Accepted for presentation at
The Third Scandinavian Conference on Image Analysis,
July, 12.-14.1983, Copenhagen, Denmark.

CLASSIFICATION OF MULTISPECTRAL SCANNER DATA BY LEARNING SUBSPACES

Kari Eloranta and Kai Mäkisara
Department of Technical Physics
Helsinki University of Technology
SF-02150 Espoo 15, Finland

INTRODUCTION

Thematic classification of remote sensing data often results in poor or moderate results even if the classifier were considered excellent. The importance of context (i.e. the neighbourhood of the pixel) in classifying individual pixels has been admitted but it has not been used due to excessive computational costs. This paper presents an application where we have used the Learning Subspace Classifier (LSM) introduced by Kohonen [1,2] for classification of satellite data. With LSM we achieve excellent classification accuracy with low computational cost at recognition phase which allows us to exploit also contextual information.

There are two quite independent problems discussed in this paper. The first is the LSM-classifier and its advantages in this application. The second is the use of contextual information. These problems are separate because the LSM-classifier simply processes vectorial data. All added information must be included in the vectors to be classified and this problem is independent of the classifier.

Some research groups have attempted to use contextual information in classification of multispectral data. There have been two major ways of doing this. The first one is the augmentation method used in this paper. Another method is to perform first classification and then try to improve the results by examining neighbouring classifications. This has been performed e.g. by relaxation methods.

This paper concentrates mostly on classification of multispectral data obtained from the spectral scanners carried by a satellite or an aeroplane. Most of the techniques mentioned in this paper are also applicable to other kinds of multispectral images, e.g. colour pictures of scenes.

THE SUBSPACE METHOD OF CLASSIFICATION

The subspace methods of classification are based on the assumption that the observations used in classification can be characterized as vectors (x) in a vector space ($\mathcal{L}$). This assumption is valid e.g. for spectral representations of signals. It is further assumed that the observation vectors from one class (C_i) are principally restricted to a subspace ($\mathcal{L}_i$) of the vector space. This leads to the classification rule:

assign x to class C_i if $\|P^i(x)\| \geq \|P^j(x)\| \quad j \neq i$

where $P^j(x)$ is projection of x on $\mathcal{L}_j$

The subspace classifiers are computationally efficient because the dimensionality of the class subspaces is usually low. Computation of the projections needed for the decision rule thus involves a small number of vector inner products and their squared sums which is a simple task for modern processors. In practice no such decisive limitations of dimensionality exist as in case of e.g. the Bayes classifier.

The differences between different subspace methods are in determination of the subspaces representing each class according to a set of teaching vectors. The "original" subspace method of Watanabe (CLAFIC, [3]) uses estimates of the class correlation matrices $C^{(i)} = E\{xx^T, \quad x \in C_i\}$. The largest eigenvalues of the correlation matrices are contributed by the principal factors of the vector distributions in C_i and so class i is represented by m_i eigenvectors corresponding to the largest eigenvalues. This choice minimizes the mean square error when the training data is represented by an m_i-dimensional subspace. The representation is statistically optimal but this does not guarantee minimal classification risk. This is especially true if the class distributions are irregular (e.g. skew or non-Gaussian).

The Learning Subspace Method invented by Kohonen [1,2] solves this problem by iteratively modifying class subspaces when misclassifications occur. The modification is performed by "rotating" the subspaces (i.e. their basis vectors) with matrix operator $(I-\alpha xx^T)$ where I stands for the identity matrix and $\alpha \in R$ for a weight factor regulating the amount of rotation. When positive the operator rotates the basis vectors towards the sample and when negative away from it. Details of the learning process and choice of are given in [2]. The number of iterations required in teaching depends somewhat on the data. Most test results show, however, that 5-20 cycles is enough.

ON REMOTE SENSING DATA

The poor classification results of remote sensing data are mainly due to the unsatisfactory statistical properties of the pixel vectors. These properties may not be well known due to lack of or small number of reliable ground truth observations. For instance atmospheric disturbances cause noticeable inhomogeneity in most large area pictures scanned from high altitudes. These facts suggest that some kind of local features (e.g. texture parameters) might be helpful in improving of classification performance. Isotropy of the picture is a nonsatisfied assumption even in a scene in state of nature. Therefore the operators extracting the neighbourhood features should be rotation invariant. Because the main goal is to classify the terrain inside the pixel area the pixel vector should not be completely neglected in forming of the feature vector. This was also confirmed by our experiments.

The behaviour of the feature vectors near border of a homogenous area should also be considered. Here the local neighbourhood of the pixel contains both information about the correct class of the pixel area and from classes of the neighbouring areas. Notice that when the feature vector can be considered as a linear combination of the feature vectors of each class the subspace classifier naturally separates the misleading information from the correct information.

AUGMENTED LOCAL FEATURES

All features tested were chosen to be radially symmetric for the reasons above. They consisted of functionals f on circles of different radius around the pixel. The area used will be called the set A and for simplicity the points are indexed with a single index. The fact that in most cases the correlation between points in a picture is an exponentially decaying function of their mutual distance suggests that the radius should be quite short. Local statistical properties as well as the form of f have effect on the optimal radius and defining of it has required an extensive set of classification experiments.

The form of the function f has mostly been restricted to a ℓ^p-function:

$$[x_A]_j = \{ \sum_i w(r) [x_{Ai}]_j^p \}^{1/p} \qquad (1)$$

Here w(r) is a piecewise constant radial weight function and the sum of the weights in A is normalized to one. The positive exponent p determines sharpness of the operation. When it is an integer the expression is the (1/p)th power of the pth statistical moment of the data in the neighbourhood. The practical experiments

showed that only small values of p or 1/p were useful. Dispersion functions have also been tried, e.g. forms like:

$$[\sigma_A]_j = \{ \sum_i [[x_{Ai}]_j - [\bar{x}_A]_j]^2 \}^{1/2} \qquad (2)$$

or

$$[\sigma_A]_j = \{ \sum_i |[x_{Ai}]_j - [\bar{x}_A]_j| \} \qquad (3)$$

where the mean value of x_i in A is denoted by $\bar{x}_A$. In practical experiments these dispersion features didn't prove to be successful at all.

Combination of x and x_A is done by concatenation which is suitable for subspace classifiers. Although it doubles the dimension of the pattern vectors producing $x' = [x^T \mid \beta x_A^T]^T$ the optimal dimensions of the subspaces still remain low and the computational effort is moderate. The scalar parameter β is of crucial importance. It determines the overall neighbourhood weight (whereas w(r) distributes it). Concatenation instead of some kind of superposition of x and x_A also enables good control over pixel and neighbourhood information separately.

An additional augmentation can be performed in order to improve the distribution of the once preprocessed feature vectors x'. The vectors corresponding to different classes can differ either in direction or in norm (or both). The subspace classifiers (like other norm-invariant classifiers) detect only differences in direction. Differences in norm can, however, be converted to differences in direction by augmenting the vectors with a simple scalar-valued augmentation function $g(\|x'\|)$. In two dimensions this can be visualized by considering two vectors inseparable by anything but their norms (fig. 1). By a suitable function of the norm we can map the one-dimensional vectors in two dimensions so that the angle between them becomes nonzero.

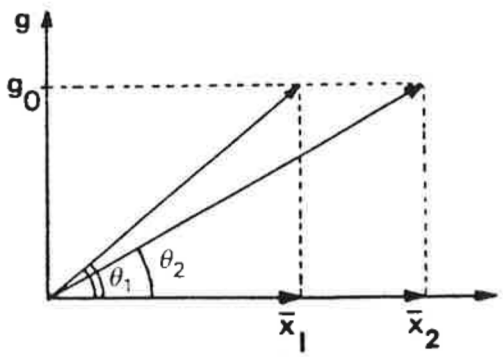

Figure 1.

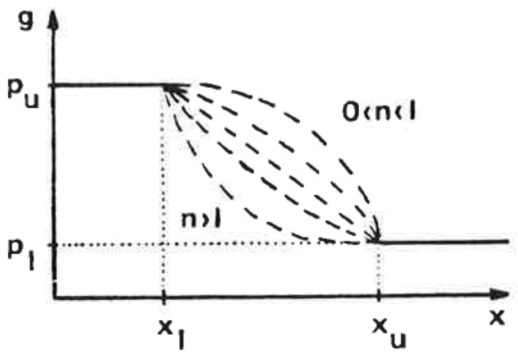

Figure 2.

Different forms of the function $g(\|x'\|)$ have been analyzed. Theoretically a piecewise concave one would be optimal in sense of producing largest uniform separation in angle. In practice the statistics of the nonseparable classes determine the form and steepness of the function. Fig. 2 illustrates the family of functions investigated most thoroughly:

$$g(x) = \begin{cases} p_u & x < x_1 \\ (p_u - p_1)\left(\dfrac{x_u - x}{x_u - x_1}\right)^n + p_1 & x_1 \leq x \leq x_u \\ p_1 & x_u < x \end{cases} \qquad (4)$$

For a constant g the optimum value is $p_u = p_1 = E\{\|x'\|\}$ over A. It is also the optimal mean value of p_1 and p_u in any case. The extreme values of the curved section (x_1 and x_u) should be determined so that $x_u - x_1$ is minimal. Classes with high norm deviation usually separate correctly without additional augmentation.

EXPERIMENTAL RESULTS

A series of classification experiments has been made using a LANDSAT-image. The picture is from wooded areas in state of nature except for some timber cutting squares and some lake and marsh areas. On basis of ground truth observations twelve classes were defined. Seven of them were clearly forest and the rest consisted of intermediate types between forest, bare areas, and marsh. Water created one class. The classes consisted of a total of 632 pixels which was divided in learning and test sets on the average proportion of three to two. The pixel vectors had four spectral components and especially the two components corresponding to visible light had poor dynamics.

The classification results obtained are summarized in table 1. Maximum likelihood (ML) classifier was used as reference method for CLAFIC and LSM. The four-dimensional data was the unpreprocessed material. The eight-dimensional data was created by using three to five adjacent circles superposed together with equal weights to form four augmentation components. It was discovered that the results had a maximum when the neighbourhood annulus was at the average distance of four to five. Optimal augmentation weight β seemed to depend linearly on the number of circles used and usually its value was between 3 and 7. The nine-dimensional data was obtained by augmenting the best eight-dimensional data with the scalar function g. The parameters of this function were not critical.

dimension of data	classification results ML	CLAFIC	LSM(learning material)	LSM(test material)
4	72%	56%	-	-
8	97%	89%	92%	88%
9	-	93%	96%	95%

Table 1. Typical classification results with learning material and test material (LSM).

The results in table 1 have been mostly obtained with learning data and so they describe the relative efficiencies of the classifiers. Results with independent test material have been obtained for LSM which was the method studied primarily.

The results show that ML-classifier gives slightly better results than LSM but the computational effort is approximately fourfold. In this connection it ought to be mentioned that by using closed disks (all circles 1...n with equal weights) instead of annuli similar results were reached. Computational aspects may prefer use of closed disks because a fast recursive preprocessing algorithm can be used [4]. In general the computational load of this kind of preprocessing with p equal to one or two is only a fraction of that of classification.

ACKNOWLEDGMENT

We want to thank prof. Teuvo Kohonen for the discussions and support leading to this study. We also want to thank the Laboratory of Land Use of Technical Research Center of Finland for providing the LANDSAT data with ground truth.

REFERENCES

[1] Kohonen, T., Nemeth, G., Bry, K.-J., Jalanko, M., and Riittinen, H., Spectral classification of phonemes by learning subspaces, Proc. 4th ICPR, Kyoto, Japan, 1978.

[2] Kohonen, T., Riittinen, H., Jalanko, M., Reuhkala, E., and Haltsonen, S., A thousand-word recognition system based on the Learning Subspace Method and Redundant Hash Addressing, Proc. 1981 IEEE ICASSP, Atlanta, GA, USA, 1981, pp. 975-978.

[3] Watanabe, S., Lambert, P.F., Kulikowski, C.A., Buxton, J.L., and Walker, R., Evaluation and selection of varibles in pattern recognition, in Computer and Information Sciences, vol. 2., ed. J.T. Tou, pp. 91-122, Academic Press 1967.

[4] Eloranta, K., Diploma Engineer's Thesis (in Finnish), Helsinki University of Technology, Dept. of Technical Physics, 1982.