "Two effective feature selection criteria for multispectral remote sensing," in *Proceedings of the 1st International Joint Conference on Pattern Recognition*, 1973, pp. 536–540.

[22] M. Hellman and J. Raviv, "Probability of error, equivocation and the chernoff bound," *IEEE Transactions on Information Theory*, vol. 16, no. 4, pp. 368–372, Jul 1970.

[23] M. Feder and N. Merhav, "Relation between entropy and error probability," *IEEE Transactions on Information Theory*, vol. 40, no. 1, pp. 259–266, Jan 1994.

[24] D. Landgrebe, "On information extraction principles for hyperspectral data: A white paper," Purdue University, West Lafayette, IN, Technical Report, School of Electrical and Computer Engineering, 1997. [Online]. Available: http://dynamo.ecn.purdue.edu/~landgreb/whitepaper.pdf

[25] J. Gualtieri and R. Cromp, "Support vector machines for hyperspectral remote sensing classification," in *Proceedings of the 27th AIPR Workshop on Advances in Computer Assisted Recognition*, Washington DC, 1998, pp. 121–132.

[26] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, Pittsburgh, Pennsylvania, United States, 1992, pp. 144–152.

[27] C. Cortes and V. N. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 1–25, 1995.

[28] B. Guo, S. R. Gunn, R. I. Damper, and J. B. Nelson, "Adaptive band selection for hyperspectral image fusion using mutual information," in *Proceedings of 8th International Conference on Information Fusion, Volume 1*, Philadelphia, PA, 2003, pp. 630–637.

Author Biographies

Baofeng Guo was born in 1973 in Xi'an, China, and received the BE degree in electronic engineering and ME degree in signal processing from Xidian University, Xi'an in 1995 and 1998, and the PhD degree in signal processing from the Chinese Academy of Sciences, Beijing, China, in 2001, respectively. From 2002 to 2004, he was a research assistant in the Department of Computer Science, University of Bristol, UK. Since 2004, he has been research fellow in the School of Electronics and Computer Science, University of Southampton, Southampton, UK. His current research interests are pattern recognition, machine learning and image processing.

Bob Damper was born in Tunbridge Wells, England, in 1948. He obtained his MSc in biophysics in 1973 and PhD in electrical engineering in 1979, both from the University of London. He also holds the Diploma of Imperial College, London, in electrical engineering. He was appointed Lecturer in electrical engineering at the University of Abertay Dundee in 1976, Lecturer in electronics at the University of Southampton in 1980, Senior Lecturer in electronics and computer science in 1989, Reader in 1998 and Professor in 2003. He has wide research interests including speech science and technology, neural computing, cognitive modeling, pattern recognition and intelligent systems engineering. Prof. Damper has published approximately 300 research articles and authored the undergraduate text *Introduction to Discrete-Time Signals and Systems*. He is a Chartered Engineer and a Fellow of the UK Institution of Engineering and Technology, a Chartered Physicist and a Fellow of the UK Institute of Physics, a Senior Member of the IEEE and an Honorary (Foreign) Member of the Yugoslav Engineering Academy.

Steve Gunn was born in Bristol, England, in 1970. He obtained his BEng (1st class honours) degree in electronic engineering in 1992 and PhD in computer vision in 1996 from the University of Southampton. He is currently a Professor in the Information: Signals, Images and Systems Research Group, School of Electronics and Computer Science at Southampton. Prof. Gunn has published over 80 research papers in the areas of computer vision and machine learning. His current research interests are in the area of sparse representations, feature selection and subspace methods for identification of salient parts of the data space for prediction. He has recently published a book entitled *Feature Extraction: Foundations and Applications*. He serves on various programme committees, and is the operational coordinator for the EU PASCAL Network of Excellence.

James Nelson was born in York, England, in 1975. He studied at Anglia Polytechnic University where he obtained a BSc in mathematics and a PhD for his work on the application of Riesz bases to signal theory. He has held research posts at Cranfield University (three years) and the University of Southampton (two years), and is currently a research associate at the University of Cambridge. His research interests include signal and image processing, wavelets, and support vector machines.