050170

050170

ON FEATURE SELECTION IN MULTICLASS PATTERN RECOGNITION*

P. June Min, D. A. Landgrebe, and K. S. Fu
Laboratory for Agricultural Remote Sensing
Purdue University
Lafayette, Indiana

Summary

A possible solution is presented to the problem of feature selection in multiclass pattern recognition when the data are distributed according to known multivariate distributions with unknown parameters.

It is shown that the feature selection in multiclass pattern recognition can be treated in terms of pair-wise constituent errors, and the degree of separability between classes is considered as the principal measure of effectiveness of a feature set. An optimum seeking procedure using the mean of a separability measure under a constraint that minimizes the maximum variance of separability measures is discussed.

The properties of the proposed feature selection techniques are demonstrated by experimental results in agricultural remote sensing.

I Introduction

Since the performance of a recognition system depends on the particular set of features chosen, the problem of feature selection is to select a set of features which minimizes the system error (probability of misrecognition) over all possible sets of features.

However, in most problems of multiclass pattern recognition (MPR), direct minimization of the system error is often impossible. An explicit analytical expression for the system error is difficult to find, and even if it may be found, the expression may be too complicated for analytical minimization. Therefore it is the objective of this paper to develop a suboptimum feature selection technique which is easy to implement.

Referring to Marill and Green[1], it can be conjectured that the ability of measurements to discriminate two classes depends on a "statistical distance" between the class distributions (at least in the case of equal covariance matrices). The formulation of such a measure is extended here to MPR in terms of parameters of the class distributions involved. This measure can be quantitatively related to system error and reflects the effectiveness of the measurements.

In the following sections, the effectiveness of a feature set is defined and the system error is approximated by the sum of pair-wise constituent errors. Since only a pair of classes is considered each time, a statistical quantity which is monotonically related to each of the pair-wise constituent errors is used as a measure of feature effectiveness and denoted as the degree of separability between classes. An optimum seeking procedure is developed by using the mean of a separability measure under a constraint that minimizes the maximum variance of separability measures.

*The work reported here was sponsored by the U.S. Department of Agriculture under contract no. 12-14-100-8926(20), and the National Science Foundation Grant GK-1970.

II Description of System Error in MPR

A feature set $\alpha^{(p)}$ is a set of p possible measurements in ensemble measurement space $\{\Omega_X;\ X = \{X_i;\ i=1,2,\ldots,m\}\}$. A vector $\underline{X} = \{X_i;\ i=1,2,\ldots,p\}$ with real components is used to denote a point in p dimensional pattern space Ω_p. Let $\{\alpha_k^{(p)};\ k=1,2,\ldots,n\}$ be a family of p-tuple feature sets in Ω_X and $\{\omega_i;\ i=1,2,\ldots,s\}$ be the finite number of classes from which pattern comes. Then, for each of the given feature sets, $R = \{R_i;\ i=1,2,\ldots,s\}$, which is determined by decision rule, is a bounded subspace in Ω_p; that is, $R\varepsilon\Omega_p$. And the error made by a decision in favor of class ω_i is $\gamma_{(i/j)}$, if true patterns come from class ω_j and the system error $\gamma(\alpha_k^{(p)}, R\varepsilon\Omega_p)$ is expressed as a linear sum of $\gamma_{(i/j)}$; that is,

$$\gamma(\alpha_k^{(p)}, R\varepsilon\Omega_p) = \sum_{i=1}^{s}\sum_{j=1}^{s} \ell_{(i/j)}\gamma_{(i/j)} \qquad (1)$$

where $\ell_{(i/j)}$ is a component of the weighted cost matrix $\underline{L} = \{\ell_{(i/j)}\};\ i,j=1,2,\ldots,s$.

In this paper, the system error for using a particular set of feature $\alpha_k^{(p)}$ is chosen as a performance index of the recognition system. The system error can be denoted by the over-all expected error probability in probabilistic MPR.

Definition I: For a given decision rule, the feature set $\alpha_k^{(p)}$ is more effective than feature set $\alpha_\ell^{(q)}$, written $\alpha_k^{(p)} \supset \alpha_\ell^{(q)}$; $p,q=1,2,\ldots,m$; $k,\ell=1,2,\ldots,n$; $p\neq q$ or $\ell\neq k$, if

$$\gamma(\alpha_k^{(p)}, R\varepsilon\Omega_p) < \gamma(\alpha_\ell^{(q)}, R\varepsilon\Omega_q) \text{ holds.}$$

The recognition scheme is normally restricted to certain specified form of decision rules. Thus, an optimum set of features is selected within the specified form of decision rule according to the above feature selection criterion.

In a two-class pattern recognition problem, the system error is the sum of two errors. However, in MPR, the system error is in general less than or equal to the sum of pair-wise constituent errors. A pair-wise constituent error (p.c.e.) of deciding the inputs from ω_i when it actually comes from ω_j can be expressed by

$$\gamma^*_{(i/j)} = \gamma_{(i/j)} + \sum_{\substack{k=1 \\ k\neq i,j}}^{s} \gamma^*_{(i/j)\cdot k} \qquad (2)$$

where $\gamma^*_{(i/j)\cdot k}$ is error made into class ω_k against class ω_j, which is in multiple error regions made with ω_i. This error comes in due to more than two

overlapped error regions.

Recent results in MPR have shown that: (a) even in the probabilistic MPR, an explicit analytical expression for the system error is difficult to determine due to more than two overlapped error regions in pattern space, and (b) since the system error is quantitatively related to the over-all p.c.e., the minimization of the system error could be approximated by minimizing the over-all p.c.e. And the convergence of this statement has been proved by means of implicit statistical expressions.*

Therefore, the problem of feature selection in MPR can be treated in terms of p.c.e. and the degree of separability between classes can be considered as the principal measure of feature effectiveness in MPR.

III Feature Selection Criterion in MPR

3.1 Measure of Separability

In this paper, a feature selection procedure is developed under the assumption of using the optimum classifier for normal distributions. However, in general, the individual distribution can be any known form and the parameters of the distributions might be estimated from a set of known samples. In many problems in MPR, a priori probabilities cannot be assigned and the optimum decision rule is to minimize the maximum probability of error. Referring to Anderson[2], the minimax decision procedure under the condition of unknown a priori probabilities is an admissible solution to the optimum procedure with known a priori probabilities. Therefore, in this paper, it is assumed that a priori probabilities are unknown.

A discriminant function between class ω_i and and class ω_j is denoted by $g_{ij}(\underline{X}\varepsilon\Omega_p)$. In probabilistic MPR, the probability of the system error using a decision rule $d(\underline{X}\varepsilon\Omega_p)$ is equal to

$$P_s[\varepsilon/d(\underline{X}\varepsilon\Omega_p)]=\sum_{i=1}^{s}\sum_{j=1}^{s}\int_{R_{(i/j)}} \ell_{(i/j)}p_j(\underline{X}\varepsilon\Omega_p)d\underline{X} \quad (3)$$

where $p_j(\underline{X}\varepsilon\Omega_p)$ is the conditional probability density function and $R_{(i/j)}$ is the region in Ω_p of which the loss in deciding that $\underline{X}$ is from class ω_i when it is actually from class ω_j.

For a given family of discriminant function $\{g_{ij}(\underline{X}\varepsilon\Omega_p);\ i=1,2,\ldots,s-1;\ j=i+1,\ldots,s\}$, the decision rule is to choose ω_i if $g_{ij}(\underline{X}\varepsilon\Omega_p) > 0$ and ω_j if $g_{ij}(\underline{X}\varepsilon\Omega_p) \leq 0$. Then each of the integration ranges in Eq. (3) becomes

$$R_{(i/j)}=R\{g_{ij}(\underline{X}\varepsilon\Omega_p) > 0\}\bigcap_{k=1}^{i-1}R\{g_{ij}(\underline{X}\varepsilon\Omega_p) > 0\} \quad (4)$$

$$\bigcap_{k=i+1}^{s} R\{g_{ij}(\underline{X}\varepsilon\Omega_p) \leq 0\}$$

and an analytical expression of this range is not easy to find.

However, suppose that a pair of classes is considered each time. Then Eq. (4) becomes simply

$$R^*_{(i/j)} = R\{g^*_{ij}(\underline{X}\varepsilon\Omega_p) > 0\} \quad (5)$$

and

$$\gamma^*_{(i/j)} = Pr\{g^*_{ij}(\underline{X}\varepsilon\Omega_p) > 0/j\} \quad (6)$$

*To be published by authors.

where $g^*_{ij}(\underline{X}\varepsilon\Omega_p)$ is denoted as the pair-wise optimum discriminant function between class ω_i and class ω_j. Let $\underline{L} = \frac{1}{s(s-1)}\{1-\delta_{ij}\}$ and then the over-all pair-wise constituent error is the weighted sum of $r_{(i/j)}$; i,j=1,2,...,s; $i\neq j$, where

$$P^*_s[\varepsilon/d(\underline{X}\varepsilon\Omega_p)]=\frac{1}{s(s-1)}\sum_{\substack{i=1\\ i\neq j}}^{s}\sum_{j=1}^{s}P^*_{ij}[\varepsilon/d(\underline{X}\varepsilon\Omega_p)] \quad (7)$$

$$P^*_{ij}[\varepsilon/d(\underline{X}\varepsilon\Omega_p)]=Pr[g^*_{ij}(\underline{X}\varepsilon\Omega_p)>0/j]+Pr[g^*_{ij}(\underline{X}\varepsilon\Omega_p)\leq 0/i].$$

When the maximum likelihood decision rule (MLDR) is used as an optimum classifier, the logarithm of the likelihood ratio is denoted as a discriminant function, and different measures of feature effectiveness have been applied to various feature selection problems (Marill and Green[1], Grettenberg[3], Kailath[4], and others). When covariance matrices are equal, the probability of the system error for two-class recognition depends on a statistical quantity, say divergence. However, when covariance matrices are unequal, there is no simple function which relates the measure to error probability, and only the upper and lower bounds, between which the error rate lies for a given value of divergence, have been given (Marill and Green[1], Kakoda and Shepp[5]). However, to develop the feature selection criterion in MPR, the most desirable condition is the monotonicity between each of p.c.e. and the statistical quantity.

Suppose that a linear discriminant function is used to discriminate a pair of classes each time. Then, since we are considering only the degree of separability between each pair of classes, the feature selection technique based on linear discriminants is admissible for that based on MLDR by means of the optimality of the minimax procedure.

Consider a family of linear discriminant functions $\{g^*_{ij}(\underline{X}\varepsilon\Omega_p);\ i=1,2,\ldots,s-1;\ j=i+1,\ldots,s\}$ where $g^*_{ij}(\underline{X}\varepsilon\Omega_p) = \underline{b}'_{ij}\underline{X}-c_{ij}$. Referring to Anderson[2], Eq. (7) becomes

$$P^*_{ij}(\varepsilon)=\{Pr[\underline{b}'_{ij}\underline{X}>c_{ij}/j]+Pr[\underline{b}'_{ij}\underline{X}<c_{ij}/i]\}$$
$$=\{2-Pr[\xi<d_j]-Pr[\xi<d_i]\} \quad (8)$$

where $d_i=\frac{\underline{b}'_{ij}\underline{\mu}_i-c_{ij}}{(\underline{b}'_{ij}\underline{\Sigma}_i\underline{b}_{ij})^{1/2}}$, $d_j=\frac{c_{ij}-\underline{b}'_{ij}\underline{\mu}_j}{(\underline{b}'_{ij}\underline{\Sigma}_j\underline{b}_{ij})^{1/2}}$

and $\xi \sim N(0,1)$. The condition that minimizes the maximum expected p.c.e. is $d_{ij}=d_i=d_j$ and then,

$$d_{ij} = \frac{\underline{b}'_{ij}(\underline{\mu}_i-\underline{\mu}_j)}{(\underline{b}'_{ij}\underline{\Sigma}_i\underline{b}_{ij})^{1/2}+(\underline{b}'_{ij}\underline{\Sigma}_j\underline{b}_{ij})^{1/2}} \quad (9)$$

The value of $\underline{b}_{ij}$, which maximizes d_{ij} is of the form

$$\underline{b}_{ij}=[\lambda_{ij}\underline{\Sigma}_i+(1-\lambda_{ij})\underline{\Sigma}_j]^{-1}(\underline{\mu}_i-\underline{\mu}_j) \quad (10)$$

where λ_{ij} is a Lagrange multiplier. λ_{ij} is calculated by solving the equation

$$\underline{b}'_{ij}[\lambda^2_{ij}\underline{\Sigma}_i-(1-\lambda_{ij})^2\underline{\Sigma}_j]\underline{b}_{ij} = 0 \quad (11)$$

with $0<\lambda_{ij}<1$. Finally, the p.c.e. between class ω_i and class ω_j becomes

$$P^{*}_{ij}(\varepsilon) = 2(1-\Pr[\xi<d_{ij}]) \qquad (12)$$

Therefore, using linear discriminant function, we are able to find a monotonic functional relation between each of p.c.e. and a statistical quantity. This quantity d_{ij} is denoted as the measure of separability between classes.

3.2 Admissible Optimum Seeking Procedure

Let $\alpha_k^{(p)}$, k=1,2,...,n, be a family of feature sets which are possible p-tuple subsets from the entire feature space Ω_X. Suppose that for each of a given measure set $\alpha_k^{(p)}$, it is possible to calculate all the separability measures $\{d^{(p)}_{ij\cdot k}; i=1,2,\ldots,s-1; j=i+1,\ldots,s\}$. Then the ensemble of measures $\{d^{(p)}_{ij\cdot k}\}$ is used to determine an optimum feature set.

Search problems occur for several reasons. The function describing the relation between the system error and the measure of feature effectiveness is analytically not sufficient and the formulation of a single condition which leads to Definition I does not exist. However, there are dependable conditions for an optimum seeking procedure to be taken as the conditions of feature selection criterion. Among them are:

(1) Maximize the mean of the separability measure for all pairs of classes; that is,

$$\max_k \left\{ \sum_{i=1}^{s-1} \sum_{j=i+1}^{s} d^{(p)}_{ij\cdot k} \right\};\ k=1,2,\ldots,n$$

(2) Maximize the minimum degree of separability; that is,

$$\max_k \{\min_{i,j} d^{(p)}_{ij\cdot k}\};\ i=1,2,\ldots,s-1;\ j=i+1,\ldots,s;\ k=1,2,\ldots,n$$

If a set of the conditions is to be satisfied to select an optimum feature set, feature selection technique automatically becomes a dynamic decision procedure. The procedure is formulated with the aid of dynamic programming and an admissible optimum seeking procedure is introduced so that the effectiveness of feature set can be optimized by the number of conditions.

Most previous work on feature selection criterion in MPR was undertaken by using the maximin condition as an index of feature selection (Grettenberg[3], Fu and Chen[6]). In this paper, the following criterion is proposed for feature selection in MPR: Maximize the mean of the separability measure for all pairs of classes under a constraint that minimizes the maximum variance of the separability measure.

IV Results of System Implementation and Experiments

4.1 Description of Data

The proposed feature selection technique in MPR has been tested on a digital computer (IBM System 360/Model 44) by performing several experiments. The data used in these tests were obtained as a part of a project in agricultural remote sensing. The objective of this project is the design of a system which can carry out agricultural surveys of crop conditions using aerospace platforms. The sensor used in this case was an airborne multiband optical-mechanical scanner. The output of this particular scanner provides 12 electrical signals, each one of which is proportional to the radiant energy from the scene in a different spectral band. The 12 bands cover the range from 0.4 to 1.0 microns in the visible and near infrared portions of the spectrum. By simultaneously sampling the output of 12 bands, one obtains a vector which contains all the spectral information available about a given resolution element on the ground.

The data can be fairly reasonably modeled by multivariate normal distributions with unequal covariance matrices. The MLDR is used as an optimum classifier and the appropriate statistical parameters are estimated from an adequate number of training samples for each class. For the experiments performed, five major crops are considered; namely, soybeans, corn, oats, wheat, and red clover. Because of the variety of crop types and growing conditions, there are a number of subclasses to be considered.

4.2 General Procedures

The tests of the proposed feature selection technique were performed by three closely related programs: Phase I - Estimation of means and covariances; Phase II - Calculation of separability measures and optimum seeking processes; and Phase III - Classification and error estimation. The flow diagram of the system is shown in Figure 1 and only the important parts of Phase III are discussed.

(1) The estimation of Lagrange multiplier λ_{ij} is carried out by the repeated application of a dynamic programming which employed a simple gradient scheme for finding a saddle point of Eq.(11).

(2) The first supervising control is applied to determine the lower limit of the error probability curve ($P^{*}_{ij}(\varepsilon)$ vs $d^{(p)}_{ij}$). If

$$d^{(p)}_{ij(\max)} = d^{(p)}_{ij}\{P^{*}_{ij}(\varepsilon) = \alpha\},$$

it is possible to bound the lower part of the error probability curve with the maximum allowable risk in the level of α. This is true since even if the degree of separability is larger than $d^{(p)}_{ij(\max)}$, the ability to discriminate is not appreciably improved. Thus essentially the degree of separability is normalized by $d^{(p)}_{ij(\max)}$. Throughout the test experiments, the value $d_{ij(\max)} = 3.00$ was used.

(3) The second supervising control is applied to impose the constraints that minimize the maximum variance of separability measures. For a given value $d^{(p)}_{ij(\min)}$, we can select $\ell(<<n)$ candidate sets from n possible sets so that

$$\min_{i,j}\{d^{(p)}_{ij}\} > d^{(p)}_{ij(\min)} \text{ for all } k=1,2,\ldots,\ell.$$

(4) The final decision is made by maximizing the mean information of separability measures.

4.3 Experimental Results

To demonstrate the properties of the analytical feature selection technique, a typical experimental result is presented. Five classes were composed by combining the several different subclasses of each of five major crops within allowable statistical unimodality. The statistical parameters were estimated from 425 samples per class and this size of training samples was considered large enough so that the estimated parameters could be unbiased. No tests were made to check the assumption of normality of distribution with unequal covariances. However, the univariate sample marginal cumulative distributions were estimated from histograms and considered, in general, as unimodal normal shapes.

After using the proposed technique to select the best feature sets from all possible combinations, the effectiveness of the sets was tested by computing the error rate of classification with 7,530 samples (about 1,500 samples per class) by the MLDR classifier. The results of this experiment in Figure 2 indicate that the optimum feature sets were selected for all the combinations. A typical example of the recognition matrix is shown in Table 1.

The following points are evident by these experimental results:

(1) For all the combinations tested, the proposed approach can be applied to select optimum feature sets.

(2) There is a smaller number of subsets which are almost as effective as the complete feature set. When the system recognition performance is lower and the risk in parameter estimation is involved, we are able to find a subset of features which gives appreciably improved performance over the complete set. Estes[7] has shown similar results with equal covariance matrices. And it appears that the optimum number of features could be determined by finding the maximum gradient point of the separability measure curve (the separability measures of the best feature sets versus the number of features).

(3) To find an absolute optimum feature set, it is necessary to evaluate the effectiveness of all the possible subsets of the given features. However, this is not, in general, possible when the number of available features is large because of the computation time required. As an alternative for selecting the best feature subset, a sequential forward selection procedure was tested along with absolute optimum selection; that is, r-tuple feature sets were constructed by adding an additional feature to the (r-1)-tuple best feature set chosen. Since the (r-1)-tuple best feature set is not necessarily a subset of the r-tuple best set, the forward sequential selection procedure is only a suboptimum procedure, but, because of the computation time required, it is a feasible solution for the large size of a given feature set.

(4) Even if there is no monotonic functional relation between the error probability and the divergence, the experiments were extended to include the selection of the best feature sets by using the divergence as the separability measure. The results are less effective than the separability measure of the minimax linear discriminant, but the time required for computation is shorter and the accuracy is acceptable for most cases. The same optimum seeking procedure was applied for divergence criterion.

V Conclusions

As far as the given data are concerned, the proposed feature selection techniques have been successfully used. In general, when the covariance matrices are unequal and over-all expected error probability is used as a performance index, it is believed that the separability measure developed by minimax linear discriminants could be a useful criterion of feature effectiveness. For equal covariance matrices, the separability measure is equivalent to the divergence, which has been successfully used for various problems. As far as the computation time required is concerned, the forward sequential selection procedure might be preferable.

References

1. Marill, T. and Green, D. M.: "On the Effectiveness of Receptor in Recognition Systems," IEEE Trans. PGIT, vol. IT-9, No. 1, pp. 11-17, January 1963.

2. Anderson, T. W. and Bahadue, R. R.: "Classification into Two Multivariate Normal Distributions with Different Covariance Matrices," Annals Math. Stat., vol. 33, No. 2, pp. 420-431, June 1962.

3. Grettenberg, T. L.: "Signal Selection in Communication and Radar System," IEEE, PGIT, vol. IT-9, pp. 265-275, October 1963.

4. Kailath, T.: "The Divergence and Bhattacharyya Distance Measures in Signal Selection," IEEE Trans. on Comm. Tech., vol. COM-15, No. 1, February 1967.

5. Kakoda, T. T. and Shepp, L. A.: "On the Best Finite Set of Linear Observables for Discriminating Two Gaussian Signals," IEEE PGIT, vol. IT-13, No. 2, April 1967.

6. Fu, K. S. and Chen, C. H.: "Sequential Decisions, Pattern Recognition and Machine Learning," TR-EE65-6, School of Electrical Engineering, Purdue University, April 1965.

7. Estes, S. E.: "Measurement Selection for Linear Discriminants Used in Pattern Classification," IBM Research Report, RJ-331, San Jose, April 1965.

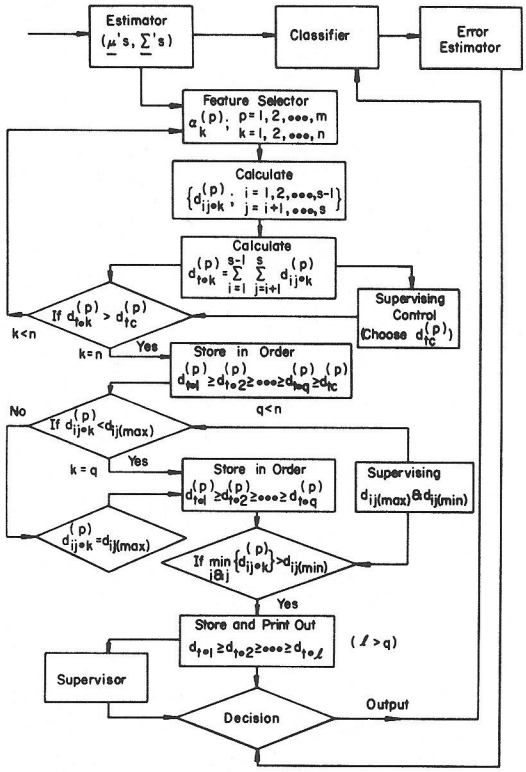
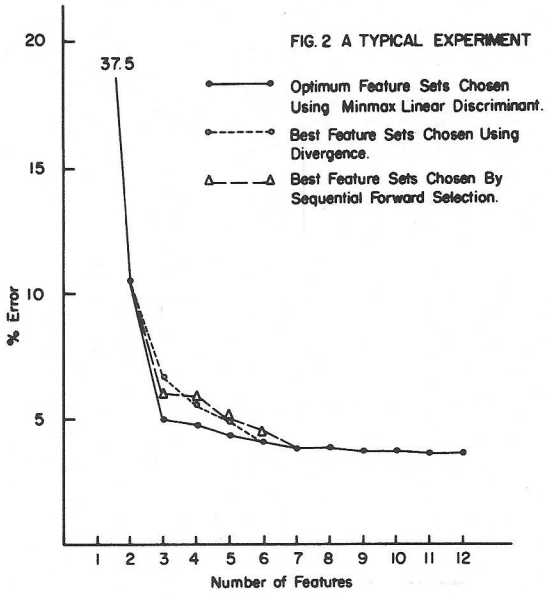


FIG. I. THE FLOW DIAGRAM OF THE SYSTEM



CLASS	NO. OF SAMPS	PCT CORCT	NO. OF SAMPLES CLASSIFIED INTO				
			SOYB	CORN	OATS	WHEAT	CLOV
SOYB	1535	96.0	1473	44	18	0	0
CORN	1476	94.0	77	1397	2	0	0
OATS	1483	90.0	7	4	1334	21	117
WHEAT	1538	97.8	2	0	32	1504	0
CLOV	1498	96.9	5	24	18		1451
TOTAL	7530	95.0	1564	1469	1404	1525	1568

TABLE I: THE RECOGNITION MATRIX OF EXPERIMENT I USING 4-TUPLE FEATURE SET $\{X_1, X_6, X_{10}, X_{11}\}$