Automatic Machine Learning of Decision Rule for Classification Problems in Image Analysis

P. Pudil, J. Novovičová* and J. Kittler
Dept. of Electronic and Electrical Engineering,
University of Surrey,
Guildford, Surrey GU2 5XH, United Kingdom

Abstract

A new method for automatic machine learning of decision rules for classification problems in image analysis is presented. The method aims at simultaneous decision rule inference and selection of discriminative features which characterize the image entities to be classified. The method is based on the approximation of class conditional densities by a mixture of parametrized densities of a special type using the EM algorithm. Its performance is tested on a classification problem involving real image data.

1 Introduction

The problem of classifying or labelling image objects or entities is one of the most common tasks in image analysis. It can be encountered at various levels of processing. Image segmentation, discrimination of objects by their shape, identification of various categories of object configurations are all examples of labelling tasks. In all such classification tasks the entity we wish to label is represented by a set of measurements. On the basis of these measurements it is assigned to one of a given number of possible categories. In this paper we focus on image labelling problems where for each class of objects we have a set of representative examples. We also make the assumption that the image objects to be classified can be segmented out so that the relevant measurements can be extracted from them.

Regardless of the nature of the classification task, its solution involves:

1. selecting the most discriminative features from image measurements that can potentially be extracted
2. learning a decision rule based on the selected features

When the form underlying the multidimensional class conditional probability densities of the measurement vector can be assumed to be parametric, it greatly simplifies the feature selection process, as well as the derivation of the decision rule. However, the parametric model will only be useful if the assumption is valid. If the samples are not from the assumed distribution, the performance of the labelling process may degrade dramatically. The reason for this is that the use of a simple parametric model such as a multivariate normal distribution can

Supported by a SERC grant GR/E 97549

*Permanently with the Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, 182 08 Prague 8, Czech Republic

BMVC 1993 doi:10.5244/C.7.2

give a very misleading description of the data, especially if the class distribution is multimodal.

The purpose of this paper is to present a method that can serve the multiple goal of:

1. learning the structure of object measurement distributions,
2. identifying the most important measurements and thus facilitate the dimensionality reduction,
3. deriving automatically a decison rule based on the selected features

for multiclass problems, even when the form underlying the class conditional probability distribution of the measurements is unknown. Such a problem arises in a number of real situations when we have the data but no other information.

It has been already stressed elsewhere (e.g. [5]) that in order to achieve an optimal overall performance in object labelling, both the classifier design and feature selection should be solved together and not separately which is often the case. However, in the case when no information about the measurement distributions is available, it is not an easy task. Though there are various nonparametric methods of classification available, none of them is problem free and universal.

The feature selection task in the absence of information regarding underlying probability structure is even more difficult. Apart from rather trivial cases when the data is governed by a simple distribution, there seems not to exist any direct method of selecting a good subset of features. Currently, the only two options left appear to be as follows:

- To use a nonparametric method of classification and to assess the quality of the feature subset indirectly by the error rate. However, in this case, the estimated error will include any structural error caused by the choice of the classifier (see e.g. Fukunaga [3]). Moreover, with increasing dimensionality this approach would become soon computationally unfeasible.

- To use a nonparametric approach to class density estimation (e.g. Parzen density estimation method) and substitute the acquired densities into the formulas for probabilistic distance measures, like Bhattacharyya distance or divergence. Then it would be theoretically possible to assess directly the quality of a feature subset by computing the corresponding probabilistic distance measure. However, since the computations would involve multiple integrations in multidimensional space, neither this approach would be possible for problems of realistic dimensionality.

So, to conclude, there is a need to find an alternative solution, which could be used at least for some real problems. Such a new approach to simultaneous feature selection and decision rule learning is presented. The method is based on approximating the unknown class conditional distributions by finite mixtures of parametrized densities of a special type. The approximation which is best in the sense of minimizing the Kullback-Leibler distances between the true and the postulated class conditional probability density functions (pdfs) mixed in the proportions in which the classes occur is used. The maximum likelihood (ML) estimates of unknown parameters of postulated class conditional pdfs are computed by the expectation-maximization (EM) algorithm (see Dempster *et al.* [1], Redner and Walker [11]). The proposed approach is especially suitable for multimodal data. The method yields the optimal feature subset of required dimensionality without the necessity to employ any search procedure and, furthermore, a pseudo-Bayes decision rule for the problem.

2 Mathematical Preliminaries

Let us assume that a pattern described by a real D-dimensional vector $\boldsymbol{x} = (x_1, x_2, \cdots, x_D)^{\mathrm{T}} \in \mathcal{X} \subset \mathcal{R}^D$ is to be classified into one of a finite set of C different classes $\Omega = \{\omega_1, \omega_2, \cdots, \omega_C\}$. The patterns are supposed to occur randomly according to some true class conditional pdfs $p^\star(\boldsymbol{x}|\omega)$ and the respective *a priori* probabilities $P^\star(\omega)$. The global pdf for vector $\boldsymbol{x}$ is then $p^\star(\boldsymbol{x}) = \sum_{\omega\in\Omega} p^\star(\boldsymbol{x}|\omega)P^\star(\omega)$ and it is independent of class.

Vector $\boldsymbol{x}$ can be then optimally classified using the Bayes minimum error rule $r(\boldsymbol{x}), r : \mathcal{X} \to \Omega$:

$$\text{if} \quad p^\star(\boldsymbol{x}|\omega_i)P^\star(\omega_i) \geq p^\star(\boldsymbol{x}|\omega_j)P^\star(\omega_j) \text{ for all } i \neq j, (\ i, j = 1, 2, \cdots, C), \tag{1}$$

then the pattern will be classified as belonging to class ω_i, i.e. $r(\boldsymbol{x}) = \omega_i$.

This classification is based on the knowledge of the components $p^\star(\boldsymbol{x}|\omega)P^\star(\omega)$, $\omega \in \Omega$ of the unconditional pdf $p^\star(\boldsymbol{x})$. Since the class conditional pdfs and the *a priori* class probabilities are seldom specified in practice, it is necessary to estimate these functions from the sets of independent labelled samples:

$$\boldsymbol{X}_\omega = \{\boldsymbol{x}_1^\omega, \boldsymbol{x}_2^\omega, \cdots, \boldsymbol{x}_{N_\omega}^\omega\}, \quad \boldsymbol{x}_k^\omega \in \mathcal{X} \subset \mathcal{R}^D, \ k = 1, \cdots, N_\omega, \quad \omega \in \Omega, \tag{2}$$

where N_ω is the number of samples from class ω.

3 Parametric Model Based on Finite Mixtures

In the case of parametric approaches to density estimation, usually a simplifying assumption about the data structure is made. As a result, instead of finding the underlying true structure in the data, a simplified and generally incorrect structure is imposed on it. This is why the practical results of estimating multivariate distributions are mostly unsatisfactory.

In our approach the following parametric model is postulated for the ω-th class conditional pdf:

$$p(\boldsymbol{x}|\omega) = \sum_{m=1}^{M_\omega} w_m^\omega F_0(\boldsymbol{x}|\boldsymbol{b}_0)F(\boldsymbol{x}|\boldsymbol{b}_m^\omega, \boldsymbol{\phi}, \boldsymbol{b}_0), \ \boldsymbol{x} \in \mathcal{X}, \ \sum_{m=1}^{M_\omega} w_m^\omega = 1, \tag{3}$$

where M_ω is the number of mixture components and w_m^ω are nonnegative weights, which is a particular case of the parametric model proposed by Grim [4]. Each component of this finite mixture includes a background distribution F_0 common to all classes, which is an important distinction from the kernel approach (see e.g. Devijver and Kittler [2]):

$$F_0(\boldsymbol{x}|\boldsymbol{b}_0) = \prod_{i=1}^{D} f(x_i|b_{0i}), \quad \boldsymbol{b}_0 = (b_{01}, b_{02}, \cdots, b_{0D}) \in \mathcal{B}^D \tag{4}$$

and a function F of the form

$$F(\boldsymbol{x}|\boldsymbol{b}_m^\omega, \boldsymbol{\phi}, \boldsymbol{b}_0) = \prod_{i=1}^{D} [\frac{f(x_i|b_{mi}^\omega)}{f(x_i|b_{0i})}]^{\phi_i}, \ \phi_i \in \{0, 1\}, \tag{5}$$

$$\boldsymbol{b}_m^\omega = (b_{m1}^\omega, b_{m2}^\omega, \cdots, b_{mD}^\omega) \in \mathcal{B}^D, \ \boldsymbol{\phi} = (\phi_1, \phi_2, \cdots, \phi_D) \in \{0, 1\}^D,$$

where $\boldsymbol{b}_0, \boldsymbol{b}_m^\omega$ and $\boldsymbol{\phi}$ are the parameter vectors. The function F is actually defined on a subspace $\mathcal{X}_l \in \mathcal{R}^l$:

$$\mathcal{X}_l = \mathcal{X}_{i_1} \times \mathcal{X}_{i_2} \times \cdots \times \mathcal{X}_{i_l}, \quad \mathcal{X}_{i_k} \subset \mathcal{R}, \quad 1 \leq i_k \leq D, \quad k = 1, \cdots, l$$

specified by nonzero binary parameters ϕ_{i_k}.

The univariate function f is assumed to be from a parametric family of probability density functions f parameterized by $b \in \mathcal{B}$, i.e. $\mathcal{F} = \{f(\xi|b), \ \xi \in \mathcal{R}, \ b \in \mathcal{B}\}$. For any choice of the binary parameters ϕ_i, which can be looked upon as *control variables*, the finite mixture (3) can be rewritten as

$$p(\boldsymbol{x}|\omega) = \sum_{m=1}^{M_\omega} w_m^\omega \prod_{i=1}^{D} [f(x_i|b_{0i})^{1-\phi_i} f(x_i|b_{mi}^\omega)^{\phi_i}]. \tag{6}$$

To our knowledge, the parametric model (3) has not yet been applied to the field of learning important image features and the decision rule. However, it will be seen later that as a result of approximating the unknown conditional distributions with the model (6) the process of feature selection and decision rule design becomes a very simple task.

4 Parameter Estimation

The problem thus remains how to estimate $p(\boldsymbol{x}|\omega)$, which is of known form but has an unknown parameters. We therefore modify our notation to write $p(\boldsymbol{x}|W_\omega, B_\omega, \boldsymbol{\phi}, \boldsymbol{b}_0)$ as the class conditional pdf for class ω:

$$p(\boldsymbol{x}|W_\omega, B_\omega, \boldsymbol{\phi}, \boldsymbol{b}_0) = F_0(\boldsymbol{x}|\boldsymbol{b}_0) \sum_{m=1}^{M_\omega} w_m^\omega F(\boldsymbol{x}|\boldsymbol{b}_m^\omega, \boldsymbol{\phi}, \boldsymbol{b}_0), \tag{7}$$

$$W_\omega = (w_1^\omega, w_2^\omega, \cdots, w_{M_\omega}^\omega), w_m^\omega \geq 0, \ \sum_{m=1}^{M_\omega} w_m^\omega = 1, \ \ B_\omega = (\boldsymbol{b}_1^\omega, \boldsymbol{b}_2^\omega, \cdots, \boldsymbol{b}_{M_\omega}^\omega),$$

$$\omega \in \Omega, \ \ \boldsymbol{b}_0 = (b_{01}, b_{02}, \cdots, b_{0D}), \ \ \boldsymbol{\phi} = (\phi_1, \phi_2, \cdots, \phi_D) \in \{0,1\}^D.$$

The estimation will be based on the labelled sample of independent observations from class ω, i.e. on $\boldsymbol{X}_\omega$ given in (2). The *a priori* probability of class ω is assumed to be known and for simplicity we denote it by $P(\omega)$.

The criterion we used for measuring the error resulting from approximating the true pdf $p^\star(\boldsymbol{x}|\omega)$ by $p(\boldsymbol{x}|W_\omega, B_\omega, \boldsymbol{\phi}, \boldsymbol{b}_0)$ is a mixture in the true proportions $P(\omega_1), \cdots, P(\omega_C)$ of the Kullback-Leibler distances between the true and the postulated class conditional pdfs of $\boldsymbol{x}$ (see Ku and Kullback [6], Maia and Fairhurst [7])

$$I(\boldsymbol{W}, \boldsymbol{B}, \boldsymbol{\phi}, \boldsymbol{b}_0) = \sum_{\omega \in \Omega} P(\omega) E_{p^\star} \{ \log \frac{p^\star(\boldsymbol{x}|\omega)}{p(\boldsymbol{x}|W_\omega, B_\omega, \boldsymbol{\phi}, \boldsymbol{b}_0)} \}, \tag{8}$$

where $\boldsymbol{W} = \{W_\omega, \ \omega \in \Omega\}$, $\boldsymbol{B} = \{B_\omega, \ \omega \in \Omega\}$.

It is known that minimizing $I(\boldsymbol{W}, \boldsymbol{B}, \boldsymbol{\phi}, \boldsymbol{b}_0)$ with respect to $\boldsymbol{W}, \boldsymbol{B}, \boldsymbol{\phi}$ and $\boldsymbol{b}_0$ is equivalent to maximizing

$$\sum_{\omega \in \Omega} P(\omega) E_{p^\star} \{ \log p(\boldsymbol{x}|W_\omega, B_\omega, \boldsymbol{\phi}, \boldsymbol{b}_0) \}, \tag{9}$$

which can be estimated directly from the samples $\boldsymbol{X}_\omega$ as

$$\sum_{\omega\in\Omega} P(\omega)\frac{1}{N_\omega}\sum_{\boldsymbol{x}\in\boldsymbol{X}_\omega}\log p(\boldsymbol{x}|W_\omega, B_\omega, \boldsymbol{\phi}, \boldsymbol{b}_0). \tag{10}$$

If we assume that the true pdf is from the family $p(\boldsymbol{x}|W_\omega, B_\omega, \boldsymbol{\phi}, \boldsymbol{b}_0)$ given by (6), then the expression (10) is the mixture of $1/N_\omega$ times the log-likelihood functions of $(W_\omega, B_\omega, \boldsymbol{\phi}, \boldsymbol{b}_0)$ generated by all the samples $\boldsymbol{X}_\omega$, $\omega \in \Omega$.

Let us denote

$$L(\boldsymbol{W}, \boldsymbol{B}, \boldsymbol{\phi}, \boldsymbol{b}_0) \;=\; \sum_{\omega\in\Omega}\frac{P(\omega)}{N_\omega}\sum_{\boldsymbol{x}\in\boldsymbol{X}_\omega}\log p(\boldsymbol{x}|W_\omega, B_\omega, \boldsymbol{\phi}, \boldsymbol{b}_0) \tag{11}$$

the log-likelihood function for $\boldsymbol{W}, \boldsymbol{B}, \boldsymbol{\phi}$ and $\boldsymbol{b}_0$. Then to determine the parameters of approximating mixtures (6), the ML estimates of parameters are found by maximization of corresponding log-likelihood function $L(\boldsymbol{W}, \boldsymbol{B}, \boldsymbol{\phi}, \boldsymbol{b}_0)$ with respect to parameters $\boldsymbol{W}, \boldsymbol{B}, \boldsymbol{\phi}$ and $\boldsymbol{b}_0$.

Unfortunately likelihood equations obtained by setting derivatives of function $L(\boldsymbol{W}, \boldsymbol{B}, \boldsymbol{\phi}, \boldsymbol{b}_0)$ to zero seem to have no explicit solution in case of mixtures. This difficulty arises because of the complex dependence of the likelihood function on the parameters to be estimated. Consequently, the alternative is to seek an approximate solution via EM algorithm (see Dempster [1], Redner and Walker [11]).

5 Feature Selection and Decision Rule Design

Our approach to the problem of selecting the subset of d features $X_d = \{x_{i_k} \mid k = 1, 2, \cdots, d; x_{i_k} \in X\}$ from the set $X = \{x_j \mid j = 1, 2, \cdots, D\}$ of D possible features representing the pattern, $d < D$, is transformed to the problem of choosing that vector $\hat{\boldsymbol{\phi}}_d$ which satisfies

$$J(\hat{\boldsymbol{\phi}}_d) = \min_{\boldsymbol{W}, \boldsymbol{B}, \boldsymbol{b}_0, \boldsymbol{\phi}_d} I(\boldsymbol{W}, \boldsymbol{B}, \boldsymbol{\phi}_d, \boldsymbol{b}_0). \tag{12}$$

That is, we attempt to find vector $\hat{\boldsymbol{\phi}}_d$ which produces the set of approximations $p(\boldsymbol{x}|\hat{W}_\omega, \hat{B}_\omega, \hat{\boldsymbol{b}}_0, \hat{\boldsymbol{\phi}}_d)$ to the class conditional pdfs $p^\star(\boldsymbol{x}|\omega)$, $\omega \in \Omega$, which is best in the sense of minimizing the Kullback-Leibler distance defined in (8) with respect to $\boldsymbol{W}, \boldsymbol{B}, \boldsymbol{\phi}_d$ and $\boldsymbol{b}_0$. Given such approximations

$$p(\boldsymbol{x}|\hat{W}_\omega, \hat{B}_\omega, \hat{\boldsymbol{b}}_0, \hat{\boldsymbol{\phi}}_d) = \prod_{j=d+1}^{D} f(x_{i_j}|\hat{b}_{0i_j}) \sum_{m=1}^{M_\omega} \hat{w}_m^\omega \prod_{k=1}^{d} f(x_{i_k}|\hat{b}_{mi_k}^\omega), \tag{13}$$
$$\omega \in \Omega, \quad 1 \le i_k \le D,$$

we may classify the observation of $\boldsymbol{x}$ according to the pseudo-Bayes decision rule:

decide that $\boldsymbol{x}$ is from class ω_l if

$$P(\omega_l)\sum_{m=1}^{M_\omega}\hat{w}_m^{\omega_l}\prod_{k=1}^{d} f(x_{i_k}|\hat{b}_{mi_k}^{\omega_l}) > P(\omega_j)\sum_{m=1}^{M_\omega}\hat{w}_m^{\omega_j}\prod_{k=1}^{d} f(x_{i_k}|\hat{b}_{mi_k}^{\omega_j}), \tag{14}$$
$$l \ne j, \quad l, j = 1, 2, \cdots, C, \quad 1 \le i_k \le D.$$

An important characteristics of this approach is that it effectively partitions the set X of all D features into two disjunct subsets X_d and $X - X_d$, where the

features from $X - X_d$ are common to all the classes and constitute the background distribution, as opposed to features $x_{i_1}, \cdots, x_{i_d}$, forming X_d, which are significant for discriminating the classes and constitute the "specific" distribution defined in (5). According to these features alone a new pattern $\boldsymbol{x}$ is classified into one of C classes and under this partition of the feature set X the Kullback-Leibler distance (8) is minimized.

As a result, the decision making based on the Bayes decision rule can be replaced in this case by the pseudo-Bayes decision rule defined by (14).

Using the EM algorithm we obtain the following algorithm for the proposed approach to feature selection and classifier design:

Step 1: Given the parameters $\boldsymbol{W}$, $\boldsymbol{B}$, $\boldsymbol{\phi}_d$ and $\boldsymbol{b}_0$ compute the weights $p(m|\boldsymbol{x},\omega)$ and $v(\boldsymbol{x}|m,\omega)$, $m = 1, 2, \cdots, M_\omega$, $\boldsymbol{x} \in \boldsymbol{X}_\omega$, $\omega \in \Omega$ according to

$$p(m|\boldsymbol{x},\omega) = \frac{w_m^\omega F(\boldsymbol{x}|\boldsymbol{b}_m^\omega, \boldsymbol{\phi}, \boldsymbol{b}_0)}{\sum_{j=1}^{M_\omega} w_j^\omega F(\boldsymbol{x}|\boldsymbol{b}_j^\omega, \boldsymbol{\phi}, \boldsymbol{b}_0)}, \tag{15}$$

$$v(\boldsymbol{x}|m,\omega) = \frac{p(m|\boldsymbol{x},\omega)}{\sum_{\boldsymbol{y}\in\boldsymbol{X}_\omega} p(m|\boldsymbol{y},\omega)}, \tag{16}$$

respectively.

Step 2: Under fixed weights (15) and (16) compute new values $\hat{\boldsymbol{W}}$ of $\boldsymbol{W}$ and $\hat{\boldsymbol{B}}$ of $\boldsymbol{B}$ by the formulas

$$\hat{w}_m^\omega = \frac{1}{N_\omega} \sum_{\boldsymbol{x}\in\boldsymbol{X}_\omega} p(m|\boldsymbol{x},\omega), \tag{17}$$

$$\hat{b}_{mi}^\omega = \arg\max_{b\in\mathcal{B}} \Big\{ \sum_{\boldsymbol{x}\in\boldsymbol{X}_\omega} v(\boldsymbol{x}|m,\omega) \log f(x_i|b) \Big\}, \; i = 1, 2, \cdots, D, \tag{18}$$

respectively.

Step 3: Given the parameters $\hat{\boldsymbol{W}}$, $\hat{\boldsymbol{B}}$ and $\boldsymbol{\phi}_d$ compute the new value $\hat{\boldsymbol{b}}_0$ of $\boldsymbol{b}_0$ by the formula

$$\hat{b}_{0i} = \arg\max_{b\in\mathcal{B}} \Big\{ \sum_{\omega\in\Omega} P(\omega) \sum_{m=1}^{M_\omega} \hat{w}_m^\omega \sum_{\boldsymbol{x}\in\boldsymbol{X}_\omega} v(\boldsymbol{x}|m,\omega) \log f(x_i|b) \Big\}, \; i = 1, \cdots, D. \tag{19}$$

If $\hat{\boldsymbol{W}} \neq \boldsymbol{W}$, $\hat{\boldsymbol{B}} \neq \boldsymbol{B}$, $\hat{\boldsymbol{b}}_0 \neq \boldsymbol{b}_0$ continue by Step 1 using the new parameters $\hat{\boldsymbol{W}}$, $\hat{\boldsymbol{B}}$ and $\hat{\boldsymbol{b}}_0$. Otherwise continue by Step 4.

Step 4: Using the parameters $\hat{\boldsymbol{W}}$, $\hat{\boldsymbol{B}}$, $\hat{\boldsymbol{b}}_0$ and the weights (16) compute the quantities $\hat{q}_{mi}^\omega$ and $\hat{Q}_i$, $m = 1, 2, \cdots, M_\omega$, $i = 1, 2, \cdots, D$, $\omega \in \Omega$ according to formulas

$$\hat{q}_{mi}^\omega = P(\omega)\hat{w}_m^\omega \sum_{\boldsymbol{x}\in\boldsymbol{X}_\omega} v(\boldsymbol{x}|m,\omega) \log \frac{f(x_i|\hat{b}_{mi}^\omega)}{f(x_i|\hat{b}_{0i})}, \tag{20}$$

$$\hat{Q}_i = \sum_{\omega\in\Omega} \sum_{m=1}^{M_\omega} \hat{q}_{mi}^\omega, \tag{21}$$

respectively. Rank $\hat{Q}_i$ so that

$$\hat{Q}_{i_1} \geq \hat{Q}_{i_2} \geq \cdots \geq \hat{Q}_{i_d} \geq \cdots \geq \hat{Q}_{i_D}, \tag{22}$$

and define for a given d

$$\hat{\phi}_{i_k} = \begin{cases} 1 \text{ for } k = 1, 2, \cdots, d, \\ 0 \text{ for } k = d+1, \cdots, D, \quad 1 \leq i_k \leq D. \end{cases}$$

If $\hat{\phi}_d \neq \phi_d$ then continue by Step 1 with $\hat{\boldsymbol{W}} = \boldsymbol{W}$, $\hat{\boldsymbol{B}} = \boldsymbol{B}$, $\hat{b}_0 = b_0$ and $\hat{\phi}_d = \phi_d$, else terminate the algorithm.

Note that in order to initialize the algorithm we should set $\phi_i = 1$ for all $i = 1, 2, \cdots, D$, i.e. $\phi = (1, 1, \cdots, 1)$. This follows both from theoretical considerations and computational reasons as the choice results in a quicker convergence of the algorithm.

It was shown in [10] that to find the vector $\hat{\phi}_d$ satisfying (12) is equivalent to finding the vector $\hat{\phi}_d$ that maximizes the criterion

$$Q(\phi) = \sum_{i=1}^{D} \phi_i \hat{Q}_i, \tag{23}$$

where $\hat{Q}_i$ is as in (21). If we rank $\hat{Q}_i$ in their descending order, as in (22), then vector $\hat{\phi}_d$ defined in Step 4 maximizes criterion (23) with respect to any other vector ϕ_d, and therefore it maximizes also criterion (8).

As it follows from the described algorithm, the problem to find a subset of d features $X_d = \{x_{i_k} \mid k = 1, 2, \cdots, d;\ 1 \leq i_k \leq D\}$ is reduced to the problem of finding a vector $\hat{\phi}_d$ for which the criterion (23) is maximized with respect to any other vector ϕ_d.

Therefore, a distinctive characteristics of our approach to feature selection is that only the operation of ranking according to (22) is required, without any search procedure, in order to obtain a required subset of d features.

It should also be noted that no knowledge of the functional form of the class conditional densities is required since it is assumed that they have the form of a mixture of densities of a special type. This assumption makes the proposed approach somewhat more realistic than the other parametric approaches. The consequence of this assumption is that it is particularly useful for the case of multimodal distributions when other feature selection methods based on distance measures (e.g. Mahalanobis distance, Bhattacharyya distance) would totally fail to provide reasonable results (see [10]). The reason is that the use of these measures incorporates the additional information that pattern vectors are multivariate Gaussian, which may not be true.

6 Experiments

A number of experiments have been conducted on real data arising from a problem concerned with the clasification of granite textures in the context of a multimedia application. In all experiments equal *a priori* class probabilities were considered and separate training and test sets were used.

A number of different images have been tested but the two colour images, the reproduction of which is shown in Fig.1, were specifically chosen since they are not well separated in the measurement space.

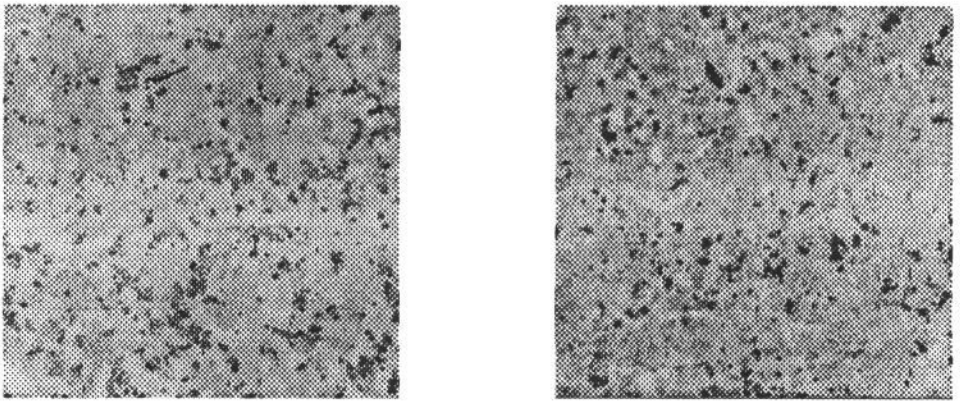

Figure 1: The two colour images used in the experiment
(a) bianco castilla (b) rosa bavenao

Each colour image was divided into two halves, the top half was used for training and the bottom half was used for classification purpose. The size of each image was 256 × 256 from which sample sub-images were selected with window of size 100 × 100 randomly placed. Then a 26-dimensional feature vector was extracted from each sub-image where the first 8 features are texture features and the remaining 18 features are colour features. The texture features were derived from the discrete cosine transform (DCT) filter of size 3 × 3 (see [12]). The colour features were gathered from the 3-dimensional histogram model of the colour texture from which the statistical description in the form of energy, entropy, local homogeneity, inertia, mean and variance were used. The sample size for both training and test sets were 1000.

The error rate of the pseudo-Bayes classifier (14) has been compared with the error rate of the Bayes classifier. Because of the high dimensionality of the original measurements of this particular experiment, when the distributions are assumed to be Gaussian, the sequential forward floating selection (SFFS) method (see Pudil *et al.* [8]) has been used to select features with the Bhattacharyya distance for multivariate normal distribution as the feature selection criterion. The SFFS algorithm has been shown to give practically the same results as the branch and bound algorithm (see [9]). The results of the classification with various sizes of feature set are depicted in Table 1.

Assumption about mixture	Pe(X_6)	Pe(X_{10})	Pe(X_{14})	Pe(X_{18})	Pe(X_{26})
2 components	0.1	0.064	0.03	0.03	0.03
3 components	0.065	0.028	0.016	0.007	0.007
4 components	0.055	0.012	0.007	0.007	0.007
Multivariate normal	0.235	0.287	0.169	0.169	0.169

Table 1: *Classifier performance on different feature subset size of image data.*

It is apparent that the results achieved by our approach are much better than those obtained by the Gaussian classifier when a multivariate normal distribution is implicitly assumed. As far as feature selection is concerned, our approach works very well since many redundant features have been detected especially with the mixture of 4 components. The higher the number of the mixture components, the better the results of feature subsets of the same size which can be clearly seen with the subsets of 10 and 14 features.

7 Conclusions

We have developed a method capable of automatic learning of decision rule based on approximating the unknown distributions by a finite mixture of the densities of a product type using the EM algorithm. The method at the same time facilitates the selection of the most important features for the classifier. The empirical results demonstrate that our approach can be superior to the method using a multivariate normal model. A higher number of mixture components is expected to perform better because of the larger flexibility in fitting the empirical pdf. Note that the training set has to be of reasonable size compared to the number of parameters of the mixture components to be estimated.

In conclusion, we have shown that it is possible to perform classification and feature selection even if neither the class conditional distributions nor their functional forms are known. The proposed method has been shown to outperform conventional methods which rely on the assumption of normality for the class distributions involved.

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.*, 39:1–38, 1977.

[2] P. A. Devijver and J. Kittler. *Pattern Recognition: A Statistical Approach.* Prentice-Hall, 1982.

[3] K. Fukunaga. *Introduction to Statistical Pattern Recognition: 2nd edition.* Academic Press, Inc., 1990.

[4] J. Grim. Multivariate statistical pattern recognition with nonreduced dimensionality. *Kybernetika*, 22(2):142–157, 1986.

[5] L. Kanal. Patterns in pattern recognition:1968-1974. *IEEE Transactions on Information Theory*, IT-20(6):697–722, November 1974.

[6] H. H. Ku and S. Kullback. Approximating discrete probability distributions. *IEEE Transactions on Information Theory*, IT-15:444–447, July 1969.

[7] M. A. G. Mattoso Maia and M. C. Fairhurst. On the use of I-Divergence for Generating Distribution Approximations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(6):661–664, November 1983.

[8] P. Pudil, J. Novovičová, and S. Bláha. Statistical approach to pattern recognition: Theory and practical solution by means of PREDITAS system. *Kybernetika*, 27:Supplement, 1–78, 1991.

[9] P. Pudil, J. Novovičová, N. Choakjarernwanit, and J. Kittler. A comparative evaluation of floating search methods for feature selection. Technical Report VSSP-TR-5/92, University of Surrey, U.K, December 1992.

[10] P. Pudil, J. Novovičová, N. Choakjarernwanit, and J. Kittler. Feature selection based on the approximation of class densities by finite mixtures of special type. *Pattern Recognition*, submitted for publication.

[11] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM J. Appl. Math.*, 26(2):195–239, April 1984.

[12] T.S.C. Tan and J. Kittler. Colour texture classification using features from colour histograms. *Proc. of the* 8^{th} *Scandinavian Conference on Image Analysis*, Tromso, Norway, 1993.