

## **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.

79-10025

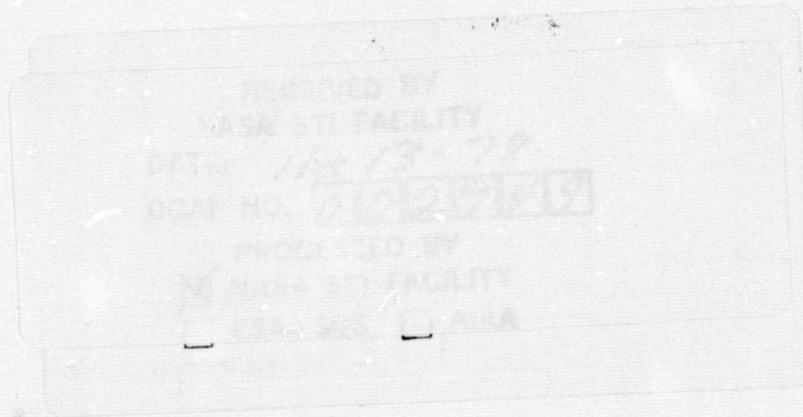
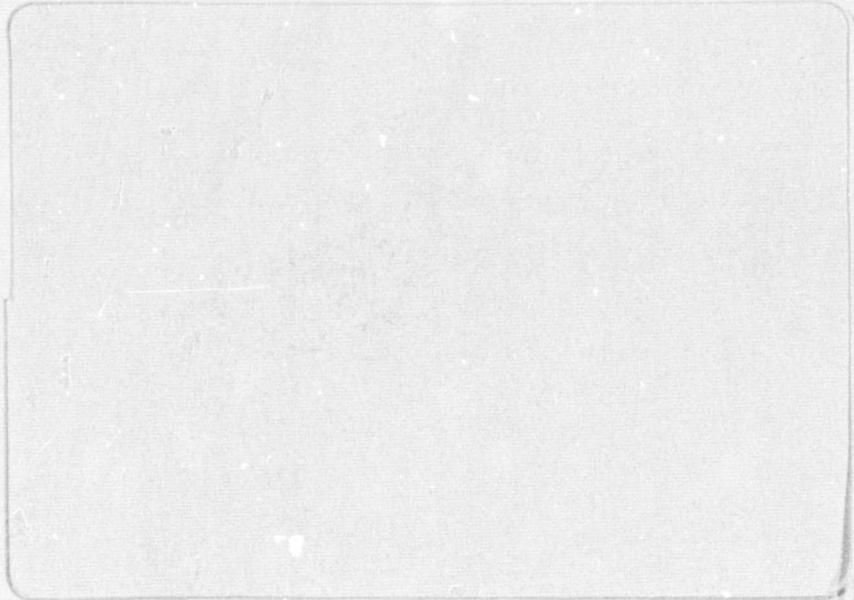
OR-157914

(E79-10025) COMPARISON OF FEATURE SELECTION  
TECHNIQUES FOR EARTH RESOURCES DATA  
(Instituto de Pesquisas Espaciais, Sao Jose)  
13 p HC A02/MF A01 CSCL 05B

N79-13436

Unclas  
00025

G3/43



CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO

**INSTITUTO DE PESQUISAS ESPACIAIS**



INDEX

	<u>Page</u>
INTRODUCTION .....	1
FEATURE SELECTION TECHNIQUES .....	2
COMPARISON OF FEATURE SELECTION TECHNIQUES .....	6
REFERENCES .....	9

LIST OF TABLES

Table I - Results of Nonparametric feature selection technique.

Table II - Comparison of Feature Selection Techniques

R. Kumar\*

INTRODUCTION

One of the problems commonly encountered in pattern recognition is the selection of effective features from a given set of measurements. The use of a large number of feature measurements increases the complexity of the size of and the computer time required by the classifier (Swain, 1972). For example, in remote sensing of earth resources and environment, the problem reduces down to the following: Given a set of  $N$  features (e.g. multispectral scanner channels), find a subset consisting of  $n$  channels which provides an optimal trade-off between classification cost and classification accuracy (Fu, 1970). For example, the SKYLAB multispectral scanner (S192) has 13 channels and generally an analyst wants to use the best four or five of these channels for classification.

The effectiveness of the features should be determined by performance of the recognition system, usually in terms of probability of correct recognition. Ideally, one would like to solve this problem by computing the probability of misclassification associated with each  $n$ -feature subset and then selecting the one giving best performance (Swain, 1972). However, it is generally not feasible to perform the required computations. Even when one assumes normal distribution, numerical integration is required which, in the multidimensional case, is impractical to carry out. Some of the techniques of feature selection are summarized below.

---

\* *Instituto de Pesquisas Espaciais (INPE) - Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - 12.200 - São José dos Campos - SP - Brazil.*

FEATURE SELECTION TECHNIQUES

Fu (1970) has used a non-parametric feature selection technique based on the direct estimation of error probability. The proposed feature selection criterion was based on the direct estimation of samples. Maximum likelihood decision rule (MLDR) was used for classification. He pointed out that a large amount of computation time is required especially when the number of classes is large. Using 7530 test samples, he applied the proposed nonparametric method of feature selection to crop classification. The results of his experiment are given in Table I. He found that all the classes are separable for most 4 feature subsets (41 sets).

TABLE I

RESULTS OF NONPARAMETRIC FEATURE SELECTION TECHNIQUE

NONPARAMETRIC METHOD			PARAMETRIC METHOD	
NUMBER OF FEATURES	BEST FEATURE SET	PERCENT ERROR	BEST FEATURE SET	PERCENT ERROR
1	$x_9$	33.8	$x_9$	37.6
2	$x_1, x_9$	3.1	$x_1, x_9$	10.6
3	$x_1, x_{10}, x_{11}$	0.1	$x_1, x_{10}, x_{11}$	5.0
4	$x_1, x_9, x_{12}$	0.0	$x_1, x_6, x_{10}, x_{11}$	4.9
	41-feature set			

Many authors have studied the linear feature-space transformation techniques to apply for the feature selection problems. For example, Watanabe (1966) introduced the feature-space compression technique based on the Karhunen-Loeve (K-L) expansion. Fu (1971) tested the feature selection technique based on generalized K-L expansion on crop classification. The results were compared with those using the parametric feature selection technique. The MLDR was used for the classifier and the appropriate statistical parameters were estimated from training samples for each class. He found that the transformed p-

dimensional feature space was less effective than the same dimensional feature subspace for all  $p$  ( $\leq N =$  total number of available features), but the difference in performance for the 4-feature subset was only 1.3 percent. The computation time required, on the other hand, was much shorter for the transformation technique.

An intermediate quantity which is related to the classification accuracy is often used as a basis for feature selection (Fu, 1971). Divergence between pattern classes has been proposed as a criteria for feature selection.

Divergence is defined for any two density functions. In the case of normal variables with unequal covariance matrices, it can be shown (Kailath, 1967) that

$$D(i,j|c_1, c_2, \dots, c_n) = \frac{1}{2} \text{tr} \left[ (\Sigma_i - \Sigma_j) (\Sigma_j^{-1} - \Sigma_i^{-1}) \right] + \frac{1}{2} \text{tr} \left[ (\Sigma_i^{-1} + \Sigma_j^{-1}) (U_i - U_j) (U_i - U_j)^T \right] \quad (1)$$

It can be shown also that the probability of misclassification is a monotonically decreasing function of divergence. Therefore, features selected according to the magnitude of divergence will imply their corresponding discriminatory power between the classes  $i$  and  $j$ . In other words, feature set  $\alpha_p$  is considered more effective than the feature set  $\alpha_\ell$  if  $D(i,j|\alpha_p) > D(i,j|\alpha_\ell)$  (Fu, 1970). Divergence is a distance measure between the two statistical distributions. It is an indirect measure of the ability of the classifier to successfully discriminate between them.

Fu (1970) assumed that feature vectors for each class were gaussianly distributed. He used the linear classification procedure based on the maximum likelihood decision rule (MLDR) for multiclass classification problem by means of minimizing the maximum probability of overall misclassification (minimax procedure, Anderson and Bahadue, 1962).

He showed that a monotonic functional relationship exists between the probability of pairwise misclassification between the classes and the separability measure. In addition, he showed that in the case of Gaussianly distributed pattern classes with equal covariance matrices, the divergence and the separability measure have a monotonic relationship. Nevertheless, it is clear that the separability measure is a more general criterion for feature effectiveness. He tested the effectiveness of the feature sets by computing the percentage of misclassification with 7530 test samples (approximately 1500 samples per class) classified by MLDR classifier and then selected the optimum feature sets from all possible combinations. He found, from experimental results, that it is possible for smaller size feature subsets to be almost as effective as the complete feature set. Thus, in many situations, selecting optimum feature subset considerably reduces the computer time required for classification, as compared to using the entire feature set, with a relatively small loss of classification accuracy.

Although divergence only provides a measure of the distance between two class densities, its use is extended to the multiclass case by taking the average over all class pairs (Fu, 1971).

If  $D_{ij}$  is the divergence between classes  $i$  and  $j$ , then the multiclass feature selection criterion is

$$D_{AVG} = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^m D_{ij} \quad (2)$$

Another strategy is to maximize the minimum pairwise divergence (Grettenberg, 1963; Fu and Chen, 1969; Kadota and Shepp, 1967; Swain, 1972) i.e., to select the feature combination which does the best job of separating the "hardest to separate" pair of classes, i.e., for example, consider a situation where there are 3 classes A, B and C

$$D_{MIN} = \text{Min} \{D_{AB}, D_{BC}, D_{CA}\} \quad (3)$$

Where  $D_{AB}$  = divergence between class A and class B

The relationship between the divergence and classification accuracy is highly nonlinear (in fact, divergence increases without bound as the class separability increases, whereas probability of correct classification must saturate at 100 percent), and it is found that widely separable classes make too much of a contribution to  $D_{AVG}$  as compared with less separable classes. As a result, in problems involving a wide range of class separabilities,  $D_{AVG}$  is not a reliable criterion for feature selection.

On the other hand,  $D_{MIN}$  is based on selecting the channels which do the best job of separating the hardest-to-separate pair of classes. Although this is certainly a reasonable strategy in many 'remote sensing of earth resources' problems, there is no guaranty that it is the optimal one.

As pointed out before, as the separability of a pair of classes increases, the pairwise divergence also increases without limit but the probability of correct classification saturates at 100 percent. A modified form of divergence, referred to as the "transformed divergence",  $D_T$ , has a behavior more like the probability of correct classification than divergence (Swain and King, 1973).

$$D_T = 2 \left[ 1 - \exp(-D/8) \right] \quad (4)$$

where  $D$  is the divergence discussed above. The saturating behavior of this function reduces the effects of widely separated classes when taking the average over all pairwise separations.  $D_{AVG}$  based on transformed divergence has been found a much more reliable criterion for feature selection than  $D_{AVG}$  based on "ordinary" divergence.

Swain et al. (1971) have shown experimentally that a separability measure referred to as the B-distance, based on Bhattacharyya's coefficient, provides a much more reliable criterion than divergence,

presumably because as a function of class separability, it behaves more like the probability of correct classification. For two densities  $p_1(x)$  and  $p_2(x)$ , the B-distance is given by

$$B = \int_x \left[ \sqrt{p_1(x)} - \sqrt{p_2(x)} \right]^2 dx \quad (5)$$

Swain and King (1973) performed an experiment to compare the separability measures divergence, transformed divergence and B-distance. Based on typical second order statistics derived from real remote sensing data, 2790 sets of Gaussianly distributed artificial data were generated: each set contained 1000 observations for each of two pattern classes in a feature space of dimensionality ranging from 1 to 6 (465 sets were generated for each dimension 1, 2, ... 6). For each set, the divergence, transformed divergence and B-distance were computed, and the actual classification error for the 2000 observations was taken as the associated probability of error. They found that both transformed divergence and B-distance are much better measures for feature selection than divergence. In addition, B-distance was found to be a slightly better measure of feature selection as compared to the transformed divergence.

#### COMPARISON OF FEATURE SELECTION TECHNIQUES

For our comparative study, aircraft multispectral scanner data (MSS) over six selected flightlines were analysed in subsets of one to twelve spectral channels covering the visible, near infrared, middle infrared and thermal infrared wavelength regions. The data of these flightlines were of good quality and free from problems such as lack of sufficient ground observations, excessive cloud cover, excessive sun angle effects etc. Black and white aerial photography and gray scale print-outs of the flightlines in the spectral channels were used to aid in locating the boundaries of the agricultural fields. Sufficient number of fields of each agricultural cover were selected carefully so that they could be assumed to be representative of the flightline.

Transformed divergence, defined in eq. (4), was used throughout this study.

Let  $D_{TAVG}^n$  and  $D_{TMIN}^n$  denote the average transformed divergence and the minimum transformed divergence, computed over all possible pairs of classes (each agricultural cover was treated as a separate class). Assuming a multivariate gaussian distribution for each class, the feature selection algorithm was used to select the best combinations of one to eleven spectral channels out of the twelve available spectral channels, using each of the following criteria of feature selection based on the values of  $D_{TAVG}^n$ :

1. Select the best subset of  $n$  ( $n=1$  to  $11$ ) spectral channels as being the one that maximizes  $D_{TAVG}^n$ , by exhaustive search of all possible combinations of  $n$  spectral channels out of the 12 available channels.
2. Select the best subset of  $n$  ( $n=1$  to  $11$ ) spectral channels using "forward feature selection". In forward feature selection, the best individual channel is selected on the first round, and then the best pair including the best one channel is selected, etc.
3. Select the best subset of  $n$  ( $n=1$  to  $11$ ) spectral channels using "backward feature selection". This method is a counterpart to forward feature selection, consisting of a sequential rejection procedure, in which one finds the "best" set of features by finding a set of  $(N-1)$  features discarding the worst one, then choosing the best set of  $(N-2)$  among the preceding  $(N-1)$  selected features, etc..

From the values of the average transformed divergence  $D_{TAVG}^n$ , the probability of correct classification ( $P_C$ ) was estimated using the curve of Swain and King (1973). Table II compares the values of  $P_C$  obtained by exhaustive search, forward feature selection and backward feature selection. It shows that forward feature selection gives almost as good results as the exhaustive search. Data of more flightlines are being analysed to check these results.

Although comparisons of feature selection techniques have been done and reported by many authors in the past, the present analysis is the first, as far as the author knows, to be done systematically on a large quantity of good quality earth resources data, covering visible, near infrared, middle infrared and thermal infrared portions of the spectrum.

The author gratefully acknowledges: the Laboratory for Applications of Remote Sensing, Purdue University, for their permission to use the multispectral scanner data, obtained under the NASA Grant No. NGL 15-005-112; Dr. Celso de Renna e Souza for his continuous encouragement and assistance and Dr. Nelson de Jesus Parada, the Director of the Instituto de Pesquisas Espaciais (INPE) for his permission to publish this work.

TABLE II

COMPARISON OF FEATURE SELECTION TECHNIQUES

NUMBER OF CHANNELS IN THE SUBSET	$P_c$ : EXHAUSTIVE SEARCH	$P_c$ : FORWARD FEATURE SELECTION	$P_c$ : BACKWARD FEATURE SELECTION
1	84.84	84.84	84.84
2	90.16	90.16	87.71
3	92.59	92.28	90.39
4	94.38	94.11	91.38
5	95.35	95.35	92.87
6	95.93	95.88	93.12
7	96.26	96.26	93.85
8	96.54	96.49	94.12
9	96.73	96.68	95.47
10	96.86	96.86	96.48
11	96.92	96.92	96.74

NOTE:  $P_c$  denotes the probability of correct classification estimated from the values of average transformed divergence using the curve of Swain and King (1973).

REFERENCES

Anderson, T.W. and Bahadur, R.R. 1962. Anr.Math, Stat. 33: 420-431.

Fu, K.S. and Chen, C.H. 1969. TR-EE-65-5, School of Electrical Engineering, Purdue University, W. Lafayette, Indiana.

Fu, K.S. 1970. IEEE Trans. Syst., Sci. & Cibern. SSC-6: 33-39.

Fu, K.S. 1971. TR-EE-71-13, School of Electrical Engineering, Purdue University, W. Lafayette, Indiana, 91 p.

Grettenberg, T.L. 1963. IEEE Trans. Information Theory IT-9: 265-275.

Kadota, T.T. and Shepp, L.A. 1967. IEEE Trans. Information Theory IT 13: 278-284.

Kailath, T. 1967. IEEE Trans. Commun. Technol. vol. COM-15: 52-60.

Swain, P.H. 1972. LARS Information Note 111572, Laboratory for Applications of Remote Sensing, Purdue University, W. Lafayette, Indiana, 40 p.

Swain, P.H. and King, R.C. 1973, "Two Effective Feature Selection Criteria for Multispectral Remote Sensing", International Joint Conference on Pattern Recognition, Washington, D.C.

Swain, P.H., Robertson, T.V., and Wacker, A.G. 1971. LARS Information Note 020871, Laboratory for Applications of Remote Sensing, Purdue University, W. Lafayette, Indiana.

Watanabe, S. 1966, "A Method of Self-Featuring Information Compression in Pattern Recognition", Proc. Bionics Symp.