Divergence Based Feature Selection for Multimodal Class Densities

Jana Novovicová, Pavel Pudil, and Josef Kittler

Abstract—A new feature selection procedure based on the Kullback J-divergence between two class conditional density functions approximated by a finite mixture of parameterized densities of a special type is presented. This procedure is suitable especially for multimodal data. Apart from finding a feature subset of any cardinality without involving any search procedure, it also simultaneously yields a pseudo-Bayes decision rule. Its performance is tested on real data.

Index Terms—Feature selection, feature ordering, mixture distribution, maximum likelihood, EM algorithm, Kullback J-divergence.

1 Introduction

In practice, when designing a pattern recognition system we often encounter situations when even the form underlying the class conditional probability distribution of the pattern is unknown. Such a problem arises in a number of real pattern recognition problems when we have the data but no other information. Apart from the rather trivial cases when the data is governed by a simple distribution which can be found from it, untill recently there seemed not to exist any direct method for selecting a good subset of features. The options left in such a case were not satisfactory ([9], [10]), becoming computationally unfeasible for problems of realistic dimensionality.

An attempt to overcome the stated difficulties was made in our preceding work (see Pudil et al. [9], [10]) where we proposed a feature selection method based on approximating the uknown class conditional probability density functions (PDFs) in the sense of minimizing the Kullback-Leibler distances between the true and the postulated densities.

However, it should be stressed that the primary goal of that approach was not to select the most discriminative features but rather the features which minimize the adopted criterion of approximation error. When features are found which are best from the point of view of approximating the unknown class distributions, we can hope that they will be good for discriminating between the classes as well. Obviously, this premise is not always substantiated. Consequently, though the method has yielded very good results in a number of applications, its performance on a problem involving correlated data was not quite satisfactory [9].

Motivated by the need to improve the performance of the feature selection process by employing a criterion more directly linked to the concept of separability between classes, a new approach to feature selection is developed in the paper. The method is based on the Kullback J-divergence between class conditional PDFs approximated by finite mixtures of parameterized densities. The maximum likelihood (ML) estimates of the unknown parameters of the postulated class conditional PDFs are computed by the expectation-maximization (EM) algorithm (see Dempster et al. [2]). The proposed approach is especially suitable for multimodal data and is restricted to two classes.

To date the use of probabilistic separability measures has been confined to the cases when the class conditional PDFs belong to a family of special parametric PDFs for which an analytical solution can be found. The method presented in this paper extends the usability of divergence criterion for feature selection to the general case of PDFs of unknown form. Furthermore, while the new method complements the recent method based on approximation ([9]) in selecting features according to their discriminatory potential more directly, it preserves all its advantages. Particularly, besides yielding the feature subset of required dimensionality without any search procedure, it also provides a pseudo-Bayes decision rule. Consequently, the problems of feature selection and classifier design are solved simultaneously.

2 Parametric Model Based on Finite Mixture

Following the statistical approach to pattern recognition, we assume that a pattern described by a real D-dimensional vector $x = (x_1, x_2, \cdots, x_D)^T \in \mathcal{X} \subset \mathcal{R}^D$ 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 $p^*(x \mid \omega)$ and the respective a priori probabilities $P^*(\omega)$. Vector x can be then optimally classified using the Bayes minimum error rule based on the knowledge of the components $p^*(x \mid \omega)P^*(\omega)$, $\omega \in \Omega$ of the unconditional PDF $p^*(x)$. 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 labeled samples: $\mathbf{X}_\omega = \{\mathbf{x}_1^\omega, \mathbf{x}_2^\omega, \cdots, \mathbf{x}_{N_\omega}^\omega\}$, $\mathbf{x}_k^\omega \in \mathcal{X} \subset \mathcal{R}^D$, $k = 1, \cdots, N_\omega$, $\omega \in \Omega$, N_ω is the number of samples from class ω.

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 often unsatisfactory.

On the other hand, when no underlying structure is assumed, nonparametric methods of estimating the class conditional PDFs which do not require any prior knowledge about the forms of these PDFs must be resorted to. However, the use of nonparametric methods gives rise to a diferent type of problems. The most common problem encountered is the excessive storage requirement for the multidimensional probability function involved (see Devijver and Kittler [3], Duda and Hart [4] for more details). Further problems concerning the kernel-based estimators are the choice of the functional form of the kernel and the choice of the window width (see Jain [7]).

In this paper we adopt an alternative approach which in terms of the required computer storage is considerably more efficient than nonparametric PDF estimation methods but at the same time it retains the capacity to reflect the local structure of the distribu-

- *J. Novovicová and P. Pudil are with the Institute of Information Theory and Automation Academy of Sciences of the Czech Republic, Pod vodarenskou vezi 4 182 08, Prague 8, Czech Republic. E-mail: {novovic, pudil}@utia.cas.cz.*
- *J. Kittler is with the Department of Electronic and Electrical Engineering, University of Surrey, Guildford Surrey GU2 5XH, United Kingdom. E-mail: j.kittler@ee.surrey.ac.uk.*

Manuscript received May 3, 1994; revised Sept. 1, 1995. Recommended for acceptance by R. Duin.
For information on obtaining reprints of this article, please send e-mail to: transactions@computer.org, and reference IEEECS Log Number P95152.

tion. In this sense it is superior to the kernel PDF estimation technique. The key idea is to use a mixture of large number of parametric components of a special type. The simplicity of the basic building block of the mixture must be compensated for by numbers. Nevertheless, we shall see that confining the kernel to a simple form has several important advantages:

1) ease of optimization,
2) local sensitivity,
3) it facilitates feature selection which is of primary consideration in our work,
4) despite the increased number of components, it results in an improved overall memory space efficiency (the number of estimated parameters grows slower than the number of parameters of more complex kernels)

In our approach the following parametric model is postulated for the ωth class conditional density function:

$$p(\mathbf{x}|\omega) = \sum_{m=1}^{M_\omega} \alpha_m^\omega p_m(\mathbf{x}|\omega) = \sum_{m=1}^{M_\omega} \alpha_m^\omega g_0(\mathbf{x}|\mathbf{b}_0) g(\mathbf{x}|\mathbf{b}_m^\omega, \mathbf{b}_0, \Phi), \ \mathbf{x} \in \mathcal{X} \quad (1)$$

where α_m^ω is a nonnegative mixing weight, $\sum_{m=1}^{M_\omega} \alpha_m^\omega = 1$, and M_ω is the number of mixture components. Each component density $p_m(\mathbf{x} \mid \omega)$ of this finite mixture includes a background PDF g_0, common to all classes, which is an important distinction from the kernel approach, and a function g of the form

$$g_0(\mathbf{x}|\mathbf{b}_0) = \prod_{i=1}^{D} f(x_i|b_{0i}) \quad (2)$$

$$g(\mathbf{x}|\mathbf{b}_m^\omega, \mathbf{b}_0, \Phi) = \prod_{i=1}^{D} \left[\frac{f(x_i|b_{mi}^\omega)}{f(x_i|b_{0i})} \right]^{\phi_i}, \ \phi_i \in \{0,1\},$$

$$\mathbf{b}_0 = (b_{01}, b_{02}, \cdots, b_{0D}), \ \mathbf{b}_m^\omega = (b_{m1}^\omega, b_{m2}^\omega, \cdots, b_{mD}^\omega) \in \mathcal{B}^D, \quad (3)$$

$$\Phi = (\phi_1, \phi_2, \cdots, \phi_D) \in \{0,1\}^D$$

where $\mathbf{b}_0$, $\mathbf{b}_m^\omega$, and Φ are the parameter vectors. The function g is actually defined on a subspace

$$\mathcal{X}_{(l)} \subset \mathcal{R}^l : \mathcal{X}_{(l)} = \mathcal{X}_{i_1} \times \mathcal{X}_{i_2} \times \cdots \times \mathcal{X}_{i_l}, \mathcal{X}_{i_k} \subset \mathcal{R},$$
$$1 \le i_k \le D, \ k = 1, \cdots, l$$

specified by nonzero binary parameters ϕ_{i_k}.

The univariate function f is assumed to be from a parametric family of PDFs parameterized by $b \in \mathcal{B}$. For any choice of the binary parameters ϕ_i, which can be looked upon as *control variables*, the finite mixture (1) can be rewritten as

$$p(\mathbf{x}|A_\omega, B_\omega, \mathbf{b}_0, \Phi) = \sum_{m=1}^{M_\omega} \alpha_m^\omega \prod_{i=1}^{D} \left[f(x_i|b_{0i})^{1-\phi_i} f(x_i|b_{mi}^\omega)^{\phi_i} \right].$$
$$A_\omega = (\alpha_1^\omega, \alpha_2^\omega, \cdots, \alpha_{M_\omega}^\omega), \quad B_\omega = (\mathbf{b}_1^\omega, \mathbf{b}_2^\omega, \cdots, \mathbf{b}_{M_\omega}^\omega). \quad (4)$$

Model (1) is a particular case of the parametric model proposed by Grim [6]. It will be seen later that as a result of approximating the unknown conditional PDFs with the model (4) the process of feature selection becomes a very simple task.

3 ML Estimation of Parameters Using EM Algorithm

We consider the ML estimation of all the uknown parameters $\mathbf{A} = \{A_\omega, \ \omega \in \Omega\}$, $\mathbf{B} = \{B_\omega, \ \omega \in \Omega\}$, and $\mathbf{b}_0$ in the parametric families $\{p(\mathbf{x} \mid A_\omega, B_\omega, \mathbf{b}_0, \Phi)\}$. The estimation will be based on the labeled sample of independent observations from class ω, i.e., on $\mathbf{X}_\omega$. The a priori probability of class ω is assumed to be known and for simplicity we denote it by $P(\omega)$.

The log-likelihood function for $\mathbf{A}$, $\mathbf{B}$, $\mathbf{b}_0$, and Φ is given under model (4) by

$$L(\mathbf{A}, \mathbf{B}, \mathbf{b}_0, \Phi) = \sum_{\omega \in \Omega} \frac{P(\omega)}{N_\omega} \sum_{\mathbf{x} \in \mathbf{X}_\omega} \log p(\mathbf{x}|A_\omega, B_\omega, \mathbf{b}_0, \Phi), \quad (5)$$

where "log" denotes the natural logarithm. Estimates of uknown parameters can be obtained as a solution of the log-likelihood equation by an iterative procedure via EM algorithm (see Dempster et al. [2], Redner and Walker [11]).

We shall use the EM algorithm to maximize the log-likelihood function (5) with respect to the parameters $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{b}_0$ *under given* Φ. The two steps of the EM algorithm of Dempster et al. [2], are specified in our case in the following: Given a current approximation $({}^c\mathbf{A}, {}^c\mathbf{B}, {}^c\mathbf{b}_0)$ of a maximizer of $L(\mathbf{A}, \mathbf{B}, \mathbf{b}_0, \Phi)$ obtain the next approximation $(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{b}}_0)$ as follows:

- **E-step:** Determine for Φ and ${}^c\Phi$ the conditional expectation

$$\mathcal{L}(\Theta, \Phi | {}^c\Theta, {}^c\Phi) = \sum_{\omega \in \Omega} \frac{P(\omega)}{N_\omega} \sum_{\mathbf{x} \in \mathbf{X}_\omega} \left\{ \sum_{m=1}^{M_\omega} p(m|\mathbf{x}, \omega) \log \left[\alpha_m^\omega g_0(\mathbf{x}|\mathbf{b}_0) g(\mathbf{x}|\mathbf{b}_m^\omega, \mathbf{b}_0, \Phi) \right] \right\}, \quad (6)$$

  where the abbreviation $\Theta = (\mathbf{A}, \mathbf{B}, \mathbf{b}_0)$ is used for simplicity and

$$p(m|\mathbf{x}, \omega) = \frac{{}^c\alpha_m^\omega g(\mathbf{x}|{}^c\mathbf{b}_m^\omega, {}^c\mathbf{b}_0, {}^c\Phi)}{\sum_{j=1}^{M_\omega} {}^c\alpha_j^\omega g(\mathbf{x}|{}^c\mathbf{b}_j^\omega, {}^c\mathbf{b}_0, {}^c\Phi)} \quad (7)$$
$$m = 1, 2, \cdots, M_\omega, \ \mathbf{x} \in \mathbf{X}_\omega, \ \omega \in \Omega.$$

- **M-step:** Choose $\hat{\Theta} = \arg\max_\Theta \{\mathcal{L}(\Theta, \Phi | {}^c\Theta, {}^c\Phi)\}$.

According to Redner and Walker [11], Grim [5], and Pudil et al. [9], for any fixed binary parameters ϕ_i and under fixed weights $v(\mathbf{x} \mid m, \omega)$ defined by

$$v(\mathbf{x}|m, \omega) = \frac{p(m|\mathbf{x}, \omega)}{\sum_{\mathbf{y} \in \mathbf{X}_\omega} p(m|\mathbf{y}, \omega)}, \quad (8)$$
$$m = 1, 2, \cdots, M_\omega, \ \mathbf{x} \in \mathbf{X}_\omega, \ \omega \in \Omega,$$

the function (6) is maximized by $\mathbf{A} = \hat{\mathbf{A}}$, $\mathbf{B} = \hat{\mathbf{B}}$, and $\mathbf{b}_0 = \hat{\mathbf{b}}_0$, where

$$\hat{\alpha}_m^\omega = \frac{1}{N_\omega} \sum_{\mathbf{x} \in \mathbf{X}_\omega} p(m|\mathbf{x}, \omega), \quad (9)$$

$$\hat{b}_{mi}^\omega = \arg\max_{b \in \mathcal{B}} \left\{ \sum_{\mathbf{x} \in \mathbf{X}_\omega} v(\mathbf{x}|m, \omega) \log f(x_i|b) \right\}, \quad (10)$$

$$\hat{b}_{0i} = \arg\max_{b \in \mathcal{B}} \left\{ \sum_{\omega \in \Omega} P(\omega) \sum_{m=1}^{M_\omega} \hat{\alpha}_m^\omega \sum_{\mathbf{x} \in \mathbf{X}_\omega} v(\mathbf{x}|m, \omega) \log f(x_i|b) \right\} \quad (11)$$
$$m = 1, 2, \cdots, M_\omega, \ i = 1, 2, \cdots, D, \ \omega \in \Omega.$$

Since the inequality $\mathcal{L}({}^c\Theta, \Phi | {}^c\Theta, {}^c\Phi) \le \mathcal{L}(\hat{\Theta}, \Phi | {}^c\Theta, {}^c\Phi)$ is satisfied, we have $L({}^c\mathbf{A}, {}^c\mathbf{B}, {}^c\mathbf{b}_0, {}^c\Phi) \le L(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{b}}_0, \Phi)$ ([11], [5]).

Like all other currently available estimation procedures, the EM algorithm does not guarantee convergence to a global maximum. Similarly the choice of the number of components in (4) influences only the quality of approximation. The frequently

discussed slow convergence of the EM algorithm in the final stages of computation is also of little importance since the corresponding changes of the criterion are usually negligible.

4 Approach to Feature Selection

In the following we shall concentrate on the two-class problem, i.e., $\Omega = \{\omega_1, \omega_2\}$.

Consider the Kullback J-divergence between two classes ω_1 and ω_2 given by (see Boekee and Van der Lubbe [1])

$$J^*(\omega_1;\omega_2) = \sum_{\omega\in\Omega} P(\omega) E_\omega \left\{ \log \frac{p^*(\mathbf{x}|\omega)}{p^*(\mathbf{x}|\Omega-\omega)} \right\}, \tag{12}$$

where E_ω denotes the mathematical expectation with respect to the class-conditional PDF $p^*(\mathbf{x}\mid\omega)$ and "$\Omega-\omega$" is the abbreviation either for ω_1 if $\omega=\omega_2$ or for ω_2 if $\omega=\omega_1$.

As this measure reflects the separability of the two classes we shall use it as a criterion for feature selection. Suppose that the PDF $p^*(\mathbf{x}\mid\omega)$ has the form $p(\mathbf{x}\mid\omega)$ defined in (4). Then $J^*(\omega_1;\omega_2)$ in (12) can be estimated from the sets $\mathbf{X}_\omega$ as

$$\bar{J}(\mathbf{A},\mathbf{B},\Phi) = \sum_{\omega\in\Omega} \frac{P(\omega)}{N_\omega} \sum_{\mathbf{x}\in\mathbf{X}_\omega} \log \frac{\sum_{m=1}^{M_\omega} \alpha_m^\omega g\left(\mathbf{x}\middle|\mathbf{b}_m^\omega, \mathbf{b}_0, \Phi\right)}{\sum_{m=1}^{M_{\Omega-\omega}} \alpha_m^{\Omega-\omega} g\left(\mathbf{x}\middle|\mathbf{b}_m^{\Omega-\omega}, \mathbf{b}_0, \Phi\right)} \tag{13}$$

It is easy to verify that by using the weights $p(m\mid\mathbf{x},\omega)$ given by (7) we can rewrite (13) in the form

$$\begin{aligned}\bar{J}(A,B,\Phi) = &\\ &+ \sum_{\omega\in\Omega} \frac{P(\omega)}{N_\omega} \sum_{\mathbf{x}\in\mathbf{X}_\omega} \left\{ \sum_{m=1}^{M_\omega} p(m|\mathbf{x},\omega) \log \frac{\alpha_m^\omega}{p(m|\mathbf{x},\omega)} \right.\\ &- \sum_{m=1}^{M_{\Omega-\omega}} p(m|\mathbf{x},\Omega-\omega) \log \frac{\alpha_m^{\Omega-\omega}}{p(m|\mathbf{x},\Omega-\omega)} \\ &+ \sum_{i=1}^{D} \phi_i \sum_{m=1}^{M_\omega} p(m|\mathbf{x},\omega) \log f\left(x_i\middle|b_{mi}^\omega\right) \\ &\left. 1 - \sum_{i=1}^{D} \phi_i \sum_{m=1}^{M_{\Omega-\omega}} p(m|\mathbf{x},\Omega-\omega) \log f\left(x_i\middle|b_{mi}^{\Omega-\omega}\right) \right\}.\end{aligned} \tag{14}$$

Denote

$$\hat{\alpha}_m^{\omega,\Omega-\omega} = \frac{1}{N_\omega} \sum_{\mathbf{x}\in X\omega} p(m|\mathbf{x},\Omega-\omega) \tag{15}$$

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

Equation (13) can be rewritten using (9), (8), (15), and (16) as

$$\begin{aligned}\bar{J}(\mathbf{A},\mathbf{B},\Phi) = &\\ &\sum_{\omega\in\Omega} P(\omega) \left\{ \sum_{m=1}^{M_\omega} \hat{\alpha}_m^\omega \log \alpha_m^\omega - \sum_{m=1}^{M_{\Omega-\omega}} \hat{\alpha}_m^{\omega,\Omega-\omega} \log \alpha_m^{\Omega-\omega} \right\} \\ &+ \sum_{i=1}^{D} \phi_i \sum_{\omega\in\Omega} P(\omega) \sum_{m=1}^{M_\omega} \hat{\alpha}_m^\omega \sum_{\mathbf{x}\in\mathbf{X}_\omega} v(\mathbf{x}|m,\omega) \log f\left(x_i\middle|b_{mi}^\omega\right) \\ &- \sum_{i=1}^{D} \phi_i \sum_{\omega\in\Omega} P(\omega) \sum_{m=1}^{M_{\Omega-\omega}} \hat{\alpha}_m^{\omega,\Omega-\omega} \sum_{\mathbf{x}\in\mathbf{X}_\omega} v^\omega(\mathbf{x}|m,\Omega-\omega) \log f\left(x_i\middle|b_{mi}^{\Omega-\omega}\right) \\ &- \sum_{\omega\in\Omega} \frac{P(\omega)}{N_\omega} \sum_{\mathbf{x}\in\mathbf{X}_\omega} \sum_{m=1}^{M_\omega} p(m|\mathbf{x},\omega) \log p(m|\mathbf{x},\omega) \\ &+ \sum_{\omega\in\Omega} \frac{P(\omega)}{N_\omega} \sum_{\mathbf{x}\in\mathbf{X}_\omega} \sum_{m=1}^{M_{\Omega-\omega}} p(m|\mathbf{x},\Omega-\omega) \log p(m|\mathbf{x},\Omega-\omega).\end{aligned} \tag{17}$$

Now substituting $\mathbf{A}=\hat{\mathbf{A}}$, $\mathbf{B}=\hat{\mathbf{B}}$ in (17) and introducing quantities

$$\begin{aligned}\hat{J}_i^\omega &= P(\omega) \left\{ \sum_{m=1}^{M_\omega} \hat{\alpha}_m^\omega \sum_{\mathbf{x}\in\mathbf{X}_\omega} v(\mathbf{x}|m,\omega) \log f\left(x_i\middle|\hat{b}_{mi}^\omega\right) \right.\\ &\left. - \sum_{m=1}^{M_{\Omega-\omega}} \hat{\alpha}_m^{\omega,\Omega-\omega} \sum_{\mathbf{x}\in\mathbf{X}_\omega} v^\omega(\mathbf{x}|m,\Omega-\omega) \log f\left(x_i\middle|\hat{b}_{mi}^{\Omega-\omega}\right) \right\} \\ \hat{J}_i &= \sum_{\omega\in\Omega} \hat{J}_i^\omega,\ i=1,2,\cdots,D,\ \omega\in\Omega,\end{aligned} \tag{18}$$

we obtain

$$\bar{J}\left(\hat{\mathbf{A}},\hat{\mathbf{B}},\Phi\right) = \bar{Q}\left(\hat{\mathbf{A}},\hat{\mathbf{B}}\right) + \sum_{i=1}^{D} \phi_i \hat{J}i \tag{19}$$

where only the second term on the right-hand side depends on Φ.

Rank $\hat{J}_i$ in their descending order

$$\left\{\hat{J}_{i_k}\right\}_{k=1}^{D},\ \hat{J}_{i_k} \geq \hat{J}_{i_{k+1}}. \tag{20}$$

and set

$$\hat{\phi}_{i_k} = \begin{cases} 1 \text{ for } k=1,2,\cdots,d, \\ 0 \text{ for } k=d+1,\cdots,D, \quad 1\leq i_k \leq D. \end{cases} \tag{21}$$

Then for vector $\hat{\Phi}_d$ of parameters $\hat{\phi}_i$ and for any vector Φ_d consisting of d 1s and $(D-d)$ 0s, the inequality

$$\bar{J}\left(\hat{\mathbf{A}},\hat{\mathbf{B}},\Phi_d\right) \leq \bar{J}\left(\hat{\mathbf{A}},\hat{\mathbf{B}},\hat{\Phi}_d\right) \tag{22}$$

is satisfied. Here and in the sequel Φ_d denotes the vector consisting of d 1s and $(D-d)$ 0s.

Our approach to the problem of selecting a 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 $\hat{\Phi}_d$ which satisfies

$$\bar{J}\left(\hat{\mathbf{A}},\hat{\mathbf{B}},\hat{\Phi}_d\right) = \max_{\Phi_d} \bar{J}\left(\hat{\mathbf{A}},\hat{\mathbf{B}},\Phi_d\right) \tag{23}$$

It follows from (19) that to find the vector $\hat{\Phi}_d$ satisfying (23) is equivalent to finding the vector $\hat{\Phi}_d$ for which the criterion

$$J(\Phi) = \sum_{i=1}^{D} \phi_i \hat{J}_i, \tag{24}$$

is maximized with respect to any other Φ_d for given $\hat{\mathbf{A}},\hat{\mathbf{B}}$, where $\hat{J}_i$ is defined as in (18). That is, we attempt to find vector $\hat{\Phi}_d$ which maximizes the estimation of the Kullback J-divergence between the approximations $p(\mathbf{x}\mid\hat{A}_\omega,\hat{B}_\omega,\hat{\mathbf{b}}_0,\Phi_d)$ and $p(\mathbf{x}\mid\hat{A}_{\Omega-\omega},\hat{B}_{\Omega-\omega},\hat{\mathbf{b}}_0,\Phi_d)$ of the class conditional PDFs $p^*(\mathbf{x}\mid\omega)$ and $p^*(\mathbf{x}\mid\Omega-\omega)$, respectively.

Using the results presented in Section 3 and in this section we propose the following algorithm for feature selection:

- Step 1: Given the parameters ${}^c\mathbf{A}$, ${}^c\mathbf{B}$, and ${}^c\Phi$ compute the weights $p(m\mid\mathbf{x},\omega)$ and $v(\mathbf{x}\mid m,\omega)$, $m=1,2,\cdots,M_\omega$, $\mathbf{x}\in\mathbf{X}_\omega$, $\omega\in\Omega$ according to (7) and (8), respectively.
- Step 2: Under fixed weights (7) and (8) compute new values $\hat{\mathbf{A}}$ of $\mathbf{A}$ and $\hat{\mathbf{B}}$ of $\mathbf{B}$ by (9) and (10), respectively.
  If $\hat{\mathbf{A}}\neq{}^c\mathbf{A}$, $\hat{\mathbf{B}}\neq{}^c\mathbf{B}$, continue by Step 1 using the new parameters $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$. Otherwise continue by Step 3.
- Step 3: Using the parameters $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$, and the weights (8), compute for the vector ${}^c\Phi$ the quantities $\hat{J}_i$, $i=1,2,\cdots,D$, accord-

ing to (18). Rank $\hat{J}_i$ so that $\hat{J}_{i_1} \geq \hat{J}_{i_2} \geq \cdots \geq \hat{J}_{i_d} \geq \cdots \geq \hat{J}_{i_D}$, and define $\hat{\Phi}_d$ for a given d according to (21).

If $\hat{\Phi}_d \neq {}^c\Phi$ then continue by Step 1 with new values $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$, and $\hat{\Phi}_d$, 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.

The proposed method possesses some unique properties which make it very useful in practice. First of all, a feature subset of a given cardinality d, where $d = 1, 2, \cdots, D$, is obtained immediately since a distinctive characteristics of our approach to feature selection is that only the operation of ranking of $\hat{J}_i$ is required, without any search procedure, in order to obtain a required subset of d features. A computationally time consuming search procedure usually associated with a selected criterion is not needed in our approach. Moreover, in practice we can get a near optimal ordering (often even an optimal one) of all the original D features after passing the Step 3 of the algorithm for the first time. Thus a near optimal feature subset of any cardinality d, where $d = 1, 2, \cdots, D$, can be obtained immediately. Subsequent computations according to the algorithm will further improve the model by adjusting all the parameters and thus increase the value of log-likelihood function, but they will not necessarily change the composition of the feature subset.

Given the approximations

$$p\left(\mathbf{x}\middle|\hat{A}_\omega, \hat{B}_\omega, \hat{\mathbf{b}}_0, \hat{\Phi}_d\right) = \prod_{j=d+1}^{D} f\left(x_{i_j}\middle|\hat{b}_{0i_j}\right) \sum_{m=1}^{M_\omega} \hat{\alpha}_m^\omega \prod_{k=1}^{d} f\left(x_{i_k}\middle|\hat{b}_{mi_k}^\omega\right), \quad \omega \in \Omega,\ 1 \leq i_k \leq D, \tag{25}$$

it can be easily seen that the background PDF g_0 may be reduced in the inequality in the Bayes decision rule. Thus we may classify the observation of $\mathbf{x}$ according to the pseudo-Bayes decision rule:

decide that x is from class ω_l if

$$P(\omega_l) \sum_{m=1}^{M_\omega} \hat{\alpha}_m^{\omega_l} \prod_{k=1}^{d} f\left(x_{i_k}\middle|\hat{b}_{mi_k}^{\omega_l}\right) = \max_{j=1,2}\left\{P(\omega_j) \sum_{m=1}^{M_\omega} \hat{\alpha}_m^{\omega_j} \prod_{k=1}^{d} f\left(x_{i_k}\middle|\hat{b}_{mi_k}^{\omega_j}\right)\right\}. \tag{26}$$

It means that a new pattern $\mathbf{x}$ is classified into one of two classes according to only d features $x_{i_1}, \cdots, x_{i_d}$.

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 joint distribution of the features from $X - X_d$ is common to both the classes and constitutes 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 the joint distribution of these features constitutes the "specific" distribution defined in (3). According to these features alone a new pattern $\mathbf{x}$ is classified into one of two classes.

4.1 Application to a Particular Type of Mixtures

If (2) and (3) are of the form

$$g_0\left(\mathbf{x}\middle|\mu_0, \sigma_0\right) = \prod_{i=1}^{D}\left[\frac{1}{\sqrt{2\pi}\sigma_{0i}} \exp\left\{-\frac{1}{2}\left(\frac{x_i - \mu_{0i}}{\sigma_{0i}}\right)^2\right\}\right] \quad \mu_{0i} \in \mathcal{R},\ \sigma_{0i} \in (0, \infty), \tag{27}$$

$$g\left(\mathbf{x}\middle|\mu_m^\omega, \sigma_m^\omega, \mu_0, \sigma_0, \Phi\right) = \prod_{i=1}^{D}\left[\frac{\sigma_{0i}}{\sigma_{mi}^\omega} \exp\left\{-\frac{1}{2}\left(\frac{x_i - \mu_{mi}^\omega}{\sigma_{mi}^\omega}\right)^2 + \frac{1}{2}\left(\frac{x_i - \mu_{0i}}{\sigma_{0i}}\right)^2\right\}\right]^{\phi_i} \tag{28}$$

$$\mu_{mi}^\omega \in \mathcal{R},\ \sigma_{mi}^\omega \in (0, \infty),\ m = 1, 2, \cdots M_\omega,\ \omega \in \Omega,$$

then (18) can be simplified as follows

$$\hat{J}_i^\omega = \frac{1}{2} P(\omega)\left\{\sum_{m=1}^{M_\Omega} \hat{\alpha}_m^\omega \log\left(\frac{1}{\hat{\sigma}_{mi}^\omega}\right)^2 - 1 + \sum_{m=1}^{M_{\Omega-\omega}} \hat{\alpha}_m^{\omega,\Omega-\omega}\left[\log\left(\hat{\sigma}_{mi}^{\Omega-\omega}\right)^2 + \left(\frac{\hat{\sigma}_{mi}^{\Omega-\omega}}{\hat{\sigma}_{mi}^{\omega,\Omega-\omega}}\right)^2 + \left(\frac{\hat{\mu}_{mi}^{\omega,\Omega-\omega} - \hat{\mu}_{mi}^{\Omega-\omega}}{\hat{\sigma}_{mi}^{\Omega-\omega}}\right)^2\right]\right\}, \tag{29}$$

where

$$\hat{\mu}_{mi}^\omega = \sum_{\mathbf{x} \in \mathbf{X}_\omega} x_i v\left(\mathbf{x}\middle|m, \omega\right), \qquad \left(\hat{\sigma}_{mi}^\omega\right)^2 = \sum_{\mathbf{x} \in \mathbf{X}_\omega} \left(x_i - \hat{\mu}_{mi}^\omega\right)^2 v\left(\mathbf{x}\middle|m, \omega\right), \tag{30}$$

and

$$\hat{\mu}_{mi}^{\omega,\Omega-\omega} = \sum_{\mathbf{x} \in \mathbf{X}_\omega} x_i v^\omega\left(\mathbf{x}\middle|m, \Omega - \omega\right), \qquad \left(\hat{\sigma}_{mi}^{\omega,\Omega-\omega}\right)^2 = \sum_{\mathbf{x} \in \mathbf{X}_\omega} \left(x_i - \hat{\mu}_{mi}^{\omega,\Omega-\omega}\right)^2 v^\omega\left(\mathbf{x}\middle|m, \Omega - \omega\right). \tag{31}$$

5 Results of Experiments

The advocated method was tested on synthetic and real data though only the results achieved on real data from texture classification and speech recognition are described in the sequel. Its performance is compared with that of the ordinary multivariate normal model and in the case of speech data also with that of the "approximation model" [9], [10]. Owing to different assumptions about the underlying probability distributions, the comparison has been made using the misclassification rate obtained from both types of classifiers, i.e., classifier based on the mixture of normal distribution (the pseudo-Bayes classifier) and the classifier based on multivariate normal distribution (the Gausian classifier). Separate training and test sets were used in all experiments. In addition, the a priori probabilities in all experiments were taken to be equal for all classes.

5.1 Texture Data

A number of different images have been tested in this experiment. Two color images (described in more detail in [9]) were specifically chosen since they are not well separated in the measurement space. They are derived from specimens of certain types of marble. From 26 features extracted altogether, eight were texture features and the remaining 18 color features. The sample size for both training and test sets were 1,000. Because of the high dimensionality of the measurements, when the distributions were assumed to be Gaussian, the sequential forward floating selection (SFFS) method was used to select features. This method has been shown to achieve quasi optimal performance at low computational costs [8]. The results of classification for various sizes of the feature set are depicted in Table 1. Here and in the sequel $Pe(X_d)$ denotes the error estimation of the classification based on feature subset X_d.

TABLE 1
PERFORMANCE OF PSEUDO-BAYES AND GAUSSIAN CLASSIFIERS FOR DIFFERENT FEATURE SUBSET SIZE OF IMAGE DATA

Model	$Pe(X_6)$	$Pe(X_{10})$	$Pe(X_{14})$	$Pe(X_{18})$	$Pe(X_{22})$	$Pe(X_{26})$
Mixture of 2 components	0.097	0.062	0.030	0.030	0.030	0.030
Mixture of 3 components	0.063	0.027	0.016	0.007	0.007	0.007
Mixture of 4 components	0.056	0.011	0.007	0.007	0.007	0.007
Multivariate normal	0.235	0.287	0.169	0.169	0.169	0.169

From Table 1, it is obvious that the assumption that the distributions are unimodal Gaussian is not appropriate. In contrast, when a mixture of normal densities is used, a much smaller error rate is obtained. As far as feature selection is concerned, our approach works very well since a substantial dimensionality reduction can be made without significantly deteriorating the classifier performance. In other words, redundant features have been detected especially with the mixture of four components. The higher the number of the mixture components, the better the results for the feature subsets of the same size. This can be clearly seen for the subsets of 10 and 14 features.

5.2 Speech Data

The data used to train, test and compare the proposed method was a set of 1,418 pattern vectors of the utterances "YES" and "NO" spoken over the public switched telephone network. Each 15-dimensional feature vector contained five segments of three features derived by low order linear prediction analysis. From this set, 798 samples were used for training and the remaining 620 samples to form the test set. Both sets contain nearly equal number of samples for each pattern class. The data was supplied by British Telecom. The results of the experiment using are shown in Fig. 1.

As reported already in [9], a detailed analysis has shown that the data lies in narrow regions along parallel hyperplanes, i.e., the data are highly correlated and unimodal. It is therefore pertinent to use the Gaussian classifier in this case. To approximate these distributions by a mixture of independent Gaussian distributions properly would require a higher number of mixture components. However, owing to a small sample size of the training set, this is not possible. For this reason the feature selection method based on PDF approximation [9] failed to provide good results in this case, as we can see from Fig. 1. In both the approaches the mixture model of two, three, four, and five components has been used.

The fact that despite the above mentioned conditions the presented new approach based on divergence provided the results comparable or even better than the "normal" approach, indicate its great potential. Note again that the "normal" approach consisted of selecting features by the sequential forward floating search method (giving in this case the identical results to the branch and bound method) and using the Gaussian classifier.

From Fig. 1 it can be seen that even in the case of divergence criterion the mixture of two independent normal densities was not sufficient to accurately approximate the unknown distributions. This insufficiency is apparent from the classification error rate which was approximately twice as high than in the case of more components. With the mixture of three, four, and five components the comparable (for five components even better) results to that of the multivariate normal model have been obtained.

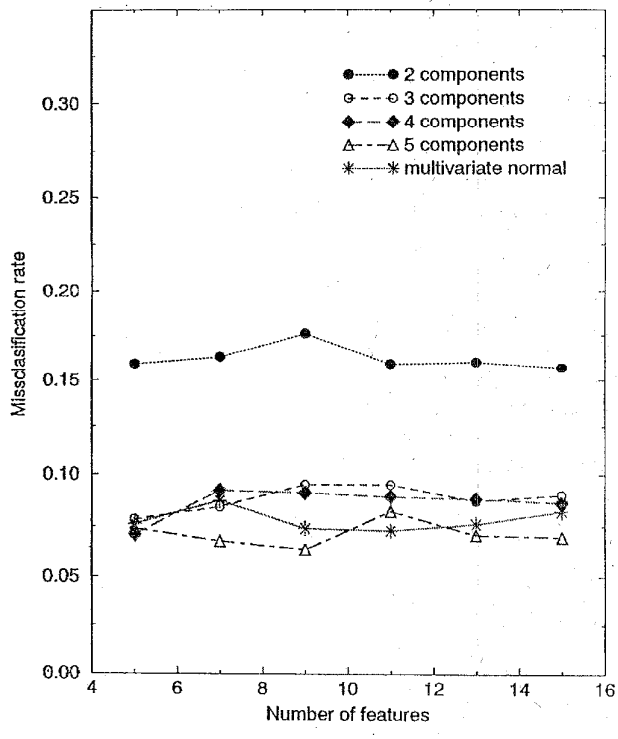

Fig. 1. Performance of pseudo-Bayes classifier.

6 CONCLUSIONS

A novel feature selection method, based on approximating the unknown class conditional distributions by finite mixtures of parameterized densities of a special type using the Kullback J-divergence as the criterion of optimality, has been developed. It is distinguished from an earlier attempt of PDF approximation [9]} in the sense that it is more directly aimed at selecting a subset of features with the highest possible discriminative potential.

The method described in this paper extends the usability of divergence criterion to the general case of PDFs of unknown form. The idea is to model these densities by a mixture of parametric densities of special type. This approach is quite realistic and 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.

As the comparison with the results achieved by "multivariate normal" approach and "approximation" approach demonstrate, the new method is more robust with respect to the form of class conditional PDFs. While the former one fails in the case of multimodal distributions and on the other hand the performance of the latter one is not too good in the case of unimodal PDFs (particularly with smaller training set sample size), the method based on divergence yields very good results in both the cases.

ACKNOWLEDGMENTS

This research was supported by the Scientific Engineering Research Council under grant GR/E 97549 and by the Czech Academy of Sciences under grant No. 275107.

REFERENCES

[1] D.E. Boekee and J.C.A. Van der Lubbe, "Some aspects of error bounds in feature selection," *Pattern Recognition*, vol. 11, pp. 353-360, 1979.
[2] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society*, vol. 39, pp. 1-38, 1977.
[3] P.A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. Englewood Cliffs, N.J.: Prentice Hall, 1982.
[4] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. New York: J. Wiley, 1973.
[5] J. Grim, "On numerical evaluation of maximum-likelihood estimates for finite mixtures of distributions," *Kybernetika*, vol. 18, pp. 173-190, 1982.
[6] J. Grim, "Multivariate statistical pattern recognition with nonreduced dimensionality," *Kybernetika*, vol. 22, pp. 142-157, 1986.
[7] A.K. Jain, "Advances in statistical pattern recognition," *Pattern Recognition Theory and Applications, Proc. NATO Advanced Study Inst.*, P. Devijver and J. Kittler, eds. Berlin-Heidelberg-New York: Springer-Verlag, pp. 1-19, 1987.
[8] P. Pudil, J. Novovicová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, pp. 1,119-1,125, 1994.
[9] P. Pudil, J. Novovicová, and J. Kittler, "Automatic machine learning of decision rule for classification problems in image analysis," *Proc. BMVC '93—Fourth British Machine Vision Conf.*, vol. 1, pp. 15-24, 1993.
[10] P. Pudil, J. Novovicová, and J. Kittler, "Simultaneous learning of decision rules and important attributes for classification problems in image analysis," *Image and Vision Computing*, vol. 12, pp. 193-198, 1994.
[11] R.A. Redner and H.F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, pp. 195-239, 1984.