A comparative evaluation of filter-based feature selection methods for hyper-spectral band selection

Bo Wu\textsuperscript{a}, Chongcheng Chen\textsuperscript{a}, Tahar Mohand Kechadi\textsuperscript{b} & Liya Sun\textsuperscript{c}

\textsuperscript{a} Key Lab of Spatial Data Mining and Information Sharing of Ministry of Education, Fuzhou University, Fuzhou, P.R. China
\textsuperscript{b} School of Computer Science and Informatics, University College of Dublin, Dublin, Ireland
\textsuperscript{c} Department of Geography, Ludwig Maximilian University of Munich, 80333, Munich, Germany


To link to this article: http://dx.doi.org/10.1080/01431161.2013.827815

PLEASE SCROLL DOWN FOR ARTICLE
A comparative evaluation of filter-based feature selection methods for hyper-spectral band selection

Bo Wu, Chongcheng Chen, Tahar Mohand Kechadi, and Liya Sun

Key Lab of Spatial Data Mining and Information Sharing of Ministry of Education, Fuzhou University, Fuzhou, P.R. China; School of Computer Science and Informatics, University College of Dublin, Dublin, Ireland; Department of Geography, Ludwig Maximilian University of Munich, 80333 Munich, Germany

(Received 31 March 2013; accepted 7 April 2013)

Band selection (dimensionality reduction) plays an essential role in hyper-spectral image processing and applications. This article presents a unified comparison framework for systematic performance comparison of filter-based feature selection models and conducts a comparative evaluation of four methods: maximal minimal associated index (MMAIQ), mutual information-based max-dependency criterion (mRMR), relief feature selection (Relief-F), and correlation-based feature selection (CFS) for hyper-spectral band selection. The evaluation is based on the performance of effectiveness, robustness, and classification accuracy, which involves five measuring indices: class separability, feature entropy, feature stability, feature redundancy, and classification accuracy. Three images acquired by different sensors were used to investigate the performance of the metrics. Experimental results show the best results for MMAIQ for all data sets in terms of used measurements, except for feature stability where mRMR and Relief-F exhibit their superiority.

1. Introduction

An advanced hyper-spectral imaging spectrometer can sample the electromagnetic spectrum at tens, hundreds, or even thousands of wavelength ranges in the visible and near infrared spectra, and thus provide a very detailed view of the spectral signature of the scene represented by particular pixels. Compared with traditional multispectral imagery, images acquired by these hyper-spectral sensors provide greater details on the spectral variation of targets, which provides a better discrimination among similar ground-cover classes (Richards 1999). However, interpretation of data from such sensors gives rise to new challenges for various applications. The availability of huge amounts of data creates several challenges for the full use of hyper-spectral images, such as data processing, memory storage, and data transfer, as well as the very large challenge of dimensionality (Liu and Motoda 1998). Direct use of all hyper-spectral bands may add to unnecessary information redundancy and significantly decrease the classification accuracy (Price, Guo, and Stiles 2002; Zhong, Zhang, and Wang 2008). Furthermore, the estimation of the parameters of certain classifiers in high-dimensional feature space is often difficult and unstable with limited samples (Jain and Zongker 1997; Foody and Mathur 2004). Therefore, it is necessary...
to develop effective and reliable dimensionality reduction techniques. It is well known that
dimensionality reduction can not only facilitate probing into the essential features, but also
contribute to accomplishing most tasks of remote-sensing application problems, such as
classification and visualization at low computational cost. As a consequence, dimension-
ality reduction techniques have long attracted attention from the remote-sensing community,
and hence a varying range of dimensionality reduction techniques have been attempted and
applied in past decades (Pudil, Novovicova, and Kittler 1994; Serpico and Bruzzone 2001;
Kavzoglu and Mather 2002; Lee and Landgrebe 2003; Li, Zhang, and Ogihara 2004; Kuo
and Landgrebe 2004; Pal 2006).

Normally, two broad categories of dimensionality reduction techniques are encoun-
tered in remote sensing: feature extraction and feature selection (Mather 2004). The feature
extraction methods provide a small set of new features that contain the greatest part of infor-
mation from original remotely-sensed data sets based on the transformation of the original
data sets. More popular, and of particular interest in this article, are feature selection meth-
ods, which select a subset of original features that captures the relevant properties of the
entire data set. In general, feature selection algorithms broadly fall into three categories:
filter, wrapper, and embedded models. The filter techniques have the advantages of easily
scaling to very high-dimensional data sets, being computationally more efficient and, most
of all, the selected features (bands) can be easily interpreted by experts. This article focuses
on filter-based feature selection techniques.

The filter-based technique relies on the general characteristics of data and evaluates
features without involving any learning algorithm. It usually consists of the following three
aspects: (1) defining a criterion function to judge the quality of each feature in terms
of its discriminating ability; (2) developing a search procedure to generate and compare
every possible feature selection solution by calculating its criterion function value for a
given feature subset; and (3) designing a strategy to eliminate relevance and redundancy
features in the process of searching the optimal feature subset if necessary. Following
these observations, a variety of feature selection methods have been developed in past
decades, ranging from chi2 (Liu and Setiono 1997), the fast correlation-based filter (FCBF)
algorithm (Yu and Liu 2004), and Gini Index (Shang et al. 2007), to relief feature selec-
tion (Relief-F) (Robnib and Kononenko 2003), mutual information-based max-dependency
criterion (mRMR) (Peng, Long, and Ding 2005), and correlation-based feature selec-
tion (CFS) (Hall and Smith 1997). Furthermore, efforts have been made to empirically
compare and evaluate feature selection algorithms with different problem settings (Li,
Zhang, and Ogihara 2004). Relief algorithms can detect conditional dependencies between
features and provide a unified view of feature estimation in regression and classifica-
tion, and they have actually been used successfully in a variety of settings. The mRMR
method was developed to identify high-dimension complex data and has been validated
as an effective technique for feature selection in a high-spatial resolution remotely-sensed
image (Wu et al. 2009). In CFS, a correlation measure is applied to evaluate the good-
ness of feature subsets based on the hypothesis that a good feature subset is the one
that contains features highly correlated to the class, yet uncorrelated to each other. This
method has also been reported to be an effective tool to select the optimal feature sub-
set from a complex remotely-sensed data set (Jose, Manuel, and James 2010). Despite
intensive efforts on feature selection, it remains a hot topic for remote-sensing research.
As various remotely-sensed data set acquisitions and new applications have emerged in
recent years, stable feature selection, optimal redundancy removal, and the exploitation
of auxiliary data are among the most fundamental and challenging problems in feature
selection.
Recently, we proposed a maximal minimal associated index (MMAIQ) feature selection method (Wu et al. 2012) to select features incrementally, satisfying the criteria of statistically maximal associations between target labels and features and minimal associations among selected features with respect to Cramer’s $V$-test value. Our previous experiments demonstrated that it has positive effects on the performance of classification. To reinforce the MMAIQ method, we performed a comparative evaluation of several feature-ranking algorithms and provided some insight into specified aspects: class separability, clustering property, the capability of redundancy removal, and algorithm stability. Moreover, it is still difficult to obtain comprehensive views on the performance of feature selection algorithms when researchers design new feature selection algorithms or pick existing algorithms. In this context, it is interesting to provide a unified comparison framework to evaluate the feature selection algorithms for hyper-spectral band selection. Therefore, our objective in this article is twofold. The first is to present a unified comparison framework for systematic performance analysis of potential feature selection algorithms. The second is performance comparison against other feature selection algorithms to further confirm the effectiveness of the MMAIQ method.

2. A unified comparison framework and experimental settings

To compare possible feature selection algorithms systematically, a unified framework of comparison and evaluation is first presented. Figure 1 illustrates the feature selection evaluation process, which is divided into two phases. In phase 1, each feature selection method generates candidate feature subsets based on respective criteria and search strategies. In phase 2, the finally selected subset is evaluated via measuring indices and is subjected to result validation by given learning algorithms.

It can be observed from Figure 1 that the framework includes four main components that are required to be specified: data used, algorithms for feature subset generation, subset evaluation metrics, and accuracy validation. In this study, three test images acquired by different sensors with low-, medium-, and high-dimensional bands were selected to express the data volume and characteristics. Considering a model’s effectiveness, representativeness,
and popularity, we selected four feature selection algorithms to carry out the comparison: MMAIQ, Relief-F, mRMR, and CFS. All of them are filter-based feature selection methods with feature redundancy removal. To evaluate the performances of the aforementioned algorithms, four measuring metrics were used: class separability, entropy, redundancy ratio, and feature stability. Class separability measures the effectiveness of feature subsets, while entropy evaluates the clustering performance of the subsets. The feature redundant rate is a quantitative measure that captures the amount of redundancy remaining in the reduced subset. The stability index, on the other hand, assesses the robustness of an algorithm to form the same feature subset with a fluctuating training data set. To validate the models’ fitting, $k$-nearest neighbor ($k$-NN) and naive bayes (NB) classifiers were adopted to account for the effectiveness of feature subsets for subsequent classification. We selected the $k$-NN and NB classifiers for classification because of their simplicity, elegance, and robustness (Webb, Boughton, and Wang 2005). Moreover, $k$-NN and NB require the user to input few parameters (Wong and Chiu 1987), and thus differences in accuracy results are predominately due to the feature subsets generated from different algorithms, having the advantage of avoiding model fitting error in comparison. Although the support vector machine (SVM) classifier usually gives good classification accuracies, it is a computationally demanding algorithm to tune parameters for high-dimensional data and/or when the training samples are larger. Therefore, SVM was not selected, since we focused on evaluating feature subsets generated by feature selection algorithms rather than the performance of a certain classifier. On the other hand, although many accuracy measures are available (Pontius and Millones 2011), we adopted overall accuracy to assess the classification accuracy, because it is the simplest and one of the most popular accuracy measures (Congalton 1991) and is recommended for use as a primary measure (Liu, Frazier, and Kumar 2007).

2.1. Data collection

Three images, acquired by Flight-line C1 (FLC1), push broom hyper-spectral imagery (PHI), and airborne visible/infrared imaging spectrometer (AVIRIS) were collected.

(1) FLC1: The data set was collected with an airborne scanner in June 1966 by FLC1, located in the southern part of Tippecanoe County, Indiana, USA. It follows a county road from the Grandville Bridge over the Wabash River just south of South River Road (West Lafayette) to near State Highway 25 (https://engineering.purdue.edu/~biehl/MultiSpec). The image contains a significant number of vegetative species or ground-cover classes; it includes many regions (e.g. fields) containing a large number of contiguous pixels from a given class, and thus facilitates quantitative results evaluation and provides available ‘ground truth’ data. A sub-scene (220 × 220 pixels) of the FLC1 image with 12 bands was tested at the spectral range of 400–1000 nm. Figure 2(a) shows the experimental FLC1 image (band 2) and the ground truth map available for the scene. From the ground truth map, it can be observed that the image was expected to be classified into six representative classes: soybeans, oats, wheat, corn, red clover, and rye. The spectral curves of these classes are shown in Figure 2(a). The labelled samples available for these classes amount to about 41,000 pixels. To compare the feature selection algorithms, 2073 (about 5%) points are stratified randomly and sampled for the experiment. The training samples are adequate according to a previous study on the relationship between training sample size and data dimensionality (Van Neil, McVicar, and Datt 2005).
This is achieved when the number of training samples per class amounts to two to four times’ bands in sample farmland area. The spatial distribution of selected samples for each class can also be observed from Figure 2(a).

(2) PHI: The data set was acquired from the Xiaqiao test site, a mixed agricultural area in Changzhou City, Jiangsu province, China, using the airborne PHI data set. The imagery was acquired at an approximate height of 2000 m with 3 m spatial resolution. A sub-scene (346 × 350 pixels) of the PHI image with 80 bands was tested at a spectral range of 417–854 nm. Figure 2(b) shows the PHI image cube and corresponding testing data set. The ground truth spectral data were collected by field spectrometer SE590 (Spectron Engineering Inc., Denver, CO, USA). The observed image was expected to be classified into seven representative classes: corn, vegetables, soil/bare land, grass, water, road, and puddle/polluted water; 5848 (about 5%) samples were randomly selected from this image for the following evaluation. The spatial distributions of selected samples for each class are also reported in Figure 2(b). The reflectance curves of the seven land-cover classes mentioned earlier are shown in Figure 2(b).

(3) AVIRIS: The third data set is the classical 220-band AVIRIS image (Figure 2(c)) taken over Indiana’s Pine test site in June 1992. It is freely available from http://dynamo.ecn.purdue.edu/~biehl/ The image is 145 × 145 pixels, containing 16 crop-type classes and a total of 10,366 label pixels. This image is a classical benchmark to validate model accuracy and constitutes a very challenging classification problem because of the strong mixture of class signatures and unbalanced number of labelled pixels per class. Since all of the pixels were intensively labelled by fieldwork, many
researchers have used these data to validate their algorithms in various aspects. In the pre-processing phase, some water absorption and noise bands were initially excluded for brevity. Thus, the final set contains a total of 185 remaining spectral bands for further analysis. Since the size of the samples in some classes is too small for training and testing, only ten classes were considered: Corn-min, Corn-notill, Grass/Pasture, Grass/Tree, hay-windrowed, Soybeans-min, Building/Grass, soybeans-clean, Soybeans-notill, and woods. The labelled samples available for these classes amount to 9275 pixels. Among them, 20% samples (about 1940 pixels) are stratified randomly and sampled for validation, and the spatial distribution of samples for each class is shown in Figure 2(c).

The list of classes and number of labelled samples for each class are summarized in Table 1. For all three testing data sets, each feature variable was pre-processed to have zero mean value and unit variance before the implementation of feature selection.

### 2.2. Feature selection methods

(1) MMAIQ: MMAIQ selects features that satisfy both the criteria of statistically maximal association between target labels and features and minimal association among selected features with respect to Cramer’s $V$-test value (Wu et al. 2012), which is defined by

$$
\max \phi(x_i) = \frac{\sum_{x_i \in S} V(x_i, y)}{\sum_{x_i, x_j \in S} V(x_i, x_j)},
$$

where $\sum_{x_i \in S} V(x_i, y)$ is the Cramer’s $V$-test coefficient between individual features $x_i$ and class $y$ and $\sum_{x_i, x_j \in S} V(x_i, x_j)$ is the Cramer’s $V$-test coefficient between features $x_i$ and $x_j$. Given the set of $(u - 1)$ features $\{S_{u-1}\}$, the $u$th search procedure adopts the methodology of the incremental algorithm, which selects the feature from set $\{X - S_{u-1}\}$ by optimizing the following criterion

$$
\max_{x_i \in X - S_{u-1}} [(u - 1) \times V(x_j, y) / \sum_{x_i, x_j \in S} V(x_i, x_j)].
$$

Given the input data of $P$ samples and
m features \( X = \{x_1, \ldots, x_m\} \) and the target classification variable \( y \), the feature selection problem is to find a subset of \( n \) features among \( m(n \leq m) \) features that optimally characterizes \( y \).

(2) Relief-F: The key idea of Relief-F is to estimate the quality of features, according to how well their values distinguish between instances that are near to each other. For this purpose, Relief-F searches for its two nearest neighbours, one from the same class \( \text{NH}(x_i, y) \) and the other from a different class \( \text{NM}(x_i, y) \), to adjust the feature weighting vector and give more weight to features that discriminate the instance from their neighbours belonging to different classes (Robnib and Kononenko 2003). Given \( P \) samples that are randomly selected from training data, the evaluation criterion of Relief-F is defined as follows:

\[
R_f(x_i) = \frac{1}{P} \sum_{i=1}^{P} \left\{ -\frac{1}{k} \sum_{x \in \text{NH}(y)} \text{diff}(x_{i,j}, x_{j,i}) + \sum_{j \in \text{NM}(x,y)} \frac{P(y)}{k} \sum_{x \in \text{NM}(x,y)} \text{diff}(x_{i,j}, x_{j,i}) \right\},
\]

where \( R_f \) is the score of \( x_i \), \( y_i \) is the class label of the sample \( x_i \) and \( P(y) \) is the probability of a sample being from class \( y \). \( x_{i,j} \) denotes the values of \( x_i \) on feature \( x_j \) and \( \text{diff} ( \cdot ) \) is the function used to calculate the difference between \( x_{i,j} \) and \( x_{j,i} \). Generally, the number of nearest neighbours defined in \( \text{NM}(x_i, y) \) and \( \text{NH}(x_i, y) \) are pre-specified by a constant \( k \).

(3) mRMR: mRMR is a supervised feature-ranking algorithm of the filter model. It selects features that are mutually far away from each other, while still having high correlation with the classification variables (Peng, Long, and Ding 2005). mRMR can be regarded as an approximation to the maximum dependency measured by conditional entropy between the joint distribution of the selected features and the classification target. The optimization criterion of mRMR is as follows:

\[
\max R(x_i) = \sum_{x_i \in S} I(x_i, y) / |S| \sum_{x_i, x_j \in S} I(x_i, x_j),
\]

where \( I(x_i, y) \) and \( I(x_i, x_j) \) denote respective mutual information between individual features \( x_i \) with label \( y \) and mutual information between features \( x_i \) and \( x_j \), respectively. \( x, y \) are multidimensional random variables; their mutual information is defined in terms of their probability density functions \( P(x), P(y), \) and \( P(x, y) \). \(|S|\) is the number of selected features.

(4) CFS: CFS is a filter-based algorithm that selects a feature subset on the basis of a correlation-based heuristic evaluation function (Hall and Smith 1997). The heuristics by which CFS measures the quality of a set of features takes into account the usefulness of individual features for predicting the class and is defined as follows:

\[
M(S) = \frac{\sum_{x_i \in S} C(x_i, y)}{\sqrt{|S| + (|S| - 1) \sum_{x_i, x_j \in S} C(x_i, x_j)}},
\]

where \( M(S) \) is the heuristic ‘merit’ of a feature subset \( S \), \( C(x_i, y) \) is the feature correlation between features \( x_i \) with class \( y \), and \( C(x_i, x_j) \) is the feature inter-correlation.
between individual features $x_i$ and $x_j$. The correlation is measured using the conditional entropy. The numerator provides an indication of how predictive of class a group of features is, whereas the denominator indicates the redundancy among the features. The evaluation criterion of CFS shows that the algorithm tends to select feature subsets that are highly predictive of the class and not predictive of each other.

2.3. Subset measuring indices

(1) Class separability: We assume that $S_w$ is the within-class scatter matrix and $S_b$ is the between-class scatter matrix. These are defined as follows:

$$S_w = \sum_{i=1}^{c} \pi_i E \left\{ (X - \mu_i)(X - \mu_i)^T \mid \omega_i \right\} = \sum_{i=1}^{c} \pi_i \Sigma_i$$

$$S_b = \sum_{i=1}^{c} (\mu_i - M_0)(\mu_i - M_0)^T$$

$$M_0 = E\{X\} = \sum_{i=1}^{c} \pi_i \mu_i,$$

where $\pi_i$ is the priori probability that a pattern belongs to class $\omega_i$, $X$ is the feature vector, and $T$ denotes the transpose operator. $\mu_i$ denotes the sample mean vector of class $\omega_i$ and $M_0$ is the mean vector of the whole data samples. $\Sigma_i$ is the sample covariance matrix of class $\omega_i$ and $E\{\bullet\}$ denotes the expectation operator. The class separability $J$ of a data set is defined by

$$J = \text{trace}(S_w^{-1}S_b).$$

(6)

Usually, a higher value of the separability criteria ensures that the classes are well separated by their scatter means.

(2) Feature entropy: Let the distance between two variables $p$ and $q$ be

$$D_{p,q} = \left[ \sum_{i=1}^{M} \left( \frac{x_{p,i} - x_{q,i}}{\max_i - \min_i} \right)^2 \right]^{1/2},$$

where $x_{p,i}$ denotes feature value for $p$ along the $i$th band, $\max_i$ and $\min_i$ are the maximum and minimum values computed over all samples in the $i$th band, and $M$ is the number of bands. Similarity between $p$ and $q$ is given by $\text{sim}(p, q) = \exp\{ -\alpha D_{p,q} \}$, where $\alpha$ is a positive constant. A possible value of $\alpha$ is $-\ln 0.5 / \overline{D}$. $\overline{D}$ is the average distance between data points computed over the whole data set. The entropy index is defined by

$$E = \sum_{p=1}^{N} \sum_{q=1}^{N} \{ \text{sim}(p, q) \times \ln[\text{sim}(p, q)] + [1 - \text{sim}(p, q)] \times \ln[(1 - \text{sim}(p, q))] \}.$$
When the data are uniformly distributed in the feature space, the entropy is maximal. When the data have well-formed clusters, their uncertainty is low and so is the entropy.

(3) Feature redundancy: Assume $s$ is the set of selected features and $X_s$ is the data containing only features in $s$. The following measurement is used for evaluating the redundancy rate of $s$:

$$R(s) = \frac{1}{|S|(|S| - 1)} \sum_{f_i, f_j \in S, i > j} c_{i,j},$$

where $c_{i,j}$ is the correlation between two features, $f_i, f_j$. The metric assesses the averaged correlation among all feature pairs, and a large value indicates that the selected features are strongly correlated and thus redundancy is expected to exist in $s$.

(4) Feature stability: To estimate the stability of a feature selection algorithm for the given data, different subsamples drawn from a training set were used to assess how different data affect a method’s preference of a feature’s importance and consequently the final set of selected features. Therefore, a measure of similarity for each pair of sets was required in order to measure stability. The similarity was measured using the Tanimoto distance between two sets:

$$T_d(s_1, s_2) = \frac{|s_1 \cap s_2|}{|s_1| + |s_2| - |s_1 \cap s_2|},$$

where $s_1$ and $s_2$ are the pair of feature subsets generated from different training folds, and $|\cdot|$ denotes the number of elements in a subset. The Tanimoto distance metric measures the amount of overlap between two sets of arbitrary cardinality. $T_d$ takes values in $[0, 1]$, where 0 means there is no overlap between two sets and 1 means the two sets are identical. To obtain the different subsamples, the $k$-fold stratified cross-validation resample technique was adopted. The feature selection algorithm outputs a feature subset for each of the training folds. The similarity of each pair of feature subsets ($k(k-1)/2$ pairs) was calculated using the Tanimoto distance, and the final stability score is simply the average of all pairs.

### 2.4. Accuracy validation

The $k$-NN classifier is a non-parametric classification technique which operates on the premise that classification of unknown instances can be done by relating the unknown to the known according to a distance/similarity function. Generally, Euclidean distance is used as the distance metric for continuous variables, and overlap metric (or Hamming distance) is used for discrete variables. Let $(x_{j1}, y_{j1}), \cdots, (x_{jk}, y_{jk})$ be the $k$ training instances whose attributes are closest to a new instance $x$. The label $x$ associated with $y$ that has the highest voting value

$$v(y) = \sum_{i=1, y_i=y}^{k} \frac{1}{d(x_{ji}, x)}.$$
The parameter \( k \) can be a user-defined constant or determined in the training process. In this article, we use the simplest version (i.e. \( k = 1 \)). In other words, a feature vector \( x \) is assigned to the class of its nearest neighbour.

On the other hand, NB is based on the Bayes rule and assumes that features are independent of each other given the class labels. Given a sample \( x_i = [x_{i1}, x_{i2}, \cdots, x_{im}]^T \), where \( T \) denotes transpose operator, for \( m \) features, the posterior probability that \( x_i \) belongs to class \( c_j \) is

\[
P(c_j | x_i) = k \prod_{i=1}^{m} P(c_i | x_i),
\]

where \( k \) is a constant coefficient and \( P(c_j | x_i) \) is the conditional probability table learned from examples in the training process.

### 3. Implementation issues

To compare the aforementioned feature selection algorithms, all were wrapped by the MATLAB7.0 platform. Among them, the MMAIQ algorithm was coded by authors on Matlab, while the other algorithms were implemented with the support of the Waikato Environment for Knowledge Analysis (WEKA) toolbox developed at the University of Waikato in New Zealand. This is a powerful, open-source Java-based machine learning software package and is available online at http://www.cs.waikato.ac.nz/ml/weka (Mark et al. 2009). In order to rule out the possibility that undocumented processing steps (where we may obtain the particular result from WEKA toolbox, but do no understanding how they work) influence the results, third party routines were also used when available in source code. The configuration of the machine used for experiments is Intel® Pentium® Dual E2180 @ 2.0 GHz and 2 GB memory. In order to guarantee a meaningful comparison between the different feature selection algorithms, the experimental protocol was designed to meet the following criteria: (1) in the stage of feature subset generation, the inputs were kept identical and only the feature selection methods varied. In this way, all features were ranked according to their selected orders; (2) to calculate each measuring index, the ranked features were incrementally used to generate a series of feature subsets; (3) for each method, the optimal parameters were carefully selected so that their best-case performance is evaluated; (4) to reduce the random effect of data sets, all experimental results are referred to the average value obtained from 10 trials according to 10-fold cross-validation on testing data sets. There are two important problems that still need to be addressed, the first being how to determine the optimal parameters for respective algorithms to achieve best possible performance; the second is how to characterize the ranked features generated from various algorithms with the specified indices, so as to evaluate their performance objectively.

#### 3.1. Determination of parameters

Since each of the algorithms has parameters that may affect the model’s performance at a certain level, we carefully selected the important parameters of each algorithm used in the experiments. The key parameter of MMAIQ is the number of intervals used to partition each feature, which can be determined by \( n \)-fold cross-validation (\( n \)-fold CV). Generally, the parameter is insensitive if there are sufficient samples in every interval (Wu et al. 2012). In our experiments, the optimal number of intervals for FLC1, PHI, and AVIRIS data sets are 7, 9, and 11 respectively. In the Relief-F algorithm, the \texttt{diff()} function used was specified by the Euclidean distance. Two parameters are required to be determined: the drawing size
which is related to the problem complexity, and the neighbouring parameter \( k \), which is related to the distance described earlier. In the following experiments, the parameter \( P \) was set to the number of training samples \( M \), so that all available samples were used. The parameter \( k \) is insensitive and is usually tuned by \( n \)-fold CV (Robnik and Kononenko 2003). In our experiments, the default value 10 gave remarkable results for all data sets. To calculate the mutual information between features and class labels in the mRMR algorithm, the Parzen-window approximation approach was used to estimate density distribution, and the pool is set to three times the feature dimension. To assess feature stability, the 10-fold CV approach is used to obtain the subsamples. In 10-fold CV approach, the overlap of training instances among the different training folds is roughly 80\%. The stability measurement defined in Equation (6) quantifies how different training sets affect feature selection.

### 3.2. Characteristic feature space

To determine which of the feature subsets \( S_1^n \) and \( S_2^n \) is superior, it is insufficient to compare the results for a specific size of the feature sets or the best result of respective algorithms obtained with different size of the features. A better way is to observe which set is recursively more characteristic for a reasonably large range \( \Omega \). Given two feature sets \( S_1^n \) and \( S_2^n \) that contain \( n \) features generated from two different algorithms, we maintain that the feature set of \( S_1^n \) is superior if the specified measurement on \( S_1^n \) is better than on \( S_2^n \). This definition can be extended recursively to the subsets of \( S_1^n \) and \( S_2^n \). Suppose, we use the first algorithm to generate a series of feature subsets in \( S_1^n \): \( S_1^1 \subset S_1^2 \subset \cdots \subset S_1^n \), similarly, the second algorithm produces another series of subsets in \( S_2^n \): \( S_2^1 \subset S_2^2 \subset \cdots \subset S_2^n \). \( S_1^n \) is recursively better than \( S_2^n \) in the range \( \Omega = [k_1, k_2] \), if for every subset in this range, the measurement on \( S_1^n \) is consistently better than on \( S_2^n \). In this article, we use the extreme case, \( \Omega = [1, m] \).

### 4. Experimental results and comparisons

Figure 3 reports the experimental results of feature selection with FLC1 data. The superior performance of the MMAIQ algorithm in terms of the above measurements can be seen from Figure 3. Specifically, in most cases, MMAIQ has the largest values in class separability, indicating that the selected bands with MMAIQ can obtain the maximum distances of between-class and in-class, thus improving the subsequent classification accuracy. In all cases, the entropy values of MMAIQ are the least, which suggest that the features of their selected bands have good clustering property. Furthermore, the feature subsets generated from MMAIQ have almost the lowest redundancy rate, showing that the feature subset selected by MMAIQ has the least redundancy and correlation. As a result, in most cases, the overall accuracies associated with NB and 1-NN classifiers are relatively higher than for other algorithms. These results are consistent and suggest an optimal performance for the MMAIQ algorithm, especially for a small number of features.

Figure 4 shows the comparison results for the PHI data set with selected features using different methods and classifiers. It will be seen from Figure 4 that MMAIQ significantly outperforms other algorithms in all measurements except for feature stability, where mRMR and Relief-F achieve relatively higher feature stability. The tendencies of plotted lines resemble those of FLC1 data. When very few bands (10) are selected, MMAIQ can generate optimal overall classification accuracies as shown in Figures 5(e) and (f), while other algorithms require many more features to attain the resembled accuracies.

The results of the above-mentioned feature selection methods with specified measurements against the number of bands with AVIRIS data are provided in Figure 5. It can be clearly seen that, in most cases, MMAIQ outperforms other algorithms, especially, when
Figure 3. Comparison of the four feature selection algorithms using measurements with the FLC1 data set. (a) Class separability, (b) feature entropy, (c) feature redundancy rate of selected features, (d) stability of selected features, (e) overall accuracy by NB classifier, and (f) overall accuracy by 1-NN classifier.

Figure 4. Comparison of the four feature selection algorithms using measurements with the PHI data set. (a) Class separability, (b) feature entropy, (c) feature redundancy rate of selected features, (d) stability of selected features, (e) overall classification accuracy by NB classifier, and (f) overall classification accuracy by 1-NN classifier.

the number of features is small. However, the overall classification accuracies for AVIRIS data were far lower than for the two previous data sets, especially for the NB classifier where the classification accuracy is higher for a smaller number of selected features, but it is still lower than for the previous two data sets and is roughly smooth when features increase with fluctuations.

It can be broadly concluded that MMAIQ is superior to other algorithms in terms of our recursive definition. However, it is difficult for one algorithm to be always superior
to another, since the values associated with the specified measurement corresponding to the two methods often overlap under different feature sets. To quantitatively evaluate the performances of the measurement entirely, we improved the metric in Chen et al. (2011) to combine used measures of the selected features and the corresponding information, to assess the whole performance.

Let $H(i)$ denote the information entropy as

$$H(i) = -\frac{i}{|S|} \ln \left( \frac{i}{|S|} \right),$$

(13)

where $i$ is the number of features selected and $|S|$ is the number of evaluating features. The metric denoted as ‘wacc’ is as follows:

$$wacc = \sum_{i=1}^{|S|} a_i \bar{H}(i),$$

(14)

where $a_i$ is the number of given measurements, which can be either class separability, entropy, stability, redundancy, or overall accuracy under $i$ features and $\bar{H}(i) = H(i) \sum_{i=1}^{|S|} H(i)$. It is clear that $\bar{H}(i)$ acts on the normalized weights. Using the ‘wacc’ metric as a holistic measurement, the entire performance of the algorithms can be quantitatively evaluated. An illustration of the ‘weights’ with 20 features is shown in Figure 6. It can be seen that low values yield relatively high values, as shown in Figure 6.

Since the metric assigns more weight for small numbers in a feature subset, ‘wacc’ can check the performance over all the dimensions and simultaneously focus on the small number of features that are preferred in the classification application.

Table 2 reports the average value in terms of ‘wacc’ measurement for all data sets, where bold values highlight the best results. It can be seen from Table 2 that for FLC1 data, MMAIQ has the largest value for class separability measurement (4.62), followed by CFS.
Figure 6. Illustration of the ‘weights’ of $I(i)$ with 20 features.

Table 2. Evaluation of five selected indices in terms of ‘wacc’ measurement.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Methods</th>
<th>‘wacc’ evaluation criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Separability</td>
</tr>
<tr>
<td>FLC1</td>
<td>MMAIQ</td>
<td>4.62</td>
</tr>
<tr>
<td></td>
<td>Relief-F</td>
<td>4.06</td>
</tr>
<tr>
<td></td>
<td>MRMR</td>
<td>3.60</td>
</tr>
<tr>
<td></td>
<td>CFS</td>
<td>4.36</td>
</tr>
<tr>
<td>PHI</td>
<td>MMAIQ</td>
<td>75.75</td>
</tr>
<tr>
<td></td>
<td>Relief-F</td>
<td>69.66</td>
</tr>
<tr>
<td></td>
<td>Mrmr</td>
<td>69.48</td>
</tr>
<tr>
<td></td>
<td>CFS</td>
<td>70.16</td>
</tr>
<tr>
<td>AVIRIS</td>
<td>MMAIQ</td>
<td>141.90</td>
</tr>
<tr>
<td></td>
<td>Relief-F</td>
<td>112.27</td>
</tr>
<tr>
<td></td>
<td>Mrmr</td>
<td>115.27</td>
</tr>
<tr>
<td></td>
<td>CFS</td>
<td>121.17</td>
</tr>
</tbody>
</table>

Note: Bold values denote the best results.

(4.36). The lowest entropy value is achieved by MMAIQ, which means that the feature subsets formed by MMAIQ have the best clustering property on average. Notably, the average feature redundancy rate of MMAIQ is only 0.367, which is much lower compared with other methods. As a consequence, MMAIQ unsurprisingly achieves the best classification results, where average overall accuracies achieve 89.9% and 90.9% for NB and 1-NN classifiers, respectively. The second best result is obtained by the CFS algorithm, where the overall accuracies of NB and 1-NN classifiers are 88.1% and 89.7%, respectively. Moreover, the highest stability value achieved by MMAIQ is 0.986 for these data, followed by mRMR and Relief-F, while that of the CFS algorithm is somewhat lower.

For PHI data, it can be seen that MMAIQ achieves the best results among all measurements except for stability, of which the largest overall accuracies of NB and 1-NN classifiers are 88.6% and 90.9%, respectively; the largest class separability value is 75.75 and lowest entropy value is 1.39, indicating that MMAIQ can provide the best classification results when applied to PHI data. The average feature redundancy in the subset of MMAIQ is
The quantitative results of AVIRIS are also reported in Table 2. It can be seen that MMAIQ still achieves the best classification accuracy in terms of ‘wacc’ measurement, of which the largest overall accuracy of the NB and 1-NN classifiers are 57.8% and 81.4%, respectively. MMAIQ also achieves the best results in separating indices, with the highest class separability value of 141.9 and lowest entropy value of 2.64. The average feature redundancy contained in the subset of MMAIQ is 0.764, the lowest value among all methods, while the redundancy rates in the subsets generated from other methods are over 0.8. Since the NB classifier is highly dependent on the assumption of independence among features, it inevitably affects performance significantly if the feature subset contains much redundancy. This indicates that the NB classifier is not applicable to the data set. On the other hand, the mRMR method obtains the best feature stability result (0.94), indicating that mRMR can generate the most stable features.

5. Discussion and conclusions

This article has presented a systematic comparison of the performance of the MMAIQ, mRMR, Relief-F, and CFS methods with five measurement indices: feature entropy, class separability, feature redundancy, feature stability, and classification accuracy. The experimental results indicate that MMAIQ can be an effective option for hyper-spectral band selection, and we summarize the findings as follows:

(1) Of the four feature selection algorithms, MMAIQ exhibits high performance on all three data sets. Specifically, it can be concluded from Table 2 that MMAIQ provides the best performance for all data sets and classifiers. The Relief-F and CFS algorithms also demonstrated their capability on hyper-spectral band selection, but Relief-F and mRMR cannot be used for high-dimensional mixture data although they work well with low- and moderate-dimensional data. In other words, it is not scalable. As shown in Figure 5, they fail completely when applied to AVIRIS data sets, where a strong mixture of class signatures occurs.

(2) In regard to class separability and entropy measurements, both are used to capture the characteristics of feature subsets. Generally, a higher value for class separability indicates that the classes are well separated by their scatter means, and lower entropy indicates that the features have good clustering property. Measuring the discrimination effectiveness and data structure of feature subsets for a given data dimension allows us to combine features appropriately and derive the ‘best’ features. As a result, these are generally in line with classification accuracy. Compared with entropy measurement, class separability appears to be more effective because of the use of labelled data.

(3) In regard to feature redundancy, MMAIQ exhibits high performance on removal of feature redundancy. In all cases, MMAIQ selects the feature subsets with lowest feature redundancy rate, which shows that it produces a concise set of features. Therefore, a relative small number of features selected by MMAIQ can generate remarkable overall accuracy. However, it must be noted that the evaluation used in this article can only measure the averaged correlation among all possible feature
pairs. Methods to assess complex redundancy for a given feature subset have not been explored yet.

(4) In regard to feature stability, the mRMR and Relief-F algorithms seem to have the best feature stability for all data sets. Compared with mRMR and Relief-F, the stability of feature selection with the MMAIQ algorithm is less satisfactory, especially for high-dimensional data sets. The main reason is that our searching scheme for MMAIQ adopts an incremental forward selection method, which may cause the selected features to be unstable because these cannot be omitted once selected. In our future work, we intend to improve stability by introducing a mechanism to reselect features by the use of a backward refining process from previously selected features.

(5) Of the classifiers, the simplest 1-NN classifier provided very good accuracy for all data sets. This indicates that after feature selection, the data expressed can be well clustered according to a given distance measure. The NB classifier has an important assumption that features should be independent of other features, and so it is not surprising that its performance was not very good when the AVIRIS data invalidated this assumption.

Acknowledgements
The authors acknowledge the research support from the National Key Technology Research and Development Programme of China under Grant No. 2013BAC08B01, the EC FP7-2009-People-IRSES under Grant 247608, and the Key Science & Technology programme of Fujian province, China (Grant Nos. 2011Y0036 and 2010I0008).

References


