Enhancing the Performance of Feature Selection Algorithms for Classifying Hyperspectral Imagery

Mukesh Kumar[1] Christopher J. Duffy Patrick M. Reed
Dept. of Civil and Environmental Engineering.
The Pennsylvania State University, PA, 16802
[1]muk139@psu.edu

***Abstract*—A method for enhancing the performance of feature selection algorithms is proposed. The proposed method is a two step process - first a feature subset is selected with optimum mutual information content and then this subset is searched to find a smaller subset, which has the best separability between classes. A subset with "optimum" mutual information content is the one which contains most of the information that is present in the rest of set. An expression has been derived to find such a subset efficiently. The two-step process is shown to reduce the search space drastically. The method is implemented with a simple Genetic Algorithm (SGA) and tested using hyperspectral remote-sensing images (acquired by AVIRIS sensor) as a data set. Theoretical result shows that the proposed method reduces the computation load by 90%. A computational efficiency to the order ~20% is obtained on the implementation of proposed method with SGA. The method is sufficiently general to be used to enhance other feature selection algorithms.**

***Keywords*-Hyperspectral remote sensing; feature selection; search algorithms; mutual information.**

I. INTRODUCTION

Hyperspectral remote sensing is very promising for gathering and analyzing information about earth surface. This is due to the dense sampling of the spectral range of the sensor. However, a major limitation on the use of hyperspectral images results from the theoretical complexity and large computational load associated with processing. As hyperspectral sensors acquire images in very narrow spectral bands, the resulting high-dimensional feature sets contain redundant information. Because of this redundancy, the number of features given as input to a classifier can be reduced without significant loss of information [1] by performing feature selection. The problem of feature selection can be easily stated as a search of sufficiently reduced subset of say , M features out of a total N (N > M) available ones, without significantly degrading the performance of the resulting classifier. This search problem is driven by a certain measure of performance or criterion function which is used to asses the validity of each feature subset. In this paper, attention is focused on enhancing the time complexity of the search algorithms. Details about criterion functions are given in [2], [3].

II. APPROACH

Four basic steps in a typical feature selection method [4] are: *generation* procedure (GP) which deals with generating different combinations of feature subsets which are evaluated for their fitness using a *evaluation* procedure (EP). If the fitness of a subset in the present iteration is better than that in the previous iteration, the subset replaces the earlier and is used as an input to *stopping* procedure (SP). The feature selection process terminates when the feature subset inputted to SP satisfies the stopping criterea function. The selected subset from the previous step is then validated in a *validation* procedure (VP). In the proposed methodology, these four steps are invoked twice. In the feature selection step 1 (FSS1), S (S<N) out of the total N bands (features) are selected. In the feature selection step 2 (FSS2) S bands are searched to obtain a final subset of M (M < S <N) bands as output. Details about the various steps in FSS1 and FSS2 are tabulated in Table 1.

TABLE I. PROCEDURAL STEPS FOR FEATURE SELECTION

Steps	Feature Selection Procedure	
	FSS1	***FSS2***
GP	SGA[a]	SGA[a]
EP	Larger Redundancy	Larger class-separability
SP	$\frac{\text{InformationIn}:S}{\text{InformationIn}:N}$ >Threshold	Maximum class-separability
VP	--	Classification Accuracy

a. Or any other search algorithm.

Following subsections describe the steps and algorithmic details of the proposed method.

A. Generation Procedure

1) SGA: Application of genetic algorithm for feature selection was proposed in [5]. The entire set of features is represented by a discrete binary space called a chromosome. Each point in this space is called a gene and it represents an individual band. The value 0 in the i-th position indicates that the i-th feature is not included in the corresponding feature set; the value 1 in the j-th position indicates that the j-th feature is included in the corresponding feature set. In each iteration of the algorithm, a number of possible solutions (called population) are generated by means of applying certain genetic operators like recombination, crossover and mutation, in a stochastic process guided by a fitness measure. The algorithm seeks to evolve optimal solutiona to the problem and is fairly robust.

This work was supported by NASA/GAPP ER020059

B. Evaluation Procedure

1) Mutual Information of a feature subset: Shannon's Entropy is the classical measure of information [6] or the uncertainty in the realization of a random variable. It is defined as

$$H(X) = - E(Log(P[X]) \quad (1)$$

where E is the expectation and P is the probability distribution of occurrence of X. Mutual information is the information shared by two variables. Mutual Information between a subset and its complement determines the amount of redundancy. The larger the information a complementary subset shares with rest of the set, the more redundant it is. Let us consider X as a set of N features defined as X = {x1,x2….xN} and Y as a subset of X with S (S < N) features defined as Y = {x1,x2……xS}. The complementary set to Y is Z. The mutual information shared between Z and Y is

$$I(Z;Y) = H(Z) - H(Z \mid Y) \quad (2)$$

H (Z|Y) is the conditional entropy or the amount of information originally in Z minus the residual information in Z after Y has been specified. Assuming a Gaussian model for the random vector X, the probability distribution of X can be written as

$$P(X)=P(Y,Z)=\frac{1}{(2\pi)^{\frac{N}{2}}\sqrt{|\Sigma|}}\exp(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)) \quad (3)$$

μ is the mean vector and Σ is the variance-covariance matrix of dimension NxN. Using Bayes' theorem [7]

$$P(Z|Y) = \frac{P(Y,Z)}{P(Y)} \quad (4)$$

and applying (3) and (4) yields

$$-\log[P(Z|Y)] = -\log[P(Y,Z)] + \log[P(Y)]$$

$$= \frac{N}{2}\log[2\pi] + \frac{1}{2}\log[|\Sigma|] + \frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu) - (\frac{s}{2}\log[2\pi] + \frac{1}{2}\log[|\Sigma_{YY}|] + \frac{1}{2}(Y-\mu_Y)^T\Sigma_{YY}^{-1}(Y-\mu_Y)) \quad (5)$$

Using (1) and (5), we get

$$H(Z|Y) = -E[\log[P(Z|Y)]]$$

$$\Rightarrow H(Z|Y) = \frac{N-s}{2}\log[2\pi] + \frac{1}{2}\log[|\Sigma \,||\, \Sigma_{YY}^{-1}|] + \frac{1}{2}E[(X-\mu)^T\Sigma^{-1}(X-\mu) - (Y-\mu_Y)^T\Sigma_{YY}^{-1}(Y-\mu_Y)]$$

We can subtract of mean μ, μ_Y from X and Y respectively without the loss of generality.

$$\Rightarrow H(Z|Y) = \frac{N-s}{2}\log[2\pi] + \frac{1}{2}\log[|\Sigma \,||\, \Sigma_{YY}^{-1}|] + \frac{1}{2}(\sum_{i,j}^{N}E[X_i(\Sigma^{-1})_{i,j}X_j] - \sum_{i,j}^{s}E[Y_i(\Sigma_{YY}^{-1})_{i,j}Y_j])$$

$$= \frac{N-s}{2}\log[2\pi] + \frac{1}{2}\log[|\Sigma \,||\, \Sigma_{YY}^{-1}|] + \frac{1}{2}(\sum_{i,j}^{N}(\Sigma^{-1})_{i,j}E[X_iX_j] - \sum_{i,j}^{s}(\Sigma_{YY}^{-1})_{i,j}E[Y_iY_j])$$

$$= \frac{N-s}{2}\log[2\pi] + \frac{1}{2}\log[|\Sigma \,||\, \Sigma_{YY}^{-1}|] + \frac{1}{2}(\sum_{i,j}^{N}(\Sigma^{-1})_{i,j}(\Sigma^{-1})_{j,i} - \sum_{i,j}^{s}(\Sigma_{YY}^{-1})_{i,j}(\Sigma_{YY}^{-1})_{j,i})$$

$$= \frac{N-s}{2}\log[2\pi] + \frac{1}{2}\log[|\Sigma \,||\, \Sigma_{YY}^{-1}|] + \frac{1}{2}(\sum_{i}^{N}(1)_{i,i} - \sum_{i}^{s}(1)_{i,i})$$

$$= \frac{N-s}{2}\log[2\pi] + \frac{1}{2}\log[|\Sigma \,||\, \Sigma_{YY}^{-1}|] + \frac{N-s}{2}$$

$$= \frac{N-s}{2}\log[2\pi e] + \frac{1}{2}\log[|\Sigma \,||\, \Sigma_{YY}^{-1}|]$$

$$\Rightarrow H(Z|Y) = \frac{N-s}{2}\log[2\pi e] + \frac{1}{2}\log[|\Sigma \,||\, \Sigma_{YY}^{-1}|] \quad (6)$$

Using (1)

$$H(Z) = -E[\log[P(Z)]] = E[\frac{N-s}{2}\log[2\pi] + \frac{1}{2}\log[|\Sigma_{ZZ}|] + \frac{1}{2}(Z-\mu_z)^T\Sigma_{ZZ}^{-1}(Z-\mu_Z)]$$

$$\Rightarrow H(Z) = \frac{N-s}{2}\log[2\pi e] + \frac{1}{2}\log[|\Sigma_{ZZ}|] \quad (7)$$

Using (2), (6) and (7) we have

$$I(Z;Y) = \frac{1}{2}\log[|\Sigma^{-1} \,||\, \Sigma_{YY} \,||\, \Sigma_{ZZ}|] \quad (8)$$

Also,

$$\Sigma = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{bmatrix} \quad (9)$$

In (8) the variance-covariance matrix Σ is calculated only once, which is before the start of GP. Eventhough Σ_{YY} and Σ_{ZZ} are the components of Σ as shown in (9), $|\Sigma_{YY}|$ and $|\Sigma_{ZZ}|$ are calculated for each generated feature subset. More is the redundancy between the feature set Z and Y, larger will be I(Z;Y). I(Z,Y) will increase as the number of redundant bands in Z increases (or number of bands in Y decreases). If I(Z,Y) in the present iteration of EP is greater than that in the previous generation, the chromosome corresponding to the feature set is assigned more fitness. For larger efficiency, only the chromosomes with total number of 1's > m_{crit} are evaluated as implied by Fig. 1.

2) Jeffries-Matusita Distance(JMD): JMD [8] is a commonly used class separability index and is a saturating transform of Bhattacharyya distance, BD_{ij} [9]. JMD between

$$JMD_{ij} = 2\sqrt{1-\exp(-BD_{ij})} \quad (10)$$

two classes i and j is given in (10). The average JMD of a feature subset is calculated by using the formula

$$JMD_{avg} = \frac{1}{N_c} \sum_{i=1}^{Nc-1} \sum_{j=1}^{Nc} JMD_{ij} \qquad (11)$$

where Nc is the number of classes and N_c is the number of class pair combinations which is equal to ${}^{Nc}C_2$.

C. Stopping Procedure

1) Criterion Function in FSS1: A threshold value of the ratio of the amount of information in feature subset Y with respect to the total information in X is the stopping criterea for FSS1. This threshold ratio corresponds to one minus the ratio of redundant information of features which if removed from the feature set will have minimal impact on the classification accuracy. A ratio close to 1 (0.95 to 1) is chosen.

2) Criterion Function in FSS2: A subset with maximum JMD_{avg} satisfies the criterion function. The corresponding subset is outputted for validation.

D. Validation Procedure

1) Classification Accuracy: The feature subset selected by FSS2 is validated by performing accuracy analysis of the classified image obtained by Gaussian Maximum Likelihood Classification (GMLC) [10] of the feature subset. The classified image and corresponding test pixels on the ground are used to form an error matrix. Diagonal elements in the confusion matrix represent observations that agree both on test and classified images. Non-diagonal elements represent those that do not agree. The Khat index [11], k is used to quantify the accuracy of classification. It uses all the elements in the error matrix and represents the proportion of agreement after chance agreement is removed from consideration. It is given by

$$k = \frac{(P_o - P_c)}{1 - P_c} \qquad (12)$$

where P_o is the overall accuracy and P_c is the chance agreement probability.

E. Search Space Reduction

The two-step proposed methodology reduces the size of the search space drastically. Considering an exhaustive search strategy,

1) Case 1: The number of possible ways of choosing a feature subset Q from the set X of size N is given by (13)

$$Nopt = \sum_{i=1}^{N} {}^{N}C_i \qquad (13)$$

2) Case 2: The number of possible ways of selecting a feature subset of minimum size m from the set X is given by (14a). Let's say that a feature subset Y of size S (where $m \le S \le N$) is selected from the previous step, then the number of ways of selecting the subset Q from Y is given by (14b)

$$Nway1 = \sum_{i=m}^{N} {}^{N}C_i \qquad (14a) \qquad Nway2 = \sum_{i=1}^{M} {}^{M}C_i \qquad (14b)$$

In both cases, subset Q is being selected from a set X. In case 2, firstly a subset Y is selected and then subset Q is selected from Y. Subset Y is a set of features that satisfy the criterion function in SP of FSS1. Thus, it is a subset with size smaller than that of X, though it carries almost all the information present in X. Reduction in search space in case 2 with respect to Case 1 is calculated in (15). Percentage reduction in search

$$\text{Reduction} = \sum_{i=1}^{N} {}^{N}C_i - \left(\sum_{i=m}^{N} {}^{N}C_i + \sum_{i=1}^{M} {}^{M}C_i\right)$$

$$= \sum_{i=1}^{m-1} {}^{N}C_i + \sum_{i=m}^{N} {}^{N}C_i - \sum_{i=m}^{N} {}^{N}C_i - \sum_{i=1}^{M} {}^{M}C_i \qquad (15)$$

Since (m,N) > M

$$\Rightarrow \text{Reduction} = \sum_{i=1}^{m-1} {}^{N}C_i - \sum_{i=1}^{M} {}^{M}C_i > 0$$

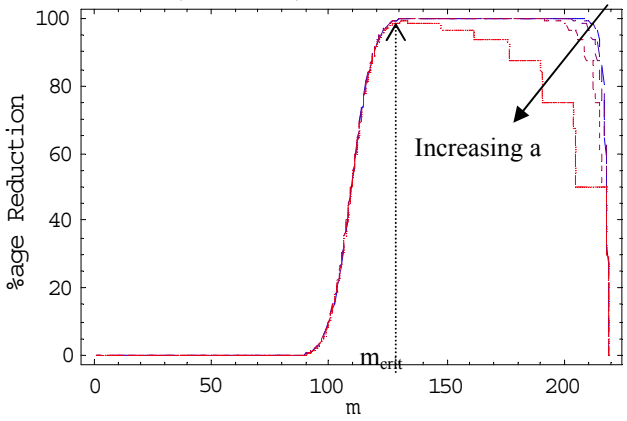

Figure 1. Search space reduction with redundancy

space for some typical values of N (= 220) and M (=m+Integer[a*(N-m)]) where 'a' varies from 0.05 to 0.95, is shown in Fig. 1. As can be seen from Fig. 1, with $m > m_{crit}$, the search space get reduced on the order ~ 90 to 100% for smaller values of a. For other values of a and m too, there is considerable decrease in search space as shown in Fig. 1. Since the computer load of most of the search algorithms are

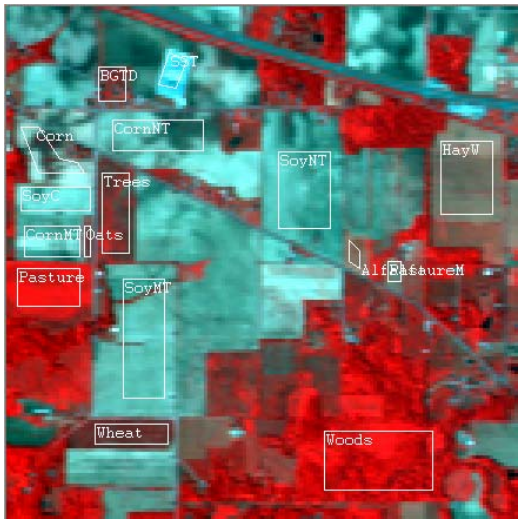

Figure 2. FCC of AOI (using band no. 51, 31 and 21)

directly proportional to the search space size, the proposed

TABLE II. ERROR MATRIX FOR TEST PIXELS

Class	Number of Samples in Class																RA
	Al	CNT	CMT	C	P	T	PM	HW	O	SNT	SMT	SC	Wh	W	BDT	SST	
Al	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100.0
CNT	0	79	17	1	0	0	0	10	0	0	6	1	0	0	1	0	68.69
CMT	0	10	51	0	0	0	0	0	0	4	2	4	0	0	0	0	71.83
C	0	0	0	36	0	3	1	1	0	0	3	0	1	0	2	0	76.59
P	0	0	0	0	59	1	0	0	0	0	0	0	0	4	0	0	92.18
T	0	0	0	2	0	71	0	1	1	0	2	0	10	0	4	0	78.02
PM	0	0	0	0	0	0	11	1	0	0	0	0	0	0	0	0	91.66
HW	2	0	0	0	0	0	2	147	0	0	0	0	0	5	0	0	94.23
O	0	0	0	0	0	0	0	0	6	0	0	0	2	0	0	0	75.0
SNT	0	0	4	0	0	0	0	0	0	148	3	8	0	0	0	0	90.79
SMT	0	2	3	1	0	1	4	3	0	19	153	8	0	6	2	0	75.74
SC	0	0	2	0	0	0	0	0	0	5	3	55	0	5	0	0	78.57
Wh	0	0	0	0	0	7	0	0	0	0	0	0	51	4	0	0	82.25
W	0	0	0	0	24	0	0	0	0	0	3	0	0	238	0	0	89.81
BDT	0	0	0	1	0	3	0	0	1	0	1	0	4	0	26	0	72.22
SST	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	100.0
RLA	80.0	86.81	66.23	87.80	71.08	82.55	61.11	90.18	75.0	84.09	86.93	72.36	75.0	90.83	74.28	100.0	

methodology will increase their computation efficiency.

F. *Experiments and Results*

1) Data Set Description: As a test of the proposed strategy, an experiment was performed on a hyperspectral set with N = 220 bands from a June 2, 1992 scene imaged by Airborne Visible-Infrared Imaging spectrometer (AVIRIS) in the Indian Pine Test site in Northwestern Indiana [12]. The scene is 145x145 pixels in size. Fig. 2 shows the false color

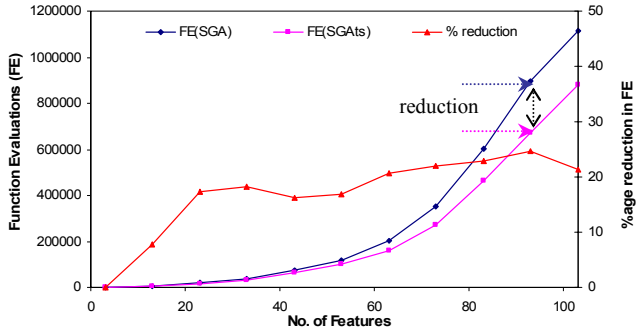

Figure 3. Computational efficiency of SGAts as compared to SGA

composite (FCC) of the area of interest (AOI) along with ground truth pixels of various classes. The ground truth samples were randomly divided into two halves - a training set and a test set.

2) Results: Experiments were carried out to assess and compare the performances of SGA and SGAts, the two step SGA using the proposed method. Fig. 3 shows the reduction in number of function evaluations (FE) in SGAts with respect to SGA is to the order ~ 20%. The scene was classified into 16 classes [12] viz. Alfalfa (Al), Corn min-till (CMT), Corn no-till (CNT), Corn (C), Pasture (P), Trees (T), Pasture-mowed (PM), Hay-windrowed (HW), Oats (O), Soy no-till (SNT), Soy min-till (SMT), Soy clean (SC), Wheat (Wh), Woods (W), Building-Driveway-Trees (BDT) and Stone-steel towers (SST). For a 15 band feature subset, the error matrix is shown in Table II. RA and RLA are the reference and reliability accuracy of each class. The overall and khat accuracy of classification for the data set is 83.41 % and 81.49% respectively.

III. CONCLUSION

As can be seen from theoretical proof in section II-B and a practical application in section II-F, the proposed method reduces the computational load of GA. Experiments being done on other search algorithms have shown the same trend.

REFERENCES

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1990.

[2] L. Bruzzone, F. Roli, and S. B. Serpico, "An extension to multiclass cases of the Jeffreys-Matusita distance," *IEEE Trans. Geosci. Remote Sensing*, vol. 33, pp. 1318–1321, Nov. 1995.

[3] L. Bruzzone and S. B. Serpico, "A tecnique for feature selection in multiclass cases," *Int. J. Remote Sensing*, vol. 21, pp. 549–563, 2000.

[4] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, pp. 131-156, 1997.

[5] W. Siedlecki and J. Sklansky, "A note on genetic algorithms for largescale feature selection," *Pattern Recognit. Lett.*, vol. 10, pp. 335–347, 1989.

[6] C. Shannon, "*A Mathematical Theory of Communication*," Bell System Technical Journal, vol. 27, pp. 379–423, 1948.

[7] T. Bayes, "An Essay Toward Solving a Problem in the Doctrine of Chances," *Philosophical Transactions of the Royal Society of London* , vol. 53, 370-418, 1764.

[8] P. H. Swain and S. M. Davis, *Remote sensing: the quantitative approach.* New York: McGraw-Hill, 1978.

[9] J.A. Richards, *An Introduction to Remote sensing digital image analysis*, Springer Verlag, 1986.

[10] A.H. Strahler, "The use of prior probabilities in maximum likelihood classification of remote sensing data," *Remote sensing of Environment*, 10,135-163,1980.

[11] R. G. Congalton and R. A. Mead, "A quantitative method to test for consistency and correctness in photointerpretation," *Photogrammetric Engineering and Remote Sensing*, 49(1):69-74, 1983.

[12] LARS, (2003, Dec.). *Documentation of Multispec* [Online],Available: www.ece.purdue.edu/~biehl/MultiSpec/documentation.html.