Kybernetika

VOLUME *34* (1998), NUMBER 4

The Journal of the Czech Society for Cybernetics and Information Sciences

Published by:
Institute of Information Theory and Automation of the Academy of Sciences of the Czech Republic

Kybernetika is a bi-monthly international journal dedicated for rapid publication of high-quality, peer-reviewed research articles in fields covered by its title.

Kybernetika traditionally publishes research results in the fields of Control Sciences, Information Sciences, System Sciences, Statistical Decision Making, Applied Probability Theory, Random Processes, Fuzziness and Uncertainty Theories, Operations Research and Theoretical Computer Science, as well as in the topics closely related to the above fields.

The Journal has been monitored in the Science Citation Index since 1977 and it is abstracted/indexed in databases of Mathematical Reviews, Current Mathematical Publications, Current Contents ISI Engineering and Computing Technology.

K y b e r n e t i k a. Volume *34* (1998) ISSN 0023-5954, MK ČR E 4902.

Published bi-monthly by the Institute of Information Theory and Automation of the Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Praha 8. — Address of the Editor: P. O. Box 18, 182 08 Prague 8, e-mail: kybernetika@utia.cas.cz. — Printed by PV Press, Pod vrstevnicí 5, 140 00 Prague 4. — Orders and subscriptions should be placed with: MYRIS TRADE Ltd., P. O. Box 2, V Štíhlách 1311, 142 01 Prague 4, Czech Republic, e-mail: myris@myris.cz. — Sole agent for all "western" countries: Kubon & Sagner, P. O. Box 34 01 08, D-8 000 München 34, F.R.G.

Published in September 1998.

KYBERNETIKA — VOLUME *34* (1998), NUMBER 4, PAGES 451 – 460

CONCEPTUAL BASE OF FEATURE SELECTION CONSULTING SYSTEM[1]

Pavel Pudil, Jana Novovičová, Petr Somol and Radek Vrňata

The paper briefly reviews recent advances in the methodology of feature selection (FS) and the conceptual base of a consulting system for solving FS problems. The reasons for designing a kind of expert or consulting system which would guide a less experienced user are outlined. The paper also attempts to provide a guideline which approach to choose with respect to the extent of a priori knowledge of the problem. The methods discussed here form the core of the software package being developed for solving FS problem. Two basic approaches are reviewed and the conditions under which they should be used are specified. One approach involves the use of the computationally effective Floating search methods. The alternative approach trades off the requirement for a priori information for the requirement of sufficient data to represent the distributions involved. Owing to its nature it is particularly suitable for cases when the underlying probability distributions are not unimodal. The approach attempts to achieve simultaneous feature selection and decision rule inference. According to the criterion adopted there are two variants allowing the selection of features either for optimal representation or discrimination.

1. INTRODUCTION

Abundance of various methods for feature selection (FS) 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, like the aim of dimensionality reduction (representation or discrimination), the original dimensionality of input data, the level of *a priori* knowledge of underlying probability structures, the size of the training set, etc.

With the aim to ease the situation, we are currently developing a software package for solving the FS 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 above named conditions, 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

[1]Supported by the grants of Czech Ministry of Education MŠMT No. VS96063, Grant Agency of the Czech Republic No. 402/97/1242 and Grant Agency of the Czech Acad. Sci. A 2075608.

are briefly described in the sequel with references to particular papers describing the methods in more detail. At the end of this paper a simplified example of the flow chart of such a consulting system is presented.

2. FEATURE SELECTION PROBLEM IN PATTERN RECOGNITION

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 $P^\star(\omega)$, that is $P^\star(\omega)$ are the "between-class" mixing proportions.

In the majority of practical cases, the dimensionality of the pattern descriptor space can be rather high. It is the consequence of the fact that in the design phase it is extremely difficult or practically impossible to evaluate directly the "usefulness" of particular descriptors or input variables. In consequence, it is desirable and important to include all the "reasonable" descriptors the designer can think of. The reason is that no subsequent mathematical processing can add information missing in the originally designed measurement set. The aim of feature selection is therefore to find a subset of d features out of original D features, where $d < D$ (if possible $d \ll D$), so as to maximize (or minimize) an adopted criterion.

Assuming that a suitable criterion function has been chosen to evaluate the effectiveness of feature subsets, feature selection is usually reduced to a search problem that detects an optimal feature subset based on the selected measure. Despite undisputable progress, the existing search techniques are not yet completely satisfactory. They either require monotonicity of the criterion function or they cope with non-monotonicity and perform well for FS problems of moderate sizes, but this property does not appear to extend properly to large scale problems [4].

Our own research and experience with feature selection (FS) 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 depend mainly on the level of our *knowledge of the problem*, thus we can talk about "knowledge-based subset selection". Hence we have attempted to extend the "batery" of available tools by developing several new methods which are briefly introduced in next pages. Each of them is suitable for different conditions as we have been trying to cover the majority of situations which can be encountered in practice.

3. BASIC SITUATION WITH RESPECT TO PROBLEM KNOWLEDGE

There are perhaps two basic classes of situations with respect to *a priori* knowledge of underlying probability density functions (pdfs):

1. *Some a priori knowledge is available*
   (At least that pdfs are unimodal). In these cases the use of some probabilistic

distance measures (like Mahalanobis, Bhattachaarya, etc.) may be appropriate as the evaluation criterion. As pointed out by Siedlecki and Sklansky [19] the error rate is even better (provided it can be reasonably computed).

New Tool: 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 Zongker and Jain [20] were the Floating Search algorithms evaluated as the most efficient 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 (see [14]).
New Tool: For these situations we have developed two new approaches aimed to cope reasonably 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.

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 (e. g. generalized versions GSFS(l), GSBS(l)). The so called nesting of feature subsets may rapidly result in suboptimality of both the SFS and SBS algorithms. This can be *partially* overcome by employing either the (l, r) or generalized (l, r) algorithms [3, 5] which involve a successive feature set augmentation and depletion process. Consequently, the resulting dimensionality in respective stages of both algorithms is fixed depending on the prespecified values of l and r. 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 a family of search strategies based on the principle of iterative search in both directions, but as opposed to the bidirectional search proposed in [19], 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* [10, 15, 16]. Although both of them switch between feature inclusion and exclusion, they are based on two different algorithms according to the dominant direction of the search:

Sequential forward floating selection (SFFS)
— the search dominantly in the forward direction

Sequential backward floating selection (SBFS)
— the search dominantly in the backward direction

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.

A more detailed description of the algorithms is given in [10, 16]. Here we present a simplified flow chart of the SFFS algorithm:

The results achieved so far on various sets of data demonstrate clearly a great potential of floating search strategies [16, 4, 20]. Floating search methods yielded in almost all the cases better results than GSFS and GSBS algorithms (which in turn gave better results than (l, r) algorithms).

The worst proved to be the Max-Min algorithm which is computationally appealing but theoretically ill-founded, as we have proved [11, 12]. The subsets found by Floating Search are practically identical with those found by a computationally tedious "exhaustive" combinatorial search. Comparison with the Branch and Bound method showed that Floating Search is much faster [16].

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. The computational efficiency allows the use of floating search even for large scale FS problems. To test the different algorithms in a large search space we also considered a document recognition problem. It involved discrimination between correct and deffective records of banking documents consisting of 360 optical measurements. The Floating Search methods outperformed in this problem the traditional suboptimal search methods and also yielded better results than the genetic algorithms, particularly for higher dimensions [4].

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 which according to a number of researchers seem to be the only legitimate criterion for feature subset evaluation.

5. FEATURE SELECTION BY MODIFIED NORMAL MIXTURES

Now we shall address the feature selection problem arising when we have no information concerning the underlying class pdfs which occurs in a number of real situations. In this section we present the modified mixtures approach to feature

selection developed in [17].
To summarized, our proposal for FS has the following features:

1. the classes are modelled as modified mixtures of normal distributions with latent structure;
2. no search procedure is required for identification of the most important feature variables and thus for facilitation of the dimensionality reduction;
3. reduces also the complexity of the corresponding Bayes decision making.

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 mixture model, the way of selecting features is different. The same holds for their purpose and optimal applicability. It is not possible to present in this paper a detailed formalized description of the approach, 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 normal mixture model with latent structure

The approach to feature selection taken here is to model the class pdfs by modified normal mixture model introduced first in [13]. We divide each class ω, $\omega \in \Omega$ into M_ω artifical subclasses, i. e. M_ω denotes the total number of different components in the ωth class mixture pdf. Let $p_m(\mathbf{x}|\omega)$ denotes the multivariate pdf of the mth component in the mixture for class ω, and let α_m^ω denote the proportion of subclass m in class ω. The "within-class" mixing proportions α_m^ω are nonnegative and satisfy the equations $\sum_{m=1}^{M_\omega} \alpha_m^\omega = 1,\ \omega \in \Omega$. In our approach to FS the ωth pdf $p^\star(\mathbf{x}|\omega)$, $\omega \in \Omega$, is approximated by a mixture pdf denoted by $p(\mathbf{x}|\omega)$, that is

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

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 $\{f(\xi|\theta) = f(\xi|\mu, \sigma),\ \xi \in \mathcal{R},\ \mu \in \mathcal{R}, \sigma^2 \in (0, \infty)\}$, with the mean μ and the variance σ^2.
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* due to the fact that the structure of the mixture (1) can be controlled by means of that 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 actual number of involved parameters can be specified by the condition $\sum_{i=1}^{D} \phi_i = \gamma$, $1 \leq \gamma \leq D$.
The parameter sets $\alpha^\omega, \theta^\omega, \theta_0, \Phi$ are unknown and can be estimated from the training sets. Suppose that the ωth training set is $\mathbf{X}_\omega$ and $|\mathbf{X}_\omega|$ is the number of training data from the ωth class. Then the log-likelihood function for the data is

$$L(\alpha, \theta, \theta_0, \mathbf{\Phi}) = \sum_{\omega \in \Omega} \frac{P(\omega)}{|\mathbf{X}_\omega|} \sum_{\mathbf{x} \in \mathbf{X}_\omega} \log p(\mathbf{x}|\alpha^\omega, \theta^\omega, \theta_0, \Phi), \tag{4}$$

$$\alpha = \{\alpha^\omega\}_{\omega \in \Omega}, \quad \theta = \{\theta^\omega\}_{\omega \in \Omega}.$$

The expectation-maximization (EM) iterative algorithm can be used to fit mixtures by maximum-likelihood. For given Φ the EM algorithm steps are

$$\begin{aligned} p(m|\mathbf{x}, \omega) &= Prob(\mathbf{x} \in m\text{th subclass of class } \omega|\mathbf{x}, \omega) \\ &= \frac{\alpha_m^\omega g(\mathbf{x}|\theta_m^\omega, \theta_0, \Phi)}{\sum_{j=1}^{M_\omega} \alpha_j^\omega g(\mathbf{x}|\theta_j^\omega, \theta_0, \Phi)}, \end{aligned} \tag{5}$$

$$\hat{\alpha}_m^\omega = \frac{1}{|\mathbf{X}_\omega|} \sum_{\mathbf{x} \in \mathbf{X}_\omega} p(m|\mathbf{x}, \omega), \quad \sum_{m=1}^{M_\omega} \hat{\alpha}_m^\omega = 1 \tag{6}$$

$$\hat{\mu}_{mi}^\omega = \frac{1}{|\mathbf{X}_\omega| \hat{\alpha}_m^\omega} \sum_{\mathbf{x} \in \mathbf{X}_\omega} x_i p(m|\mathbf{x}, \omega), \quad (\hat{\sigma}_{mi}^\omega)^2 = \frac{1}{|\mathbf{X}_\omega| \hat{\alpha}_m^\omega} \sum_{\mathbf{x} \in \mathbf{X}_\omega} (x_i - \hat{\mu}_{mi}^\omega)^2 p(m|\mathbf{x}, \omega) \tag{7}$$

$$\hat{\mu}_{0i} = \sum_{\omega \in \Omega} P(\omega) \sum_{m=1}^{M_\omega} \hat{\alpha}_m^\omega \hat{\mu}_{mi}^\omega, \quad (\hat{\sigma}_{0i})^2 = \sum_{\omega \in \Omega} P(\omega) \sum_{m=1}^{M_\omega} \hat{\alpha}_m^\omega \left[(\hat{\sigma}_{mi}^\omega)^2 + (\hat{\mu}_{mi}^\omega - \hat{\mu}_{0i})^2 \right]. \tag{8}$$

5.2. Feature Selection for best approximation

Two methods have been derived depending on different criteria for features evaluation [8, 17]. The first method selects a feature subset X_d by choosing $\mathbf{\Phi}_d$ (i. e.

parameter vector $\mathbf{\Phi}$ 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* [9, 17] for FS the criterion we use for measuring the error resulting from approximating the true pdf $p^{\star}(\mathbf{x}|\omega)$ by $p(\mathbf{x}|\alpha_{\omega}, \theta_{\omega}, \theta_0, \Phi)$ for all $\omega \in \Omega$ 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 $\mathbf{x}$.

The "approximation" method has the following interesting characteristics:

1. owing to the convenient form of the postulated model the "contribution" $Q(\mathbf{\Phi})$ 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: $Q(\mathbf{\Phi}) = \sum_{i=1}^{D}\phi_i\hat{Q}_i$ (formulas for contributions $\hat{Q}_i$ can be found in [9, 17])
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. Is 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 [8, 9] 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 (see e. g. [2]) 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 at the moment to two classes. The two interesting characteristics specified above for the approximation method hold for the divergence method too [8].

5.4. Properties of approximation and divergence methods

An important characteristic of our 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 distri-

bution 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.

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 inequality in the Bayes decision rule. Thus we may classify the observation of $\mathbf{x}$ according to the sample Bayes decision rule applied in a lower-dimensional subspace corresponding to the selected subset :
decide that $\mathbf{x}$ is from class ω_l if

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

6. PROTOTYPE OF FS CONSULTING SYSTEM

Though it is certainly too premature to claim the real development of a knowledge guided approach to FS, 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 or consulting system. It should guide a user according to the degree of available knowledge to the semi-automatic choice of the most suitable method. A simplified flow chart of the "prototype" Feature Selection Consulting System is presented in the next page. Obviously, there are some research issues yet to be pursued, like a possible combination of genetic algorithms and floating search or future developments of approximation approach taking into account certain peculiarities which may occur in practice etc.

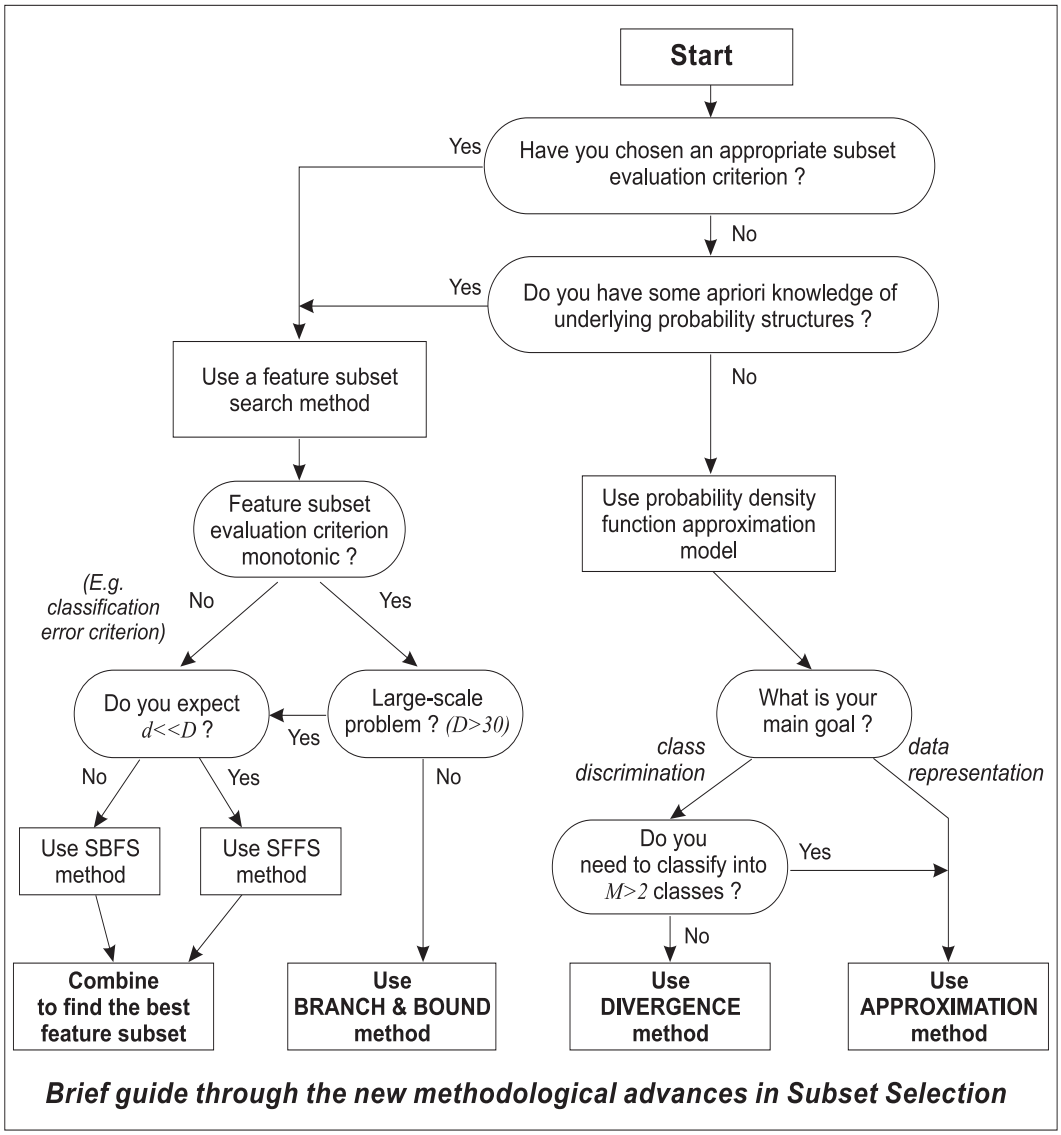
A software package integrating the described methods (plus other methods currently used) and the FS Consulting System is being developed for both MS Windows and Unix platforms.

(Received December 18, 1997.)

REFERENCES

[1] E. Backer and J. A. De Schipper: On the Max–Min approach for feature ordering and selection. In: Seminar on Pattern Recognition, 2.4.1., Liege University, Sart–Tilman 1977

[2] D. E. Boekee and J. C. A. Van der Lubbe: Some aspects of error bounds in feature selection. Pattern Recognition *11* (1979), 252–360.



Brief guide through the new methodological advances in Subset Selection

[3] P. Devijver and J. Kittler: Pattern Recognition: A Statistical Approach. Prentice 1982.

[4] F. J. Ferri, P. Pudil, M. Hatef and J. Kittler: Comparative study of techniques for large–scale feature selection. In: Pattern Recognition in Practice IV (E. S. Gelsema and L. N. Kanal, eds.), Elsevier 1994, pp. 403–413.

[5] J. Kittler: Feature selection and extraction. Handbook of Pattern Recognition and Image Processing (T. Y. Young and K. S. Fu, eds.), Academic Press, New York 1986, pp. 60–81.

[6] P. M. Narendra and K. Fukunaga: A Branch and Bound Algorithm for feature subset

selection. IEEE Trans. Computers *C-26* (1977), 917–922.
[7] J. Novovičová, P. Pudil and J. Kittler: Feature selection based on divergence for empirical class densities. In: Proc. of the 9th Scandinavian Conf. on Image Analysis, Uppsala 1995.
[8] J. Novovičová, P. Pudil and J. Kittler: Divergence based feature selection for multimodal class densities. IEEE Trans. Pattern Recognition Machine Intelligence *18* (1996), 2, 218–223.
[9] J. Novovičová and P. Pudil: Feature selection and classification by modified model with latent structure. In: Dealing With Complexity: Neural Network Approach. Springer Verlag, Berlin 1997, pp. 126–140.
[10] P. Pudil, S. Bláha and J. Novovičová: PREDITAS – software package for solving pattern recognition and diagnostic problems. In: Proc. BPRA 4th Internat. Conference on Pattern Recognition, Cambridge (J. Kittler, ed.), Springer–Verlag, Berlin 1988, pp. 146–152.
[11] P. Pudil, J. Novovičová, N. Choakjarernwanit and J. Kittler: An analysis of the Max–Min approach to feature selection. Pattern Recognition Lett. *14* (1993), 11, 841–847.
[12] P. Pudil, J. Novovičová, N. Choakjarernwanit and J. Kittler: The Max–Min approach to feature selection: Its foundations and potential. Indian J. Pure Appl. Math. *24* (1994), 11, 69–81.
[13] P. Pudil, J. Novovičová and J. Kittler: Automatic machine learning of decision rule for classification problems in image analysis. In: Proceedings of BMVC '93 – the 4th British Machine Vision Conference, 1993.
[14] P. Pudil, J. Novovičová and J. Kittler: Simultaneous learning of decision rules and important attributes for classification problems in image analysis. Image and Vision Computing *12* (1994), 3, 193–198.
[15] P. Pudil, F. Ferri, J. Novovičová and J. Kittler: Floating search methods for feature selection with nonmonotonic criterion functions. In: Proc. of the 12th IAPR Intern. Conf. on Pattern Recognition, Jerusalem 1994, IEEE Comp. Society Press, pp. 279–283.
[16] P. Pudil, J. Novovičová and J. Kittler: Floating search methods in feature selection. Pattern Recognition Lett. *15* (1994), 1119–1125.
[17] P. Pudil, J. Novovičová, N. Choakjarerwanit and J. Kittler: Feature selection based on the approximation of class densities by finite mixtures of special type. Pattern Recognition *28* (1995), 9, 1389–1397.
[18] P.Pudil, J.Novovičová and F. J. Ferri: Methods of dimensionality reduction in statistical pattern recognition. In: Proceedings of the IEEE European Workshop CMP'94, Prague 1994, Institute of Information Theory and Automation, pp. 185–198.
[19] W. Siedlecki and J. Sklansky: On automatic feature selection. Internat. J. Pattern Recognition and Artificial Intelligence *2* (1988), 2, 197–220.
[20] D. Zongker and A. Jain: Algorithms for feature selection: An evaluation. In: Proceedings of 13th International Conference on Pattern Recognition, Vienna 1996, Vol. II, Track B, pp. 18–22.

Ing. Pavel Pudil, CSc., RNDr. Jana Novovičová, CSc., Mgr. Petr Somol, and Ing. Radek Vrňata, Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 182 08 Praha 8 and Laboratory of Faculty of Management, Jindřichův Hradec. Czech Republic.
e-mails: pudil,novovic,somol,vrnata@utia.cas.cz