Feature Selection for Classification of Hyperspectral Remotely Sensed data using NSGA-II

Mukesh Kumar

Graduate Research Assistant
Dept. of Civil Engg.
Penn State University
State College, PA 16802

Abstract

This paper summarizes the implementation and performance of Nondominated Sorting Genetic algorithm (NSGA-II) [2] for feature selection of remotely sensed hyperspectral imagery. Two step processes have been followed. In first step, a feature subset is selected with optimum spectral and texture information content resulting in a smaller space to be searched in the next step. In the second step, a single objective search algorithm is used to obtain a final smaller subset out of the features already selected (with optimum information content), which have best separability between the classes. Classes are obtained by classifying the subset bands using maximum likelihood classification algorithm. Method of spectral and textural information evaluation of images, genotypic representation of our algorithm, classification methodology and separability criteria between classes have been discussed. Also discussed is the reason for choice of NSGA-II and a strategy to extract optimum results from it.

1 INTRODUCTION

Processing of images obtained from Hyperspectral satellite sensors like Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) which generates large amount of data require demanding computational resources. These remotely sensed data are widely used for thematic map generation. Thematic maps are produced through the process of digital image classification. The cost and complexity of classification depends on the number of features (N_b, image bands at different frequencies) and band size, while classification accuracy depends on factors like type of classifier used, noise in the bands and the information carried by each band. With increasing number of bands, cost of classification increases exponentially though accuracy saturates after increase to a certain number of bands. So, one have to do feature subset selection which attempts to select a minimally sized subset of bands, $L_b<N_b$, such that separability between classes is optimized over that subset [8].

If the original feature subset contains N_b number of bands, then the number of competing subsets will be 2^{N_b}. This is too large a number even for medium-sized N_b. For state-of-the art hyperspectral sensors like AVIRIS with 220 bands, doing an exhaustive search is computationally infeasible. Among many, searching strategy like Best-first search have been used for searching for an accurate subset [5]. Pei *et al.* [9] used Simple Genetic Algorithm (SGA)/ (K- nearest neighbor) KNN hybrid approach for feature selection. Clustering using KNN algorithm for large number of large subsets is computationally intensive. Algorithm suggested in this paper, first derives a smaller subset $M_b(<N_b)$ with optimum information content and less noise before application of any clustering or classification method.

2 THE APPROACH

In remote sensing, both textural and spectral information should be used in order to improve the accuracy of classification [6]. Texture information content is quantified by entropy, which is a measure of degree of disorder or heterogeneity in an image. Entropy image for all bands is obtained. Principal Component Transform (PCT) of both spectral and entropy bands is done separately. This produces uncorrelated components that explain maximum amount of variance possible by linear transformation of spectral and texture bands respectively. Absolute value of correlation coefficient(called factor loadings) of band i with principal component(PC) j tells how important that band i is to component j. Associated information(AI) carried by each band is calculated. Only those PC's are used in AI calculation which cumulatively carries almost all information (~99.5 %). Doing this, components carrying very less information(supposed to be noise) are discarded. The objective is to optimize cumulative texture and spectral associated

information(CTAI and CSAI) of a band combination while using minimum number of bands as possible. A tradeoff curve between these conflicting objectives are obtained by NSGA-II. An optimum feature subset is selected by analyzing the trade off curve. Here selection entirely depends on the preferences of the decision maker for various objectives. Optimum subset selected, basically carries most of the information of the entire set. So only this subset can be searched (instead of entire set) to look for a collection of bands that will yield high classification accuracy upon classification as other bands not included in this subset are redundant and don't contribute in classification. Out of this optimum subset, our goal is to identify a smaller subset of bands with improved class separability between classes and hence raised classification accuracy. Separability is quantified by Brightness Value Overlap Index (BVOI) which measures the degree of overlap among classes [7]. Smaller the BVOI, better is the classification. Classes(various land cover types) are obtained by Gaussian Maximum Likelihood Classifier (GMLC). Figure 1 shows the schematic chart of the entire process.

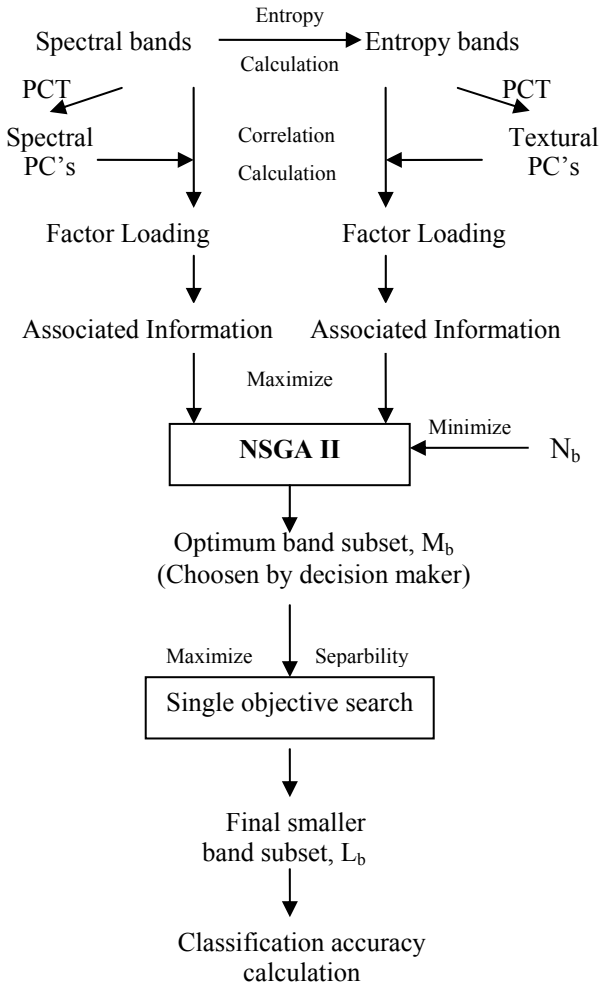

Figure 1

Classification accuracy of the final subset band selected is evaluated by Khat index [1], also called kappa index of agreement. Method to calculate various functions and algorithms have been discussed in the following subsection.

2.1 ENTROPY

Texture measures the spatial distribution of pixel value variations with in a band. One method to asses texture is by calculation of entropy by Grey Level Difference Histogram (GLDH) Method.

2.1.1 GLDH

Let $g(x, y)$ represents a image band in spatial domain. For any displacement function $\delta = (\Delta x, \Delta y)$, where Δx and Δy are integers values indicating the amount of displacement in x and y directions respectively. The difference function $g_\delta(x, y)$ is defined as

$$g_\delta(x, y) = \left| g(x, y) - g(x + \Delta x, y + \Delta y) \right|$$

Probability density function is defined as

$$p(i/\delta) = p(g_\delta(x, y) = i)$$

where, i is the gray value, which ranges from 1, 2, 3, .., N_g where N_g is the maximum quantization value(in the problem under consideration, it is 255). Using density function, entropy can be calculated as

$$entropy = \sum_{i=0}^{N_g - 1} p(i/\delta) \log p(i/\delta)$$

2.2 PCT

This transform produces a new uncorrelated vector space in which data has most variance along its first axis, the next largest variance along a second mutually perpendicular axis and so on. Principal Components are calculated in two steps [4]. Firstly $n \times n$ covariance matrix from the n bands is derived. In second step , eigen values and eigen vectors of the covariance matrix which are related as

$$\Sigma_Y = \Phi^T \Sigma_X \Phi = \begin{vmatrix} \lambda_1 0 \ldots\ldots\ldots\ldots\ldots\ldots 0.. \\ \ldots\ldots \lambda_2 \\ \ldots\ldots\ldots\ldots \lambda_3 \\ . \\ .0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots \lambda_n \end{vmatrix}$$

is computed. Here Σ_X is covariance matrix of original image bands, Σ_Y is uncorrelated covariance matrix of uncorrelated bands, Φ is eigen vector matrix and λ_i's are eigenvalues such that $\lambda_i > \lambda_j$ for i > j. Eigen values are axes of the vector space and variances of the PC's their length. The percent of total variance carried by the each principal component can be calculated using the formula

$$\text{var}_i = \frac{\lambda_i}{\sum_{k=1}^{n} \lambda_k} \times 100$$

Correlation of a band with principal component is given by

$$\rho_{ij} = \frac{e_{ij}\sqrt{\lambda_j}}{\sqrt{\text{var}_i}}$$

where ρ_{ij} is correlation for band i and component j, e_{ij} is eigenvector for band i and principal component j. Associated information carried by a particular band is defined by

$$AI_i = \sum_{j=1}^{K} \rho_{ij} * \text{var}_i$$

where K is number of principal componets that carries sufficient information. Cumulative spectral/textural associated information is calculated by

$$CSAI = \sum_{i=1}^{r} Spectral\_AI_i$$

$$CTAI = \sum_{i=1}^{r} Textural\_AI_i$$

where r is the number of bands being considered.

2.3 NSGA II

Selecting an optimum subset of features, M_b with sufficient spectral and textural information is a multi-objective optimization problem. The three objectives considered are maximization of spectral and textural information while minimizing number of bands. (Pareto Optimal) Solutions to this problem can be expressed in terms of their superiority to the rest of solutions in search space when all objectives are considered, but they may be inferior to other solution in atleast one objective. This concept is also called Pareto dominance. In absence of any other information like degree of preference of one objective over other, one solution cannot be said better than other. So one needs as many optimal solutions as possible to help in final decision making. These solutions can be obtained by NSGA II. NSGA II also outperforms other algorithms for multiobjective optimization with its lower computational complexity and elitism property [14] and [3]. For reducing the computational effort and automating the parameter specification process in NSGA II, simplifications were done based on [10] and [11]. Chromosome length was taken equal to the number of bands used in the experiment. In the first run of NSGA-II, a small population size is used. Population size is doubled with each successive run to evolve nondominated solutions. Increase of population with each step is stopped when the threshold percentage change in number of nondominated individuals for two successive run specified by Δ_{ND} is reached. Δ_{ND} is given by

$$\Delta_{ND} < 100 \frac{\left| NDI_n - NDI_{n-1} \right|}{NDI_{n-1}}$$

where NDI_n and NDI_{n-1} are the number of non dominated individuals in run n and n-1 respectively. Run length is estimated to be twice the binary string length [10]. This estimate assumes that NSGA-II converges as fast as the system undergoing pure binary selection because of additional selection pressure due to elitism. Crossover probability P_c is derived from disruption boundary relationship [13] which is

$$P_c \le \frac{s-1}{s}$$

where s is the number of individuals participating in tournament selection. In NSGA II, child is selected by binary tournament selection, so s=2. Probability of mutation, P_m is taken as inverse of population [12].

2.4 SINGLE OBJECTIVE SEARCH

Finding a feature subset, L_b so that the classes have maximum separability between them is a single objective optimization problem. SGA was used to find the final optimum subset. To obtain the best parameter (viz. population size and number of generations) setting, following steps were followed:

- Fitness values corresponding to different population sizes are obtained at some large number of generations. This is based on the assumption that atleast at that population size (P), the solution has converged more than at other population sizes.
- At population P, SGA is run for number of generations till the solution has converged.

Probability of mutation, P_m is taken as inverse of population. Binary tournament selection was used.

2.5 BVOI

In this method, the range of pixel values within a class is compared, with histogram for all classes with in the bands used for classification. Accumulated percentage of all pixel values having pixel values ranging from minimum to maximum for each class is determined by

$$F_{m,n} = \frac{\sum_{i=Min_{m,n}}^{Max_{m,n}} f(X_{i,m})_n}{\sum_{n-1}^{N} \left(\sum_{i=Min_{m,n}}^{Max_{m,n}} f(X_{i,m})_n \right)}$$

where $F_{m,n}$ is accumulated frequency of class n in band m, $f(X_{i,m})_n$ is frequency of pixel value X_i within a class *n* of band m, $Min_{m,n}$ and $Max_{m,n}$ are minimum and maximum pixel value in class n and band m respectively. BVOI is calculated by

$$BVOI = \frac{F_{k,m}}{F_k}$$

where

$$F_{km} = \sum_{n=1}^{N} F_{m,n}$$
$$F_k = \sum_{n=1}^{N} (\frac{1}{M} \sum_{m=1}^{M} F_{m,n})$$

Band subset with least BVOI value denotes maximum separability between classes.

2.6 CLASSIFICATION

A classified image is needed before separability between classes can be evaluated. Classes have been obtained by supervised classification of the band subset. In supervised classification, user attempts to locate the specific sites in the remotely sensed data that represent homogeneous areas of known features (land cover types). These areas are commonly represented as training sites because the spectral characteristics of these known sites are used to train the classier for mapping the entire image. Maximum Likelihood Classification is the most common supervised classification method used with remote sensing image data.

2.6.1 GMLC

The GMLC method assumes that specified values of training samples are statistically distributed according to a multivariate normal probability density function. Though in real world nothing may be normally distributed, in practice, however it is found that the assumption of normality holds reasonably well. The probability that a pixel x_k, taken over n bands, being allocated to class i is $p(X_k|i)$.

$$p(X_k|i) = \frac{1}{(2\pi)^{n/2}|V_i|^{\frac{1}{2}}} \exp[-\frac{1}{2}(X_k - U_i)^T V_i^{-1} (X_k - U_i)]$$

where

$$X_k = \begin{bmatrix} x_{k1} \\ x_{k2} \\ . \\ . \\ x_{kn} \end{bmatrix} \qquad U_i = \begin{bmatrix} u_{i1} \\ u_{i2} \\ . \\ . \\ u_{in} \end{bmatrix}$$

and

$$V_i = \begin{bmatrix} v_{i11} & v_{i12} & . & . & . & v_{i1n} \\ v_{i21} & v_{i22} & . & . & . & v_{i2n} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & v_{ihm} & . & . \\ v_{in1} & v_{in2} & . & . & . & v_{inn} \end{bmatrix}$$

Here u_{im} and v_{ihm} are

$$u_{im} = \frac{1}{q_i} \sum_{k=1}^{q_i} x_{km} \qquad \text{m=1,2,……n}$$

$$v_{ihm} = \frac{1}{q_i - 1} \sum_{k=1}^{q_i} (x_{kh} - u_{ih})(x_{km} - u_{im}) \quad \text{h=1,2….n}$$

m=1,2….n

q_i is total number of pixels in class i.

2.6.2 Accuracy analysis by Khat index

Classified image and corresponding training pixels on the ground is used to form a confusion matrix. A confusion matrix is a square array of size $m \times m$ (m is the number of classes) in which numbers are laid out in rows and columns that give the number of sample pixels assigned to a particular class relative to the actual class as verified on the ground. Diagonal elements represent observations that agree both on training and classified images and non-diagonal elements represent that do not agree. Khat index k, uses all the elements in the confusion matrix and it is a proportion of agreement after chance agreement is removed from consideration.

$$k = \frac{(P_o - P_c)}{(1 - P_c)}$$

where k is kappa index , P_o is overall accuracy given by

$$P_o = \frac{\sum_{i=1}^{m} x_{ii}}{\sum_{i=1}^{m} \sum_{j=1}^{m} x_{ij}}$$

and P_c is

$$P_c = \frac{(x_{i+} x_{+i} + ...x_{+m} x_{m+}}{N^2}$$

where $x_{i+}.....,x_{m+}$ are row total, $x_{+i}.....,x_{+m}$ are column total of confusion matrix. N is

$$N = \sum_{i=1}^{m} \sum_{j=1}^{m} x_{ij}$$

For individual classes, the Khat index is calculated using the formula

$$k_i = \frac{Nx_{ii} - x_{+i}x_{i+}}{Nx_{i+} - x_{i+}x_{+i}}$$

3 IMPLEMENTATION

Algorithm is implemented on a set of 22 bands from Jun2 1992 AVIRIS data set[1] of a mixed agriculture/forestry landscape in the Indian Pine Test Site in Northwestern Indiana. It contains 145 rows by 145 columns of pixels. The 22 bands were band no. 1+10*i, i=0,1,2….21 of the original data set. Figure 2 shows the false color composite (FCC) of the area of interest (AOI) prepared using the band numbers 141, 71 and 31 respectively. It also shows the training pixels chosen for various classes using MultiSpec software. In figure 2, training pixels are the one, inside various rectangles (in white), for respective classes.

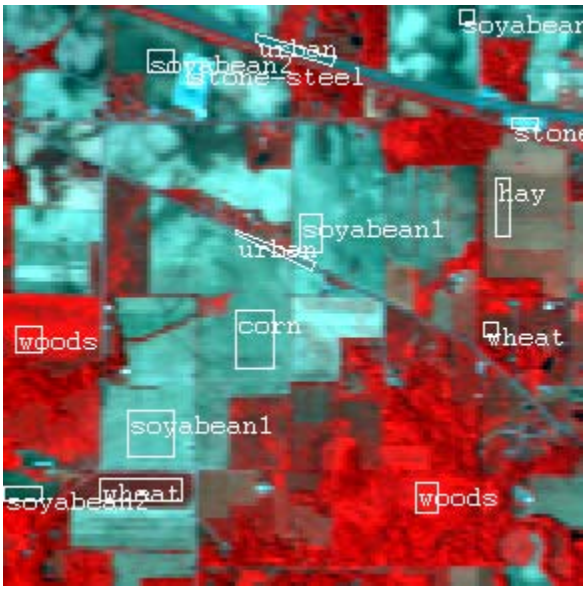

Figure 2

PC's of the 22 spectral/texture bands are calculated. 99.5% of spectral information was carried by first 9 PC's and 99.3%of texture information was carried by first 15 PC's out of 22. Associated spectral/texture information of each band with PC's (cumulatively having 99.5% information) is calculated. Optimum band subset is to be obtained from 2^{22} band combinations. A binary coded chromosome of length 22 was constructed. If the bit was one, then that feature was used in calculation of cumulative spectral/textural information content. Using NSGA II, a tradeoff curve between maximum spectral and texture information, and number of features considered is obtained. To help in decision making, variables on the three axes were expressed as band fraction (i.e. ratio of number of bands used in evaluation to maximum number of bands which is 22), ratio of CTAI to maximum CTAI possible (i.e. when all the bands are considered) and ratio of CSAI to maximum CSAI. Decision maker can select the number of feature according to his/her preferences about various objectives. The tradeoff curve is shown below

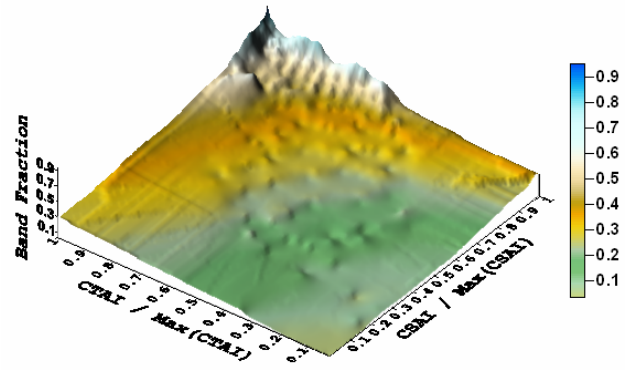

Figure 3

For further experimentation in this problem, the author as a decision maker, choose 10 bands with cumulative 85% and 93% of cumulative texture and spectral AI with respect to the maximum cumulative texture/spectral AI of any non dominated band combinations. Here decision was made based on the reservations of author to have ratio of CTAI and CSAI to their maximum possible, more than a particular threshold (85%, one can set the threshold taking in consideration the dominance of textural/spectral property of the bands). Projecting the non-dominated optimal points along band fraction-CSAI/Max (CSAI) plane, as can be seen in figure 4, the decrease in CSAI fraction (on x axis) is very less compared to decrease in band fraction (on y-axis) up to band fraction=0.4545.

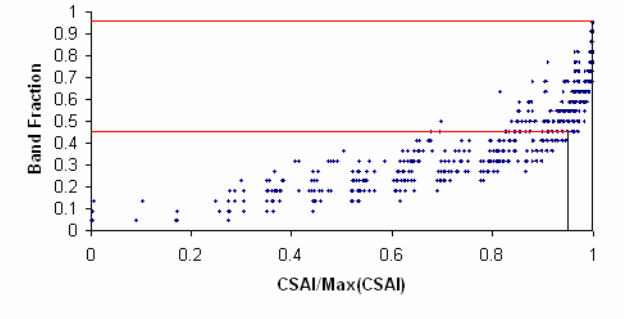

Figure 4

[1] AVRIS have 210 bands in all for each scene. Only 22 were used due to their availability but these are sufficient to demonstrate an analysis of the method suggested.

Similarly drawing 2-dimensional plots for the other objective i.e. the case when non-dominated pareto optimal points are projected in plane band fraction-CTAI/Max (CTAI), decrease on x-axis is 15% with 55% decrease on y-axis (shown in figure 5). In figure 4 and 5, the points plotted are non-dominated points. Some of them seem to be dominated because they have been projected on a 2-dimensional plane. In the above experimentation, Δ was set to 10%, implying that close approximation of true

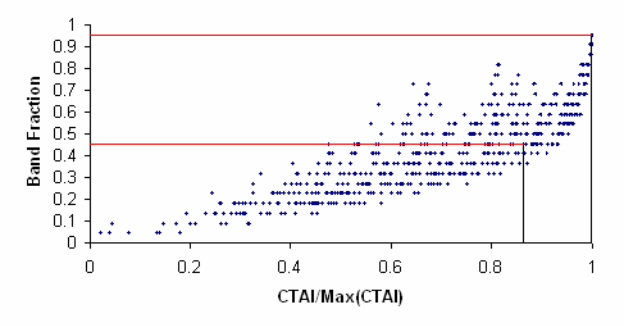

Figure 5

pareto front is being sought. Δ decreased with each successive run as shown in figure 6. After 6 runs, $\Delta \approx 8$ %. Population size and number of generations for this run was 640 and 50 respectively. Total of 563 non dominated unique subset combinations were obtained which were used to draw the tradeoff curve. After having selected the 10 bands, now the aim is to find a smaller subset out of this with good separability. This is a single objective search problem. Search space is only 2^{10}. SGA have been used to derive the final optimum band subset that has largest separability between classes. This step seems trivial because the number of combinations to be searched here is very small and an exhaustive search could have been done. But if the experiment was started with all the bands of AVIRIS(i.e. 220) instead of 22, then decision maker would have to settle for a larger subset with optimum information content after NSGA II step, then usage of SGA in this step will be of utmost importance. Chromosome of 10 gene length was constructed. Again, if the bit was one, then the feature was used in classification.

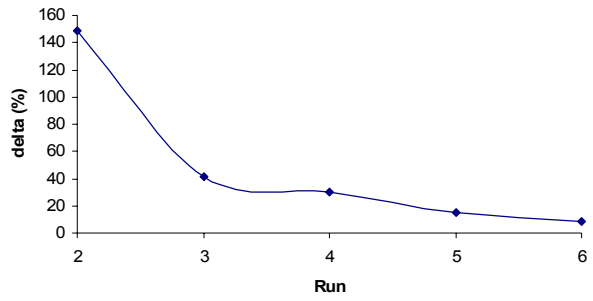

Figure 6

For finding the best parameter setting, steps discussed in section 2.4 were followed. For number of generations = 60, fitness value(BVOI) is obtained at various population sizes as shown below

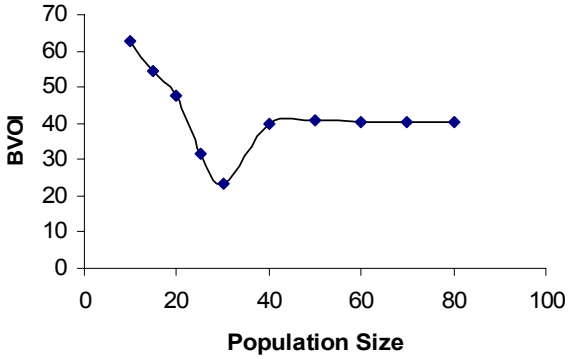

Figure 5

As is seen in figure 5, at population size=30 BVOI have least magnitude. At this population SGA was run till the solution converged i.e. till generation=100, as shown in figure 6. This is due to the fact that as number of generations increase, fitter individuals take over the population resulting.

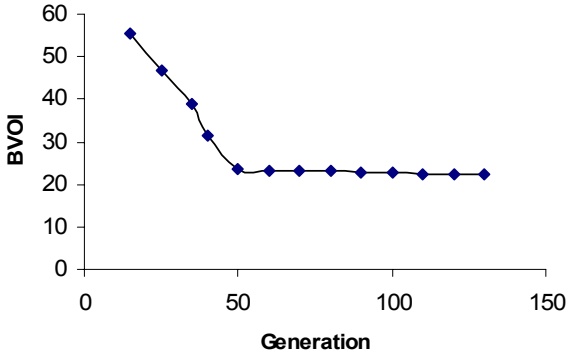

Figure 6

Highest separability was found for subset with band 11, 21 and 81.

4 RESULTS

The following classified image was obtained using the selected bands.

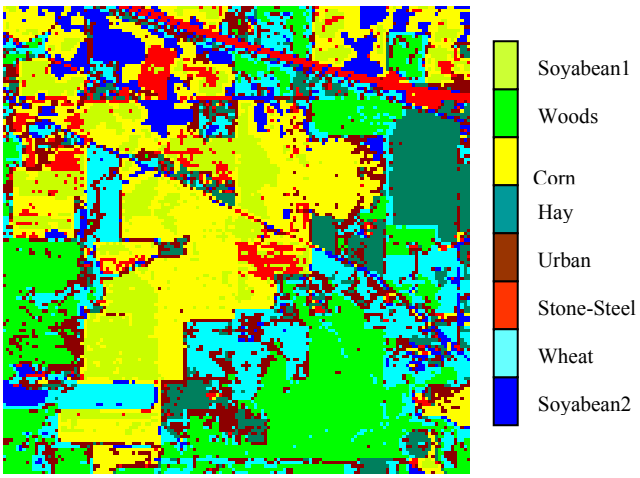

Figure 7

The image got classified into eight prominent classes viz. Soyabean1 (S1), Woods (WD), Corn (C), Hay (H), Urban (U), Stone-Steel (S), Wheat (W) and Soyabean2 (S2). Class Soyabean1 differed from Soyabean2 in the way tillage practice was employed in the two land covers, since the amount of residue from previous vegetation varies. From the classified image, it is seen that AOI appears to be about 2/3rd agriculture and 1/3rd woods. Two linear patterns are seen running from southeast to northwest direction. These are highway (U.S. 52 & U.S. 231) and a major secondary road (Jackson highway). Confusion matrix between training and classified pixels is listed in table 1.

Table 1: Training class performance confusion matrix

CLASS	NUMBER OF SAMPLES IN CLASS								PA
	S	S1	C	W	H	U	S2	WD	
S	33	0	1	0	0	11	0	0	73.3
S1	1	182	21	0	0	0	0	0	89.2
C	2	19	129	0	0	0	0	0	86.0
W	0	0	0	138	0	2	0	2	97.2
H	0	0	0	0	60	0	0	0	100
U	6	0	5	3	0	60	0	0	81.1
S2	1	0	0	0	0	70	90	0	91.8
WD	0	0	0	0	0	0	0	97	100
UA	76.7	90.5	82.7	97.9	100	75	100	98	

In table 1, UA is user's accuracy or the probability that the pixel of a particular class actually belongs to that class on the ground. PA is producer's accuracy or the accuracy with which the maximum likelihood classifier classified the image. Overall accuracy calculated was 90.7 % and Kappa statistics= 89 %. Overall accuracy (in red line) and respective accuracy of classification of the various classes (as bars) are shown below

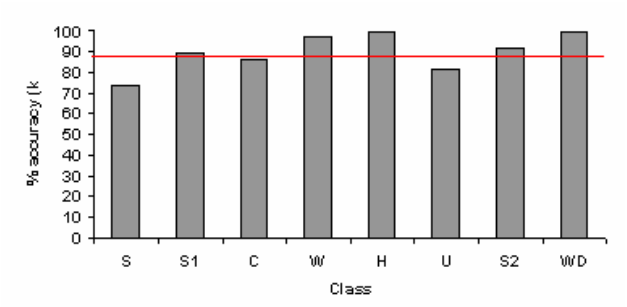

Figure 8

5 CONCLUSIONS

This paper presents experimental results applying NSGA II for feature selection of hyperspectral remotely sensed imagery. Genetic algorithms have earlier been applied to feature selection. However this work presents a new strategy to make the whole process computationally efficient. Also the accuracy of classification obtained was high (~90%). A novel way of breaking down problem into two steps was followed. This enables in getting rid of noise from the bands and also in decreasing the size of the feature subset to be searched before it is used for *feature selection for classification* problem. The paper also discusses the method to optimize the parameters of NSGA II.

Acknowledgments

The author wish to thank Dr. P. Reed, PSU for his suggestions and guidance through out the project. Thanks also to V.Devireddy for numerous fruitful discussions.

References

[1] Bishop, Y. N. M., S. E. Fienberg. and P. W. Holland, 1975, Discrete Multivariate Analysis: Theory and Practise. MIT Press, Cambridge, Massachusetts.

[2] Deb, K., A. Pratap, S. Agarwal and T. Meyarivan 2000 A Fast and Elitist Multi-Objective Genetic Algorithm: NSGA-II.

[3] Deb, K., L. Thiele, M. Laumanns, and E. Zitzler, 2001, Scalable Test Problems for Evolutionary Multi-Objective Optimization, Computer Engineering and Networks Laboratory Report (TIK-112), Department of Electrical Engineering, Swiss Federal Institute of Technology, Zurich, Switzerland.

[4] Jensen, J.R., 1996, Introductory Digital Image Processing, Printice-Hall, Englewood Cliffs, New Jersy.

[5] Kohavi, R., 1995, Wrappers for Performance Enhancement and Oblivious Decision Graphs. PhD thesis, Stanford University.

[6] Lee, J.H. and W.D. Philpot, 1991, Spectral Texture Matching: A classifier for Digital Imargery. IEEE Transaction on Geoscience and Remote Sensing, Vol. 29 No. 4:545-554.

[7] Ma, Z. and C. E. Olson, 1989, A Measurement of Spectral Overlap among Cover Types. *Photogrammetric Engineering and Remote Sensing*, Vol. 55(10), 1441-1444.

[8] Narendra, P.M. and K. Fukunaga, 1977, A branch and bound algorithm for feature selection, IEEE Transactions on Computers, C-26(9):917-922.

[9] Pei, M., E.D. Goodman and W.F. Punch, June 1995, Genetic Algorithms for classification and feature extraction. Annual Meeting, Classification Society of North America.

[10] Reed, P., B. Minsker, and D. E. Goldberg, 2000 Designing a competent simple genetic algorithm for search and optimization. *Water Resources Research*, 36(12): 3757-3761.

[11] Reed, P., 2002 *Striking the balance: Longterm groundwater monitoring design for multiple conflicting objectives*, Doctoral dissertation, University of Illinois at Urbana-Champaigne, 4: 53-70.

[12] Schaffer, J. D., R. A. Caruana, L. J. Eshelman, and R. Das, 1989, A study of control parameters affecting online performance of genetic algorithms for function optimization, In Schaffer, J. D. (Ed.), Proceedings of the Third International Conference on Genetic Algorithms, p. 51-60, Morgan Kaufmann, San Mateo, CA.

[13] Thierens, D., 1995, *Analysis and design of genetic algorithms*, Doctoral dissertation, Katholieke Universiteit Leuven, Leuven, Belgium.

[14] Zitzler, E., K. Deb, and L. Thiele, 2000, Comparison of multiobjective evolutionary algorithms: empirical results. Evolutionary Computation, 8(2), 173-195.