

REGULARIZATION IN IMAGE RECOGNITION: THE PRINCIPLE OF DECISION RULE SMOOTHING

O. Seredin¹, V. Mottl²

¹ Tula State University, Tula, Russia,

² Computing Center of the Russian Academy of Science, Moscow, Russia

A technique of training regularization in image recognition is considered. The main idea underlying this investigation consists in overcoming the small sample size problem by adding an subsidiary a priori information on interrelation among the image elements.

1. Introduction

It is typical for data analysis problems that several features measured at an object are not independent of each other. Image and signal recognition are glowing examples of this kind of data analysis problems.

First, in this kind of problems, the number of features is essentially greater than the number of objects in a training set of typical size. Second, the specificity of feature registration usually implies their linear $\mathbf{x} = (x_t, t = 1, \dots, T)$, or spatial $\mathbf{x} = (x_{ts}, t = 1, \dots, T, s = 1, \dots, S)$ ordering. Third, the hypothesis that the features are “smooth” along the order axes is suitable in many cases, i.e., close values are typical for closely related features $(x_t, x_{t \pm \delta})$ or $(x_{ts}, x_{t \pm \delta', s \pm \delta''})$. As for signal analysis, smoothly ordered features are usually the result of measuring the same physical value along an axis with a sufficiently small step. In the case of images, smoothness of features implies minor difference between brightness values at adjacent elements of the pixel grid.

The large number of features leads to the small sample problem well known in data analysis – a decision rule which exactly separates the objects of the training set has poor extrapolation properties. Two ways of solving this problem may be distinguished, namely, reduction of the feature space dimensionality, usually, by feature selection [1], and imposing constraints on the class of decision rules as a means of training regularization [2]. In this paper, we keep to the latter approach. To improve the prediction power of a pattern recognition algorithm, we suggest to take into account the known specificity of the respective kind of objects, namely, the ordering of features and their smoothness.

In many algorithms, the result of training is the vector of coefficients of a discriminant hyperplane immediately in the linear space of features or in a secondary linearized space. The idea of regularization we propose in this paper is utilization of the fact that any coefficient of the discriminant hyperplane is associated with some feature. If the positions of two features in the structure of an object (immediately successive values of a signal or adjacent pixels of an image) are close to each other, the corresponding coefficients of the discriminant hyperplane must also not be too different. Thereby, among the entire set of decision rules correctly fitting the trainer’s data, we choose only the subset that satisfies some *a priori* constraint.

In paper [3], this regularization principle is applied to the signal recognition problem in terms of the popular support vector algorithm (SVM). In this paper, we evolve this technique as applied to images.

2. The idealized image recognition problem

The problem of face verification is a typical two-class images recognition problem. Some examples of normalized and scaled face images are shown in Figure 1.



Fig. 1.

The gray level values of the elements of the pixel grid serve as the features of an image: $\mathbf{x} = (x_{ts}; t = 1, \dots, T, s = 1, \dots, S) \in R^n$, $n = TS$. Even if we deal with relatively small images, the dimension of the resulting feature space is so huge that no training set can exceed it in size. For instance, if $T = 60$ and $S = 40$, the feature space dimension will be $n = 2400$.

If we apply the principle of linear decision rule, it will be mathematically expressed as a 2400-dimensional direction vector which is, actually, an image:

$$\mathbf{a} = (a_{ts}; t = 1, \dots, T, s = 1, \dots, S) \in R^n, \quad \mathbf{a}^T \mathbf{x} + b = \sum_{t=1}^T \sum_{s=1}^S a_{ts} x_{ts} + b \begin{cases} > 0 \rightarrow \text{class 1,} \\ < 0 \rightarrow \text{class 2.} \end{cases} \quad (1)$$

A pictorial representation of the optimal discriminant hyperplane, to be strict, its direction vector, computed by Vapnik's SVM principle [4] for the above ten images (Figure 1) is shown in Figure 2.

The coefficients of the discriminant hyperplane shown in Figure 2 are not sufficiently smooth. At the same time, as we believe, it is just ignoring minor individual details of the training-set images what essentially contributes to good prediction properties of a recognition technique. A face image basically consists of quite large areas of forehead, eyes, cheeks, nose, and too excessive attention to inessential details (birthmarks, wrinkles, face expression) does not improve the prediction.



Fig. 2.

3. The mathematical principle of smoothness-based training regularization

Let $\{(\mathbf{x}_j, g_j), j = 1, \dots, N\}$, $\mathbf{x}_j = (x_{ts,j}, t = 1, \dots, T, s = 1, \dots, S)$, $g_j \in \{1, -1\}$ be a training set from a universe containing two classes of images. The most popular SVM criterion of finding the optimal discriminant hyperplane $\mathbf{a} = (a_{ts}; t = 1, \dots, T, s = 1, \dots, S) \in R^n$ (1) for two subsets set of objects whose convex hulls do not intersect lead to the well known quadratic programming problem in the feature space [4]:

$$\begin{cases} \mathbf{a}^T \mathbf{a} + C \sum_{j=1}^N \delta_j = \sum_{t=1}^T \sum_{s=1}^S a_{ts}^2 + C \sum_{j=1}^N \delta_j \rightarrow \min, \\ g_j (\mathbf{a}^T \mathbf{x}_j + b) = g_j \left(\sum_{t=1}^T \sum_{s=1}^S a_{ts} x_{ts,j} + b \right) \geq 1 - \delta_j, \delta_j \geq 0, j = 1, \dots, N. \end{cases} \quad (2)$$

It is convenient to solve this problem in the dual form with respect to the nonnegative Lagrange multipliers $(\lambda_j, j=1, \dots, N)$ at the inequality constraint associated with the objects of training set:

$$\begin{cases} \sum_{j=1}^N \lambda_j - \frac{1}{2} \sum_{j=1}^N \sum_{k=1}^N \left(g_j g_k \sum_{t=1}^T \sum_{s=1}^S x_{ts,j} x_{ts,k} \right) \lambda_j \lambda_k \rightarrow \max, \\ \sum_{j=1}^N g_j \lambda_j = 0, \quad 0 \leq \lambda_j \leq C/2, \quad j=1, \dots, N. \end{cases} \quad (3)$$

The direction vector of the optimal discriminant hyperplane \mathbf{a} (discriminant image) is the linear combination of the support training-set feature vectors (support images) $a_{ts} = \sum_{j: \lambda_j > 0} g_j \lambda_j x_{ts,j}$ with coefficients $g_j \lambda_j$ produced by non-zero Lagrange multipliers $\lambda_j > 0$. The discriminant image completely specifies the recognition rule applicable to any new image (1).

In this work, our aim is to modify the standard SVM criterion with the purpose to incorporate the available *a priori* information on the sought discriminant image, namely, the assumption that its mutually adjacent elements must not differ significantly from each other.

To formalize the notion of “close” elements of the pixel grid, we consider the Euclidian distance between pair of pixels in the discrete image plane $d_{ts,t's'} = \sqrt{(t-t')^2 + (s-s')^2} \geq 0$, and introduce, on its basis, a nonnegative proximity function $p_{ts,t's'} \geq 0$. The form of this function is rather arbitrary and is to be tried experimentally. Some examples of proximity functions which define nonzero proximity only for adjacent elements are shown in Figure 3.

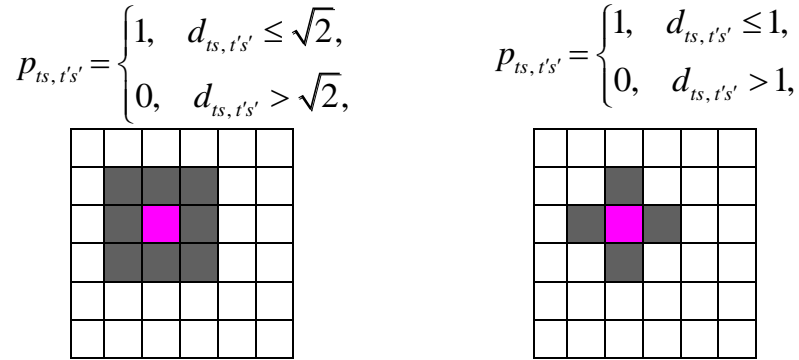


Fig. 3.

We define the regularized training criterion as including an additional penalty upon the difference between spatially close coefficients of the discriminant image (discriminant hyperplane) in the following form:

$$\begin{aligned} & \sum_{t=1}^T \sum_{s=1}^S a_{ts}^2 + \gamma \sum_{t=1}^T \sum_{s=1}^S \sum_{t'=1}^T \sum_{s'=1}^S p_{ts,t's'} (a_{ts} - a_{t's'})^2 + C \sum_{j=1}^N \delta_j = \\ & \sum_{t=1}^T \sum_{s=1}^S a_{ts}^2 + \gamma \sum_{t=1}^T \sum_{s=1}^S \sum_{t'=1}^T \sum_{s'=1}^S b_{ts,t's'} a_{ts} a_{t's'} + C \sum_{j=1}^N \delta_j \rightarrow \min. \end{aligned} \quad (4)$$

Here $\mathbf{B} = (b_{ts,t's'})$ is the matrix $(TS \times TS)$ responsible for the smoothness of the discriminant image

$$\mathbf{B} = 2 \begin{pmatrix} -p_{11} + \sum_{j=1}^{TS} p_{1j} & \cdots & -p_{1TS} \\ \vdots & \ddots & \vdots \\ -p_{TS1} & \cdots & -p_{TS TS} + \sum_{j=1}^{TS} p_{TS j} \end{pmatrix},$$

and the parameter $\gamma \geq 0$ sets the regularization degree. The dual optimization problem corresponding to (4) will have the form



$$\begin{cases} \sum_{j=1}^N \lambda_j - \frac{1}{2} \sum_{j=1}^N \sum_{k=1}^N \left(g_j g_k \sum_{t=1}^T \sum_{s=1}^S \sum_{t'=1}^T \sum_{s'=1}^S f_{ts, t's'} x_{ts, j} x_{ts, k} \right) \lambda_j \lambda_k \rightarrow \max \\ \sum_{j=1}^N g_j \lambda_j = 0, \quad 0 \leq \lambda_j \leq C/2, \quad j = 1, \dots, N, \end{cases} \quad (5)$$

where $\mathbf{F} = (\mathbf{I} + \gamma \mathbf{B})^{-1} = (f_{ts, t's'})$ is the dual regularization matrix ($TS \times TS$).

4. Experimental study

For the experimental study, we used a data set from the well known collection BioID Face DB (<http://www.humanscan.de>). For the two-class recognition problem considered in this paper, we chose images of two person. Some examples of these images are shown in Figure 1. The data base contains 236 face images of the selected two persons. We conducted 100 experiments, in each of which 10 images were randomly chosen as the training set, 5 images of the first and the second person. The remaining 226 images we used for testing. The averaged result is presented in Table 1.

Table 1.

Usual SVM training – smoothness coefficient $\gamma = 0$		
An example of the optimal discriminant hyperplane:		Average error rate in the test sets: 8,7%
Regularized SVM training – smoothness coefficient $\gamma = 10$		
An example of the optimal discriminant hyperplane:		Average error rate in the test sets: 5,7%

Of special interest is the choice of the regularization parameter $\gamma \geq 0$. It is seen from (2) that the regularized training criterion turns into the standard one if $\gamma = 0$. We studied the dependence of the error rate in test set and the leave-one-out error in the training set from the values of

γ . The results are shown in Figure 4. It is seen that positive penalty values lead to improving the extrapolation properties.



Fig. 4.

5. Conclusion

A new method of overcoming the small-sample problem in image recognition is proposed. The idea of the method is incorporating a penalty on the non-smoothness of decision rule coefficients into the standard SVM training criterion. Experiments with face images has shown that the proposed modification essentially improves the prediction power of the SVM recognition rule.

This work is supported by the Russian Foundation for Basic Research, Grants 05-01-00679, 06-01-08042, 06-01-00412, 06-07-89249, and INTAS Grant 04-77-7347.

References

- 1 Guyon I., Elisseeff A., An Introduction to Variable and Feature Selection, *Journal of Machine Learning Research* 3 (2003) 1157-1182.
- 2 Juwei Lu, K.N. Plataniotis, A.N. Venetsanopoulos, Regularization Studies of Linear Discriminant Analysis in Small Sample Size Scenarios with Application to Face Recognition, *Pattern Recognition Letter*, vol. 26, issue 2, pp. 181-191, 2005.
- 3 O.S. Seredin, S.D. Dvoenko, O.V. Krasotkina, and V.V. Mottl, Machine Learning for Signal Recognition by the Criterion of Decision Rule Smoothness. *Pattern Recognition and Image Analysis*, Vol. 11, No. 1, 2001, pp. 87-90.
- 4 Vapnik, V., *Statistical Learning Theory*. New York: Wiley, 1998.