A Combined Support Vector Machines Classification Based on Decision Fusion

Mathieu Fauvel [*◇], Jocelyn Chanussot [*]

Jon Atli Benediktsson [◇]

[*]Signals & Images Laboratory - LIS Grenoble
Grenoble National Polytechnical Institute - INPG
BP 46 - 38402 St Martin d'Heres - FRANCE
E-mail: {mathieu.fauvel, jocelyn.chanussot}@lis.inpg.fr

[◇]Department of Electrical and Computer Engineering
University of Iceland
Hjardarhagi 2-6, 107 Reykjavik, ICELAND
E-mail: benedikt@hi.is

Abstract— **Decision fusion for classification of hyperspectral data from urban area is addressed. Classical classification algorithms are based on the spectral signature of the individual classes. For urban area, where classes could be defined in accordance with the shape of the structure, these methods have a major drawback: no spatial information are contained in the spectrum. A new method has been proposed that considers the spatial content, but it reduces the spectrum to a small number of bands and does not exploit the spectral richness of the hyperspectral data. In this paper, we propose to use both approaches, and then fuse them. The data are first pre-processed to extract some spatial information. Using Support Vector Machines (SVMs), the data are classified. Finally, according to the property of SVMs outputs, we propose to fuse the results using three different operators. Results are presented on real hyperspectral data from urban area. The proposed approach is positively compared to the results obtained by each of the classifiers used separately.**

I. Introduction

Hyperspectral images are now fully available. Many approaches have been defined to handle the characteristics of high dimensional data. Algorithms, such as the Decision Boundary Feature Extraction (DBFE), focus on finding a subspace projection of the original space using spectral class characteristics, then a statistical classifier is usually applied, [1] (*e.g.,* the maximum likelihood classifier). However, these methods concentrate the analysis only on the spectral data, while spatial contents is not used. Many image processing algorithms can be used on individual images, but they limit the spectral information to only one channel. Recently, a principal component analysis step followed by a morphological processing step were used to create the Extended Morphological Profile (EMP) [2]. This approach extracts spatial information from the data and it is well suited for the analysis of urban area images. The EMP was classified with a neural network. This approach gave good results in terms of classification accuracies. However, PCA reduces the whole spectrum to a few bands and the richness of the hyperspectral data is not fully used. While the use of all the spectrum without morphological feature extraction does not provide information about the size, the shape or the orientation of the structure. For an urban area context, both kinds of information are needed to allow a fine classification.

In this paper, we propose to fuse the results obtained by a separate use of the spectral data and the Extended Morphological Profile. Each data are processed by SVMs classifiers. The SVMs where chosen according to their strong capability to deal with remote sensing data [3], [4]. The result from each classifiers are aggregated according to the intrinsic characteristic of the SVMs outputs:

- outputs are not bounded,
- outputs are signed numbers.

Classical fuzzy fusion operators, such as T-norm, T-conorm or symmetrical sum [5], cannot deal with signed data. For the fusion, we must define operators which use sign as information. In this paper, we suggest three different operators. First, a modified version of the max operator, namely the *absolute maximum* decision rule is applied. Second, the agreement of classifiers is suggested. The agreement is seen as the probability of the outputs of each classifier. Third, a rule based on the majority voting, which was initially used for multiclass SVMs is investigated.

In the following, we start with a brief introduction of SVMs (Sec. II). Then, in relation to the nature of the classifiers' outputs, the fusion operators are presented and the fusion scheme is detailed (Sec. III). The proposed method is applied on real hyperspectral remote sensing data of an urban area and results are given Sec. IV. Finally, conclusions are drawn in Sec. V.

II. Support Vector Machines

A. Linear SVM

For a two-class problem in a n-dimensional space $\mathbb{R}^n$, we assume that l training samples, $\mathbf{x}_i \in \mathbb{R}^n$, are available with their corresponding labels $y_i = \pm 1$, $S = \{(\mathbf{x}_i, y_i) \mid i \in [1, l]\}$. The

SVM method consists of finding the hyperplane that maximizes the margin (see Fig. 1), *i.e.*, the distance to the closest training data points in both classes. Noting $\mathbf{w} \in \mathbb{R}^n$ as the vector normal to the hyperplane and $b \in \mathbb{R}$ as the bias, the hyperplane H_p is defined as

$$\langle \mathbf{w}, \mathbf{x} \rangle + b = 0, \ \forall \mathbf{x} \in H_p \tag{1}$$

where $\langle \mathbf{w}, \mathbf{x} \rangle$ is the inner product between $\mathbf{w}$ and $\mathbf{x}$. If $\mathbf{x} \notin H_p$ then $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$ is the distance of $\mathbf{x}$ to H_p. The sign of f corresponds to decision function $y = sgn\left(f(\mathbf{x})\right)$. The optimal parameters $(\mathbf{w}, b)$ are found by solving

$$\min \left[\frac{\|\mathbf{w}\|^2}{2} + C \sum_{i=1}^{l} \xi_i \right] \tag{2}$$

subject to

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i, \ \xi_i \geq 0 \ \forall i \in [1, l] \tag{3}$$

where the constant C control the amount of penalty and ξ_i are *slack* variables which are introduced to deal with misclassified samples (see Fig. 1). This optimization task can be solved through its Lagrangian dual problem.

$$\max_{\alpha} \quad \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$
$$\text{subject to} \quad 0 \leq \alpha_i \leq C \quad \forall i \in [1, l] \tag{4}$$
$$\sum_{i=1}^{l} \alpha_i y_i = 0.$$

Finally:

$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i. \tag{5}$$

The solution vector is a linear combination of some samples of the training set, whose α_i is non-zero, called *Support Vectors*. The hyperplane decision function can thus be written as:

$$y_u = sgn\left(\sum_{i=1}^{l} y_i \alpha_i \langle \mathbf{x}_u, \mathbf{x}_i \rangle + b \right) \tag{6}$$

where $\mathbf{x}_u$ is an unseen sample.

B. Non-linear SVM

Using the so-called *Kernel Trick*, one can generalize SVMs to non-linear decision functions. This way, the classification capability is improved. The idea is as follows. Via a non-linear mapping Φ, data are mapped onto a higher dimensional space F

$$\begin{aligned} \Phi : \ \mathbb{R}^n &\rightarrow F \\ \mathbf{x} &\mapsto \Phi(\mathbf{x}). \end{aligned} \tag{7}$$

The SVM algorithm can now be simply considered with the following training samples: $\Phi(S) = \{(\Phi(\mathbf{x}_i), y_i) \mid i \in [1, l]\}$.
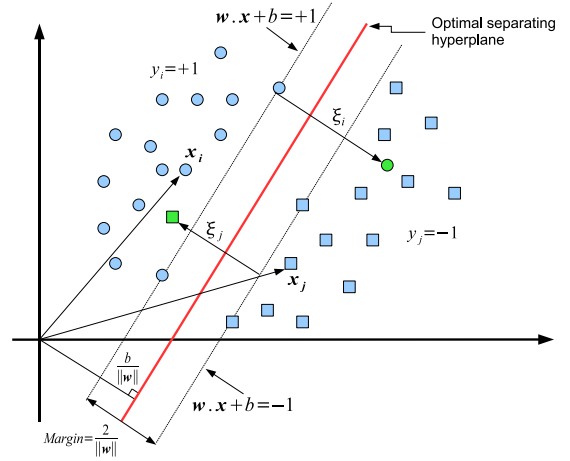
Fig. 1. Classification of non-linearly separable case by SVMs. There is one non separable vector in each class.

It leads to a new version of (6) where the scalar product is now: $\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$. Hopefully, for some kernels function k, the extra computational cost is reduced to:

$$\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle = k(\mathbf{x}_i, \mathbf{x}_j). \tag{8}$$

The kernel function k should fulfill Mercers' conditions [6]. Using kernels, it is possible to work implicitly in F while all the computation are done in the input space.

Classical kernels in remote sensing are the polynomial kernel and the Gaussian radial basis function:

$$k_{poly}(\mathbf{x}_i, \mathbf{x}_j) = [(\mathbf{x}_i \cdot \mathbf{x}_j) + 1]^p. \tag{9}$$

$$k_{gauss}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left[-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2 \right]. \tag{10}$$

C. Multiclass SVMs

SVMs are designed to solve binary problems where the class labels can only take two values: ± 1. For a remote sensing application, several classes are usually of interest. Various approaches have been proposed to address m-class problems [6]. They usually split the problem into a set of binary classifiers before combining them. The *one against all* classification strategy splits the problem into m binary sub-problems (class 1 against the others, class 2 against the others ...). The selected class is the one which gets the highest positive result. The *one versus one* classification strategy creates $m(m-1)/2$ binary sub-problems (class 1 against class 2, class 1 against class 3 ...). Then, results are combined following a majority voting scheme. This approach has shown to be more suitable for large problems [7]. Even though the number of the used classifiers is large, the whole classification problem is decomposed into much simpler ones. Therefore, this approach was used in our experiments.

III. DECISION FUSION

As explained in the previous section, the SVM's decision function returns the sign of the distance to the hyperplane. For the fusion scheme, it is more useful to have access to the belief of the classifier rather than the final decision [8]. For SVMs, it is possible to get the distance to the hyperplane, thanks to a simple change in (6). For a given sample, the more the distance to the hyperplane, the more reliable the label. That is the basis of the *one against all* strategy [6]. For the combination process, we choose to fuse this distance. We consider that the most reliable source is the one that gives the highest *absolute* distance.

In this paper, we first used the *absolute maximum* decision rule. For an m-source problem $\{S_1, S_2, \ldots, S_m\}$, where $S_1 = d_{ij}^1$ is the distance provided by the first SVM classifier which separates class i from j, this decision rule is defined as follows:

$$S_f = AbsMax(S_1, \ldots, S_m) \qquad (11)$$

where $AbsMax$ is the set of logical rules:

$$
\begin{aligned}
&\textbf{if}(|S_1| > |S_2|, \ldots, |S_m|) \quad \textbf{then} \quad S_1 \\
&\textbf{else if}(|S_2| > |S_1|, \ldots, |S_m|) \quad \textbf{then} \quad S_2 \\
&\qquad\qquad\qquad \vdots \\
&\textbf{else if}(|S_m| > |S_1|, \ldots, |S_{m-1}|) \quad \textbf{then} \quad S_m.
\end{aligned}
\qquad (12)
$$

The second operator considered takes into account the agreement of the classifier. Each distance is multiplied by the maximum probability associated to the two considered classes. Then, the absolute maximum is used to fuse the results. The probabilities are simply computed by [9]:

$$p_i = \frac{2}{m(m-1)} \sum_{j=0, j \neq i}^{m} I(d_{ij}) \qquad (13)$$

where I is the indicator function $I(x) = 1$ if $x \geq 0$ else $I(x) = 0$. For the fusion, the absolute distance is used as in (11), where source S_k is weighted by the corresponding p^k:

$$S_f = AbsMax\left(\max(p_i^1, p_j^1)S_1, \ldots, \max(p_i^m, p_j^m)S_m\right). \qquad (14)$$

The third operators is the one that is used to combine classifiers in the *one versus one* strategy. If we have two SVM classifiers, and apply each of them on a datasets with the same number of classes, each classifier builds $m(m-1)/2$ binary classifiers and uses majority voting. Thus, we propose to build a new set of classifiers, containing $m(m-1)$ classifiers. Then, we apply a classical majority voting scheme.

Finally, the fusion is done as follows. First we extract for each classifier the distance to the hyperplane for each sample. Then, using one of the three operators, the data are fused. For the operators based on absolute maximum, a majority voting is done. For all the operators the winning class is selecting as the one which has the highest number of votes.

No	Class Name	Samples Train	Test
1	Asphalt	548	6304
2	Meadow	540	18146
3	Gravel	392	1815
4	Tree	524	2912
5	Metal Sheet	265	1113
6	Bare Soil	532	4572
7	Bitumen	375	981
8	Brick	514	3364
9	Shadow	231	795
	Total	3921	400002

IV. EXPERIMENT

The proposed approach has been tested on real hyperspectral data. The image data were collected by means of the Reflective Optics System Imaging Spectrometer (ROSIS-03) optical sensor. The flight over the University of Pavia, Italy, was operated by the Deutschen Zentrum fur Luft - und Raumfahrt (DLR, the German Aerospace Agency) in the framework of the HySens project, managed and sponsored by the European Union. According to specifications the number of bands of the ROSIS-03 sensor is 115 with a spectral coverage ranging from 0.43 to 0.86 μm. The data are of very fine spatial resolution (1.3m per pixel). The image used here was 610×340 pixels and 9 classes were defined, and 103 bands were available. The original image is shown in Fig. 2.(a). See Table I for a description of the class of interest and of the training and testing set. 3 principal components were selected and the morphological profile was made of 10 openings/closings by reconstruction. The image was first classified with the spectral data (103 bands) and then with the EMP (63 bands). Gaussian kernels were used for each experiment. The parameters (C, γ) of the SVMs were tuned using a five-fold cross validation. The results were combined following the classification scheme previously defined. The accuracies in terms of classification are listed in Table II. The overall accuracy (OA) is the percentage of correctly classified pixels whereas the average accuracy (AA) represents the average of class classification accuracies. Kappa coefficient is another criterion classically used in remote sensing classification to measure the degree of agreement and takes into account the correct classification that may have been obtained "by chance" by weighting the measured accuracies. *Per Class* classification accuracy has been also reported. Classification map for the absolute maximum fusion operator is presented Fig. 2.(b).

As can be seen from the Table, the fusion step using absolute maximum improves the classification accuracies. The highest overall accuracies as well as the highest average accuracies and the highest Kappa value were achieved when the absolute maximum and probability were used conjointly. By comparing the global accuracies (OO, OA, Kappa), it is clear that the use of probabilities does not help so much in the fusion process of these

	Spect.	PCA+EMP	Abs. Max.	A.M.+Prob.	Maj. Vot.
OA	80.99	85.22	89.56	**89.65**	86.07
AA	88.28	90.76	93.61	**93.70**	88.49
Kappa	76.16	80.86	86.57	**86.68**	81.77
Class 1	83.71	**95.36**	93.18	93.02	93.98
Class 2	70.25	80.33	83.89	83.96	**85.34**
Class 3	70.32	**87.61**	82.13	82.23	64.94
Class 4	97.81	98.37	**99.67**	**99.67**	**99.67**
Class 5	99.41	**99.48**	**99.48**	99.41	**99.48**
Class 6	**92.25**	63.72	91.21	91.83	61.55
Class 7	81.58	**98.87**	96.99	97.22	93.01
Class 8	92.59	95.41	96.39	96.41	**98.83**
Class 9	96.62	97.68	**99.58**	**99.58**	**99.58**

experiments. The use of the majority voting rule does not improve the results compared to those obtained with the EMP. Regarding the per class accuracies, it is interesting to note that the best per class results are *only* in three cases provided by the normalized absolute maximum rule. However, all the accuracies are higher than 82% and for each class, the accuracy is close to the highest obtained accuracy. Regarding the computing time, the majority voting is the combination rule that leads to the shortest processing while the absolute maximum approach requires slightly more time. Assessing the probabilities increases the computing time.

V. CONCLUSION

Decision fusion for SVMs classifier has been discussed. Three operators based on the main characteristics of the outputs of SVMs were proposed. The operators were based on the assumption that the absolute distance to a hyperplane gives good information about agreement of classifiers. In experiments, the proposed approach outperformed each of the individual classifiers in terms of overall accuracies. The use of the absolute maximum operator lead to a significant improvement in terms of classification accuracy.

It is noteworthy that other operators are able to use sign as an informative feature. The classical *mean* or MYCIN rules [5] are examples of possible operators. Unfortunately, for a two-source problem, such operators have the same influence on the sign of the fused data as the absolute maximum. Thus, in our case (majority voting) lead to the same results.

In this paper, only one type of kernel was used. One possible extension of the proposed method is to include other sources using different kernels. Polynomial kernels, which are known to perform well on complex data, could be investigated.

The good performance of the proposed combination scheme is interesting because it uses no information about the reliability of the source. A topic of a future research is to use a more advance fusion scheme, that takes into account the performance of the classifiers such as in [8].
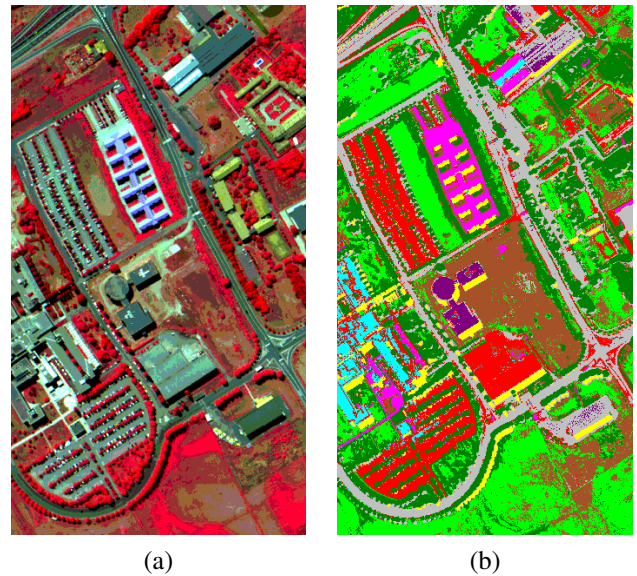

(a)   (b)

Fig. 2. Rosis University Area. (a): false colors original image, (b): classification map. Classes description: asphalt, meadow, gravel, tree, metal sheet, bare soil, bitumen, brick, shadow.

REFERENCES

[1] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. New Jersey: John Wiley and Sons, 2003.

[2] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sensing*, vol. 42, no. 3, pp. 480–491, Mar. 2005.

[3] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sensing*, vol. 42, pp. 1778–1790, Aug. 2004.

[4] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "Evaluation of kernels for multiclass classification of hyperspectral remote sensing data," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, May 2006.

[5] I. Bloch, "Information combination operators for data fusion: A comparative review with classification," *IEEE Trans. Syst., Man, Cybern. A*, vol. 26, no. 1, pp. 52–67, Jan. 1996.

[6] B. Scholkopf and A. J. Smola, *Learning with Kernels, Support Vector Machines, Regularization, Optimization and Beyond*. Cambridge: MIT Press, 2002.

[7] C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Networks*, vol. 13, pp. 415–425, Mar. 2002.

[8] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "Decision fusion for the classification of urban remote sensing images," *IEEE Trans. Geosci. Remote Sensing*, Accepted for publication.

[9] T. Wu, C. Lin, and R. Weng, "Probability estimates for multiclass classification by pairwise coupling," *Journal of Machine Learning*, vol. 5, pp. 975–1005, Aug. 2004.