Feature Selection Expert - User Oriented Approach

Methodology and Concept of the System

P. Pudil, J. Novovičovà, P. Somol, R. Vrňata

Inst. of Information Theory and Automation
Academy of Sciences of the Czech Republic
Department of Pattern Recognition
182 08 Prague 8, Czech Republic

Abstract. The paper describes the methodology for feature selection and the concept of a user-oriented software package (FS Expert) for feature selection with a consulting system integrated into the package. It attempts to provide a guideline which approach to choose with respect to the extent of a priori knowledge of the problem. The methods implemented in FS Expert are based mostly on the methodology developed by the authors, though it is being built as an open system.

1 Introduction

Abundance of various methods for feature selection can be found in the literature, however, for somebody in need of choosing the proper method for his particular problem, it is rather difficult to do so. The optimal choice depends certainly on a number of conditions. Certainly different methods will be appropriate for optimal data representation and for discrimination. Also the original dimensionality of the problem at hand, the level of *apriori* knowledge of the form of underlying probability structures and the availability of a reasonably sized training set will influence the choice.

With the aim to ease the situation, we are currently developing in our research team a software package to solve semi-automatically the Fubset Selection problem. It will be equipped with a kind of expert or consulting system which should guide a less experienced user through the methods included into the package. With respect to the level of problem knowledge, the user will arrive to the particular method fitting best his knowledge of the problem at hand. Though a number of currently available methods will be included, the core of the package will be formed by the novel methods we have developed ourselves. These methods are briefly described in the sequel. At the end of this paper we are presenting a simplified example of the flow chart of such a consulting system.

[0] Supported by the Grant No. A2075703 of the Academy of Sciences, by the Grant No. 402/97/1242 of the Czech Grant Agency and by the Grant of the Ministry of Education No. VS 96063 of the Czech Republic.

2 Feature Selection Problem in Statistical PR

Following the statistical approach to pattern recognition, we assume that a pattern or object described by a real D-dimensional vector $\mathbf{x} = (x_1, x_2, \cdots, x_D)^{\mathrm{T}} \in \mathcal{X} \subset \mathcal{R}^D$ 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 probability density functions (pdfs) $p^{\star}(\mathbf{x}|\omega)$ and the respective *a priori* probabilities $P^{\star}(\omega)$. 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 training sets of samples with known classification. The aim of feature selection is to find a subset of d features out of D original ones so as not to deteriorate the performance of the classifier.

Our own research and experience with feature selection has led us to the conclusion that there exists no unique optimal approach to the problem. Some approaches are more suitable under certain conditions, and different approaches are more appropriate under other conditions. These conditions are related to the level of our **knowledge of the problem**, thus we can talk about "knowledge-based subset selection". Hence we have attempted to extend the "battery" of available tools by developing several new methods, each of them suitable for different conditions, trying to cover the majority of situations which can be encountered in practice. These methods are introduced in the next pages.

3 Basic Situation with Respect to Problem Knowledge

There are perhaps two basic classes of situations with respect to *a priori* knowledge of the form of underlying probability structures:

1. Some apriori knowledge is available (at least that pdfs are unimodal)
In these cases the use of some probabilistic distance measures (like Mahalanobis, Bhattacharyya, etc.) may be appropriate as the evaluation criterion. As pointed out by [2] the error rate is even better (provided it can be reasonably computed).

For this type of situations we have developed a family of Floating Search algorithms which yield close to optimal solution, are computationally effective (facilitating FS in high dimensional problems) and do not require the fulfilment of monotonicity condition. In a recent comparative study of currently available subset search strategies carried out by Jain [4] were the Floating Search algorithms evaluated as the most powerful ones.

2. No a priori knowledge is available
We cannot even assume that pdfs are unimodal (or suspect they are multimodal). The only source of available information is provided by the training data. Feature selection in such a case becomes a very challenging problem. The early solutions of this problem suffer from serious shortcomings. For these situations we have developed two new approaches aimed to cope reasonably well in such circumstances, conceptually very different from those mentioned above. They are based on approximating unknown conditional pdfs by finite mixtures of a special type and are discussed later.

4 Floating Search Methods

Various search strategies are used to find the subset of features optimizing an adopted criterion, once this criterion has been chosen. They range from simple but popular ones, like sequential forward (SFS) and sequential backward (SBS) selection, to more sophisticated but computationally more difficult ones. The so called "nesting of feature subsets" may rapidly result in deteriorating performance of both the SFS and SBS algorithms. This can be *partially* overcome by employing either the so-called (l, r) (in one step include l and exclude r features) or generalized (l, r) algorithms ([1]). Unfortunately, there is no theoretical way of predicting the values of l and r so as to achieve the best feature set.

To counteract these problems we developed recently a novel family of search strategies based on the principle of iterative search in both directions, but as opposed to the bidirectional search proposed in [2], exploiting a flexible level of repeated backtracking. Instead of fixing the values of l and r, these values are allowed to "float", i.e. to flexibly change so as to approximate the optimal solution as much as possible. Consequently, the resulting dimensionality in respective intermediate stages of the algorithm is not changing monotonously but is actually "floating" up and down. Because of this "floating" characteristics, the two methods have been named **floating search methods** ([11]). Although they both switch between feature inclusion and exclusion, they are based on two different algorithms according to the dominant direction of the search:

Sequential forward floating search (SFFS)

- the search dominantly in the forward direction

Sequential backward floating search (SBFS)

- the search dominantly in the forward direction

A more exact description of the algorithms is given in [11]. A simplified flow chart of the SFFS algorithm is given in Fig.1:

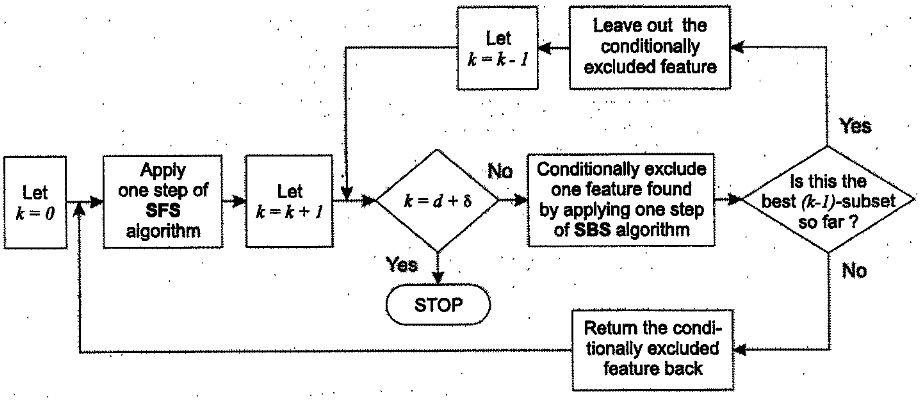

Let $k=0$
Apply one step of SFS algorithm
Let $k=k+1$
$k=d+\delta$
No
Yes
STOP
Conditionally exclude one feature found by applying one step of SBS algorithm
Is this the best $(k-1)$-subset so far ?
Yes
No
Leave out the conditionally excluded feature
Let $k=k-1$
Return the conditionally excluded feature back

Fig. 1. Flow Chart of SFFS algorithm

The terminating condition $k = d + \delta$ in the flowchart means that in order to fully utilize the properties of SFFS search we should not stop the algorithm immediately after it reaches for the first time the dimensionality d. By leaving it to float up and back a bit further, the potential of the algorithm is better utilized and a subset of cardinality d outperforming the first one achieved is usually found. In practice we can let the algorithm either go up to the original dimensionality D, or if it is too large, then the value of δ can be determined heuristically (e.g. according to the value of maximum number of backtracking steps prior to reaching d first time).

Unlike the (l, r) and generalized (l, r) algorithms in which factors such as the net change in the size of the current feature set, and especially the amount of computational time, are governed by the values of l and r, the SFFS and SBFS methods are not restricted by these factors. By means of conditional "floating down and up" both the methods are freely allowed to correct wrong decisions made in the previous steps so as to approximate the optimal solution as much as possible.

The results achieved so far on various sets of data demonstrate clearly a great potential of floating search strategies ([11, 3, 4]). Generally, though of heuristic nature, Floating Search methods provide either the optimal or a close to optimal solution, but also require much less computational time than the Branch and Bound method and most other currently used suboptimal strategies. Moreover, as opposed to the branch and bound method, the floating search methods are also tolerant to deviations from monotonic behaviour of the feature selection criterion function. It makes them particularly suited in conjunction with nonmonotonic FS criterion like the error rate of the classifier.

5 Feature Selection by Modified Gaussian Mixtures

Now we shall address the feature selection problem arising **when we have the data but no other information** which occurs in a number of real situations. Though various nonparametric methods of classification are available, none of them is problem free and universal.

In order to facilitate the solution under these conditions, we have developed a completely new approach that can serve the multiple goal of:

1. learning the structure of multivariate distributions,
2. identifying the most important variables and thus facilitating the dimensionality reduction,
3. deriving automatically a decision rule based on the selected features

The approach is based on approximating the unknown class conditional distributions by finite mixtures of parametrized densities of a special type. In terms of the required computer storage it is considerably more efficient than nonparametric pdf estimation methods.

As far as the FS problem is concerned, there are two different methods with respect to the criterion employed. Though both methods exploit a common basic approximation model, the way of selecting features is completely different. The same holds for their purpose and optimal applicability. Since it is not possible to present a detailed formalized description of the approach in this paper, the reader is referred to respective original sources. However, we attempt to provide at least an outline of the common statistical model used for both the methods and then the respective methods of FS.

5.1 Modified Gaussian Mixture Model

The approach to feature selection taken here is to approximate the ωth class density of $\mathbf{x}$ by the following modified Gaussian model with latent structure or modified latent subclass Gaussian model ([10, 7, 6]):

$$p(\mathbf{x}|\omega) = \sum_{m=1}^{M_\omega} \alpha_m^\omega g_0(\mathbf{x}|\theta_0) g(\mathbf{x}|\theta_m^\omega, \theta_0, \Phi), \quad \mathbf{x} \in \mathcal{X}, \tag{1}$$

where α_m^ω is the mixing probability for the mth subclass, $\sum_{m=1}^{M_\omega} \alpha_m^\omega = 1$. The function g_0 is a nonzero "background" density common to all classes and functions $g(\mathbf{x}|\theta_m^\omega, \theta_0, \Phi)$ include structural parameters ϕ_i:

$$g_0(\mathbf{x}|\theta_0) = \prod_{i=1}^{D} f_i(x_i|\theta_{0i}), \quad g(\mathbf{x}|\theta_m^\omega, \theta_0, \Phi) = \prod_{i=1}^{D} \left[\frac{f_i(x_i|\theta_{mi}^\omega)}{f_i(x_i|\theta_{0i})}\right]^{\phi_i}, \tag{2}$$

$$\theta_0 = \{\theta_{0i}\}_{i=1}^D, \quad \theta_m^\omega = \{\theta_{mi}^\omega\}_{i=1}^D, \quad \phi_i = \{0, 1\}, \quad \Phi = \{\phi_i\}_{i=1}^D.$$

The univariate function f is assumed to be from a family of univariate normal densities $\{f(\xi|\theta) = f(\xi|\mu, \sigma), \xi \in \mathcal{R}, \ \mu \in \mathcal{R}, \sigma^2 \in (0, \infty)\}$. Our model is based on the idea to posit a common "background" normal density for all classes and to express each class pdf as a mixture of a product of this "background" density with a class-specific function defined on a subspace of the feature vector space. This subspace is chosen by means of the parameters ϕ_i and the same subspace of $\mathcal{X}$ for each component density is used in all classes. Any specific univariate function $f_i(x_i|\theta_{mi}^\omega)$ is substituted by the respective "background" density $f_i(x_i|\theta_{0i})$ whenever ϕ_i is zero. In this way the binary parameters ϕ_i can be looked upon as *control variables* since the structure of the mixture (1) can be controlled by means of these parameters.

For any choice of ϕ_i the finite mixture (1) can be rewritten by using (2) as

$$p(\mathbf{x}|\alpha^\omega, \theta^\omega, \theta_0, \Phi) = \sum_{m=1}^{M_\omega} \alpha_m^\omega \prod_{i=1}^{D} [f_i(x_i|\theta_{0i})]^{(1-\phi_i)} [f_i(x_i|\theta_{mi}^\omega,)]^{\phi_i}, \tag{3}$$

$$\alpha^\omega = \{\alpha_m^\omega\}_{m=1}^{M_\omega}, \quad \theta^\omega = \{\mu_m^\omega,\}_{m=1}^{M_\omega}$$

Setting some $\phi_i = 1$, we replace the function $f_i(x_i|\theta_{0i})$ in the product in (3) by $f_i(x_i|\theta_{mi}^\omega)$ and introduce a new independent parameter θ_{mi}^ω in the mixture (3). The number of involved parameters is specified by $\sum_{i=1}^D \phi_i = \gamma, \ 1 \le \gamma \le D$. The parameter sets $\alpha^\omega, \theta^\omega, \theta_0, \Phi$ are unknown and can be estimated from the training sets. $\mathbf{X}_\omega$. The EM ("Expectation-Maximization") iterative algorithm ([12]) provides a convenient method for maximizing the likelihood function.

5.2 Feature Selection for Approximation

Two methods have been derived depending on different criteria for features evaluation ([10, 7, 6]). The first method selects a feature subset X_d by choosing Φ_d (i.e. parameter vector Φ restricted to have just d components equal to 1 and $D - d$ components equal to 0) such that the best approximation is obtained. In the **"approximation" method** for FS the criterion we use 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 densities of $\mathbf{x}$. A simplified float chart of that method can be found in [8].
The "approximation" method has the following interesting characteristics:

1. owing to the convenient form of the postulated model the "contribution" $Q(\Phi)$ of a feature subset to the chosen criterion is the sum of individual contributions
$\hat{Q}_i = \sum_{\omega \in \Omega} \sum_{m=1}^{M_\omega} P(\omega) \hat{\alpha}_m^\omega \log \frac{(\hat{\sigma}_{0i})^2}{(\hat{\sigma}_{mi}^\omega)^2}$, which can be assessed independently for each feature: $Q(\Phi) = \sum_{i=1}^{D} \phi_i \hat{Q}_i$
2. only the operation of ranking of individual feature contributions is therefore required (without any search procedure) in order to obtain a required subset of d features.

Though the "approximation method" yielded very good results in many problems ranging from image analysis to classification, we should be aware of the fact that the features are selected with respect to their "approximation" or "representation" quality, which may not in particular cases coincide with their "discriminative" quality. Consequently, the method is particularly convenient for the problems of multivariate data representation in a lower-dimensional space or pattern interpretation. It is also applicable to multiclass problems. For the cases when the discrimination between classes is the primary goal, the following "divergence" method has been developed.

5.3 Feature Selection for Discrimination

In order to select those features that are most useful in describing differences between two possible classes, we have developed another method ([7, 6]) for feature selection. Similarly to the "approximation" method it utilizes the same general model for approximating unknown class conditional pdfs by finite mixtures of parametrized densities (1). However, in this case the Kullback's J-divergence between two classes defined in terms of a posteriori probabilities (or equivalently the Kullback-Leibler measures of discriminatory information between two classes mixed in the proportions in which the classes truly occur) is used as the appropriate evaluation criterion. The goal of the method is to maximize the divergence discrimination, hence the name of "divergence" method.

The proposed approach is especially suitable for multimodal data and is restricted to two classes. The two interesting characteristics specified above for the approximation method hold for the divergence method too ([7]).

5.4 Properties of Approximation and Divergence Methods

An important characteristic of our approach is that it effectively partitions the feature 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 all the classes and constitutes the background distribution, as opposed to the features forming X_d, which are significant for discriminating the classes. The joint distribution of these features constitutes the "specific" distribution defined in (2). According to these features alone, a new pattern $\mathbf{x}$ is classified into one of C classes.

It has been proved that just under this partition of the set X the following holds:

1. For the "approximation" method: the Kullback-Leibler distance between the true and the postulated class conditional pdfs is minimized.
2. For the "divergence" method: the Kullback J-divergence is maximized.

For those interested in clasification problems it may be of interest that our approach is not only a classification procedure but also a data reduction tool. The modified mixture (3) reduces also the computational complexity of the corresponding Bayes decision rule. We can represent multiclass data by d features, where $d < D$ if $\phi_i = 1$ for $i = 1, \cdots d$. Given the approximations $p(\mathbf{x}|\hat{\alpha}_\omega, \hat{\theta}_\omega, \hat{\theta}_0, \hat{\Phi}_d)$ it can be easily seen that the background pdf g_0 may be reduced in the Bayes decision rule. Thus we may classify the observation of $\mathbf{x}$ according to the pseudo-Bayes decision rule: decide that $\mathbf{x}$ is from class ω_l if

$$P(\omega_l) \sum_{m=1}^{M_\omega} \hat{\alpha}_m^{\omega_l} \prod_{k=1}^{d} f_i(x_{i_k}|\hat{\theta}_{mi_k}^{\omega_l}) = \max_{j=1,\cdots,C} \{P(\omega_j) \sum_{m=1}^{M_\omega} \hat{\alpha}_m^{\omega_j} \prod_{k=1}^{d} f_i(x_{i_k}|\hat{\theta}_{mi_k}^{\omega_j})\}.$$

We use the term "pseudo-Bayes" as the rule is applied in a lower-dimensional subspace corresponding to the selected subset and, furthermore, the true conditional pdfs are in the formulas substituted by their approximations. It means that a new feature vector $\mathbf{x}$ is classified into one of the classes according to only d features $x_{i_1}, \cdots, x_{i_d}$.

6 Feature Selection Expert System

Though it is certainly too premature to claim a real development of the knowledge guided approach to subset selection, the presented methods can serve as a starting point to achieve such an ambitious goal. To implement it, we are developing an integrated environment which would incorporate a whole family of methods together with a sort of expert system. It should guide a user according to his goals and the degree of the problem knowledge he has to the semi-automatic choice of the most suitable method. We hope that the flow chart of the "prototype" Subset Selection Guide presented in Figure 2. will give the reader the idea how it is being built.

The flow-chart and the questions are certainly very simplified here. Nevertheless, we would like to draw readers' attention to the fact that in a great number

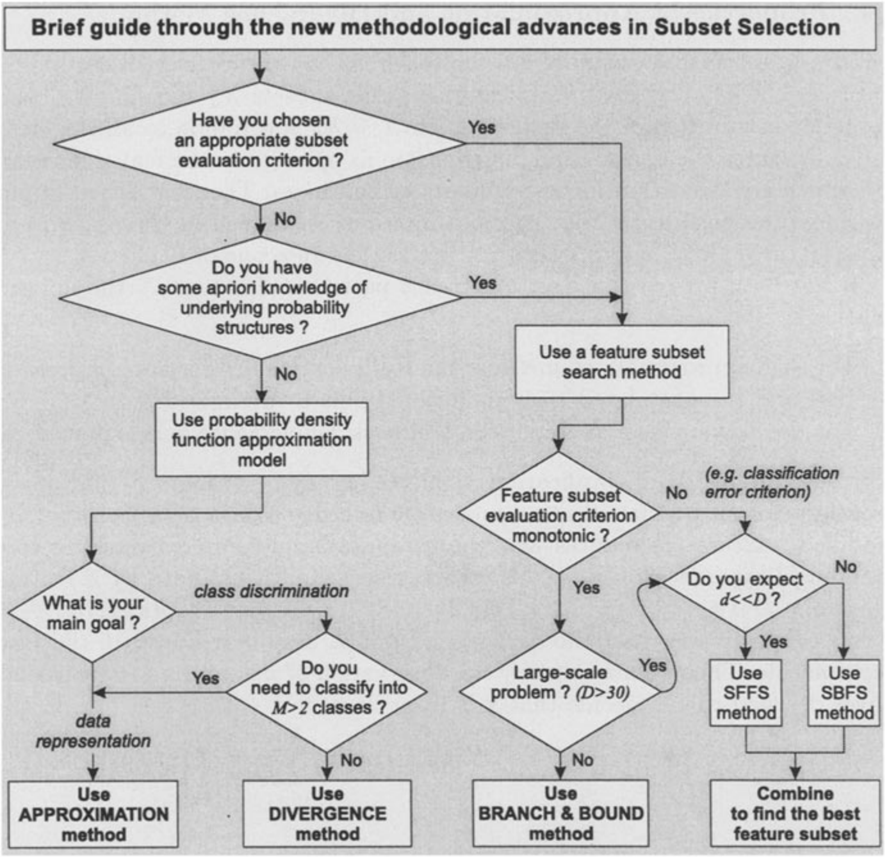

Brief guide through the new methodological advances in Subset Selection
Have you chosen an appropriate subset evaluation criterion ?
Yes
No
Do you have some apriori knowledge of underlying probability structures ?
Yes
No
Use a feature subset search method
Use probability density function approximation model
Feature subset evaluation criterion monotonic ?
No
(e.g. classification error criterion)
Do you expect d<<D ?
No
What is your main goal ?
class discrimination
Yes
Yes
Do you need to classify into M>2 classes ?
Yes
Large-scale problem ? (D>30)
Yes
Use SFFS method
Use SBFS method
data representation
No
No
Use APPROXIMATION method
Use DIVERGENCE method
Use BRANCH & BOUND method
Combine to find the best feature subset

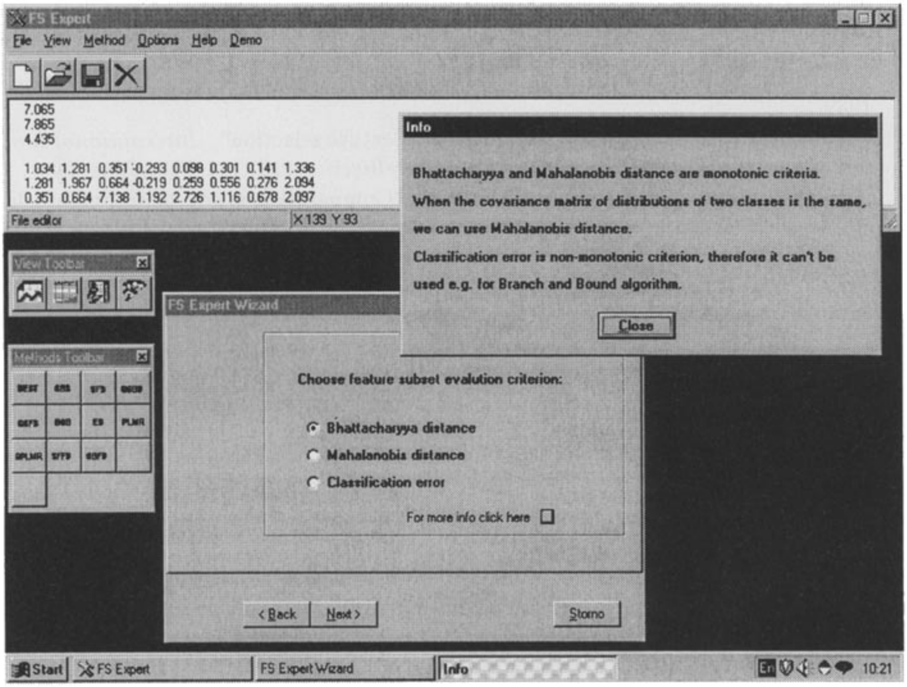
Fig. 2. Flow chart of "prototype" Subset Selection Guide

of real situations there exists a generally accepted and appropriate criterion to evaluate the selected subset for a given problem (e.g. for multiple stepwise regresion analysis). In all these cases the subset search techniques may be used as depicted in the flow-chart.

The FS Expert system is provided by a context sensitive help. A user will be guided by a rule-based consulting system which is realized in the form of so called wizards. Having in mind less experienced users or teaching purposes, the help modul contains sample files with the description of various FS methods and examples of their use. To demonstrate how the system looks like, two snapshots of the FS Expert screens are given in Figures 3 and 4.

Finally, we should stress that the system is being built as an open one. Accordingly it will not contain only our own new or recently developed methods, but eventually also other methods.

Obviously, there are some research issues yet to be pursued, like a possible combination of genetic algorithms and floating search, however, we believe that the research community may benefit already from the currently achieved results.



```
7.065
7.865
4.435

1.034 1.281 0.351 -0.293 0.098 0.301 0.141 1.336
1.281 1.967 0.664 -0.219 0.259 0.556 0.276 2.094
0.351 0.664 7.138 1.192 2.726 1.116 0.678 2.097
```

Bhattacharyya and Mahalanobis distance are monotonic criteria.

When the covariance matrix of distributions of two classes is the same, we can use Mahalanobis distance.

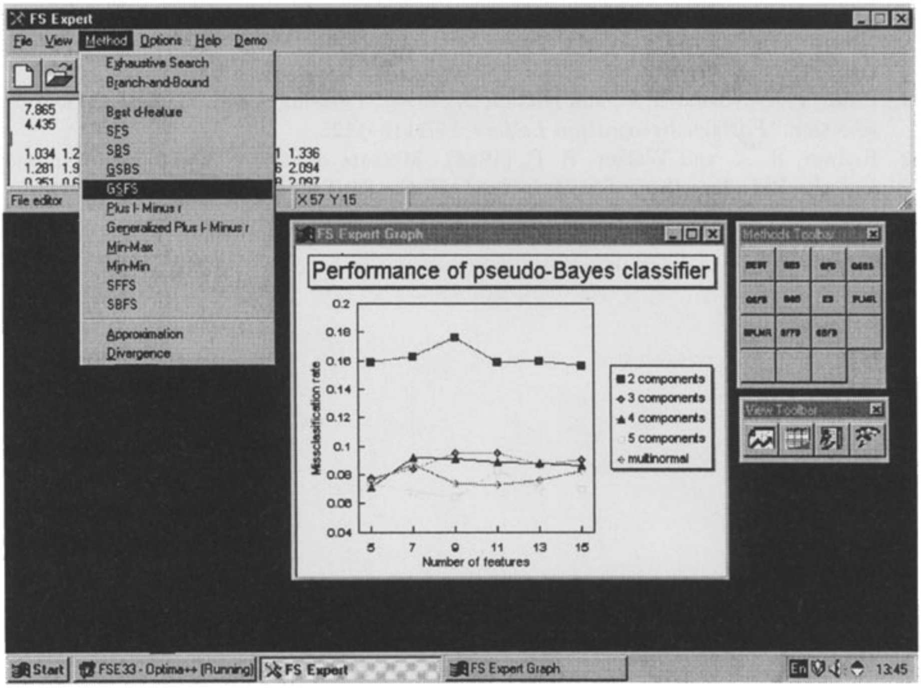
Classification error is non-monotonic criterion, therefore it can't be used e.g. for Branch and Bound algorithm.

Choose feature subset evalution criterion:

● Bhattacharyya distance
○ Mahalanobis distance
○ Classification error

For more info click here □

Fig. 3. FS Expert Snapshot No.1



```
7.865
4.435

1.034 1.2                                1 1.336
1.281 1.9                                6 2.094
0.351 0.6                                8 2.097
```

Fig. 4. FS Expert Snapshot No.2

References

1. P.Devijver and J.Kittler. *Pattern Recognition: A Statistical Approach.* Prentice, 1982.
2. W.Siedlecki and J.Sklansky. "On automatic feature selection", *International Journal of Pattern Recognition and Artificial Intelligence*, 2(2):197–220, June 1988.
3. F.J.Ferri, P.Pudil, M.Hatef and J.Kittler. "Comparative study of techniques for large-scale feature selection", In E. S. Gelsema, L. N. Kanal, eds, *Pattern Recognition in Practice IV*, pp.403–413. Elsevier, 1994.
4. A.Jain and D.Zongker. "Feature Selection: Evaluation, Application and Small Sample Performance", *IEEE Transactions on PAMI*, 19(2): 153–158, 1997.
5. P.Pudil, J.Novovičová and J.Kittler. "Floating Search Methods in Feature Selection. *Pattern Recognition Letters"*, 15 (11), 1119–1125, 1994.
6. Novovičová, J. and Pudil, P. (1997). *Dealing With Complexity: Neural Network Approach*, chapter Feature Selection and Classification by Modified Model with Latent Structure. Springer Verlag.
7. Novovičová, J., Pudil, P., and Kittler, J. (1996). Divergence based feature selection for multimodal class densities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18:218–223.
8. Pudil, P. and Novovičová, J. (1998). Novel methods for subset selection with respect to problem knowledge. *IEEE Intelligent Systems - Special Issue on Feature Transformation and Subset Selection*,March/April 1998: 66–74.
9. Pudil, P., Novovičová, J., Choakjarernwanit, N., and Kittler, J. (1993). An analysis of the max-min approach to feature selection. *Pattern Recognition Letters*, 14(11):841–847.
10. Pudil, P., Novovičová, J., Choakjarernwanit, N., and Kittler, J. (1995). Feature selection based on the approximation of class densities by finite mixtures of special type. *Pattern Recognition*, 28(9):1389–98.
11. Pudil, P., Novovičová, J., and Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, 15:1119–1125.
12. Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM J. Appl. Math.*, 26(2):195–239.