A RANDOM FOREST CLASS MEMBERSHIPS BASED WRAPPER BAND SELECTION CRITERION : APPLICATION TO HYPERSPECTRAL

Arnaud Le Bris (a), Nesrine Chehata (b,c), Xavier Briottet (d), Nicolas Paparoditis (a)

(a) Université Paris-Est, IGN/SR, MATIS, 73 avenue de Paris, 94160 Saint Mandé , France
(b) IRD/UMR LISAH El Menzah 4, Tunis, Tunisia
(c) Bordeaux INP, G
E, EA 4592, F-33600, Pessac, France
(d) ONERA, The French Aerospace Lab, 2 avenue Edouard Belin, BP 74025, 31055 Toulouse Cedex 4, France

ABSTRACT

Hyperspectral imagery generates huge data volumes, consisting of hundreds of contiguous and often highly redundant spectral bands. Difficulties are caused by this high dimensionality. Feature selection (FS) is a possible strategy to reduce the number of bands, consisting in selecting the most relevant bands for a classification problem. It is adapted to the design of superspectral sensor dedicated to specific applications. FS is an optimization problem involving both a metric (that is to say a FS score or criterion measuring the relevance of feature subsets) to optimize and an optimization strategy. In this paper, a wrapper FS score based on Random Forests (RF) and taking into account RF class membership measures was proposed. It was compared to a state-of-the-art wrapper FS score (classification Kappa obtained by RF). Both were then evaluated quantitatively considering both classification performance reached applying different classifiers. An qualitative analysis was also performed to consider the stability/regularity of the selected features along the spectrum. Even though the quantitative evaluation showed little differences between the two tested FS criteria, there seemed to be a trend in favour of the proposed criterion. Taking into account the measures of class membership provided by a RF classifier slightly improved results, regularizing feature selection.

***Index Terms*—** Random Forests, Classification, Feature selection, Confidence

1. INTRODUCTION

Hyperspectral imagery generates huge data volumes, consisting of hundreds of contiguous and often highly redundant spectral bands. Difficulties are caused by this high dimensionality. First, the Hughes phenomenon can occur when classifying such data, even though modern classifiers such as Support Vector Machines (SVM) or Random Forests (RF) are less sensitive to these problems. Second, important computing times are required to process high dimensionality data. Feature selection (FS) is a possible strategy to reduce the number of bands. It consists in selecting the most useful bands, i.e. the most relevant ones for a problem. It has the advantage not to loose the physical meaning of the selected bands. It is also adapted to the design of superspectral sensor, that is to say to identify the most suitable spectral bands, to design superspectral cameras dedicated to specific applications (e.g. classification of urban land cover and material maps). Hence, both the most suitable number of spectral band and the most useful parts of the spectrum have to be identified thanks to supervised band selection.

1.1. Supervised band selection: state-of-the-art and proposed idea

Feature Selection (FS) is an optimization problem involving both a metric (that is to say a FS score or criterion measuring the relevance of feature subsets) to optimize and an optimization strategy. Supervised FS methods and criteria can be differentiated between “filter”, “wrapper” and “embedded”.

Filter methods compute relevance scores independently from any classifier. Many filter FS methods rank features according to an importance score, as the ReliefF method. In supervised cases, separability measures can be used to identify the sets of features making it possible to best separate classes, using Bhattacharrya or Jeffries-Matusita measures [1]. High order statistics from information theory, e.g. mutual information, can also be used to select the best feature sets [2].

For **wrapper** methods, the relevance score associated to a feature set corresponds to the classification accuracy obtained using this feature set. Examples of such approaches can be found in [2, 3] using SVM classifier or [4] using random forests.

Embedded FS methods are also related to a classifier, but usesfeature relevance scores different from classification accuracy. SVM-RFE [5] considers the influence of the different features in a SVM model. Random forests classifier gives another measure of feature importance [6], estimated as the difference between prediction accuracy before and af-

978-1-4799-7929-5/15/$31.00 ©2015 IEEE

ter permuting the features. Other embedded methods do not calculate a score for each feature, but for sets of features. For instance, [7] use generalization performance, e.g. SVM margin, as separability measure to rank sets of features.

Another issue for band selection is the optimization method: an exhaustive search of the best set of features is often impossible, especially for wrappers because of high computing times. Hence, heuristics have been proposed to find a near optimal solution without visiting the entire solution space. They can be differentiated into sequential (or incremental) and stochastic ones.

To put it in a nutshell, on the one hand, feature selection wrapper methods for classification often rely on traditional classification error rates (overall or average accuracies, Kappa,...) taking into account only final hard labellisation. On the other hand, filter separability scores (e.g. Bhattacharrya distance) measure the theoretical ability of a feature subset to well discriminate classes.

However, for each sample to classify, most classifiers provide not only the best label, but also a set of values measuring the degree of membership to the different possible classes, making it possible to assess the confidence of the obtained labellisation. Thus, it can be interesting to use such information as a FS criterion. That is to say to try to select the subset of features, making it possible to achieve a classification highly confident when samples are well classified and very uncertain for bad classified samples. This paper presents such FS criterion, based on the Random Forests classifier.

2. PROPOSED FEATURE SELECTION CRITERION

2.1. Random Forests

Random Forests (RF) [6] is a modification of bagging applied with decision trees. It can achieve a classification accuracy comparable to boosting [6], or SVM [8]. It does not require assumptions on the distribution of the data, which is interesting when different types or scales of input features are used. It was successfully applied to remote sensing data such as multispectral data, hyperspectral data, or multisource data. This ensemble classifier is a combination of tree predictors built from T multiple bootstrapped training samples. For each node of a tree, a subset of features is randomly selected. Then, the best feature with regard to Gini impurity measure is used for node splitting. For classification, each tree gives a unit vote for the most popular class at each input instance and the final label is determined by a majority vote of all trees.

Thus, for each sample to classify, the number of votes obtained by each possible label can be used as a class membership measure. Besides, it is provided by random forests at no additional computational cost.

Let $\mathcal{C} = \{c_1, ..., c_{nc}\}$ be the set of possible classes and $v(\mathbf{x}, c)$ the number of votes obtained by class c when classifying sample **x**. A class membership score m can then be obtained by normalizing the number of votes by the number of trees, i.e. $m(\mathbf{x}, c) = \frac{v(\mathbf{x},c)}{T}$

2.2. A feature selection score taking into account RF class membership measures

Let $\mathcal{X} = \{(\mathbf{x}_i, y_i)\}_{1 \leq i \leq n}$ be a set a groundtruth samples $\mathbf{x}_i$ and their associated true label y_i. A possible feature selection score $\mathcal{R}$ taking into account class membership measures and thus classification confidence can be defined as :

$$\mathcal{R}(\mathcal{X}) = \sum_{i=1}^{n} \delta(y_i, c(\mathbf{x}_i)).m(\mathbf{x}_i, c(\mathbf{x}_i))$$

with $\delta(i, j) = \{-1 \text{ if } i \neq j \text{ and } 1 \text{ otherwise}\}$, and $c(\mathbf{x})$ the label given to **x** by the classifier.

This score has the advantage to measure the ability to well classify the test samples for a given feature set and the separability between classes. Indeed, the more the samples are well classified, the more the score increases. The more the classifier is confident for well classified samples, the more the score increases. The more the classifier is confident for bad labelled samples, the more the score decreases.

3. EXPERIMENTS AND RESULTS

The proposed FS criterion was compared to an alternative FS score to optimize : the classification Kappa (κ) obtained by RF. Their relevance was assessed considering both classification accuracies reached applying different classifiers and the stability/regularity of the selected features.

Band selection was performed optimizing the FS scores using a genetic algorithm (GA). The process was launched several times to benefit from the stochastic behaviour of the algorithm and to get several near optimal solutions [9]. FS was performed for fixed numbers of bands. In our experiments, this happened to be 5 bands (the number of bands from which classification performance is no more greatly improved when adding new bands for both data sets) and 15 bands (that is to say more features than really required).

The proposed algorithms were tested on the VNIR ROSIS hyperspectral "Pavia City Center" and the VNIR-SWIR AVIRIS hyperspectral "Salinas" scenes [1]. Training and testing sample sets included respectively 100 and 500 samples per class for feature selection, while for evaluation training sample sets were limited to 50 samples per class and testing sample sets included all remaining pixels.

[1] http://www.ehu.es/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes

3.1. Classification accuracies reached using the selected feature subsets

The selected band subsets were then quantitatively evaluated according to the classification error rates obtained using several classifiers.

Evaluation using other classifiers different from RF

In order to be able to draw conclusions about the relevance of the FS criteria independently from RF classifier, selected band subsets were also evaluated considering the classification error rates obtained using two other classifiers : an optimized RBF kernel SVM and a maximum likelihood (ML) classifier assuming gaussian class models classifier. Indeed, FS can be used to design superspectral sensors dedicated to specific application, and thus the used wrapper FS score must not be too dependent to its associated classifier. Classification Average Accuracies (AA) and κ were considered.

Accuracies reached using either the proposed FS score or the κ as FS criterion are presented in table 1. Even though they are very similar, it can be observed that they are slightly increased using the proposed score. This is more explicit selecting more features (e.g. 15) than required.

Evaluation using RF

Selected band subsets were evaluated according to RF classification. Classification Average Accuracies (AA) and κ were considered in association to several indices taking into account class memberships measures provided by the classifier : the measure proposed in this paper for FS, the RF margin [6] (defined as the difference between the two best class membership measures) and the entropy of the membership measures [10]. Each of these indices was computed on all testing samples and synthesized into a overall value using the same formula as the FS score proposed in this paper. An average of these indices onfor true positive (TP) samples was also calculated.

These indices reached using either the proposed FS score or the κ as FS criterion are presented in table 2. There are very little differences between the two tested FS criteria. κ and AA reached using RF classifier for evaluation are very similar whatever the FS criterion, and contrary to previous evaluation using SVM or ML, no real trend in favour of the proposed FS criterion can be observed, even when selecting more bands than required.

On the opposite, the other indices taking into account class membership measures are generally improved using the proposed criterion instead of the κ coefficent as FS criterion.

3.2. Stability/regularity of the selected features

As described in [9], band importance profiles (fig.1) can be derived from intermediate results of a GA feature selection. Indeed, GA visits many band subsets almost as good as the final solution. Thus, the occurrence at which a band has been selected by GA among the intermediate best band subsets obtained at each generation can be used as a band importance measure. In the case of hyperspectral data, as the contiguous bands are correlated, such band importance profile should be regular (not too noisy). Its regularity is also related to the stability of the solutions obtained according to a FS criterion. This analysis remained qualitative. For Pavia data set, band importance profiles obtained selecting 5 bands are displayed on figure 1. It can then be noted that the importance of the first bands (e.g. band 3) is less important using the proposed criterion than using κ FS criterion, which is positive since these bands are noisy. To some extent, the overall shape of the profile obtained using the proposed criterion appears also smoother.

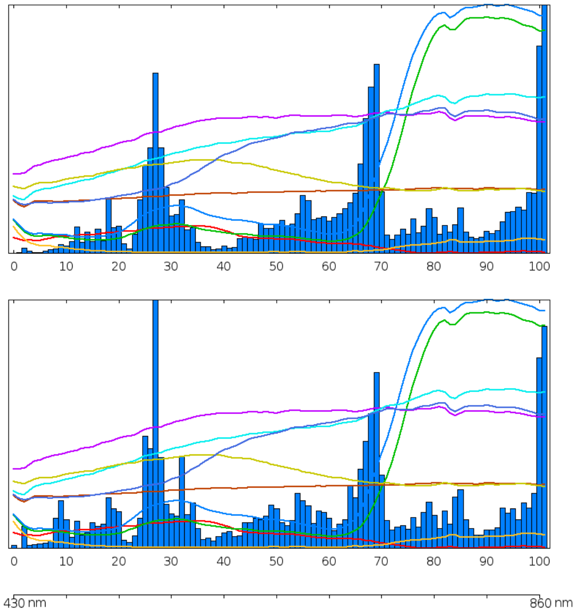

Fig. 1. GA band importances obtained selecting 5 bands for Pavia data set, according to the proposed FS score (left) and RF classification κ accuracy (right) as FS criterion. (x-axis = band numbers ; y-axis = band importances (histogram) ; The spectra of the different classes are also displayed).

4. CONCLUSION

Even though the quantitative evaluation showed little differences between the two tested FS criteria, there seemed to be a trend in favour of the proposed criterion : taking into account the class membership measures provided by a RF classifier slightly improves results and tends to regularize feature selection. This result was confirmed by a more obvious trend when selecting more features than required. Indeed, according to evaluation using classifiers different from RF, it seemed to select band subsets that are more general and less dependent to RF. Similar results were obtained with other data sets.

Table 1. Quantitative evaluation, assessing classification performance reached using rbf svm and ml classifiers

		Pavia				Salinas			
		SVM		ML		SVM		ML	
Nb bands	FS score	κ (%)	AA (%)	κ (%)	AA (%)	κ (%)	AA (%)	κ (%)	AA (%)
5	Proposed score	94.96	92.18	94.29	91.73	87.02	93.99	84.35	93.33
5	κ	94.84	92.16	94.17	91.63	86.82	93.93	84.19	93.26
15	Proposed score	95.46	93.17	94.03	91.16	87.82	94.67	84.73	93.82
15	κ	95.19	92.84	93.76	90.81	86.85	94.18	83.70	93.32

Table 2. Quantitative assessment of classification performance reached using RF classifier.

Nb bands	FS score	κ (%)	AA (%)	score (%)	score.TP (%)	margin (%)	margin.TP (%)	entropy	entropy.TP
		Pavia							
5	Proposed score	91.66	89.58	85.68	94.63	83.03	90.37	0.105	0.156
5	κ	92.19	89.81	84.91	94.49	82.21	90.06	0.103	0.157
15	Proposed score	92.67	90.52	85.46	93.65	82.13	88.70	0.136	0.189
15	κ	92.50	90.45	84.72	93.02	81.00	87.61	0.149	0.205
		Salinas							
5	Proposed score	81.85	90.87	64.19	89.40	61.34	80.36	0.089	0.252
5	κ	81.80	90.81	64.02	89.28	61.13	80.20	0.090	0.255
15	Proposed score	82.77	91.72	64.78	87.86	61.13	77.82	0.120	0.295
15	κ	82.81	91.78	64.41	87.20	60.44	76.71	0.134	0.313

5. REFERENCES

[1] S. B. Serpico and G. Moser, "Extraction of spectral channels from hyperspectral images for classification purposes," *IEEE TGRS*, vol. 45, no. 2, pp. 484–495, Feb. 2007.

[2] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Transactions on Neural Networks*, vol. 20, no. 2, pp. 189–201, Feb. 2009.

[3] S. Li, H. Wu, D. Wan, and J. Zhu, "An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine," *Knowledge-based Systems*, vol. 24, pp. 40–48, 2011.

[4] R. Díaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, vol. 7, no. 3, pp. 1–13, 2006.

[5] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 289–422, 2002.

[6] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[7] M. Pal, "Margin-based feature selection for hyperspectral data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 11, pp. 121–220, 2009.

[8] M. Pal, "Random forest classifier for remote sensing classification," *International Journal of Remote Sensing*, vol. 26, no. 1, pp. 217–222, 2005.

[9] A. Le Bris, N. Chehata, X. Briottet, and N. Paparoditis, "Identify important spectrum bands for classification using importances of wrapper selection applied to hyperspectral," in *Proc. of the 2014 International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM'14)*, Nov. 2014.

[10] A. Joshi, F. Prikli, and N. Papanikopoulos, "Multiclass activelearning for image classification," in *Proc. of CVPR*, 2009.