Featureless Pattern Recognition in an Imaginary Hilbert Space

Vadim Mottl, Oleg Seredin, Sergey Dvoenko
Tula State University
Lenin Ave. 92, 300600 Tula, Russia
mottl@atm.tsu.tula.ru

Casimir Kulikowski, Ilya Muchnik
Rutgers University
P.O. Box 8018, Piscataway, NJ 08855, USA
kulikows@cs.rutgers.edu

Abstract

The featureless methodology is applied to the class of pattern recognition problems in which the adopted pairwise similarity measure possesses the most fundamental property of inner product to form a nonnegative definite matrix for any finite assembly of objects. It is proposed to treat the set of all feasible objects of recognition as a subset of isolated points in an imaginary Hilbert space. This idea is applied to the problem of determining the membership of a protein given by its amino acid sequence (primary structure) in one of preset fold classes (spatial structure) on the basis of measuring the likelihood that two proteins have the same evolutionary origin by way of calculating the so-called alignment score between two amino acid sequences, as it is commonly adopted in computational biology.

1. Introduction

There exists a wide class of applications in which it is easy to evaluate some numerical characteristics of pairwise relationship between any two objects, but it is hard to indicate a set of rational individual attributes of objects that could form the axes of a feature space. As an alternative to the feature-based methodology of pattern recognition, R. Duin and his colleagues [1] proposed a featureless approach, in which objects are assumed to be represented by appropriate measures of their pairwise similarity or dissimilarity.

It is just this idea we use here as the basis of a pattern recognition technique for the class of applications in which the similarity measure forms a nonnegative definite matrix for any finite set of objects and, so, possesses the most fundamental property of inner product.

In this work, we lean upon the heuristic hypothesis that the set of all feasible objects may be considered as a subset of isolated points in an imaginary linear space (Hilbert space) in which the role of inner product is played by the similarity between any two objects. In case of two classes, such an assumption allows for treating the sought-for decision rule of pattern recognition as a discriminant hyperplane immediately in the Hilbert space of objects.

As a glowing example of such an application, we consider the problem of protein fold class recognition, i.e. allocating a protein, given by the primary chemical structure of its polymerous molecule as a sequence of amino acid residues from the alphabet of 20 amino acids existing in nature, over a finite set of typical spatial structures, each associated with a specific manner in which the primary amino acid chains fold in space under a combination of numerous physical forces [3].

It is common practice in computational biology to measure the proximity between two amino acid chains $\omega' = (a_1, \ldots, a_{K'})$ and $\omega'' = (a_1, \ldots, a_{K''})$ as the logarithmic likelihood ratio $\mu(\omega', \omega'')$ of the main hypothesis that both of them originate from the same unknown protein as result of independent successions of local evolutionary mutations versus the null hypothesis that the chains are completely occasional combinations over the alphabet of 20 amino acids. The generally accepted way of measuring such a likelihood ratio is the so-called alignment of two amino acid sequences, namely, finding the minimum number of deletions, insertions and substitutions of single amino acids that turn the sequences into each other [4] (Fig. 1).

By its nature, the logarithmic likelihood ratio may take as positive as well negative values and, having been calculated for an amino acid sequence with itself, gives different positive values for different proteins. It appears natural to consider the alignment score $\mu(\omega', \omega'')$ as inner product of two proteins in an imaginary linear space. We checked this hypothesis for an assembly of 396 proteins by way of calculating all the eigenvalues of the alignment score matrix. The fact that all the eigenvalues turned out to be positive is a

```
ω′ : TNPGNASSTTTTKPTTTS-----RGLKTINETDPCIKNDSC
ω″: GS----ATSTPATSTTAGTKLPCVRNKTDSNLQSCNDTIIE
i = 12      34567  ···
```

Figure 1. Fragment of an aligned pair of amino acid chains.

This work is supported by the Russian Foundation of Basic Research and Rutgers University Strategic Opportunity Allocation Project on Bioinformatics.

reason in favor of the theoretical applicability of the principle of featureless pattern recognition in a Hilbert space to the problem of protein fold classification.

2. Hilbert space of classified objects and optimal discriminant hyperplane

Let the set Ω of all feasible objects under consideration $\omega \in \Omega$ is partitioned into two classes $\Omega_1 = \{\omega \in \Omega\colon\ g(\omega) = 1\}$ and $\Omega_2 = \{\omega \in \Omega\colon\ g(\omega) = -1\}$ by an unknown indicator function $g(\omega) = \pm 1$. The main idea of the approach to pattern recognition we consider here 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 $c\omega \in \Omega$, $c \in \mathsf{R}$, 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 $(\omega', \omega'') = (\omega'', \omega') \in \mathsf{R}$ 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 $(\omega, \omega) \geq 0$, $(\omega, \omega) = 0$ if and only if $\omega = \phi$, and gives the norm $\|\omega\| = (\omega, \omega)^{1/2} \geq 0$.

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 for speaking 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 $d(\omega|\vartheta, b) = (\vartheta, \omega) + b$, where $b \in \mathsf{R}$ is a constant, may be used as decision rule $\hat{g}(\omega)\colon\ \Omega \to \{1, -1\}$ of judging on the hidden class-membership of an arbitrary object $\omega \in \Omega$, might it really exist or not:

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

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

Let the observer be submitted a classified training set of objects $\Omega^* = \{\omega_1, \ldots, \omega_N\} \subset \Omega$, $g_1 = g(\omega_1), \ldots, g_N = g(\omega_N)$. Parameters of the discriminant hyperplane $\vartheta \in \Omega$ and $b \in \mathsf{R}$ (1) should be chosen so that the training objects would be classified correctly with a positive margin $\xi > 0$:

$$d(\omega_j|\vartheta, b) = (\vartheta, \omega_j) + b \begin{cases} \geq \xi \text{ when } g(\omega_j) = 1, \\ \leq -\xi \text{ when } g(\omega_j) = -1. \end{cases}$$

The same reasoning as in [2] leads to the conclusion that the maximum margin is provided by choosing the direction element $\vartheta \in \Omega$ and threshold $b \in \mathsf{R}$ from the condition $\|\vartheta\|^2 \to \min$, $g_j[(\vartheta, \omega_j) + b] \geq 1$, $j = 1, \ldots, N$, or, in case the training set is linearly inseparable, $\|\vartheta\|^2 + C\sum_{j=1}^{N} \delta_j \to \min$, $g_j[(\vartheta, \omega_j) + b] \geq 1 - \delta_j$, $\delta_j \geq 0$.

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

$$\begin{cases} \|\vartheta\|^2 + C\sum_{j=1}^{N} \delta_j \to \min, \\ g_j[(\vartheta, \omega_j) + b] \geq 1 - \delta_j,\ \delta_j \geq 0,\ j = 1, \ldots, N. \end{cases} \qquad (2)$$

3. Basic assembly of objects and choice of the norm of the direction element

As we shall see below, the structure of the decision rule inferred from the training set essentially depends on the kind of the norm of the direction element whose squared value $\|\vartheta\|^2$ is to be minimized in accordance with the training criterion (2). The formula

$$\|\vartheta\| = (\vartheta, \vartheta)^{1/2} \qquad (3)$$

does not exhaust the variety of ways in which the norm may be defined. In this Section, we introduce the notion of a basic assembly of objects which does not coincide with the training set in the general case and serves as basis for, at least, one more kind of norm meant to carry some additional information on the distribution of objects in the Hilbert space, different from that carried by the trainer's information on the class memberships of objects.

In many applications, acquisition of information on class membership of objects presents a considerable difficulty, therefore, forming a sufficiently large training set is quite problematic. At the same time, it is much easier to collect an unclassified assembly of objects for which pairwise inner products can be calculated. As examples of such applications may serve, in particular, the problem of protein fold class recognition as well as practically all problems of medical diagnosis.

Let $\Omega^0 = \{\omega_1^0, \ldots, \omega_n^0\} \subset \Omega$ be an unclassified assembly of objects called the basic assembly. In particular, the training set may be part of the basic assembly $\Omega^* \subset \Omega^0$ or coincide with it $\Omega^* = \Omega^0$, but

these are different sets in the general case. If we denote by $\mathbf{M} = \left((\omega_i^0, \omega_j^0),\, i,j = 1,\ldots,n\right)$ the matrix of inner products within the basic assembly and by $\mathbf{x}(\vartheta) = \left((\vartheta,\omega_1^0)\cdots(\vartheta,\omega_n^0)\right)^T \in \mathsf{R}^n$ the vector of inner products of element $\vartheta \in \Omega$ with basic objects $\omega_i^0 \in \Omega^*$, then function

$$\|\vartheta\| = \left(\left(\mathbf{M}^{-1}\mathbf{x}(\vartheta)\right)^T \mathbf{M}^{-1}\mathbf{x}(\vartheta)\right)^{1/2} \qquad (4)$$

will possess all properties of norm in the Hilbert space Ω. Here vector $\mathbf{M}^{-1}\mathbf{x}(\vartheta)$ is vector of coefficients $(a_1(\vartheta)\cdots a_n(\vartheta))^T$ that form the projection $\sum_{i=1}^n a_i(\vartheta)\,\omega_i^0$ of element $\vartheta \in \Omega$ onto the subspace spanned over the basic assembly $\Omega^0 = \{\omega_1^0,\ldots,\omega_n^0\} \subset \Omega$.

Training by criterion (2) with the "native" version of norm (3) is aimed at finding the shortest admissible direction element $\vartheta \in \Omega$, and, so, all orientations of the discriminant hyperplane in the Hilbert space are equally preferable. On the contrary, it can be shown [5] that training with norm (4) is equivalent to a pronounced preference in the 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. 2).

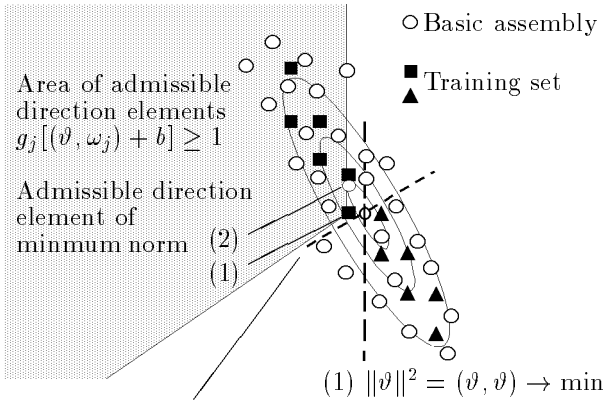

Figure 2. Minimum norm of the direction element of the discriminant hyperplane in the Hilbert space as criterion of training for two versions of norm.

4. Structure of decision rules

The structure of decision rules for both kinds of norm (3) and (4) immediately follows from the dual formulation of the quardratic programming problem (2). The only specificity is that argument $\vartheta \in \Omega$ is element of the Hilbert space but not a vector, as usually, therefore, for minimization of the respective Lagragian we have to use the notion of Frechet differential [6] instead of that of gradient. In case of norm (3), the Frechet differential of the real-valued function $f(\vartheta) = (\vartheta,\vartheta)\colon\ \Omega \to \mathsf{R}$ is the element of the Hilbert space $2\vartheta \in \Omega$, and for norm (4) the Frechet differential of the function $f(\vartheta) = \left(\mathbf{M}^{-1}\mathbf{x}(\vartheta)\right)^T \mathbf{M}^{-1}\mathbf{x}(\vartheta)$ will be linear combination of basic objects $2\sum_{i=1}^n a_i(\vartheta)\,\omega_i^0$ with elements of vector $\mathbf{M}^{-1}\mathbf{x}(\vartheta)$ as coefficients.

Training by criterion (2) with norm (3), what means absense of preferred orientation of the discriminant hyperplane in the Hilbert space, is equivalent to solving the dual quadratic programming problem

$$\begin{cases} 2\sum_{j=1}^N \lambda_j - \sum_{j=1}^N \sum_{k=1}^N g_j g_k (\omega_j,\omega_k)\lambda_j\lambda_k \to \max, \\ \sum_{j=1}^N g_j\lambda_j = 0,\ 0 \leq \lambda_j \leq (1/2)C,\ j = 1,\ldots,N. \end{cases}$$

Those of resulting Lagrange multipliers $\lambda_j \geq 0$ at objects of the training set $\Omega^* = \{\omega_j,\, j = 1,\ldots,N\}$ which differ from zero $\lambda_j > 0$ point at support objects forming the direction element of the optimal dicriminant hyperplane as their linear combination $\vartheta = \sum_{j:\,\lambda_j>0} g_j\lambda_j\omega_j$ and, so, constitute the discriminant function applicable to any new object $\omega \in \Omega$ (1)

$$d(\omega|\vartheta,b) = \sum\nolimits_{j:\,\lambda_j>0} g_j\lambda_j(\omega,\omega_j) + b. \qquad (5)$$

Threshold b is determined here by the symmetricity condition

$$b = (1/2)\left[\min_{j:\,g_j=1}(\vartheta,\omega_j) - \max_{j:\,g_j=-1}(\vartheta,\omega_j)\right] \qquad (6)$$

In this case, the basic assembly does not participate in training and, respectively, in the discriminant function (5) which leans upon inner products of the new object with only support objects of the training set.

But in case of norm (4), which provides a pronounced preference in favor of direction elements oriented along the major inertia axis of the basic assembly, we come to another dual problem

$$\begin{cases} 2\sum_{j=1}^N \lambda_j - \sum_{j=1}^N \sum_{k=1}^N g_j g_k \mathbf{x}^T(\omega_j)\mathbf{x}(\omega_k)\lambda_j\lambda_k \to \max, \\ \sum_{j=1}^N g_j\lambda_j = 0,\ 0 \leq \lambda_j \leq (1/2)C,\ j = 1,\ldots,N, \end{cases}$$

where $\mathbf{x}(\omega) = \left((\omega,\omega_1^0)\cdots(\omega,\omega_1^0)\right)^T$ is vector formed by inner products of an object $\omega \in \Omega$ with elements of the basic assembly. The resulting direction element of the optimal discriminant hyperplane is linear combination of basic objects $\vartheta = \sum_{i=1}^n c_i\,\omega_i^0$ with coefficients determined by positive Lagrange multipliers and inner products of support objects of the training set with elements of the basic assembly $c_i = \sum_{j:\,\lambda_j>0} g_j\lambda_j\,(\omega_j,\omega_i^0)$. This solution results in the discriminant function

$$d(\omega|\vartheta,b) = \sum\nolimits_{j:\,\lambda_j>0} g_j\lambda_j\mathbf{x}^T(\omega)\mathbf{x}(\omega_j) + b$$

where threshold b can be found by the formula analogous to (6) with $\mathbf{x}^T(\omega)\mathbf{x}(\omega_j)$ instead of (ω,ω_j).

Just as previously, only support obects of the training set occur in this discriminant function, but, in contrast to (5), each of them is represented by its inner products with elements of the basic assembly, and, so, all the basic objects participate in training as well as in decision making.

5. Experimental results

Experiments on fold class recognition were conducted with the collection of amino acid sequences of 396 protein domains grouped into 51 fold classes [5]. As the initial data set served the matrix 396×396 of pair-wise alignment scores considered as matrix of inner products of respective protein domains (ω_i, ω_j) in the imaginary Hilbert space.

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 linear decision rule of recognition.

For each of the 1275 class pairs, the training sample consisted of all protein domains making the respective two classes. The size of the training sample varied from $N = 1$ for pairs of small classes to $N = 59$ in two greatest classes.

We applied the technique of pattern recognition with preferred orientation of the discriminant hyperplane along the major inertia axis of the basic assembly $\Omega = \{\omega_1^0, \ldots, \omega_n^0\}$, which was formed by 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 alignment scores with other members of the respective class. The quadratic programming problem (2) with norm (4) was solved for each of 1275 class pairs in its dual formulation.

A way of empirical estimating the quality of the decision rule immediately from the training set offers the well-known leave-one-out procedure [2], which was used in each of 1275 experiments for evaluating the separability of the respective two fold classes. 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%.

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 protein domains, i.e. the dimensionality of the projectional feature space, increases essentially. For this experiment, we took all the protein domains of the collection as basic ones $n = 396$.

The experiment was conducted with 7 selected fold classes different by their size and averaged separability from other classes. As a result, the extension of the basic assembly from $n = 51$ to $n = 396$ improved the averaged separability of the class pairs that participated in the experiment from 63.5% to 76.6%.

6. Conclusions

Within the bounds of the featureless approach to pattern recognition, the main idea of this work is treating the pair-wise similarity measure of objects of recognition as inner product in an imaginary Hilbert space, into which really existing objects may be mentally embedded as a subset of isolated points. In the practical problem of protein fold class recognition, to embed the discrete set of known proteins into a continuous Hilbert space, we propose to consider as inner product the pair-wise alignment score of amino acid chains, which is commonly adopted in bioinformatics as their biochemically justified similarity measure.

References

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

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

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

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

[5] V. Mottl, S. Dvoenko, O. Seredin, C. Kulikowski, I. Muchnik. *Alignment scores in a regularized support vector classification method for fold recognition of remote protein families*. DIMACS Technical Report 2001-01. Rutgers University, USA.

[6] A.N. Kolmogorov, S.V. Fomin. *Introductory Real Analysis*. Prentice-Hall, Englewood Cliffs, 1970.