Design of Nearest Prototype Signal Classifiers

W. A. GARDNER, MEMBER, IEEE

Abstract—**The mean-squared-error measure of quality is used as a basis for a general nearest prototype signal-classification methodology. Canonical signal features for this methodology are identified. A consistency requirement is proposed and used to develop a general approach for**

Manuscript received October 30, 1979; revised March 24, 1980.
The author is with the Signal and Image Processing Laboratory, Department of Electrical and Computer Engineering, University of California, Davis, CA 95616.

0018-9448/81/0500-0368$00.75

determining appropriate class prototypes in discriminant space. It is shown that the class indicator, which is a commonly used class prototype in pattern recognition applications, will often violate the consistency requirement. The general results are used to obtain a solution to a previously posed complexity-performance trade-off problem for matched-filter-tapped-delay-line receivers for serial signal classification in an M-ary data transmission system.

I. Introduction

As discussed in [1], the general approach to signal-classifier design that is based on minimizing mean-squared error (MSE) leads to two alternative signal-classification rules: one employs the mode of the minimum-MSE estimate of the posterior distribution, and the other employs the mean. A specific objective of this correspondence is to provide a general solution to the problem of determining when the mean, rather than the mode of the estimated posterior distribution, should be employed. This question was first raised in [2], where it took the form of a complexity–performance trade-off problem for matched-filter-tapped-delay-line receivers for serial signal classification in an M-ary data transmission system. The general formulation of this question, which is answered in this correspondence, was put forth in [1], where an answer to only a special case relating to [2] was given. The question arises in a natural way from the results in [3] and [4] on the characterization of linear minimum-MSE receivers for digital data transmission.

More generally, the objective of this correspondence is to employ the MSE measure of quality as a basis for a general nearest prototype signal classification methodology, to identify canonical signal features for this methodology, and to propose a consistency requirement that leads to a general approach for determining appropriate class prototypes for this methodology. These results together with the results in [5] (on MSE, signal-to-noise ratio (SNR), and other second-order measures of quality) provide an integrated approach to the design of a complete signal classifier, i.e., feature extraction and discriminant–functional design tailored to fit a minimum-distance discrimination rule.

The notation used here is the same as that in [5]; in particular, capitals are used for random quantities and lower case letters are used for samples of random quantities.

II. Nearest Prototype Signal Classification

Let Δ denote the vector of random class indicators

$$\Delta \triangleq \{\Delta(C_1), \Delta(C_2), \cdots, \Delta(C_M)\}. \tag{1}$$

Samples, $\delta(C_i)$, of $\Delta(C_i)$ take on values of either 0 or 1:

$$\delta(C_i) = \begin{cases} 1, & y \in C_i \\ 0, & y \notin C_i, \end{cases} \tag{2}$$

where y is a received waveform and C_i is the class of all sample paths y of the random waveform Y that can occur when the ith of M possible signals is transmitted. The problem of detecting which of the M signals was transmitted corresponds to the problem of classifying y into one of the M classes $\{C_i\}$. This is usually accomplished by transforming the waveform y into an N-tuple of numbers x, called a *discriminant vector*, and then using a decision rule to partition the discriminant space into M decision regions that are in one-to-one correspondence with the M classes $\{C_i\}$ (and M possible transmitted signals).

A common choice for the transformation that maps y into x is a *generalized linear* transformation that minimizes the MSE, $E\{\|X-\Delta\|^2\}$, between X and the vector of random indicators [1], [5] (where $\|\cdot\|$ denotes Euclidean norm). In this case the minimum-MSE X is denoted by $\hat{\Delta}$. As shown in [1], $\hat{\Delta}(C_i)$ is mean-square equivalent to the generalized linear minimum-MSE estimator of the random posterior probability $P[C_i|Y]$:

$$\hat{\Delta}(C_i) = \hat{P}[C_i|Y]. \tag{3}$$

A discrimination rule that is consistent with the use of the minimum-MSE design criterion for deriving its input discriminants is a minimum-distance rule. That is, if the discriminant vector is derived so as to minimize the mean square distance between it and an ideal discriminant (e.g., the vector of class indicators), then it would appear to be appropriate to employ the ideal discriminants as class prototypes in discriminant space and to classify the received waveform y as belonging to the class C_i whose prototype is closest to the computed discriminant. In this correspondence we formalize this approach and propose a general method for prescribing appropriate class prototypes in discriminant space.

The most frequently recommended random prototype vector in the pattern recognition literature is the vector of class indicators Δ. The M samples of this random prototype vector are of the form

$$\delta = \{0, \cdots, 0, 1, 0, \cdots, 0\}, \tag{4}$$

where the 1 is the ith position when y is in class C_i. Observe that the class-conditional vector $\Delta|C_i$ is given by (4) and is therefore nonrandom. The corresponding minimum-distance discrimination rule is

$$\left\{ \min_j \{\|\hat{\delta} - \delta|C_j\|\} = \|\hat{\delta} - \delta|C_i\| \right\} \Leftrightarrow \{\text{classify } y \text{ in } C_i\}. \tag{5}$$

It follows from (3) that (5) is equivalent to the decision rule

$$\left\{ \max_j \{\hat{P}[C_j|y]\} = \hat{P}[C_i|y] \right\} \Leftrightarrow \{\text{classify } y \text{ in } C_i\}, \tag{6}$$

which chooses the mode of the estimated posterior distribution (see Theorem 1).

In spite of the popularity of the random prototype vector Δ in pattern recognition applications, it has been shown by example that Δ is not appropriate for some important signal-detection applications [1], [4], [6] because even when *class variability* (e.g., additive noise) vanishes, classification performance is unacceptable for linear discriminant functionals in these applications. It is also shown in [1], [4], [6] that for these applications, use of an even simpler prototype vector yields acceptable classification performance. The *estimation-theorists' approach*[1] *to signal detection* that is recommended in [1] to circumvent the inappropriateness of Δ is referred to here as the *nearest prototype signal-classification methodology* and is characterized by the classification rule

$$\left\{ \min_j \{\|\hat{s} - s|C_j\|\} = \|\hat{s} - s|C_i\| \right\} \Leftrightarrow \{\text{classify } y \text{ in } C_i\}, \tag{7}$$

where s is a sample of an N-tuple S of random variables whose class-conditional values $S|C_j$ are (like those of Δ, (4)) nonrandom. For example, S could be a vector of random signal-modulation parameters. Specifically, for amplitude-shift keying (ASK) and discrete pulse-amplitude modulation (PAM), S could be the random scalar ($N=1$) amplitude parameter; for phase-shift keying (PSK), S could be either the random scalar ($N=1$) phase parameter or the random 2-tuple of in-phase and quadrature amplitude parameters ($N=2$); for frequency-shift keying (FSK), S could be the random scalar ($N=1$) frequency parameter. In all cases, S could also be chosen to be $S=\Delta(N=M)$.

Using the characterization

$$s = \theta\delta, \tag{8}$$

where the $N \times M$ nonrandom matrix θ has the (nm)th element

$$\theta_{nm} \triangleq s_n|C_m, \tag{9}$$

[1] This term was borrowed from [7, p. 387].

and s_n is the nth element of $\boldsymbol{s}$, we obtain the characterization

$$\hat{\boldsymbol{s}} = \boldsymbol{\theta}\hat{\boldsymbol{\delta}} \tag{10}$$

for the generalized linear minimum-MSE estimator for $\boldsymbol{S}$ [5]. This characterization suggests a new interpretation of $\hat{\boldsymbol{\delta}}$ as a *canonical second-order-optimal feature vector* (rather than a discriminant vector) and an interpretation of $\boldsymbol{\theta}$ as a set of N linear discriminant functionals whose outputs comprise a discriminant vector $\hat{\boldsymbol{s}}$ which is to be used in the minimum-distance rule (7). The feature vector $\hat{\boldsymbol{\delta}}$ is canonical in the sense that every minimum-MSE (and other second-order-optimal [5]) discriminant vector $\hat{\boldsymbol{s}}$ can be obtained from $\hat{\boldsymbol{\delta}}$ with the linear transformation (10). This role of $\hat{\boldsymbol{\delta}}$ is illustrated for matched-filter-tapped-delay-line receivers for serial detection of M-ary data signals in [3] and [4]. The alternative interpretation in which $\hat{\boldsymbol{\delta}}$ is the discriminant vector and $\boldsymbol{\theta}$ is interpreted as a weight matrix in the weighted-norm distance measure which defines the discriminator (7) is less viable because when the appropriate $\boldsymbol{\theta}$ is not of full rank (as discussed in the following) the distance measure is not a valid metric; i.e., it does not distinguish between vectors contained in certain subspaces of discriminant space.

Use of (3), (9), and (10) yields

$$\hat{\boldsymbol{s}} = \sum_{m=1}^{M} (\boldsymbol{s}|C_m)\hat{P}[C_m|\,\boldsymbol{y}], \tag{11}$$

from which it follows that $\hat{\boldsymbol{s}}$ is the mean of the estimated posterior distribution of $\boldsymbol{S}$. Thus rule (7) chooses the closest prototype $\boldsymbol{s}|C_i$ to the *mean* of $\boldsymbol{S}$ with respect to the estimated posterior distribution; whereas rule (6) (and rule (5)) chooses the *mode*. These two rules coincide when $\boldsymbol{S} = \boldsymbol{\Delta}$, i.e., when $\boldsymbol{\theta}$ is the identity matrix.

Having the characterization (10), we are now in a position to propose a formal approach to the prescription of appropriate class prototypes $\{\boldsymbol{s}|C_j\}_1^M$, which is equivalent to the prescription (design) of the matrix $\boldsymbol{\theta}$. The basis for the approach is the imposition of the following *consistency requirement*:[2]

$$\left[\hat{\boldsymbol{s}}|C_j\right]_{V=0} = \boldsymbol{s}|C_j, \qquad j = 1,2,\cdots,M, \tag{12}$$

where V represents *variability*. That is, when class variability vanishes, the estimate of each class prototype must be identical to that prototype. An implicit definition of the condition $V=0$ is given by the necessary and sufficient conditions

$$\text{(i)} \quad [\,\boldsymbol{y}|C_i]_{V=0} = \text{unique waveform},$$

$$\text{(ii)} \quad [\,\boldsymbol{y}|C_i]_{V=0} \neq \left[\,\boldsymbol{y}|C_j\right]_{V=0}, \qquad \text{for} \quad i \neq j. \tag{13}$$

An explicit definition of the condition $V=0$ depends on the particular signal-classification problem. For example, for one-shot detection of known signals in additive noise (W), $V = W$; and for serial detection with additive noise and intersymbol interference (ISI), $V = \{W, \text{ISI}\}$. Thus V represents "channel randomness" in an abstract sense. Substitution of (8) and (10) into (12) yields the *prototype design equation*

$$\boldsymbol{\theta}\cdot\boldsymbol{\psi} = \boldsymbol{\theta}, \tag{14}$$

where the $M \times M$ matrix $\boldsymbol{\psi}$ has an (ij)th element defined by

$$\psi_{ij} \triangleq \left[\hat{\boldsymbol{\delta}}\,(C_i)|C_j\right]_{V=0}. \tag{15}$$

Substitution of [5, eq. 32] into (15) yields the explicit formula

$$\psi_{ij} = p_i\Big\{1 + \Big(\big\{\left[\boldsymbol{z}|C_j\right]_{V=0} - \boldsymbol{m}(\boldsymbol{Z})\big\},\ \boldsymbol{k}(\boldsymbol{Z})^{-1}\{\boldsymbol{m}(\boldsymbol{Z}|C_i) - \boldsymbol{m}(\boldsymbol{Z})\}\Big)_{\Lambda}\Big\}, \tag{16}$$

where

$$\boldsymbol{Z} = \boldsymbol{g}(\boldsymbol{Y}), \tag{17}$$

and $(\cdot,\cdot)_{\Lambda}$ denotes an inner product in feature space, and $\boldsymbol{m}(\boldsymbol{Z})$ and $\boldsymbol{k}(\boldsymbol{Z})$ denote mean vector and covariance operator, respectively (cf. [5]). Thus once the (waveform) feature extraction transformation $\boldsymbol{g}(\cdot)$ has been prescribed, and variability has been explicitly defined for the problem at hand, the matrix $\boldsymbol{\psi}$ is completely specified. There is still some freedom in the choice of prototypes, however, since (14) does not possess a unique solution.

One of the most important characteristics of $\boldsymbol{\psi}$ is its rank (R_{ψ}), since this sets an upper limit on the rank of any $\boldsymbol{\theta}$ that satisfies (14): $R_{\theta} \leq R_{\psi} \leq M$. An important conclusion which follows immediately is that the conventional prototype vector $\boldsymbol{s} = \boldsymbol{\delta}$ violates the consistency requirement (14) if $\boldsymbol{\psi}$ is not the identity matrix. An application where $\boldsymbol{\psi}$ is not of full rank (and therefore is not the identity) is the detection of linearly dependent M-ary signals (e.g., ASK, PSK, APK, etc.) in additive noise using a linear receiver ($\boldsymbol{g}(\cdot)$ = identity transformation) [1]. As a matter of fact, when $\boldsymbol{\psi}$ is of less than full rank, receiver complexity can be significantly reduced simply because only $N = R_{\psi} < M$ statistics need to be computed. For example, as shown in [3] and briefly discussed in [2], the number of tapped-delay lines in an adaptive receiver for high-speed data transmission can be significantly reduced from NM to N^2 ($N = R_{\psi}$) by choosing $\boldsymbol{\theta}$ to be $N \times M$ rather than $M \times M$. Specifically, for APK signaling $N = 2$, and with four phases and two amplitudes, $M = 8$; thus only four tapped-delay lines rather than 16 are needed. These results answer the query in [2]: "It would be interesting to know whether the more complex structure [which uses $R_{\theta} = M$ when $R_{\psi} < M$] yields any significant advantage [in probability-of-error performance]." The answer is an emphatic *no*. Rather, the more complex structure (with $R_{\theta} > R_{\psi}$) yields a higher probability of error as discussed[3] (in terms of examples only) in [1], [4], and [6].

When $\boldsymbol{\psi}$ is of full rank it is helpful to know under what conditions a proposed set of prototypes $\{\boldsymbol{s}|C_j\}_1^M$ (e.g., values of a vector of signal modulation parameters) is equivalent to the indicator vectors $\{\boldsymbol{\delta}|C_j\}_1^M$ in the sense that the two classification rules (5) and (7) are equivalent. In the remainder of this correspondence we compare these two rules with each other as well as with the maximum-estimated-posterior-probability rule (6).

The norm used in rules (5) and (7) as referred to in the following theorem can be any norm induced by an inner product, not just the Euclidean norm (e.g., a weighted norm); this is emphasized by use of the modifier *Hilbert*.

Theorem 1

i) The nearest-indicator (Hilbert) rule (5) is equivalent to the maximum-estimated-probability rule (6).

ii) The rules (5) and (6) are equivalent to the nearest prototype (Hilbert) rule (7) if and only if the class prototypes $\{\boldsymbol{s}|C_j\}_1^M$ are mutually equidistant.

The norm used in rules (5) and (7) as referred to in the following theorem need not be induced by an inner product (e.g., any of the l^p norms can be used); this is emphasized by use of the modifier *Banach*.

Theorem 2

i) For binary classification the nearest indicator (Banach) rule (5) is equivalent to the maximum-estimated-probability rule (6).

ii) The rules (5) and (6) are equivalent to the nearest prototype (Banach) rule (7) if and only if the class prototypes $\{\boldsymbol{s}|C_j\}_1^2$ are distinct.

These two general equivalences are related since every two distinct vectors constitute a mutually equidistant set. Since non-

[2] This consistency requirement is conceptually related to the *well-structured condition* employed in empirical discriminant analysis [8], [9].

[3] The results referred to in [1, ref. [17]] are contained herein.

distinct prototypes would never be used in practice, rules (5), (6), and (7) are always equivalent for binary classification. Proofs of the preceeding theorems are given in the Appendix.

III. Summary

The results in this correspondence can be summarized as follows. The signal classification rule (6) classifies the received signal y as belonging to the class C_i whose estimated posterior probability is largest; i.e., this rule employs the mode of the estimated posterior distribution. The signal classification rule (7) classifies y as belonging to the class C_i for which the class prototype $s|C_i$ is closest to the mean $\hat{S}$ of the prototypes $\{s|C_j\}_1^M$, with respect to the estimated posterior distribution. We can conclude that in general the mean, rather than the mode, of the estimated posterior distribution should be used, since the mode yields an inconsistent (cf. (12)) rule except in the special case for which the matrix of class-conditional estimated class indicators evaluated at zero variability (15) is the identity matrix, whereas a set of class prototypes for which the mean yields a consistent rule can always be found by solving the linear matrix equation (14) for the matrix θ (cf. (9)). As a further result, necessary and sufficient conditions for equivalence among classification rules based on the mean of the estimated posterior distribution (7), but using different sets of class prototypes, and rules (5) and (6) are determined in Theorems 1 and 2.

Since the minimum-MSE discriminant-design criterion employed in this correspondence and in [1] is equivalent to the maximum-SNR discriminant-design criterion [5], and both of these criteria can be characterized in terms of a waveform scatter ratio [5], then the consistency requirement proposed herein for discriminator design based on MSE (i.e., for the selection of appropriate class prototypes) is applicable, more generally, to discriminator design based on other second-order measures of quality, as discussed at length in [5]. The results in this correspondence, together with the results in [5], provide an integrated approach to the design of a complete signal classifier, i.e., feature extraction and discriminant-functional design tailored to fit a minimum-distance (nearest prototype) discrimination rule.

Appendix
Proofs of Theorems

Proof of Theorem 1 on Multiclass Discrimination

Expansion of the Hilbert norm in (7) in terms of inner products yields the following equivalent criterion for classification:

$$\max_j \left\{ (\hat{s}, s|C_j) - \tfrac{1}{2}\|s|C_j\|^2 \right\}. \tag{A1}$$

Substitution of (10) and [5, (32)] yields

$$\max_j \left\{ \sum_{m=1}^{M} (s|C_m, s|C_j)\hat{P}[C_m|y] - \tfrac{1}{2}\|s|C_j\|^2 \right\}, \tag{A2}$$

which can be reexpressed as

$$\max_j \left\{ \tfrac{1}{2} \sum_{m=1}^{M} \left(\|s|C_j\|^2 + \|s|C_m\|^2 - d_{jm}^2 \right)\hat{P}[C_m|y] - \tfrac{1}{2}\|s|C_j\|^2 \right\}, \tag{A3}$$

where d_{jm} is the distance between $s|C_j$ and $s|C_m$. Use of [1, appendix A],

$$\sum_{m=1}^{M} \hat{P}[C_m|y] = 1, \tag{A4}$$

reduces (A3) to

$$\max_j \left\{ \sum_{m=1}^{M} \|s|C_m\|^2 \hat{P}[C_m|y] - \sum_{m=1}^{M} d_{jm}^2 \hat{P}[C_m|y] \right\}, \tag{A5}$$

which is equivalent to

$$\min_j \left\{ \sum_{m=1}^{M} d_{jm}^2 \hat{P}[C_m|y] \right\}. \tag{A6}$$

This criterion is independent of the class prototypes $\{s|C_j\}_1^M$ if and only if d_{jm} is independent of j and m (for $j \neq m$), i.e., if and only if $\{s|C_j\}_1^M$ are mutually equidistant. In this case (A6) becomes

$$\min_j \left\{ d^2 \sum_{m \neq j} \hat{P}[C_m|y] \right\}, \tag{A7}$$

which upon substitution of (A4), becomes

$$\max_j \left\{ \hat{P}[C_j|y] \right\}. \tag{A8}$$

This criterion is identical to that used in rule (6). Furthermore, since $\{\delta|C_j\}_1^M$ are mutually equidistant, rule (5) is equivalent to rule (7).

Proof of Theorem 2 on Binary Discrimination

Substitution of (10) into (7) for $M = 2$ yields

$$\|\hat{P}[C_1|y]s|C_1 + \hat{P}[C_2|y]s|C_2 - s|C_1\| \underset{C_2}{\overset{C_1}{\lessgtr}} \|\hat{P}[C_1|y]s|C_1 + \hat{P}[C_2|y]s|C_2 - s|C_2\|. \tag{A9}$$

Use of (A4) in (A9) yields

$$|\hat{P}[C_2|y]|\|s|C_1 - s|C_2\| \underset{C_2}{\overset{C_1}{\lessgtr}} |\hat{P}[C_1|y]|\|s|C_1 - s|C_2\|. \tag{A10}$$

By hypothesis, the vectors $s|C_1$ and $s|C_2$ are distinct; thus the nonzero norm can be cancelled in (A10), and (A4) can then be used to obtain the equivalent rule

$$\hat{P}[C_2|y] \underset{C_2}{\overset{C_1}{\lessgtr}} \hat{P}[C_1|y], \tag{A11}$$

in which the absolute value operation has been removed. This rule is identical to rule (6) for $M = 2$. Furthermore, by letting $s = \delta$, it follows that (5) is equivalent to (7). Hence all three rules, (5), (6), and (7), are equivalent for binary classification.

References

[1] W. A. Gardner, "Structurally constrained receivers for signal detection and estimation," *IEEE Trans. Commun.*, vol. COM-24, pp. 578–592, June 1976. Eratta: the following errors in [1] should be noted. 1) Interchange δ and {following (3). 2) Change first + in (42) to −. 3) Interchange "not H_i" and "not H_j" in (73). 4) Replace m_x in (80) with closest X_i to m_x. 5) Insert "finite" before "mean" preceding (A1). 6) Replace $\delta(X/y)$ with $\delta(X)$ in (A2) and following (A4). 7) Change 535–544 to S34–S46 in reference 18 of [1].

[2] A. R. Kaye and D. A. George, "Relationships between LMS compensators for intersymbol interference in linear and nonlinear modulation schemes," *IEEE Trans. Inform. Theory*, vol. IT-19, p. 244, Mar. 1973.

[3] W. A. Gardner, "The structure of least mean square linear estimators for synchronous M-ary signals," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 240–243, Mar. 1973.

[4] ——, "An equivalent linear model for marked and filtered doubly stochastic Poisson processes with application to MMSE linear estimation for synchronous M-ary optical data signals," *IEEE Trans. Commun.*, vol. COM-24, pp. 917–921, Aug. 1976.

[5] ——, "A unifying view of second order measures of quality for signal classification," *IEEE Trans. Commun.*, vol. COM-28, pp. 807–816, June 1980. Errata: The following error in [5] should be noted. The second set of quantities, p_i, $\boldsymbol{m}_i$ and $\boldsymbol{k}_i$, in the third from last line on p. 814 should be replaced with $\tilde{p}_i$, $\tilde{\boldsymbol{m}}_i$ and $\tilde{\boldsymbol{k}}_i$.

[6] ——, "Comparison of estimation-based classifiers," in *Abstracts of Papers Presented at IEEE Int. Symp. on Inform. Theory*, Ronneby, Sweden, June 21–24, 1976, p. 44.

[7] J. Ziv and M. Zakai, "Some lower bounds on signal parameter estimation," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 386–391, May 1969.
[8] L. Fisher and J. W. Van Ness, "Admissable discriminant analysis," *J. Am. Stat. Ass.*, vol. 68, pp. 603–607, Sept. 1973.
[9] J. Rubin, "Optimal classification into groups: An approach for solving the taxonomy problem," *J. Theor. Bio.*, vol 15, pp. 103–44, 1967.