DIMACS Technical Report 2001-01
January 2001

Alignment scores in a regularized support vector classification method for fold recognition of remote protein families

by

Vadim Mottl, Sergey Dvoenko, Oleg Seredin

Department of Cybernetics
Tula State University
Lenin Ave. 92, 300600 Tula, Russia
mottl@atm.tsu.tula.ru

Casimir Kulikowski, Ilya Muchnik

Department of Computer Science
Rutgers University
P.O. Box 8018, Piscataway, NJ 08855
kulikows@cs.rutgers.edu

DIMACS is a partnership of Rutgers University, Princeton University, AT&T Labs-Research, Bell Labs, Bellcore and NEC Research Institute.

DIMACS is an NSF Science and Technology Center, funded under contract STC-91-19999; and also receives support from the New Jersey Commission on Science and Technology.

Abstract

One of fundamental principles of molecular biology says that the primary structure of a protein, i.e. sequence of amino acid residues forming its polypeptide chain, carries an essential amount of information for unambiguous establishing its spatial structure. Despite the fact that each protein has its own spatial structure, it is typical phenomenon that the fold pattern remains basically the same within large groups of evolutionarily allied proteins, so that the "number" of essentially different spatial structures is much less than that of known proteins. Since spatial structures are classified in that or other manner, the estimation of the spatial structure of a given protein reduces to its allocating over a finite set of classes, i.e. the problem falls into the competence area of pattern recognition.

The traditional methodology of pattern recognition presupposes that the object whose class-membership is to be recognized is represented by vector of some numerical features and is considered as a point in the respective linear vector space. However, the actual diversity of amino acid properties that may play an important part in forming the spatial structure of a protein is so immensely rich, that the choice of suitable numerical properties makes a special problem which is the key one here. We consider here an alternative featureless approach to recognition of spatial structure of proteins. It is proposed to judge about the membership of a protein in one of the classes of spatial structures immediately on the basis of measuring the proximity of its amino acid chain to those of some other proteins whose spatial structure is known.

To infer the decision rule of recognition from a training sample of proteins of known structure, we apply the traditional support vector method of machine learning on the basis of treating amino acid chains as elements of a Hilbert space, i.e. linear space with inner product, in what role the pairwise alignment scores are used. The inevitable difficulty of the small size of training samples is overcome by a special regularization technique that makes use of some available a priori information on the sought-for decision rule.

The proposed approach to the problem of fold class recognition illustrated by results of processing a collection of 396 mutually distant protein domains of 51 fold classes chosen from the SCOP database.

Contents

1. Introduction

The problem of finding the spatial structure of a protein represented by its primary amino acid sequence is a challenge posed by nature to the bioinformatical community.

On the one hand, the necessity of such algorithms is dictated by the fact that application of usual physical techniques of magnetic resonance and X-ray analysis is problematic in most cases. Although the number of proteins whose spatial structure is known ever grows, the gap between the number of known amino acid sequences and that of known spatial structures is increasing dramatically.

On the other hand, the existence theorem is proved by nature itself, because it has been never observed that an amino acid chain had more than one spatial structure. One of fundamental principles of molecular biology says that the primary structure of a protein carries a sufficient amount of information for unambiguous establishing its spatial structure.

Each protein has its specific spatial organization which does not coincide with organization of any other protein. The main principle of establishing the spatial structure of a given protein from its amino acid chain consists in finding, for the given chain, the most appropriate structure from a bank of known structures and their fragments. For each amino acid residue in the chain forming a protein of a known structure, the vector of some quantitative features is evaluated which are assumed to be responsible for the spatial position of this residue in the three-dimensional structure. The succession of such features along the amino acid chain is called the profile of this structure. The same features are evaluated for the amino acid chain of the new protein, whereupon the succession obtained is compared with profiles of known structures by alignment of positions in this succession and in the respective profile with respect to eventual insertions and deletions. Such a principle named threading [1] is fraught with enumeration of a large number of known structures.

However, despite the uniqueness of the spatial structure of each protein, it is the usual case that large groups of evolutionary allied proteins have very similar spatial structures. In this sense, there exist ”much less” spatial structures than primary ones.

Of course, the classification of spatial structures is a problem which is not simple, but once a version of classification is accepted, the problem of assigning a protein given by its amino acid chain to a class of spatial structures falls into the competence area of machine learning and, more specifically, pattern recognition.

However, the modern machine learning theory is unable to propose a completely ready instrument for solving this problem because of its deep specificity.

First, a protein is represented by its amino acid sequence whose length may vary essentially, whereas the machine learning theory considers only the case when an object is characterized by a vector of numerical features of a fixed dimensionality.

Second, as the number of classes may run into dozens and hundreds, the spatial structure is known for a relatively small part of proteins, so that each class can be represented in a training sample of a reasonable size only by a small number of proteins.

Third, on the force of the fact that classes of spatial structure exist only conventionally, any classification of spatial structures will always be incomplete in the sense that at any moment it may turn out to be convenient to set off a new class.

The traditional methodology of pattern recognition presupposes that the object, whose class-membership is to be recognized, in this case, the primary amino acid sequence of a protein, is represented by vector of some numerical features and is considered as a point in the respective linear vector space. In particular, the primary structure of a protein may be represented by frequencies with which amino acids of the polar, neutral and hydrophobic type occur in it along with frequencies of their pairwise occurrence at neighboring positions in the chain [2].

The machine learning algorithm analyses a training sample of proteins whose spatial structure is known, and forms such a system of hyperplanes in the feature space which would separate, as accurately as possible, the points that correspond to proteins of different classes. It is expected that new proteins, that did not participate in training, will have the structure of just that class which is associated with the region of the feature space where their feature vectors fall.

This principle of machine learning aimed at inferring from a training sample a reliable decision rule applicable to all feasible objects of recognition is called training by examples and is based on the so-called compactness hypothesis, in accordance with which it is expected that points close to each other in the feature space tend to be produced by objects of the same class.

However, the actual diversity of amino acid properties that may play an important part in forming the spatial structure of a protein is so immensely rich, that the choice of suitable properties of amino acids makes a special problem which is the key one here.

In this Report, we consider an alternative methodology of protein fold class recognition based on the so-called featureless approach to pattern recognition proposed by R. Duin and colleagues [3,4,5] originally for handprinted character recognition. As applied to proteins, the essence of the methodology consists in judging about the membership of a protein in one of the fold classes immediately from the measured proximity of its amino acid sequence to those of some other proteins whose spatial structures are known, without designing any special features that would reflect structure-forming properties of amino acids in the polypeptide chains.

As the set of experimental data, we used the collection of proteins selected by Dr. Sun-Ho Kim from Lawrence Berkley National Laboratory. The collection contains 396 remote protein domains of 51 fold classes chosen from the SCOP database (Table 1.1). The principle of selection was to provide as a little similarity of amino acid sequences within each family as possible. As the threshold, the similarity degree of 27% measured by the Fasta3 pair-wise alignment procedure was accepted, because the lesser threshold did not allow to find representatives for some fold classes.

Such a principle of selection resulted in protein domain families that are not equal in size. The smallest class *Beta-Lactamase / D-ala carboxypeptidase* contains only 3 domains, whereas three largest classes *Immunoglobin beta – sandwich*, *TIM – barrel* and *Ferredoxin* consist of, respectively, 31, 28 and 20 domains.

As proximity measure of two proteins, we use the pairwise alignment score of their amino acid chains. In its turn, the alignment score is dealt with as inner product of the respective proteins in an imaginary Hilbert space. Just as in the classical pattern recognition, we lean upon the compactness hypothesis, which, however, is treated immediately but not through the intervening notion of a linear feature space.

Nevertheless, we use here the principle of machine learning designed in the early 90s by V. Vapnik just for pattern recognition in linear feature spaces and named support vector method [6,7]. The main idea of this method fits to the best to the featureless pattern recognition because it rests on an inner-product-like similarity measure between entities whose class-membership is to be recognized rather than on a set of their features.

Table 1.1. Dr. Kim's collection of proteins.

	Fold class	Size		Fold class	Size
1	Globin	12	27	Flavodoxin	9
2	Cytochrome C	7	28	Adenine nucleotide alpha hydroclase	4
3	Four-helical bundle	8	29	Rossmann-fold domains	14
4	Ferritin	8	30	Thiamin-binding	3
5	4-gelical cytokines	11	31	P-loop containing NTP hydrolases	9
6	EF Hand	13	32	Thioredoxin fold	9
7	Cyclin	4	33	Restriction endonucleases	5
8	Cytochrome P450	5	34	Ribonuclease H motif	9
9	Immunoglobin beta – sandwich	31	35	Phosphoribosyltransferases (PRTases)	3
10	Common fold of difteria toxin / transcription factors / cytochrome	5	36	S-adenosyl-L-methionine-dependent methyltransferases	5
11	Cupredoxins	9	37	Alpha / beta-Hydrolases	12
12	C2 domain	3	38	Phosphorylase / hydrolase	5
13	Viral coat and capsid proteins	15	39	Periplastic binding protein I	7
14	Crystallins / protein S / yeast killer toxin	5	40	Periplastic binding protein II	7
15	Galastose-binding domain	4	41	Lysozyme	4
16	ConA lectins / glucanases	8	42	Cysteine proteinases	4
17	OB-fold	17	43	Beta-Grasp	8
18	Beta-Trefoil	5	44	Cystatin	7
19	Reductase / isomerase / elongation factor	4	45	Ferredoxin	20
20	Trypsin serine proteases	6	46	Zincin	7
21	Acid proteases	5	47	N-terminal nucleophile aminohydrolases	4
22	PH domain	7	48	ADP-ribosylation	4
23	Lipocalings	6	49	C-type lectin	6
24	Double-stranded beta-helix	6	50	Protein kinases (PK), catalytic core	4
25	Barrel-sandwich hybrid	6	51	Beta-Lactamase / D-ala carboxypeptidase	3
26	TIM-barrel	28			

2. Support vector method of machine learning for pattern recognition in a linear feature space

2.1. The optimal discriminant hyperplane

The technique of featureless pattern recognition we consider below in this Report essentially rests on the classical support vector method designed for machine learning in a linear feature space. Therefore, to start with, we outline here the fundamental notions of the classical theory [7] that will serve below as a basis for the featureless approach.

We shall assume that each object is characterized by index of its hidden membership in one of two classes $g \in \{1, -1\}$, where indices 1 and –1 denote, respectively, the first and the second class, as well as by the observable vector of its features $\mathbf{x} = (x_1 \cdots x_n)^T \in \mathsf{R}^n$ considered as a point in the n-dimensional real space. The decision rule meant to estimate the class-membership of an object is to be chosen within the family of linear discriminant functions

$$d(\mathbf{x} \,|\, \mathbf{a}, b) = \mathbf{a}^T \mathbf{x} + b = \sum_{j=1}^{N} a_j x_j + b \,, \qquad \hat{g}(\mathbf{x} \,|\, \mathbf{a}, b) = \begin{cases} 1, & d(\mathbf{x} \,|\, \mathbf{a}, b) > 0, \\ -1, & d(\mathbf{x} \,|\, \mathbf{a}, b) < 0. \end{cases} \tag{2.1}$$

Thus, the decision rule is a hyperplane $d(\mathbf{x} \,|\, \mathbf{a}, b) = \mathbf{a}^T \mathbf{x} + b = 0$ that partitions the vector space R^n into two decision regions, respectively, in favor of the first and the second class.

Let the training sample consist of N objects $(\mathbf{x}_j, g_j)$, $j = 1,...,N$, each of which is represented by its feature vector $\mathbf{x}_j = (x_{j1} \cdots x_{jn})^T \in \mathsf{R}^n$ and class-membership index $g_j \in \{1, -1\}$. The main idea of training is to find such a hyperplane in R^n that the points $\mathbf{x}_i$ belonging to different classes would fall at different sides of it:

$$d(\mathbf{x}_j \,|\, \mathbf{a}, b) = \mathbf{a}^T \mathbf{x}_j + b = \begin{cases} > 0 \text{ when } g_j = 1, \\ < 0 \text{ when } g_j = -1. \end{cases}$$

The training sample is said to be linearly separable if such a hyperplane exists, and linearly inseparable in the opposite case (Fig. 2.1). In any case, it is required to find the hyperplane that

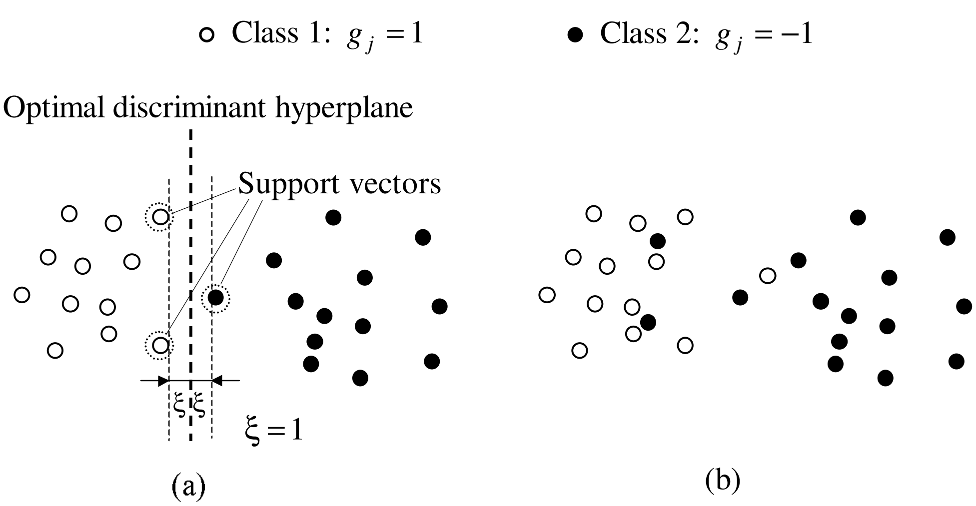

Figure 2.1. A linearly separable (a) and a linearly inseparable (b) training sample.

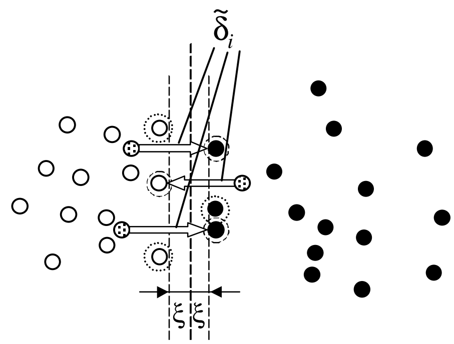

Figure 2.2. Shifting the "nuisance" vectors that prevent linear separability of the training sample.

would separate the subsamples of both classes as good as possible.

Let us consider first the situation when classes are linearly separable. Then there exists a family of hyperplanes $d(\mathbf{x}\,|\,\mathbf{a},b)=\mathbf{a}^T\mathbf{x}+b=0$ such that

$$\begin{aligned} d(\mathbf{x}_j\,|\,\mathbf{a},b)&=\mathbf{a}^T\mathbf{x}_j+b\geq\xi \text{ when } g_j=1,\\ d(\mathbf{x}_j\,|\,\mathbf{a},b)&=\mathbf{a}^T\mathbf{x}_j+b\leq-\xi \text{ when } g_j=-1,\end{aligned} \tag{2.2}$$

where $\xi>0$. In other words, each of such hyperplanes is characterized by the margin 2ξ between the classes (Fig. 2.1,a). One of them, for which $\xi\to\max$ and the conditions (2.2) and $\mathbf{a}^T\mathbf{a}=1$ are met, is called the optimal discriminant hyperplane.

Dividing the constraints in (2.2) by ξ leads to the equivalent constraints

$$d\left(\mathbf{x}_j\Big|\frac{1}{\xi}\mathbf{a},\frac{1}{\xi}b\right)=\frac{1}{\xi}\mathbf{a}^T\mathbf{x}_j+\frac{1}{\xi}b\geq 1 \text{ when } g_j=1,$$
$$d\left(\mathbf{x}_j\Big|\frac{1}{\xi}\mathbf{a},\frac{1}{\xi}b\right)=\frac{1}{\xi}\mathbf{a}^T\mathbf{x}_j+\frac{1}{\xi}b\leq -1 \text{ when } g_i=-1.$$

Using symbols $\mathbf{a}$ and b instead of $\frac{1}{\xi}\mathbf{a}$ and, respectively, $\frac{1}{\xi}b$, we come to the equivalent expressions

$$d(\mathbf{x}_j\,|\,\mathbf{a},b)=\mathbf{a}^T\mathbf{x}_j+b\geq 1 \text{ when } g_j=1,$$
$$d(\mathbf{x}_j\,|\,\mathbf{a},b)=\mathbf{a}^T\mathbf{x}_j+b\leq -1 \text{ when } g_j=-1,$$

or, in a more compact form,

$$g_j(\mathbf{a}^T\mathbf{x}_j+b)\geq 1,\quad j=1,...,N.$$

The problem of building the optimal discriminant hyperplane is equivalent to that of maximization of ξ or, what is the same, minimization of $\|\mathbf{a}\|^2=\mathbf{a}^T\mathbf{a}$, because the squared norm of the "new" vector $\mathbf{a}$ equals $\mathbf{a}^T\mathbf{a}=(1/\xi^2)\tilde{\mathbf{a}}^T\tilde{\mathbf{a}}=(1/\xi^2)$ when $\tilde{\mathbf{a}}^T\tilde{\mathbf{a}}=1$, i.e. when the squared norm of the "old" vector is fixed. As a result, we come to the objective function

$$\mathbf{a}^T\mathbf{a}\to\min,\quad g_j(\mathbf{a}^T\mathbf{x}_j+b)\geq 1,\quad j=1,...,N. \tag{2.3}$$

This criterion is the standard quadratic programming problem that can be solved by one of well-known standard methods.

However, such an approach becomes senseless in case the classes are inseparable (Fig. 2.1,b) because the constraints (2.2) and, hence, (2.3) are no longer compatible. To design an analogous criterion for such training samples, we use the following trick. Let it be impossible to find vector $\mathbf{a}$ that would satisfy inequalities (2.2) with a positive margin ξ, or, what is the same, inequalities $g_j(\mathbf{a}^T\mathbf{x}_j+b)\geq\xi$ with $\mathbf{a}^T\mathbf{a}=1$. But it is always possible to provide this by admitting nonnegative shifts

$$g_j(\mathbf{a}^T\mathbf{x}_j+b)\geq\xi-\tilde{\delta}_j,\ \ \xi>0,\ \ \tilde{\delta}_j\geq 0\,.$$

The shifts $\tilde{\delta}_j$ should be as small as possible and, at the same time, make the margin as large as possible $\xi\to\max$ or $1/\xi\to\min$, what may be achieved by a compromise criterion, for instance $1/\xi+C\sum_{j=1}^N\tilde{\delta}_j\to\min$ under the constraint $\mathbf{a}^T\mathbf{a}=1$ with a sufficiently large positive coefficient C meant to give preference to the minimization of shifts[1]. If we introduce the denotation $\delta_j=\tilde{\delta}_j/\xi$ and treat $\mathbf{a}$ as the normalized vector $\frac{1}{\xi}\mathbf{a}$, we shall come to the criterion $1/\xi^2+C\sum_{j=1}^N\delta_j\to\min$, where $1/\xi^2=\mathbf{a}^T\mathbf{a}$, under constraints $g_j(\mathbf{a}^T\mathbf{x}_j+b)\geq 1-\delta_j$ and $\delta_j\geq 0$.

The idea of this trick consists in mental shifting, to the side of the "correct" class $\delta_j>0$, those of objects, which fall into the wrong region, and in leaving untouched $\delta_j=0$ those ones, which do not prevent "pushing" a hyperplane between subsamples of the first and the second class (Fig. 2.2).

So, we come to the following formulation of the generalized problem of finding the optimal discriminant hyperplane that covers both the separable and inseparable case:

$$\begin{cases}\mathbf{a}^T\mathbf{a}+C\sum\limits_{j=1}^N\delta_j\to\min,\\ g_j(\mathbf{a}^T\mathbf{x}_j+b)\geq 1-\delta_j,\quad \delta_j\geq 0,\quad j=1,\ldots,N.\end{cases}\tag{2.4}$$

This is a quadratic programming problem with the number of arguments equal to dimension of the feature space plus the size of the training sample, and the number of constraints in the form of inequalities equal to the doubled size of the training sample.

However, the dual formulation of this problem is more convenient here, because its solution points out, in an explicit form, the so-called support elements of the training sample, which are the only ones that participate in forming the optimal discriminant hyperplane.

2.2. The dual formulation of the training problem

It is important to mark that the decision will always be at the boundary of the search area. This means that the inequalities $g_j(\mathbf{a}^T\mathbf{x}_j+b)\geq 1-\delta_j$ in (2.4) turn into equalities

---

[1] In Vapnik's works, also other ways of penalizing the shifts are considered, in particular, $C\sum_{j=1}^N\delta_j^2$ and $C\left(\sum_{j=1}^N\delta_j\right)^2$.

$g_j(\mathbf{a}^T\mathbf{x}_j + b) = 1 - \delta_j$ for some vectors $\mathbf{x}_j$. It is just these objects of the training sample which V. Vapnik calls support vectors, because only these vectors form the direction vector of the discriminant hyperplane as their linear combination (Fig. 1a and 2). To make sure that it is so, let us study the Lagrangian of the problem (2.4):

$$L(\mathbf{a}, b, \lambda_1,...,\lambda_N, \delta_1,...,\delta_N, \mu_1,...,\mu_N) = \frac{1}{2}\mathbf{a}^T\mathbf{a} + \frac{C}{2}\sum_{j=1}^{N}\delta_j - \\ - \sum_{j=1}^{N}\lambda_j\left[g_j(\mathbf{a}^T\mathbf{x}_j + b) - 1 + \delta_j\right] - \sum_{j=1}^{N}\mu_j\delta_j, \tag{2.5}$$

where $\lambda_j \geq 0$, $\mu_j \geq 0$, $j = 1,...,N$, are nonnegative Lagrange multipliers. The solution of the problem is the saddle point of the Lagrangian

$$L(\mathbf{a}, b, \lambda_1,...,\lambda_N, \delta_1,...,\delta_N, \mu_1,...,\mu_N) \to \min \quad \text{with respect to } \mathbf{a}, b, \delta_1,...,\delta_N$$

without any constraints and

$$L(\mathbf{a}, b, \lambda_1,...,\lambda_N, \delta_1,...,\delta_N, \mu_1,...,\mu_N) \to \max \quad \text{with respect to } \lambda_1,...\lambda_N, \mu_1,...,\mu_N \tag{2.6}$$

under constraints $\lambda_j \geq 0, \mu_j \geq 0, \quad j = 1,...,N$.

The former of these conditions gives

$$\nabla_{\mathbf{a}} L(\mathbf{a}, b, \lambda_1,...,\lambda_N, \delta_1,...,\delta_N, \mu_1,...,\mu_N) = 0, \quad \frac{\partial L(\mathbf{a}, b, \lambda_1,...,\lambda_N, \delta_1,...,\delta_N, \mu_1,...,\mu_N)}{\partial b} = 0,$$

$$\frac{\partial L(\mathbf{a}, b, \lambda_1,...,\lambda_N, \delta_1,...,\delta_N, \mu_1,...,\mu_N)}{\partial \delta_j} = 0, \quad j = 1,...,N,$$

whence we obtain

$$\mathbf{a} = \sum_{j=1}^{N}\lambda_j g_j \mathbf{x}_j, \tag{2.7}$$

$$\sum_{j=1}^{N}\lambda_j g_j = 0, \tag{2.8}$$

$$\lambda_j + \mu_j = \frac{C}{2}, \quad j = 1,...,N. \tag{2.9}$$

Mark that $0 \leq \lambda_j \leq C/2$ и $0 \leq \mu_j \leq C/2$, because $\lambda_j \geq 0, \mu_j \geq 0$ and $\lambda_j + \mu_j = C/2$.

Substitution of (2.7) into the condition (2.6) turns the latter one into an objective function that is dual relative to the original objective function (2.4)

$$W(b, \lambda_1,...,\lambda_N, \delta_1,...,\delta_N, \mu_1,...,\mu_N) = \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}(g_j g_k \mathbf{x}_j^T\mathbf{x}_k)\lambda_j\lambda_k + \frac{C}{2}\sum_{j=1}^{N}\delta_j - \\ \sum_{j=1}^{N}\sum_{k=1}^{N}(g_j g_k \mathbf{x}_j^T\mathbf{x}_k)\lambda_j\lambda_k - \left(\sum_{j=1}^{N}\lambda_j g_j\right)b + \sum_{j=1}^{N}\lambda_j - \sum_{j=1}^{N}\lambda_j\delta_j - \sum_{j=1}^{N}\mu_j\delta_j,$$

where, on the force of inequalities (2.8) and (2.9), only Lagrange multipliers $\lambda_j \geq 0$, $j = 1,...,N$, are active arguments.

Thus, we come to the following dual formulation of the problem of building the optimal discriminant hyperplane as a quadratic programming problem:

$$\begin{aligned} & W(\lambda_1,...,\lambda_N)=\sum_{j=1}^{N}\lambda_j-\frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}(g_j g_k \mathbf{x}_j^T\mathbf{x}_k)\lambda_j\lambda_k \to \max, \\ & \sum_{j=1}^{N}\lambda_j g_j=0, \quad 0\le\lambda_j\le\frac{1}{2}C,\ j=1,...,N. \end{aligned} \tag{2.10}$$

Here the optimal values $\lambda_j \ge 0$ are Lagrange multipliers under constraints $g_j(\mathbf{a}^T\mathbf{x}_j+b)\ge 1-\delta_j$, each of which is associated with the respective element of the training sample $\mathbf{x}_i$, so that the positive of them $\lambda_j>0$ point at active constraints $g_j(\mathbf{a}^T\mathbf{x}_j+b)=1-\delta_j$.

According to (2.7), the direction vector of the optimal discriminant hyperplane is a linear combination of, generally speaking, all the vectors making the training sample $\mathbf{a}=\sum_{j=1}^{N}\lambda_j g_j\mathbf{x}_j$, but in its forming participate, actually, only those of them whose Lagrange multipliers differ from zero $\lambda_j>0$. It is just these vectors which are support ones, because only they "support" the optimal discriminant hyperplane. In the case of a linearly separable training sample, these are marginal points of the subsamples of the first and second class, by which the latter ones are faced to each other (Fig. 2.1,a). If the training sample is inseparable, the situation is somewhat more complicated, and along with the shifted "nuisance" points shown in Fig. 2.2, also some "good" marginal points of both subsamples will be support ones.

To complete finding the optimal discriminant hyperplane, it remains to find parameter b that is responsible for its parallel translation along the direction vector.

The shifts of elements of the training sample equal zero if and only if the respective Lagrange multipliers are positive $\mu_j>0$, i.e. if $\lambda_j<C/2$ on the force of (2.9). In other terms,

$$\delta_j>0 \text{ if and only if } \lambda_j=\frac{1}{2}C. \tag{2.11}$$

For the positive λ_j the condition $g_j(\mathbf{a}^T\mathbf{x}_j+b)=1-\delta_j$ holds. If $0<\lambda_j<C/2$, then $\delta_j=0$ and, hence, $g_j(\mathbf{a}^T\mathbf{x}_j+b)=1$. Thus, those of elements of the training sample for which $0<\lambda_j<C/2$ satisfy the conditions $b=1-\mathbf{a}^T\mathbf{x}_j$ when $g_j=1$ and $-b=1+\mathbf{a}^T\mathbf{x}_j$ when $g_j=-1$, that give, in their turn,

$$\sum_{\substack{j:\,g_j=1\\0<\lambda_j<C/2}} b = \sum_{\substack{j:\,g_j=1\\0<\lambda_j<C/2}} 1 - \sum_{\substack{j:\,g_j=1\\0<\lambda_j<C/2}} \mathbf{a}^T\mathbf{x}_j, \qquad \sum_{\substack{j:\,g_j=-1\\0<\lambda_j<C/2}} (-b) = \sum_{\substack{j:\,g_j=-1\\0<\lambda_j<C/2}} 1 + \sum_{\substack{j:\,g_j=-1\\0<\lambda_j<C/2}} \mathbf{a}^T\mathbf{x}_j .$$

Let the training sample contain $\tilde{N}_1$ objects of the first class and $\tilde{N}_{-1}$ objects of the second class, then we have

$$b=1-\frac{1}{\tilde{N}_1}\sum_{\substack{j:\,g_j=1\\0<\lambda_j<C/2}}\mathbf{a}^T\mathbf{x}_j, \qquad -b=1+\frac{1}{\tilde{N}_{-1}}\sum_{\substack{j:\,g_j=-1\\0<\lambda_j<C/2}}\mathbf{a}^T\mathbf{x}_j .$$

Subtracting the second equality from the first one gives the formula for b:

$$b = -\frac{1}{2}\left(\frac{1}{\tilde{N}_1} \sum_{\substack{j:\, g_j=1 \\ 0<\lambda_j<C/2}} \mathbf{a}^T \mathbf{x}_j - \frac{1}{\tilde{N}_{-1}} \sum_{\substack{j:\, g_j=-1 \\ 0<\lambda_j<C/2}} \mathbf{a}^T \mathbf{x}_j \right). \quad (2.12)$$

Now, that we have found both parameters of the discriminant hyperplane $\mathbf{a} \in \mathsf{R}^n$ and $b \in \mathsf{R}$, the training is completed, and the discriminant function (2.1) is ready for being applied to the feature vectors $\mathbf{x} \in \mathsf{R}^n$ of new objects to make decisions of their class-memberships $\hat{g}(\mathbf{x}) = \pm 1$. Mark that, in accordance with (2.7) and (2.12), the discriminant function deals only with inner products of the feature vector of the new object with those of the support objects of the training sample, but not with feature vectors themselves:

$$\begin{cases} d(\mathbf{x}) = \sum_{j:\lambda_j>0} g_j \lambda_j \mathbf{x}^T \mathbf{x}_j + b, \\ b = -\frac{1}{2}\left(\frac{1}{\tilde{N}_1} \sum_{j:\lambda_j>0} \sum_{\substack{k:\, g_k=1 \\ 0<\lambda_k<C/2}} g_j \lambda_j \mathbf{x}_j^T \mathbf{x}_k - \frac{1}{\tilde{N}_{-1}} \sum_{j:\lambda_j>0} \sum_{\substack{k:\, g_k=-1 \\ 0<\lambda_k<C/2}} g_j \lambda_j \mathbf{x}_j^T \mathbf{x}_k \right). \end{cases} \quad (2.13)$$

It is just this aspect of the support vector method which will be exploited below in the problem of protein fold class recognition.

Although the solution of the dual problem (2.10) does not give the shift values δ_j for the elements of the training sample, it is easy to find them from (2.11) and the active constraints $g_j(\mathbf{a}^T \mathbf{x}_j + b) = 1 - \delta_j$ once the parameters of the optimal discriminant hyperplane $\mathbf{a}$ and b are found:

$$\delta_j = \begin{cases} 0 \ \text{ when } \ \lambda_j < \frac{1}{2}C, \\ 1 - g_j(\mathbf{a}^T \mathbf{x}_j + b) \ \text{ when } \ \lambda_j = \frac{1}{2}C. \end{cases} \quad (2.14)$$

It is clear that in the case of a linearly separable training sample all the shifts will equal zero.

Let us show now that the maximum value of the dual objective function $\max_{\lambda_1,\dots,\lambda_N} W(\lambda_1,\dots,\lambda_N)$ equals the half of the minimum value of the original criterion (2.4), whence it will be easy to find the final margin ξ of separating the points of the first and the second class by the optimal discriminant hyperplane after the "nuisance" points are shifted. Indeed, the equalities $g_j(\mathbf{a}^T \mathbf{x}_j + b) + \delta_j = 1$ are held for the active constraints with zero values of Lagrange multipliers. Hence, the first sum in the objective function $W(\lambda_1,\dots,\lambda_N)$ (2.10) at the point of its maximum value may be represented in the form

$$\sum_{j=1}^N \lambda_j = \sum_{j=1}^N \lambda_j \big(g_j(\mathbf{a}^T \mathbf{x}_j + b) + \delta_j \big) = \sum_{j=1}^N \lambda_j g_j \mathbf{a}^T \mathbf{x}_j + \left(\sum_{j=1}^N \lambda_j g_j \right) b + \sum_{j=1}^N \lambda_j \delta_j .$$

Here the first summand is the minimum possible value of the squared norm of the direction vector $\mathbf{a}^T \mathbf{a} = 1/\xi^2$ on the force of (2.7), the second summand equals zero on the force of (2.8), and, on the force of (2.9), the last summand has the form $\frac{C}{2} \sum_{j=1}^N \delta_j - \sum_{j=1}^N \mu_j \delta_j$, where only the first sum may differ from zero, because $\mu_j = 0$ if $\delta_j > 0$ and conversely. The second sum in (2.10), on the force of (2.8) again, equals $(\mathbf{a}^T\mathbf{a})/2 = 1/(2\xi^2)$. Thus, $\max_{\lambda_1,\ldots,\lambda_N} W(\lambda_1,\ldots,\lambda_N) = \frac{1}{2}\left(\frac{1}{\xi^2} + C\sum_{j=1}^{N}\delta_j\right)$, what gives

$$\frac{1}{\xi^2} = 2\max_{\lambda_1,\ldots,\lambda_N} W(\lambda_1,\ldots,\lambda_N) - C\sum_{j=1}^{N}\delta_j .$$

The fact that the direction vector of the discriminant hyperplane (2.7) and (2.21) is determined only by the support elements of the training sample $\lambda_j > 0$ makes the essence of the method of machine learning called the support vector method. There is a lot of philosophy in the literature about how support vectors should be interpreted and why this method has become so popular in the last few years. One of the reasons is that the training problem reduces to the standard quadratic programming problem (2.4) or (2.10), and that the ratio of the number of support vectors to the size of the training sample is a good estimate on the error rate of the discriminant hyperplane when applied to the general population.

3. Similarity-based featureless protein fold class recognition

3.1. The pair-wise alignment score of two amino acid chains as their inner product in an imaginary Hilbert space

The fact that features themselves do not occur in the decision rule inferred from a training sample by the support vector method (2.13) leads to the suggestion that a pair-wise similarity measure of proteins which would possess the properties of inner product is required for fold class recognition rather than a set of features.

It appears to be enough to assume that there exists a linear space with inner product into which all proteins may be mapped, at least, mentally. This imaginary linear space may be "richer" than the set of all proteins, i.e. contain "redundant" points, but the value of the inner product for pairs of points that correspond to really existing proteins should coincide with the accepted similarity measure.

Linear spaces with inner product are called Hilbert spaces, therefore, we call such an approach to pattern recognition featureless recognition in a Hilbert space.

When studying the structure and properties of proteins, one of commonly used instruments is the characteristic of mutual similarity of two amino acid sequences $\omega' = (a_1,\ldots,a_N)$ and $\omega'' = (b_1,\ldots,b_K)$ given by an appropriate pair-wise alignment procedure (Fig. 3.1). Procedures of such a kind lean upon a preset similarity matrix for all 210 pairs of 20 amino acids. Such matrices are called substitution matrices and characterize each amino acid pair (a,b) by logarithmic

```
ω′ : TNPGNASSTTTTKPTTTS-----RGLKTINETDPCIKNDSCTG
ω″ : GS----ATSTPATSTTAGTKLPCVRNKTDSNLQSCNDTIIEKE
 i = 12    34567  ···
```

Figure 3.1. Fragment of an aligned pair of amino acid chains from the protein family *Envelope glycoprotein GP120* in the database Pfam.

ratio of, first, the probability of their independent occurrence in two amino acid chains p_{ab} as result of evolutionary substituting the same unknown amino acid c in a common ancestor chain, and, second, the product of general probabilities q_a and q_b of their occurrence in arbitrary sequences [8]:

$$s(a,b) = \log \frac{p_{ab}}{q_a q_b}. \tag{3.1}$$

The log likelihood ratio $s(a,b)$ is positive if the probability that these two amino acids have a common ancestor is greater than the product of their general probabilities, equals zero in the indifferent case, and is negative if the hypothesis of their common origin is less likely than that of the null hypothesis of their independent occasional appearance.

There are several versions of substitution matrices [8,9,10], but each of them is result of observations in large sets of proteins aligned in that or other manner by experienced biologists in accordance with their intuition based, in its turn, on that or other model of evolution.

It appears natural to interpret the log likelihood ratio for two amino acids $s(a,b)$ (3.1) as experimentally registered outward exhibition of the actual proximity of their hidden properties. Let these properties be expressed by some hidden vectors $\mathbf{y}_a$ and $\mathbf{y}_b$ for which the notion of inner product is defined $(\mathbf{y}_a, \mathbf{y}_b)$, then the structure of (3.1) suggests the idea to consider $s(a,b)$ as a rough measure of it: $s(a,b) \cong (\mathbf{y}_a, \mathbf{y}_b)$.

The numerical measure of the proximity of two proteins represented by their amino acid chains is determined as the greatest possible sum of $s(a_{j_i}, b_{k_i})$ over all related pairs of amino acids (j_i, k_i), $i = 1,2,3,...$, in a pair-wise alignment with respect to some penalties posed on the presence and length of gaps (Fig. 3.1):

$$\mu(\omega', \omega'') = \sum_i s(a_{j_i}, b_{k_i}) + (\text{gap length penalties}).$$

By analogy to a single summand, the score of the alignment as a whole may also be interpreted as inner product of the respective combined feature vectors of two proteins $\mu(\omega', \omega'') \cong (\mathbf{x}_{\omega'}, \mathbf{x}_{\omega''})$ in an imaginary linear feature space. The greater the positive value of the similarity, the more "synchronous" are some essential properties of amino acids along the polypeptide chain, the zero value says about full lack of agreement what corresponds to the notion of orthogonality, and a negative value should be interpreted as "opposite phases" of amino acid properties along the chains.

However, this is not more than a cursory analogy. For an accurate justification of the hypothesis that there exists a Hilbert space in which the set of proteins could be embedded, we should show that the score matrix of any finite assembly of proteins tends to be nonnegative definite, i.e. is a Gram matrix or can be approximated by such a matrix.

We checked this hypothesis for an assembly of 396 proteins (Section 1, Table 1.1) by way of calculating all the eigenvalues of the score matrix obtained by the pair-wise alignment procedure Fasta3 [11,12] with the substitution matrix Blossum55. All the eigenvalues turned out to be positive (Fig. 3.2).

The conclusion suggests itself that the pair-wise similarity measure determined by the procedure Fasta3 possesses properties having much in common with those of inner product. This circumstance should be considered as a reason in favor of the theoretical applicability of the principle of featureless pattern recognition in a Hilbert space to the problem of protein fold class recognition.

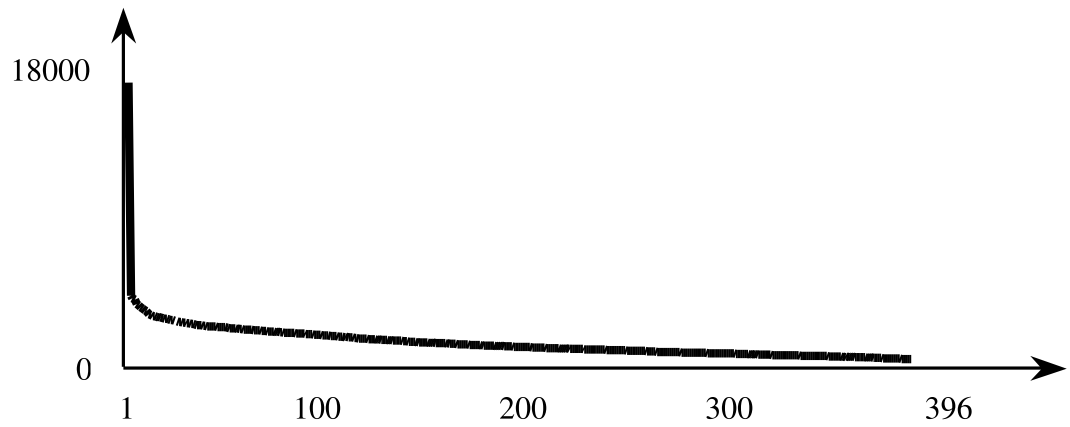

Figure 3.2. Eigenvalues of Dr.Kim's collection of proteins: $\lambda_{\max} = 16621$, $\lambda_{\min} = 304$; all eigenvalues are positive.

3.2. Hilbert space of classified objects and parametric family of linear decision rules

Let the set Ω of all feasible objects under consideration $\omega \in \Omega$ is partitioned into two classes $\Omega_1 = \{\omega \in \Omega : g(\omega) = 1\}$ and $\Omega_{-1} = \{\omega \in \Omega : g(\omega) = -1\}$ by an unknown indicator function $g(\omega) = \pm 1$. The main idea of the featureless approach to pattern recognition consists in treating the set Ω as a Hilbert space in which the linear operations and inner product are defined in an arbitrary manner under the usual constraints:

(1) addition is symmetric and associative $\omega' + \omega'' = \omega'' + \omega' \in \Omega$, $\omega' + (\omega'' + \omega''') = (\omega' + \omega'') + \omega'''$;

(2) there exists an origin $\phi \in \Omega$ such that $\omega + \phi = \omega$ for any element $\omega \in \Omega$;

(3) there exists the inverse elements $(-\omega) + \omega = \phi$ for any $\omega \in \Omega$;

(4) multiplication by a real coefficient $c\,\omega \in \Omega$, $c \in \mathrm{R}$, is associative $(cd)\omega = c(d\omega)$ and $1\omega = \omega$ for any $\omega \in \Omega$;

(5) addition and multiplication are distributive $c(\omega' + \omega'') = c\,\omega' + c\,\omega''$, $(c+d)\omega = c\omega + d\omega$;

(6) inner product of elements is symmetric $(\omega', \omega'') = (\omega'', \omega') \in \mathrm{R}$ and linear $(\omega, \omega' + \omega'') = (\omega, \omega') + (\omega, \omega'')$, $(\omega, c\,\omega') = c\,(\omega, \omega')$;

(7) inner product of an element with itself possesses the properties $(\omega, \omega) \ge 0$, $(\omega, \omega) = 0$ if and only if $\omega = \phi$ and gives the norm $\|\omega\| = (\omega, \omega)^{1/2} \ge 0$.

It is not meant that all the elements of the Hilbert space Ω do exist in reality. We consider really existing objects as making a subset $\tilde{\Omega}$ of isolated points in Ω, whereas all the remaining elements are nothing else than products of our imagination. It is just the extension of $\tilde{\Omega}$ to Ω what allows saying about "sums" of really existing objects and their "products" with real-valued coefficients.

It is assumed that even if an element of the Hilbert space $\omega \in \Omega$ really exists $\omega \in \tilde{\Omega} \subset \Omega$, it cannot be perceived by the observer in any other way than through its inner products (ω, ω') with other really existing elements $\omega' \in \tilde{\Omega} \subset \Omega$. If $\vartheta \in \Omega$ is a fixed element of the Hilbert space, an imaginary one in the general case, the real-valued linear discriminant function $d(\omega \mid \vartheta, b) = (\omega, \vartheta) + b$, where $b \in \mathsf{R}$ is a constant, may be used as decision rule $\hat{g}(\omega): \Omega \to \{1, -1\}$ of judging on the hidden class-membership of an arbitrary object $\omega \in \Omega$, might it really exist or not:

$$d(\omega \mid \vartheta, b) = (\omega, \vartheta) + b \begin{cases} > 0 \to \hat{g}(\omega) = 1, \\ < 0 \to \hat{g}(\omega) = -1. \end{cases} \tag{3.2}$$

Here the element $\vartheta \in \Omega$ plays the role of the direction element of the respective discriminant hyperplane in the Hilbert space $(\omega, \vartheta) + b = 0$.

However, we have, so far, no constructive instrument of choosing the direction element $\vartheta \in \Omega$ and, hence, the decision rule of recognition, because, just as any element of Ω, it can be defined only by its inner products with some other fixed elements that exist in reality.

Let the observer have chosen an assembly of really existing objects $\Omega^0 = \{\omega_1^0, ..., \omega_n^0\} \subset \Omega$, called the basic assembly, which is not assumed to be classified, in the general case, and, therefore, it is not yet a training sample. The basic assembly will play the role of a finite basis in the Hilbert space that defines an n-dimensional subspace $\Omega_n(\omega_1^0, ..., \omega_n^0) = \left\{\omega \in \Omega : \omega = \sum_{k=1}^{n} a_k \omega_k^0 \right\} \subset \Omega$.

We shall restrict our consideration to only those discriminant hyperplanes whose direction elements belong to $\Omega_n(\omega_1^0, ..., \omega_n^0)$, i.e. can be expressed in the form of linear combinations

$$\vartheta(\mathbf{a}) = \sum_{k=1}^{n} a_k \omega_k^0, \quad \mathbf{a} \in \mathsf{R}^n. \tag{3.3}$$

The respective parametric family of discriminant hyperplanes will be completely defined by inner products of their elements with the objects of the basic assembly: $\big(\omega, \vartheta(\mathbf{a})\big) + b = \sum_{k=1}^{n} a_k (\omega, \omega_k^0) + b = 0$.

Mark that if $\big(\omega, \vartheta(\mathbf{a})\big) + b = 0$ then $(\omega, \omega_k^0) = 0$ for all $\omega_k^0 \in \Omega^0$. This means that by choosing the direction vectors in accordance with (3.3) we restrict our consideration to only those discriminant hyperplanes which are orthogonal to the subspace spanned over the basic assembly of objects.

Each element of the Hilbert space $\omega \in \Omega$ is associated with the totality of its inner products with basic elements, which we shall consider as its real-valued “feature vector”

$$\mathbf{x}(\omega) = \big(x_1(\omega) \cdots x_n(\omega)\big)^T \in \mathsf{R}^n, \quad x_k(\omega) = (\omega, \omega_k^0). \tag{3.4}$$

The basis $\Omega^0 = \{\omega_1^0, ..., \omega_n^0\}$ is not assumed to be complete in the Hilbert space Ω, therefore, an arbitrary element $\omega \in \Omega$ cannot be represented as linear combinations of elements of the basic assembly, but the feature vector $\mathbf{x}(\omega)$ completely determines the projection of the element $\omega \in \Omega$ onto this subspace.

There exist many elements having the same feature vector $\mathbf{x}(\omega)$ and, so, the same projections. But, on the force of orthogonality of discriminant hyperplanes with direction elements (3.3) to $\Omega_n(\omega_1^0, ..., \omega_n^0)$, all such elements will be assigned the same class $\hat{g}(\omega) = \pm 1$ by the linear decision rule (3.2) with $\vartheta = \vartheta(\mathbf{a})$ of the form (3.3). Therefore, we shall call the features (3.4) projectional features of the Hilbert space elements.

We have come to a parametric family of decision rules of pattern recognition in a Hilbert space that lean upon projectional features of objects:

$$d\big(\mathbf{x}(\omega)\,|\,\mathbf{a},b\big) = \mathbf{a}^T\mathbf{x}(\omega) + b \begin{cases} > 0 \rightarrow \hat{g}(\omega) = 1, \\ < 0 \rightarrow \hat{g}(\omega) = -1, \end{cases} \quad \mathbf{a} \in \mathsf{R}^n, \quad \omega \in \Omega .$$

By its structure, such a family of decision rules completely corresponds to the family of linear decision rules (2.1) in Section 2.1. Thus, the notion of projectional features reduces, at least, superficially, the problem of featureless pattern recognition in a Hilbert space to the classical problem of pattern recognition in a usual linear space of real-valued features.

Indeed, let the observer be submitted a classified training sample of objects $\Omega^* = \{\omega_1,...,\omega_N\} \subset \Omega$, $g_1 = g(\omega_1),...,g_N = g(\omega_N)$, that does not coincide, in the general case, with the basic assembly $\Omega^0 = \{\omega_1^0,...,\omega_n^0\}$. The observer has no other way of perceiving them than to calculate their inner products with objects of the basic assembly, what is equivalent to evaluating their projectional features

$$\mathbf{x}(\omega_j) = \big(x_1(\omega_j)\cdots x_n(\omega_j)\big)^T = \big((\omega_j,\omega_1^0)\cdots(\omega_j,\omega_n^0)\big)^T \in \mathsf{R}^n .$$

The parameters of the discriminant hyperplane $\mathbf{a} \in \mathsf{R}^n$ and $b \in \mathsf{R}$ should be chosen so that the training objects would be classified correctly with a positive margin $\xi > 0$:

$$d\big(\mathbf{x}(\omega_j)\,|\,\mathbf{a},b\big) = \mathbf{a}^T\mathbf{x}(\omega_j) + b \geq \xi \;\text{ when }\; g(\omega_j) = 1,$$

$$d\big(\mathbf{x}(\omega_j)\,|\,\mathbf{a},b\big) = \mathbf{a}^T\mathbf{x}(\omega_j) + b \leq -\xi \;\text{ when }\; g(\omega_j) = -1.$$

This desire is completely the same as that expressed by conditions (2.2), and everything said in Section 2.1 appears to be applicable to the featureless case.

If there is no guarantee that the training sample is linearly separable, the “nuisance” objects should be ignored by mental “shifting” them towards their native classes. So, we come to the formulation of the training problem that is practically the same as (2.4):

$$\begin{cases} \|\,\vartheta(\mathbf{a})\,\|^2 + C\sum_{j=1}^{N}\delta_j \rightarrow \min, \\ g_j\big(\mathbf{a}^T\mathbf{x}(\omega_j) + b\big) \geq 1 - \delta_j, \quad \delta_j \geq 0, \quad j = 1,...,N. \end{cases}$$

However, in contrast to (2.4), the norm of the direction element of the sought-for hyperplane can be understood, at least, in two ways, namely whether as that of an element of the Hilbert space $\|\,\vartheta(\mathbf{a})\,\|^2 = \big(\vartheta(\mathbf{a}),\vartheta(\mathbf{a})\big)$, $\vartheta(\mathbf{a}) \in \Omega$, or as the norm of the direction vector in the space of projectional features $\|\,\vartheta(\mathbf{a})\,\|^2 = \|\,\mathbf{a}\,\|^2 = \mathbf{a}^T\mathbf{a}$, $\mathbf{a} \in \mathsf{R}^n$. These two versions of norm lead to techniques of training and decision making that essentially differ in their structure.

3.3. Support vector method of machine learning from a classified training sample in a Hilbert space

Absence of a priori preferences regarding the direction element of the discriminant hyperplane in the Hilbert space. Let us consider first the “native” version of norm, namely, that measured immediately in the Hilbert space: $\|\,\vartheta(\mathbf{a})\,\|^2 = \big(\vartheta(\mathbf{a}),\vartheta(\mathbf{a})\big)$. In this case, the training technique is aimed at minimizing the norm of the direction element, and, so, all orientations are equally preferable.

In accordance with (3.3)

$$\left(\vartheta(\mathbf{a}),\vartheta(\mathbf{a})\right)^2=\sum_{j=1}^{n}\sum_{k=1}^{n}(\omega_j^0,\omega_k^0)a_j a_k=\mathbf{a}^T\mathbf{M}\mathbf{a},$$

where $\mathbf{M}$ is matrix $(n\times n)$ formed by inner products of basic elements $\omega_1^0,...,\omega_n^0$:

$$\mathbf{M}=\begin{pmatrix}(\omega_1^0,\omega_1^0) & \cdots & (\omega_1^0,\omega_n^0)\\ \vdots & \ddots & \vdots\\ (\omega_n^0,\omega_1^0) & \cdots & (\omega_n^0,\omega_n^0)\end{pmatrix}.$$

Thus, we come to the following version of the training problem:

$$\begin{cases}\mathbf{a}^T\mathbf{M}\mathbf{a}+C\sum_{j=1}^{N}\delta_j\to\min,\\ g_j\left(\mathbf{a}^T\mathbf{x}(\omega_j)+b\right)\geq 1-\delta_j,\quad \delta_j\geq 0,\quad j=1,...,N.\end{cases}\tag{3.5}$$

The Lagrangian of this quadratic programming problem has the form:

$$L(\mathbf{a},b,\lambda_1,...,\lambda_N,\delta_1,...,\delta_N,\mu_1,...,\mu_N)=\frac{1}{2}\mathbf{a}^T\mathbf{M}\mathbf{a}+\frac{C}{2}\sum_{j=1}^{N}\delta_j-$$
$$-\sum_{j=1}^{N}\lambda_j\left[g_j\left(\mathbf{a}^T\mathbf{x}(\omega_j)+b\right)-1+\delta_j\right]-\sum_{j=1}^{N}\mu_j\delta_j,$$

where, just as in (2.5), $\lambda_j\geq 0$ and $\mu_j\geq 0$, $j=1,...,N$, are nonnegative Lagrange multipliers. The condition

$$L(\mathbf{a},b,\lambda_1,...,\lambda_N,\delta_1,...,\delta_N,\mu_1,...,\mu_N)\to\min\quad\text{with respect to }\ \mathbf{a},b,\delta_1,...,\delta_N$$

gives the same equalities (2.8) and (2.9), but instead of (2.7) we have

$$\mathbf{M}\mathbf{a}=\sum_{j=1}^{N}\lambda_j g_j\mathbf{x}(\omega_j)\quad\text{and}\quad \mathbf{a}=\mathbf{M}^{-1}\sum_{j=1}^{N}\lambda_j g_j\mathbf{x}(\omega_j).\tag{3.6}$$

The same manipulations as in Section 2.2 result in the dual formulation of the training problem:

$$\begin{gathered}W(\lambda_1,...,\lambda_N)=\sum_{j=1}^{N}\lambda_j-\frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}\left(g_j g_k\mathbf{x}^T(\omega_j)\mathbf{M}^{-1}\mathbf{x}(\omega_k)\right)\lambda_j\lambda_k\to\max,\\ \sum_{j=1}^{N}\lambda_j g_j=0,\quad 0\leq\lambda_j\leq\frac{1}{2}C,\ j=1,...,N.\end{gathered}\tag{3.7}$$

The Lagrange multipliers $\lambda_1\geq 0,...,\lambda_N\geq 0$, to be exact, those of them, which differ from zero $\lambda_j>0$, form the decision rule of recognition applicable to any object $\omega\in\Omega$ in terms of its projectional features $\mathbf{x}(\omega)=\left((\omega,\omega_1^0)\cdots(\omega,\omega_n^0)\right)^T\in\mathsf{R}^n$ and those of the support elements of the training sample $\mathbf{x}(\omega_j)=\left((\omega_j,\omega_1^0)\cdots(\omega_j,\omega_n^0)\right)^T\in\mathsf{R}^n$, $\omega_j\in\mathbf{\Omega}^*=\{\omega_1,...,\omega_N\}$ (remember, here $\Omega^0=\{\omega_1^0,...,\omega_n^0\}$ is the basic assembly):

$$d\left(\mathbf{x}(\omega)\right)=\sum_{j:\lambda_j>0}g_j\lambda_j\mathbf{x}^T(\omega)\mathbf{M}^{-1}\mathbf{x}(\omega_j)+b\begin{cases}>0\ \to\ \hat{g}(\omega)=1,\\ <0\ \to\ \hat{g}(\omega)=-1.\end{cases}\tag{3.8}$$

The constant b can be found in the same way as in the case of a linear feature space (2.12) with respect to (3.6).

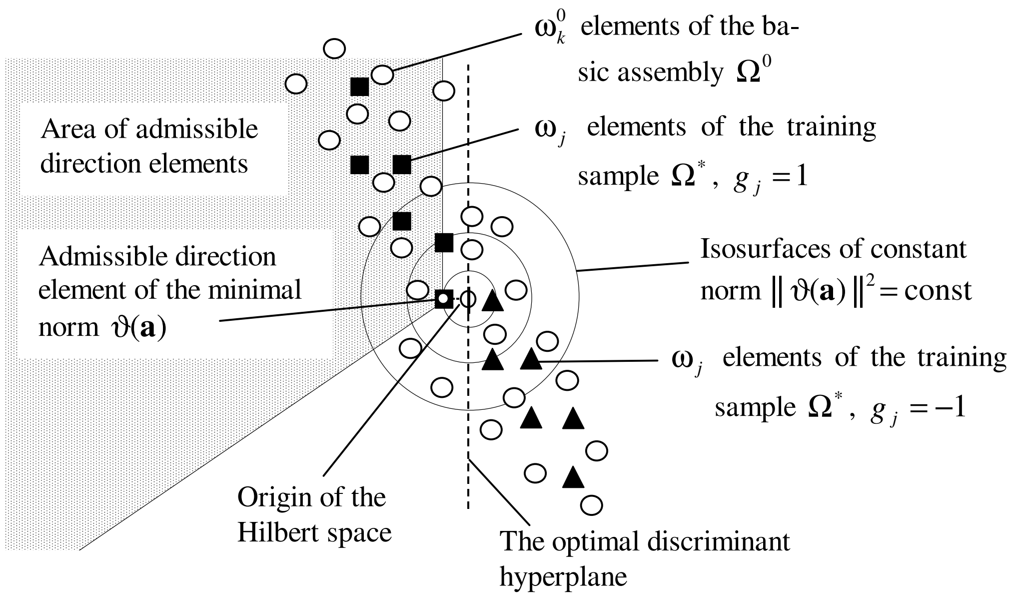

Figure 3.1. Minimal norm of the direction element in the original Hilbert space as criterion of training. There are no preferences regarding its orientation, and the criterion ignores the fact that the basic assembly is stretched in the space.

It should be marked that the decision rule (3.8) depends on projection feature vectors $\mathbf{x}(\omega_j) = \big((\omega_j, \omega_1^0) \cdots (\omega_j, \omega_n^0)\big)^T$ of only those objects in the training sample for which $\lambda_j > 0$, and it is just these vectors that are called support ones. Thus, the decision rule is formed by all objects of the basic assembly and support objects of the training sample.

Fig. 3.1 gives a schematic illustration of what happens in the Hilbert space of objects if the norm of the direction element subject to minimization is measured in the original Hilbert space $\| \vartheta(\mathbf{a}) \|^2 \to \min$. In this case, all orientations of the sought-for direction element are equally preferable, and the training technique does not pay attention to the fact that the basic assembly of objects is stretched in space.

In the particular case when the training sample coincides with the basic assembly, so that $N = n$ and $\Omega^* = \Omega^0 = \{\omega_1, ..., \omega_N\}$, $g_1 = g(\omega_1), ..., g_N = g(\omega_N)$, the matrix of inner products of objects making the training sample will play the role of $\mathbf{M}$:

$$\mathbf{M} = \begin{pmatrix} (\omega_1, \omega_1) & \cdots & (\omega_1, \omega_N) \\ \vdots & \ddots & \vdots \\ (\omega_N, \omega_1) & \cdots & (\omega_N, \omega_N) \end{pmatrix}.$$

The projectional features of objects in the training sample $\mathbf{x}_j = \mathbf{x}(\omega_j) = \big((\omega_j, \omega_1) \cdots (\omega_j, \omega_N)\big)^T \in \mathsf{R}^N$, $j = 1, ..., N$, will be formed by columns, or, what is the same, rows of the symmetric matrix $\mathbf{M}$ $(N \times N)$, so that $\mathbf{M} = (\mathbf{x}_1 \cdots \mathbf{x}_N)$.

Mark that, in this case, it follows from

$$\mathbf{M}^{-1}\mathbf{M} = \mathbf{M}^{-1}(\mathbf{x}_1 \cdots \mathbf{x}_N) = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix} = \mathbf{I}$$

that $\mathbf{M}^{-1}\mathbf{x}_k$ is column vector, all elements of which equal zero except the kth element that equals unity. Hence, the quadratic form $\mathbf{x}_j^T\mathbf{M}^{-1}\mathbf{x}_k$ that occurs in the objective function of the dual problem (3.7) has the same value as the kth element of vector $\mathbf{x}_j$, i.e. (ω_j, ω_k). Thus, the dual problem becomes especially simple:

$$W(\lambda_1,...,\lambda_N) = \sum_{j=1}^{N}\lambda_j - \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}\big(g_j g_k(\omega_j,\omega_k)\big)\lambda_j\lambda_k \to \max,$$

$$\sum_{j=1}^{N}\lambda_j g_j = 0, \quad 0 \le \lambda_j \le \frac{1}{2}C, \ j = 1,...,N.$$

Especially simple will be also the decision rule of recognition, because we have in (3.8) $\mathbf{x}^T(\omega)\mathbf{M}^{-1}\mathbf{x}(\omega_j) = (\omega, \omega_j)$ for an arbitrary object $\omega \in \Omega$, and, so,

$$d(\omega) = \sum_{j:\lambda_j > 0} g_j\lambda_j(\omega,\omega_j) + b \begin{cases} > 0 \ \to \ \hat{g}(\omega) = 1, \\ < 0 \ \to \ \hat{g}(\omega) = -1. \end{cases}$$

The decision rule leans upon inner products of the new object with only those objects of the training sample which are marked by non-zero values of Lagrange multipliers. Such objects, that should be called support ones, participate in the decision rule immediately but not through any features.

It is just this situation that was considered by R. Duin and colleagues [3,4,5], and it is just these authors who introduced the notion of support objects in featureless pattern recognition in addition to that of support vectors in the traditional formulation of the pattern recognition problem.

Preferred orientation of the direction element of the discriminant hyperplane along the major inertia axis of the basic assembly. The alternative way in which the norm of the sought-for direction element may be measured is the norm $\|\mathbf{a}\|^2 = \mathbf{a}^T\mathbf{a}$ of the coefficient vector of its representation in the space R^n of projectional features (3.3). In this case the problem is no different from that formulated for the usual linear feature space (2.4) with $\mathbf{x}_j = \mathbf{x}(\omega_j)$

$$\begin{cases} \mathbf{a}^T\mathbf{a} + C\sum_{j=1}^{N}\delta_j \to \min, \\ g_j\big(\mathbf{a}^T\mathbf{x}(\omega_j) + b\big) \ge 1 - \delta_j, \quad \delta_j \ge 0, \quad j = 1,...,N, \end{cases} \tag{3.9}$$

and, in the dual formulation, from (2.10):

$$\begin{aligned} & W(\lambda_1,...,\lambda_N) = \sum_{j=1}^{N}\lambda_j - \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}\big(g_j g_k \mathbf{x}^T(\omega_j)\mathbf{x}(\omega_k)\big)\lambda_j\lambda_k \to \max, \\ & \sum_{j=1}^{N}\lambda_j g_j = 0, \quad 0 \le \lambda_j \le \frac{1}{2}C, \ j = 1,...,N. \end{aligned} \tag{3.10}$$

In accordance with (2.7), the optimal decision rule of recognition will have the structure

$$d\big(\mathbf{x}(\omega)\big)=\sum_{j:\lambda_j>0} g_j\lambda_j \mathbf{x}^T(\omega)\,\mathbf{x}(\omega_j)+b\begin{cases}>0 \;\rightarrow\; \hat{g}(\omega)=1,\\ <0 \;\rightarrow\; \hat{g}(\omega)=-1,\end{cases}\qquad(3.11)$$

in which the projectional feature vectors of the new object $\mathbf{x}(\omega)=\big((\omega,\omega_1^0)\cdots(\omega,\omega_1^0)\big)^T$ and of the support objects in the training sample $\mathbf{x}(\omega_j)=\big((\omega_j,\omega_1^0)\cdots(\omega_j,\omega_1^0)\big)^T$ occur in absolutely the same manner as in (2.13).

It is easy to see that if $\vartheta\in\Omega$ and $\omega\in\Omega$ are two arbitrary elements of the Hilbert space Ω then the distance from ω to its projection onto the beam determined by the element ω equals $(\omega,\omega)-\dfrac{(\omega,\vartheta)^2}{(\vartheta,\vartheta)}$. We are going to show that the criterion $\mathbf{a}^T\mathbf{a}\rightarrow\min$ among direction elements with the same squared norm $\big(\vartheta(\mathbf{a}),\vartheta(\mathbf{a})\big)=\mathbf{a}^T\mathbf{M}\mathbf{a}=\text{const}$ is equivalent to the criterion $\sum_{j=1}^n\big(\omega_j,\vartheta(\mathbf{a})\big)^2\rightarrow\max$. This will mean that training by criterion $\mathbf{a}^T\mathbf{a}\rightarrow\min$ prefers discriminant hyperplanes whose direction elements are close to the major inertia axis of the basic assembly $\Omega^0=\{\omega_1^0,\dots,\omega_n^0\}$.

First of all, let us make sure that the condition $\mathbf{a}^T\mathbf{a}\rightarrow\min$ under the constraint $\big(\vartheta(\mathbf{a}),\vartheta(\mathbf{a})\big)=\mathbf{a}^T\mathbf{M}\mathbf{a}=\text{const}$ is provided by the major eigenvector of $\mathbf{M}$. Indeed, the Lagrangian for this minimization problem has the form

$$L'(\mathbf{a},\beta')=\frac{1}{2}\mathbf{a}^T\mathbf{a}-\frac{1}{2\beta'}(\mathbf{a}^T\mathbf{M}\mathbf{a}-\text{const}),$$

and the equation

$$\nabla_{\mathbf{a}}L(\mathbf{a},\beta')=\mathbf{a}-\frac{1}{\beta'}\mathbf{M}\mathbf{a}=\mathbf{0}$$

leads to the condition $\mathbf{M}\mathbf{a}=\beta'\mathbf{a}$ that, in its turn, gives the equality $\mathbf{a}^T\mathbf{M}\mathbf{a}=\beta'\mathbf{a}^T\mathbf{a}=\text{const}$. Because here $\mathbf{a}^T\mathbf{a}\rightarrow\min$, it is inevitable that $\beta'\rightarrow\max$, what means that the latter equality is provided by the greatest eigenvalue of matrix $\mathbf{M}$.

On the other hand, the major eigenvector of $\mathbf{M}$ satisfies the condition $\mathbf{a}^T\mathbf{M}\mathbf{M}\mathbf{a}\rightarrow\max$ under the constraint $\mathbf{a}^T\mathbf{M}\mathbf{a}=\text{const}$. To show this, it is enough to differentiate the Lagrangian

$$L''(\mathbf{a},\beta'')=\frac{1}{2}\mathbf{a}^T\mathbf{M}\mathbf{M}\mathbf{a}-\frac{1}{2}\beta''(\mathbf{a}^T\mathbf{M}\mathbf{a}-\text{const})$$

with respect to $\mathbf{a}$, what will lead to the equality

$$\nabla_{\mathbf{a}}L(\mathbf{a},\beta'')=\mathbf{M}\mathbf{M}\mathbf{a}-\beta''\mathbf{M}\mathbf{a}=\mathbf{M}(\mathbf{M}\mathbf{a}-\beta''\mathbf{a})=\mathbf{0}\;\text{ if }\;\mathbf{M}\mathbf{a}=\beta''\mathbf{a}.$$

So, the requirement $\mathbf{a}^T\mathbf{M}\mathbf{M}\mathbf{a}\rightarrow\max$ is equivalent to the requirement $\beta''\mathbf{a}^T\mathbf{M}\mathbf{a}\rightarrow\max$, which, on the force of the constraint $\mathbf{a}^T\mathbf{M}\mathbf{a}=\text{const}$, is provided by the greatest eigenvalue of $\mathbf{M}$.

Thus, if $\mathbf{a}^T\mathbf{a}\rightarrow\min$ under the constraint $\big(\vartheta(\mathbf{a}),\vartheta(\mathbf{a})\big)=\mathbf{a}^T\mathbf{M}\mathbf{a}=\text{const}$ then $\mathbf{a}^T\mathbf{M}\mathbf{M}\mathbf{a}\rightarrow\max$. Here $\mathbf{a}\mathbf{M}=\mathbf{a}^T(\mathbf{x}_1\cdots\mathbf{x}_N)$, where $\mathbf{x}_j$ is a single column of the symmetric matrix $\mathbf{M}$, and $\mathbf{a}^T\mathbf{M}=(\mathbf{M}\mathbf{a})^T$, so that $\mathbf{a}^T\mathbf{M}\mathbf{M}\mathbf{a}=\sum_{j=1}^n\mathbf{a}^T\mathbf{x}_j\mathbf{x}_j^T\mathbf{a}=\sum_{j=1}^n(\mathbf{x}_j^T\mathbf{a})^2$. In its turn, $\mathbf{x}_j^T\mathbf{a}=\sum_{i=1}^n a_i(\omega_j,\omega_i)=\big(\omega_j,\sum_{i=1}^n a_i\omega_i\big)=\big(\omega_j,\vartheta(\mathbf{a})\big)$, and, so, $\mathbf{a}^T\mathbf{M}\mathbf{M}\mathbf{a}=\sum_{j=1}^n\big(\omega_j,\vartheta(\mathbf{a})\big)^2$.

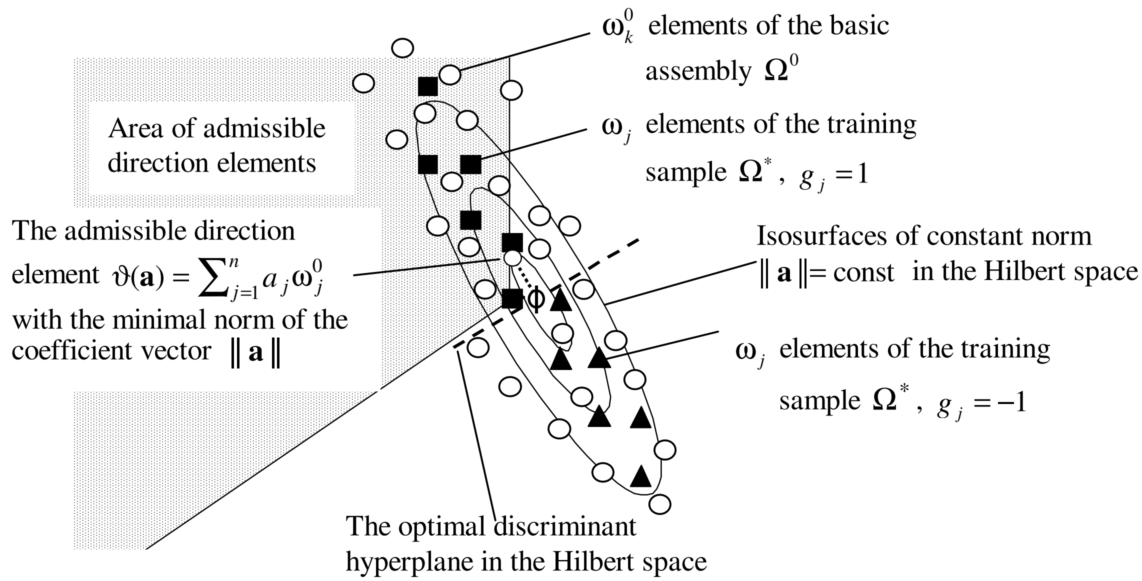

Figure 3.2. Minimal norm of the direction vector of the discriminant hyperplane in the space of projectional features as criterion of training. In the original Hilbert space, the discriminant hyperplanes are preferred whose direction elements are oriented along the major inertia axis of the basic assembly.

This means that if $\mathbf{a}^T\mathbf{a} \to \min$ under the constraint $(\vartheta(\mathbf{a}), \vartheta(\mathbf{a})) = \mathbf{a}^T\mathbf{M}\mathbf{a} = \text{const}$ then $\sum_{j=1}^{n} (\omega_j, \vartheta(\mathbf{a}))^2 \to \max$, and, so, $\vartheta(\mathbf{a})$ tends to be close to the major inertia axis of the basic assembly. It is just this what we wanted to show.

Thus, training by criterion (2.4), (2.10). i.e. without any preferences in the space of projectional features, is equivalent to a pronounced preference in the original Hilbert space in favor of direction elements oriented along the major inertia axis of the basic assembly of object. As a result, the discriminant hyperplane in the Hilbert space tends to be orthogonal to that axis (Fig. 3.2).

This is out of significance if the region of major concentration of objects in the Hilbert space is equally stretched in all directions. But such indifference is rather an exclusion than a rule. It is natural to expect the distribution of objects be differently extended in different directions, what fact will be reflected by the form of the basic assembly and, then, by the training sample. In this case, a reliable decision rule of recognition exists only if objects of two classes are spaced just in one of the directions where the extension is high. Therefore, it appears reasonable to escape discriminant hyperplanes oriented along the basic assembly even if the gap between the points of the first and the second class in the training sample has such an orientation (Fig. 3.1), and prefer transversal hyperplanes (Fig. 3.2). It is just this preference that is expressed by the training criterion $\mathbf{a}^T\mathbf{a} \to \min$ in contrast to $(\vartheta(\mathbf{a}), \vartheta(\mathbf{a})) = \mathbf{a}^T\mathbf{M}\mathbf{a} \to \min$.

3.4. Estimation of the statistical stability of the decision rule inferred from a training sample: The leave-one-out procedure

The quality of the decision rule, in this case, the discriminant hyperplane in the Hilbert space $(\omega, \vartheta(\mathbf{a})) + b = 0$ or in the space of projectional features $\mathbf{a}^T \mathbf{x}(\omega) + b = 0$, is assessed by the probability of errors $\mathrm{P}[\hat{g}(\mathbf{x}(\omega)) \neq g(\omega)]$ under the assumption that objects $\omega \in \Omega^*$ occur in accordance with the probabilistic properties of the general population. However, these properties are unknown to the observer, and, so, the quality of the decision rule cannot be estimated in such a straightforward way.

The portion of objects that fall at the wrong side of the optimal discriminant hyperplane in the training sample itself is on no account a correct estimate of the decision rule quality, because it was found as that providing the minimum value of the criterion which immediately depends on the number of errors. The more so, a training sample that contains the same or a lesser number of objects than the dimensionality of the feature space will always be separable by a hyperplane independently of how the trainer allocated the objects over two classes, but it does not mean that this hyperplane will give no errors when applied to new objects that did not participate in training. In particular, such a situation is typical for featureless pattern recognition, because the size of the training sample will practically always be less than that of the basic assembly, and, so, than the dimensionality of the space of projectional features.

The probability of errors can be estimated only from a sufficiently large examination sample from the same general population that the training sample was taken from.

On the other hand, the greater the training sample, the better it reflects the general population, the higher is the quality of the decision rule inferred from it, therefore, it is always preferable to use for training the entire available amount of classified objects without dividing it into the training and the examination sample.

The well-known leave-one-out procedure [7] offers a way of empirical estimating the quality of the decision rule immediately from the training sample (Fig. 3.3). One of the objects of the full training sample containing N objects is left out at the stage of training, and the decision rule inferred from the remaining $N-1$ objects is applied to the left-out one. If the result of recognition coincides with the actual class given by the trainer, this fact is registered as success at the stage of examination, otherwise an error is fixed. Then the control object is returned to the training sample, another one is left out, and the experiment is run again. Such a procedure is applied to all the objects of the training sample, and the number of errors N' is calculated. The ratio of the number of misclassified objects to the sample size $p = N'/N$ is considered as an estimate on the error rate of the decision rule inferred from the full sample would it be applied to the general population.

The decision rule of recognition, in this case, the discriminant hyperplane $\mathbf{a}^T \mathbf{x}(\omega) + b = 0$, is inferred from a finite training sample $\Omega^* = \{\omega_1, ..., \omega_N\} \subset \Omega$, $g_1 = g(\omega_1), ..., g_N = g(\omega_N)$, which should be considered as a random set whose probabilistic properties are determined by those of the general population. Thus, the parameters of the optimal discriminant hyperplane are random values $\mathbf{a}(\Omega^*)$, $b(\Omega^*)$, and so will be also its quality expressed by the error probability $\mathrm{P}[\hat{g}(\mathbf{x}(\omega)) \neq g(\omega)]$ which exists but is unknown to the observer. The mathematical expectation $\mathrm{M}\{\mathrm{P}[\hat{g}(\mathbf{x}(\omega)) \neq g(\omega)]\}$ is a characteristic of the "quality" of the training method.

But the estimate of this probability $p = N'/N$ obtained by the leave-one-out method will be a random value, too. It is well-known fact [13] that this is an unbiased estimate of the error

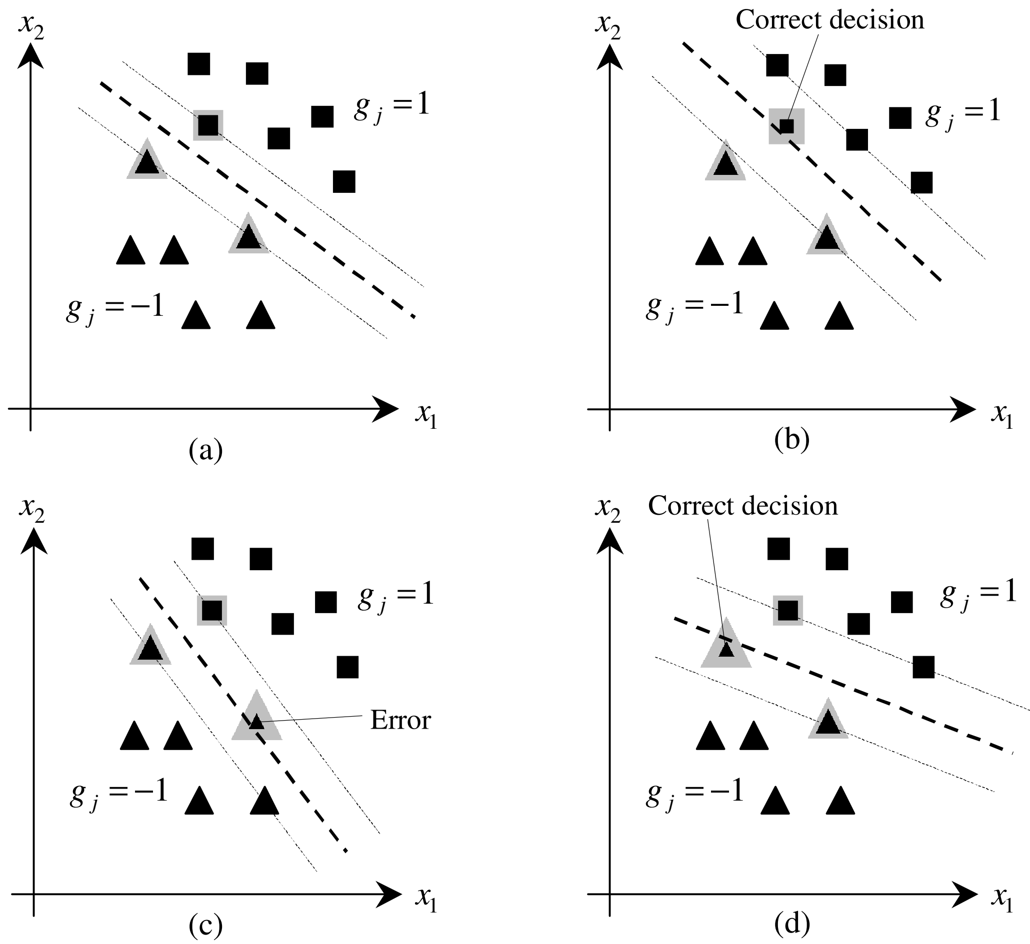

Figure 3.3. The discriminant hyperplane inferred from the whole training sample (a), and the leave-one-out procedure applied to the three support objects (b), (c) and (d).

probability, so that $\mathrm{M}(p) = \mathrm{M}\{\mathrm{P}[\hat{g}(\mathbf{x}(\omega)) \neq g(\omega)]\}$. It is just this sense in which the result of the leave-one-out procedure should be interpreted.

In the general case, the leave-one-out procedure leads to a large amount of calculations, because when the training sample consists of N objects, the training algorithm is to be run N times. However, if the support vector method is applied, there is no need to start it anew for each left-out object. The matter is that the discriminant hyperplane is determined only by support objects, to be exact, by their projectional features in the case of featureless pattern recognition, and removing any other object cannot change it. Because only those objects can be non-support ones which are classified correctly by the optimal discriminnat hyperplane that remains the same if one of them is left out, no modification of the training sample on the cost of be non-support objects can change the result of classification.

Thus, it is enough to apply the leave-one-out procedure only to the objects proclaimed support ones by the training procedure applied to the full training sample, as it is shown in Fig. 3.3, since only these objects may be classified wrongly having been left out of training.

4. Regularization of the training process

4.1. The curse of a small sample

The necessity of inferring a decision rule of recognition from a training sample of a small size, or, what is the same, in the feature space of a large dimensionality, is typical for the majority of applications. When the number of features is large, it is very likely that the number of errors at the stage of training will be small, but extrapolational properties of the training technique become doubtful, because the decision rule is bound to specific peculiarities of the given training sample and poorly reflects the properties of the general population as a whole. The main idea of training consists not in retaining the whole amount of information contained in the training sample but in minimizing the attention paid to occasional details.

The necessity of overcoming the small size of the training sample inevitably arises when solving the problem of protein fold class recognition. The quest for as a detailed characterization of an amino acid chain as it is possible in terms of the vector of its similarities to some fixed basic chains, leads to the desirability of enlarging the size of the basic assembly and, hence, the dimensionality of the space of projectional features. At the same time, the number of known proteins of a certain fold class is, as a rule, very small, and, hence, so is also the size of the training sample.

If the number of objects in the training sample equals to or is less than the dimensionality of the space of features, in this case, the projectional ones, the errorless discriminant hyperplane exists for any spatial distribution of the objects of two classes. However, the linear separability of a small training sample is deceptive, because it does not mean that there exists a hyperplane giving an acceptable error rate on the general population as a whole.

The lack of information on the properties of the general population which is contained in the training sample may be replaced only in case there exists another source of such information, a priori one relative to the sample. Methods of correcting the training procedure on the basis of taking into account some a priori information in addition to the experimental one contained in the training sample are called methods of regularization of training.

In this Report, we consider a way of regularization that follows from the very idea of featureless pattern recognition in an imaginary Hilbert space of objects. The way being proposed consists in decreasing the freedom of varying the direction element of the discriminant hyperplane, when adjusting it to the training sample, on the basis of taking into account a specificity of projectional features of proteins.

Actually, training by criterion $\mathbf{a}^T\mathbf{a} \to \min$ instead of $\big(\vartheta(\mathbf{a}), \vartheta(\mathbf{a})\big) = \mathbf{a}^T\mathbf{M}\mathbf{a} \to \min$ in Section 3.3 is nothing else than a regularization method that makes use of the information on the form of the region of major concentration of objects in the Hilbert space. This information is taken from the basic assembly and, so, should be considered as a priori information relative to the training sample. Since the basic assembly, in contrast to the training sample, usually consists of quite a large number of objects, it represents well enough the hidden general probability distribution in the Hilbert space. In case that distribution is almost degenerate in some directions, it is reasonable to prefer discriminant hyperplanes of transversal orientation even if the training sample suggests the longitudinal one as it is shown in Fig. 3.1 and 3.2.

In the next Section, we consider another source of a priori information that may be drawn from the basic assembly of objects before processing the training sample. The respective regularization method follows from the very nature of projectional features, namely, from the suggestion that the closer are two objects of the basic assembly, the less should be the difference between the coefficients of their participating in the direction element of the discriminant hyperplane (3.3).

4.2. The smoothness principle of regularization in the space of projectional features

The discriminant hyperplane in the space of projectional features is an expression of the linear decision rule that compares a linear combination of features $\sum_{j=1}^{n} a_j x_j(\omega) = \sum_{j=1}^{n} a_j(\omega, \omega_j^0)$ with a threshold $-b$. The coefficients a_j each of which is associated with an object of the basic assembly $\omega_j^0 \in \Omega^0$ form a vector $\mathbf{a} \in \mathsf{R}^n$ that indicates the direction in the space of projectional features along which the two classes to be separated are displaced relative each other. Finding this direction is just the main task of the training procedure.

In the feature space of an arbitrary nature, there are no a priori preferences in favor of that or other mutual arrangement of classes, and the only source of information on the sought-for direction is the training sample. But in the space of projectional features different directions are not equally probable, and it is just this fact that underlies the regularization principle considered here.

The elements of the projectional feature vector of an object $\omega \in \Omega$ are its scalar products with objects of the basic assembly $\mathbf{x}(\omega) = \left(x_1(\omega) \cdots x_n(\omega)\right)^T \in \mathsf{R}^n$, $x_k(\omega) = (\omega, \omega_k^0)$, $\omega_k^0 \in \Omega^0 \subset \Omega$. The basic objects, in their turn, are considered as elements of the same linear Hilbert space and, so can be characterized by their mutual proximity. If two basic objects ω_j^0 and ω_k^0 are close to each other, the respective projectional features do not carry essentially different information on objects of recognition $\omega \in \Omega$, and it is reasonable to assume that the coefficients a_j and a_k in the linear decision rule should also take close values. Therein lies the a priori information on the direction vector of the discriminant hyperplane that is to be taken into account in the process of training.

In fact, the coefficients a_j are functions of basic points in the Hilbert space $a_j = a(\omega_j^0)$, and the regularization principle we have accepted consists in the a priori assumption that this function should be smooth enough. It is just this interpretation that impelled us to give such a principle of regularization the name of smoothness principle.

It remains only to decide how the pair-wise proximity of basic objects should be quantitatively measured. For instance, scalar products $\mu_{jk} = (\omega_j, \omega_k)$ might be taken as such a measure. Then, the a priori information on the sought-for direction element can be easily introduced into the training criterion $J(\mathbf{a}, \delta_1, ..., \delta_N) = \mathbf{a}^T\mathbf{a} + C\sum_{j=1}^{N} \delta_j \to \min$ (2.4) as an additional quadratic penalty $J(\mathbf{a}, \delta_1, ..., \delta_N) = \mathbf{a}^T\mathbf{a} + C\sum_{j=1}^{N} \delta_j + J'(\mathbf{a})$:

$$J'(\mathbf{a}) = \alpha \sum_{j=1}^{n} \sum_{k=1}^{n} \mu_{jk}(a_j - a_k)^2 = \alpha\, \mathbf{a}^T\mathbf{B}\mathbf{a}, \quad \mathbf{B} = 2\begin{pmatrix} -\mu_{11} + \sum_{j=1}^{n} \mu_{1j} & \cdots & -\mu_{1n} \\ \vdots & \ddots & \vdots \\ -\mu_{n1} & \cdots & -\mu_{nn} + \sum_{j=1}^{n} \mu_{nj} \end{pmatrix},$$

where parameter $\alpha > 0$ presets the intensity of regularization. Inserting such a penalty reduces to substituting the quadratic form with identity matrix $\mathbf{a}^T\mathbf{a}$ by quadratic form $\mathbf{a}^T(\mathbf{I} + \alpha\mathbf{B})\mathbf{a}$. As a result, we come to a modified formulation of the training problem:

$$\begin{cases} \mathbf{a}^T(\mathbf{I}+\alpha\mathbf{B})\mathbf{a}+C\sum_{j=1}^{N}\delta_j \to \min, \\ g_j(\mathbf{a}^T\mathbf{x}_j+b)\geq 1-\delta_j, \quad \delta_j\geq 0, \quad j=1,...,N. \end{cases} \tag{3.12}$$

Because the size of the training sample N is, as a rule, less than the dimensionality of the projectional feature space n, the subsamples of the first and the second class will most likely be linearly separable. On the force of this circumstance, when solving the quadratic programming problem (2.4) without regularizing penalty, the optimal shifts of objects will equal zero $\delta_j=0$, $j=1,...,N$. After introducing the regularization penalty, the errorless hyperplane may turn out to be unfavorable from the viewpoint of a priori preferences expressed by matrix $\mathbf{B}$ with sufficiently large coefficient α. In this case, the optimal hyperplane will sacrifice, if required, the correct classification of some especially nuisance objects of the training sample, what will result in positive values of their shifts $\delta_j>0$ in (3.12).

By its mathematical structure, the training problem (3.12) coincides with (3.6), therefore, its dual formulation will look like (3.7):

$$\begin{aligned} & W(\lambda_1,...,\lambda_N)=\sum_{j=1}^{N}\lambda_j-\frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}(g_jg_k\mathbf{x}_j^T(\mathbf{I}+\alpha\mathbf{B})^{-1}\mathbf{x}_k)\lambda_j\lambda_k \to \max, \\ & \sum_{j=1}^{N}\lambda_jg_j=0, \quad 0\leq\lambda_j\leq\frac{1}{2}C,\ j=1,...,N. \end{aligned} \tag{3.13}$$

The direction vector of the optimal discriminant hyperplane in the space of projectional features is expressed by the folowing formula that is analogous to (3.6):

$$\mathbf{a}=(\mathbf{I}+\alpha\mathbf{B})^{-1}\sum_{j=1}^{N}\lambda_jg_j\mathbf{x}_j$$

Formulas for parameter b of the optimal hyperplane $\mathbf{a}^T\mathbf{x}+b=0$ and shifts of "nuisance" objects δ_j, $j=1,...,N$, remain the same as in the classical case (2.12) and (2.14).

5. Experimental results

5.1. Fold class recognition "one against one"

Experiments on fold class recognition were conducted with the collection of amino acid sequences of 396 protein domains grouped by Dr. Sun-Ho Kim into 51 fold classes (Table 1.1, Section 1). As the initial data set served the matrix 396×396 of pair-wise alignment scores obtained by alignment procedure Fasta3 and considered as Gram matrix of scalar products of respective protein domains (ω_j,ω_k), $j,k=1,...,396$, in an imaginary Hilbert space.

In the first series of experiments described in this Section, we solved the problem of pair-wise fold class recognition by the principle "one against one". There are $m=51$ classes in the collection and, so, $m(m-1)/2=1275$ class pairs, for each of which we found a decision rule of recognition and checked it by the leave-one-out procedure.

As the basic assembly $\Omega^0=\{\omega_1^0,...,\omega_n^0\}$, we took amino acid chains of 51 protein domains, $n=51$, one from each fold class. As representatives of classes, their "centers" were chosen, i.e. the protein domains that gave the maximum sum of pair-wise scores with other members of the respective class. Thus, each protein domain was represented by a 51-dimensional vector of its projectional features in accordance with (3.4).

For each of the 1275 class pairs, the training sample consisted of all protein domains making the respective two classes (Table 1.1). Thus, the size of the training sample varied from $N=7$ for pairs of small classes, such as (50) *Protein kinases, catalic core* and (51) *Beta-Lactamase*, to $N=59$ in two greatest classes (9) *Immunoglobin beta - sandwich* and (26) *TIM-barrel*.

We applied the technique of pattern recognition with preferred orientation of the discriminant hyperplane along the major inertia axis of the basic assembly in the Hilbert space. The quadratic programming problem (3.9) was solved for each of 1275 class pairs in its dual formulation (3.10), and the Lagrange multipliers $\lambda_1,...,\lambda_N$ found by the solver were used in the linear decision rule of recognition (3.11).

In each of 1275 experiments, the separability of the respective two fold classes was estimated by the leave-one-out procedure in accordance with the technique described in Section 3.4. Two rates were calculated for each class pair, namely, the percentage of correctly classified protein domains of the first and the second class. As the final estimate of the separability, the worst, i.e. the least, of these two percentages was taken.

As a result, the separability was found to be not worse than:

100% in 9% of all class pairs (completely separable class pairs),
90% in 14% of all class pairs,
80% in 32% of all class pairs,
70% in 53% of all class pairs.

The separability of 26 classes from more than one half of other classes is not worse than 70%. One class, namely, (50) *Protein kinases (PK), catalytic core*, showed its complete separability from all the classes.

On a data set of a lesser size, we checked how the pair-wise separability of fold classes will change if the number of basic proteins, i.e. the dimensionality of the projectional feature space, increases essentially. For this experiment, we took all the proteins of the collection as basic ones, so that the dimensionality of the projectional feature space became $n=396$.

The same truncated data set was used for studying how the separability of classes is affected by normalization of the alignment scores between amino acid chains, what is equivalent to projection of respective points of the imaginary Hilbert space onto the unit sphere. If (ω',ω'') is inner product of two original points of the Hilbert space associated with the respective two protein domains ω' and ω'', the inner product of their projections $\overline{\omega}'$ and $\overline{\omega}''$ onto the unit sphere will be $(\overline{\omega}',\overline{\omega}'')=\dfrac{(\omega',\omega'')}{\sqrt{(\omega',\omega')}\sqrt{(\omega'',\omega'')}}$. We used these values, instead of (ω',ω''), as similarity measure of protein domain pairs for fold class recognition.

For this series of experiments, we selected 7 fold classes different by their size and averaged separability from other classes. The chosen classes that contain in sum 85 protein domains are shown in Table 5.1.

The results are presented in Table 5.2. As we see, the extension of the basic assembly improved the separability of the class pairs that participated in the experiment. As to the normalization of the alignment score, it led to an improvement with the small basic assembly and practically did not change the separability with the enlarged one.

Table 5.1. Seven fold classes selected for the additional series of experiments.

	Fold class	Size	Averaged separability from other classes
1	Globin	12	73.4 %
3	Four-helical bundle	8	70.8 %
4	Ferritin	8	60.4 %
5	4-gelical cytokines	11	66.0 %
10	Common fold of difteria toxin / transcription factors / cytochrome	5	65.2 %
12	C2 domain	3	8.2 %
26	TIM-barrel	28	52.6 %

Table 5.2. Averaged pair-wise separability of seven fold classes in four additional experiments.

Size of the basic assembly n	Averaged separability	
	Original score matrix (ω', ω'')	**Normalized score matrix** $(\overline{\omega}', \overline{\omega}'')$
51	63.5 %	69.4 %
396	76.6 %	75.3 %

5.2. Effects of regularization

Experimental study of effects of regularization were conducted with the same truncated data set (Table 5.1) that had been used in experiments presented in the previous Section.

We examined how the smoothness principle of regularization, expressed by the modified quadratic programming problem (3.12) and its dual formulation (3.13), improves the separability of fold classes "one against one" within the selected part of the collection. The separability of each of 21 pairs of classes was estimated by the leave-one-out procedure several times with different values of the regularization coefficient α. Each time, the separability of a class pair was measured by the worst percentage of correct decisions in the first and the second class, whereupon the averaged separability over all 21 class pairs was calculated for the current value of α.

Such a series of experiments was carried out twice, with original and normalized alignment scores. The dependence of the separability on the regularization coefficient in both series is shown in Fig. 5.1, and the final estimates are summarized in Table 5.3.

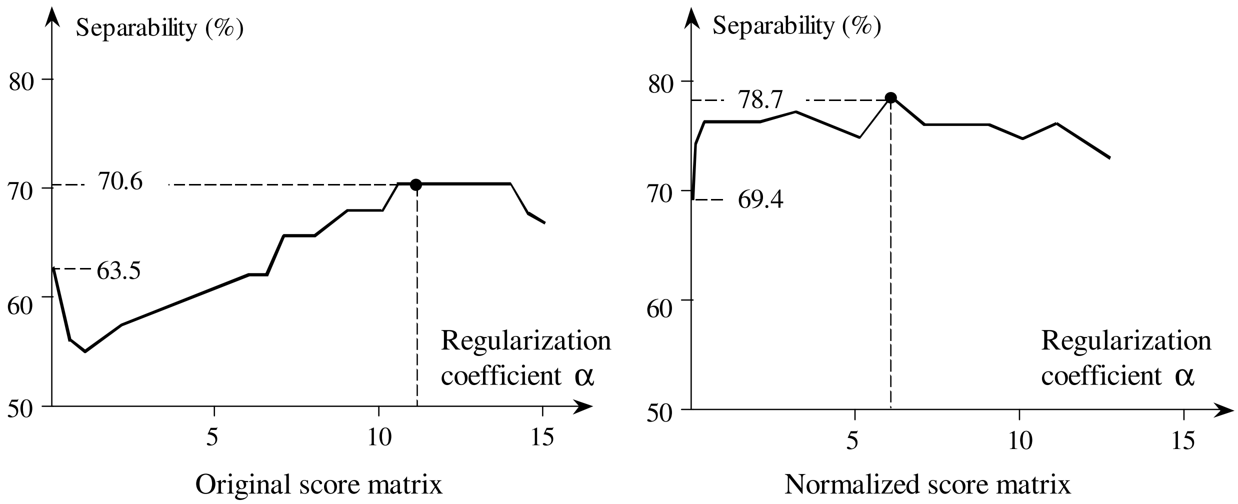

Figure 5.1. Dependence of the averaged pair-wise separability over 21 fold class pairs on the regularization coefficient.

Table 5.2. Averaged pair-wise separability of seven fold classes.

	Training without regularization, $\alpha=0$	Regularized training with the optimal value of α
Original score matrix (ω', ω'')	63.5 %	70.6 % ($\alpha=11$)
Normalized score matrix $(\overline{\omega}', \overline{\omega}'')$	69.4 %	78.7 % ($\alpha=6$)

In both series of experiments, a marked improvement of the separability is gained. The quality of training grows as the regularization coefficient increases, however, the improvement in not monotonic. A slight drop in separability with further increase in the coefficient after the maximum is attained arises from a too deep roughness of the decision rule adjustment.

6. Conclusions

This Report pursues two aims, a theoretical and a practical one, if viewing from the standpoint of the theory of pattern recognition. *The theoretical aim* of the Report is to study possible modifications of the classical support-vector approach to pattern recognition as applied to practical problems when the only way of perceiving an object of recognition is measuring its proximity to other objects and, so, there is no linear space of explicit features in which objects might be represented as vectors. *The practical aim* is to find ways of utilizing these modifications in solving pattern recognition problems that arise in predicting the spatial structure of proteins from their amino acid sequences immediately on the basis of the compactness hypothesis without highly problematic search for numerical features responsible for structure-forming properties of primary chains.

As to *the theoretical aim*, two methodological inventions are advanced in the Report. As the main idea of the featureless approach to pattern recognition, it is proposed to use an inner-product-like similarity measure between any two objects and to embed the variety of all feasible objects of recognition into a more rich set of elements of an imaginary linear space. Since it is not assumed that such a linear space has a finite dimensionality, it belongs to the most general class of inner product spaces called Hilbert spaces. It turned out that the classical support vector method, originally designed for pattern recognition in spaces of a finite number of preset features, is practically completely adequate to featureless problems interpreted in such a way.

Another methodological idea which led to a slight modification of the support vector method is that of regularization of the training process with the purpose of overcoming the curse of a small training sample which is an inherent feature of the featureless pattern recognition. The training objects cannot be represented in any other way than by values of their proximity to some other objects forming the so-called basic assembly that, in contrast to the training sample, needn't be classified. Since it is, as a rule, preferable to include the training objects in the basic assembly, the size of the training sample turns out practically always to be not greater than the number of numerical characteristics of each object. As a result, the freedom of varying the decision rule of recognition when adjusting it to the training sample is too poorly limited by the insufficient amount of information carried by a too small set of classified objects, and the inferred decision rule reflects rather the specificity of the training sample than properties of the general population as a whole. It is proposed to replace this lack by a priori information contained in the matrix of mutual similarity of basic objects.

For achieving *the practical aim* of the work on the basis of theoretical suggestions, it was required, first of all, to choose an appropriate similarity measure of amino acid sequences that represent the primary chemical structure of proteins. It is proposed to consider the set of all feasible amino acid sequences as a subset of a Hilbert space, at least, an imaginary one, in which the pair-wise alignment score of any two sequences plays the role of their inner product. The respective modification of the support vector method is applied to the problem of building a classifier of 51 protein fold families within a collection of 351 mutually remote proteins. Some preliminary results of this study are presented in the Report.

The idea of featureless pattern recognition on the basis of treating entities to be allocated over a finite number of classes as elements of a Hilbert space invites further research because it raises more questions than we have answered here.

The version of the support vector method taken in this work as a basis for featureless pattern recognition presupposes only a rigid decision on the hidden class-membership of an object in the sense that each object is always associated with one and only one class. At the same time, it is typical for many practical problems that there are boundary object which cannot be classified strictly. In this connection, it appears to be reasonable to provide evaluating posterior probabilities of the membership of an object in all classes at once, so that it would be possible to rule out some of them and reduce the ambiguity to the remaining ones. Such an option is especially important for the problem of protein fold class recognition, because the spatial structures of different proteins are never completely equivalent, and their initial classification is rather a conventional one.

The idea of treating the pair-wise alignment score of any two amino acid sequences as their inner product in a Hilbert space, even if an imaginary one, is sound only in case the matrix of such scores is positive definite for any assembly of proteins. It turned out to be so with 396 proteins we used as material for experiments, but there is no guarantee that this property will hold also for other assemblies. Therefore, it appears to be important to find such similarity measures of amino acid sequences that would be, simultaneously, natural from the biochemical standpoint and consistent with the idea of mathematical treating them as inner products in a linear space. In particular, a pilot analysis of the assembly of proteins based on multiple alignment of some groups of them appears to be a promising way of designing such similarity measures.

Within the bounds of the classical approach to pattern recognition, when an object is represented independently of other objects by vector of its individual features, each particular feature is meant to carry information of a specific kind, so that the feature vector as a whole is a many-sided characteristic of the object. In the featureless approach, though, at least, in its version considered in this Report, only one similarity measure is stipulated, that represents only one point of view at the relationship between objects of recognition. Such a one-sided characteristic may turn out to be insufficient for a reliable classification of objects. However, the presence of several similarity measures will cause the problem of overcoming their mutual contradiction. Development of a theoretical framework for coordinating several featureless classifiers based on different similarity measures, in particular, for protein fold class recognition, is a glowing topic that invites further investigations.

Refertences

[1] Fetrow J.S., Bryant S.H. New programs for protein tertiary structure prediction. *Biotechnology*, Vol. 11, April 1993, pp. 479-484.

[2] Dubchak I., Muchnik I., Mayor C., Dralyuk I., Kim S.-H. Recognition of a protein fold in the context of the SCOP classification. *Proteins: Structure, Function, and Genetics*, 1999, 35, 401-407.

[3] Duin, R.P.W, De Ridder, D., Tax, D.M.J. Featureless classification. *Proceedings of the Workshop on Statistical Pattern Recognition*, Prague, June 1997.

[4] Duin, R.P.W, De Ridder, D., Tax, D.M.J. Experiments with a featureless approach to pattern recognition.

[5] Duin, R.P.W, Pekalska, E., De Ridder, D. Relational discriminant analysis. *Pattern Recognition Letters*, Vol. 20, 1999, No. 11-13, pp. 1175-1181.

[6] Cortes C., Vapnik V. Support-vector networks. *Machine Learning*, Vol. 20, No.3, 1995.

[7] Vapnik V. *Statistical Learning Theory*. John-Wiley & Sons, Inc. 1998.

[8] Durbin R., Eddy S., Krogh A., Mitchison G. *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1988.

[9] Dayhoff M.O., Schwartz R.M., Orcutt B.C. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure* (Dayhoff M.O., ed.), National Biomedical Research Foundation, Washington D.C., 1978, Vol. 5, Suppl. 3, pp. 345-352.

[10] Henikoff S., Henikoff J.G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the USA*, 1992, Vol. 89, pp. 10915-10919.

[11] Pearson W. R. and Lipman D. J. Improved tools for biological sequence analysis. *PNAS*, 1988, 85, 2444- 2448.

[12] Pearson W. R. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods in Enzymology*, 1990, 183, 63-98.

[13] Vapnik V.N. *Estimation of Dependences Based on Empirical Data*. Springer, New York, 1982.