NON-REDUNDANT FEATURE SELECTION OF MULTI-BAND REMOTELY SENSED IMAGES FOR LAND COVER CLASSIFICATION

Sergios Petridis, Eleni Charou and Stavros J. Perantonis
Computational Intelligence Laboratory,
Institute of Informatics & Telecommunications,
National Center of Scientific Research "Demokritos", Greece

This article presents a novel algorithm for feature selection applied to the land cover classification based on multi-band remotely sensed images. The algorithm is based on information theory, greedily select features with no redundant information, avoiding however evaluations in the joint space of features. Experimental comparison with texture features extracted from 14 bands shows the efficiency of the algorithm with respect to plain greedy feature selection

1 Introduction

In this article, we examine the efficiency of a novel information-theoretic based feature selection technique for selecting suitable bands of multispectral images, used for landscape classification.

In the first section we briefly review feature selection and describe its utility in the context of remotely-sensed data. In the second section we present the *GreeNRed* algorithm, which effectively selects useful and non-redundant features with respect to a specific classification task. In the third section we describe experiments done using multispectral images obtained from the Compact Airborne Spectrographic Imager (CASI).

2 Background

Multispectral sensors have been widely used since the 1960's. These sensors collect spectral data up to 20 bands and observe Earth's surface providing useful information for Environmental monitoring. The increasing availability of multispectral data and images has enriched us with better data for environmental monitoring. Currently, images obtained from multispectral sensors such as CASI are used for forest fire campaigns, agricultural and forestry activities and inventories, protected area, analysis of the quality of coastal waters, land use, etc. Multispectral data undoubtedly possess a rich amount of information. Nevertheless, redundancy in information among the bands opens and area for research 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 ([4]).

Generally speaking, classification performance depends on four factors: class separability, training sample size, dimensionality, and classifier type [9]. To improve classification performance, our attention in the paper is focused on dimensionality reduction. Dimensionality reduction can be achieved in two different ways [13]. The first approach is to select a small subset of features which could contribute to class separability or classification criteria. This dimensionality reduction process is referred to as feature selection or band selection. The other approach is to use all the data from original feature space and map the effective features and useful information to a lower-dimensional subspace. This method is referred to as feature extraction.

In feature selection, features that do not contribute to the discrimination of classes can be removed by assessment of specific criteria. Feature selection can not be performed indiscriminately. Methods must be devised that allow the relative worth of features to be assessed in a quantitative and rigorous way. A procedure commonly used is to determine the separability of different classes ([11]). Separability is a measure of probabilistic distance or within classes. The separability commonly used in feature selection is Mahalanobis, Divergence, Transformed Divergence, Bhattacharya, or Jeffries-Matusita, etc ([10]).

A separability analysis can be performed on the training data to estimate the expected error in the classification for various feature combinations ([12]). Supposing that the number of spectral bands is n, the problem of feature selection is to select the optimal subset of m with $m < n$. The number of feature combinations that need to be considered equals $\frac{n!}{(n-m)!m!}$. This number is too large for hyper spectral data and yields low efficiency in computation. Some algorithms such as the Branch-and-Bound algorithm, Sequential Backward and max-Min Feature selection which can determine the optimal or suboptimal feature set have been proposed to reduce the computational burden ([13]). For an overview of these approaches, see [5] or [6]. The algorithm presented here can be considered to be a member of the Forward Sequential Selection algorithm family.

3 The Greenred Algorithm

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 one band, whereas 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 each 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].

3.1 Locally Sufficient Features

The central concepts on which the algorithm is based are the concepts of redundancy and local sufficiency, expressed via information measures. Now, given two features x_1, x_2 and the class variable Ω, all considered as random variables, 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. 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.

By considering the mutual information criterion, the algorithm gains two important benefits. First, mutual information is closely connected to the optimal misclassification error, or Bayes error, by means of lower and upper bounds

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

The lower bound is known as the Fano inequality.

Second, the selected features are optimal, regardless of the specific classifier that will be used later for the classification process. This implies a clear distinction between the feature selection and classification processes, that allows a liberty 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.

3.2 Greedy Feature Selection

An important characteristic of the algorithm is its greedy nature with respect to the number of features to be found. At each step, features are examined, one by one, for their suitability for discriminating the classes in the region of the observation space which is not yet covered, and the best one is chosen. This is repeated until enough features are found.

The greedy approach offers three advantages. First, it allows us to control the number of features to be selected, as features are selected, by inspection of their classification ability. 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, and most importantly, 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 will 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 incremental algorithm, the local sufficiency regions are required to be 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

Algorithm 1 Greedy Forward Selection Sufficient Feature Procedure

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

probable that the set of sufficient features selected will include most of the optimal sufficient features.

The greedy procedure is outlined in Algorithm **1**.

3.3 Sample-Based Sufficiency Region Specification

The implementation of the local sufficiency feature search with the greedy approach described above requires a way of finding the limits of sufficiency regions. The algorithm we propose effectively deals with this issue by indirectly specifying the regions by means of the samples in a smooth way. Namely, each region is determined as a set of weights $\{w^p\}$ corresponding to the samples $\{x^p\}$, $p = 1 \ldots P$. A weight equal to zero for some sample, means that the region around this sample is not included in the considered region, whereas a weight equal to one, implies maximum inclusion of the region around the sample to the considered region.

This way of specifying regions has two important advantages. First, it provides a smooth partitioning of the space, which gives the algorithm robustness. Second, it allows for implicitly defining the regions, without the need of denoting the limits in terms of feature coordinates. Thus, it allows the implicit presence of the limits, even when evaluating the local suitability of features which are not by themselves, or only by themselves, defining these regions.

The last observation 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,

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

Algorithm 2 The *GreeNRed* Feature Selection Algorithm

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

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 "forest", "urban area" etc).

3.4 Outline of the algorithm

The algorithm is outlined in Algorithm **2**. In words, 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

	TopMap	CASI	Common
North	4677010.00	4675200.25	4674990
South	4663000.00	4663300.25	4666850
East	459000.00	458200.00	457540
West	453000.00	455750.00	455980

	TopMap	CASI	Common
Resol.	30 x 30	3.5 x 3.5	5 x 5
Pixels	467 x 200	3400 x 700	1628 x 312

Table 1: Map Coordinates and Image Sizes

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. 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.

4 Experiments

4.1 Preparation of the data

The images come from the project of "Parc Natural de la Garrotxa" and date from 23/07/98. The Institute Cartografic de Catalunya provides us with CASI flight images together with a tagged image (topographic map) of a region around the city of Olot. The flight images came in 14 bands, whereas there were 8 different tags. Some of the tags were not sufficiently represented on the image, and hence a merging of classes, resulting in 5 different classes, was done.

The coordinates of the flight image on which we focused and the tagged image are different. Since the coordinates where known, we made an alignment of images. The topological map has been cropped, since it corresponded to a wider region. Tables 4.1, 4.1 and 4.1 show respectively the maps coordinates, the digital image sizes for each feature and the possible values for the classes.

4.2 Texture Features

Raw intensities are usually not sufficient for successful land use classification. To this end, for each band, and for each pixel, we additionally evaluated a number of

	Tag Name
1	Forest
2	Continuous urban fabric
3	Non-irrigated arable land
4	Industrial or commercial units
5	Green urban areas

Table 2: The Categories Tags

texture features on a sliding window centered around each image pixel.

The features we use are 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. In our setting, we found that a window of 32×32 pixels together with 8 distinct levels of intensity and 2 pixels distance in both horizontal and vertical axes gave the best results.

Based on the co-occurrence matrix, the extracted features were :

- the *energy* , $sum_{a_1,a_2} P^2_{\phi,d}(a_1, a_2)$
- the *entropy*, $\sum_{a_1,a_2} P_{\phi,d}(a_1, a_2) \log P_{\phi,d}(a_1, a_2)$
- the *contrast*, $\sum_{a_1,a_2} (a_1 - a_2)^2 P_{\phi,d}(a_1, a_2)$
- the *inverse different moment* $\sum_{a_1,a_2,a_1 \neq a_2} \frac{P_{\phi,d}(a_1,a_2)}{(a_1-a_2)^2}$ and
- the *correlation* $\frac{\sum_{a_1,a_2}\left[a_1 a_2 P_{\phi,d}(a_1,a_2)\right]-\mu_{a_1}\mu_{a_2}}{\sigma_{a_1}\sigma_{a_2}}$ where $\mu_{a_1}, \mu_{a_2}, \sigma_{a_1} and \sigma_{a_2}$ are 1st and 2nd order -based statistcis of the co-occurrence matrix.

Hence there are in all 10 texture 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, we end up with a feature vector for each pixel composed by a total of 11×14 different features.

5 Results

The evaluation of the feature selection procedure was conducted by measuring the classification rate of a plain K-NN classifier. K-NN was chosen because of its simplicity and because it makes no assumption over the underlying pdf of the data. The training and a test sets were formed by randomly extracting 4000 pixels from the

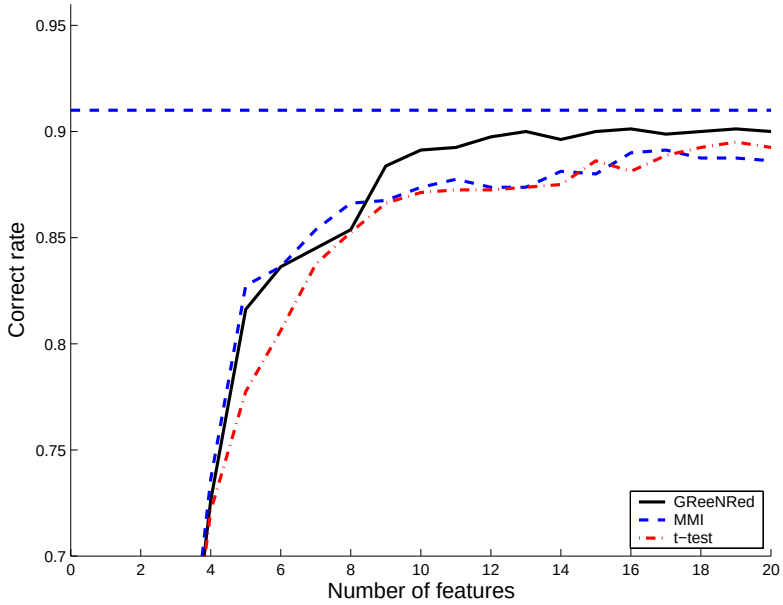

Figure 1: Performance of the *GreeNRed* algorithm. The straight line denotes the correct rate obtained using the whole set of features

image. 80% of the selected pixels were used for training and 20% for testing. In the selection procedure we made sure that each class was equally well represented.

Figure 5 shows the performance of the *GreeNRed* algorithm together with the performance of a plain greedy search of features using mutual information and the well-known t-test statistic. Furthermore, the straight line shows the performance of the K-NN using the full set of features, which was 91%. The x-axis shows the number of features used to train and test the K-NN. Only the 20-best features for each method are shown, since for more than to 20 features, there is no significant difference of the methods. The y-axis shows the correct rate obtained using K-NN, with $K = 3$. For the *GreeNRed* and simple greedy method, mutual information approximation was done using only up to 2nd order statistics to speed up the computation and to increase robustness, as it is discussed in [8]. Furthermore, mutual information was computed as an average of mutual information evaluated for every pair of classes.

Experimenting with the dataset revealed that there exist an important overlap of information among the features, which is, however "spread" among the features. Hence, since no feature combinations are done, there doesn't exist a compact set of features that contains *all* the classification information. However, as it is seen from the graph, one can attain 90% correct classification rate by keeping the first 13 fea-

tures found by the *GreeNRed* algorithm. Notice that plain greedy feature selection as well as the t-test statistic do not attain the *GreeNRed* performance.

Acknowledgement

The authors are grateful to the *Institute Cartographic de Catalunya* for providing the experimental data. They would also like to thank Dr. Ioannis Pratikakis for useful discussions during the preparation of the paper.

References

[1] Robert B. Ash. *Information Theory*. Dover Publications, 1990.

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

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

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

[5] 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.

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

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

[8] Sergios Petridis and Stavros J. Perantonis. On the relation between discriminant analysis and mutual information for supervised linear feature extraction. *Pattern Recognition*. submitted.

[9] Hsien P.F. and D. Landgrebe. PhD thesis, 1998.

[10] Schowengerdt R.A. *Remote Sensing" Models and Methods for Image Processing*. Academic Press, 1997.

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

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

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