INNOVATIVE FEATURE SELECTION USED IN MULTISPECTRAL IMAGERY CLASSIFICATION FOR WATER QUALITY MONITORING

E. Charou[a*], S. Petridis[a], M. Stefouli[b], O. D. Mavrantza[c], and S. J. Perantonis[a]

[a]Computational Intelligence Laboratory, Institute of Informatics and Telecommunications,
National Center of Scientific Research "Demokritos"-(exarou, petridis,sper)@iit.demokritos.gr

[b]Institute of Geology and Mineral Exploration, Messogeion 70, 115 27, Athens, Greece - mst99@otenet.gr

[c]Laboratory of Remote Sensing, School of Rural and Surveying Engineering,
National Technical University of Athens, Greece - rannia@survey.ntua.gr

KEY WORDS: Remote Sensing, ASTER, Feature Selection, Classification, Algorithms, Land cover

ABSTRACT

This paper concerns the evaluation of a feature selection and classification techniques for land cover classification and potential monitoring of temporal changes. For the purposes of our study, an ASTER satellite image was used, acquired on 5 October, 2001, with 14 spectral bands resampled at the spatial resolution of 15 m. The study area concerns the geographic area of the water basin of Lake Vegoritis, that is located in the northern part of Greece. A variety of industrial and agricultural activities take place in the Vegoritis lake basin, which result to constant lowering of the lake water table and the change of land use, and subsequently, lead to degrading of the lake environment.
Besides the intensity level of the original bands, several features were created including spectral indices (e.g. NDVI), band ratios and products among selected bands, and Haralick texture features together with their second order combinations. The total of the features was used as input to a novel feature selection process, the Greedy Non-Redundant (*GreeNRed*) feature selection algorithm. This algorithm is based on information theory, and greedily selects features with no redundant information. The algorithm succeeded in keeping the complexity at low levels by restraining evaluations in one-dimensional feature spaces, whereas non-redundancy is achieved by a boosting-like sample weighting.
After the feature selection process different categories of classification methods were applied, namely, K-nearest neighbors and support vector machine. Classification accuracy assessment followed in order to derive the best classification method and consequently, to give further feedback as far as the performance of the feature selection algorithm is concerned and the usefulness of machine learning algorithms for land cover classification as a prerequisite for the assessment of the extended lake environment.

1 INTRODUCTION

The use of multispectral sensor technology has given significant rise to Earth observation and monitoring, because it provided a quite rich amount of information for forest fire monitoring, agricultural and forestry activities and inventories, protected area management, analysis of the quality of coastal waters, land use classification, etc. Nevertheless, redundancy in information among the bands, opens provides the opportunity to explore the optimal selection of bands for analysis. Theoretically, using images with more bands should increase automatic classification accuracy. However, this is not always the case. As the dimensionality of the feature space increases subject to the number of bands, the number of training samples needed for image classification has to increase too. If training samples are insufficient for the need, parameter estimation becomes inaccurate. The classification accuracy first grows and then declines as the number of spectral bands increases, which is often referred to as the Hughes phenomenon (Hughes, 2003).

In this work the efficiency of a novel information-theoretic based feature selection technique for selecting suitable bands of multispectral images used for land cover classification, was examined. This algorithm was applied on ASTER multispectral data of the water basin of Lake Vegoritis, Greece. After the feature selection process, different classification methods were applied and assessed (e.g. artificial neural networks and Support Vector Machine algorithms) in order to give further feedback concerning the feature selection algorithm, and to assess the usefulness of machine learning algorithms for land cover classification for estimating the extended environment for water quality monitoring purposes.

2 METHODOLOGY

2.1 Study Area and Data Used

For the implementation and evaluation of the Feature Selection algorithm, an ASTER multispectral image was used. The satellite image was acquired on 5 October, 2001, 9:33:34, includes 14 spectral bands and has a spatial resolution of 15 m for the VIS/NIR (3 used out of 4 bands), 30 m for the SWIR (6 bands) and 90 m for the TIR spectral range (5 bands). The image covered a part of the lake basin of Vegoritis, NW Greece (Figure 1). The river basin of the extended geographic area of Vegoritis includes four (4) inland lakes, while agricultural and mining activities are taking place. In the subset area only two of them are portrayed. Current agricultural practices may affect the space and time variability of the lake dynamics, sediment transport, water pollution (point and non-point releases of pollutants, discharge of wastewaters, etc). Therefore, monitoring of land cover and its dynamic change is a prerequisite for water quality assessment.

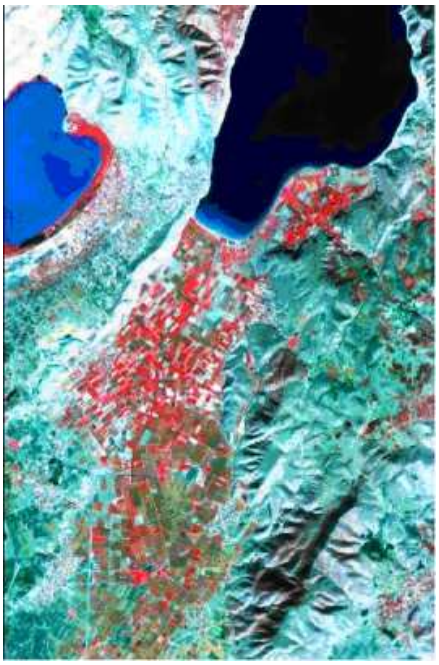

Figure 1: RGB composite of the ASTER image of the study area (Lake basin of Vegoritis). On Red, the NIR band (band 3N) of ASTER is displayed, on Green, the Red band of ASTER (band 2) is displayed and on Blue the green band (band 1) of ASTER is displayed.

2.2 Data Preparation and Pre-processing

In the pre-processing stage, all ASTER bands were resampled to 15m spatial resolution, so as to achieve the same pixel size, and further geodetically transformed into the Transverse Mercator Projection using the Hellenic Geodetic Reference System (HGRS'87). The image was cropped so as to keep the geographic area of interest that contained distinct land cover classes.

The land cover classes selected as tags for input to the feature selection algorithm were coded as followed: 30: pastures and shrubland (3 subclasses), 40: non- irrigated and irrigated arable land (5 subclasses), 50: permanent crops(3 subclasses) and 90: lakes (2 subclases), i.e. 13 different classes in total. The training set selection based on these classes was performed by delineating of small polygonal regions from each identified land cover type using Photointerpretation. 1500 training and validation samples were obtained for input to the GreeNRed algorithm.

Raw intensities are usually not sufficient for successful land cover classification. Therefore, for each band, and for each pixel, we additionally evaluated a number of textural and other features on a sliding window centered on each image pixel. The list of features used, together with their definitions, are shown in Table 1. In particular, the textural features were calculated as functions of the co-occurrence matrix $P_{\phi,d}(a_1, a_2)$, which is a matrix describing how frequently pixels with intensities a_1 and a_2 appear in the window of size $h \times w$ around the pixel, with a specified distance d in direction ϕ between them. Hence, there are in total 10 textural features per band (5 for vertical and 5 for horizontal displacement for the formation of the co-occurrence matrix). Adding the pixel intensity value as an 11th feature for each band, as well as the band combination features, we end up with a feature vector for each pixel composed by a total of 11×14 different features. For illustration purposes, a projection of a subset of the aquired samples in the NIR- REDplane is shown in figure 2.2, where only the four main classes are distinguished. Notice that even though the plane is informative, classes have a relatively big overlap.

Band Combinations			
NDVI	$\frac{\text{NIR} - \text{RED}}{\text{NIR} + \text{RED}}$	*R14*	$\frac{\text{SWIR1}}{\text{SWIR4}}$
DIFF	$\text{NIR} - \text{RED}$	*R15*	$\frac{\text{SWIR1}}{\text{SWIR5}}$
SDIFF	$\sqrt{\text{NIR} - \text{RED}}$	*R16*	$\frac{\text{SWIR1}}{\text{SWIR6}}$
TH	$\frac{\text{TIR1}}{\text{TIR4}}$	*R31*	$\frac{\text{SWIR3}}{\text{SWIR1}}$
Texture			
energy	$\sum_{i,j} P(i,j)^2$		
entropy	$\sum_{i,j} P(i,j) \cdot \log P(i,j)$		
contrast	$\sum_{i,j} (i-j)^2 \cdot P(i,j)$		
inv. diff. moment	$\sum_{i,j,i \neq j} \frac{P(i,j)}{1+(i-j)^2}$		
correlation	$\frac{\sum_{i,j}[i \cdot j \cdot P(i,j)] - \mu_i \cdot \mu_j}{\sigma_i \cdot \sigma_j}$		

Table 1: Original set of features. The RED, NIR, TIR1 . . . TIR4 and SWIR1 . . . SWIR6 refer to the corresponding Aster bands, whereas μ_i, μ_j, σ_i and σ_j are 1st and 2nd order -based statistics of the co-occurrence matrix $P = P^2_{\phi,d}$.

2.3 The *GreeNRed* algorithm : Description and Implementation

The *GreeNRed* (*GREEdy Non REDundant*) feature selection algorithm is an information - theoretic based algorithm that efficiently searches for a minimal set of features of non overlapping information. The feature efficiency is measured as the mutual information between the feature and the classification variable. In the context of multi-band remotely sensed images, each feature, denoted henceforth by x_i, corresponds to the intensity value of a band or to some band combination or textural feature previously evaluated, as explained in the previous section. The classification variable, denoted henceforth by Ω, corresponds to the land usage.

The main characteristic of the algorithm is that it focuses not only on finding useful features, but also on ensuring that the selected features are as much "independent" as possible, i.e. they don't contain overlapping information concerning a specific classification task. This is important, since it allows for further reducing the total number of features to be selected. Most importantly, the search for independent features is done by evaluations in individual feature space, without needing to consider at all their joint space, thus ensuring algorithm robustness and efficiency.

In the following we will assume that the reader is familiar with information theory, and especially with Shannon entropy and mutual information. For an introduction to information theory see, for instance, (1) or (3).

2.3.1 Locally Sufficient Features The algorithm is based on the concepts of redundancy and local sufficiency, expressed via information measures. Given two features x_1, x_2 and the class

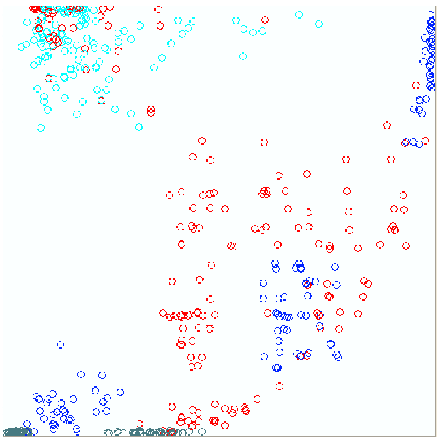

Figure 2: Projection of samples in the RED-NIR plane.

variable Ω, feature x_2 is said to be *locally redundant* with respect to x_1 in the region A of the observation space, if

$$\mathcal{I}_A(x_1, x_2; \Omega) = \mathcal{I}_A(x_1; \Omega)$$

i.e. the mutual information of the joint features with the class equals the information of the first feature with the class. Notice that region A doesn't refer to a geographical region but to a region defined via feature values in the joint feature space of all features.

Extending this concept for many features $\{x_i\}$, we call feature i *locally sufficient* at A with respect to features $j \neq i$, if

$$\mathcal{I}_A(\{x_j\}, x_i; \Omega) = \mathcal{I}_A(x_i; \Omega)$$

Local sufficiency implies that, in the specific region, we can discard all but one feature without loss of discriminative information. The aim of the algorithm is to effectively partition the observation space in a suitable way, such that a minimum number of sufficient features can cover the whole region of interest.

Using the mutual information criterion has two advantages. First, mutual information is closely connected to the optimal misclassification error, or Bayes error, by means of lower and upper bounds

$$\frac{\mathcal{H}(\Omega|\mathsf{X}) - 1}{\log(K-1)} < P_e(\mathsf{X}, \Omega) < \frac{1}{2}\mathcal{H}(\Omega|\mathsf{X}) \qquad (1)$$

The lower bound is known as the Fano inequality.

Second, the selected features are optimal, regardless of the specific classifier that is to be used later for the classification process. This implies a clear distinction between the feature selection and classification processes, that allows a freedom of choice of a more or less sophisticated classifier, whose training is likely to be greatly facilitated by the reduction of the input space dimension.

Mutual information has been used in the past as a criterion for feature selection (2), (7), though its use may be considered as limited because of complexity and lack of robustness in its evaluation via numerical methods. However, our algorithm minimizes the implications of these issues by considering only one-dimensional mutual information evaluations with the class, which make evaluations both robust and linear with respect to the number of samples.

Algorithm 1 Greedy Sufficient Feature Selection Procedure

$F \leftarrow \{\mathsf{x}_i\}_{i=1}^n$, $S \leftarrow \emptyset$
$A \leftarrow \mathcal{X}$, $j \leftarrow 1$
repeat
  $\mathsf{x}_j \leftarrow \text{argmax}_{\mathsf{x}_i \in F} \mathcal{I}_A(x_i; \Omega)$
  $F \leftarrow F \setminus \mathsf{x}_j$, $S \leftarrow S \cup \mathsf{x}_j$
  $A_j = \{\boldsymbol{x} : \boldsymbol{x} \in A, i(x_j; \Omega) > i_{suf}\}$, $A_j^c = A / A_j$
  $A \leftarrow A_j^c$, $j = j + 1$
until $A_j^c < A_\epsilon$ or $j = M$

2.3.2 Greedy Feature Selection The proposed algorithm is greedy in respect to the number of features to be found. At each step, features are examined, one by one, in respect to their suitability for discriminating the classes in the region of the observation space not yet covered by previously found features, and the best one is chosen. This is repeated until enough features are found. A formal description of the proposed greedy procedure is outlined in Algorithm 1.

The greedy approach offers three advantages. First, it allows us an adaptive control of the number of features to be selected, by inspection of the classification ability of those already selected. Second, it ensures linear algorithmic complexity with respect to the number of features to be selected. This complexity is far more satisfying than an exhaustive search of all the feature combinations. Finally, in this particular algorithm, the greedy approach guarantees effectiveness and self-containment of the features. Indeed, one should notice that not only should the selected features be locally sufficient but, also, the total of the selected features should determine the limits of the sufficiency regions, since otherwise the discrimination information would be lost. This can be better seen by denoting local mutual information as

$$\mathcal{I}_A(x_i; \Omega) = \mathcal{I}(x_i; \Omega | \mathsf{X} \in A)$$

which implies that local mutual information exists only by knowledge of the sufficiency regions.

Thus, at each step of the algorithm, the local sufficiency are always implicitly defined via the feature to be selected and the already selected features, which guarantees that the limits are indeed defined by the selected features.

As a price to pay for these benefits, it should be stressed that the greedy search is not guaranteed to provide the optimal minimum feature set, although it is very probable that the set of sufficient features selected will include most of the optimal sufficient features.

2.3.3 Soft sufficiency regions The implementation of the local sufficiency feature search with the greedy approach described above requires a way of finding sufficiency regions. Here, we propose to indirectly specifying the covering of regions by means of soft inclusion of samples in them. Namely, a region is defined as a set of weights $\{w^p\}$ having 1-to-1 correspondence with the samples $\{x^p\}$, $p = 1 \ldots P$. When $w^p = 0$, the corresponding sample x^p is not included in the considered region, whereas when $w^p = 1$, a sample is maximally included. Intermediate values are interpretated as "soft inclusion" of samples, indicating that the region around those samples is partially covered.

Instance-based soft sufficieny regions defintion has two important advantages. First, it provides a smooth partitioning of the space, increasing the robustness of the algorithm. Second, it allows for implicitly defining the regions, without the need of denoting the limits in terms of feature coordinates. Thus, the limits are implicit presence, even when evaluating local suitability of features,

Algorithm 2 The *GreeNRed* Feature Selection Algorithm

$D \leftarrow \{\boldsymbol{x}^p\}_1^P$
$\boldsymbol{w} = [\cdots \frac{1}{P} \cdots], \boldsymbol{w} \in \mathbb{R}^n$
$F \leftarrow \{x_i\}_{i=1}^N$, $S \leftarrow \emptyset$.
repeat
  $\forall \mathsf{x}_i \in F, \boldsymbol{x}^p \in D, I_{ip} \leftarrow \mathcal{I}_W({x_i}^p; \Omega)$
  $\forall \mathsf{x}_i \in F, I_i \leftarrow \sum_p I_{ip}$
  $\hat{\mathsf{X}} \leftarrow \text{argmax}_{x_i} I_i$
  $F \leftarrow F \setminus \hat{\mathsf{X}}, S \leftarrow S \cup \hat{\mathsf{X}}$
  $\forall w^p \in W, w^p \leftarrow 1 - \max_{i \in F} I_{ip}$.
  $\forall w^p \in W, w^p \leftarrow w^p / \sum_{i=1}^P w^p$
until enough features are selected

which may not necessarily be involved in the definition of the regions. This is a key observation for evaluating local mutual information with the class in one dimension : Mutual information is evaluated as

$$I_A(x_i; \Omega) = I_w(x_i; \Omega),$$

i.e region A is specified by weights, and,

$$\begin{aligned} \mathcal{I}_w(x_i, \Omega) &= \mathcal{H}_w(x_i) - \mathcal{H}_w(x_i|\Omega) \\ &= -\int p_w(x_i) \log p_w(x_i) + \\ &\quad \sum_k \int p_w(x_i|\omega_k) \log p_w(x_i|\omega_k) \end{aligned} \quad (2)$$

where p_w is a parzen estimate of the probability density function evaluated as

$$p_w(\boldsymbol{x}) = \sum_{\text{all}p} w^p \mathcal{N}(\boldsymbol{x}|\boldsymbol{x}^p, \sigma^p)$$

and

$$p_w(\boldsymbol{x}|\omega_k) = \sum_{p \mapsto \Omega_k} w^p \mathcal{N}(\boldsymbol{x}|\boldsymbol{x}^p, \sigma^p)$$

where the σ is automatically adjusted for each sample, $\mathcal{N}(\cdot, m, \sigma)$ denotes the normal probability density function with mean m and standard deviation σ, and $\{\omega_k\}$ are the set of values the classification variable takes (i.e 'lake', "grass" etc).

2.3.4 The algorithm The algorithm is outlined in Algorithm **2**. It consists of the following steps: The set of features under consideration and the selected features are initialized to contain all the features and no feature respectively. Each sample is initially given a weight $\frac{1}{P}$, where P are the number of samples. Then for each feature to be selected the following are done.

1. The suitability of each feature under consideration is evaluated as the mutual information of the feature with the class variable, given the weighted samples
2. The best feature is added to the selected features and removed from the features under consideration
3. The cover of the region with respect to the classification is evaluated as the local mutual information at the sample. New weights are given to the samples, according to how "uncovered" they are from the already selected features. The weights are normalized, so that they sum up to 1.

When the "covering" of the region is judged adequate, according to the local mutual information around the samples, the algorithm stops.

Feature Set	3-NN	SVM
14 B	0.87	0.90
14 B + Cmb	0.82	0.88
14 Bds + 5R	0.85	0.90
14 Bds + 5R + Cmb	0.86	0.90
GreeNRed	**0.89**	**0.92**

Table 2: Generalisation accuracy of the *K-NN* and *SVM* algorithms. Each of the 4 first rows corresponds to the classification accuracy achieved by a manually constructed feature subset. The last row corresponds to the feature subsect selected by *GreeNRed*. It can be seen that for both the *K-NN* and the *SVM* algorithms, the *GreeNRed* feature subset achieved better performance.

3 CLASSIFICATION AND ACCURACY ASSESSMENT

In order both to obtain feedback regarding the features relative for land cover and to evaluate the *GreeNRed* algorithm, we applied it to the feature space described in section 2.2. The evaluation aimed at comparing the generalisation performance of classification techinques with and without the *GreeNRed* feature selection. The classifiers used were the K-nearest neighbors (*K-NN*) and the linear support vector machine (*SVM*) which belong to different families, namely memory-based classifiers and linear classifiers. The sample set was formed by randomly choosing 500 points of the area under consideration where the 13 classes of land use were equally represented.

Table 2 presents the classification accuracies achieved with *K-NN* and *SVM* using 4 different feature ad-hoc assembled sets as well as the feature set constructed by the *GreeNRed* algorithm. The four subsets correspond to (a) the 14 original aster bands, (b) the 14 orignal band together with their second order combinations (119 features), (c) the orginal bands plus the ad-hoc feature combinations, as described in section 2.2 (19 features) and (d) the total of features with their second order combinations (299 features). Texture features are ommited from the tables since they have proven non-significant for the classification process when tested with varying number of graying levels, angle and displacment step. Furthermore, the "*GreeNRed* " feature set has been prduced by feeding the *GreeNRed* algorithm with the totality of features.. The 2 columns show the average correct rates of 10-fold cross validation sets, using 90% of the data for training and the remaining 10% for testing.

As a first observation, notice that the *SVM* achieved better overall performance for the task than the *K-NN* classifier. Most importantly, however notice that, in both cases, *GreeNRed* manages to increase generalisation performance. Notice that a big number of features doesn't guarantee an increase in classification accuracy, since these features may be irrelevant to classification and hence act as noise to the classification process. In figure 3, a function of the performance of the *K-NN* in respect to the number of features extracted by the *GreeNRed* algorithm is presented. In contrast, the performance of the classifier when a similar greedy but not redundancy-aware search was conducted.

4 CONCLUSION

From the examination of the results of the applied feature selection algorithm, the following conclusions were derived: The texture features did not provide significant results probably because of the relatively low spatial resolution of the ASTER image (15m), where patterns were prevented from being recognized. On

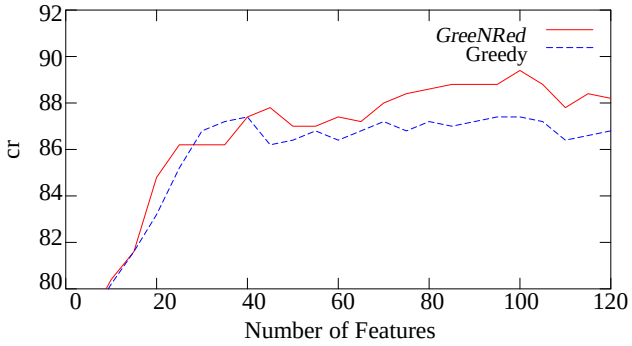

Figure 3: *GreeNRed*-greedy comparaison with the 3-NN algorithm

the other hand, second order combinations of bands (e.g. band products, etc.) turned out to provide more information, although some of them were considered as noise. At this point, the use of features selected by the GreeNRed algorithm improved the classification accuracy, as it kept the features that carried all useful information and ignored those that added noise to the dataset. From the analysis of the results presented in Table 2, it was concluded that, the used classification algorithms provided results of very high accuracy (90%), especially the SVM algorithm. However, the classification results were further improved by using participating features from the GreeNRed algorithms (92%), rather than using all combinatorial features (90%).

ACKNOWLEDGEMENTS

This work was based on methodology partially conducted under the ”INTerWatPol” Research Program. This Program is funded by the General Secretariat of Research and Technology, Greece under the Bilateral Cooperation Project between Turkey and Greece.

REFERENCES

Robert B. Ash. *Information Theory*. Dover Publications, 1990.

Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, Jul 1994.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, inc, 1991.

G.F Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Trans. Inform. Theory*, 14:55–63, 2003.

Anil Jain and Douglas Zongler. Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153–158, Feb 1997.

Mineichi Kudo and Jack Sklansky. Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, 33:25–41, 2000.

Nonjun Kwak and Chong-Ho. Input feature selection for classification problems. *IEEE Transactions on Neural Networks*, 13:143–159, 2002.

Hsien P.F. and D. Landgrebe. PhD thesis, 1998.

Schowengerdt R.A. *Remote Sensing” Models and Methods for Image Processing*. Academic Press, 1997.

J.A. Richards. *Remote Sensing Digital Image Analysis: An introduction*. Springer-Verlag Berlin Heidelberg, Second Edition, 1997.

P.H. Swain and Shirley M.D. *Remote Sensing: the Quantitative Approach*. McGRAW W-HILL, 1978.

T.Y. Young and K.S. Fu. *Handbook of Pattern Recognitions and Image Processing*. College of Engineering, University of Miami,. Coral Gables, Florida, 1986.

Petridis, S., E. Charou and S. J. Perantonis, 2003 Non-Redundant Feature Selection of Multi-Band Remotely Sensed Images for Land Cover Classification, In *Proceedings of the 2003 Tyrrhenian International Workshop on Remote Sensing*