Abstract: In this paper we propose a new feature extraction scheme for hyperspectral images based on mutual information. Relevance of extracted feature set to class label has been measured by average of mutual information between each of them and class label and Redundancy of them is measured by average of mutual information between each pair of them. Based on relevance of features and redundancy between them, we propose a cost function that maximize relevance of extracted features and simultaneously minimize redundancy between them. This cost function has been already used for feature selection. In this paper we will find the parameters of an optimal linear mapping by optimizing the proposed cost function with respect them. Linear methods are attractive due to their simplicity. Because of nonlinear and nonconvex relation between proposed cost function and the parameters, we use genetic algorithm for optimization. Mutual information accounts for higher order statistics, not just for second order as PCA and LDA do. Hence mutual information is a better criterion for hyperspectral images because they have higher order statistics than two. Our classification results for AVARIS data shows proposed method has better performance over PCA and LDA.

Keywords: Hyperspectral Image, Classification, Feature Extraction, Mutual Information, and Genetic Algorithm.

1. Introduction

Hyperspectral imagery consists of measurement of a larger number of closely spaced narrow bands that cover visible and near infrared portion of electromagnetic spectrum. These images are used successfully for recognition and analyzing land cover types. The main problem for classification of these images is “curse of dimensionality” or “Hughes effect” that is the ratio of training samples to the number of bands is small, typically less than ten. Under this condition, parameter estimating of the classifier is less reliable and the probability of error is high. Generating of training samples is very time consuming and expensive. Hence, we have to reduce the dimension of feature space to increase this ratio. Band selection and feature extraction are two possibilities.

There are strong correlations between adjacent bands. Hence the effective dimensionality of feature space is less than number of bands. This is the main reason for reducing bands.

In feature selection algorithms, we try to select a number of bands that have little correlation with each other and each of them has relevant information to class labels. Instead of correlation (linear dependency), we can use total independency for selecting bands. Mutual information is a criterion for measuring dependency between two random variables. Hence, we can replace correlation coefficient between two bands with mutual information between them and gain higher accuracy in classification.

Obviously discriminatory information is distributed over bands and with canceling a band, we loss its new information. Hence, feature extraction is a better scenario than band selection. In feature extraction, we find a mapping from available space to a lower dimension space with little loss of information. In general this mapping is nonlinear. Feature extraction methods can divide to supervised and unsupervised one. In unsupervised feature extraction, we try to extract the manifold of data in available feature space regardless of their class labels. Usually the dimension of this manifold is smaller than the number of bands. Principal component analysis (PCA) and multidimensional scaling (MDS) are classical techniques for this purpose [1], but they are suitable for linear manifolds. The nonlinear structure of hyperspectral data is invisible to these. One extension of these algorithms is by using kernel methods. Kernel PCA is an extension of PCA that can extract manifolds with nonlinear structures [2]. Results of kernel methods highly depend on selected kernel and can be very bad with selecting inappropriate kernel for a particular dataset. Other algorithms for extracting of nonlinear manifolds such as ISOMAP, LDE, and Hessian Eigenmaps fall into the general area of manifold learning [3]. Main problem of manifold learning algorithms is treatment of sparse and clustered data such as hyperspectral data.

One of the most important algorithms for supervised feature extraction is linear discriminate analysis (LDA)
[4]. LDA aims to map an original feature space to another space with as little less in discriminatory information using maximizing Fisher ratio. One extension of LDA is kernel linear discriminate analysis (KLDA) based on kernel methods [5]. As mentioned before, results of these methods highly depend on selected kernel and selecting of an appropriate kernel for a particular dataset via a data-driven approach is a very difficult task.

Independent component analysis (ICA) is also a linear algorithm for extracting independent features. Feature extraction using ICA is equivalent to minimum redundancy criterion that both of them attempt to extract independent features. ICA does not consider relevance of features to class labels and is an unsupervised feature extraction method. Hence its results are weaker than our proposed method. Band selection was done using ICA by H. Du, et. al. [6].

In this paper we use maximum relevance, minimum redundancy cost function instead Fisher ratio. By optimizing this cost function, we can extract features with highest relevance to the target class and with minimal redundancy. Hence, this method is a combination of supervised and unsupervised feature extraction scenarios. Then we optimize this cost function with respect to the parameters of a linear mapping by genetic algorithm. In [8], K. Torkokola has used only maximum relevance criterion for feature extraction.

Mutual information accounts for higher order statistics, not just for first and second order as Fisher ratio do. Hence, for datasets with higher order statistics such as hyperspectral data, this leads to maximally clustered data.

In section two, we introduce the proposed criteria and discuss its optimization. In section three, we present results of classification for AVARIS data based on proposed criteria and compare them with PCA and LDA. Conclusion is presented in section 4.

2. MAXIMUM RELEVANCE, MINIMUM REDUNDANCY CRITERIA FOR FEATURE EXTRACTION

In this paper the original features are represented by \( x \) and extracted ones by \( y \). Maximum relevance is to extract features that satisfy

\[
\text{max } R = \frac{1}{|c|} \sum_{c \in C} I(c, y_i)
\]

\( S \) represents the extracted feature set and \(|s|\) represents the number of extracted features. \( C \) represents class label, \( i \)th training sample and \( c \) is mutual information between them. \( I(c, y_i) \) is zero (minimum) if \( c \) and \( y_i \) are independent and is maximum if they have functional dependency i.e., \( c=f(y_i) \). Hence \( I(c, y_i) \) can be interpreted as relevance of feature \( y_i \) to class label \( c \) and \( D \) is the average relevance of feature set \( S \).

It is likely that extracted features according to equation (1) could have rich redundancy. When two features highly depend on each other, one of them can be removed without any change in the discriminatory power of feature set. Therefore the following minimal redundancy condition can be added to select mutually exclusive features.

\[
\text{min } R = \frac{1}{|c|} \sum_{c \in C} I(c, y_i)
\]

2. MAXIMUM RELEVANCE, MINIMUM REDUNDANCY CRITERIA FOR FEATURE EXTRACTION

In this paper the original features are represented by \( x \) and extracted ones by \( y \). Maximum relevance is to extract features that satisfy

\[
\text{max } R = \frac{1}{|c|} \sum_{c \in C} I(c, y_i)
\]

\( S \) represents the extracted feature set and \(|s|\) represents the number of extracted features. \( C \) represents class label, \( i \)th training sample and \( c \) is mutual information between them. \( I(c, y_i) \) is zero (minimum) if \( c \) and \( y_i \) are independent and is maximum if they have functional dependency i.e., \( c=f(y_i) \). Hence \( I(c, y_i) \) can be interpreted as relevance of feature \( y_i \) to class label \( c \) and \( D \) is the average relevance of feature set \( S \).

It is likely that extracted features according to equation (1) could have rich redundancy. When two features highly depend on each other, one of them can be removed without any change in the discriminatory power of feature set. Therefore the following minimal redundancy condition can be added to select mutually exclusive features.

\[
\text{min } R = \frac{1}{|c|} \sum_{c \in C} I(c, y_i)
\]

2. MAXIMUM RELEVANCE, MINIMUM REDUNDANCY CRITERIA FOR FEATURE EXTRACTION

In this paper the original features are represented by \( x \) and extracted ones by \( y \). Maximum relevance is to extract features that satisfy

\[
\text{max } R = \frac{1}{|c|} \sum_{c \in C} I(c, y_i)
\]

\( S \) represents the extracted feature set and \(|s|\) represents the number of extracted features. \( C \) represents class label, \( i \)th training sample and \( c \) is mutual information between them. \( I(c, y_i) \) is zero (minimum) if \( c \) and \( y_i \) are independent and is maximum if they have functional dependency i.e., \( c=f(y_i) \). Hence \( I(c, y_i) \) can be interpreted as relevance of feature \( y_i \) to class label \( c \) and \( D \) is the average relevance of feature set \( S \).

It is likely that extracted features according to equation (1) could have rich redundancy. When two features highly depend on each other, one of them can be removed without any change in the discriminatory power of feature set. Therefore the following minimal redundancy condition can be added to select mutually exclusive features.

\[
\text{min } R = \frac{1}{|c|} \sum_{c \in C} I(c, y_i)
\]

2. MAXIMUM RELEVANCE, MINIMUM REDUNDANCY CRITERIA FOR FEATURE EXTRACTION

In this paper the original features are represented by \( x \) and extracted ones by \( y \). Maximum relevance is to extract features that satisfy

\[
\text{max } R = \frac{1}{|c|} \sum_{c \in C} I(c, y_i)
\]

\( S \) represents the extracted feature set and \(|s|\) represents the number of extracted features. \( C \) represents class label, \( i \)th training sample and \( c \) is mutual information between them. \( I(c, y_i) \) is zero (minimum) if \( c \) and \( y_i \) are independent and is maximum if they have functional dependency i.e., \( c=f(y_i) \). Hence \( I(c, y_i) \) can be interpreted as relevance of feature \( y_i \) to class label \( c \) and \( D \) is the average relevance of feature set \( S \).

It is likely that extracted features according to equation (1) could have rich redundancy. When two features highly depend on each other, one of them can be removed without any change in the discriminatory power of feature set. Therefore the following minimal redundancy condition can be added to select mutually exclusive features.

\[
\text{min } R = \frac{1}{|c|} \sum_{c \in C} I(c, y_i)
\]
After having observed feature y, the uncertainty of class label is conditional entropy as

$$H(C|y) = - \int_y p(y)(\sum_c p(c) \log(P(c))) dy$$  \hspace{1cm} (8)

The difference between prior entropy and conditional entropy is amount of information that y brings about class label c. Hence mutual information is defined as

$$I(c,y) = H(C) - H(C|y) = \sum_c \int_y p(c,y) \log \frac{p(c,y)}{p(c)p(y)} dy$$  \hspace{1cm} (9)

Hence mutual information is a good measure of relevance of a feature to class label. Mutual information can also be seen as the Kullback-Leibler (KL) divergence between $p(c,y)$ and $p(c)p(y)$. In general for two densities $f(y)$ and $g(y)$, this is defined as

$$KL(f,g) = \int_y f(y) \log \frac{f(y)}{g(y)} dy$$  \hspace{1cm} (10)

Hence, we have

$$I(c,y) = KL(p(c,y), p(c)p(y))$$  \hspace{1cm} (11)

Hence, mutual information is KL distance between joint distribution and the product of marginal distributions. With this interpretation of mutual information, we can replace mutual information with any distance between two distributions without any change in results. One choice is quadratic mutual information defined as

$$l_q(c,y) = \sum_c \int_y (p(c,y) - p(c)p(y))^2$$  \hspace{1cm} (12)

Using a Parzen density estimator with Gaussian kernel for estimation of $p(c,y)$ and $p(y)$, we have

$$p(c_p,y) = \frac{1}{N} \sum_{j=1}^{J_p} G(y - y_{jp}, \sigma^2 I), \hspace{0.5cm} p = 1, ..., N_c$$

$$p(c_p) = \frac{J_p}{N}$$

$$p(y) = \sum_c p(c,y) = \frac{1}{N} \sum_{p=1}^{N_c} \sum_{j=1}^{J_p} G(y - y_{jp}, \sigma^2 I) = ...$$

$$\frac{1}{N} \sum_{l=1}^{N} G(y - y_l, \sigma^2 I)$$

$$G(y, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}} \frac{1}{\Sigma^{\frac{d}{2}}} \exp \left( -\frac{1}{2} y^T \Sigma^{-1} y \right)$$  \hspace{1cm} (13)

$J_p$ is the number of training samples in $p$th class and $N$ is total number of training samples.

Inserting (13) to (12) then to (1), we will have [8]

$$D = \frac{1}{N^2} \sum_{p=1}^{N_c} \sum_{p=1}^{J_p} \sum_{p=1}^{J_p} G(y_{jp} - y_{jp}, 2\sigma^2 I) + ...$$

$$\frac{1}{N^2} \left( \sum_{p=1}^{N_c} \frac{J_p^2}{N} \right) \sum_{p=1}^{N_c} \sum_{p=1}^{J_p} G(y_{jp} - y_{jp}, 2\sigma^2 I) + ...$$

$$\frac{1}{N^2} \sum_{p=1}^{N_c} \sum_{p=1}^{J_p} \sum_{p=1}^{J_p} G(y_{jp} - y_{jp}, 2\sigma^2 I)$$  \hspace{1cm} (14)

$y_{jl}$ is $j$th sample of $p$th class. $N_c$ is the number of classes and $\sigma$ is an arbitrary window size for parzen estimator.

There are many estimators for estimating $I(y_j, y_j)$. In this paper we use from the below estimator [9]
Where \( K_1(.) \) and \( K_2(.) \) are separable kernel functions. In this paper, we use Gaussian kernels with the same variance. \( y_{in} \) is \( n \)th sample of \( i \)th feature. Computation complexity of this estimator is \( O(N^3) \). Inserting (15) to (2), we will estimate redundancy of feature set \((R)\).

Now, replacing \( y \) with \( Wx \) in \( \phi(D,R) \), we will have a cost function of \( W \) and training samples \( \{x_i\}_{i=1}^m \). Now \( W \) optimum can be computed by solving the below optimization problem:

\[
W^* = \arg \max_{W} \phi(D,R)
\]

\[
\text{S.t.} \quad WW^T = I
\]

(16)

Regarding the nonlinear and nonconvex nature of cost function with respect to parameters, gradient based optimization algorithms have possibility of trapping in local minimums. We use genetic algorithm for solving this optimization problem. \( GA \) starts with a fixed population of candidate solutions and each of the candidates is evaluated with a fitness function that is a measure of the candidate potential as a solution to the problem. The fitness function maps an individual of the population to a scalar. Genetic operators like selection, crossover and mutation are implemented to simulate the natural evolution. A population, usually presented by a binary string is modified by the probabilistic application of the genetic operators from one generation to the next. The fitter individual has more chance for reproduction in next generation proportional to their rank. In uniform crossover, each gene of children comes with equal probability from one of the parents.

Orthornormal constraint on \( W \) is accomplished by orthonormalizing \( W \) after each generation by Gram-Schmidt procedure. We repeat this process until the variation of a norm of matrix \( W \) is smaller than a preselected threshold.

3. EXPERIMENTAL RESULTS

The hyperspectral dataset used in our experiments is a scene taken over NW Indiana’s Indian Pine by the AVIRIS sensor. From the 220 spectral channels acquired by the AVIRIS, 20 channels were discarded because of effect of atmospheric phenomena. In Fig. 1, bands 16, 29, and 55 are used for red, green, and blue colours, respectively, and a pseudocolor image of the scene is constructed. The available ground truth map which includes 16 different land-cover classes was used to generate a set of 2449 training samples for learning the classifiers and also a set of 7919 test samples for assessing their accuracies. PCA and LDA used as linear feature extraction algorithms in different experiments and are compared with proposed algorithm.

Fig. 2 represents Scatter plot for the first three features extracted by PCA for classes 1, 2, and 5. It is clear that extracted features by PCA have maximum variance but distributions are highly overlapped.

Fig. 3 represents Scatter plot for the first three features extracted by LDA. It is clear that extracted features by LDA have less overlap than ones extracted by PCA. Hence classification by them has higher accuracy as is clear in Fig. 7.

Fig. 4 represents Scatter plot for the first three features extracted by proposed algorithm. It is clear extracted features by proposed algorithm have less overlap than ones extracted by PCA and LDA. Hence classification by proposed algorithm has higher accuracy as is clear in Fig. 7.

Fig. 5 represents class map for classification by the first thirty extracted features by LDA and a maximum likelihood (ML) classifier and Fig. 6 represents this for proposed algorithm. It is clear proposed algorithm works very well and the produced class map is satisfying.

Fig. 7 represents classification accuracy for features extracted by PCA, LDA, and proposed method. It is clear extracted features by proposed algorithm have less overlap than ones extracted by PCA and LDA. Hence classification by proposed algorithm has higher accuracy as is clear in Fig. 7.

Fig. 5 represents class map for classification by the first thirty extracted features by LDA and a maximum likelihood (ML) classifier and Fig. 6 represents this for proposed algorithm. It is clear proposed algorithm works very well and the produced class map is satisfying.

Fig. 7 represents classification accuracy for features extracted by PCA, LDA, and proposed method. It is clear that proposed method outperform PCA and LDA that are algorithms based on first and second order statistics. It can be seen that classification accuracy of proposed algorithm is more than PCA and LDA.
Figure 1. Pseudocolor image of AVIRIS data (Left), available ground truth map for the 16 classes (right).

Figure 2. Scatter plot of the first three features extracted by PCA for classes 1, 2, and 5.

Figure 3. Scatter plot of the first three features extracted by LDA for classes 1, 2, and 5.

Figure 4. Scatter plot of the first three features extracted by proposed algorithm for classes 1, 2, and 5.

Figure 5. Classification results of LDA by ML classifier.

Figure 6. Classification results of proposed algorithm by ML classifier.
4. CONCLUSION

In this paper, we proposed a feature extraction method for hyperspectral images based on optimizing maximum relevance, minimum redundancy criteria. These criteria are based on mutual information between extracted features and class labels for measuring relevance of them to class labels and mutual information between each pair of extracted features for measuring the redundancy level of extracted feature set. Mutual information accounts for higher order statistics, not just for second order statistics and is suitable for hyperspectral data that have higher order statistics than two. Then we find an optimum linear mapping by optimizing proposed cost function with respect to its parameters. Proposed method can produce features with distributions that have little overlap and can lead to higher classification accuracy. Reported results for proposed method show better performance compared to PCA and LDA that use only first and second order statistics. Because of nonlinear and nonconvex relation between cost function and parameters, we use genetic algorithm for optimizing the criterion and finding the optimum parameters.

References