The Feature Extraction Procedure for Pattern Recognition with Learning Using Genetic Algorithm

EDWARD PUCHALA
Wroclaw University of Technology
Faculty of Electronics
Wyb.Wyspianskiego 27,50-370 Wroclaw
POLAND

ALEKSANDER REWAK
Wroclaw University of Technology
Faculty of Electronics
Wyb.Wyspianskiego 27,50-370 Wroclaw
POLAND

Abstract: The paper deals with the extraction of features for statistical pattern recognition. In particular, the case of recognition with learning is considered. Bayes probability of correct classification is adopted as the extraction criterion. The problem with incomplete probabilistic information is discussed and Bayes-optimal feature extraction procedure is presented in detail. As method of solution of optimal feature extraction a genetic algorithm is proposed. A numerical example demonstrating quality of proposed algorithm to solve feature extraction problem is presented. .

Key–Words: Genetic algorithm, Feature extraction, Bayes approach

1 Introduction

Feature dimension reduction has been an important and long-stading research problem in statistical pattern recognition. In general, dimension reduction can be defined as a transformation from original high-dimensional space to low-dimensional space where an accurate classifier can be constructed.

There are two main methods of dimensionality reduction ([2], [6]): *feature selection* in which we select the best possible subset of input features and *feature extraction* consisting in finding a transformation (usually linear) to a lower dimensional space. Although feature selection preserves the original physical meaning of selected features, it costs a great degree of time complexity for an exhaustive comparison if a large number of features is to be selected. In contrast, feature extraction is considered to create a new and smaller feature set by combining the original features. We shall concentrate here on feature extraction for the sake of flexibility and effectiveness [7]. There are many effective methods of feature extraction. One can consider here linear and nonlinear feature extraction procedures, particularly ones which ([4], [5]):

1. assume underying Gaussian distribution in the data ([6], [7], [8]),
2. utilize nonparametric sample-based methods when data cannot be described with the Gaussian model ([9]),
3. minimize the empirical probability of Bayes error ([6], [10]),
4. maximize the criteria for the information values of the individual features (or sets of features) describing the objects ([4], [5], [11]).

For the purpose of classifiation, it is sensible to use linear feature extraction techniques which is considered as a linear mapping of data from a high to a low-dimensional space, where class separability is approximately preserved. Construction of linear transformation is based on minimization (maximization) of proper criterion in the transformed space. In other words, in order to define a linear transformation one should determine the values of the transformation matrix components as a solution of an appropriate optimization problem.

As it seems, the Bayes probability of error (or equivalently, the Bayes probability of correct classification) i.e. the lowest attainable classification error is the most appropriate criterion for feature extraction procedure. Unfortunately, this criterion is very complex for mathematical treatment, therefore researches have restored to other criteria like various functions of scatter matrices (e.g. Fisher criterion) or measures related to the Bayes error (e.g. Bhattacharyya distance).

In this paper we formulate the optimal feature extraction problem adopting the Bayes probability of correct classification as an optimality criterion. Since this problem cannot be directly solved using analytical ways (except simple cases including for example multivariate normal distribution), we propose to ap-

ply genetic algorithm (GA), which is very-well known heuristic optimization procedure and has been successfully applied to a broad spectrum of optimization problems, including many pattern recognition and classification tasks [12], [13].

The contents of the paper are as follows. In section 2 we introduce necessary background and formulate the Bayes-optimal feature extraction problem. In section 3 and 4 optimization procedures for the cases of complete probabilistic information and recognition with learning are presented and discussed in detail. Section 5 describes nonparametric algorithms which were applied to find optimal solution. Section 6 presents solution of extraction problem via GA. Finally, conclusions are presented in section 7.

2 Preliminaries and the Problem Statement

Let us consider the pattern recognition problem with probabilistic model. This means that n-dimensional vector of features describing recognized pattern $x = (x_1, x_2, ..., x_n)^T \in \mathcal{X} \subseteq \mathcal{R}^n$ and its class number $j \in \mathcal{M} = \{1, 2, ..., M\}$ are observed values of a pair of random variables $(\mathbf{X}, \mathbf{J})$, respectively. Its probability distribution is given by *a priori* probabilities of classes

$$p_j = P(\mathbf{J} = j),\ j \in \mathcal{M} \tag{1}$$

and class-conditional probability density function (CPDFs) of $\mathbf{X}$

$$f_j(x) = f(x/j),\ x \in \mathcal{X},\ j \in \mathcal{M}. \tag{2}$$

In order to reduce dimensionality of feature space let consider linear transformation

$$y = Ax, \tag{3}$$

which maps n-dimensional input feature space $\mathcal{X}$ into m-dimensional derivative feature space $\mathcal{Y} \subseteq \mathcal{R}^m$, or - under assumption that $m < n$ - reduces dimensionality of space of object descriptors. It is obvious, that y is a vector of observed values of m dimensional random variable $\mathbf{Y}$, which probability distribution given by CPDFs depends on mapping matrix A, viz.

$$g(y/j; A) = g_j(y; A),\quad y \in \mathcal{Y},\ j \in \mathcal{M}. \tag{4}$$

Let introduce now a criterion function $Q(A)$ which evaluates discriminative ability of features y, i.e. Q states a measure of feature extraction mapping (3). As a criterion Q any measure can be involved which evaluates both the relevance of features based on a feature capacity to discriminate between classes or quality of a recognition algorithm used later to built the final classifier. In the further numerical example the Bayes probability of correct classification will be used, namely

$$Q(A) = Pc(A) = \int_{\mathcal{Y}} \max_{j \in \mathcal{M}} \{p_j\, g_j(y; A)\}\, dy. \tag{5}$$

Without any loss of generality, let us consider a higher value of Q to indicate a better feature vector y. Then the feature extraction problem can be formulated as follows: for given *priors* (1), CPDFs (2) and reduced dimension m find the matrix A^* for which

$$Q(A^*) = \max_A Q(A). \tag{6}$$

3 Optimization Procedure

In order to solve (6) first we must explicitly determine CPDFs (4). Let introduce the vector $\bar{y} = (y, x_1, x_2, ..., x_{n-m})^T$ and linear transformation

$$\bar{y} = \bar{A}\, x, \tag{7}$$

where

$$\bar{A} = \begin{bmatrix} & A & \\ - & - & - \\ I & | & 0 \end{bmatrix} \tag{8}$$

is a square matrix $n \times n$. For given y equation (7) has an unique solution given by Cramer formulas

$$x_k(y) = \mid \bar{A}_k(y) \mid \cdot \mid \bar{A} \mid^{-1}, \tag{9}$$

where $\bar{A}_k(y)$ denotes matrix with k-th column replaced with vector $\bar{y}$. Hence putting (9) into (2) and (4) we get CPDFs of $\bar{y}$ ([3]):

$$\bar{g}_j(\bar{y}; A) = J^{-1} \cdot f_j(x_1(\bar{y}), x_2(\bar{y}), \cdots, x_n(\bar{y})), \tag{10}$$

where J is a Jacobian of mapping (7). Integrating (10) over variables $x_1, ..., x_{n-m}$ we simply get

$$g_j(y; A) = \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} ... \int_{\mathcal{X}_{n-m}} \bar{g}_j(\bar{y}; A) dx_1 dx_2 ... dx_{n-m}. \tag{11}$$

Formula (11) allows one to determine class-conditional density functions for the vector of features y, describing the object in a new m-dimensional space. Substituting (11) into (5) one gets a criterion defining the probability of correct classification for the objects in space $\mathcal{Y}$:

$$Q(A) = Pc(A) = \int_{\mathcal{Y}} \max_{j \in \mathcal{M}} \left\{ p_j \cdot \int_{\mathcal{X}_1} ... \int_{\mathcal{X}_{n-m}} J^{-1} \times \right.$$

$$\left. \times f_j(x_1(\bar{y}), \cdots, x_n(\bar{y}))\ dx_1 ... dx_{n-m} \right\} dy =$$

$$= \int_{\mathcal{Y}} \max_{j \in \mathcal{M}} \left\{ p_j \cdot \int_{\mathcal{X}_1} ... \int_{\mathcal{X}_{n-m}} J^{-1} f_j(\mid \bar{A}_1(y) \mid \times \right.$$

$$\left. \mid \bar{A} \mid^{-1}, \cdots, \mid \bar{A}_n(y) \mid \cdot \mid \bar{A} \mid^{-1})\ dx_1 ... dx_{n-m} \right\} dy. \tag{12}$$

Thus, the solution of the feature extraction problem (6) requires that such matrix A^* should be determined for which the Bayes probability of correct classification (12) is the maximum one.

Consequently, complex multiple integration and inversion operations must be performed on the multidimensional matrices in order to obtain optimal values of A. Although an analytical solution is possible (for low n and m values), it is complicated and time-consuming. Therefore it is proposed to use numerical procedures. For linear problem optimization (which is the case here) classic numerical algorithms are very ineffective. In a search for a global extremum they have to be started (from different starting points) many times whereby the time needed to obtain an optimal solution is very long. Thus it is only natural to use the parallel processing methodology offered by genetic algorithms ([14]).

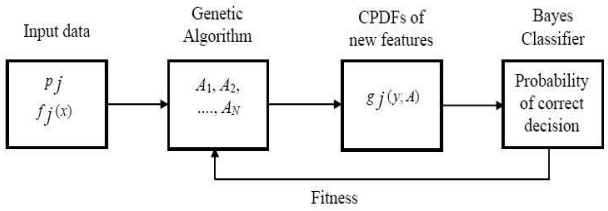

Figure 1: GA-based Bayes-optimal feature extractor

Fig. 1 shows the structure of a GA-based feature extractor using Bayes probability of correct classification as an evaluation criterion. The GA maintains a population of transformation matrices A. To evaluate each matrix in this population, first the CPDFs (11) of features y in transformed space must be determined and next probability of Bayes correct classification (12) is calculated. This accuracy, i.e. fitness of individual is a base of selection procedure in GA. In other words, the GA presented here utilizes feedback from the Bayes classifier to the feature extraction procedure.

4 The Case of Recognition with Learning

It follows from the above considerations that an analytical and numerical solution of the optimization problem is possible. But for this one must know the class-conditional density functions and the *a priori* probabilities of the classes. In practice, such information is rarely available. All we know about the classification problem is usually contained in the so-called learning sequence:

$$S_L(x) = \{(x^{(1)}, j^{(1)}), (x^{(2)}, j^{(2)}), ..., (x^{(L)}, j^{(L)})\}. \tag{13}$$

Formula (13) describes objects in space $\mathcal{X}$. For the transformation to space $\mathcal{Y}$ one should use the relation:

$$y^{(k)} = A \cdot x^{(k)};\ \ k = 1, 2, ..., L \tag{14}$$

and then the learning sequence assumes the form:

$$S_L(y) = \{(y^{(1)}, j^{(1)}), (y^{(2)}, j^{(2)}), ..., (y^{(L)}, j^{(L)})\}. \tag{15}$$

The elements of sequence $S_L(y)$ allow one to determine (in a standard way) the estimators of the *a priori* probabilities of classes p_{jL} and class-conditional density functions $f_{jL}(x)$. Then the optimization criterion assumes this form:

$$Q_L(A) = Pc_L(A) = \int_{\mathcal{Y}} \max_{j \in \mathcal{M}} \left\{ p_{jL} \cdot \int_{\mathcal{X}_1} ... \int_{\mathcal{X}_{n-m}} \times \right.$$

$$\left. \times J^{-1} f_{jL}(x_1(\bar{y}), \cdots, x_n(\bar{y}))\ dx_1 ... dx_{n-m} \right\} dy. \tag{16}$$

Alternatively, in case of recognition with learning, the criterion (16) can be estimated nonparametrically by first estimating CPDFs of features y on the base of samples (15) (e.g. using either k-NN or Parzen procedures [1], [2]) and then classifying the available samples according to the empirical Bayes rule. The number of samples misclassified by the algorithm is counted and the error estimate is obtained by dividing this number by the total number of training samples.

The next section presents a method for nonparametric estimation based on the Parzen procedure.

5 Non-complete Probabilistic Information - Nonparametric Estimation

It is apparent that in order to estimate optimization criterion (16) one should first estimate CPDFs for features y. Parzen's procedure was used for this purpose. According to the procedure's assumptions function $f_{jL}(y)$ can be written as follows:

$$f_{jL}(y) = \frac{1}{K_j} \sum_{k=1}^{K_j} \frac{1}{h_j(K_j)} \mathcal{H}\left[\frac{y - y_j^k}{h_j(K_j)}\right], \qquad (17)$$

where
K_j - the number of objects of learning sequences $S_L(y)$ belonging to class j,
$h_j(K_j)$ - a positive number satisfying the conditions:

$$\lim_{K_j \to \infty} h_j(K_j) = 0, \ \lim_{K_j \to \infty} K_j [h_j(K_j)]^m = \infty, \qquad (18)$$

$\mathcal{H}$ - kernel of Parzen estimator

Let us assume that the estimator kernel is the Gaussian function:

$$\mathcal{H}\left[\frac{y - y_j^k}{h_j(K_j)}\right] = \frac{1}{\sqrt{2\pi}} exp\left[-\frac{(y - y_j^k)^2}{\sqrt{2} h_j(K_j)}\right]. \qquad (19)$$

After transformations and the expansion of function (19) into a Taylor series one gets:

$$\sum_{k=1}^{K_j} \mathcal{H}\left[\frac{(y - y_j^k)}{h_j(K_j)}\right] = exp\left[-\frac{y^2}{2h_j(K_j)^2}\right] \sum_{s=0}^{\infty} c_{js} y^s, \qquad (20)$$

where

$$c_{js} = \frac{1}{s! h_j(K_j)^{2s}} \sum_{k=1}^{K_j} (y_j^k)^s exp\left[-\frac{(y_j^k)^2}{2h_j^2(K_j)}\right]. \qquad (21)$$

Substituting formula (21) into (16) one gets a criterial function whose value depends on the values of the elements of learning sequence $S_L(y)$. The latter values in turn depend on (via transformation (3) the values of the elements of matrix A. In order to determine the elements of matrix A an optimization procedure should be carried out. Because of the complex form of the criterial function it is proposed to use the genetic algorithm with proper problem encoding.

Table 1: Results of experiments (E- experiment number, NG - number of generations, a_{ik} - values of matrix A elements, $Pc_L(A)$ - probability of Bayes correct classification)

E	1	2	3	4	5
$Pc_L(A)$	0.91	0.94	0.96	0.96	0.97
a_{11}	0.561	0.647	0.518	0.854	0.251
a_{12}	-0.758	3.287	0.214	-0.788	2.227
a_{13}	0.768	0.255	0.317	0.519	0.247
a_{21}	0.138	1.550	0.176	0.205	0.317
a_{22}	0.932	-1.033	0.957	1.110	-0.504
a_{23}	0.119	0.698	0.037	-0.030	0.360
NG	11	20	6	7	13

6 Recognition with Learning - Solution via GA

In order to assess the quality of the proposed method of reducing feature vector dimensionality for recognition with learning an numerical example was considered. First learning sequence $S_L(x)$, containing elements from 3 classes ($M = 3$), was generated. It was assumed that $dim(x) = n = 3$. The following parameters of multivariate normal distribution generator were adopted:

$p_1 = p_2 = p_3 = 1/3$ - *a priori* probabilities of classes,

$cov_1 = cov_2 = cov_3 = \mathbf{1}$ - unit covariance matrices for the particular classes,

$m_1 = [0, 0, 0]^T, m_2 = [0, 1, 0]^T, m_3 = [1, 0, 1]^T$ - vectors of mean feature values for the particular classes.

In all the classes CPDFs were estimated using Parzen's estimator with a kernel in the form of a Gaussian function.

The aim of the reduction process was to reduce dimensionality to level $dim(y) = m = 2$. Therefore optimization boiled down to determining the best, in the sense of criterion (16), values of the elements of matrix A ($dim(A) = 3 \times 2$). The following genetic algorithm parameters were assumed:

- the number of populations - 10,
- the maximum number of generations - 20.

Five experiments were carried out. The results are shown in Table 1, containing matrix A element values, the corresponding criterion (16) values and the number of the generation after which the solution was found.

7 Conclusions

Feture extraction is an important task in any practical example that involves pattern classification. In this paper we formulate the optimal feature extraction problem with the Bayes probability of correct classification as an optimality criterion. Since this problem, in general case, cannot be directly solved using analytical methods, we propose to apply genetic algorithm, which is effective heuristic optimization procedure and has been successfully applied to a wide range of practical problems. This proposition leads to the distribution-free Bayes-optimal feature extraction method, which can be applied in the case of recognition with learning. A numerical example demonstrates that the GA is capable to solve this optimization problem.

Many questions of GA application in proposed procedure of feature extraction are still open, e.g. the proper choice of the appropriate GA model, especially the choice of GA control parameters and investigation of their influence on result of optimization process. Our related works are underway and the results will be reported in the near future.

References:

[1] Devroye L., Gyorfi P., Lugossi G.: A Probabilistic Theory of Pattern Recognition, Springer Verlag, New York, 1996

[2] Duda R., Hart P., Stork D.: Pattern Classification, Wiley-Interscience, New York, 2001

[3] Golub G., Van Loan C.: Matrix Computations, Johns Hopkins University Press, 1996

[4] Guyon I., Gunn S., Nikravesh M, Zadeh L.: Feature Extraction, Foundations and Applications, Springer Verlag, 2004

[5] Park H., Park C., Pardalos P.: Comparitive Study of Linear and Nonlinear Feature Extraction Methods - Technical Report, Minneapolis, 2004

[6] Fukunaga K.: Introduciton to Statistical Pattern Recognition, Academic Press, 1990.

[7] Hsieh P., Wang D., Hsu C.: A Linear Feature Extraction for Multiclass Classification Problems Based on Class Mean and Covariance Discriminant Information, IEEE Trans. on PAMI, Vol. 28 (2006) 223-235

[8] Loog M., Duin R., Haeb-Umbach R.: Multiclass Linear Dimension Reduction by Meighted Pairwise Fisher Criteria, IEEE Trans. on PAMI, Vol. 23 (2001) 762-766

[9] Kuo B., Landgrebe D.: A Robust Classification Procedure Based on Mixture Classifiers and Nonparametric Weighted Feature Extraction, IEEE Trans. on GRS, Vol. 40 (2002) 2486-2494

[10] Buturovic L.: Toward Bayes-Optimal Linear Dimension Reduction, IEEE Trans. on PAMI, Vol. 16 (1994) 420-424

[11] Choi E., Lee C.: Feature Extraction Based on the Bhattacharyya Distance, Pattern Recognition, Vol. 36 (2002) 1703-1709

[12] Raymer M., Punch W. at al.: Dimensionality Reduction Using Genetic Algorithms, IEEE Trans. on EC, Vol. 4 (2002) 164-168

[13] Rovithakis G., Maniadakis M., Zervakis M.: A Hybrid Neural Network and Genetic Algorithm Approach to Optimizing Feature Extraction for Signal Classification, IEEE Trans. on SMC, Vol. 34 (2004) 695-702

[14] Goldberg D.: Genetic Algorithms in Search, Optimization and Machine Learning. Adison-Wesley, New York, 1989