A NEW NON-PARAMETRIC FEATURE SELECTION CRITERION—'EFFECT'*

BELUR V. DASARATHY

(School of Automation, Indian Institute of Science, Bangalore-560012)

Received on March 27, 1973 and in Revised form on September 28, 1973

ABSTRACT

The problem of feature selection in the field of pattern classification of multi-dimensional feature vectors is essentially one of dimensionality reduction. A new non-parametric technique for choosing an optimum subset of features from the given feature set is proposed in this study. This technique is based on the concept of inter-class and intra-class distances and tests conducted reveal the efficacy of this new Effective Figure of Merit Criterion—'EFFECT'.

Key Words: Feature Selection, Pattern Recognition, Non-Parametric Methods, Estimated Probability of Misrecognition.

1. INTRODUCTION

Numerous parametric and non-parametric techniques of feature selection have been proposed in the Literature [1-4] from time to time. Many of these methods are based on the concept of inter-class and intra-class distances of the different pattern classes. Deuser [5] proposed one such technique called the Hybrid Multispectral Feature Selection Criterion. Tests conducted using this criterion on Iris data [6] brought out the unsuitability of the technique for implementation in multispectral feature selection problems of interest to remotely sensed earth resources data analysis and the like. Results of these preliminary tests on comparison with the results of actual classification of the test data using non-parametric recursive algorithms (such as the Ho–Kashyap Algorithm [7] and maximum likelihood classification methods [8]) revealed that the Hybrid Figure of Merit obtained for different feature subsets were highly disproportionate to the corresponding actual classification efficiencies obtained using the same feature subsets. The Hybrid Figure of Merit in addition had no lower and upper bounds

---

* Presented at the National Systems Conference held in June 1973, at the Indian Institute of Science, Bangalore.

on the actual values and this proves to be a drawback when a choice has to be made among the possible feature subsets. Such a comparison between feature subsets is most effective if the figure of merit is based on a normalized scale. This lack of normalized scale of merit together with the major conceptual shortcoming of not accounting for the inter-feature correlations arising within the feature subset brought out the need for a new criterion of feature selection. Investigations carried out towards this end has led to a new and more meaningful criterion for feature selection which takes into account the effect of inter-feature correlations. This is conceptually significant in that one of the basic aims of transformation techniques employed in feature selection problems is to obtain an orthogonal system of coordinates (which form the components of the feature vector) minimizing the cross correlation between features, *i.e.*, to choose such (transformation of) features as will diagonalize the covariance matrices of the training sample sets of the different pattern classes. The new figure of merit is bounded and rates the different feature subsets on a universal 0–1 scale, thus allowing a direct comparison amongst these different feature subsets. Results derived by application of this new criterion to Iris test data have verified the superiority of this criterion over the Hybrid Figure of Merit Criterion. Also, as a further evaluation of this new criterion, comparison relating to a parametric feature selection method (based on Gaussian assumptions) was made using the same test data. Results showed a fair agreement between the two approaches relative to the actual classification efficiencies, with the new criterion showing a far better correlation than the parametric method (which possibly could be attributed to the doubtful validity of the Gaussian assumptions). These are presented and discussed in the sequel in detail.

2. ANALYSIS

Let C_i $(i = 1, m)$ represent the 'm' pattern classes into which any given feature vector x, of maximum dimension 'n', is to be assigned. The purpose of this or any other feature selection criterion is to determine the optimal subset of r features $(r \leqslant n)$ out of the possible $(2^n - 1)$ feature subsets for the ensuing classification (decision) process. Let $x(j,i,p)$ represent the j-th feature $(j \leqslant r)$ of the p-th training sample of class C_i $[p \leqslant L(i)$ where $L(i)$ is the total number of training samples from class $C_i]$. To appreciate the physical significance of this non-parametric feature selection criterion, consider two pattern classes C_1 and C_2. Assuming that two training samples from each of these classes are available, one can visualize that the decision or classification process becomes more efficient with increase

in the inter-class (between samples of the two different classes) Euclidean distance in the corresponding feature space. On the other hand, an increase in the intra-class (between samples of the same class) Euclidean distances tends to decrease the efficiency of classification. Therefore, the optimum subset of features is one that maximizes the inter-class distance while minimizing the intra-class distance in the Euclidean feature space. An obvious candidate criterion is, hence, the difference of inter-class and intra-class distance. This non-parametric method (being based on no specific parametric assumption about the distributions) can be made more meaningful physically by attaching suitable weights to these distances to account for possible differences in the ease of measuring (and hence the cost) of certain features over some others or such other physical or problem dependent considerations. This is the basis for Deuser's criterion, which can be written as:

Hybrid Figure of Merit of a set of r features

$$H = \sum_{j=1}^{r} SUM(j)$$

$$= \sum_{j=1}^{r} [SUM1(j) - a(j)\, SUM2(j)]$$

Here

$$SUM1(j) = \sum_{i=2}^{m} \sum_{i_1=1}^{i-1} k_{ii_1} \sum_{p=1}^{L(i)} \sum_{p_1=1}^{L(i_1)} [x(j, i, p) - x(j, i_1, p_1)]^2$$

$$SUM2(j) = \sum_{i=1}^{m} k_i \sum_{p=2}^{L} \sum_{p_1=1}^{p-1} [x(j, i, p) - x(j, i, p_1)]^2$$

k_{ii_1}, k_i, $a(j)$ are the appropriate weights defined from physical constraints of the problem.

The particular subset of r features which maximises this criterion was considered as the best 'r' features. (A particularly interesting exercise in FORTRAN coding was called for to automatically consider all possible combinations). However, this does not allow for a direct comparison over different sized feature subsets (r_1, $r_2 \ldots n$), *i.e.*, over all the possible subsets of features. To overcome this and other deficiencies, attempts were made to define a new criterion and this is discussed in the sequel.

A careful analysis of Deuser's criterion reveals that no consideration is given to the effect of interfeature correlation, which increases the scatter of the classes in the feature space and hence is a significant factor in problems with highly correlated features. Also, no attempt at defining a bounded universal scale of merit is made in his analysis, thus lacking a necessary and basic ingredient of any figure of merit concept. To meet these conceptual omissions, the following new criterion is proposed.

3. Effective Figure of Merit Criterion

Let $J = \{J_i : i = 1, \ldots n\}$ be the set of features under consideration.

Let $Q(J)$ be the complement of $\mathcal{P}(J)$ with respect to Φ where $\mathcal{P}(J)$ is the power set (set of all subsets) of J and Φ is the null set.

We define B the best feature subset of J as

$$B = S \ni \underset{S \epsilon Q(J)}{E(S)} \text{ is maximum,}$$

where, $E(S)$: The Effective Figure of Merit of a feature subset

$S = \{s_j : j = 1, \ldots r\}$ $r = m(S)$ is given by

$$\mathrm{E}(\mathrm{S}) = [P(S)/(1 \cdot 0 + C(S)]^{\frac{1}{2}}$$

$$P(S) = \left[1 \cdot 0 - \prod_{j=1}^{r} (1 \cdot 0 - F(s_j))\right]$$

$$C(S) = SUM3(S) / \sum_{j=1}^{r} (SUM1(s_j) - SUM2(s_j)); \quad r \geqslant 2$$

$$= 0 \qquad ; \quad r = 1$$

$$F(s_j) = (SUM1(s_j) - SUM2(s_j))/SUM1(s_j).$$

Here,

$$SUM1(s_j) = \sum_{i_1=2}^{m} \sum_{i_2=1}^{i_1-1} K_{i_1 i_2} \sum_{p_1=1}^{L(i_1)} \sum_{p_2=1}^{L(i_2)} (x(s_j, i_1, p_1) - x(s_j, i_2, p_2))^2$$

$$SUM2(s_j) = \sum_{i=1}^{m} K_i \sum_{p_1=2}^{L(i)} \sum_{p_2=1}^{p_1-1} (x(s_j, i, p_1) - x(s_j, i, p_2))^2$$

$$SUM3(S) = \sum_{i=1}^{m} \sum_{p=1}^{L(i)} \sum_{j_1=2}^{r} \sum_{j_2=1}^{j_1-1} |d_{j_1} d_{j_2}|$$

$$d_j = x(s_j, i, p) - \left[\sum_{p=1}^{L(i)} x(s_j, i, p)/L(i)\right]$$

An estimate of the probability of misrecognition can be determined as $EPOMR(S) = (m - 1)\left(1 - E^2(S)\right)/m$.

This effective figure of merit unlike Deuser's Hybrid Figure of Merit, allows comparison over different sizes of subsets, and more proportionate comparison within the combinations possible of a given subset size. It can even represents a rational basis for comparing different feature selection problems. As a further evaluation of this criterion, comparison with a suitable parametric feature selection method was thought of. The existing parametric feature selection technique based on the assumption of Gaussian distributions and equal covariance matrices for the pattern classes, was considered here. The high classification efficiency obtained for the test data through maximum likelihood classification procedures based on Gaussian assumptions indicated that the Gaussian assumption may be justifiable to an extent. The statistical analysis of the Iris data showed that the sample covariance matrices for the two classes were unequal, but close enough to justify the use of the above parametric approach provided some suitable modifications to the method could be made to account for the difference between the covariance matrices of the two classes of Iris data. Two alternatives were considered: (i) Average the covariance matrices and computing the inverse of this averaged covariance matrix for further analysis, (ii) Computing the inverses of the covariance matrices of the two classes individually and averaging the inverse matrices for further analysis.

Let M_i, K_i; $(i = 1, 2)$ be the mean vectors and covariance matrices of the classes 1 and 2 respectively. Under Gaussian assumption, the divergence criterion, for two class problem with equal covariance matrices K, can be written as

$$D(C_1, C_2) = (M_1 - M_2)^T K^{-1} (M_1 - M_2)$$

and the features are selected on the basis of the magnitude of D for the different feature subsets. Here, the covariances matrices being unequal, the measure is not necessarily optimal. However, an equivalent inverse covariance matrix can be defined in view of the fact K_1 and K_2 are not very different in their values and towards this end two possibilities are considered:

(i) $K_e^{-1} = [(K_1 + K_2)/2]^{-1}$: inverse of the average,

(ii) $K_e^{-1} = (K_1^{-1} + K_2^{-1})/2$: average of the inverses.

is computed through both these approaches. The results of experiments along these directions with Iris test data is presented and discussed in the next section.

4. Discussion of Results of Experiments with Iris Data

The two class Iris data [6] consisting of 100 four-dimensional training samples was used as a base for this experiment designed to make a comparative study of the Hybrid and Effective Figure of Merit. The Hybrid and Effective Figures of Merit were computed using the Iris data for all the possible 15 feature combinations. To obtain a basis for comparative evaluation of these two figures of merit, the actual classification efficiencies under each of these 15 feature combinations were estimated using a new non-parametric method of feature classification. (This technique can be used either independently or as an extension to existing techniques such as the *Ho*–Kashyap algorithm. The classification results listed in Table I were obtained using the new method as an extension to *Ho*–Kashyap algorithm. It was found that the classification efficiencies, under this approach, turned out to be considerably higher than under either the *Ho*–Kashyap algorithm without extension or the maximum likelihood approach with the assumption of Gaussian distributions, for all feature combinations).

A careful perusal of Table I brings to light the relative superiority of the new Effective Figure of Merit over the Hybrid Figure of Merit. This point can be dramatically brought home by considering, for example, the cases of features 3 and 4. The Hybrid Figure of Merit for feature 3 is about 350 per cent of that for feature 4 thus indicative of a very high superiority of feature 3 over feature 4. But the fact, as can be seen from actual classification, is that not only the feature 3 is not superior to feature 4 by this order of magnitude, but actually is even slightly less efficient than feature 4. This indeed is very clearly broughtout by the new figure of merit. Similarly, comparison of the results among other cases of interest such as features 1 and 4 reinforce the superiority of the new figure of merit. The bounded (0–1) scale of merit of the new method, in addition to allowing good relative evaluation of different feature subsets, does also give beforehand a fair measure (a sort of lower bound) of the classification efficiencies that can be expected in the corresponding feature space. Indeed except for the case of feature subsets (1, 2, 4) and (1, 4) the order of merit as derived from the Effective Figure of Merit criterion is exactly the same as that derived by actual classification of the data and even there the measure of classification

TABLE I

A comparative study of the new Effective figure of merit with Deuser's hybrid figure of merit and modified (parametric) divergence criterion using Iris data

Features used	Effective figure of merit	Estimated probability of mis-recognition	Estimated classification on efficiency (%)	Actual observed classification efficiency (%)	Deuser's hybrid figure of merit	Divergence D (C_1, C_2) using parametric feature selection criteria	
						$K_e^{-1} = \bar{K}_e^{\,1}$	$K_e^{-1} = K^{-1}$
1, 2, 3, 4	0·9834	0·0165	98·35	98	6564·92	14·219	15·853
1, 3, 4	0·9833	0·0166	98·34	98	6460·88	17·207	19·207
2, 3, 4	0·9794	0·0204	97·96	98	5502·17	22·974	24·445
3, 4	0·9770	0·0227	97·73	97	5398·13	25·109	26·848
1, 2, 4	0·9466	0·0520	94·80	96	2391·78	18·262	29·925
1, 4	0·9399	0·0583	94·17	95	2287·74	21·496	34·128
1, 2, 3	0·9346	0·0633	93·67	97	5339·22	11·637	16·956
1, 3	0·9234	0·0737	92·63	96	5235·88	11·533	16·315
2, 4	0·9178	0·0788	92·12	95	1329·03	12·378	22·856
4	0·9020	0·0932	90·68	94	1224·99	14·759	26·107
2, 3	0·8961	0·0985	90·15	94	4277·17	23·012	29·307
3	0·8742	0·1179	88·21	93	4173·13	22·054	27·718
1, 2	0·7017	0·2538	74·62	75	1166·79	3·976	4·048
1	0·6267	0·3036	69·64	73	1062·75	3·976	4·259
2	0·4164	0·4133	58·67	63	104·04	0·712	0·740

efficiency derived from the criterion compares favourably with the actual classification efficiency. This is of value in comparing various feature selection and classification problems which are of a similar nature. Table II shows the results of the new feature selection technique as applied to Iris data, wherein the various feature subsets are ordered according to the values of the corresponding effective figure of merit. Depending on the ease (and hence the cost) of measurement of different sizes of feature subsets, the optimum size of feature subset can be chosen. Then, using Table II, the best feature subset within the chosen subset size may be selected.

TABLE II

Results of feature selection for Iris data using the effective figure of merit criterion

Feature subsets listed in the descending order of merit			
Subset dimension	1	2	3
Figures in parentheses indicate the best classification efficiencies obtained in the corresponding feature subspace	4 (94%)	3, 4 (97%)	1, 3, 4 (98%)
	3 (93%)	1, 4 (95%)	2, 3, 4 (98%)
		1, 3 (96%)	
	1 (73%)	2, 4 (95%)	1, 2, 4 (92%)
		2, 3 (94%)	
	2 (63%)	1, 2 (75%)	1, 2, 3 (97%)

Table 1, also gives the divergence values $D(C_1, C_2)$ computed using the psuedo equivalent inverse covariance matrices as detailed earlier. While the divergence values are relatively lower for subsets which have significantly lower actual classification efficiencies, comparison of divergence values between subsets, which are not too far apart in terms of their actual classifica-

tion efficiencies, does not reflect well on the reliability of this criterion. This may possibly be due to the doubtful validity of the Gaussian assumptions. Further the results in Table I show that averaging the covariance matrices before computing the inverse gives far better ordering of the different subsets. This is also more meaningful from physical and mathematical considerations.

One can also notice that agreement between the Effective Figure of Merit Criterion and the modified version of the Divergence Criterion is good for single features and fair for feature subsets of two dimensions. However, as the size of the feature subset increases, the agreement becomes poorer. This probably is due to the fact that errors due to assumption of equal covariance matrices and averaging become more and more predominent as the size of the feature subsets increase. Also, the validity of the assumption of multivariate Gaussian distribution tend to reduce with increasing dimensionalities. However, the non-parametric Effective Figure of Merit Creterion being independent of such assumptions has the same over all reliability. Thus the eflectiveness and reliability of the new creterion is clearly demonstrated by these tests conducted with the well known Iris data.

References

1. Wee, W. G. .. On feature selection in a class of distribution-free pattern classifiers. *IEEE Trans. on Information Theory*, IT–16, 1970, pp. 47–55.

2. Fu, K. S., Min, P. J. and Li, T. J. On feature selection in pattern recognition. *Proc. of the 6th Annual Allerton Conference on Circuits and Systems Theory*, 1968, pp. 10–19.

3. Min, P. J., Landgrebe, D. A. and Fu, K. S. On feature selection in multiclass pattern recognition. *Proc. of the 2nd Annual Princeton Conference on Information Sciences and Systems*, 1968, pp. 453–457.

4. *IEEE Trans. on Computers*: Special Issue on feature extraction and selection in Pattern recognition, 1971, C–20, No. 9.

5. Deuser, L. M. .. A hybrid multispectral feature selection criterion. *IEEE Conference Record of the Symposium on Feature Extraction and Selection in Pattern Recognition*, 1970, pp. 223.

6. Fischer, R. A. .. The use of multiple measurement in taxonomic problem. *Ann. Eugenics*, 1936, 7, 179–187.

7. Ho, Y. C. and Kashyap, R. L. A class of itirative procedures for non-linear inequalities. *J. SIAM Control*, 1966, 4, 112–115.

8. ——— and Agrawala, A. K. On pattern classification algorithms—introduction and survey. *IEEE Trans. on Automatic Control*, 1968, AC–13, pp. 676–690.