

# USING A LABORATORIAL HYPERSPECTRAL IMAGE FOR THE EVALUATION OF FEATURE REDUCTION METHODS FOR THE CLASSIFICATION OF HIGH DIMENSIONAL DATA

Yasser Maghsoudi\*, Jung-il Shin\*, Sun-hwa Kim and Kyu-sung Lee  
ymaghsoudi@yahoo.com , Jungil79@inhaian.net , ksung@inha.ac.kr

Department of Geoinformatic engineering, Inha university, S. Korea

Commission VII, WG VII/3

**KEY WORDS:** Hyperspectral Sensing, Feature Selection, Feature Extraction, Classification

## ABSTRACT:

These rapid advances in hyperspectral sensing technology have made it possible to collect remote sensing data in hundreds of bands. However, the data analysis methods which have been successfully applied to multispectral data are often limited to achieve satisfactory results for hyperspectral data. The major problem is the high dimensionality, which deteriorates the classification due to the Hughes phenomenon. In order to avoid this problem a feature reduction process is inevitable. There are currently many different methods for feature reduction process in hyperspectral data. The feature selection methods pick the most informative features and discard the redundant features from the total set of features. Feature extraction methods, on the other hand, transform a large amount of information into a small number of transformed features. The decision boundary feature extraction (DBFE) and nonparametric weighted feature extraction method (NWFE) are two important approaches for feature extraction. Another group of feature reduction algorithms are based on the theory of multiple classifiers. Thus far, many different methods for the feature reduction process have been proposed but the validation of these algorithms has not yet been done on an appropriate image dataset. The main goal of this study is to have a good evaluation of these different feature reduction algorithms based on a laboratorial hyperspectral data. Selection of classes for the simulated target was based on the challenging point of different algorithms which are classifying targets with very similar spectral characteristics, targets with different shapes, targets with high different spectral characteristics or targets with high spatial variability. In this respect following the aforesaid criteria 22 classes were considered in the final simulated target. The feature reduction methods were compared using the test image. The consistency between the various methods is discussed as well as the implication of feature reduction on image classification.

## 1. INTRODUCTION

Recently, advances in hyperspectral remote sensing have provided a powerful tool for the monitoring of the earth's surface. The resulting high dimensional data, in one hand, can provide a better discrimination of the spectral classes. On the other hand, it poses some new challenges to image classification. Hence, the data analysis methods which have been successfully applied to multispectral data in the past, often fail to achieve satisfactory results for hyperspectral data as well. The major problem in applying the traditional image classification methods into hyperspectral data is the high dimensionality, which deteriorates the classification due to the Hughes phenomenon (Hughes, 1968). In order to avoid this problem the number of training samples must increase as the number of spectral bands increases. As the number of training samples available is usually limited, the increase in the number of features can bring about a significant increase in the classification error. An approach to mitigate these problems is feature reduction. Feature reduction refers to the process of reducing the dimensions of the feature vector while preserving the needed information. The feature reduction methods generally fall into feature selection and feature extraction.

The main goal of the feature selection methods is to pick the most informative features and discard the redundant features from the total set of features. Recently, there have been a large number of algorithms proposed for the purpose of feature

selection (Bajcsy and Groves, 2004; Sheffer and Ultchin, 2003, Serpico and Bruzzone, 2000). The sequential methods are probably the most common group of methods. These methods are ranging from sequential forward selection (SFS) and sequential backward selection (SBS) methods (Kittler, 1986). SFS starts from an empty set. It iteratively generates new feature sets by adding one feature which is selected by some evaluation function. SBS, on the other hand, starts from a complete set and generates new subsets by removing a feature selected by some evaluation function. The main problem of these two algorithms is that the selected features can not be removed (SFS) and the discarded features can't be reselected (SBS). To overcome these problems Pudil et al. (1994) proposed the floating versions of SFS and SBS. Sequential forward floating search algorithms (SFFS) can backtrack unlimitedly as long as it finds a better feature subset. SFFS is the backward version.

Genetic feature selectors are a series of feature selection methods which use genetic algorithm to guide the selection process (Siedlecki and Sklansky, 1989). In genetic feature selection each feature subset is represented by a chromosome which is binary string including 0's and 1's, which corresponds to a discarded or selected features respectively. New chromosomes are generated using crossover, mutation and reproduction operators. Ferri et al. (1994) compared SFS, SFFS, and the genetic algorithm methods on data sets with up to 360 dimensions. Their results showed that SFFS gives good

performance even on very high dimensional problems. They showed that the performance of genetic algorithm, while comparable to SFFS on medium-sized problems, degrades as the dimensionality increases.

Steepest ascent (SA) search algorithm was proposed for feature selection in hyperspectral data (Serpico and Bruzzone, 2000). If  $n$  is the total number of features and  $m$  is the desired number of features, SA is based on the representation of the problem solution by a discrete binary space, which is initialized with a random binary string containing  $m$  "1" and  $(n - m)$  "0". Next, it searches for constrained local maximas of a criterion function in such space. A feature subset is a local maximum of the criterion function if the value of that feature subset criterion function is greater than or equal to the value the criterion function takes on any other point of the neighborhood of that subspace. They also proposed fast constrained (FC) search algorithm which is the computationally reduced version of SA. Unlike the SA for which the exact number of steps is unknown in advance, FC Search method exhibits a deterministic computation time. A comparative study of feature reduction techniques showed that FC is always faster than or as fast as SA (Serpico et al., 2003). Further, SA and FC methods allowed greater improvements than SFFS.

The basic idea of the feature extraction methods, on the other hand, is to transform a large amount of information into a small number of transformed features. There are many feature extraction methods. They fall into two categories, parametric and nonparametric feature extractions (Fukunaga, 1990). Discriminant analysis feature extraction (DAFE) is probably the first method used for feature extraction. Since it uses the mean vector and covariance matrix of each class, it is also called the parametric feature extraction method (Fukunaga, 1990). The purpose of DAFE is to find a transformation matrix such that the class separability of transformed data is maximized. Usually between-class over within-class scatter matrices (Fisher criteria) is used to formulate the criteria of class separability. Approximated pairwise accuracy criterion Linear Dimension Reduction (aPAC-LDR) uses a weighted fisher criterion which is defined as the weighted contributions of individual class pairs according to the Euclidian distance of respective class means (Duin and Haeb-Umbach, 2001). These two methods, though distribution free, works well with normal-like distributions. Since the rank of within-class scatter matrix is the number of classes  $(N) - 1$ , so only  $N-1$  features can be extracted using these methods. Nonparametric methods were proposed to solve the problems of the parametric approaches. Nonparametric Discriminant Analysis (NDA) proposes a nonparametric definition of the between class scatter matrix (Fukunaga, 1990). To improve the DAFE, this method uses local information. The Nonparametric Weighted Feature Extraction method (NWFE) is developed to solve limitations of the so-called methods (Kuo and Landgrebe, 2004). NWFE takes advantages of desirable characteristics of pervious methods, while avoiding their shortcomings. The main idea of NWFE is putting different weights on each sample to compute the "weighted means" and defining new nonparametric between-class and within-class scatter matrices to obtain more than  $N-1$  features.

There also have been many researchers that applied the idea of multiple classifiers for the classification of hyperspectral data. Multiple "data sources" were also used for the classification of hyperspectral data (Benediktsson and Kanellopoulos, 1999). Based on the correlation of the input bands, they split the hyperspectral data into several smaller data sources. Kumar et al. (2001) developed a pairwise feature extraction. They

decomposed a  $c$ -class problem into two-class problems. For each pair they extract features independently, and a Bayesian classifier is learned on each feature set. The outputs of all those classifiers are then combined to determine the final decision of a pixel.

Lack of a good ground truth data has always been a crucial problem for the evaluation of the feature reduction methods. Considering the challenging points of different algorithms i.e. classifying targets with very similar spectral characteristics, targets with different shapes, targets with high different spectral characteristics and targets with high spatial variability, a laboratorial hyperspectral image was taken which contains 22 classes. In this paper a comparative evaluation of some of classical and also recently proposed feature reduction algorithms is carried out using hyperspectral data obtained by a portable hyperspectral camera.

## 2. METHODOLOGIES

### 2.1 Feature Selection Algorithms

Feature selection algorithms involve both a search strategy and an evaluation function (Jain and Zongker, 1997). The aim of the search algorithm is to generate subsets of features from the original feature space  $X$  and the evaluation function compares these feature subsets in terms of discrimination. We assume that the original feature space has  $n$  features and the aim of feature selection is to select a good subset of  $m$  features.

In this study all considered feature selection methods employed Jeffries-Matusita (JM) distance, which is an inter-class measure, as a criterion for the evaluation of feature subsets. The JM distance is as follows:

$$JM = 2 \sum_{n=1}^N \sum_{m>n}^N JM_{mn} \quad (1)$$

$$JM_{mn} = \sqrt{2(1 - e^{-b_{mn}})} \quad (2)$$

$$b_{mn} = \frac{1}{8} (M_m - M_n)^T \left( \frac{C_m + C_n}{2} \right)^{-1} (M_m - M_n) + \frac{1}{2} \ln \left( \frac{|C_m + C_n|}{\sqrt{|C_m| |C_n|}} \right) \quad (3)$$

where  $N$  is the number of classes,  $b_{ij}$  is the Bhattacharyya distance between class  $i$  and  $j$  and  $M_i$  and  $C_i$  are the mean vector and covariance matrix of the class  $i$  respectively.

**2.1.1 Genetic Feature Selector (GFS):** In a genetic algorithm a possible solution of the problem under consideration is represented by a chromosome which is binary string including 0's and 1's, which corresponds to a discarded or selected features respectively. In the initialization step of the algorithm a set of chromosomes are created randomly. The actual set of chromosomes is called the population. A population with the size of 100 is used in this study. The fitness function, which is JM distance in our study, is defined to represent the quality of the solution given by a chromosome.

The tournament selection was adopted in this study. Pairs of individuals are picked at random from the population (Goldberg and Deb, 1991). Whichever has the higher fitness is copied into a mating pool (and then both are replaced in the original population). This is repeated until the mating pool is full. In the selection process the two chromosomes with the highest value of fitness are copied to the mating pool without any tournament. When the mating pool in full, the crossover and mutation

operators are applied to the chromosomes in the mating pool. A single point crossover is used in this study: one crossover point is considered, the offspring gene from beginning of chromosome to the crossover point is copied from one parent, and the rest is copied from the second parent. The probability of the crossover was set to 90 %. The mutation operator is applied to all new chromosomes produced by the crossover operator. It randomly alters each gene with a small probability (In this study 0.002). Mutation provides a small amount of random search, and helps ensure that no point in the search space has a zero probability of being examined.

The crossover and mutation operators are applied to the selected parents until there are enough offspring for the generation of a new population. At this time the old population is replaced by the new one (Figure 1). If the value of the maximum fitness in the population is the same in 50 generations in a row the algorithm is terminated.

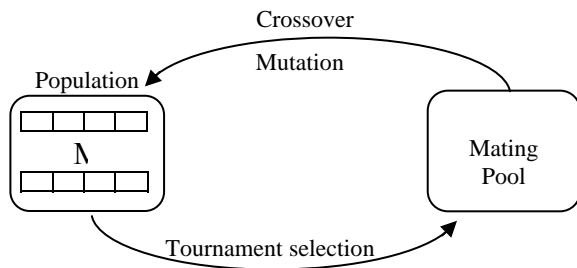


Figure 1. The scheme of the genetic algorithm used in the study

The genetic approach, here, incorporates an appropriate penalty function to force the algorithm finish with the prespecified number of features. Without using this penalty function the algorithm would tend to select the highest number of features. Adopting this penalty function, the following fitness function has been defined.

$$J(X_{d'}) = JM(X_{d'}) \cdot e^{-\frac{(d'-d)}{\sigma}} \quad (4)$$

In which  $d$  is the required number of features and  $d'$  is the number of features selected in each iteration.

If the value of the maximum fitness in the population is the same in 50 generations in a row the algorithm is terminated. The features of the chromosome with the highest fitness value are the final result and are used for the classification.

**2.1.2 Fast Constrained Search (FCS) Algorithm:** Fast constrained (FC) search algorithm is the computationally reduced version of SA. Unlike the SA for which the exact number of steps is unknown in advance, FC Search method exhibits a deterministic computation time. The method is described algorithmically as follows (Serpico and Bruzzone, 2000):

START from an initial feature subset  $S_0$  composed of  $m$  features selected from  $X$   
 Set the current feature subset  $S_k$  to  $S_0$   
 Compute the complementary subset  $D_k$  of  $S_k$

FOR each element  $s_i \in S_0$   
 FOR each element  $S_j \in D_k$   
 Generate  $S_{ij}$  by exchanging  $s_i$  for  $s_j$  in  $S_k$   
 Compute the value  $J(S_{ij})$  of the criterion function  
 CONTINUE  
 Set  $J_{i,max}$  to the maximum of  $J(S_{ij})$  obtained by exchanging  $s_i$  for any possible  $s_j$   
 IF  $J_{i,max} > J(S_k)$ , THEN update  $S_k$  by the exchange  $s_j$  to  $s_i$  that provided  $J_{i,max}$   
 Compute the complementary subset  $D_k$  of  $S_k$   
 ELSE leave  $S_k$  unchanged  
 CONTINUE

**2.2 Feature Extraction Algorithms**

**2.2.1 Discriminant Analysis Feature Extraction (DAFE):**

This method find a transformation matrix  $A$  such that the class separability of transformed data  $Y=ATX$  is maximized. The fisher criterion is used as the optimization function. In doing so, the within-class and between-class scatter matrices are defined as follows (Fukunaga, 1990):

$$S_w = \sum_{i=1}^N P_i \Sigma_i$$

$$S_b = \sum_{i=1}^N P_i (\mathbf{m}_i - \mathbf{m}_o)(\mathbf{m}_i - \mathbf{m}_o)^T = \sum_{i=1}^{N-1} \sum_{j=i+1}^N P_i P_j (\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T \quad (5)$$

$$\mathbf{m}_o = \sum_{i=1}^N P_i \mathbf{m}_i$$

where  $N$  is the number of classes,  $\mathbf{m}_0$  represents the expected vector of the mixture distribution and  $P_i$  and  $\mathbf{m}_i$  are the prior probability and mean vector of the class  $i$ , respectively. Upon computation of these matrices, the fisher ration can be calculated. Finally, the optimal features will be toward the eigenvectors of the fisher ratio.

**2.2.2 Nonparametric Weighted Feature Extraction (NWFE):**

The main idea of NWFE is putting different weights on each sample to compute the “weighted means” and defining new nonparametric between-class and within-class scatter matrices to obtain more than  $N-1$  features. In NWFE, the nonparametric between-class scatter matrix for  $L$  classes is defined as:

$$S_b = \sum_{i=1}^N \frac{P_i}{N-1} \sum_{j=1, j \neq i}^N \sum_{k=1}^{N_i} \lambda_k^{(i,j)} (x_k^{(i)} - M_j(x_k^{(i)}))(x_k^{(i)} - M_j(x_k^{(i)}))^T \quad (6)$$

where  $x_k^{(i)}$  refers to the  $k^{\text{th}}$  sample from class  $i$ ,  $N$  is the number of classes,  $N_i$  is training sample size of class  $i$ ,  $p_i$  denotes the prior probability of class  $i$ . The scatter matrix weight  $\lambda_k^{(i,j)}$  is a function of  $x_k^{(i)}$  and  $M_j(x_k^{(i)})$ , and defined as:

$$\lambda_k^{(i,j)} = \frac{\text{dist}(x_k^{(i)}, M_j(x_k^{(i)}))^{-1}}{\sum_{l=1}^{N_i} \text{dist}(x_k^{(i)}, M_j(x_l^{(i)}))^{-1}} \quad (7)$$

where  $\text{dist}(a,b)$  denotes the Euclidean distance from a to b.  $M_j(x_k^{(i)})$  denotes the weighted mean  $x_k^{(i)}$  in class j and defined as:

$$M_j(x_k^{(i)}) = \sum_{l=1}^{N_i} W_l^{(i,j)} x_l^{(i)} x_l^{(j)} \quad (8)$$

Where

$$W_l^{(i,j)} = \frac{\text{dist}(x_k^{(i)}, x_l^{(j)})^{-1}}{\sum_{l=1}^{N_i} \text{dist}(x_k^{(i)}, x_l^{(j)})^{-1}} \quad (9)$$

The nonparametric within-class scatter matrix is defined as:

$$S_w = \sum_{i=1}^N p_i \sum_{k=1}^{N_i} \lambda_k^{(i,j)} (x_k^{(i)} - M_i(x_k^{(i)}))(x_k^{(i)} - M_i(x_k^{(i)}))^T \quad (10)$$

Similarly, the optimal features will be toward the eigenvectors of the fisher ratio.

### 2.3 Multiple Classifiers

Most of the feature selection algorithms mentioned in the literature seek only one set of features that distinguish among all the classes simultaneously. This, in one hand, can increase the complexity of the decision boundary between classes. On the other hand, considering one set of features for all the classes requires a large number of features. As stated in section 1, in the literature there are several approaches make use of the idea of multiple classifiers for the classification of hyperspectral data.

In this study we examined a new method to create multiple classifiers. The main idea of the method is that from the huge number of spectral bands in hyperspectral data there are some bands which can discriminate each class better than the others. In order to find the best features for each of the classes we applied a feature selection process.

The method is explained as follows: first of all, the feature selection process is applied for the first class; hence, the most appropriate features for discriminating the first class from the others is selected. Next, the most discriminative features for the second class are selected by using the same procedure for the second class. This process is repeated until all the feature subsets for all classes are selected.

Finally a combination schema is used to combine the outputs of the individual classifiers. Since the outputs of classifiers, here, is a list of probabilities for each class, measurement level methods can be used to combine the classifiers outputs (Kittler

et al., 1998). The most commonly used measurement level methods are mean and product combination rules which perform the same classification in most cases. In the case of independent feature spaces, however, the product combination rule outperforms the mean rule (Tax et al., 2000) and hence it was applied as the combination method in this study. According to product combination rule the pixel  $x$  is assigned to the class  $c_i$  if:

$$\prod_{j=1}^M p(x_j / c_i) = \max_{k=1}^N \left[ \prod_{j=1}^M p(x_j / c_k) \right] \quad (12)$$

in which  $N$  is the number of classes and  $M$  is the number of classifiers. In our case  $N = M$ .

## 3. EXPERIMENTS AND RESULTS

### 3.1 Dataset Description

The laboratorial hyperspectral image was taken by a ground-based hyperspectral camera (SOC700). This hyperspectral camera acquires data in 120 spectral bands from 400nm to 900nm with 12bit radiometric resolution.

We made a ground truth consists of 22 surface materials (classes).

In the selection of different targets we considered the challenging points of different algorithms i.e. classifying targets with very similar spectral characteristics, targets with different shapes, targets with high different spectral characteristics and targets with high spatial variability (table 1).

Class number	Material
1~6	Similar green colored paper from same producer
7~8, (2)	Green colored paper from different producer
9	Clay (dry soil)
10	Sand paper (black)
11~16	Various colored paper from same producer
17	Broad green leaf
18	Fabric (green)
19	Gypsum
20	Wood board
21	Plastic (green)
22	Black colored paper

Table 1. 22 materials (classes) and criteria of pseudo ground truth target.

Most of the classes in this laboratorial image are quite homogeneous. However, in real cases the targets have a fair amount of spatial variability. To mitigate this problem, a Gaussian noise model  $N(0, \sigma)$  was employed to add an appropriate level of noise to the generated image. The standard deviation  $\sigma$  is obtained by

$$\sigma = \beta \frac{Max - Min}{\rho} \tag{14}$$

in which  $\rho$  and  $\beta$  are between class and between band variability coefficients respectively. Max and Min are the maximum and minimum gray level value in each class respectively. Using this formulation, it is guaranteed that, on one hand, the classes have different levels of variability in each band. On the other hand, the level of variability is varied among different bands. You can see the ground targets in figure 2(a) Figure 2.a and 2.b show band 110 of the data before and after adding variability.

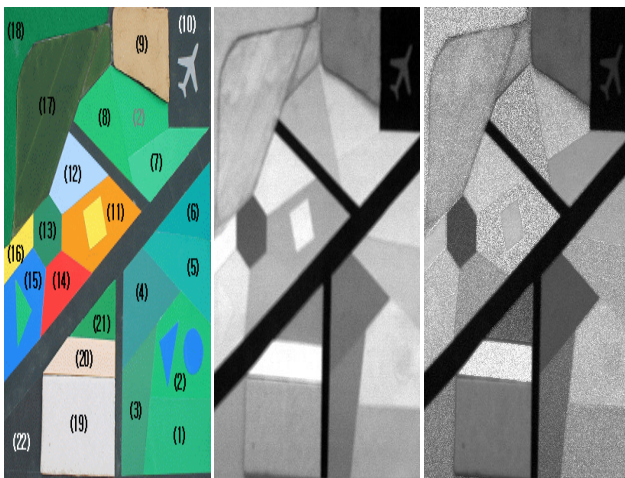


Figure 2. Pseudo target (left), band 118 of original image (middle), and band 118 after adding variability (right).

The training and test samples were selected by stratified random sampling method. The number of selected samples is proportional to the area of each class. The larger the area of each class the higher the number of the samples. Table 2 shows number of training and test sample for each class.

Class	Training	Test	Class	Training	Test
C01	32	1239	C12	33	670
C02	38	1936	C13	33	423
C03	33	899	C14	33	577
C04	33	1013	C15	32	1000
C05	33	861	C16	31	414
C06	32	542	C17	63	2967
C07	31	827	C18	32	940
C08	33	842	C19	38	1926

Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	overall
NWFE	53.8	68.4	95.1	95.6	80.6	95.9	94.2	74.2	97.8	96.0	96.0	94.8	87.5	95.8	91.5	92.8	90.1	72.3	97.8	94.3	91.2	87.1	86.9
DAFE	43.7	62.7	93.9	95.6	78.0	96.3	94.8	63.2	98.4	96.6	86.1	94.6	90.3	95.7	92.4	93.0	82.4	74.9	97.7	94.2	89.9	78.7	85.4
FCS	68.9	68.0	84.0	92.2	71.8	90.2	87.7	78.7	95.6	91.5	95.8	93.6	75.4	93.4	86.6	86.7	84.9	81.6	95.1	94.3	77.0	82.8	85.6
GFS	48.3	52.6	83.1	93.0	77.4	87.3	84.5	74.6	95.1	89.2	95.2	92.7	71.4	82.3	83.8	86.2	83.7	73.2	95.8	94.8	74.2	83.8	82.0
CBFS	68.5	70.4	89.4	93.0	73.8	91.7	89.7	80.6	96.0	91.8	96.7	93.1	77.9	93.4	91.3	87.7	88.3	85.8	95.3	94.5	91.0	81.9	87.4

Table 3. Maximum likelihood classification accuracy (%) with 8 extracted and selected features for each class.

Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	overall
NWFE	62.5	67.4	95.9	95.5	86.6	95.6	94.4	72.1	97.8	95.7	95.8	95.1	87.0	96.2	93.2	92.8	78.8	85.6	97.7	94.3	90.7	89.4	87.9
DAFE	51.3	63.1	93.1	95.2	86.2	96.3	95.0	66.6	98.2	96.4	86.0	94.6	90.3	96.4	92.5	92.5	82.6	80.2	97.7	92.5	90.1	79.3	86.3

C09	33	1166	C20	32	599
C10	86	3666	C21	34	365
C11	32	1025	C22	30	668
Total training sample : 807			Total test sample : 24564		

Table 2. List of classes training and testing sample sizes used in the experiments.

### 3.2 Experimental Results

Experiments are carried out to assess the performance of the two feature selection algorithms i.e. GFS and FCS feature selector, two feature extraction algorithms i.e. DAFE and NWFE and our class-based feature selection methodology using the Maximum likelihood classifier.

#### 3.2.1 First experiment: finding the best number of features:

As the first experiment, the two feature selection techniques as well as the two feature extraction methods were applied to the dataset. The difference in algorithms performance, in terms of classification accuracy, as a function of the number of features used, is visualized in figure 3.

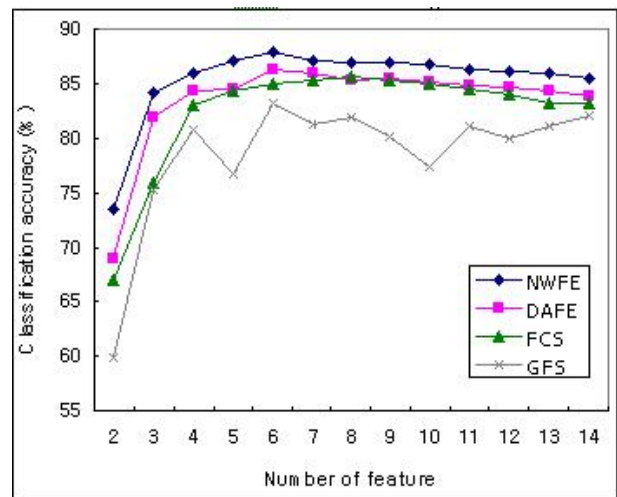


Figure 3. Comparison of classification accuracy of two feature selection method and two feature extraction methods using different number of features

Figure 3 presents the comparison of classification accuracies of two feature selection and two feature extraction methods respectively when using different number of features (from 2 to 14). The total accuracies first improve as more features are

FCS	70.6	68.3	82.8	91.5	71.2	90.2	84.8	78.7	95.1	89.9	89.3	93.4	74.0	89.6	87.7	82.9	85.7	80.7	94.5	94.5	80.0	85.0	85.0
GFS	62.8	68.4	80.3	82.0	72.8	88.4	83.2	75.7	94.2	90.0	90.2	93.3	72.6	87.9	87.2	81.6	79.2	79.5	94.5	94.5	78.4	79.0	83.3
CBFS	69.70	65.84	90.24	91.91	81.09	91.16	86.78	80.97	95.64	90.42	96.79	93.43	76.00	91.18	90.72	85.75	85.73	87.49	94.82	94.32	84.93	83.08	86.5

Table 4. Maximum likelihood classification accuracy (%) with 6 extracted and selected features for each class.

added into feature subsets but it then deteriorates (Hughes phenomenon). The highest accuracies are obtained when using 8 features in FCS, 6 features in genetic algorithm and 6 features in DAFE and NWFE.

**3.2.2 Second experiment: classification accuracy comparison among the considered methods:** upon finding the best number of features for each feature reduction method, the Maximum likelihood classifier is employed to the best selected features in each method. Here, we used 6 and 8 features in classification accuracy comparisons. Table 3 and 4 show the classification accuracy for each class in different methods using 8 and 6 features respectively.

As can be inferred from figure 3 and also table 3 and 4 the feature extraction methods provide better accuracy than feature selection methods. This can arise from the influence of adding noise to the original image for the purpose of variability. The feature extraction methods try to get rid of this noise by transforming the original bands to the new transformed space whilst the selection methods do not tackle with this added noise and they only select the good bands from the available spectral bands.

In feature extraction methods, the NWFE performed better than the DAFE. It is an effective method both in terms of accuracy and time complexity.

In the feature selection methods the FCS method performed better than genetic feature selector. On the other hand, the time complexity of the genetic method is also higher than FCS method. As can be seen from figure 3 the classification results provided by genetic method have fluctuating behaviour. This can be interpreted as a consequence of the random nature of this method in forming the population as well as the production of offspring in each generation. The FCS method is a straightforward method for feature selection. It is both accurate and fairly fast and thus was selected as the basis in our proposed class-based method for feature selection. The results provided by our proposed CBFS method are better than other feature selection methods. As can be seen from table 4 our proposed method provides more than 5 percent higher accuracy than genetic method and about 2 percent better accuracy than FCS method using 8 features. Using 6 features (table 5), the CBFS method has again provided better accuracies than other feature selection methods. The results of the CBFS method is quite comparable with feature extraction method. Comparison of accuracies in each class in table 4 and 5 demonstrates that the CBFS method provides better accuracy in almost all the classes than other methods. This can be interpreted as a consequence of local nature of the CBFS method. Different regions in the image are classified with different sets of features and this can increase the accuracy in almost all the classes, though at the cost of an increase in execution time.

**4. CONCLUSIONS**

In the present paper we made an evaluation of some of the feature reduction algorithms based on a laboratorial

hyperspectral data. A ground truth data based on the challenging points of different methodologies was generated. At the first experiment the best number of features for each feature reduction method was found. The highest accuracies are obtained when using 8 features in FCS, 6 features in genetic algorithm and 6 features in DAFE and NWFE. In the second experiment we made a class by class comparison by applying the Maximum Likelihood to the best number of features obtained in the first experiment.

The feature extraction methods, first, try to get rid of the noise, added to the original image as variability, by transforming the original bands to a new transformed space and hence outperformed the feature selection methods. In feature extraction algorithms the NWFE method performed better than DAFE. In feature selection methods the FCS method was proved to be an effective feature selection method and hence was employed as the basis for our proposed class-based feature selection. The results of the proposed CBFS method was superior to other feature selection methods i.e. the genetic algorithm and FCS method. The results were also quite comparable to feature extraction methods.

**ACKNOWLEDGEMENT**

This research was supported by the Defence Acquisition Program Administration and Agency for Defence Development, Korea, through the Image Information Research Center at Korea Advanced Institute of Science & Technology under the contract UD070007AD.

**REFERENCES**

Bajcsy, P. and Groves, P., 2004. Methodology for hyperspectral band selection. *Photogrammetric Engineering & Remote Sensing*, 70(7), pp. 793–802.

Benediktsson, J. A., and Kanellopoulos, I., 1999. Classification of multisource and hyperspectral data based on decision fusion. *IEEE Transaction on Geoscience and Remote Sensing*, 37, pp. 1367–1377.

Duin, R. P. W. and Haeb-Umbach, R., 2001. Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 23, pp. 762-766.

Ferri, F., Pudil, P., Hatef, M., and Kittler, J., 1994. *Comparative Study of Techniques for Large Scale Feature Selection. Pattern Recognition in Practice IV*, Elsevier Science B.V., pp. 403-413.

Fukunaga, K., 1990. *Introduction to Statistical Pattern Recognition*. Academic Press, San Diego.

Goldberg, D.E. and Deb, K., 1991. *A comparative analysis of selection schemes used in genetic algorithms. Foundations of Genetic Algorithms*, Morgan Kaufmann, pp. 69-93.

Hughes, G.F., 1968. On the SUM accuracy of statistical pattern recognizers. *IEEE Transaction on Information Theory*, pp.55-63.

Jain, A., Zongker, D., 1997. Feature selection: evaluation, application and small sample performance. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19, pp. 153-158.  
Kittler, J., 1986. *Feature selection and extraction. Handbook of Pattern Recognition and Image Processing*. Academic Press, pp. 60-81.

Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J., 1998. On combining classifiers. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 20, pp. 226–239.

Kumar, S., Ghosh, J., and Crawford, M. M., 2001. Best basis feature extraction algorithms for classification of hyperspectral data. *IEEE Transaction on Geoscience and Remote Sensing*, 29(7), pp. 1368–1379.

Kuo, B-C. and Landgrebe, D. A., 2004. Nonparametric Weighted Feature Extraction for Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 42(5), pp 1096-1105.

Pudil, P., Novovicova, J., and Kittler, J., 1994. Floating search methods in feature selection. *Pattern Recognition Letters*, 15, pp. 1119–1125.

Serpico, S. B. and Bruzzone, L., 2000. A new search algorithm for feature selection in hyperspectral remote sensing images. *IEEE. Transaction on Geoscience and Remote Sensing, Special Issue on Analysis of Hyperspectral Image Data*, 39(7), pp. 1360-1367.

Serpico, S. B., D’Inca, M., Melgani, F., and Moser, G., 2003. Comparison of feature reduction techniques for classification of hyperspectral remote sensing data. In: *Proceedings of SPIE—Image and Signal Processing for Remote Sensing VIII*, Vol. 4885, pp. 347–358.

Sheffer, D. and Ultchin, Y., 2003. Comparison of band selection results using different class separation measures in various day and night conditions. In: *Proceedings of SPIE Conference Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery IX*, Vol. 5093, pp. 452-461.

Siedlecki, W. and Sklansky, J, 1989. A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters*, 10, pp. 335–347.

Tax, D., Van Breukelen, M., Duin, R. P. W., and Kittler, J., 2000. Combining multiple classifiers by averaging or by multiplying?. *Pattern Recognition.*, 33, pp. 1475–1485.

