DESIGN OF OPTIMAL FEATURE EXTRACTORS BY MATHEMATICAL PROGRAMMING TECHNIQUES

Rui J. P. de Figueiredo

Department of Electrical Engineering and
Department of Mathematical Sciences
Rice University, Houston, Texas 77001

June, 1976

Technical Report EE-7608

(Preprint of an article from the volume ARTIFICIAL INTELLIGENCE AND PATTERN RECOGNITION, edited by C. H. Chen, to be published by Academic Press, New York)

Abstract: In an automatic pattern recognition system, the processor that selects and measures features of the data, on the basis of which classification is made, is called a "feature selector" or "feature extractor". This paper presents a mathematical programming approach for the design of a feature extractor.

DESIGN OF OPTIMAL FEATURE EXTRACTORS BY MATHEMATICAL PROGRAMMING TECHNIQUES*

by

Rui J. P. de Figueiredo
Rice University, Houston, Texas 77001

1. Introduction

The "feature extraction" operation plays a very significant role in the functioning of a pattern recognition system. For this reason, considerable attention has been devoted to the feature extraction problem in the pattern recognition literature (see, for example, [1] and [2] and the references therein). However, most of the existing techniques for feature extractor design rely on the maximization of some average distance measure amongst pattern classes in the "feature (transformed) space."

More recently, it was proposed by the author [3] that the performance of the entire pattern recognition system ought to be taken into account when selecting the optimal feature extraction transformation. According to this point of view, the structure of the desired feature extractor would be tuned to the classification strategy adopted in a given problem. In particular, if the classification strategy were Bayes, the optimal feature extraction transformation would be the one that would minimize, over a suitable class of admissible transformations the Bayes risk (probability of misclassification) in the feature space. The problem thus posed becomes essentially a mathematical programming problem.

In what follows, we formulate precisely the above problem in a framework sufficiently general to permit the use of statistical and/or linguistic considerations. Then, for the case in which the classification strategy is Bayes, we discuss the important design considerations and cite some specific results.

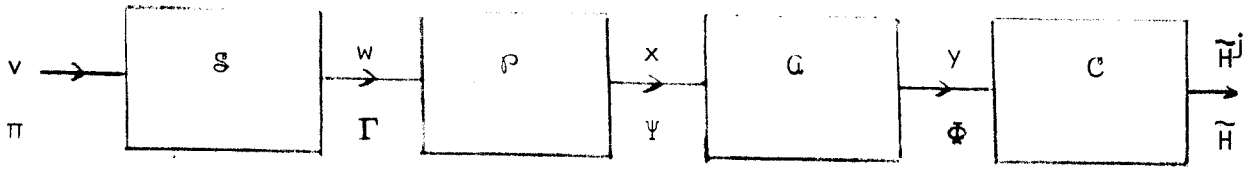

Fig. 1. Block Diagram of a Pattern Recognition System.

*Supported by the AFOSR Grant 75-2777.

2. Mathematical Formulation

The general configuration of a pattern recognition system is depicted in Fig. 1. It consists of four blocks representing respectively a system $\mathcal{S}$ of sensors, a preprocessor $\mathcal{P}$, a feature extractor $\mathcal{Q}$, and a classifier $\mathcal{C}$. In an actual hardware implementation, some of these blocks may overlap or be combined into a single unit.

The sensing system $\mathcal{S}$ converts the excitation from a "pattern" v being perceived by $\mathcal{S}$ into some form of raw data w. This data is usually contaminated by "distortion" and "noise" introduced by the sensing devices and the environment. The preprocessor $\mathcal{P}$ simply removes, to the extent possible, this distortion and noise from w by means of some "cleaning" (filtering, enhancement, restoration, . . .) operation. Thus the output x from $\mathcal{P}$ is what may be called "clean data" or "preprocessed signal". The feature extractor $\mathcal{Q}$ then measures the values of a set of variables pertaining to x called "features". Hopefully these variables contain most of the information needed for recognition purposes. Finally, if y denotes the output of $\mathcal{Q}$, the classifier $\mathcal{C}$ assigns y to some pattern class H^j, and thus the recognition operation is completed.

It may be remarked in passing that, in some pattern recognition literature (see, for example, [1], p.7), the preprocessor $\mathcal{P}$ and the feature extractor $\mathcal{Q}$ are considered to be one and the same entity. Here, we make the distinction between $\mathcal{P}$ and $\mathcal{Q}$, in the sense that $\mathcal{P}$ performs a "signal processing" operation on the raw data with the objective of essentially optimizing the signal-to-noise ratio without necessarily taking the ultimate use of the signal into account (many of the so-called "picture processing" papers deal with this problem); on the other hand, the objective of $\mathcal{Q}$ is to provide measurements on the (filtered) signal solely for the purpose of recognition.

Let us now attempt to formulate precisely the problem under consideration.

Let π denote the set of all patterns v to be sensed by $\mathcal{S}$. Assume that π may be partitioned into M pattern classes $H^1, \ldots, H^M$, and let the set $\{H^1, \ldots, H^M\}$ be denoted by H. Each member of H is to be classified as belonging to some H^j.

Let Γ, Ψ, Φ and $\tilde{H}$ denote the sets of outputs from respectively $\mathcal{S}$, $\mathcal{P}$, $\mathcal{Q}$, and $\mathcal{C}$.

As done earlier in this section, members of Γ, Ψ, and Φ will be denoted respectively by w, x, and y. Any given output from $\mathcal{C}$ is a statement saying that a pattern v being perceived by $\mathcal{S}$ belongs to some class H^j. We will

denote such a statement simply by $\tilde{H}^j$. The set of all $\tilde{H}^j$, j=1, ..., M, constitutes the set $\tilde{H}$ mentioned above.

In order to describe the operation of $\mathcal{S}$, $\mathcal{P}$, $\mathcal{Q}$, and $\mathcal{C}$ we introduce respectively the maps

$$S: \pi \rightarrow \Gamma, \tag{1}$$

$$P: \Gamma \rightarrow \Psi \tag{2}$$

$$A: \Psi \rightarrow \Phi \tag{3}$$

$$C: \Phi \rightarrow \tilde{H} \tag{4}$$

The operation of the pattern recognition system may now be expressed by

$$\tilde{H}^j = C(A(P(S(v)))) \tag{5a}$$

$$\underset{=}{\Delta} K(v). \tag{5b}$$

It ought to be noted that in (5a) the superscript j on $\tilde{H}^j$ is generic, that is, depending on v, $\tilde{H}^j$ could be any one of the statements $\tilde{H}^1$, ..., $\tilde{H}^M$.

In terms of the above notation then, we will call a given set $\{\pi, H\}$ a "pattern structure". Also, given any "pattern recognition system", we will refer to it by the set of maps {S, P, A, C}, or simply by the corresponding composition map K, which describes it.

In our formulation, we will assume, as in most practical cases, that S is fixed because of hardware constraints, and so is P since the design of P is conditioned by the structure of S, as we pointed out earlier.

The same is not true with regard to the two remaining maps A and C Clearly, A is not fixed since our very objective is to select a $A^* \in \chi$ which is optimal with respect to all maps A belonging to an admissible class χ.

Usually the classification strategy is selected beforehand. However, since the domain of C depends on A, the structure of the classifier itself will vary with A. To signal this fact we will replace the symbol C by C_A.

One final consideration is the inclusion of training sets in our problem formulation. Such sets constitute the main source of information on a given pattern structure, on the basis of which a recognition system

for that structure can be designed. We will denote by $\mathcal{A}^j$ the available training set pertaining to the pattern class H^j, j=1, ..., M. Elements of $\mathcal{A}^j$ will be denoted by u^j, and when necessary to distinguish these elements among themselves we will number them with subscripts, thus u^j_1, u^j_2, ..., $u^j_{N_j}$, where N_j is the total number of training samples in $\mathcal{A}^j$.

In a conventional way, we will assume that the elements in $\mathcal{A}^j$, j=1, ..., M, are independent (with respect to some underlying probability measure), and that each $\mathcal{A}^j$ is partitioned into two subsets: a subset $\mathcal{A}^j_C$ is used in the construction of the structure of the classifier, and a subset $\mathcal{A}^j_A$ used in the design of the map A. Let

$$\Lambda = \mathcal{A}^1 \cup \mathcal{A}^2 \cup \ldots \cup \mathcal{A}^M \tag{6}$$

and define the partition of Λ into Λ_C and Λ_A where

$$\Lambda_C = \mathcal{A}^1_C \cup \mathcal{A}^2_C \cup \ldots \cup \mathcal{A}^M_C, \tag{7}$$

$$\Lambda_A = \mathcal{A}^1_A \cup \mathcal{A}^2_A \cup \ldots, \cup \mathcal{A}^M_A. \tag{8}$$

To indicate that the structure of the classifier is based on Λ_C we will replace C_A by $C_{A\Lambda_C}$.

Let a function ξ from $\tilde{H}\times\tilde{H}$ to the reals be defined by

$$\xi(\tilde{H}^j, \tilde{H}^k) = 1-\delta_{jk}, \quad j,k = 1, \ldots, M, \tag{9}$$

where δ_{jk} = Kronecker delta.

Then the total cost (risk or probability) of misclassification when all the training samples in Λ are presented to a pattern recognition system $\{S, P, A, C_{A\Lambda_C}\}$ is

$$Q(A) = \sum_{j=1}^{M} \sum_{u^j_i \in \mathcal{A}^j_A} \alpha_{jk(i)}\, \xi(\tilde{H}^j, C_{A\Lambda_C}(A(P(S(u^j_i))))), \tag{10}$$

where the nonnegative constants $\alpha_{jk(i)}$ are suitable "cost" weights. The

subscript $k(i)$ on $\alpha_{jk(i)}$ is the superscript on the output $\tilde{H}^k$ of the pattern recognition system with u_i^j as input, i.e. $\tilde{H}^k = C_{A\Lambda_C}(A(P(S(u_i^j))))$. (11)

The optimal feature extractor design problem now reduces to the following:

Problem. Let all the symbols be defined as above. For the purpose of finding a system $\{S, P, A, C\}$ to recognize a pattern structure (π, H), suppose that you are given S, P, Λ, a classification strategy C, a set of weights α_{jk}, $j,k=1, \ldots, M, j\neq k$, and a class χ of admissible maps A. Find a $A^* \in \chi$ which minimizes $Q(A)$ defined by (10) over all $A \in \chi$.

At this point, the following observations about our formulation are in order:

(i) In selecting the feature extraction map, we are considering the performance of the entire pattern recognition system.

(ii) Except for the mild measurability condition assumed on the training sets needed to justify our criterion functional (10), no restrictions have been imposed on the ranges and domains of the four maps constituting our pattern recognition system. So our formulation may be used when the pattern recognition system under consideration is described either by operators in linear vector spaces as in statistical pattern recognition [1], or by the formal language approach [4].

(iii) We have developed our general formulation to the point where, with the addition of details pertaining to a specific application, the feature extractor design problem becomes a very well-defined nonlinear programming problem which can be readily solved by any one of the standard algorithms available in the literature [5].

3. Specific Considerations and Results

3.1. The structure of π, Γ, Ψ and Φ

A given pattern recognition application would determine the structures of the sets π, Γ, Ψ, and Φ. In many applications, Γ, Ψ, and Φ are linear spaces, the dimensions of Γ and Ψ being high and that of Φ low. For this reason, the selection of A is sometimes called the "dimensionality reduction" problem.

From now on, we will asuume that Φ and Ψ are linear spaces.

3.2. The characterization of χ

One important consideration in the design of the optimal A* is the characterization of the class χ of admissible transformations A. A very general class that we propose is that of continuous functions from Ψ to Φ. We denote such a class by χ_C. Provided the domain of each $A \in \chi_C$ is assumed compact, any member of χ_C can be represented to any desired degree of accuracy by a polynomic operator [6] [7]. In particular, if Ψ and Φ are finite-dimensional with dimensions n and m respectively, a member of χ may be approximated by m multivariate polynomials p_i, $i = 1, \ldots, m$, expressing the components of $y = (y_1, \ldots, y_m) \in \Phi$ in terms of the components of $x = (x_1, \ldots, x_n) \in \Psi$, thus

$$y_i = A_i(x) = p_i(x_1, \ldots, x_n), \quad i = 1, \ldots, m. \tag{12}$$

A subclass of such χ_C is the class χ_{CL} of linear transformations consisting of all mxn matrices.

The various discrete transforms (e.g. Fourier, Walsh, ...) that have been used in digital data processing are linear transformations and may be used as intermediate vehicles in composing a subclass of χ_{CL}.

For example, suppose we pick for this purpose the discrete Fourier transform which we denote by D. Then D is a linear transformation from Ψ to $\tilde{\Psi}$, the span of an appropriate discrete Fourier transform basis elements. The dimension of $\tilde{\Psi}$ is the same as that of Ψ and hence equal to n. It is well-known that in some pattern recognition problems, events pertaining to different pattern classes are better separated in the transformed domain $\tilde{\Psi}$. For the purpose of dimensionality reduction then, one would follow the transform operation by some other appropriate linear operation L. For example, L could select m of the spectral components of the

transformed data vector, the choice of these components being such that those m components containing the maximum amount of information needed for recognition are selected. In this case, we would construct a subset of χ_{CL} consisting of all maps A = LD, different A's corresponding to different L's (different choices of m spectral components).

3.3. The Bayes risk as Criterion Functional

If the pattern structure to be recognized is modeled probabilistically, then the criterion functional (10), with appropriate interpretation, may be viewed as the Bayes risk (probability of misclassification).

Thus in the probabilistic case, for j=1, ..., M, let $f_Y(y/H^j, A)$ denote the probability density function for the random vector $Y = (Y_1, \ldots, Y_m)$ conditioned on the pattern class H^j and on the transformation A. Here we consider the elements $y = (y_1, \ldots, y_m) \in \Phi$ as realization of the random vector Y conditioned on one of the pattern classes. Also, let P_j denote the prior probability for H^j and β_{ij} the cost of classifying a y arising from H^j as pertaining to H^i. Then the Bayes risk is:

$$Q(A) = \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j\neq i}}^{M} P_j \, \beta_{ij} \int_{\Omega_i(A)} f_Y(y, H^j, A)\, dy, \tag{13}$$

where $\Omega_i(A)$ is the Bayesian decision region in Φ for H^i.

If

$$\beta_{ij} = 1 - \delta_{ij}, \quad i, j=1, \ldots, M, \tag{14}$$

(13) reduces to the probability of misclassification

$$Q(A) = \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j\neq i}}^{M} P_j \int_{\Omega_i(A)} f_Y(y/H^j, A)\, dy. \tag{15}$$

Under (14), the decision regions $\Omega_i(A)$ appearing in (15) are defined by

$$\Omega_i(A) = \left\{ y\in\Phi : g_{ij}(y,A) > 0 \quad j\neq i,\ j=1, \ldots, M \right\},\ i = 1, \ldots, M, \tag{16}$$

where

$$g_{ij}(y, A) = P_i \, f_Y(y/H^i, A) - P_j \, f_Y(y/H^j, A). \tag{17}$$

We assume that the functions f_Y satisfy conditions that allow $\Omega_i(A)$ to be well defined (see [3]).

A number of ways of estimating the probability of error from the training samples have been reviewed by Toussaint [8].

However, a new way of estimating the error in the feature space has been proposed and studied by Sagar and the author [9]. This approach essentially considers the discriminant functions g_{ij}, $i = 1, \ldots, M-1$, $j = 2, \ldots, M$, $j>i$, as random variables.

For given $f_Y(y/H^j, A)$, $j = 1, \ldots, M$, we define the conditional distributions F_j, $j = 1, \ldots, M$, as follows: For any given $(M-1)$ real numbers $\gamma_1, \ldots, \gamma_{M-1}$, we have*

$$F_1(\gamma_1, \gamma_2, \ldots, \gamma_{M-1} / A) = \text{Prob}\{ y: g_{12}(y,A) \leq \gamma_1, \ldots, g_{1M}(y,A) \leq \gamma_{M-1} / y \in H^1\}, \quad (18\text{-}1)$$

$$F_2(\gamma_1, \ldots, \gamma_{M-1} / A) = \text{Prob}\{ y: -g_{12}(y\ A) < \gamma_1, g_{23}(y\ A) \leq \gamma_2, \ldots, g_{2M}(y,A) \leq \gamma_{M-1} / y \in H^2\}, \quad (18\text{-}2)$$

. . .

$$F_M(\gamma_1 \ldots, \gamma_{M-1}/ A) = \text{Prob}\{ y: -g_{1M}(y,A) < \gamma_1, -g_{2M}(y,A) < \gamma_2, \ldots, -g_{M-1,M}(y,A) < \gamma_{M-1} / y \in H^M\}. \quad (18\text{-}M)$$

It now follows that the expression (15) for the probability of misclassification is equivalent to**

$$Q(A) = \sum_{j=1}^{M} P_j F_j(0, 0, \ldots, 0/ A). \quad (19)$$

In writing an estimate $\hat{Q}(A)$ of (19) on the basis of training samples, we first use the samples in Λ_C to obtain estimates $\hat{f}_Y(y/H^j, A)$ of the functions $f_Y(y/H^j, A)$ by means of the Parzen kernel [10], [11]. This in turn leads to estimates $\hat{g}_{ij}(y,A)$ of $g_{ij}(y,A)$ via equation (17). We use the training samples in Λ_A to obtain the estimates $\hat{F}_j(., \ldots, ./A)$ of the distributions $F_j(.,\ldots,./A)$ defined by equations (18-1) through (18-M) and hence the estimates $\hat{F}_j(0,\ldots,0/A)$, appearing in (19).

*$y \in H^j$ $\longleftrightarrow$ y arises from a pattern belonging to H^j.

**Strictly speaking, some of the arguments of F_j, $j \neq 1$, in (19) should be written $0-$ instead of 0, to show that we are referring to left-hand limits corresponding to the strict inequalities in equations (18).

By means of the Kiefer-Wolfowitz [12] generalization of the Kolmogorov [13] and Smirnov [14] theory, one can derive bounds on the error in the estimate $\hat{Q}(A)$ of the probability of error $Q(A)$. Details of this study will appear elsewhere [9].

3.4. Optimal Dimensionality of the Feature Space

The optimal dimension m of the feature space Φ is intimately related to the size of the training set Λ. This is because, for sets Λ_A and Λ_C of fixed sizes, the errors in the estimates $\hat{g}_{ij}$ and $\hat{F}_j$ of the functions g_{ij} and F_j decrease with m(that is, the lower the m the closer the $\hat{g}_{ij}$ and $\hat{F}_j$ to g_{ij} and F_j); on the other hand, the probability of error increases with decreasing m, because of the loss of information by reduction of dimensionality. This indicates that the optimal dimension m should be the one that corresponds to the best compromise between the aforementioned two competing effects.

While some papers have appeared previously [15] [16] [17] on the dimensionality-versus-sample-size problem, we have studied this problem in the context of the feature extraction problem using the developments mentioned in the preceding section, and these results appear in [9] and [18].

3.5. The Mathematical Programming Approach

From all the preceding considerations it is clear that the problem of optimal feature extractor design stated at the end of section 2 is a well-defined nonlinear programming problem with the criterion functional given by (10) or (19) to be minimized, and the constraints specified by the properties of the given pattern structure to be recognized, and by the class χ over which the optimization is to be carried out.

Typically, in any given application, a complete study and implementation of this approach would require: (a) the study of conditions for the existence and uniqueness of the optimal A*; (b) development of efficient convergent algorithms for the determination of A*; (c) programming and testing of these algorithms on a computer with simulated and real data bases.

All these phases have been carried out by the author and his associates for Gaussian pattern structures and for the class of linear feature extractors in [3] [18] [19] [20]. Also, phases (a) and (b) of the study of the general nonGaussian nonlinear case has nearly been completed by A. Sagar and the author, and these results will appear in future.

4. Conclusion

A mathematical programming approach has been described for the design of a processor for feature extraction in pattern recognition.

The main consideration in the development of the design algorithm is the optimization of the recognition capability of the system taking into account the realistic constraints appearing in a particular application.

REFERENCES

[1] Meisel, W.S. Computer-Oriented Approaches to Pattern Recognition. Academic Press, New York, 1972.

[2] Chen, C.H., "On Information and distance measures, error bounds, and feature selection", Information Sciences, 10, 159-173 (1976).

[3] de Figueiredo, R.J.P., "Optimal linear and nonlinear feature extraction based on the minimization of the increased risk of misclassification," Rice University Institute for Computer Services and Applications Technical Report #275-025-014 (June, 1974) and Proceedings of the Second Joint International Conference on Pattern Recognition, Copenhagen, Denmark, August, 1974.

[4] Fu, K.S., Syntactic Methods in Pattern Recognition. Academic Press, New York, 1974.

[5] Luenberger, D.G., Introduction to Linear and Nonlinear Programming. Addison-Wesley, Reading, Mass., 1973.

[6] Prenter, P.M., "A Weierstrass theorem for real normed linear space", Bull. Amer. Math. Soc., 75, 860-862 (1969).

[7] Prenter, P.M., "On polynomial operators and equations " in Nonlinear Functional Analysis and Applications. (L. Rall, Editor). Academic Press, New York, 1971.

[8] Toussaint, G.T., "Bibliography on estimation of misclassification", IEEE Trans. on Info. Theory, IT-20, 472-479 (1974).

[9] Sagar, A. and de Figueiredo, R.J.P., to be published.

[10] Cacoullos, T. "Estimation of a multivariate density", Annals of the Institute of Statistical Mathematics (Tokyo), 18, 179-189 (1966).

[11] Bennett, J.O., de Figueiredo, R.J.P., and Thompson, J.R., "Classification by means of B-spline potential functions with application to remote sensing", in Proc. of the Sixth Southeastern Symposium on System Theory, Louisiana State University, Baton Rouge, La., February, 1974.

[12] Kiefer, J. and Wolfowitz, J., "On deviations of the empirical distribution functions of vector chance variable", Trans. Am. Math. Soc., 87, 173-186 (1958).

[13] Kolmogorov, A.N., "Determinazione empirica di una legga di distribuzione", Giornale Instit. Ital. Attuari, 4, 83 (1933).

[14] Smirnov, N., "Sur les ecarts de la courbe de distribution empirique", Mat. Sbornik, 48, 3 (1939).

[15] Highleyman, W.H., "The design and analysis of pattern recognition experiments," Bell Syst. Tech. Journal, 41, 723-744 (1962).

[16] Hughes G.F., "On the mean accuracy of statistical pattern recognizers", IEEE Trans. on Information Theory, IT-14, 55-63 (1968).

[17] Kanal, L.N. and Chandrasekaran, B., "On dimensionality and sample size in pattern recognition", Pattern Recognition, 3, 225-234 (1971).

[18] Starks, S.A., de Figueiredo R.J.P., and Van Rooy, D.L., "An algorithm for optimal single linear feature extraction from several Gaussian pattern classes", To appear in the Intl. Journal of Computer and Information Sciences, vol.6, No.1.

[19] de Figueiredo, R.J.P., "Feature extraction techniques for classification and identification of spectral signatures", Proc. of the 1976 Milwaukee Symposium on Automatic Computation and Control, pp. 303-304 1976.

[20] de Figueiredo, R.J.P., Pau, K.C., Sagar, A.D., Starks, S.A. and Van Rooy, D.L., "An algorithm for extraction of more than one optimal linear feature from several Gaussian pattern classes", Rice University ICSA Technical Report No. 275-025-026 (EE Tech. Report No. 7604), April 1976. (To appear in the Proc. of the 3rd. Joint Intl. Conference on Pattern Recognition, Coronado. California Nov. 1976).