A novel approach to hyperspectral band selection based on spectral shape similarity analysis and fast branch and bound search

Shijin Li a,∗, Jianbin Qiu a, Xinxin Yang a, Huan Liu b, Dingsheng Wan a, Yuelong Zhu a

Article history:
Received 2 April 2013
Received in revised form 8 June 2013
Accepted 13 July 2013
Available online 6 August 2013

Keywords:
Hyperspectral remote sensing
Feature selection
Spectral clustering
Key point of time series
Conditional mutual information
Branch and bound search

ABSTRACT

With the development of hyperspectral remote sensing technology, the spectral resolution of the hyperspectral image data becomes denser, which results in large number of bands, high correlation between neighboring bands, and high data redundancy. It is necessary to reduce these bands before further analysis, such as land cover classification and target detection. Aiming at the classification task, this paper proposes an effective band selection method from the novel perspective of spectral shape similarity analysis with key points extraction and thus retains physical information of hyperspectral remote sensing images. The proposed approach takes all the bands of hyperspectral remote sensing images as time series. Firstly, spectral clustering is utilized to cluster all the training samples, which produces the prototypical spectral curves of each cluster. Then a set of initial candidate bands are obtained based on the extraction of key points from the processed hyperspectral curves, which preserve discriminative information and narrow down the candidate band subset for the following search procedure. Finally, filtering contiguous bands according to conditional mutual information and branch and bound search are performed sequentially to gain the optimal band combination. To verify the effectiveness of the integrated band selection method put forward in this paper, classification employing the Support Vector Machine (SVM) classifier is performed on the selected spectral bands. The experimental results on two publicly available benchmark data sets demonstrate that the presented approach can select those bands with discriminative information, usually about 10 out of 200 original bands. Compared with previous studies, the newly proposed method is competitive with far fewer bands selected and a lower computational complexity, while the classification accuracy remains comparable.

1. Introduction

Hyperspectral imaging is concerned with the measurement, analysis, and interpretation of spectra acquired from a given scene (or specific object) at a short, medium or long distance by an airborne or satellite sensor (Plaza et al., 2009). Due to the dense sampling of the wavelength region from 0.4 to 2.5 μm with more than two hundred spectral channels, hyperspectral images have a better discrimination among similar ground land cover classes than traditional multispectral scanners (Landgrebe, 2003). However, these tens of or hundreds of spectral bands will result in high redundancy and great amount of computation in hyperspectral image classification. It is of great value to reduce the dimension (bands) and retain or even enhance the classification accuracy.

In the literature, there are two categories of methods for the purpose of hyperspectral data dimensionality reduction: feature extraction and feature selection. The main purpose of feature extraction is trying to mitigate the Hughes phenomenon and to find the features which can describe the separability between classes (Kuo and Chang, 2007). Linear Discriminant Analysis (LDA) is one of the most popular methods, which usually suffers from the curse of dimensionality when dealing with hyperspectral pixel classification with hundreds of bands. In addition, because every band of the hyperspectral data has its own corresponding image, the way of feature extraction that the high-dimensional feature space is mapped to a low-dimensional space by linear or nonlinear transformation, could not keep the original physical interpretation of the image. As the spectral distance between the two adjacent bands in the hyperspectral data is only 10 nm and the correlation between them is so high, there is considerable redundancy, which should be largely reduced by the feature selection or band selection methods so that the classification efficiency can be greatly improved, with a comparable or even higher accuracy.

Feature selection techniques generally involve both a search algorithm and a criterion function (Serpico and Moser, 2007).
Recently, band selection based on search algorithms has received a lot of attention. Serpico and Bruzzone (2001) proposed a suboptimal search strategy suitable for hyper-dimensional feature selection, which was based on the search for constrained local extremes in a discrete binary space. Bazi and Melgani (2006) put forward a genetic algorithm-based system to optimize the Support Vector Machine (SVM) classifier accuracy for hyperspectral imagery, which determined the best SVM parameters and detected the best discriminative features (bands) in a completely automatic way. Sotoca and Pla (2006) presented a band selection method using correlation among bands based on Mutual Information (MI) and deterministic annealing optimization. Guo et al. (2008) found that selection based on the simple criterion of only retaining features with high associated MI values may be problematic when the features were highly correlated. And they proposed a fast feature-selection scheme based on a ‘greedy’ optimization strategy.

Observing high correlation between the adjacent bands, band grouping and selection had been proposed by several researchers. Sun and Gao (2000) proposed that the original bands were separated into groups based on fuzzy set theory and then reduced the dimension by selecting a representative band or their linear fusion, and the final band subset could be found out based on rough set theory. Huang and He (2005) presented a feature weighting method for band selection, which was based on the pairwise separability criterion and matrix coefficients analysis through Principal Component Analysis (PCA). Another method was proposed by Wang et al. (2005), that all the bands were separated into subspace according to the correlation between adjacent bands, and selected the bands which had the best sensitivity of SVM from every subspace. According to Peng et al. (2005), in feature selection, it has been recognized that the combination of individually good features do not necessarily lead to good classification performance.

Since there are hundred of bands in the hyperspectral imagery, the search space directly based on the original band space will be too huge and thus incur high computational overhead. Recent years have witnessed several novel approaches to band selection or feature extraction for hyperspectral image classification. To reduce the search space, Li et al. (2011) have proposed a hybrid band selection strategy based on genetic algorithm and support vector machine (GA-SVM), which forms a wrapper to search for the best combination of bands with higher classification accuracy and the computational cost of the genetic algorithm is alleviated by band grouping constraints. Martínez-Usó et al. (2007) exploited the band correlation through a hierarchical clustering-based algorithm, to minimize the intra-cluster variance and maximize the inter-cluster variance. They chose the bands in an unsupervised way. Serpico and Moser (2007) put forward a procedure to extract a set of spectral channels of variable bandwidths and spectral positions from the hyperspectral image, in which each spectral channel was obtained by averaging a group of contiguous channels of the original data. Mojaradi et al. (2009) proposed a hybrid feature extraction and feature selection method based only on class spectra, called prototype space feature extraction, in which the spectra channels were clustered by fuzzy c-means. Cariou et al. (2011) addressed the problem of unsupervised band reduction in hyperspectral remote sensing imagery, by using an information theoretic criterion to automatically separate the sensor’s spectral range into disjoint sub-bands without ground truth knowledge. Su et al. (2012) proposed semi-supervised dimensionality reduction approach to clustering bands based on orthogonal projection divergence.

Along the line of feature (band) selection for hyperspectral image classification, there are still several issues to resolve. On the one hand, since the original band number reaches 200, traditional search-based feature selection methods, such as branch and bound search, sequential forward search, and sequential forward floating search, are incapable of achieving satisfactory results directly on such a huge search space. On the other hand, although there have emerged some new efforts to tackle the aforementioned huge search space issue, the resultant chosen bands are still outnumbered and the time cost is very high. There is still much space for improvements to further reduce the number of selected bands and alleviate the time complexity of the algorithm.

In this paper, we put forward an innovative band selection method from the perspective of spectral shape similarity analysis, in which the reflectance characteristics of a given pixel response are regarded as time series. Hyperspectral image pixel classification is transformed into shape classification by recognizing the shapes of the spectral curves. The proposed method is made up of two stages. In the first stage, key points of prototypical spectral curves of every cluster are extracted and then those bands that contain discriminative information are picked out, which narrow down the candidate band subset of hyperspectral image data and generate the candidate bands for further processing. The sample data is grouped into clusters by spectral clustering algorithm with the help of visual assessment of cluster tendency, then the initial candidate bands set are obtained based on the extraction of landmarks from the time series. In the second stage, filtering contiguous bands according to conditional mutual information and branch and bound search on a relatively smaller search space are performed sequentially to get the optimal band combination. The experimental results on two publicly available benchmark data sets demonstrate that the newly proposed approach can select those bands with discriminative information. Compared with previous studies, the presented method is competitive with far fewer bands selected and also with a lower computational complexity. In addition, the newly proposed approach guarantees the classification accuracy; sometimes it is even higher than those previous methods. The flow chart of the method proposed in this paper is shown in Fig. 1.

The rest of the paper is arranged as follows. Section 2 describes the process of generating prototypical hyperspectral curves via spectral clustering, and the cluster number is estimated by visual assessment of cluster tendency. Section 3 introduces the concept of key points of time series, and a set of candidate bands are extracted according to the prototype curves obtained in Section 2. In Section 4, two post processing procedures are depicted to further reduce the number of finally chosen bands. Section 5 presents the experimental results on two benchmark data sets, as well as the comparisons with other state-of-the-art works. And finally, concluding remarks are given in the last section.

2. Prototype curves generation based on spectral clustering

The new idea of band selection for hyperspectral image presented in this study is to find all bands with discriminative information and those bands are retained as features to distinguish different classes, resulting in dimension reduction. However, by observing the original data of training samples, it is easy to find that the trends of spectral curves are complex, fluctuating frequently and having quite a lot of noise. In order to alleviate this situation, we will firstly generate the prototype curves of the training samples by spectral clustering, which will provide fewer “time series” for the processing in the next section.

2.1. Spectral clustering algorithm

Spectral clustering is based on the spectral graph theory, which tackles the classification problem as optimally partitioning of an
undirected graph. The method establishes an affinity similarity matrix with training data, and then spectral decomposition is performed on the corresponding Laplacian matrix of it. There are many variants of spectral clustering algorithms. We utilize the Ng–Jordan–Weiss (NJW) algorithm as follows (Ng et al., 2002).

1. Construct the similarity matrix $W$ of training samples $\{s_1, s_2, ..., s_n\}$, $W_{ij} = \exp(-d^2(s_i, s_j)/2\sigma^2)$, where $d(s_i, s_j)$ is the Euclidean distance of two samples $s_i$ and $s_j$, and $\sigma$ is the variance of the distance. The number of clusters $k$ is given beforehand. Then the diagonal of $W$ is set to 0, i.e., $W_{ii} = 0$;
2. Define $D$ to be the diagonal matrix whose $(i,i)$-element is the sum of $W$'s $i$-th row, and construct the matrix $L = D^{-1/2}WD^{-1/2}$;
3. Find $v_1, v_2, ..., v_k$ the $k$ largest eigenvectors of $L$, and form the matrix $V = [v_1, v_2, ..., v_k]$, $V \in \mathbb{R}^{n \times k}$, by stacking the eigenvectors in columns;
4. Normalize $V$, making each of $V$'s rows have unit length and thus obtaining $U$;
5. Cluster the rows of $U$ into $k$ clusters via k-means or any other clustering algorithms;
6. Assign the original training sample $s_i$ to cluster $j$ if and only if row $i$ of the matrix $U$ was assigned to cluster $j$.

Compared with traditional clustering algorithm, spectral clustering can cluster sample space of any shape and guarantee convergence to the global optimal solution. It has several advantages over conventional clustering algorithms. First, kernel techniques are used to project the data into a high-dimensional feature space in which the clusters can be more spatially distinct and compact. Second, only a distance matrix is needed in spectral clustering, no matter what number of dimensions of the original sample is. Finally, theoretical analysis shows that spectral decomposition can reveal the block structure of the affinity matrix, which is related to the number of intrinsic clusters.

However, like K-means, spectral clustering also needs to set the cluster number beforehand. Hence, this study proposes that the cluster number is roughly estimated with the approach of visual assessment of cluster tendency (VAT).

2.2. Estimating the number of clusters

In the process of clustering of hyperspectral image data, without prior knowledge, it is very difficult to set the cluster number properly. In Yu et al. (2010), large range search with Davies–Bouldin index is proposed. If the range of clusters is set too small, it is likely that the optimal cluster number is missed; otherwise it will increase the computational complexity in a large search range. In the spectral clustering algorithm, how to choose a suitable range of cluster number becomes an issue. This study leverages visual assessment of cluster tendency (VAT) approach (Huband et al., 2005) to estimate the rough number of clusters.

The VAT approach represents the pair-wise dissimilarity information about the set of samples $O = \{o_1, ..., o_N\}$ as a square digital image with $N \times N$ pixels, after the samples are suitably reordered so that the image is better able to highlight potential cluster structure.

Usually, we have a matrix of dissimilarity of training samples, $R = [R_{ij}]$, where $R_{ij}$ is the pairwise dissimilarity measure $d(o_i, o_j)$ between samples $o_i$ and $o_j$ for $1 \leq i, j \leq N$. According to Huband et al. (2005), The VAT procedure is described as follows:

1. Construct the dissimilarity matrix $R$, whose size is $N \times N$, and the array $P$ is used to store the reordering found by VAT;
2. Let $K=\{1, 2, ..., N\}$, select $i,j \in \text{arg}\min_{(i,j) \in K} \min(R_{ij})$;
   Set $P(1) = i$; $I(1) = \{i\}$; and $J(1) = K - \{i\}$.
3. For $t=2, ..., N$:
   Select $i,j \in \text{arg}\min_{(i,j) \in K} \min(R_{ij})$;
   Set $P(t) = j$; Replace $I(\{i\})$ and $J(\{j\})$ with $I(\{i\} \cup J(\{j\}))$ and $J(\{j\} \cup I(\{i\}))$;
   End
4. Reorder the dissimilarity matrix to generate a new matrix $\tilde{R}, \tilde{R} = [\tilde{R}_{ij}] = [R_{IJ}]$;
5. Display $\tilde{R}$ as an intensity image, scaled so that max $\tilde{R}_{ij}$ corresponds to white and 0 to black.

The algorithm could reorder dissimilarity of the training data, making the similar samples positioning along the diagonal. In the image of the reordered matrix, the cluster structure is made apparent, if there are any. So the number of clusters could be visually estimated from the VAT image. The resulting image of VAT on the hyperspectral data of Washington DC Mall is shown in Fig. 2.
in which it can be observed that there are about 8 to 12 blocks along
the diagonal. Hence, in spectral clustering, we can set the range of
the cluster number \( k \) to 8–12 and choose the best one according to
certain cluster validity index. In this study, we choose the cluster
number \( k \) with minimum Davies-Bouldin index (Davies and Bouldin,
1979), in a much smaller search range.

3. Candidate band selection based on time series analysis

After spectral clustering, several prototype hyperspectral
curves are obtained. However, our goal is to select the best band
combination to achieve dimensionality reduction with satisfactory
classification accuracy. In this study, we will firstly generate the
initial candidate band set through time series analysis. The final
optimal band combination will be obtained by branch and bound
search, preceded by conditional mutual information filtering, all of
which will be depicted in Section 4.

3.1. Time series analysis and key points

Time series is a sequence of data recorded at different time. The
hyperspectral remote sensing image data is treated as time series
in this section, in which each band is regarded as a time stamp.
So the cluster center of each group of hyperspectral data in the
previous section can be processed in the same manner as time
series. For instance, there are 4428 training samples in the data set
of Washington DC Mall, which has 191 valid bands. These training
samples can be analyzed as 4428 time series and each time series
owns 191 “time” points. Fig. 3 gives one normalized sample in the
Washington DC Mall data set, with the band as the horizontal
time axis.

Perng et al. (2000) proposed to match time series with key
point sequence, being able to greatly improve the similarity search
speed. Zhou and Wu (2007) come up with the concept of
important point of time series, which is applied to extract the
trend of time series. Key points and important points both carry
the shape information of time series, which could be used to
perform time series classification. The classification of hyperspec-
tral data can also be seen as “shape recognition” of different class
of spectral curves. This study proposes that time series analysis
technique could be “borrowed” to band selection, and the bands
corresponding to key points can be retained. Then the dimension
can be reduced while the classification accuracy and efficiency
may be improved.

Figs. 4 and 5 give the time series graphs based on the fifth and
the seventh class of samples in the Washington DC Mall data set.
There are 136 samples and 74 samples in those two classes respectively. It can be observed from Figs. 4 and 5 that, the shapes and trends of the samples in the same class are very similar, and for some bands, even the numerical range is alike, which also reflects high consistency between hyperspectral data in the same class. These trends and shape can be used to distinguish different classes of spectral curves.

The new idea of band selection of hyperspectral image presented in this study is to find all bands with discriminative information and these bands are regarded as features to distinguish different classes. All bands with discriminative information of every class are gathered together and the other bands are discarded, resulting in dimension reduction in the first stage. However, by observing the time series in Figs. 4 and 5, it is also easy to find that the trends of spectral curves are complex, fluctuating frequently and having quite a lot of noise, which will lead to many difficulties to extract those important points in the following procedures. Thus, in the previous section, all the samples are clustered into several groups and alleviate the complexities to some extent.

3.2. Extracting key points (important bands)

Key points can reflect the trend characteristics of time series, and it is helpful to analyze time series. The key points are defined as follows:

For time series \( X = \{X(t_1), \ldots, X(t_n)\} \), the \( q \)th key point of \( X \) is described as \( X^q = X(t_p) \), \( P \subseteq\{1, 2, \ldots, n\} \) and \( P_q \) represents the \( q \)th key point's position in time series \( X(t_p) \) is the data point which meets the conditions as defined in Eqs. (1) and (2):

\[
\begin{align*}
&\left[\left|X(t_p - 1) - X(t_p)\right| \leq \left|X(t_p - 1) - X(t_p)\right| + \left|X(t_p - 1) - X(t_p)\right|\right]\cup\left[\left|X(t_p + 1) - X(t_p)\right|\right] \\
&\left[\left|X(t_p - 1) - X(t_p)\right| \geq \left|X(t_p + 1) - X(t_p)\right|\right]\cup\left[\left|X(t_p + 1) - X(t_p)\right|\right]
\end{align*}
\]

(1)

(2)

For the time series data in which noise has been removed, it only needs to extract key points in the time series according to the definition in (1) or (2). If the data satisfy the conditions, the point is considered as important and thus retained, otherwise that point is discarded. This process continues until all points are processed. Therefore, all retained bands corresponding to the key points are those with discriminative “shape” information, so the dimension can be reduced according to these important bands.

For time series analysis, key points of each curve form the macro trend feature and can be fed into the following classification procedure. However, for the purpose of hyperspectral image classification, after extracting key points of prototype spectral curves, all key points are combined to form a candidate set, which act as the initially selected bands and will undergo further processing. Taking the Washington DC Mall data set as an instance, there are totally 65 bands after the above processing. For hyperspectral data, due to large number of bands, there is still high correlation and rich redundancy between those important bands, the number of bands can be further reduced.

4. Postprocessing to further reducing bands

In our preliminary work (Yu et al., 2010), it is proposed to remove redundant bands heuristically in the adjacent 5-band interval after selecting important bands according to time series key point analysis. This study further tackles this issue by thresholding conditional mutual information and employing the branch and bound algorithm to search for the optimal band combination. The branch and bound method is a kind of efficient feature combination search algorithm, which can get the optimal combination suitable for the problem without so many features. Previous studies (Pudil et al., 1994; Dy, 2008) found that for the branch and bound search method, it became impractical when the total number of features was more than 50; therefore this study comes up with preliminary filtering based on the conditional mutual information between bands in order to reduce the computation time for the following branch and bound search algorithm.

4.1. Conditional mutual information-based filtering

In feature selection, it has been recognized that the combination of individually good features do not necessarily lead to good classification performance. Because of high correlation between bands in hyperspectral image, those redundant bands must be removed. And mutual information can measure the correlation and redundancy in a proper way. When mutual information between two bands is high, it means there is high correlation between them. Conversely, redundancy is small and low degree of correlation existed between the two bands. And conditional mutual information is just the opposite to the previous statement for mutual information, which can not only reflect the degree of nonlinear correlation like mutual information but also measure relationships between features and the target labels (Novovicova et al., 2007). Therefore, this study applies conditional mutual information to measure the correlation between bands.

For two discrete random variables \( X \) and \( Y \), which have \( \Phi \) and \( \Psi \) alphabets and their joint probability density function is \( p(x, y) \), \( x \in \Phi, y \in \Psi \), the joint entropy of \( X \) and \( Y \) is defined as:

\[
H(X,Y) = -\sum_{x \in \Phi} \sum_{y \in \Psi} p(x,y) \log p(x,y)
\]

(3)

When certain variables \( Y \) are known and others \( X \) are not, the remaining uncertainty is measured by the conditional entropy:

\[
H(Y|X) = -\sum_{x \in \Phi} \sum_{y \in \Psi} p(x,y) \log p(x|y)
\]

(4)

The mutual information is usually used to measure the nonlinear correlation between two random variables and it is defined as in (Novovicova et al., 2007):

\[
I(X,Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y)
\]

(5)

If we assume that two bands are represented with two discrete variables \( X \) and \( Y \) and the class is represented with a discrete variables \( C \), when \( Y \) is given the conditional mutual information of \( X \) and \( C \) is defined as:

\[
I(C,X|Y) = H(X|Y) - H(X|C,Y) = I(C,Y) - I(C,X,Y)
\]

(6)

Due to high correlation between neighboring bands, there is no necessity to calculate condition mutual information of any two bands. The computational complexity could be reduced as long as just calculating conditional mutual information between neighboring bands. With a suitable threshold according to the range of the conditional mutual information between the adjacent bands, if the conditional mutual information of any two bands is smaller than the threshold, it means there is high redundancy between them and one of them can be discarded. Still taking Washington DC Mall data as an example, Fig. 6 gives the conditional mutual information of the neighboring important bands.

As can be seen from Fig. 6 that the range of conditional mutual information between adjacent important bands is quite large and the maximum reaches 0.45 while the minimum is only 0.05. One of the two bands can be discarded if the value is too small, which
imply that there exist high redundancy between these two bands. Taking the tradeoff of the classification accuracy and the searching efficiency in the following search algorithm into account, we choose 0.15 as the threshold, and remove one of the bands whose conditional mutual information is lower than the threshold, and also smaller mutual information with the class labels.

4.2. Branch and bound search

Currently, the only optimal feature selection algorithms are exhaustive search and the branch and bound algorithm and their variants (Nakariyakul and Casasent, 2007). Exhaustive search could find optimal feature subset by a criterion function, but it must enumerate all the feature combinations. With the increasing of feature dimension, the number of combination grows exponentially, so exhaustive search is only suitable for low-dimension problem. This study employs branch and bound algorithm to search for the final band combination. The basic idea is to search all feasible solution space of an optimization problem with a monotonic criterion function. It avoids exhaustively exploring the entire search space, which is achieved by organizing the search so that many subsets that are guaranteed to be sub-optimal are pruned. This study adopts the adaptive branch and bound algorithm recently put forward by Nakariyakul and Casasent (2007) to search for the best band combination, as it is currently one of the most computationally efficient branch and bound algorithms.

In the implementation, the resultant candidate band set from the previous step of each sample, along with their labels, are fed into the branch and bound algorithm. Given the number of chosen bands $M$, the branch and bound algorithm returns the optimal band combination. Since we do not know the exact number of the final bands, we have to search the optimal $M$ in certain range. For different data set, we have tried several Ms, and the one with the highest classification accuracy is reported as the final best combination. More details will be given in the experimental section.

5. Experimental results and analysis

After the branch and bound algorithm searches the final band combination, we perform classification to validate the optimality. In this study, the adopted classifier is SVM, as it is one of the most competitive classifiers for small samples problem. Since it is very common nowadays, we omit the principles in the text and use the LIBSVM package (Chang and Lin, 2011).

5.1. Data sets

In this study, our experimental validation is conducted on two publicly available benchmark data sets, namely the Washington DC Mall data set and the Indian Pine data set. In the former data set, two hundred and ten bands are collected in the 0.4–2.4 μm region of the visible and infrared spectrum. The water absorption bands are then deleted, resulting in 191 channels. The dataset contains 1280 scan lines with 307 pixels in each scan line, which has been studied extensively in the literature (Landgrebe, 2003; Benediktsson et al., 2005; Dundar and Landgrebe, 2004; Kuo and Chang, 2007). The Washington DC Mall data set is available in the student CD-ROM of (Landgrebe, 2003). The second data set is acquired in the 0.4–2.5 μm region by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor and made up of a 145 × 145 pixel portion of the northwest Indiana’s Indian Pines in June 1992 (Landgrebe, 2003; Serpico and Bruzzone, 2001; Serpico and Moser, 2007; Guo et al., 2008). Not all of the 220 original bands are employed in the experiments since 18 bands are affected by atmosphere absorption phenomena and are consequently discarded. Hence, the original dimensionality for the Indian Pine data set is 202 (Serpico and Moser, 2007). As stated by Serpico and Moser, the subdivision of the ground truth into training and test data is not performed randomly but defining spatially disjoint training and test fields for each class to reduce as much as possible the correlation between the samples used to train the system and those employed to test its performances. The Washington DC Mall data set has been manually annotated in (Landgrebe, 2003) and there are 137 patches of 7 classes with ground truth. We pick those pixels from the patches with even number as training set, while the rest patches with odd number as test set. The MultiSpec project file provided on the student CD-ROM (Landgrebe, 2003) is used to select training/test samples in our experiments. For this data set, there are 4428 training samples and 3651 test samples. For the Indian Pine data set, we adopt the same partition of training/test sets as (Serpico and Moser, 2007), in which there are nine classes and the number of training samples is 5012 and 3728 for the test samples.

5.2. Experimental results of the proposed method

5.2.1. The results on two benchmark data sets

For the Washington DC Mall data set, 9 clusters are obtained after spectral clustering. Then 65 important bands are selected through important point extraction. In the post processing stage, 29 bands are retained after the thresholding of conditional mutual information between neighboring bands. Finally, branch and bound algorithm searches for the optimal combination of 8 to 12 bands out of the 29 bands. Figs. 7 and 8 illustrate the time series of class 5 and 7 from the the Washington DC Mall data set, with those bands out of the 29 bands. Figs. 7 and 8 illustrate the time series of class 5 and 7 from the the Washington DC Mall data set, with those bands out of the 29 bands. Figs. 7 and 8 illustrate the time series of class 5 and 7 from the the Washington DC Mall data set, with those bands out of the 29 bands.
To note, all the results of classification accuracy in this study are obtained by the SVM classifier. From Table 1, it can be firstly observed that the important bands really contain sufficient discriminative information for classification, which is supported by the fact that the difference between the classification accuracies by the all bands and all the initial key points (bands) is not significant (although the latter is a bit lower). Secondly, satisfactory performances are achieved on both data sets, after the post processing by our proposed method in this study. For the Washington DC Mall data set, only 9 bands are selected, but the accuracy is higher than that with all the bands. For the Indian Pine data set, we choose 12 bands out of the original 202 bands, and the accuracy reaches 82.22%, which is marginally lower than that with all the bands. However, it is higher than the preliminary method in (Yu et al., 2010).

In order to test the statistical significance of the difference among the methods listed in Table 1 on the Indian Pine data set (Since all the methods achieve their accuracies over 99% on the Washington DC Mall data set with different number of bands utilized, we omit the comparison of statistical significance on that data set.), the |z| values in the McNemar’s test are given in Table 2 (at the 5% level of significance).

From Table 2, several interesting observations can be found. Firstly, the accuracy of our proposed method is not statistically different from that with all the bands, although ours is a bit lower than it (82.22% vs. 83.28%). However, we just select 12 bands from the original 202 bands. It can be concluded that the accuracies are statistically in the same level, but ours is more efficient with far less bands. For 5% level of significance, the corresponding |z| value is 1.96.

The normal test statistic

\[ z = \frac{f_{12} - f_{21}}{\sqrt{f_{12} + f_{21}}} \]  

where \( f_{12} \) denote the number of samples misclassified by method 1 but not by method 2, and \( f_{21} \) the number of samples misclassified by method 2 but not by method 1. For 5% level of significance, the corresponding |z| value is 1.96. When a |z| value is greater than this quantity, it means that the two methods to be compared have significant performance discrepancy. Table 2 gives the |z| values between the methods listed in Table 1 on the Indian Pine data set (Since all the methods achieve their accuracies over 99% on the Washington DC Mall data set with different number of bands utilized, we omit the comparison of statistical significance on that data set.).
fewer bands chosen. Secondly, the accuracy of the preliminary method in (Yu et al., 2010) is statistically lower than both ours and that with all the bands, which indicate that it really needs improvement from (Yu et al., 2010). Thirdly, the accuracy with all the initial key points is statistically different from that of our proposed method, which proves that there is a quite a lot of redundancy in those important bands and further feature selection is necessary to prune those redundant bands. And finally, the accuracy with all the initial key points is also statistically different from that with all the bands, which means there is some information loss during this processing and more efforts will be needed to retain more informative candidate bands.

Table 1 only gives the best results on the two data sets by our method. In fact, we do not know the real optimal number of bands (ground truth) and have conducted small range search as other previous studies in the literature (Serpico and Moser, 2007; Guo et al., 2008). Figs. 9 and 10 show the accuracies when selecting the optimal combination of bands with numbers varying from 8 to 12 bands, on the two test data sets respectively.

5.2.2. The sensitivity analysis of the parameters

There are 3 groups of parameters in our method. The first hyperparameter is the cluster number in the spectral clustering procedure. In this study, the Davis–Bouldin index (Davies and Bouldin, 1979) is employed to estimate the cluster number. After the visual analysis by VAT, the search range of the cluster number is narrowed down. Hence, the cluster number is determined automatically. The second hyperparameter is the threshold of conditional mutual information for the post processing of the neighboring bands. This parameter is heuristically set and will be discussed in detail later. The third set of hyperparameters is the ones of the SVM in the last stage. In this study, we adopt the grid search strategy provided in the LIBSVM package, although there are some new model selection methods in the literature (Anguita et al., 2012; Parrado-Hernández et al., 2012). The kernel function used is the radial basis function (RBF function). The two SVM parameters (i.e. C and γ) are selected based on 5-fold cross-validation during the training phase, and the search range for C is $[2^{-3}, 2^{10}]$, and $[2^{-8}, 2^{7}]$ for γ. Next we give the sensitivity analysis of the second parameter.

For the second hyperparameter, we have tried three different values, i.e., 0.1, 0.15, and 0.2. Table 3 gives the detailed results with different thresholds on the two benchmark data sets. Note the reported time is on a typical PC configured with CPU 2.0 GHz, 512 M RAM, and running Windows XP.

From Table 3, it can be found that the threshold controls the tradeoff of time complexity and the final accuracy. If the threshold is small, there will be more bands retained and the accuracy is higher, while the search time is also long. If the threshold is large, the number of retained bands is small and the final accuracy is lower, but with short search time. We choose 0.15 to report the results as it can give better tradeoff of search time and classification accuracy.

5.3. Comparisons

5.3.1. Comparison I

Guo et al. (2008) put forward a fast greedy search-based feature selection algorithm. Their method chose the bands one by one until the number of bands reached the desired number and the chosen bands must contribute to category separability as much as possible, but low correlation with the previously selected bands. The greedy strategy is approximated by setting up a spectral window with a certain bandwidth w. If the bands are within the chosen band window, their conditional mutual information is approximated by the corresponding mutual information with a factor b. However, Guo et al. did not specify these two parameters or tell the readers how to choose them in their paper. In this study, the best accuracy of their algorithm is obtained by enumerating values of w and b in certain ranges. The value of w is selected in the following numbers: 4, 6, 8, 10, and the value of b are got by selecting from the range of 0.1–0.9, with a step of 0.1. Twenty bands are selected according to the values of these two parameters, and then classification accuracy with the same test data is calculated. For the Indian Pine data set, the experimental results are shown in Table 4.

From Table 4, it can be found that the highest accuracy 66.71% is achieved when $w=8$, $b=0.7$. Hence, in order to get the best results of Guo et al., w and b take these two values in the following comparison experiment. And then 10, 20, 30, 40, 50, 60, 70, 80 bands are selected respectively, in the same manner as in (Guo et al., 2008). Then the corresponding classification accuracies are computed on the same test data set. To note, because the number of bands has been reduced to 42 after reducing redundant bands using the conditional mutual information filtering presented in the previous section, we can only choose 10, 20, 30, 40 bands respectively. The detailed comparison is shown in Fig. 11.

Meanwhile, for the Washington DC Mall data set, we also compare the method in (Guo et al., 2008) and ours. Since we have already reduced the number of bands to 29 before the branch and bound search, we can only select 10, 20, 29 bands respectively. The detailed results are given in Fig. 12.

From Figs. 11 and 12, it can be observed that the performance of our method is much better than the fast feature selection method proposed in (Guo et al., 2008) on the two benchmark data sets, with small number of chosen bands. Moreover, the classification accuracy does not always increase with the growing of the number of bands selected, which indicates that there are redundancy among the hyperspectral bands and it is meaningful and necessary to select bands before classifying the pixels in hyperspectral image.
5.3.2. Comparison II

The classification accuracy is 99.51% with 7 bands selected, using the method proposed in this study on the Washington DC Mall data set. Compared with the hybrid genetic algorithm presented in (Li et al., 2011), in which the classification accuracy is 98.25% when searching the same amount of bands, we can see that the presented method is more effective. On the other hand, the GA-SVM method in (Li et al., 2011) has a higher computational complexity: it needs about 300 min to accomplish one hundred iterations. However, the method proposed in this study needs only about 30 min to finish the whole processing, with the same configuration of a general personal computer. The branch and bound algorithm takes most of the time in the whole procedure. It varies with the number of chosen bands in the final stage. For the Washington DC Mall data set, it spends about 20 min selecting 10 bands from the remaining 29 bands. If the branch and bound algorithm is directly carried out on the set of initial important bands, the computational cost will be even higher. Thus, it can also be concluded that the filtering by conditional mutual information is helpful.

For the Indian Pine data set, the classification accuracy is 82.22% when selecting only 12 bands. It is marginally lower than the accuracy with all the original 202 bands (83.28%). According to Table 2, they are at the same level. Although the classification accuracy does not improve, the dimension or bands number is greatly reduced. We only choose about 6% percentage from the original 202 bands to achieve same level of accuracy. Serpico and Moser (2007) proposed three search algorithms for band selection and conducted experiments on the same Indian Pine data set as ours. They have employed the MAP classifier in their paper, but we have utilized the SVM classifier. To compare fairly, we conduct the train/test experiments on the final bands searched by their methods, but using the same classifier of SVM as employed in this study. The highest accuracy of their method is 86.64%, with 26 s-bands selected. To note, their finally chosen bands are not from the original 202 bands, but the s-bands, which are obtained by averaging a group of contiguous channels of the hyperspectral image, thus containing much more information with actually more bands involved.

In summary, our method is more competitive on the above two benchmark data sets, with far fewer bands selected, lower computational cost, and satisfactory classification accuracies.

6. Conclusions

Due to the special characteristics of hyperspectral remote sensing image data, shape similarity analysis approach is borrowed to narrow down candidate bands and achieve discriminative information preservation. This paper has put forward a novel band selection method based on time series key point extraction, with spectral clustering as the preprocessing step, filtering with conditional mutual information and branch and bound search algorithm as the post processing steps. Comprehensive experiments on two benchmark data sets have proved that compared with the existing methods, the proposed method is competitive with lower computational cost, far fewer bands selected, but with same level of classification accuracies.

There are at least two purposes of feature selection, one is to reduce the number of features and boost the computational efficiency, while improving the classification accuracy could be another one. In this study, we are interested in finding the

### Table 3

<table>
<thead>
<tr>
<th>The threshold</th>
<th>Washington DC Mall</th>
<th>Indian Pine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The retained bands</td>
<td>Finally chosen bands</td>
</tr>
<tr>
<td>0.10</td>
<td>55</td>
<td>9</td>
</tr>
<tr>
<td>0.15</td>
<td>29</td>
<td>9</td>
</tr>
<tr>
<td>0.20</td>
<td>21</td>
<td>9</td>
</tr>
</tbody>
</table>

### Table 4

<table>
<thead>
<tr>
<th>b (%)</th>
<th>0.1 (%)</th>
<th>0.2 (%)</th>
<th>0.3 (%)</th>
<th>0.4 (%)</th>
<th>0.5 (%)</th>
<th>0.6 (%)</th>
<th>0.7 (%)</th>
<th>0.8 (%)</th>
<th>0.9 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>w=4</td>
<td>60.43</td>
<td>59.76</td>
<td>60.92</td>
<td>63.71</td>
<td>63.62</td>
<td>63.28</td>
<td>66.68</td>
<td>66.01</td>
<td>66.42</td>
</tr>
<tr>
<td>w=6</td>
<td>60.54</td>
<td>60.97</td>
<td>62.45</td>
<td>61.37</td>
<td>64.32</td>
<td>64.79</td>
<td>66.15</td>
<td>66.52</td>
<td>66.01</td>
</tr>
<tr>
<td>w=8</td>
<td>62.98</td>
<td>64.02</td>
<td>62.23</td>
<td>63.52</td>
<td>62.01</td>
<td>62.29</td>
<td>66.71</td>
<td>66.04</td>
<td>66.47</td>
</tr>
<tr>
<td>w=10</td>
<td>64.62</td>
<td>60.76</td>
<td>62.93</td>
<td>62.90</td>
<td>62.69</td>
<td>62.85</td>
<td>62.55</td>
<td>66.28</td>
<td>66.01</td>
</tr>
</tbody>
</table>
minimum number of bands with satisfactory classification accuracies. Due to the high correlation and high redundancy among the hyperspectral bands, traditional feature selection methods cannot be directly utilized. In the pattern recognition literature, branch and bound search algorithm is capable of producing the optimal feature subset. However, it can only cope with not so many features (usually less than 30 features). Therefore, we tackle the problem from the novel perspective of spectral shape classification and have proposed to extract the important points (bands) from those prototypical spectral curves generated by spectral clustering and reduce the number of bands in the first stage. From the results in Tables 1 and 2, it is observed that this process incurs somewhat information loss. In the future, more efforts will be directed to pick out more informative bands in the original hyperspectral bands. In this study we have preprocessed the hyperspectral data to cater for the branch and bound algorithm in the last step. That is just one of the two sides of the coin. How to modify the branch and bound search algorithm and improve its search efficiency taking into account of the characteristics of hyperspectral data is also one of the concerned issues of further research interests.

Acknowledgments

Part of this work is finished when the first author is visiting Arizona State University. The authors would like to thank the anonymous reviewers for their constructive comments. The Washington DC Mall data set are obtained from the student CD-ROM which accompanies Prof. Landgrebe's book. The authors are also grateful to Prof. Serpico and Dr. Moser for providing us the training/test samples of Indian Pine data set. This work is partially funded by National Natural Science Foundation of China (Grant no. 61172002 and 51079040), the Key Technology R&D Program of Jiangsu Province (BE2012179).

References


