and upon applying the vector calculus identity

$$\nabla \cdot (\psi \bar{A}) = \bar{A} \cdot \nabla \psi + \psi \nabla \cdot \bar{A}$$

to the first term, (13) follows.

Corollary: The divergence of $\hat{s}(x)$ equals its variance at x; i.e.,

$$\nabla \cdot \hat{s}(x) = V(x). \tag{14}$$

This corollary relates spatial variations in $\hat{s}(x)$ to conditional expectations at x; and, in addition, it enables us to state that the CME is completely specified by its variance function $V(x)$. This statement is a consequence of Helmholtz's theorem, which states that a vector is completely specified by its divergence and curl. (Recall that as $\hat{s}(x)$ is conservative, $\nabla \times \hat{s}(x) = 0$.)

The primary value of Property 3 and its corollary is that they relate both the likelihood ratio and CME to a conditional variance function of the signal. Hence, by viewing $L(x)$ as a potential function and $\hat{s}(x)$ as a conservative vector field, the mathematics of potential theory can be applied to solving problems in signal detection theory.

III. Conclusions

This correspondence has noted a fundamental property relating optimum detection and CME for random signals in white Gaussian noise for discrete-time processes, and has discussed the role of the estimation–correlation operation in forming an optimum decision statistic.

By viewing the log-likelihood ratio as a potential function, the CME of the signal was shown to constitute a conservative vector field. This concept was used to show the intimate connection between spatial variation (divergence) of the CME and the conditional signal variance. These results suggest that the mathematics of potential theory might play an important role in furthering the theory of signal detectability.

C. P. Hatsell[3]
Dep. Elec. Eng.
USAF Inst. Technol.
Wright-Patterson AFB, Ohio
L. W. Nolte
Dep. Elec. Eng.
Duke Univ.
Durham, N.C. 27706

[3] Formerly with the Department of Electrical Engineering, Duke University, Durham, N.C.

Note on Optimal Selection of Independent Binary-Valued Features for Pattern Recognition

Abstract—**Given a set of conditionally independent binary-valued features, a counter example is given to a possible claim that the best subset of features must contain the best single feature.**

Recently, Elashoff *et al.*[1] showed that for optimal selection of a subset of independent binary-valued features, the features generally may not be evaluated independently. Specifically, an example is given[1] in which, given three independent variables x_1, x_2, and x_3 such that $\varepsilon(x_1) < \varepsilon(x_2) < \varepsilon(x_3)$, where $\varepsilon(x_i)$ is the error probability when the ith variable alone is used, the first and third variables are jointly better than the first and second variables. In other words, $\varepsilon(x_1,x_3) < \varepsilon(x_1,x_2)$, where $\varepsilon(x_i,x_j)$ is the probability of error when the ith and jth variables are used together. In this note, the results of Elashoff *et al.* are carried one step further, and it is shown that the best pair of variables need not contain the best single variable.

Let there be two equiprobable pattern classes C_1 and C_2, and let $\alpha_i = P(x_i = 1 \mid C_1)$ and $\beta_i = P(x_i = 1 \mid C_2)$, $i = 1,2,3$, where $P(x_i = 1 \mid C_j)$ is the conditional probability that the ith variable takes on the value ONE conditioned on the jth pattern class. As in Elashoff *et al.*, let the following assumptions be made:

1) $\varepsilon(x_1) < \varepsilon(x_2) < \varepsilon(x_3)$;
2) $\alpha_i < \beta_i$, $i = 1,2,3$;
3) $\beta_1 - \alpha_1 > \beta_2 - \alpha_2 > \beta_3 - \alpha_3$.

For simplicity of notation, let $l_i = (\beta_i - \alpha_i)$, $h_i = \frac{1}{2}(1 - \alpha_i - \beta_i)$, and $D_{ij} = |h_i| - |h_j|$. It is shown by Elashoff *et al.* that for two conditionally independent variables x_i and x_j, the minimum error probability is given by

$$\varepsilon(x_i,x_j) = \tfrac{1}{2}[\varepsilon(x_i) + \varepsilon(x_j) - l_i\,|h_j| - l_j\,|h_i|], \tag{1}$$

where $\varepsilon(x_k) = \frac{1}{2}[1 - (\beta_k - \alpha_k)]$ for $k = i,j$. From (1) and conditions 1), 2), and 3) above, it can easily be shown that for $\varepsilon(x_1,x_2) < \varepsilon(x_2,x_3)$, a sufficient condition is given by

$$|h_1| > |h_3| \tag{2}$$

and a necessary and sufficient condition is given by

$$D_{31} < \frac{1}{2}\left(\frac{l_1 - l_3}{l_2}\right)(1 + 2\,|h_2|). \tag{3}$$

Consider as an example, three features x_1, x_2, and x_3 chosen so as to violate (3) such that $\varepsilon(x_1) < \varepsilon(x_2) < \varepsilon(x_3)$. Such examples are not difficult to find. The probabilities of the three features conditioned on the two pattern classes are given by

$$\alpha_1 = 0.10 \qquad \alpha_2 = 0.05 \qquad \alpha_3 = 0.01$$
$$\beta_1 = 0.90 \qquad \beta_2 = 0.80 \qquad \beta_3 = 0.71.$$

Substituting these figures into (1) yields the following results:

$\varepsilon(x_1) = 10$ percent $\qquad \varepsilon(x_1,x_2) = 8.25$ percent
$\varepsilon(x_2) = 12.5$ percent $\qquad \varepsilon(x_1,x_3) = 6.9$ percent
$\varepsilon(x_3) = 15$ percent $\qquad \varepsilon(x_2,x_3) = 5.875$ percent.

From these results it is observed that, although all pairs of features are better than the best single feature, the pair consisting of the two worst single features is much better than the pair consisting of the two best single features. Furthermore, the best pair does not contain the best single feature; in fact, the best pair is made up of the worst single features.

Godfried T. Toussaint
Dep. Elec. Eng.
Univ. British Columbia
Vancouver, B.C., Canada

Manuscript received August 7, 1970; revised January 27, 1971. This research was supported in part by the National Research Council of Canada under Grant NRC A-3308 and by the Defence Research Board of Canada under Grant DRB 2801-30.

[1] J. D. Elashoff, R. M. Elashoff, and G. E. Goldman, "On the choice of variables in classification problems with dichotomous variables," *Biometrika*, vol. 54, 1967, pp. 668–670.

Comments on "A Modified Figure of Merit for Feature Selection in Pattern Recognition"

In a recent correspondence [1] a modification of the conventional mutual-information effectiveness criterion for feature selection in pattern recognition was described. However, there seems to be some confusion between selecting a subset of features and selecting features individually. This apparent confusion may confuse the reader further

Manuscript received August 15, 1970; revised January 27, 1971.