Experimental Study of *a priori* Preferences for Decision Rules in Hilbert Spaces of Recognition Objects[1]

O. S. Seredin

Tula State University, pr. Lenina 92, Tula, 300600 Russia
e-mail: seredin@uic.tula.ru

Abstract—Experimental results confirming the earlier published theoretical statements about the usefulness of the application of secondary projection features in the presence of *a priori* information on the distribution of objects from a general population are given. The use of projection features may improve classification quality when the training samples are sparse.

INTRODUCTION

The authors of [1, 2] suggested the concept of the so-called featureless approach to pattern recognition, which consists in the idea of training algorithms with use in training only the matrix of proximities (similarities) between objects from the training sample. Such an approach is useful when it is difficult to specify a set of meaningful characteristics of objects but, at the same time, there is an objective measure of similarity or dissimilarity between objects.

One of the methods of the featureless approach supposes training in the space of so-called secondary, or projection, features, each of which is a measure of similarity to an object from a distinguished set of nonclassified objects (basis population). It was theoretically shown in [3] that such a training method can be used for regularizing the decision rule in recognition in the presence of *a priori* information about the mutual arrangement of the classes to be recognized. This paper describes an experimental confirmation of the idea of employing the useful information contained in the basis population of objects.

TRAINING IN THE SPACE OF PROJECTION FEATURES

Let Ω be the set of all objects $\omega \in \Omega$ under consideration classified into two subsets $\Omega_1 = \{\omega \in \Omega: g(\omega) = 1\}$ and $\Omega_{-1} = \{\omega \in \Omega: g(\omega) = -1\}$, $\Omega_1 \cup \Omega_{-1} = \Omega$, $\Omega_1 \cap \Omega_{-1} = \varnothing$ by some unknown indicator function $g(\omega) = \pm 1$. Our approach, as opposed to the pattern recognition problem in the classical setting, does not assume the possibility of measuring any observable features $\mathbf{x}(\omega) = (x_1(\omega), \ldots, x_n(\omega))$ on objects that would make it possible to apply the training methods developed for vector feature spaces. However, for any two objects $\omega' \in \Omega$ and $\omega'' \in \Omega$, a numerical characteristic $\mu(\omega', \omega'')$ of their similarity can be measured. Thus, the space Ω of recognition objects can be treated as a Hilbert space. Under certain constraints, we can treat the measure $\mu(\omega', \omega'')$ of similarity between objects of the space Ω as the inner product of elements in the Hilbert space; i.e., we can assume that $\mu(\omega', \omega'') = (\omega', \omega'')$.

One way of training in such a problem is as follows. We suggest fixing a finite set $\Omega^0 = \{\omega_1^0, \ldots, \omega_n^0\} \subset \tilde{\Omega} \subset \Omega$ called a basis assembly. In the general case, we do not assume that the elements of the basis assembly are classified (this is not a training sample). The basis population plays the role of a finite basis in the Hilbert space; it determines the n-dimensional subspace $\Omega_n(\omega_1^0, \ldots, \omega_n^0) = \{\omega \in \Omega: \omega = \sum_{k=1}^{n} a_k \omega_k^0\} \subset \Omega$. Thus, each element $\omega \in \Omega$ in the Hilbert space is assigned a set of similarity measures, which is considered as a real-valued vector of secondary or projection features:

$$\mathbf{x}(\omega) = (x_1(\omega), \ldots, x_n(\omega))^T \in R^n, \; x_k(\omega) = (\omega, \omega_k^0).$$

In [3, 4], it is shown that training in the space of projection features is equivalent to the expression of preferences, which are related to the tendency of the directing element of a partitioning hyperplane to be close to the principal inertia axis of the basis population of objects, in the initial Hilbert space. As a result, the partitioning hyperplane tends to be orthogonal to this axis.

The aforesaid considerations are inessential if the domain where the objects of the Hilbert space largely accumulate does not have a prevailing orientation. Such an indifference is an exception rather than a rule. It is natural to assume that the distribution of objects is stretched differently in different directions. This must be reflected in the basis assembly and, of course, in the training sample. We believe that the assumption that the set of objects from one class is approximately equally stretched in all directions is natural. Mathematically,

[1] This work was supported by State scientific and technical program of the Russian Federation "Promising Information Technologies" and the Russian Foundation for Basic Research.

Received October 29, 2002

this assumption is expressed as follows: if we form the matrix of pairwise mutual proximities having the properties of an inner product for some set of objects from the same class, then all eigenvalues of this matrix are approximately the same. Now, suppose that the set under consideration contains objects from both the first and second classes. Suppose also that we do not know which objects in this set belong to the first and second classes, but we know that the classes can be linearly partitioned. Note that the basis assembly is the very same set of objects of two classes. According to our assumption, each class forms a domain close to a sphere; therefore, both classes together must form a domain stretched in the direction in which the classes are distant from each other. The eigenvalues of the matrix of pairwise mutual proximities between objects forming this domain must differ substantially. The eigenvector corresponding to the maximum eigenvalue indicates the stretch direction.

Thus, under the assumptions made above, in training, it is natural to give priority to partitioning hyperplanes almost orthogonal to the principal eigenvector of the basis population of objects. For this reason, it seems expedient to ignore the partitioning hyperplanes oriented along the basis population, even if the gap between the objects from the first and second classes has such an orientation, and prefer "transverse" hyperplanes.

The preference of the decision rules whose directing elements are close to the inertia axis of the basis can be used as a method for fighting the small sample curse (a method for stabilizing or regularizing the decision rule in recognition); it consists in employing additional information, possibly *a priori* and not reflected in the training sample, in the construction of the decision rule.

AN EXPERIMENTAL STUDY

To verify the theoretical statements, we considered two kinds of features:

(i) the feature vector $\mathbf{x}_j = (x_1, x_2, \ldots, x_n)^T \in \mathrm{R}^n$ of the jth object is a point in n-dimensional real space;

(ii) the feature vector $\mathbf{x}_j^* = (\mathbf{x}_1^T\mathbf{x}_j, \mathbf{x}_2^T\mathbf{x}_j, \ldots, \mathbf{x}_N^T\mathbf{x}_j)^T \in \mathrm{R}^N$ of the jth object is the vector of pairwise inner products with the initial feature vectors of objects from the basis population.

In the second case, we modeled the problem of featureless pattern recognition. As the basis assembly, the training sample was taken.

A decision rule $\hat{g}(\mathbf{x})$ was chosen in the class of linear functions

$$d(\mathbf{x}|\mathbf{a}, b) = \mathbf{a}^T\mathbf{x} + b = \sum_{j=1}^{N} a_j x_j + b,$$

$$\hat{g}(\mathbf{x}|\mathbf{a}, b) = \begin{cases} 1, & d(\mathbf{x}|\mathbf{a}, b) > 0 \\ -1, & d(\mathbf{x}|\mathbf{a}, b) < 0. \end{cases}$$

The hyperplane $d(\mathbf{x}|\mathbf{a}, b) = \mathbf{a}^T\mathbf{x} + b = 0$ partitions the vector space R^n (in the first case) or R^N (in the second case) into two domains of decisions in favor of the first and second classes. However, we emphasize that, in the second case, the dimension of the coefficient vector $\mathbf{a}^* = (a_1, a_2, \ldots, a_N)^T$ equals N, because the dimension of the vector of projection features equals the number of objects in the training sample; i.e., $\mathbf{x}_j^* \in \mathrm{R}^N$.

To construct a decision rule, we used Vapnik's method of support vectors [5], which determines the optimal partitioning hyperplane in the sense that it maximizes gaps between samples.

To define the notion of the stretch of the distribution of objects in the feature space, we denote the minor and major half-axes of the ellipse bounding the domain in the two-dimensional space inside which the features of objects from the training sample are randomly formed on the basis of the uniform distribution law by a and b and the sides of the conventional right-angled parallelepiped bounding the domain in the three-dimensional space by a, b, and c. In the three-dimensional case, we assume that $a = c$. The stretch of the population of objects is the ratio $k = b/a$.

To establish the advantage of the pattern recognition training algorithm using the projection features $\mathbf{x}_j^* = (\mathbf{x}_1^T\mathbf{x}_j, \mathbf{x}_2^T\mathbf{x}_j, \ldots, \mathbf{x}_N^T\mathbf{x}_j)^T \in \mathrm{R}^N$ over the training method using features of the form $\mathbf{x}_j = (x_1, x_2, \ldots, x_n)^T \in \mathrm{R}^n$ when the stretch coefficient is large enough, we performed series of model experiments on pattern recognition training with the use of the method of support vectors for various values of the stretch coefficient k. Objects of each class in the training sample are uniformly distributed in the bounded two- or three-dimensional space of features. The feature space was assumed to be partitioned in advance by a fixed (preset) partitioning hyperplane into two equal domains of features of objects from the first and second classes. In the course of the experiment, by considering a small number of objects from the training sample (2–10 in each class), we tried to model the situation of the lack of training material (the problem of a small sample).

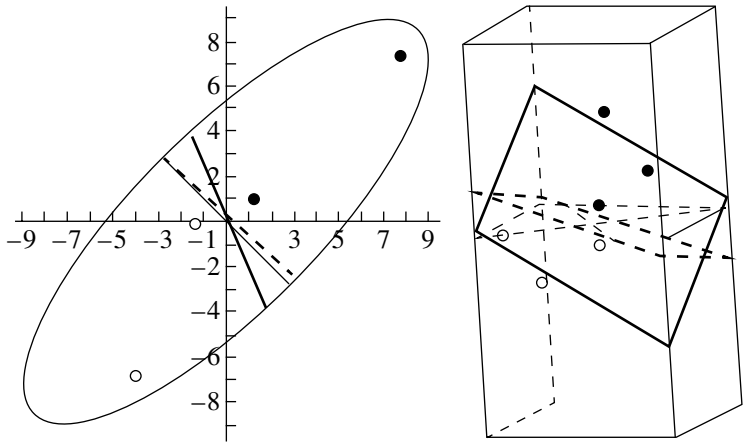

Fig. 1. Optimal partitioning hyperplanes for features of the first (solid line) and second (dashed line) kinds. The stretch coefficient of the sample was $k = 3$. The recognition error was 4.03% in the first case and 1.34% in the second case for the two-dimensional space; for the three-dimensional space, the errors were 9.47 and 4.37%, respectively.

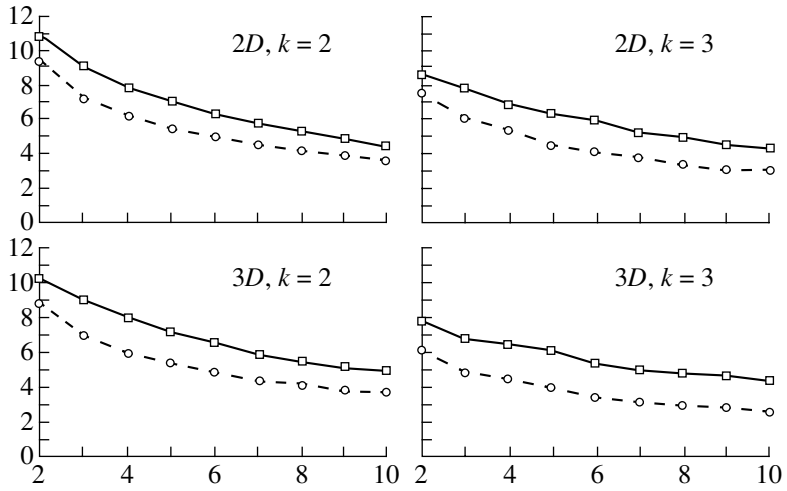

Fig. 2. Experimental results. The percentage of recognition error for various k values and various dimensions of the initial space as a function of the number of objects of each class in the training sample. Input (solid line) and projection (dashed line) features are used.

The classification error was estimated as the ratio of the area (in the two-dimensional case) or volume (in the three-dimensional case) incorrectly attached to domains of objects of the first and second classes to the total area (volume) of the bounded feature space (Fig. 1).

To determine the tendency of the influence of the choice of characteristics of the form $\mathbf{x}_j = (x_1, x_2, \ldots, x_n)^T \in \mathrm{R}^n$ or $\mathbf{x}_j^* = (\mathbf{x}_1^T\mathbf{x}_j, \mathbf{x}_2^T\mathbf{x}_j, \ldots, \mathbf{x}_N^T\mathbf{x}_j)^T \in \mathrm{R}^N$ as features of objects on the construction of a partitioning hyperplane, series of 1000 experiments each in the two- and three-dimensional initial feature spaces for various numbers of objects in both classes and at various values of the stretch coefficients of the training population were performed. The results of the experiments are presented in Fig. 2.

CONCLUSIONS

The experiments confirmed the earlier theoretical conjecture that the quality of recognition improves when the classifier is constructed in the space of projection features. This approach can be used both in the case of the featureless concept and in the framework of the classical, "feature," recognition. Under the *a priori* assumption that the general population is stretched along the distribution of classes, the use of projection features may improve the quality of classification.

REFERENCES

1. Duin, R.P.W., de Ridder, D., and Tax, D.M.J., Featureless Classification, *Proc. Workshop on Statistical Pattern Recognition*, Prague, 1997.

2. Duin, R.P.W., Pekalska, E., and de Ridder, D., Relational Discriminant Analysis, *Pattern Recogn. Lett.*, 1999, vol. 20, nos. 11–13, pp. 1175–1181.

3. Seredin, O.S., Methods and Algorithms for Featureless Pattern Recognition, *Cand. Sci. Dissertation*, Moscow, 2001.

4. Mottl, V.V., Dvoenko, S.D., Seredin, O.S., Kulikowski, C.A., and Muchnik, I.B., Featureless Regularized Recognition of Protein Fold Classes in a Hilbert Space of Pairwise Alignment Scores As Inner Products of Amino Acid Sequences, *Pattern Recognit. Image Anal.*, 2001, vol. 11, no. 3, pp. 597–615.

5. Vapnik, V., *Statistical Learning Theory*, New York: Wiley, 1998.