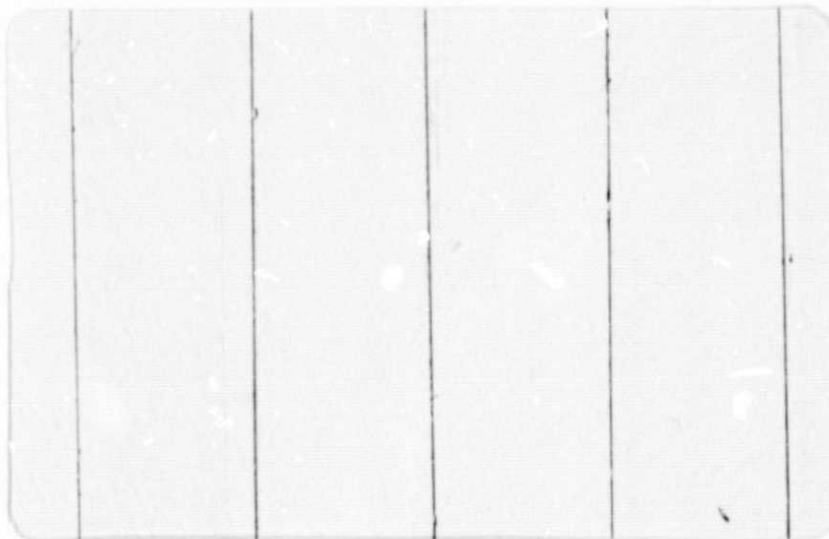# General Disclaimer

One or more of the Following Statements may affect this Document

- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.

- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.

- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.

- This document is paginated as submitted by the original source.

- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

Produced by the NASA Center for Aerospace Information (CASI)

ICSA

INSTITUTE FOR COMPUTER SERVICES AND APPLICATIONS

RICE UNIVERSITY

Classification Improvement by
Optimal Dimensionality Reduction
When Training Sets
Are of Small Size

by

S. A. Starks*, R. J. P. de Figueiredo*[+]
and D. L. Van Rooy°

*Elec. Eng. Dept., [+]Math. Sciences Dept.,
°I. C. S. A.
Rice University

April, 1976

ABSTRACT:

When the sizes of the training sets are small, classification in a subspace of the original data space may give rise to a smaller probability of error than the classification in the data space itself. This is because the gain in the accuracy of estimation of the likelihood functions used in classification in the lower dimensional space (subspace) offsets the loss of information associated with dimensionality reduction (feature extraction). To test this conjecture, a computer simulation was performed. A number of pseudo-random training and data vectors were generated from two four-dimensional Gaussian classes. An algorithm previously described (ICSA Technical Report #275-025-022, EE Technical Report #7520) was used to create an optimal one-dimensional feature space on which to project the data. When the sizes of the training sets were small, classification of the data in the optimal one-dimensional space was found to yield lower error rates than the one in the original four-dimensional space. Specifically, depending on the sizes of the training sets, the improvement ranged from 11% to 1%.

I. Introduction:

In real pattern recognition systems, the situation often arises that the classifier as well as the feature extractor must be designed with a limited number of training samples. As a result, in certain cases, the estimates of the class conditional statistics which are used to determine the classification strategy are poor.

When Gaussian statistics are assumed and the dimension of the raw data is n, then the n elements of the mean vector $\bar{x}^j$ for class H^j as well as the $\left(n \times (n+1)\right)/2$ independent elements of the covariance matrix $\tilde{R}^j$ for class H^j are estimated using the formulas:

$$\bar{x}^j = \frac{1}{N_j} \sum_{x_i \,\epsilon\, S_j} x_i \,, \tag{1}$$

$$\tilde{R}^j = \frac{1}{N_j - 1} \sum_{x_i \,\epsilon\, S_j} (x_i - \bar{x}^j)(x_i - \bar{x}^j) \,, \tag{2}$$

where S_j is the training set representing the class H^j and N_j is the total number of training vectors x_i in S_j.

It is well known that the uncertainties of these estimates decrease monotonically with increasing N_j and decreasing n [2]. The number of training samples necessary to obtain a non-singular estimate for the covariance matrix must be greater than or equal to $n + 1$. However, in order to obtain a really good estimate of the covariance as well as the mean often several times this number of training samples are needed [2].

When the ratio N_j / n $(j = 1, \ldots, M)$ (where M = total number of classes), tends to infinity, classification results obtained using all available features are superior to those results obtained using any transformation of the original space into a lower dimensional space. However, when Gaussian pattern classes are present and the ratio N_j / n $(j = 1, \ldots, M)$ is small, the feature extraction method presented in [1] can be of great value. When these conditions are satisfied, one can sometimes obtain better classification performance by using the optimal single linear Gaussian feature than by using all n features. This is so because when the dimensionality of the data is reduced to unity, the estimate of the mean in the reduced space is a one-dimensional estimate rather than an n-dimensional estimate. Similarly, the estimate of the class conditional covariance is merely the one-dimensional variance estimate rather than the $n \times n$ dimensional covariance matrix given by (2). Essentially, then the ratio N_j / m (where m is the dimension of the space in which classification is made) is increased with the reduction of dimensionality from n to $m = 1$. Hence the uncertainties in the mean and covariance estimates are reduced. This gain in accuracy in estimation may offset the loss of information resulting from the dimensionality reduction. Thus, in certain cases, results from classification obtained using our optimal single linear Gaussian feature can give rise to a lower probability of error than those obtained using all available features. Numerical results from the computer simulation described in the following section do indeed attest to this fact.

II. Numerical Results:

To verify the preceding argument, the following test procedure was conducted. A number of pseudo-random data vectors from two four-dimensional Gaussian classes were generated. N of these samples from each class were used to compose a training set from which the class conditional statistics given by (1) and (2) were obtained. Using these estimates, the optimal single linear Gaussian feature was found. The remainder of the pseudo-random data vectors were transformed using the optimal single linear Gaussian feature and were classified in the reduced space. Classification on these same samples were also performed in the untransformed space. The classification performances, which is the ratio of the number of samples classified properly to the total number of classifications made, were computed for each method and are listed in Table I. These results clearly show that one can improve classification using the optimal single linear Gaussian feature for small values of N/n. At higher values of N/n, one may even obtain comparable classification performance.

Number of Training Samples (N)	Classification Performance	
	Optimal Single Linear Gaussian Feature	All 4 Available Features
5	.590	.485
10	.610	.600
20	.610	.600
30	.605	.630
40	.590	.630
50	.610	.640

Table I Classification Performance for
Varying Sizes of Training Sets

III. Conclusions :

In the test case presented, it is readily noted that for low values of N / n, the classification performance obtained using the optimal single linear Gaussian feature exceeds that obtained using all available features. Similar results were found by classifying using subsets of all available features in [3]. Thus for low values of N / n, one realizes certain advantages from using this approach. First, a reduction in computer storage and mathematical computation is achieved. More importantly, one may improve the performance of the classifier.

The effect of having a small number of vectors in the training set on other algorithms ought to be explored.

REFERENCES

[1] 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", Rice University
Institute for Computer Services and Applications Technical
Report #275-025-022 (EE Technical Report #7520), November,
1975.

[2] Duda, R.O. and Hart, P.E., Pattern Classification and Scene
Analysis, Wiley, N.Y., 1973, pp. 67-68.

[3] Fu, K.S., Landgrebe, D.A., and Phillips, T.L., "Information
Processing of Remotely Sensed Agricultural Data," Proc. IEEE,
57, April, 1969, pp. 639-653.