USE INTERMEDIATE RESULTS OF WRAPPER BAND SELECTION METHODS: A FIRST STEP TOWARD THE OPTIMIZATION OF SPECTRAL CONFIGURATION FOR LAND COVER CLASSIFICATIONS

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) EA 4592 G&E, ENSEGID-IPB, University of Bordeaux, 1, allée F. Daguin, 33607 Pessac Cedex, France
(d) ONERA, The French Aerospace Lab, 2 avenue Edouard Belin, BP 74025, 31055 Toulouse Cedex 4, France

ABSTRACT

Intermediate results of two state-of-the-art wrapper feature selection approaches (GA and SFFS) associated to a classifier (linear SVM) applied to hyperspectral data sets were used to derive information about band importance for specific land cover classification problems. The impact of the number of selected bands on classification accuracy was obtained thanks to SFFS, while a band importance measure was derived from intermediate sets of bands tested by GA. Such results are a first step toward the identification of the most suitable spectral bands to design superspectral camera systems dedicated to specific applications (e.g. classification of urban land cover and material maps).

Index Terms— Feature selection, Classification, Support Vector Machines, Sequential Forward Floating Search, Genetic algorithm, Hyperspectral, Sensor design

1. INTRODUCTION

1.1. Feature selection for sensor design

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 [1] except when very few training data is available. Second, important computing times are required to process high dimensionality data. Third, storing data requires huge volumes. Last, displaying high dimensionality imagery can be necessary. Thus, two strategies makes it possible to reduce the number of features. On the one hand, feature extraction methods reformulate and sum up original information. On the other, feature selection (FS) selects the most useful bands, i.e. the most relevant ones for a problem. The latter has advantages compared to feature extraction: first, it makes it possible not to lose the physical meaning of the selected bands. Last but not least, it is adapted to the design of superspectral sensor. Indeed, our final work aims at identifying the most suitable spectral bands (both position in spectrum and width), 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 band selection.

1.2. Band selection: state-of-the-art

Feature selection (FS) methods and criteria can be differentiated between “filter”, “wrapper” and “embedded”. Supervised and unsupervised ones can be differentiated too.
Filter methods compute relevance scores independently from any classifier. Some filter FS methods rank features according to a score of importance, as the ReliefF method. Other calculate importance measures from a feature extraction (e.g. PCA or LDA) [3]. 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 [4, 5, 6, 7] or other measures, such as Minimum Estimated Abundance Covariance [8]. High order statistics from information theory, e.g. mutual information, can also be used to select the best feature sets, either in unsupervised [9] or supervised [10, 11] cases. The Orthogonal Projection Divergence [12] is another measure of correlation between bands.
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 [11, 13, 14] using SVM classifier, [15] using maximum likelihood classifier, [16] using random forests.
Embedded FS methods are also related to a classifier, but use feature relevance scores different from classification accuracy. SVM-RFE [17] considers the influence of the different features in a SVM model. It has been extended to multi-kernel SVM [18]. Random forests classifier gives another

measure of feature importance [19], estimated as the difference between prediction accuracy before and after permuting the features. Other embedded methods do not calculate a score for each feature, but for sets of features. For instance, [20, 21] use generalization performance, e.g. SVM margin, as separability measure to rank sets of features.

Nevertheless, hybrid approaches involving several criteria belonging to different categories often exist, as for instance in [11] or [13] where FS is based on a wrapper method respectively guided or associated to filter criteria (mutual information between selected bands, and between the ground truth).

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. Incremental search strategies includes the Sequential Forward Search (SFS) or its opposite the Sequential Backward Search (SBS). Variants such as Sequential Forward Floating Search (SFFS) or Sequential Backward Floating Search (SBFS) [22], or Steepest Ascent (SA) [7] have also been proposed. Several stochastic optimization strategies have also been used for feature selection, including Genetic Algorithms (GA) [13, 11, 20, 14], Particle Swarm Optimization (PSO) [8], clonal selection [15], ant colony [23] or even simulated annealing [6].

Thus, many FS methods have been proposed in literature. However, FS (and especially wrapper) is often considered as a first step in a classification workflow. Some methods even performed both band selection and optimization of classifier parameters jointly as [14]. Wrapper results are considered as obtaining the best classification accuracies for a problem, but also as sometimes lacking of generality and being too dependent to the associated classifier. As our final work aims at designing superspectral sensors dedicated to specific classification problems, this paper intends to use the intermediate results of two wrapper feature selection methods in order to draw conclusions about band importance, to define the most important parts of the spectra for a specific problem.

2. PROPOSED METHODS

As land cover classification is our final goal, two wrapper approaches (SFFS and GA) optimizing classification accuracies were used in this paper. This made it possible to examine sets of bands, instead of individual bands. Intermediate results of these methods will then be considered.

Both use linear one-against-one SVM classifier. Indeed this classifier directly implicitly takes “ratios” between bands into account, which would not have been done when using separability measures. Besides, using a linear kernel with a fixed cost parameter instead of a more complex kernel is a way to avoid possible overfitting, caused by both feature selection and kernel optimization.

2.1. Sequential Forward Floating Search (SFFS) associated to SVM

The first band selection method used in this study was SFFS associated to a linear SVM classifier. SFFS is a state-of-the-art sequential optimization heuristic proposed by [22]. It is reminded below in pseudo-code, with C the original set of all bands, S the set of selected bands and p the maximum number of bands to select.

```
**Initialization :** Find band $c \in C$ maximizing classification accuracy: $S \leftarrow \{c\}$
**while** ($\#S < p$)
  Find band $c \in C \setminus S$ so that $S \cup \{c\}$ maximizes classification accuracy: $S \leftarrow S \cup \{c\}$
  Question $S$: find band $s \in S$ so that $S \setminus \{s\}$ maximizes classification accuracy. It means that $s$ is less important for classification than the other bands of $S$, since removing it decreases the classification accuracy less.
  **if** $s \neq c$ **then** $S \leftarrow S \setminus \{s\}$ **endif**
**endwhile**
```

SFFS provides useful intermediate results. Indeed, it selects the “best” sets of bands for different dimensions, starting from 1. Thus, it makes it possible to observe the evolution of classification accuracy, depending on the number of selected bands, and then to decide how many bands are necessary to obtain suitable results. Other sequential methods, such as SVM-RFE [17] could also provide such information. However, contrary to RFE, SFFS question at each step the selected set of bands, obtained at previous step. Therefore, SFFS makes it possible to see whether the number of selected bands has a strong influence on the position of the selected bands along the spectrum.

2.2. Genetic algorithm associated to SVM

The second band selection method used in this study was a Genetic Algorithm (GA) associated to a linear SVM classifier. GA is a family of stochastic optimization heuristics. The variant used in our experiments is described below, with p the number of bands to select :

```
**Initialization** ($t = 0$) **:** Randomly generate a population of $N$ sets of $p$ bands
**while** max number of generations unreached ($t < tmax$)
  $t \leftarrow t + 1$
  Calculate the score of each set of $p$ bands of the current population
  Keep only the $n$ ($n < N$) best sets of bands according to these scores. Let $S(t)$ be this population.
  Generate a new population of $N$ sets of $p$ bands from these remaining sets of bands :
  **for all** new set of bands **do**
```

```
        Create a new set of $p$ bands by crossing two parents randomly chosen in $S(t)$. Random mutations can also occur with a fixed probability, randomly replacing a selected band by another band, in order to avoid to stay in a local minimum.
    endfor
endwhile
```

This approach has advantages for our problem. First, only the best solution is usually kept, while GA has visited many other candidates. Many of them have scores quite similar to the score of the best solution. Therefore, it appeared interesting to use these intermediate results to determine which bands are often selected in these intermediate solutions. Thus, the proposed method consist in associating a score $I(c)$ to each band c , measuring the occurrence at which it has been selected by GA among the different n best sets of bands obtained at each generation: $I(b) = \sum_t \sum_{S(t)} \delta(b, S(t))$ where $\delta(b, S) = 1$ if $b \in S$, 0 otherwise.

To increase robustness, GA-SVM can be launched several times (i.e. different initializations/mutations) and over several training/testing sets randomly extracted from the whole data set. The proposed importance score is calculated for each of these results. Finally, the mean and the standard deviation of these scores are considered for each band: the mean gives the importance associated to each band, while the standard deviation is a way to detect potentially "unstable" good features.

Last, intermediate results of GA could also be useful to detect repetitive patterns of selected bands. This time, the interesting information would not be the occurrence of detection of individual bands, but of sequences of bands.

3. TESTS AND RESULTS

The proposed algorithms were tested on the VNIR hyperspectral Pavia Center data set [1]. Training and testing sets included respectively 100 and 500 samples per class. The score to optimize was the overall accuracy. Even if average accuracy or F-score of the worst classified class could also be relevant, it must be kept in mind that both training and testing data included the same number of samples per class.

Information brought by SFFS-SVM about the number of useful bands is illustrated on fig. 1: classification accuracy tends to not increase a lot using more than 5-6 bands. This might be related to the fact that the classes are quite general, and additional bands do not bring much information compared to classic multispectral data, which is confirmed by the position of the most important features according to GA-SVM.

Thus, GA-SVM was launched to select 5 bands. Feature importance brought by GA-SVM are presented on fig. 2. Groups of adjacent important features are clearly visible, corresponding to the most important parts of the spectrum for this classification. The positions of these groups seem coherent with the spectra (just near or during changes in spectrum slope). The standard deviation of the importances gave another interesting information: bands 3 to 5 appeared important (which seemed confirmed by the spectra) but standard deviation of their importance is high, corresponding to the fact they are quite noisy.

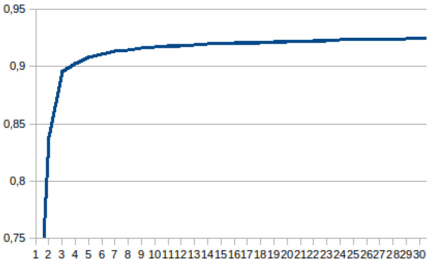

Fig. 1. Classification overall accuracy depending on the number of selected bands: on this data set, classification accuracy tends to not increase a lot using more than 5-6 bands

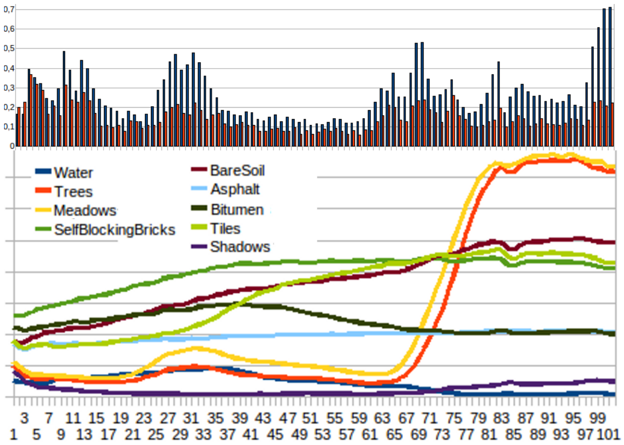

Fig. 2. Top row: GA-SVM band importances (blue: mean importance ; red: standard deviation). Second row: spectra of classes of the Pavia data set

4. CONCLUSION

These results are a first step toward the optimization of spectral configuration for specific land cover classifications, identifying the most relevant parts of the spectrum. Further works will cross them with information about correlation between bands, both to try to optimize the width of spectral bands and to identify whether one band is really more important than the others among clusters of highly correlated bands.

The occurrence score can also be modified to take into account how the selected bands influence the model among the set of bands they belong to (for instance accumulating SVM-RFE weights associated to the selected bands among

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

the different band sets provided by GA). Another possible improvement mentioned above is to try to identify several patterns/sequences of bands, among selected sets of bands.
Further experiments will also be carried out on a data set dedicated to urban material map, with a more detailed legend than for the Pavia data set and then for which hyperspectral imagery may be more relevant.

5. REFERENCES

[1] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE TGRS*, vol. 43, no. 6, pp. 1351–1362, June 2005.

[2] M. Pal and G.M. Foody, "Feature selection for classification of hyperspectral data by SVM," *IEEE TGRS*, vol. 48, no. 5, pp. 2297–2307, 2010.

[3] C.-I. Chang, Q. Du, T.-L. Sun, and M.L.G. Althouse, "A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification," *IEEE TGRS*, vol. 37, no. 6, pp. 2631–2641, Nov. 1999.

[4] L. Bruzzone and S. B. Serpico, "A technique for feature selection in multiclass problem," *IJRS*, vol. 21, no. 3, pp. 549–563, 2000.

[5] M. Herold, M. E. Gardner, and D. A. Roberts, "Spectral resolution requirements for mapping urban areas," *IEEE TGRS*, vol. 41, no. 9, pp. 1907–1919, Sept. 2003.

[6] S. De Backer, P. Kempeneers, W. Debruyn, and P. Scheunders, "A band selection technique for spectral classification," *IEEE GRSL*, vol. 2, no. 3, pp. 319–232, 2005.

[7] 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.

[8] H. Yang, Q. Du, and G. Chen, "Particle swarm optimization-based hyperspectral dimensionality reduction for urban land cover classification," *IEEE JSTARS*, vol. 5, no. 2, pp. 544–554, Apr. 2012.

[9] A. Martínez-Usó, F. Pla, J. Martínez Sotoca, and P. García-Sevilla, "Clustering-based hyperspectral band selection using information measures," *IEEE TGRS*, vol. 45, no. 12, pp. 4158–4171, Dec. 2007.

[10] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, July 1994.

[11] 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.

[12] Q. Du and H. Yang, "Similarity-based unsupervised band selection for hyperspectral image analysis," *IEEE GRSL*, vol. 5, no. 4, pp. 564–568, Oct. 2008.

[13] 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.

[14] L. Zhuo, J. Zheng, F. Wang, X. Li, A. Bin, and J. Qian, "A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine," *IAPRS*, vol. 37, no. B7, pp. 397–402, July 2008.

[15] L. Zhang, Y. Zhong, B. Huang, J. Gong, and P. Li, "Dimensionality reduction based on clonal selection for hyperspectral imagery," *IEEE TGRS*, vol. 45, no. 12, pp. 4172–4186, Dec. 2007.

[16] 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.

[17] 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.

[18] D. Tuia, G. Camps-Valls, G. Matasci, and M. Kanevski, "Learning relevant image features with multiple-kernel classification," *IEEE TGRS*, vol. 48, no. 10, pp. 3780–3791, Oct. 2010.

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

[20] H. Fröhlich, O. Chapelle, and B. Schölkopf, "Feature selection for support vector machines by means of genetic algorithms," in *Proc. of the 15th IEEE International Conference on Tools with Artificial Intelligence*, 2003, pp. 142–148.

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

[22] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, pp. 1119–1125, Nov. 1994.

[23] S. Zhou, J.P. Zhang, and B.K. Su, "Feature selection and classification based on ant colony algorithm for hyperspectral remote sensing images," in *Proc. of CISP'09*, Oct. 2009, pp. 1–4.