The Bayes-Optimal Feature Extraction Procedure for Pattern Recognition Using Genetic Algorithm

Marek Kurzynski, Edward Puchala, and Aleksander Rewak

Wroclaw University of Technology, Faculty of Electronics, Chair of Systems and Computer Networks, Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
{marek.kurzynski, edward.puchala, aleksander.rewak}@pwr.wroc.pl

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

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]).

S. Kollias et al. (Eds.): ICANN 2006, Part I, LNCS 4131, pp. 21–30, 2006.

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 apply 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 numerical example for which both analytical way and genetic algorithm were applied to find optimal solution. Finally, conclusions are presented in section 6.

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 $\mathcal{Y}$:

$$Q(A) = Pc(A) = \int_{\mathcal{Y}} \max_{j\in\mathcal{M}} \left\{ p_j \cdot \int_{\mathcal{X}_1}\int_{\mathcal{X}_2} ... \int_{\mathcal{X}_{n-m}} J^{-1} \times \right.$$

$$\left. \times f_j(x_1(\bar{y}), x_2(\bar{y}), \cdots, x_n(\bar{y}))\, dx_1\, dx_2\, ...\, dx_{n-m} \right\} dy =$$

$$= \int_{\mathcal{Y}} \max_{j\in\mathcal{M}} \left\{ p_j \cdot \int_{\mathcal{X}_1}\int_{\mathcal{X}_2} ... \int_{\mathcal{X}_{n-m}} J^{-1} \times \right.$$

$$\left. \times f_j(|\, \bar{A}_1(y)\, | \cdot |\, \bar{A}\, |^{-1}, \cdots, |\, \bar{A}_n(y)\, | \cdot |\, \bar{A}\, |^{-1})\, dx_1\, dx_2\, ...\, dx_{n-m} \right\} dy. \tag{12}$$

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]).

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

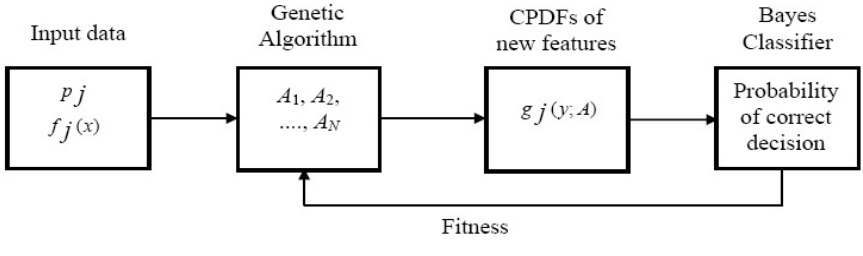

Fig. 1. GA-based Bayes-optimal feature extractor

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:

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

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:

$$Q_L(A) = Pc_L(A) = \int_{\mathcal{Y}} \max_{j \in \mathcal{M}} \left\{ p_{jL} \cdot \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} ... \int_{\mathcal{X}_{n-m}} J^{-1} \times \right.$$

$$\left. \times f_{jL}(x_1(\bar{y}), x_2(\bar{y}), \cdots, x_n(\bar{y}))\, dx_1\, dx_2 ... dx_{n-m} \right\} dy. \tag{16}$$

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 numerical example illustrating proposed approach to Bayes-optimal feature extraction problem.

5 Numerical Example

Let consider two-class pattern recognition task with equal *priors* and reduction problem of feature space dimension from $n = 2$ to $m = 1$. Input feature vector is uniformly distributed and its CPDFs are as follows:

$$f_1(x) = \begin{cases} 0.5 & \text{for } 0 \le x_1 \le 2 \text{ and } 0 \le x_2 \le x_1, \\ 0 & \text{otherwise,} \end{cases} \tag{17}$$

$$f_2(x) = \begin{cases} 0.5 & \text{for } 0 \le x_1 \le 2 \text{ and } x_1 \le x_2 \le 2, \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

Feature extraction mapping (3) has now the form

$$y = [a, 1] \cdot [x_1, x_2]^T = a \cdot x_1 + x_2 \tag{19}$$

and problem is to find such a value a^* which maximize criterion (12).

To illustrate the behavior of the GA as solution method of optimal feature extraction problem, we solve this example in threefold manner: (1) directly, according to the analytical procedure presented in section 3, (2) using GA and assuming that complete probabilistic information is given and (3) using GA procedure for the case of classification with learning.

1. Complete probabilistic information - analytical solution

Since Jacobian of (7) is equal to 1 hence from (9) and (10) for $j = 1, 2$ we get

$$\bar{g}_j(\bar{y}, a) = f_j(x_1, y - a \cdot x_1). \tag{20}$$

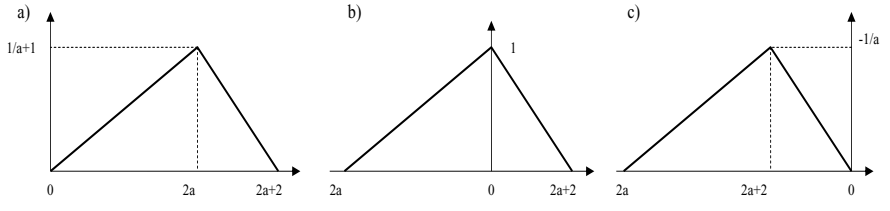

Fig. 2. Illustration of example

The results of integrating (20) over x_1, i.e. CPDFs (11) for $a \ge 1, -1 \le a \le 1$ and $a \le -1$ are presented in Fig.2. a), b) and c), respectively.

Finally, from (5) we easy get:

$$Pc(a) = \begin{cases} \frac{a+1}{4a} & \text{for } a \ge \mid 1 \mid, \\ \\ \frac{a+1}{4} & \text{for } a \le \mid 1 \mid . \end{cases} \tag{21}$$

The chart demonstrating the Bayes probability of misclassification $P_e(a) = 1 - P_c(a)$ depending on parameter a of feature extraction mapping is depicted in Fig.3. The best result $P_e(a^*) = 0$ (or equivalently $P_c(a^*) = 1$) is obtained for $a^* = -1$.

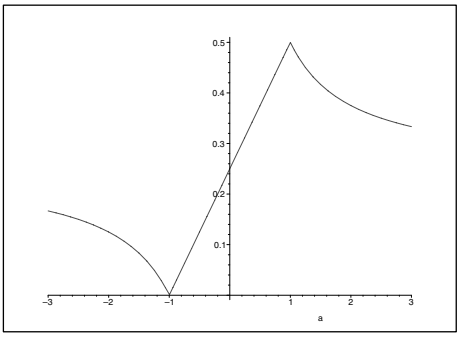

Fig. 3. Probability of misclassification

2. Complete probabilistic information - solution via GA

In order to find parametr a^*, the GA was applied, which was proceeded as follows:

- *Coding method* - Binary representation has been widely used for GA analysis. In our task, the value of parametr a was directly coded to the chromosome. It means, that a is represented by a binary string length:

$$Length = log_2[(a_{max} - a_{min})/\triangle a], \tag{22}$$

  where a_{max}, a_{min} and $\triangle a$ denote the maximum value, the minimum value and the resolution of a, respectively. To avoid irregularities we decided to put $a_{max} = 32.536$, $a_{min} = -33.0$ and $\triangle a = 0.001$ which gave the length of chromosome $Length = 16$ bits (genes).
- *The fitness function* - The Bayes probability of correct classification (12).
- *Initialization* - GA needs an initial individual population to carry out parallel multidirectional search of optimal solution. The initial population of chromosomes with which the search begins was generated randomly. The size of population after trials was set to 40.
- *Selection* - The probability of selecting a specific individual can be calculated by using the individuals fitness and the sum of population fitness. In this research a roulette wheel approach was applied. Additionally, an elitizm policy, wherein the best individuals from the current generation is copied directly to the next generation, was also used for fast convergence.
- *Crossover* - The crossover process defines how genes from the parents have been passed to the offspring. In each generation a standard two-point crossover was used and probability of crossover was equal to 1.

– *Mutation* - The mutation process simulates the natural disturbance during crossover. It was a bit-by-bit operation made with probability 0.01.
– *Stop procedure* - evolution process was terminated after 300 generations. In fact, the fitness value usually converged within this value. Fig. 4. shows the fitness change against generation number in one run of GA.

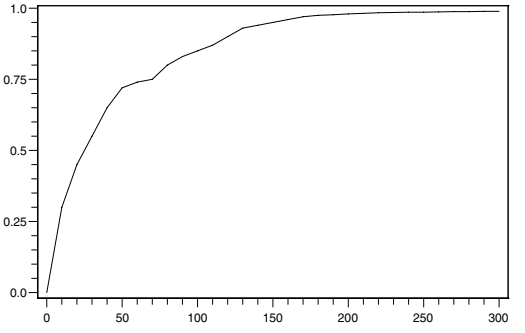

Fig. 4. The example of the course of the fitness value vs. number of generation

Table 1. Results of genetic algorithm applied to the example

Trial	Complete Information	Recognition with Learning		
		L=100	L=200	L=300
1	-0.987	-0.946	-1.075	-0.983
2	-0.983	-0.951	-1.048	-0.971
3	-0.974	-1.102	-0.957	-1.036
4	-1.078	-0.938	-0.939	-0.947
5	-1.023	-1.093	-0.962	-1.056
6	-0.991	-1.084	-1.053	-1.043
7	-0.975	-0.956	-0.961	-1.032
8	-1.052	-1.066	-0.972	-0.953
9	-0.979	-0.941	-1.042	-1.023
10	-1.044	-1.076	-0.951	-1.041
Best	-0.991	-0.956	-0.962	-1.023
Mean	-1.008	-1.015	-0.996	-1.009
SD	0.036	0.07	0.049	0.039

To compare the optimal solution and the performance of GA, ten independent runs of GA were carried out for different random initial populations. The results are shown in Table 1. The values depicted in the Table are those of the best solution obtained at the end of a GA trial.

3. Recognition with learning - solution via GA

For evaluation of GA performance in the case of recognition with learning, three experiments were made on computer generated data with different number of learning patterns ($L = 100, 200, 300$, respectively). Patterns were generated according to the CPDFs (17) and (18) using Maple 10 environment. In each experiment *priors* were calculated in standard way and CPDFs were estimated using Parzen estimator with uniform kernel function [1]. Next, GA was applied as a method of solution of optimization problem presented in section 4. GA was used with the same control parameters as in the prevoius case and the number of trials was qual to 10. The results are depicted in Table 1.

Table 1 contains also the best result, the mean value and standard deviation for each case where GA was applied. Results demonstrate that the proposed GA method can reach value of parametr of extraction mapping (19) very close to optimal solution a^*.

6 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 both in the case of complete probabilistic information and in the case of recognition with learning. A numerical example demonstrates that the GA is capable to solve this optimization problem for both cases.

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