KYBERNETIKA — VOLUME 34 (1998), NUMBER 4, PAGES 411 – 416

ON SELECTING THE BEST FEATURES IN A NOISY ENVIRONMENT

JAN FLUSSER[1] AND TOMÁŠ SUK

This paper introduces a novel method for selecting a feature subset yielding an optimal trade-off between class separability and feature space dimensionality. We assume the following feature properties: (a) the features are ordered into a sequence, (b) robustness of the features decreases with an increasing order and (c) higher-order features supply more detailed information about the objects. We present a general algorithm how to find under those assumptions the optimal feature subset. Its performance is demonstrated experimentally in the space of moment-based descriptors of 1-D signals, which are invariant to linear filtering.

1. INTRODUCTION

Selection of a subset of a large set of features which is "optimal" in some sense is an essential task on the field of pattern recognition. Usually there are two opposite requirements working against one another: the subset should be "small enough" to reduce significantly the dimension of the feature space and, on the other hand, it should provide sufficient object representation and discriminative power. There have been published numerous feature selection methods in the last three decades, we refer to classical monographs [1] and [2] for a survey.

The problem formulation we are dealing with in this paper is slightly different. We assume the features are *ordered* and having the following property: robustness to noise in the original data decreases with the increasing feature order whereas its ability to supply detailed information about the objects increases. Central moments of signals and images, all moment-based invariants and some of differential invariants and Fourier descriptors behave exactly in that way.

We consider only subsets formed by first p members of the feature sequence. Thus, the problem of the optimal feature selection is restricted to searching for an optimal order p_{opt}.

There have not been many publications devoted to the problem formulated above. Mostafa and Psaltis [3] and Teh and Chin [4] proved that image moments satisfy our assumption. Liao and Pawlak analyzed the one-class problem of noisy image

[1]This work has been supported by the grant No. 102/96/1694 of the Grant Agency of the Czech Republic and by the grant No. 4178–3 of the Ministry of Health.

representation by moments in [5] and [6]. They defined the optimal order p_{opt} according to the minimum-reconstruction-error criterion and they showed that such an optimum exists and that it decreases with increasing noise variance.

In this paper, we consider objects belonging to *different* classes. We try to optimize the *discrimination power* of the features (i.e. the separability of the classes) which obviously differs from its reconstruction ability.

2. A TWO–CLASS PROBLEM

Consider a problem of two pattern classes w_1 and w_2 with mean vectors $\mathbf{m_1}, \mathbf{m_2}$ and covariance matrices $\mathbf{C_1}, \mathbf{C_2}$, respectively, in a p-dimensional space of features $S_1, \ldots, S_p$. Define the *mean covariance matrix* $\mathbf{C}$ as

$$\mathbf{C} = P(w_1)\,\mathbf{C_1} + P(w_2)\,\mathbf{C_2},$$

where $P(w_i)$ is an a priori probability of w_i and assume that $\mathbf{C}$ is non-singular.

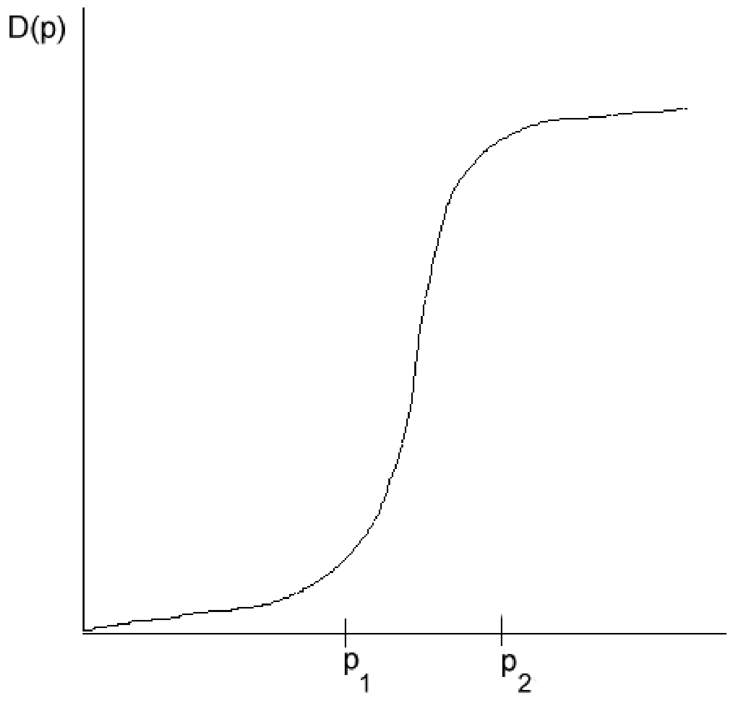

Fig. 1. Typical behaviour of the Mahalanobis distance between two classes. p_1 – threshold of separability; p_2 – threshold of noise sensitivity.

Under these conditions, we can take the *Mahalanobis distance*

$$D(p) = (\mathbf{m_1} - \mathbf{m_2})\mathbf{C^{-1}}(\mathbf{m_1} - \mathbf{m_2})'$$

as the measure of the class separability.

The typical course of the Mahalanobis distance under the assumptions given in Section 1 as a function of p is depicted in Figure 1. The features of order less than p_1 have not enough discriminative power to separate the given classes. The features of order greater than p_2 do not contribute significantly to the class separability because of their high sensitivity to noise. Thus, we consider $(S_{p_1}, \ldots, S_{p_2})$ as the optimal feature subset. Clearly, the values of p_1 and p_2 depend on the given classes. If the classes are "well-separable" by lower-order features, then p_1 is close to one. Similarly if they are "well-separable" by higher-order features, p_2 becomes very high. On the other hand, if the classes are "similar" to each other, then p_2 may be equal to p_1 and those classes may be non-separable by any feature subset.

In practice we do not search for p_1 because it is usually small and the dimensionality reduction by p_1 is not significant. Moreover, most features we are dealing with are calculated recursively. To calculate the feature S_p of the object, one has to know all lower-order features first.

3. AN OPTIMIZATION PROCEDURE

In this Section, we present a numeric algorithm which finds the optimal feature number defined above.

1. Inputs:
   $f_1^{(k)}, \ldots, f_{n_k}^{(k)}$ – training patterns of w_k, $k = 1, 2$,
   ε – user defined tolerance parameter.
2. Set $p = 1$; $ind = 0$;
3. for $k = 1 : 2$
   $$P(w_k) = n_k/(n_1 + n_2);$$
   end;
4. for $k = 1 : 2$
   Estimate $\mathbf{m_k}$ and $\mathbf{C_k}$;
   end;
5. $\mathbf{C} = P(w_1)\,\mathbf{C_1} + P(w_2)\,\mathbf{C_2}$;
6. $D(p) = (\mathbf{m_1} - \mathbf{m_2})\mathbf{C^{-1}}(\mathbf{m_1} - \mathbf{m_2})'$;
7. if $(p \leq 2)$ then goto Step 8;
   else if $(D(p) - D(p-2) \geq 2\varepsilon)$ then $ind = 1$;
   else if $(ind = 1)$ then $p_{opt} = p - 2$; STOP
   end;
   end;
   end;
8. $p = p + 1$;
   goto Step 4.

Step 7 is a key point of the algorithm. The logical variable *ind* tells us whether the current p is greater than p_1 ($ind = 1$). Without this indicator the algorithm would stop in most cases at the beginning giving a false result $p_{opt} < p_1$. The stop condition says that for two consecutive values of p the average increment of the Mahalanobis distance $D(p)$ must be less than ε.

4. NUMERICAL EXPERIMENTS

To demonstrate the previous considerations as well as the optimization procedure experimentally, we employ one kind of moment invariants of 1-D signals. (We could use, however, any features meeting our assumptions.)

In the following text, by *signal* we understand any absolutely integrable function $f(t)$ which is non-zero on bounded support and the integral of which is non-zero.

Let us define for any odd p the feature $S_p(f)$ by the following recursive formula:

$$S_p(f) = \mu_p^{(f)} - \frac{1}{\mu_0^{(f)}} \sum_{n=1}^{(p-1)/2} \binom{p}{2n} S_{p-2n}(f) \cdot \mu_{2n}^{(f)},$$

where $\mu_p^{(f)}$ is the central moment of the signal $f(t)$.

It was proved in [7] that each S_p is an invariant with respect to blur, that means its value does not change when the signal is convolved with any symmetric impulse response $h(t)$, i.e. $S_p(f) = S_p(f * h)$. Due to this property, these invariants are very powerful features for recognizing signals filtered by an unknown linear system [8].

First, let us demonstrate that the blur invariants do meet our assumption about the decreasing robustness. Let g be a noisy version of f, i.e. $g = f + n$, where n denotes zero-mean Gaussian noise. The robustness of the invariants can be characterized by their relative error

$$r_p = \frac{|S_p(g) - S_p(f)|}{S_p(f)}$$

(high relative error indicates low robustness).

The course of the mean value of r_p as the function of the order of the invariants is depicted in Figure 2. In this experiment, 100 realizations of the noisy signal were generated to estimate $E(r_p)$. Signal-to-noise ratio was always equal to 5 dB. The increasing character of $E(r_p)$ demonstrates the decreasing robustness of the invariants with respect to additive random noise.

Now two classes of 1-D digital signals are given, each of them containing 30 elements. We applied the above algorithm (with the tolerance parameter $\varepsilon = 0.05$) to find the optimal number of the blur invariants for class separation. Figure 3 shows the course of the Mahalanobis distance between the classes depending on the number of the invariants used. The algorithm yields the result $p_{opt} = 6$. Thus, the optimal subset consists of blur invariants S_3, S_5, S_7, S_9, S_{11} and S_{13} in this case (S_1 is useless because $S_1 = 0$ everywhere).

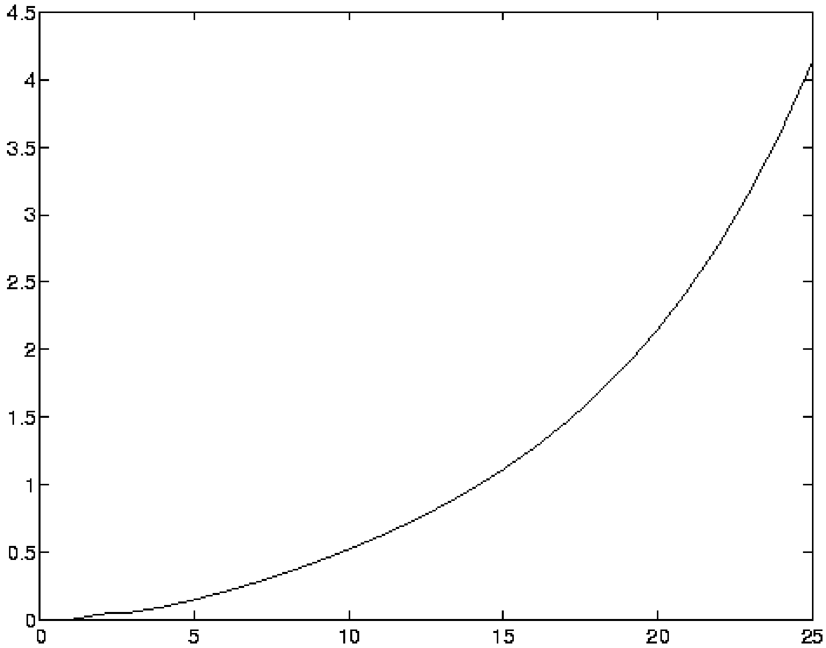

Fig. 2. Robustness of the blur invariants with respect to additive random noise. Horizontal axis: the order of the invariant; vertical axis: relative error (mean value over 100 runs).

5. CONCLUSION

In this paper, we have introduced a method for selecting the optimal feature subset for class separability in a noisy environment. The method works for any ordered feature sets which meet the two following assumptions: robustness of the features decreases with the increasing order and the higher-order features supply more detailed information about the objects. The performance of the presented algorithm has been demonstrated experimentally in the space of the moment-based descriptors of 1-D signals, which are invariant to linear filtering.

We have defined the optimal number of features p_{opt} as the highest order, which contributes "significantly" to class separability measured by Mahalanobis distance. If we use more than p_{opt} features, the class separability cannot be worse but computational cost increases. On the other hand, p_{opt} might be in some cases too high to employ all features up to the order p_{opt} in practice. Moreover, the algorithm may be numerically unstable for high p because covariance matrix $\mathbf{C}$ may become ill-conditioned. Thus, we should define some threshold values of p or of Mahalanobis distance $D(p)$. Once one of these thresholds is exceeded the algorithm should stop even if p_{opt} has not been reached.

(Received December 18, 1997.)

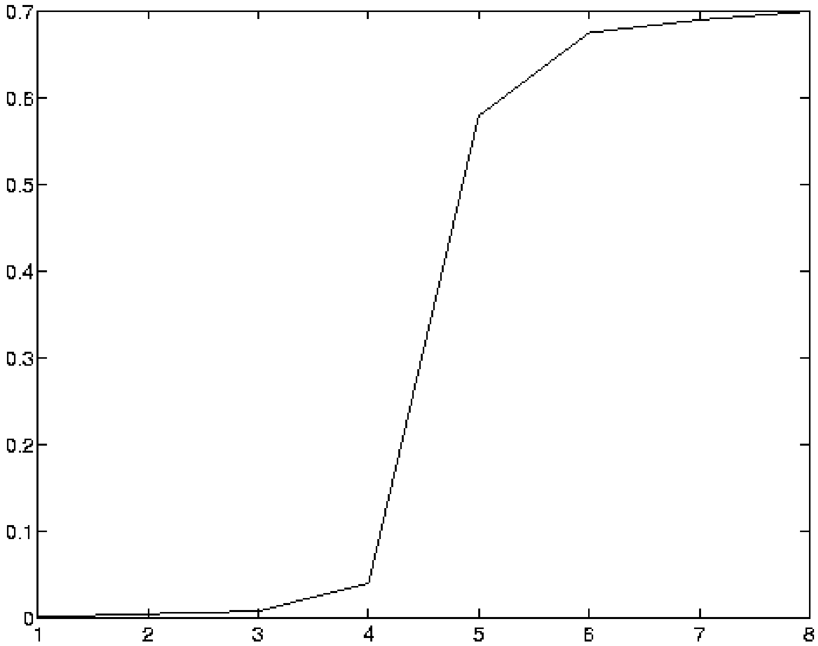

Fig. 3. The Mahalanobis distance in the space of blur invariants between two classes of signals depending on the order of invariants used.

REFERENCES

[1] K. Fukunaga: Introduction to Statistical Pattern Recognition. Academic Press, New York 1972.

[2] P. A. Devijver and J. Kittler: Pattern Recognition: A Statistical Approach. Prentice Hall, London 1982.

[3] Y. S. Abu–Mostafa and D. Psaltis: Recognitive aspects of moment invariants. IEEE Trans. Pattern Anal. Mach. Intell. *6* (1984), 698–706.

[4] C. H. Teh and R. T. Chin: On image analysis by the methods of moments. IEEE Trans. Pattern Anal. Mach. Intell. *10* (1988), 496–512.

[5] M. Pawlak: On the reconstruction aspects of moment descriptors. IEEE Trans. Inform. Theory *38* (1992), 1698–1708.

[6] S. X. Liao and M. Pawlak: On image analysis by moments. IEEE Trans. Pattern Anal. Mach. Intell. *18* (1996), 254–266.

[7] J. Flusser and T. Suk: Invariants for recognition of degraded 1-D digital signals. In: Proc. 13th ICPR, Vienna 1996, vol. II, pp. 389–393.

[8] J. Flusser and T. Suk: Classification of degraded signals by the method of invariants. Signal Processing *60* (1997), 243–249.

Ing. Jan Flusser, CSc. and Ing. Tomáš Suk, CSc., Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 182 08 Praha 8. Czech Republic.
e-mails: flusser@utia.cas.cz, suk@utia.cas.cz