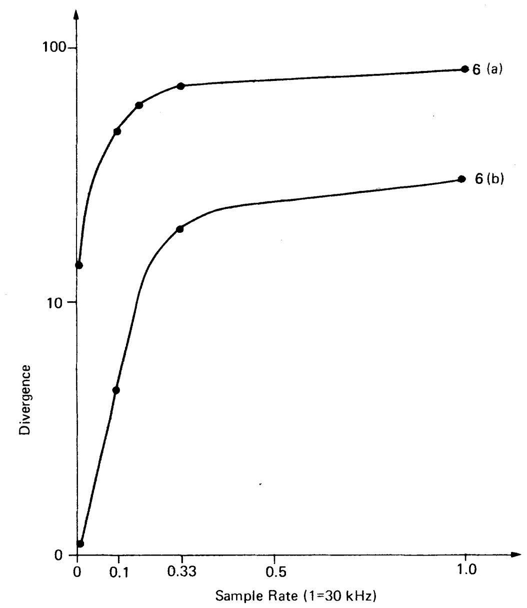

Fig. 7. Divergence versus sample rate.

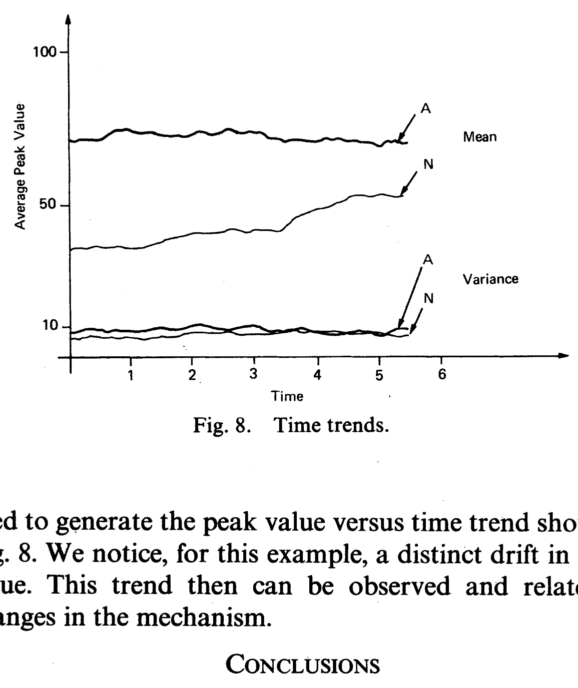

Fig. 8. Time trends.

used to generate the peak value versus time trend shown in Fig. 8. We notice, for this example, a distinct drift in mean value. This trend then can be observed and related to changes in the mechanism.

Conclusions

Signature analysis, using second-order statistical data analysis for pattern recognition of time signals, for mechanical diagnostics of small mechanisms has been studied. A rationale for this method of analysis was established and one condition for successful diagnostics, namely signal-event independence, was discussed. As a result, preprocessing, feature extraction, and recognition algorithms were formulated and applied to diagnosing mechanical subfunction failures relating to mechanical impacts in a small cyclic mechanism. Experimental data tend to validate second-order statistical assumptions for the mechanisms under study. A significant bonus to those who develop diagnostic systems is the design aids and trend tracking techniques which become readily available when second-order statistics are used. The applicability of divergence and the separation of waveform instability into short- and long-term variations resulted from applying these tools.

References

[1] D. L. George, "Noise and vibration in Q.C. design, and maintenance," *Sound and Vibration*, Oct. 1967, pp. 21–26.
[2] A. Hannavy, "New detection devices help predict potential failure," *Prod. Eng.*, Mar. 27, 1967, pp. 37–44.
[3] F. J. Lavoid, "Signature analysis: Product early warning system," *Mach. Des.*, Jan. 23, 1969, pp. 151–160.
[4] J. Page "Recognition of patterns in jet engine vibration signals," *IEEE Comput., Conf. Dig.*, Sept. 1967, pp. 102–105.
[5] K. D. Roberson and B. D. Olson, "The value of maintenance," presented at the Federal Aviation Administration Reliability and Maintenance Symp., Oklahoma City, Okla., Nov. 8, 1967, Paper AD664382.
[6] R. A. Rogers, "Integration of airborne data recorders and ground-based computers for engine maintenance purposes," presented at the FAA R & M Symp., Oklahoma City, Okla., Nov. 8, 1967, Paper AD664383.
[7] R. B. Tatge, "Acoustical techniques for machinery diagnostics," presented at the Acoustical Society of America Conf., Ottawa, Canada, May 21–24, 1968.
[8] H. G. Zerkle, "Development and test of the engine analyzer system," Wright Patterson Air Force Base, Dayton, Ohio., Aug. 1967, Paper AD665158.
[9] P. W. Becker, *Recognition of Patterns Using the Frequencies of Occurrence of Binary Words.* Copenhagen, Denmark: Polyteknisk, 1968, 216 pps.
[10] T. Pavidis, "A linguistic model for waveform analysis," Princeton University, Princeton, N. J., Mar. 1969, Paper AD684360.
[11] R. H. Peterson and R. L. Hoffman, "A new technique for dynamic analysis of acoustical noise," *IBM J. Res. Develop.*, vol. 9, May 1965, pp. 205–209.
[12] V. T. Rhyne, "A digital system for enhancing the fetal electrocardiogram," *IEEE Trans. Bio-Med. Eng.*, vol. BME-16, Jan. 1969, pp. 80–86.
[13] C. R. Trimble, "What is signal averaging," *Hewlett-Packard J.*, vol. 19, Apr. 1968, pp. 2–16.
[14] N. J. Nilsson, *Learning Machines, Foundations of Trainable Pattern-Classifying Systems.* New York: McGraw-Hill, 1965.
[15] K. Fukunaga and T. F. Krile, "Calculation of Bayes' recognition error for two multivariate Gaussian distributions," *IEEE Trans. Comput.*, vol. C-18, Mar. 1969, pp. 220–229.
[16] ——, "A minimum distance feature selection criteria," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-14, Sept. 1968, pp. 780–782.
[17] T. Kailath, "The divergence and Bhattacharyya distance measures in signal detection," *IEEE Trans. Commun. Technol.*, vol. COM-15, Feb. 1967, pp. 52–60.
[18] T. Marill and D. M. Green, "On the effectiveness of receptors in recognition systems," *IEEE Trans. Inform. Theory*, vol. IT-9, Jan. 1963, pp. 11–17.

A Direct Method of Nonparametric Measurement Selection

A. WAYNE WHITNEY

Abstract—A direct method of measurement selection is proposed to determine the best subset of *d* measurements out of a set of *D* total measurements. The measurement subset evaluation procedure directly employs a nonparametric estimate of the probability of error given a finite design sample set. A suboptimum measurement subset search procedure is employed to reduce the number of subsets to be

Manuscript received October 29, 1970; revised March 26, 1971. This work was supported in part by the Rome Air Development Center, Griffiss AFB, Rome, N. Y., under Contract F 30602-69-C-0227. A preliminary version of this note was presented at the IEEE Symposium on Feature Extraction and Selection in Pattern Recognition, Argonne, Ill., October 5–7, 1970.

The author is with the General Electric Company, Syracuse, N. Y. 13201.

evaluated. The primary advantage of the approach is the direct but nonparametric evaluation of measurement subsets, for the *M* class problem.

Index Terms—Feature space evaluation, measurement (feature) selection, nearest neighbor rule, pattern classification.

INTRODUCTION

A basic problem in signal and pattern classification is to determine which measurements (features) should be employed for minimum error in classification. A large number of pattern classification problems involve classification of patterns into one of M classes, which are defined only by a limited number of labeled representative patterns. In general, additional knowledge may be available concerning the phenomenon that produced the patterns but a complete and accurate problem model is usually lacking.

Suppose that a vector representation for the patterns has been chosen (including possibly useless and dependent measurements), and the set of labeled representative patterns has been mapped into a set of representative measurement vectors. This finite set of N representative measurement vectors, corresponding to the N representative patterns, is called the design sample set. This design sample set is assumed to be statistically representative of the underlying probability distributions of the measurement vectors and is used to infer the corresponding M decision regions. If these decision regions were designed to minimize the probability error with complete knowledge of the underlying probability distributions, then inclusion of an additional measurement could improve but not degrade the performance of the optimum decision rule. In practice these decision regions are designed on the basis of a fixed number of design samples, and it has been observed both theoretically and experimentally that the performance of the resulting classification procedure does not always improve as the dimensionality of the measurement vector is increased. In some cases, the performance even deteriorates with increasing dimensionality [1], [4], [7]. This possibility of decrease in performance, with increase in system implementation complexity, places a premium on selection and evaluation of measurements.

Various indirect techniques have been suggested for evaluation and selection of measurements for classification procedures. These techniques range from information theoretic measures [9], [10] to the projection techniques of principal component analysis [5]. A major disadvantage of these techniques is that the criterion employed to determine the best measurements or projections has no simple relationship to the probability of error [13], except for the case of special probability distributions [10], [12]. Also, any functional dependence between an indirect measure of performance and probability of error is dependent on the form of the underlying distributions.

A Direct Method of Nonparametric Measurement Subset Evaluation

A central problem in the evaluation of a measurement subset is to estimate the minimum probability of error associated with the subset, given a fixed number of design samples. Early techniques of estimation employed either a resubstitution method or a holdout method of the design samples. The resubstitution method uses the total design sample set to design the decision rule and simply resubstitutes the same set into the designed decision rule to estimate the performance. This technique has been found to be misleading in the sense that the method gives too optimistic an estimate of the probabilities of misclassification [8]. The holdout method of performance evaluation uses a subset of the total sample set to design the decision rule and uses the remaining samples as a test set to estimate the performance. This holdout method does not make efficient use of the samples, since a reduced sample size does not permit either accurate design or evaluation of the classification technique [8].

The evaluation technique employed will be called the "leave-one-out" method (U method [8]). The essential idea of this technique is to design the decision rule using $N-1$ samples of the total N samples and apply the decision rule to the one remaining sample. This process is repeated for all partitions of size $N-1$ for the design sample set and size one for the test set. The probability of error is then estimated by the ratio of the test samples incorrectly classified to the total number of samples classified. This procedure has the advantage of an unbiased estimate of the probability of error based on $N-1$ design samples, but has the disadvantage of requiring the decision rule to be designed N times for each evaluation.

The nonparametric classification technique to be used in the evaluation procedure is the k nearest neighbor rule, which assigns an unidentified sample to the class most heavily represented among the k nearest samples of the N design samples. Let R^* be the Bayes' probability of error and $R(k)$ be the asymptotic probability of error associated with the $k-NN$ rule. When the number of design samples N becomes large, the probability of error of a single nearest neighbor rule is bounded by

$$R^* \leq R(1) \leq 2R^*(1 - R^*)$$

for the two-class problem [3]. This asymptotic bound is independent of the metric used to measure the distance between samples. These inequalities may be inverted to obtain asymptotic bounds for an estimate of Bayes' risk using the $k-NN$ rule and the "leave-one-out" method:

$$\left[1 - \sqrt{1 - 2R(1)}\right]/2 \leq R^* \leq R(1).$$

A similar bound could be obtained for R^* in terms of $R(k)$ with $k>1$, $M=2$, but this bound cannot be represented in closed form since the original bound was not in closed form. Bounds on $R(k)$ have also been demonstrated for $k=1$, and $M>2$ [3]. The concept of using the $k-NN$ risk to bound the Bayes risk seems to have been arrived at independently by Cover [2], Hart [6], and Whitney [14]. A fundamental practical limitation of this evaluation procedure is that the estimate $\hat{R}(k)$ may differ from the asymptotic value of risk $R(k)$ for any finite number N of design samples. If $\hat{R}(k)$

differs greatly from $R(k)$, then the N design samples do not provide enough information to accurately estimate the minimum probability of error of the underlying distributions. Therefore, it is doubtful that any other measurement subset evaluation procedure would be more accurate unless additional assumptions can be employed. When applying indirect measures of performance based on a finite design sample, one is also confronted with the variation of the estimated indirect measure from the expected value of the indirect measure as well as the possible relationship between the expected indirect measure value and the minimum probability of error.

A basic concept in using the $k-NN$ decision rule as an evaluation tool is that its performance should be approximately equal to the performance attainable by any decision rule that has a simple implementation. The efficiency of using the $k-NN$ decision rule is that only $N-1$ distances need to be computed and searched for the k smallest distances to perform the "leave-one-out" method of evaluation on one partition. If the $N(N-1)/2$ distances may be stored, then only the appropriate $N-1$ distances need be searched for the k smallest distances for each of the N partitions. These computations can also be reduced by appropriate choice of the metric and use of a fast sorting procedure specifically designed to find the k smallest out of N distances.

A Suboptimum Measurement Subset Search Procedure

Suppose that, out of a list of D total measurements, we wish to select a subset of d measurements for a classification system. Using simple combinatorial arguments, the total number of subsets of size d that can be considered for selection out of D possible measurements is given by

$$\binom{D}{d} = \frac{D!}{(D-d)!d!}.$$

It is easily seen that the number of subsets to be evaluated can become very large for modest measurement selection problems. For example, to select the best subset of ten measurements out of twenty possible measurements requires evaluation of 184 756 subsets.

One approach to reduce the number of subsets evaluated is to use a suboptimum search procedure. One logical choice for this suboptimum search procedure [11], [4] is to select the best single measurement first, then the best pair is selected where the pair includes the best single measurement selected. This process is continued by selecting a single measurement that appears to be best when combined with the previously selected subset of measurements. Hence, the number of subsets searched to find a subset of d measurements from D possible measurements is given by

$$\sum_{i=1}^{d} (D-i+1) = d[D-(d-1)/2].$$

Thus, for the previous example of selecting ten measurements out of twenty possible measurements requires the evaluation of 155 subsets, which is three orders of magnitude less than the number of subsets that must be evaluated by the exhaustive search procedure. Although this technique employs a suboptimum search procedure, it has the following advantages. First, by considering each measurement individually at the start of the search, the evaluation procedures should have their greatest accuracy due to high ratio of number of design samples to dimension of measurement subset. Second, by considering which measurement is best when adjoined to the previously selected subset, the procedure is responsive to dependencies between measurements as a function of sample classification. In addition to the final subset selection, all subsets within the constraints of the search of d or less measurements have been evaluated, which is essential to the ranking of these subsets and to plotting the performance as a function of measurement subset size. This permits an analyst to make tradeoffs between system complexity and performance. The measurement subset evaluation of subsets of size two, three and four measurements should be very informative in choosing coordinate vector projections for interactive computer graphics classification techniques.

Additional computational reduction can be realized by storing the $N(N-1)/2$ distances between design samples in the previously selected subspace. Suppose that the Euclidean metric is used to measure the distance between samples in the first $d-1$ coordinates, where

$$\rho_{d-1}(x, y) = \left[\sum_{i=1}^{d-1} (x_i - y_i)^2\right]^{1/2}$$

is the distance between x and y. Since the square of the Euclidean distance may be used to rank the distance between samples, the squared distance between x and y in the d-dimensional space is given by

$$[\rho_d(x, y)]^2 = [\rho_{d-1}(x, y)]^2 + (x_d - y_d)^2$$

where the squared distance $[\rho_{d-1}(x, y)]^2$ is stored from the previously selected subset. A similar procedure can be used for the Manhattan metric. These computation shortcuts in combination with those mentioned previously permit accurate and efficient measurement selection using the $k-NN$ rule and the "leave-one-out" method for problems with a modest size design sample set and a modest number of measurements. In problems involving a large number of potentially useful measurements, the subset search procedure may be modified such that only a reasonable number of the best individual measurements are retained after the individual measurements are evaluated.

A Measurement Selection Example

A simple example is given to illustrate some of the concepts of the proposed measurement selection procedure. Consider a four class problem involving a total of ten measurements. Let each mean of the four classes in the measurement space be a vertex of a regular four point simplex in a three-dimensional subspace with the center of the simplex at the origin such that $\|u_i - u_j\| = \sqrt{2}$, $i \neq j$. The three coordinates of the means in this subspace were embedded in the ten-dimensional space as the third, sixth, and ninth coordinates, respectively, Then, independent, identically distributed samples of a zero-mean ($\sigma = 0.5$) normal random variable were added to each of the ten coordinates. Using this construction procedure, 25 design

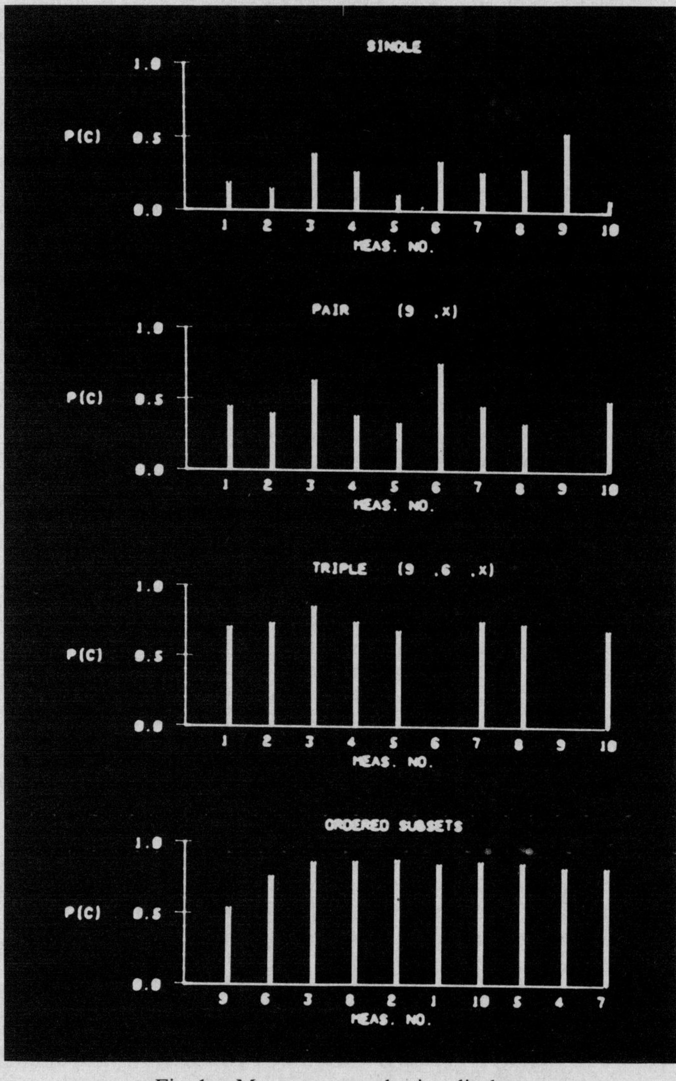

Fig. 1. Measurement selection display.

samples were generated for each of the 4 classes. The resulting set of 100 vectors were used as the design sample set.

Applying the nonparametric measurement selection procedure ($k = 1$) to the above design sample set, a measurement selection display was generated and is illustrated in Fig. 1. This nonparametric measurement selection procedure is to be included in the interactive computer graphics system for the design of signal classification systems [15].

In this measurement selection display, the height of the bar or the vertical line indicates the probability of correct classification associated with the measurement subset and the evaluation procedure. Note that the uppermost portion of this display indicates that the estimated best single measurement is the ninth measurement, with measurements number three and six being only slightly inferior. The second portion of the display indicates the estimated performance of each pair of possible measurements, including measurement number nine as one of the pair. Note that the pair (9, 3) and (9, 6) have the best performance, while all other pairs evaluated are approximately equal in performance to measurement nine alone. The third portion of the display indicates the estimated performance of all triples where the triple includes the pair (9, 6). The final portion of this display indicates the order in which the measurements were selected by the suboptimum subset search procedure and the resulting estimated performance. This part of the display also tells the analyst that the triple consisting of measurements 9, 6, and 3 contain most of the classification information. It should also be noted that measurements 8 and 2 slightly improve the performance while the remaining measurements degrade the classification performance. In actual classification applications, the plot of performance as a function of subset size might not attain its maximum value for such a small measurement subset and may decrease more rapidly after a certain subset size. This plot of optimum estimated performance as a function of the subset size for the subsets searched allows the analyst to perform tradeoffs between system complexity and performance of the pattern classification system.

Conclusions

A direct method of nonparametric measurement selection is proposed to determine the best subset of d measurements out of a set of D total measurements using a suboptimum measurement subset search procedure. An example is given to illustrate the measurement selection procedure.

References

[1] C. D. Allais, "The selection of measurements for prediction," Stanford Electron. Lab., Stanford, Calif., Nov. 1964, Rep. SEL-64-115, Paper AD-456 770.

[2] T. M. Cover, "Learning and pattern recognition," in *Methodologies of Pattern Recognition*, S. Watanabe, Ed. N. Y.: Academic Press, 1969, pp. 111–132.

[3] T. M. Cover and P. E Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-13, Jan. 1967, pp. 21–27.

[4] S. E. Estes, "Measurement selection for linear discriminants used in pattern classification," Ph.D. dissertation, Stanford Univ., Stanford, Calif., 1964; also IBM Corp., San Jose, Calif., Res. Rep. RJ-331, Apr. 1965.

[5] K. S. Fu, *Sequential Methods in Pattern Recognition and Machine Learning*. N. Y.: Academic Press, 1968.

[6] P. E. Hart, "An asymptotic analysis of the nearest neighbor decision rule," Stanford Electron. Lab., Stanford, Calif., May 1966, Rep. SEL-66-016, Paper AD-806 836.

[7] L. Kanal and B. Chandrasekaran, "On dimensionality on sample size in statistical pattern classification," in *1968 Proc. Nat. Electron. Conf.*, Chicago, Ill., pp. 2–7.

[8] P. A. Lachenbruch and M. R. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, Feb. 1968, pp. 1–11.

[9] P. M. Lewis, "The characteristic selection problem in recognition systems," *IRE Trans. Inform. Theory*, vol. IT-8, Feb. 1962, pp. 171–178.

[10] T. Marill and D. M. Green, "On the effectiveness of receptors in recognition systems," *IEEE Trans. Inform. Theory*, vol. IT-9, Jan. 1963, pp. 11–17.

[11] R. M. Miller, "Statistical prediction by discriminant analysis," *Meteorological Monographs*, vol. 4, Ch. 2 and Appendix A, Oct. 1962.

[12] E. A. Patrick and F. P. Fischer, II, "Nonparametric feature selection," *IEEE Trans. Inform. Theory*, vol. IT-15, Sept. 1969, pp. 577–584.

[13] D. L. Tebbe and S. J. Dwyer, III, "Uncertainty and the probability of error," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-14, May 1968, pp. 516–518.

[14] A. W. Whitney, "Performance and implementation of the k nearest neighbor decision rule with incorrectly identified training samples," Ph.D. dissertation, Univ. Missouri, Columbia, Mo., Jan. 1967.

[15] A. W. Whitney and W. E. Blasdell, "Study of computer graphics and signal classification applications," Rome Air Development Center, Griffiss AFB, Rome, N. Y., Contract F 30602-69-C-0227, RADC-TR-70-148, Final Rep.