

A NEW DECISION CRITERION FOR FEATURE SELECTION

Application to the Classification of Non Destructive Testing Signatures

(EUSIPCO European Signal Processing Conference. Rhodes. Greece. 1998)

Latifa OUKHELLOU*, Patrice AKNIN*, Hervé STOPPIGLIA**, Gérard DREYFUS**

* INRETS, LTN - 2 Avenue Malleret-Joinville. 94114 Arcueil. France
Tel : 33 1 47 40 73 37; Fax : 33 1 45 47 56 06; e-mail : aknin@inrets.fr

** ESPCI, Laboratoire d'Électronique - 10 Rue Vauquelin. 75005 Paris. France
Tel : 33 1 40 79 45 41; Fax : 33 1 40 79 44 25; e-mail : Gerard.Dreyfus@espci.fr

ABSTRACT

This paper describes a new decision criterion for feature selection (or descriptor selection) and its application to a classification problem. The choice of representation space is essential in the framework of pattern recognition problems, especially when data is sparse, in which case the well-known *curse of dimensionality* appears inevitably [1]. Our method associates a ranking procedure based on Orthogonal Forward Regression with a new stopping criterion based on the addition of a random descriptor. It is applied to a non destructive rail diagnosis problem that has to assign each measured rail defect to one class among several ones.

1. INTRODUCTION

In pattern recognition problems, perception is generally performed by one or several sensors whose raw outputs are not appropriate as classifier inputs; a preprocessing must be performed in order to transform the data representation space and to make the assignment of each object (or signal) to one of several predetermined classes easier. Indeed, classification performances depend critically on the relative locations of the classes in representation space : if the classes are widely apart, the design of the classifier is easy and its performances are likely to be good. Therefore, the design of a pattern recognition system usually requires a tradeoff between preprocessing complexity and classifier complexity : a suitable preprocessing may greatly facilitate classification by making the design of the classifier simple.

The first stage of preprocessing is a parameterization procedure which replaces the signal by a set of descriptors, having hopefully a high discriminative power.

This parameterization procedure may produce a higher number of descriptors than necessary. Therefore, it must be followed by a reduction of the dimensionality of the input data. This second stage of preprocessing reduces the computing cost and the complexity of the classifier structure.

In this paper, we describe an original two-step method for dimensionality reduction. In the first step, the descriptors

are ranked in order of decreasing relevance, as estimated from the contribution of each input to the classifier outputs. This is achieved by an Orthogonal Forward Regression procedure. The second step determines the final number of relevant descriptors, using a new decision criterion based on the addition of a random descriptor.

2. DESCRIPTOR RANKING

The selection of an optimal subset of descriptors among a larger set of candidate descriptors requires either an exhaustive search, or the use of a "branch and bound" algorithm [2], which are generally too costly (2^p possible subsets for a p parameter problem). Therefore, we have to resort to seeking a satisfactory, but possibly suboptimal solution.

The ranking method that we use is inspired from linear regression. Each observation is described by p descriptors $[x_{1i}, x_{2i}, \dots, x_{pi}]$ and one output. For example, in a two-class problem, the output is set to 0 or 1 depending on the class.

For a given set of n observations, the complete linear regression model can be written in matrix form :

$$\underline{Y} = \underline{X}\underline{P} + \underline{\varepsilon} \quad \text{with}$$
$$\underline{X} = \begin{bmatrix} \underline{X}_1^t & \underline{X}_2^t & \dots & \underline{X}_p^t \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{p1} \\ x_{12} & & \\ \vdots & & \\ x_{1n} & & x_{pn} \end{bmatrix} \begin{matrix} \leftarrow \text{observations} \\ \\ \\ \uparrow \text{descriptors} \end{matrix}$$

\underline{X}_i is the n -vector whose components are the value of descriptor i for each element of the data set.

$\underline{Y} = [y_1 \ y_2 \ \dots \ y_n]^t$ is the n -vector of outputs

$\underline{\varepsilon} = [\varepsilon_1 \ \varepsilon_2 \ \dots \ \varepsilon_n]^t$ is the residual modeling error.

\underline{P} is the p -vector of parameters; if it is the least-squares solution of the regression problem, it is the orthogonal projection of the output vector onto the subspace generated by the n observations.

Thus, descriptor ranking is carried out according to the contribution of each descriptor to the model output. However, because of the correlation between descriptors, their own contribution cannot be estimated directly: The ranking procedure must include an orthogonalization step that ensures an independence of the descriptors.

An adaptation of the Gram-Schmidt orthogonalization procedure allows us to iteratively rank the p descriptors [3]:

- at the first step, the relevance of each descriptor \underline{X}_i is estimated by measuring the angle between itself and the model output. The descriptor with the lowest absolute angle (maximum cosines) is ranked first :

$$\underline{X}_{i_1} \text{ such as } \cos(\underline{X}_{i_1}, \underline{Y})^2 = \max_{1 \leq i \leq p} \left[\frac{(\underline{X}_i^t \cdot \underline{Y})^2}{\|\underline{X}_i\|^2 \|\underline{Y}\|^2} \right]$$

- at the second step, the remaining descriptors, and the output, are orthogonalized with respect to the first descriptor \underline{X}_{i_1} :

$$\underline{X}_k^{(2)} = \underline{X}_k - \frac{\underline{X}_{i_1}^t \underline{X}_k}{\underline{X}_{i_1}^t \underline{X}_{i_1}} \underline{X}_{i_1} \quad \text{and} \quad \underline{Y}^{(2)} = \underline{Y} - \frac{\underline{X}_{i_1}^t \underline{Y}}{\underline{X}_{i_1}^t \underline{X}_{i_1}} \underline{X}_{i_1}$$

The procedure is continued by choosing the descriptor which has the lowest angle with the projected output.

$$\underline{X}_{i_2} \text{ such as } \cos(\underline{X}_{i_2}, \underline{Y}^{(2)})^2 = \max_{1 \leq i \leq p-1} \left[\frac{(\underline{X}_i^{(2)t} \cdot \underline{Y}^{(2)})^2}{\|\underline{X}_i^{(2)}\|^2 \|\underline{Y}^{(2)}\|^2} \right]$$

- The procedure terminates when all descriptors are ranked.

Note that this method is applicable to any model which is linear with respect to its parameters, but not necessarily linear with respect to the descriptors, since the X 's may be nonlinear functions of the descriptors.

3. DECISION CRITERIA : CHOICE OF THE SUBSET OF DESCRIPTORS

3.1 Principle of the method

Once the descriptors are ordered, the dimension p_r of the descriptor subset must be defined. Usual criteria such as the Akaike Information Criterion can be used [4]. However, these criteria are efficient only if the number n of observations is much larger than the number p of candidate descriptors.

A new criterion, based on the addition of a random descriptor, has been proposed [5]. The basic idea is to rank all descriptors, including a random one, with the orthogonalization method. Then, the descriptors which are ranked below the random descriptor are considered as irrelevant. In practice, the rank of a random descriptor is itself a random variable, so that it is necessary to generate a large number of random descriptors in order to make a good estimation of the cumulative distribution function of its rank.

Then, an hypothesis test is used to set the subspace dimension p_r . Considering the null hypothesis "the random descriptor rank is higher than p_r " and the alternative hypothesis "the random descriptor rank is lower than p_r ". We have chosen to set the significance level α associated to the following event : the null hypothesis is retained whereas the alternative hypothesis is true. The subspace dimension is derived directly from the equation $\mathbf{F}_p = \alpha$ where \mathbf{F}_p is the cumulative distribution function of the random descriptor rank; so, the probability that a random descriptor is more relevant than one of the p_r descriptors ranked first is $\alpha\%$ (figure 1).

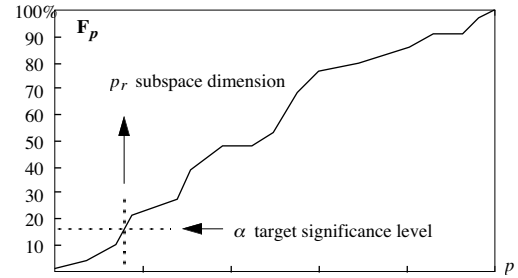


Figure 1 : Cumulative distribution function of random descriptor rank and choice of subspace dimension

3.2 Practical implementation

The implementation of the decision criterion described above involves the repeated use of the Gram-Schmidt procedure for each random descriptor generated. This can be too costly, especially if the candidate descriptor number is high.

One way to reduce the amount of computation consists of generating a large set of random descriptors and ranking all of them in a single Gram-Schmidt procedure.

We start with two sets of descriptors :

- the original candidate descriptors $\{\underline{X}_i\}_{1 \leq i \leq p}$
- the random descriptors $\{\underline{R}_i\}_{1 \leq i \leq p_a}$ with $p_a \ll p$

At each step k of the ranking procedure, the best descriptor \underline{X}_{i_k} is selected among the candidate descriptors; then, we determine the set A of random descriptors whose contribution to the output is more relevant than the \underline{X}_{i_k} contribution. The ratio of the cardinal of A to the total number p_a of random descriptors, is an estimation of the probability density function f_k of the event : a random descriptor is more relevant than the k th selected descriptor. The probability \mathbf{F}_k that a random descriptor is more relevant than one of the k descriptors ranked first is simply derived from the cumulative sum :

$$\mathbf{F}_k = \sum_{i=1}^k f_i$$

For the next step, the random descriptors belonging to A are deleted from the random descriptor set; the remaining random descriptors and the unranked non-random descriptors

are orthogonally projected onto the subspace orthogonal to the k th selected descriptor. The procedure terminates when all random descriptors are ranked. Figure 2 summarizes the algorithm implementation.

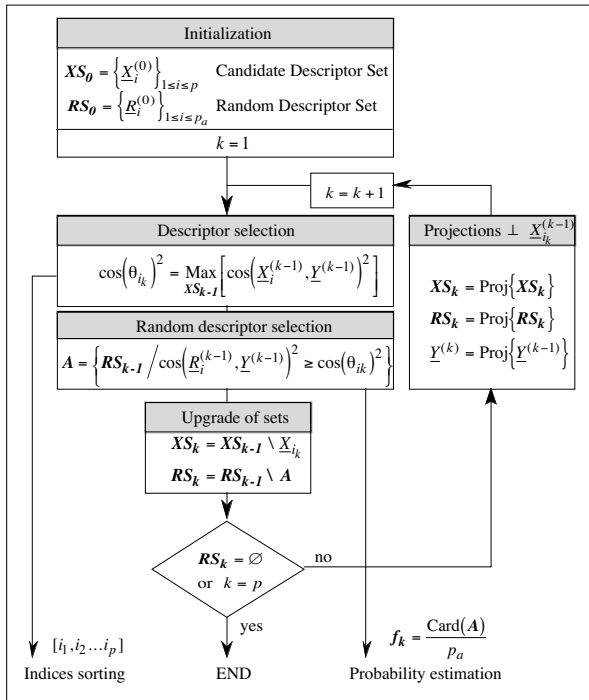


Figure 2 : Algorithm chart

4. APPLICATION TO THE CLASSIFICATION OF 2D ELECTROMAGNETIC SIGNATURES

4.1 Presentation of the application

This method has been applied to the descriptor selection of eddy current signatures in the railway domain. In this application, an electromagnetic multisensor detects rail defects, and each defect must be assigned to one of 4 classes : external cracks, welded joints, joints before switch and spallings. A data base of 140 defects was available for the design of the classifier [6].

Four complex signals are delivered by the device and a modified Fourier parameterization procedure generates about 100 descriptors [7]. Because of noise, descriptor redundancy, and data sparsity, a classifier cannot take into account all descriptors. It is therefore necessary to select the most relevant descriptors among the initial ones.

4.2 Partitioning approach to classification

This selection was carried out with a partitioning approach of the K-class initial problem. In fact, the linear regression modelling is well suited to a 2-class problem. Therefore, we choose to split the global classification task into several elementary tasks [8]. One way of doing this is to generate K sub-problems; each of them being the separation of one class from all others (figure 3); each classifier delivers an

estimation of the probability of the unknown pattern to belong to a given class.

We design each sub-classifier as very simple neural network consisting of a single neuron with a sigmoidal activation function. Thus, each sub-classifier performs a linear separation of input space into two regions.

This approach has the following advantages: (i) the feature selection can be performed independently for each sub-problem, so that each classifier has only the specific inputs that are relevant to the problem that it has to solve, (ii) the learning phases of each sub-classifier are independent, which gives us a better control than a global learning.

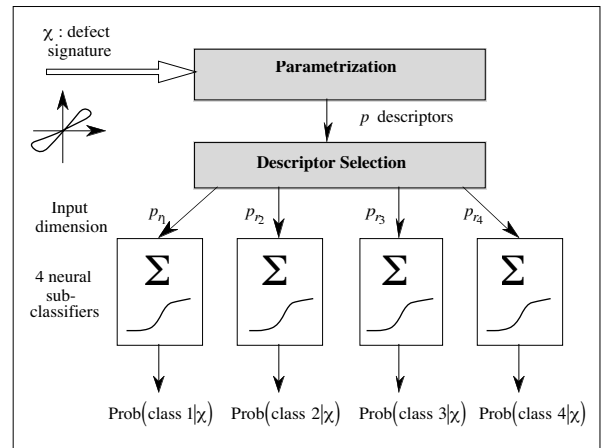


Figure 3 : Partitioning approach to classification

4.3 Results

Figure 4 presents the evolution of the squared cosines between the ordered descriptors and the output of the first sub-classifier dedicated to the separation of the first class from the others. The ranking was carried out with the modified Gram-Schmidt procedure.

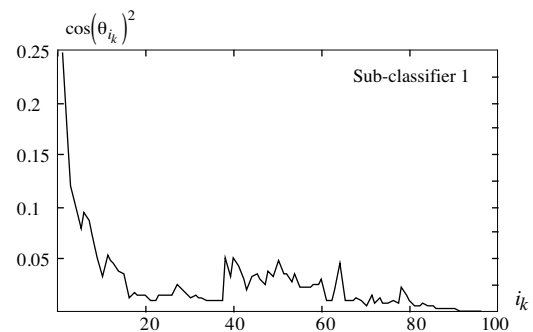


Figure 4 : Evolution of squared cosines during the Gram-Schmidt orthogonalization procedure

Note that the course of the squared cosines as a function of the subspace descriptor dimension is not monotonous decreasing. Because of the successive projections operated by the Gram-Schmidt orthogonalization procedure, the k th cosine is estimated in a $(n-k)$ dimension subspace. At each

step, the working subspace is different; therefore, the successive cosine values cannot be compared.

The feature selection has been carried out as indicated in section 3. Figure 5 presents the probability density function and the cumulative distribution function of the first sub-problem. With about 10% of significance level, the reduction of dimensionality is very important since the final input dimension decreases to $p_{r1}=15$ compared to the initial dimension $p=100$.

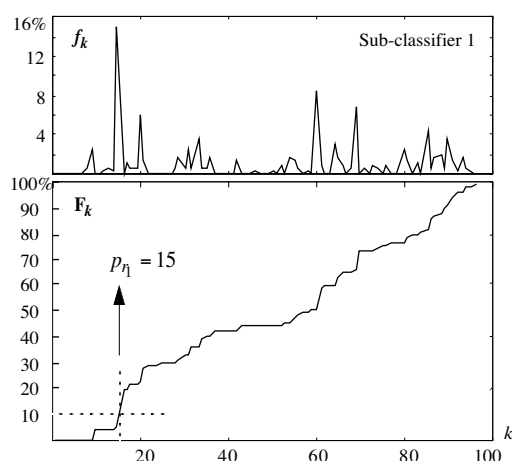


Figure 5 : Probability density function and cumulative distribution function of random descriptor rank

For each sub-classifier, such a small subset of the original descriptors was determined by our selection procedure. We find respectively 15, 15, 8 and 9 descriptors with a significance level $\alpha=10\%$.

This feature selection leads to performances up to 95% of correct classification with these elementary sub-classifiers. Obviously, the feature selection with random descriptor addition can be applied to more complex classifier structures [9].

5. CONCLUSION

The curse of dimensionality is a major problem when designing statistical classifiers (neural networks or other) from small data bases : the dimension of the classifier input must be much lower than the number of training patterns in order to produce statistically significant decision boundaries.

The new decision criterion proposed in this paper is an attractive way of building models with the smallest possible number of inputs; it has two attractive features : it is an intuitive - but principled - approach to the descriptor selection procedure, and its computer implementation is simple.

The present paper has described an application of this method to a classification problem with high performances, but the method is of general relevance to any nonlinear

modeling problem; the procedure has also been applied successfully to the modeling of chemical properties of molecules [10].

Keywords :

Feature Selection, Classification, Pattern Recognition, Gram-Schmidt Orthogonalization, Akaike Information Criterion, Neural Networks, Non Destructive Evaluation

References :

- [1] : See for instance: Bishop C.M. *Neural Networks for Pattern Recognition*. Ed Clarendon Press. 1995
- [2] : Kittler J. *Feature Selection and extraction. Handbook of pattern recognition and image processing*. Ed. Academic Press. 1986
- [3] : Chen S. Billings S.A. Luo W. *Orthogonal least squares methods and their application to non-linear system identification*. Int J. Control vol 50 n°5 pp. 1873-1896. 1989
- [4] : Akaike H. *A new look at the statistical model identification*. IEEE Trans. on Automatic Control. vol 19 n°6 pp.716-723. 1974
- [5] : Stoppiglia H. *Méthodes Statistiques de Sélection de Modèles Neuronaux. Applications Financières et Bancaires*. Thèse de l'Université Pierre et Marie Curie, Paris. 1997.
- [6] : Oukhellou L. Aknin P. Perrin J.P. *Specific sensor and classifier of defaults of rail head for railway systems*. 8th IFAC Symposium on Transportation Systems. Chania. Grèce. 1997
- [7] : Oukhellou L. Aknin P. *Modified Fourier Descriptors : A new parametrization of eddy current signatures applied to the rail defect classification*. III intern. Workshop on Advances in Signal Processing for Non Destructive Evaluation of Materials. Quebec. 1997
- [8] : Price D. Knerr S. Personnaz L. Dreyfus G. *Pairwise neural network classifiers with probabilistic outputs*. Neural Information Proc. Systems. 1994
- [9] : Oukhellou L. Aknin P. *Optimization of radial basis function network for classification tasks*. Intern. Conference on Neural Networks and their Applications. Marseille. 1998
- [10] : Duprat A.F. Huynh T. Dreyfus G. *Towards a Principled Methodology for Neural Network Design and Evaluation in QSAR; Application to the Prediction of LogP*. J. Chem. Inf. Comput. Sci., in press (1998).

Acknowledgements :

The research concerning the Rail Defect Classification is supported by the French Research Ministry in the framework of PREDIT program.