Practical Applications of Sequential Pattern Recognition Techniques

R. Newman
School of Electrical Engineering, Purdue University, Lafayette, Indiana

B. Reisine
School of Electrical Engineering, Purdue University, Lafayette, Indiana

PRACTICAL APPLICATIONS OF SEQUENTIAL PATTERN RECOGNITION TECHNIQUES

R. Newman
School of Electrical Engineering
Purdue University
Lafayette, Indiana

B. Reisine
School of Electrical Engineering
Purdue University
Lafayette, Indiana

ABSTRACT

Experiments involving sequential recognition techniques and feature ordering schemes were performed on 23 feature samples of vowel spectra and 12 feature samples of remotely sensed agricultural crop data. Since each experiment dealt with two pattern classes, Wald's sequential probability ratio test was used. The test was implemented with both fixed and time-varying stopping boundaries. Feature ordering was accomplished by both dispersion analysis and the divergence criterion.

INTRODUCTION

A pattern recognition system consists of a feature extractor and a classifier (see Figure 1). The feature extractor makes measurements of salient characteristics of the input patterns. These are called feature measurements and based on them, the classifier assigns each input pattern to one of the possible pattern classes. We are concerned with those classifiers that are sequential in nature. That is, those that utilize the feature measurements one at a time in performing the classification. The advantages of sequential techniques are realized when the cost of taking feature measurements is high or the speed of classification is important.

TECHNIQUES

The problem of classifying patterns from two classes is formulated here as a statistical decision problem. N feature measurements, denoted by $X_1, X_2, \ldots, X_N$, are given for each pattern. The two pattern classes are called ω_1 and ω_2. For each pattern class ω_j, $j = 1,2$, it is assumed that the probability density function of this feature vector $X, p(X| \omega_j)$, is known. A discriminant function,

$$D_i(X) = \log P(\omega_i)p(X | \omega_i) \quad i = 1,2 \qquad (1)$$

is now defined which can easily be implemented by a Bayes classifyer. When $D_i(X) > D_j(X)$, $i,j = 1,2$, then X is said to be in class ω_i.

When $p(X | \omega_i)$ $i = 1,2$, is a multivariate Gaussian density function with mean vector M_i and covariance matrix K_i, i.e.,

$$p(X | \omega_i) = \left[(2\pi)^{N/2} |K_i|^{1/2}\right]^{-1} \exp \left[-\frac{1}{2} (X - M_i)^T K_i^{-1} (X - M_i)\right], \; i = 1,2 \qquad (2)$$

then the above discriminant function yields

$$D_i(X) = \log P(\omega_i) - \frac{1}{2} \log |K_i| - \frac{1}{2}(X - M_i)^T K_i^{-1}(X - M_i) \qquad (3)$$

This is the discriminant function used as the samples to be classified are assumed to be Gaussian in nature. In all recognition schemes used in this paper, the training procedure has been to compute M_i and K_i from the first 75 samples of each class.

For each sample to be classified, D_1 and D_2 were computed. If $D_1(X) - D_2(X)$ was positive, the sample was placed in class 1 and if negative in class 2.

In the above procedure it is necessary to use N measurements from each pattern to be classified. Quite often this is inconvenient (because of cost or time consumption) and it becomes desirable to have a scheme using less feature measurements. When there are only two pattern classes to be recognized, Wald's sequential probability ratio test (SPRT) can be applied. Here the feature measurements can be taken one at a time. At the nth stage of the sequential process, that is, after the nth feature measurement is taken, the classifier computes the sequential probability ratio

$$\lambda_n = \frac{p_n(X | \omega_1)}{p_n(X | \omega_2)} \qquad (4)$$

where $p(X | \omega_i)$, $i = 1,2$ is the (multivariate n-dimensional) conditional probability density function of X for pattern class ω_i. λ_n is then compared with two stopping boundaries A and B. If $\lambda_n \geq A$, then the decision is that X is in class ω_1, and if $\lambda_n \leq B$, then the decision is that X is in class ω_2. If $B < \lambda_n < A$ then an additional feature measurement will be taken and the process will proceed to the (n+1)st stage (see Figure 2). The two stopping boundaries are related to the error (misrecognition) probabilities by the following expressions

$$A = \frac{1 - e_{21}}{e_{12}} \quad \text{and} \quad B = \frac{e_{21}}{1 - e_{12}} \tag{5}$$

where e_{ij} is the probability of deciding X is in class ω_1, when actually X is in class ω_2 is true, $i,j, = 1,2$.

The above sequential probability ratio test can be generalized to the case where the stopping boundaries become time varying instead of remaining constant. Let $g_1(n)$ and $g_2(n)$ be either constants or monotonically nonincreasing and nondecreasing functions of n, respectively. The classifier continuously takes measurements as long as the sequential probability ratio λ_n lies between $e^{g_1(n)}$ and $e^{g_2(n)}$, that is, the sequential process continues by taking additional feature measurements as long as

$$e^{g_2(n)} < \lambda_n < e^{g_1(n)}, \quad n = 1,2,\ldots \tag{6}$$

If $\lambda_n \geq e^{g_1(n)}$, then the decision is that X is in class ω_1 and if $\lambda_n \leq e^{g_2(n)}$, then the decision is that X is in class ω_2. If $g_1(n)$ and $g_2(n)$ are constants it is easily seen that the standard Wald's SPRT can be considered as a special case of this modified SPRT. The fact that, in general, $g_1(n)$ and $g_2(n)$ can be made functions of n enables the sequential classifier to be designed such that the expected number of features measurements in reaching a terminal decision and the probability of misrecognition may be controlled in advance.

Here, the following two functions have been used:

$$g_1(n) = a'\left(1 - \frac{n}{N}\right)^r$$

$$g_2(n) = -a'\left(1 - \frac{n}{N}\right)^r \tag{7}$$

where $0 < r \leq 1$, $a' > 0$, and N is the prespecified number of feature measurements where the truncation occurs and the classifier is forced to reach a terminal decision (see Figure 3). The value of a' was obtained from the relation

$$e'_{12} = e^{-a'}\left[1 + \frac{ra'E'_1(n)}{N}\right] \tag{8}$$

where $E'_1(n)$ is the expected member of features to be used for the modified SPRT when X is in class 1. $E'(n)$ was obtained by using the average number of features found in the truncated case.

DIVERGENCE CRITERION FOR FEATURE ORDERING

The divergence between two pattern classes, whose samples are assumed to be Multivariate Gaussian Distributed, with equal covariance matrices, can be expressed as [1]:

$$J(\omega_1,\omega_2) = (M_1 - M_2)^T K^{-1}(M_1 - M_2) \tag{9}$$

M_1 and M_2 are the mean vectors and K^{-1} is the inverse of the covariance matrix. In the present study, the covariance matrices for the two classes were unequal, so the average was used.

The above relationship was used to find optimal feature subsets of 1,2,...,n-1 features from feature sets of 2,3,...,n features respectively, in the following way (see Figure 4):

1. The divergence between classes, for all n possible n-1 feature subsets of the original feature set, was computed.

2. The subsets corresponding to the largest value of divergence was chosen as optimal and the feature, which was deleted from the original set to form this subset, was placed at the end of the list of ordered features.

3. The divergence between classes, for all n-1 possible n-2 feature subsets of the new feature set, was computed.

4. The subset corresponding to the largest value of divergence was chosen as optimal and the feature, which was deleted to form this subset, was placed next to last on the ordered feature list.

5. This procedure was continued until finally the best subset consisting of one feature was selected and the ordered feature list was completed.

It should be noted that this technique is restrictive in that any feature discarded at a given level cannot be a member of a smaller optimal subset. To overcome this limitation, the divergence between classes would have to be computed for ${}_nC_{n-1}$ possible subsets to find optimal subsets of 1,2,...,n-1 features.

DISPERSION ANALYSIS FOR FEATURE ORDERING [3]

Vowel data was obtained (see Acknowledgement) in which the features had been ordered via a linear transformation. The technique utilized can be described as follows:

1. A sample covariance matrix was computed as an approximation to the true covariance matrix of the data.

2. The eigenvalues and corresponding eigenvectors of this matrix were determined and the eigenvectors were subsequently normalized and arranged according to the descending order of their associated eigenvalues. The resulting set of vectors constituted a generalized Kahunen-Loeve coordinate system.

3. Each input feature vector was transformed by this coordinate system and the first fifteen components were retained.

EXPERIMENTAL RESULTS AND DISCUSSION

The data employed in the described recognition schemes consisted of vowel samples and remotely sensed agricultural crop data.

Vowel samples were obtained by band-pass filtering recorded utterances of the form /hə CVC/, of each vowel, in each of the 23 consonantal environments, at 35 frequencies covering the range of the speech spectrum (250 to 10,000 hz). Filter outputs were rectified, smoothed, sampled and subsequently passed through an analog-to-digital converter. The data for each vowel in its final form consisted of a deck of IBM cards, each one containing the amplitude, in decimal form, of each filter output during a particular sample interval. In all instances of the use of untransformed vowel data, 23 features were selected, corresponding to the first 23 filter outputs, These features were chosen because they cover the range of possible values for the 1st and 2nd formant frequencies of the vowels under consideration (see Table 4).

Agricultural crop data was obtained by the detection and recording of both reflected and emitted electromagnetic radiation energy, from specific earth surface areas. Spatially scanning radiometers were used to obtain relative measurements of energy from the ground in 12 discrete spectral bands. The first ten bands encompassed visible wave lengths and the last two covered the reflective infrared portion of the spectrum. The data, first recorded on a 12-channel magnetic tape, underwent analog-to-digital conversion, and was then formatted and recorded on a data storage tape. IBM cards were punched from this tape such that each crop was characterized by a deck of cards, each one containing the relative energy in each spectral band, for a particular region of ground surface.

It can be seen (Table 1) that for the data under consideration, high recognition accuracies are achieved with the Bayes fixed sample size classifier. However, the Bayes classifier necessitates the use of all feature measurements, which is quite inefficient.

Table 2 shows the results of applying a sequential recognition technique and feature ordering. In the case of ɛ and æ the number of features required for classification has decreased 80.4% from the fixed sample size classifier without ordering and 83.6% with ordering by divergence while recognition accuracy has dropped no more than 1.2%.

The results for /ɑ/ and /ɔ/ in Table 2 are a conclusive demonstration of the value of feature ordering. In the unordered sequential case, the number of features used is 81% less than in the fixed sample size case. However, along with this is a decrease in accuracy of 27.3%. This startling decrease is understandable if one considers the rank of the various features (filter outputs) after ordering. The 1st, 2nd, 3rd, and 4th features were placed 4th, 9th, 20th, and 23rd. Thus the first four features on which the classifier was trained and based most decisions were found to be poor and/or misleading. The classification accuracy using ordered features is seen to be comparable to that of the fixed sample case, while the average number of features is down 95.6% for the divergence criterion and 90.1% for dispersion analysis.

In the case of the agricultural crop data (Table 2) there is again a substantial reduction in the number of features without much loss of accuracy. This result also indicates that the use of ordering in conjunction with sequential techniques is more powerful than the use of sequential techniques alone.

Through considerations of Table 3 and Figure 3, it is seen that the use of time-varying stopping boundaries causes a trade-off between the accuracy of recognition and the average number of features used to reach a decision. For values close to one, the average number of features may be reduced with only a slight drop in accuracy.

CONCLUSION

In reviewing the experimental results obtained in this study, it is reasonable to conclude that the use of sequential recognition techniques and feature ordering schemes are necessary for the efficient utilization of feature information.

Use of the time-varying stopping boundaries in equation (7) is not optimal. There is a need for further research to devise selection procedures for a', r, and even the functional form of the boundaries.

ACKNOWLEDGEMENT

We wish to thank both the Speech Research Group and the Laboratory of Agricultural Remote Sensing at Purdue, for the vowel data and crop data respectively.

We would also like to thank Dr. K. S. Fu for supplying the foundation from which this work developed.

This work was supported by NSF under Grant No. GK-1970.

REFERENCES

(1) T. Marill and D. M. Green, "On the Effectiveness of Receptors in Recognition Systems," IEEE Trans. Inform. Theory, 9, 1963.

(2) G. W. Hughes and J. F. Hemdal, "Speech Analysis," Purdue University Tech. Report, TR-EE 55-9, 1965.

(3) G. W. Hughes, A. S. House, and K. P. Li, "Research on Word Spotting," Clearinghouse, Dept. of Commerce, pp. 20-34, 1969.

(4) K. S. Fu, D. A. Landgrebe, and T. L. Phillips, "Information Processing of Remotely Sensed Agricultural Data," Proceedings IEEE, Vol. 57, April 1969, pp. 639-653.

BIBLIOGRAPHY

(1) K. S. Fu, Sequential Methods in Pattern Recognition and Machine Learning, Academic Press, 1968.

(2) J. L. Flanagan, Speech Analysis, Synthesis and Perception, Springer-Verlag, 1965.

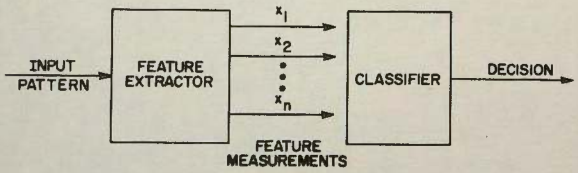

FIGURE 1. A PATTERN RECOGNITION SYSTEM.

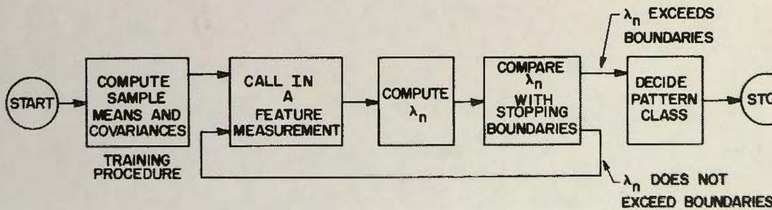

FIGURE 2. FLOW DIAGRAM OF A SEQUENTIAL PATTERN RECOGNITION SYSTEM.

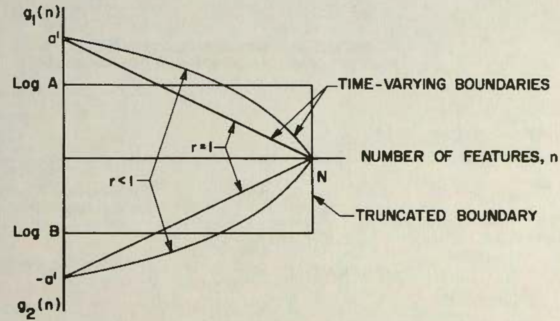

FIGURE 3. DIAGRAM OF TRUNCATED AND TIME-VARYING STOPPING BOUNDARIES.

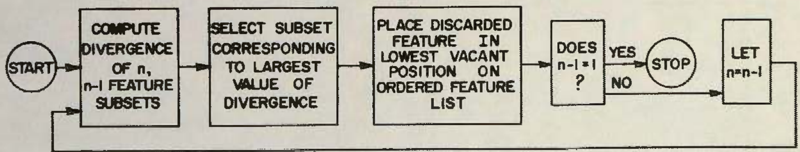

FIGURE 4. FLOW DIAGRAM OF FEATURE ORDERING BY THE DIVERGENCE CRITERION.

TABLE 1.
BAYES FIXED SAMPLE SIZE RESULTS

CLASSES	NUMBER OF SAMPLES	% ACCURACY	NUMBER OF FEATURES USED
|ɛ|,|æ|	171	96.5	23
|ɑ|,|ɔ|	166	99.0	23
CORN SOY BEANS	370	87.5	12

TABLE 2.
SEQUENTIAL RESULTS WITH TRUNCATED STOPPING BOUNDARIES

CLASSES	FEATURE ORDERING	NUMBER OF SAMPLES	% ACCURACY	AVERAGE NUMBER OF FEATURES
|ɛ|,|æ|	UNORDERED	171	97.0	4.59
|ɛ|,|æ|	DIVERGENCE CRITERION	171	95.3	3.78
|ɑ|,|ɔ|	UNORDERED	166	71.7	4.32
|ɑ|,|ɔ|	DIVERGENCE CRITERION	166	98.8	1.00
|ɑ|,|ɔ|	DISPERSION ANALYSIS	132	98.5	2.27
CORN SOY BEANS	UNORDERED	370	90.0	10.313
CORN SOY BEANS	DIVERGENCE CRITERION	370	85.1	3.43

TABLE 3.

SEQUENTIAL RESULTS WITH TIME VARYING STOPPING BOUNDARIES

r	CLASSES	FEATURE ORDERING	NUMBER OF SAMPLES	% ACCURACY	AVERAGE NUMBER OF FEATURES
1.00	|ɑ|, |ɔ|	UNORDERED	166	71.7	4.20
0.50	|ɑ|, |ɔ|	UNORDERED	166	72.9	4.36
0.25	|ɑ|, |ɔ|	UNORDERED	166	72.9	4.43
1.00	|ɑ|, |ɔ|	DISPERSION ANAL.	132	96.2	1.66
0.50	|ɑ|, |ɔ|	DISPERSION ANAL.	132	96.9	1.71
0.25	|ɑ|, |ɔ|	DISPERSION ANAL.	132	96.9	1.72
1.00	CORN SOY BEANS	UNORDERED	370	79.2	8.98
0.50	CORN SOY BEANS	UNORDERED	370	88.4	10.28
0.25	CORN SOY BEANS	UNORDERED	370	89.7	11.00

TABLE 4.

FEATURES SELECTED[2]

FILTER NUMBER	CENTER FREQUENCY OF FILTER	FILTER NUMBER	CENTER FREQUENCY OF FILTER
1	286	13	1070
2	317	14	1157
3	368	15	1290
4	428	16	1425
5	473	17	1560
6	526	18	1713
7	585	19	1901
8	643	20	2087
9	707	21	2316
10	780	22	2550
11	864	23	2814
12	966		