Kybernetika

STATISTICAL APPROACH TO PATTERN RECOGNITION

Theory and Practical Solution by Means of PREDITAS System

PAVEL PUDIL, JANA NOVOVIČOVÁ, SVATOPLUK BLÁHA

ACADEMIA
PRAHA

Statistical pattern recognition has developed into a self-contained mature discipline which represents one of two principal approaches to the solution of pattern recognition problems. The concept of pattern is to be looked upon in a rather broad sense since pattern recognition methods have been successfully applied to very diverse application domains. The final goal in statistical pattern recognition is classification, i.e. assignment of a pattern vector to one of a finite number of classes, where each class is characterized by a probability density function on the measured features. In statistical pattern recognition which is sometimes regarded also as a geometric approach, a pattern vector is viewed as a point in the multidimensional space defined by the features. Though the classification is the primary goal of pattern recognition there is a number of related problems requiring careful attention when solving problems from real life. Among these problems, a thorough consideration of which proves to be crucial for the overall performance of any pattern recognition system, the following problems belong: evaluation of the training set quality, feature selection and dimensionality reduction, estimation of classification error, iterative corrections of individual phases of the solution according to the results of testing, and finally the problem of interconnecting feature selection with the classifier design as much as possible. An attempt to provide a complex solution of all these interconnected problems has resulted in the design of PREDITAS (Pattern REcognition and DIagnostic TAsk Solver) software package. It is a combination of both theoretically based and heuristic procedures, incorporating as much as possible requirements and suggestions of specialists from various application fields. Theoretical background of the employed methods and algorithms, together with their reasoning and some examples of applications is presented.

REFERENCES

[1] T. W. Anderson, An Introduction to Multivariate Statistical Analysis. John Wiley, New York 1958.

[2] G. Biswas, A. K. Jain and R. Dubes: Evaluation of projection algorithms. IEEE Trans. Pattern Anal. Machine Intell. *3* (1981), 701—708.

[3] S. Bláha, P. Pudil and R. Pecinovský: Classification by sequential discriminative rule and its optimization by measure of discriminative power. In: Proceedings of DIANA — Conf. of Discr. Anal., Cluster Anal. and Others Methods on Data Class, Liblice 1982, pp. 277—284.

[4] S. Bláha and P. Pudil: A general approach to diagnostic problem solution by means of pattern recognition. Problems Control Inform. Theory *13* (1984), 3, 192—208.

[5] S. Bláha and P. Pudil: The PREDITAS system and its use for computer-aided medical decision making. In: Medical Decision Making: Diagnostic Strategies and Expert Systems (J. H. Van Bemmel, F. Grémy, J. Zvárová, eds.), North-Holland, Amsterdam 1985, pp. 215—218.

[6] S. Bláha, J. Novovičová and P. Pudil: Solution of Pattern Recognition Problem by Means of the PREDITAS Program System. Part I.: Dichotomic Classification — Theoretical Background, Research Report ÚTIA ČSAV No. 1549, Prague 1988.

[7] S. Bláha, J. Novovičová and P. Pudil: Solution of Pattern Recognition Problem by Means of the PREDITAS Program System. Part II.: Feature Selection and Extraction Principles and Used Methods. Research Report ÚTIA ČSAV No. 1555, Prague 1988.

[8] S. Bláha, J. Novovičová an P. Pudil: Solution of Pattern Recognition Problem by Means of the PREDITAS Program System. Part III: Sample-Based Classification Procedures. Research Report ÚTIA ČSAV No. 1593, Prague 1989.

[9] S. Bláha, P. Pudil and F. Patočka: Program system PREDITAS and its application in geology (in Czech). In: Proceedings of International Symposium Mathematical Methods in Geology, Příbram 1989, pp. 6—17.

[10] C. K. Chow: An optimum character recognition system using decision functions. IRE Trans. Electronic Computers *EC-6* (1957), 6, 247—254.

[11] C. K. Chow: On optimum recognition error and reject tradeoff. IEEE Trans. Inform. Theory *IT-16* (1970), 1, 41—46.

[12] T. M. Cover and J. M. Van Campenhout: On the possible orderings in the measurement selection problem. IEEE Trans. Systems, Man Cybernet. *7* (1977), 657—661.

[13] H. P. Decell and L. T. Guseman: Linear feature selection with applications. Pattern Recognition *11* (1979) 55—63.

[14] P. A. Devijver and J. Kittler: Pattern Recognition — A Statistical Approach. Prentice-Hall, Engelwood Cliffs 1982.

[15] R. Dubes and A. K. Jain: Clustering metodology in exploratory data analysis. In: Advances in Computers 19, Academic Press, New York 1980.

[16] R. O. Duda and P. E. Hart: Pattern Classification and Scene Analysis. J. Wiley, New York 1973.

[17] B. Efron: Bootstrap methods. Another look at the jackknife. Ann. Statist. *7* (1979), 1—26.

[18] B. Efron: The Jackknife, the Bootstrap and Other Resampling Plans. Society for Industrial and Applied Mathematics, Philadelphia 1982.

[19] R. A. Fisher: The use of multiple measurements in taxonomic problems. Ann. Eugenics *7* (1936), Part II, 179—188.

[20] R. A. Fisher: Statistical Methods for Research Workers. Hafner, New York 1963.

[21] D. H. Foley: Considerations of sample and feature size: IEEE Trans. Inform. Theory *IT-18* (1972), 5, 618—626.

[22] K. S. Fu: Applications of Pattern Recognition (K. S. Fu, ed.) CRC Press 1982.

[23] K. S. Fu: A step towards unification of syntactic and statistical pattern recognition. IEEE Trans. Pattern Anal. Machine Intell. *8*, (1986), 398—404.

[24] K. Fukunaga and R. R. Hayes: Effects of sample size in classifier design. IEEE Trans. Pattern Anal. Machine Intell. *11* (1989), 8, 873—885.

[25] K. Fukunaga and R. R. Hayes: Estimation of classifier performance. IEEE Trans. Pattern Anal. Machine Intell. *11* (1989), 10, 1087—1101.

[26] N. Glick: Sample-based classification procedures derived from density estimators. J. Amer. Statist. Assoc. *67* (1972), 116—122.

[27] L. F. Guseman, Jr. and H. F. Walker: On minimizing the probability of misclassification for linear feature selection. JSC International Technical Note JSC-08412, Johnson Space Center, Houston, Texas, August 1973.

[28] L. F. Guseman, Jr. and H. F. Walker: On Minimizing the Probability of Misclassification for Linear Feature Selection: A Computational Procedure. The Search for Oil. Marcel Dekker, New York 1975.

[29] L. F. Guseman, Jr., B. C. Peters, Jr. and H. F. Walker: On minimizing the probability of misclassification for linear feature selection. Ann. Statist. *3* (1975), 661.
[30] D. J. Hand: Recent advances in error rate estimation. Pattern Recognition Lett. *4* (1986), 335—346.
[31] M. M. Kalayeh and D. A. Landgrebe: Predicting the requirement number of training samples. IEEE Trans. Pattern Anal. Machine Intell. *5* (1983), 664—667.
[32] L. Kanal: Patterns in pattern recognition 1968—1974. IEEE Trans. Inform. Theory *IT-18* (1974), 618—626.
[33] L. Kanal and B. Chandrasekar: On dimensionality and sample size in statistical pattern classification. Pattern Recognition *3* (1971), 225—234.
[34] P. A. Lachenbruch. Discriminant Analysis. Hafner Press, London 1975.
[35] P. A. Lachenbruch and R. M. Mickey: Estimation of error rates in discrimining analysis. Technometrics *10* (1968), 1, 1—11.
[36] P. M. Lewis: The characteristic selection problem in recognition systems. IRE Trans. Inform. Theory *8* (1962), 171—178.
[37] W. Malina: On an extended Fisher criterion for feature selection. IEEE Trans. Pattern Anal. Machine Intell. *3* (1981), 611—614.
[38] T. Marill and D. M. Green: On the effectivness of receptors in recognition systems. IEEE Trans. Inform. Theory *9* (1963), 1, 11—17.
[39] P. M. Narendra and K. Fukunaga: A branch and bound algorithm for feature subset selection. IEEE Trans. Comput. *26* (1977), 917—922.
[40] N. J. Nilsson: Learning Machine — Foundations of Trainable Pattern Classifying Systems. McGraw-Hill, New York 1965.
[41] R. Pecinovský, P. Pudil and S. Bláha: The algorithms for sequential feature selection based on the measure of discriminative power. In: Proceedings of DIANA — Conf. on Discr. Anal., Cluster Anal. and Others Methods on Data Class., Liblice, 1982, pp. 277—284.
[42] P. Pudil and S. Bláha: Evaluation of the effectiveness of features selected by the methods of discriminant analysis. Pattern Recognition *14* (1981), Nos. 1—6, 81—85.
[43] P. Pudil and S. Bláha: A global approach to the solution of situation recognition. In: Fourth Formator Symposium on Mathematical Methods for Analysis of Large-Scale Systems, Liblice, May, 1982 (J. Beneš, L. Bakule, eds.). Academia, Praha 1983, pp. 405—418.
[44] P. Pudil, S. Bláha and J. Novovičová: PREDITAS — software package for solving pattern recognition and diagnostic problems. In: Pattern Recognition — Proceedings of BPRA 4th Internat. Conf. on Pattern Recognition, Cambridge 1988 (J. Kittler, ed.). (Lecture Notes in Computer Science 301.) Springer-Verlag Berlin—Heidelberg—New York 1988, pp. 146—152.
[45] P. Pudil, S. Bláha and Z. Pertold: Significance analysis of geochemical data for rock type discrimination by means of PREDITAS system (in Czech). In: Proceedings of International Symposium Mathematical Methods in Geology, Příbram 1989, pp. 119—125.
[46] G. Sebestyen: Decision Making Processes in Pattern Recognition. MacMillan, New York 1962.
[47] G. T. Toussaint: Bibliography on estimation of misclassification. IEEE Trans. Inform. Theory *IT-20* (1974), 4, 472—479.
[48] S. Watanabe: Karhunen-Loève expansion and factor analysis. In: Trans. Fourth Prague Conf. on Information Theory, 1965. Academia, Prague 1967, pp. 635—660.
[49] W. G. Wee: Generalized inverse approach to adaptive multiclass pattern classification. IEEE Trans. Comput. *17* (1968), 1157—1164.

1. INTRODUCTION

Pattern recognition, originated by Rosenblatt's pioneer work on perception, has extended a long time ago into other directions and developed into a self-contained discipline, though it utilizes many theoretical results from probability theory, mathematical statistics and theory of decision functions.

However, to identify pattern recognition with any of these earlier established disciplines would be a great misunderstanding, which unfortunately is not very unusual. This misunderstanding follows from the fact that pattern recognition is often considered only as a decision problem. Though the solution of this decision problem itself represents really the most essential part of the overall problem, the complex solution is far more complicated.

There are many other partial but very important problems which are not treated sufficiently in the literature on pattern recognition, namely the quality of so called training set, optimal data representation in a lower-dimensional subspace, testing of the complete sequence of computer programs in practice, quality of decision making evaluation, and finally corrections of individual phases according to the results of testing.

Before proceeding further, we should attempt to clarify what is actually understood by pattern recognition. First of all the notion of pattern itself needs to be specified. In their widest sense, patterns are the means by which we interpret the world. In our childhood we learnt to distinguish the visual patterns of persons and things, the aural patterns of speech and music, the tactile patterns of cold and warmth, patterns of the senses. Later on we were able to refine the details of our pattern recognition and to abstract our sensory discrimination to indirect patterns. While some patterns, like typed or handwritten characters, the electromagnetic signals of radar, ECG or EEG signals, have a physical manifestation, other patterns have only an abstract existence, e.g. patterns in social or economic data.

Having clarified how the world pattern is to be interpreted, we can discuss the question: What is the pattern recognition process? When a human glances at a printed page and recognizes character after character, he is utilizing his past accumulated experience which has been somehow transformed into fixed decision rules. He is generally unable to explain those rules, however, their existence is out of question. The mechanism of arriving at these rules is quite obvious. It has been necessary to expose him in the "learning phase" to samples of respective characters and to tell him which they were. In other words to provide him with so called labelled samples from which he developed a decision rule.

We can see that there are two aspects to pattern recognition – developing a decision rule and using it. The actual recognition occurs in the stage of using the rule.

We have to be aware of the fact that the pattern recognition problem can be defined only if we state clearly what we wish to decide, by defining the pattern classes. A pattern recognition problem thus begins with class definitions, which should be

given a thorough consideration since in certain applications the choice of class definition may greatly influence the overall performance of recognition system.

With computers becoming widely used, the field of pattern recognition ceased to be the process specific exclusively to the human brain. However, there is an essential difference between the human's recognition and "mathematical" (statistical) pattern recognition by means of computers. Methods employed in mathematical pattern recognition are based on relatively simple concepts. As opposed to human's recognition, their possible success does not depend upon sophistication relative to the human, but upon the computer's ability to store and process large numbers of data in an exact manner and above all upon the computer's ability to work in high dimensions. This is an essential priority over the human's abilities since the humans can do very sophisticated things in three dimensions or less, but begin to falter when dealing with higher dimensions.

In mathematical or computerized recognition we can speak about pattern recognition techniques. They assign a physical object or an event to one of several prespecified categories. Thus, a pattern recognition system can be viewed as an automatic decision rule which transforms measurements on a pattern into class assignments.

The patterns themselves, recognized by these techniques, could range from geological sites or agricultural fields in a remotely sensed image from satellites to a speech waveform; the associated recognition or classification problems are to label an agricultural field as wheat or non-wheat or a geological site as containing or not containing certain ore deposits, and to identify the spoken word, respectively. Patterns are described by measurements made on the pattern or by features derived from these measurements. The physical interpretation of features varies with applications. While in speech recognition the so called linear predictive coefficients computed from the speech waveform are used as features, in medical diagnostic and recognition problems we often use directly a set of data on a patient as features, without any preliminary computation or transformation.

The development of pattern recognition techniques together with growing availability of computers has lead to extending the application of pattern recognition to new domains. To mention just some of them, industrial inspection, document processing, remote sensing, personal identification belong to this category. A detailed account of many of these applications is given in [22].

There are essentially two basic approaches to pattern recognition. In a geometric or statistical approach, a pattern is represented in terms of M features or properties and viewed as a point in M-dimensional space. The first task is to select those features such that pattern vectors of different categories will not overlap in the feature space (this ideal cannot be generally reached, so a near optimum solution is sought). Given certain labelled sample patterns from each pattern class (training samples), the objective is to establish decision boundaries in the feature space which would separate as much as possible patterns belonging to different classes. Either the direct

statistical approach can be adopted when the decision boundaries are determined by the statistical distributions of the patterns (known or estimated), or a "non-statistical" approach is utilized. In the latter case, first the functional form of the decision boundary is specified (linear, piecewise linear, quadratic, polynomial, etc.) and afterward the best decision boundary of the specified functional form is found.

The choice of features is crucial for the quality of pattern recognition. Participation of an expert from the corresponding application area in designing the primary set of features is absolutely essential since the choice of features is data dependent. Mathematical processing can help to remove the redundant information of the decision problem, but just by itself it obviously cannot supply any completely "new" information not supplied by the expert designing the original set of features.

Beside the statistical approach to pattern recognition there is another approach being studied. In many recognition problems it is more appropriate to consider a pattern as being composed of simple subpatterns, which can again be built from yet simpler subpatterns, etc. The simplest subpatterns can be characterized by the interrelationships among those primitives. The need to define primitives and rules of constructing patterns from these primitives arises. Since an analogy is drawn between the structure of patterns and the syntax of a language, the whole approach has been named the *syntactic* pattern recognition. To make the analogy more explicit we can state that the patterns are viewed as sentences of a language, primitives are viewed as the alphabet of the language, and generating of sentences is governed by a grammar. The main advantage of the syntactic or structural approach is that, in addition to classification, it also provides a description how the given pattern is constructed from the primitives.

However, the implementation of a syntactic approach is not without difficulties either. While neither the statistical approach is free of difficulties, it is better understood and is based upon well established elements of statistical decision theory. This is perhaps the reason why most commercially used pattern recognition systems utilize decision-theoretic approaches. An interesting attempt to combine both the approaches has been made by Fu [23] who has introduced the notion of attributed grammars which unifies the syntactic and statistical pattern recognition.

Now we shall describe the organization of the paper:

After explaining the general structure of PREDITAS system resulting from the endeavour to meet the requirements and conditions imposed by practice, the common conventions are introduced. In Chapter 2 we give an outline of the elementary statistical decision theory and we show how the theory provides statistical model for pattern generating mechanisms and the corresponding optimal classification process. Optimality can be achieved only when the statistical distributions involved are known. We also present some ways of designing discriminant functions. Chapter 3 is devoted to the theoretical background of feature selection and extraction methods as well as to the comparison of feature selection and extraction from the practical

viewpoint. Basic properties of so called search algorithms are discussed too. Chapter 4 deals with the principle of feature selection used in the PREDITAS system as well as with the employed search procedures. Chapter 5 addresses the important and difficult question of estimating the performance of pattern recognition. Chapter 6 deals with sample-based classification rules which are from the perspective of statistical decision theory substitutes for unknown optimal rules. Finally the stepwise sample-based decision procedure for two class classification problem and the determination of the optimal dimensionality used in PREDITAS system is treated in Chapter 7. A brief account of the software tools and the architecture of PREDITAS is presented in Chapter 8, together with a more detailed discussion of a completely solved problem from the field of medical diagnostics. The paper concludes with the list of solved problems accompanied by their brief characteristics.

1.1 General characteristics of PREDITAS system

Growing demands for solving a great variety of practical problems of a decision-making character, which cannot be solved for some reason analytically, stimulated the development of a general purpose software package for their solution, which would aim to be as far as possible universal and problem free. A statistical pattern recognition approach to the solution of this class of problem has been adopted, in accordance with general practice.

Since our principal application field has been medicine, where the well established term "diagnostic problem" [4], [43] is commonly used, we have incorporated this term into the name of the software package "PREDITAS" which stands for Pattern REcognition and DIagnostic TAsk Solver [5]. Though a general medical diagnostic problem can be formulated in terms of statistical pattern recognition, we use the term "diagnostic" explicitly in order to stress the formal analogy.

There exists a broad spectrum of statistical pattern recognition methods that differ widely in their assumptions about data set structure, in their computational complexity, etc. Our goal has not been to develop a software package that would be based on completely new original methods, but rather to design a practicable system, incorporating certain new ideas into the already known theoretical results.

Due to the impossibility of an exact mathematical formulation of each phase of the overall solution, our approach is a combination of both theoretically based and heuristic procedures. It is the result of a long experience in solving various pattern recognition problems in close cooperation with specialists in various application fields [9], [45], whose valuable opinions and comments have been incorporated in the present version of the PREDITAS system.

1.1.1 Requirements and constraints imposed by practice

Frequent discussions with potential users of the software package have resulted in a number of constraints, imposed by the needs and requirements of practice.

A system with ambitions to be useful for everyday practice should fulfill these requirements, and should be designed with respect to the corresponding constraints. Without claiming completeness, let us state briefly at least the principal conditions [44] to be met:

1) The resultant solution supplied to the user should be as simple as possible, computable quickly even on a very small personal computer, easily interpretable and finally, perhaps most importantly, it should leave some space for the user himself, allowing him to take the final decision and thus to participate in the solution.

2) A reject option, equivalent to not taking a decision in the case where the probability of misclassifying a certain pattern would be higher then an apriori set limit, should be included in the system. It should also be possible to set these limits independently for both classes in terms of probabilities of misclassification, or in terms of admissible average risk if the latter is preferred.

3) The sample size is usually too small for a given dimension of original data. It results in the necessity of feature extraction or selection.

4) The dimension of original measurement can be rather high, in quite a few applications it was of the order of one hundred. A feature extraction or selection algorithm should be able to cope with this dimension.

5) Components of original multidimensional measurement vectors cannot be supposed to be statistically independent. This fact must be respected by the feature selection algorithm.

6) Experts from application fields prefer decision-making on the basis of meaningful interpretable features to decision-making by means of some mathematical abstractions. This issue is especially sensitive in most medical applications. Priority has therefore been given to feature selection instead of feature extraction, even if the latter can generally yield a slightly more discriminative set of features. Moreover, only feature selection enables one to exclude completely some of the original measured variables and thus to cut the cost of data acquisition.

7) The system should not only provide a list of features ordered according to their significance for classification (respecting of course complicated multidimensional statistical relations among them), but should help as well in choosing the optimum number of features for classification (the dimensionality of classifier). Since the answer to this partial but important problem cannot be always unique and theoretically justified, the user should be provided at least with some heuristic procedure allowing him to assess a few possible solutions offered by some system and to choose the most suitable one according to his own scale of preferences.

8) The algorithmic and computational complexity of the system of programs should not prevent the finding of a solution within a reasonable time limit even for high dimensions.

9) Finally, the two principal stages of the solution – feature selection and deriva-

tion of the classification rule should be interconnected as much as possible and not solved independently, which is usually the case. This condition has not been raised by practice, but it follows from the experience of leading experts in the pattern recognition field (Kanal [32]).

In connection with these conditions there is one point to be mentioned. In a multiclass problem, optimum feature subsets for discrimination of two particular classes are generally not the same for different pairs of classes. For this reason only a dichotomic problem is considered in the PREDITAS system. In case of a multiclass problem a decision system having a hierarchical structure can be built.

The philosophy and the corresponding architecture of the PREDITAS system have resulted from the endeavour to fulfill the above named conditions. These conditions have greatly influenced not only the algorithmic realization but also the underlying theory employed in PREDITAS, which will be described in the Chapter 2 through 7.

1.1.2 General structure of PREDITAS system and block-diagram of solution

As we have just mentioned above, in our opinion it is therefore desirable to adopt some basic principle unifying from a certain point of view all the main phases of the solution. This unifying principle should be incorporated properly into the solution of individual problems, respecting of course the differences among them.

In order to illustrate more clearly what we mean by this unifying principle, it will be convenient to survey briefly the individual phases of the problem and their solution (see Fig. 1).

In further discussion we shall follow the block diagram of the complex solution according to our conception (Fig. 1).

The complex solution starts with the primary data collection. This phase is generally considered only as a purely technical problem and is left to the specialists in the particular application field. Since the quality of data is crucial for the possibility of successfull solution and the data should meet certain requirements of pattern recognition, in our opinion mathematicians and computer specialists should participate already in this phase of the problem.

The phase of preliminary data processing includes the formation of so called training set. This set usually consists of a number of samples with known classification, where the samples or patterns are generally represented by M-dimensional vectors.

Here a very important problem arises, namely how to evaluate the quality of the training set.

This quality can be assessed from different points of view. The most important one among them is the concept of "representativeness" of the training set, because the derived decision rule will be used outside its definition domain, i.e. the training set. Such an extrapolation of the decision rule is possible only if the training set, as a finite size sample from the basic set, is representative with respect to the basic set. In the opposite case the satisfactory solution of the whole problem cannot be expected.

However, the problem of evaluating the representativeness of the training set has not yet been satisfactorily solved.

Generally the quality of the training set should be tested according to a certain criterion and when not found satisfactory, it has to be improved before proceeding in the solution further.

The following block denoted the "data analysis" includes a basic statistical analysis of data components and the evaluation of their significance. Already in this phase we can often get some very interesting results, which are useful for the particular application field. Now the two most essential phases of the diagnostic problem follow – feature selection and derivation of classification rule. Pattern recognition is usually identified with these two stages of the solution. Because of their importance we shall treat them in more detail in the next two sections.

Feature extraction or selection generally denotes the process of finding the most important characteristics from given data vectors which are optimal with respect

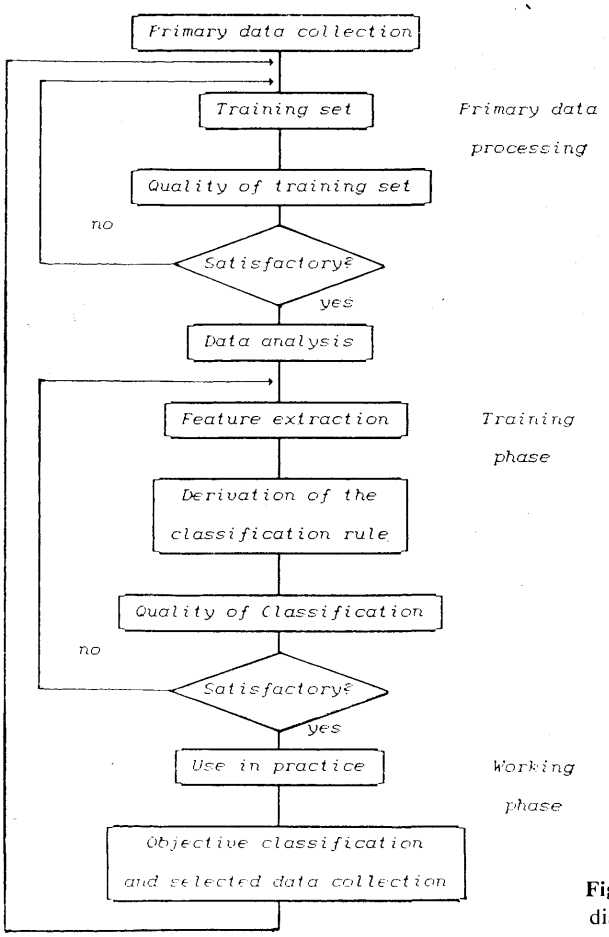

Fig. 1. Block diagram of diagnostic problem solution.

to a certain criterion. Formally it means an optimal mapping of the *pattern space* into *feature space*; the dimensionality of the latter should be lower. Feature extraction methods find a subset of features which are functions of the original ones, while feature selection methods find a subset of the original components [7].

This problem is usually solved quite independently from the derivation of the decision (classification) rule, though both these phases should be closely connected, as Kanal [32] states in his survey paper. The main reason is that the classification significance of the feature can be evaluated only by the differences of classification errors obtained in the cases when the particular feature was or was not used for discrimination.

The third of the main problems consists in determination of a decision rule [6], [7], supposing that the set of the features characterizing the elements is given. Solution of both these problems. determination of the decision rule and feature reduction should be based on a common unifying principle. In other words, the decision rule should be in accordance with the criterion used for feature selection. In the opposite case the selected features may not be optimal for the decision rule.

Finally the complete solution continues by testing in practical applications and if necessary again by returning to the original phase.

As it is not possible to solve all the three main phases as one problem, it seems to be useful to incorporate the feedback into the algorithm of solution and to solve the whole problem by an iterative way.

The use of the *measure of discriminative power* [42] is the unifying principle we mentioned a while ago. It can be used even in the phase of testing the quality of the training set, where the convergence of this measure serves as the criterion of quality and reliability. The best subset of features maximizing the measure of discriminative power is selected in the phase of feature selection and this measure is used for the derivation of classification rule as well.

Thus in our conception the whole solution represents an iterative process with a number of feedback loops and with some underlying unifying principle, which in our case is the measure of discriminative power. By the global approach [43] to diagnostic problem solution we understand just this iterative process with an unifying principle. Generally any other suitable unifying principle may be used, but its incorporation into the overall solution is essential.

1.2 Notational conventions

1.2.1 Common conventions

(3.2) ... Eq. (2) in Chap. 3.
$\hat{=}$ ... means "is equal to by definition"
iff ... if and only if
$\Rightarrow$ ... implies

$\in$ ... is in

$\subset$ ... inclusion

$\forall$ ... for all

$a \mid b$ ... a conditioned on given b

a/b ... a divided by b

$\binom{a}{b}$ ... $\dfrac{a!}{b!\,(a-b)!}$, binomial coefficients

$\bigcup_{i=1}^{c} \Omega_i$ ... means "union of sets $\Omega_1, \ldots, \Omega_c$"

$\log a$ ... natural logarithm of a

1.2.2 Notation for observation vectors

x ... scalar

x_i ... subscribed scalar

$\mathbf{x}$ ... column vector (if $\mathbf{x}$ is m-dimensional, $\mathbf{x} = (x_1, \ldots, x_m)^T$

$\mathbf{x}_i$ ... subscribed vector

$\mathbf{A}, \boldsymbol{\Sigma}$ ... matrices

$\mathbf{x}^T, \mathbf{A}^T$ ... $\mathbf{x}$, $\mathbf{A}$ transpose

$\mathbf{A}^{-1}$ ... inverse matrix

$\mathrm{tr}\,(\mathbf{A})$ ... trace of the matrix $\mathbf{A}$

$|\mathbf{A}|$ ... determinant of the matrix $\mathbf{A}$

$\mathbb{R}$ ... one-dimensional Euclidean space

$\mathbb{R}^m$ ... m-dimensional Euclidean space

Ω ... m-dimensional feature space; $\mathbf{x} \in \Omega$. Usually $\Omega = \mathbb{R}^m$.

X, Y, F ... sets, elements of which are scalar or, sometimes scalars

$\emptyset$ ... the empty set

ω_i ... ith class (or category)

$\int g(\mathbf{x})\,d\mathbf{x}$ ... multivariate and multiple integration is performed over the entire Ω space

1.2.3 Notation for decision rules

The difference in writing between a random vector and its realization is omitted.

$p(\mathbf{x})$ ... probability density function of random vector $\mathbf{x}$

$p_i(\mathbf{x})$ ... probability density function of $\mathbf{x}$ given that it is from class ω_i. Also called ith conditional probability density

P_i ... a priori probability of class ω_i

$P(\omega_i \mid \mathbf{x})$ ... probability that a pattern with observation $\mathbf{x}$ belongs to class ω_i. Using previous definitions,

$$P(\omega_i \mid \mathbf{x}) = \frac{P_i p_i(\mathbf{x})}{\sum_i P_i p_i(\mathbf{x})}.$$

$\mathsf{E}\{\cdot\}$ ... mathematical expectation

$\mathsf{E}_i\{\cdot\}$ ... mathematical expectation with respect to class ω_i

$\boldsymbol{\mu}_i$ ... mean vector of $\mathbf{x}$ in class ω_i

$\boldsymbol{\Sigma}_i$ ... covariance matrix of $\mathbf{x}$ in class ω_i

c ... number of classes

$d(\mathbf{x})$ ... decision (or classification or discrimination) rule assigns $\mathbf{x}$ to class $d(\mathbf{x})$

$d(\mathbf{x}) = d_i$ ... decision rule d assigns $\mathbf{x}$ to class ω_i

$d(\mathbf{x}) = d_0$ ... decision rule d rejects to classify $\mathbf{x}$

Ω_i ... decision rule d induces $c + 1$ sets $\{\Omega_0, \Omega_1, \ldots, \Omega_c\}$, where $\mathbf{x} \in \Omega_i$ iff $d(\mathbf{x}) = d_i$

$\mathscr{D}$ ... collection of all decision rules

L_{ij} ... loss or cost incurred when the decision is $d(\mathbf{x}) = d_i$ and the true class is ω_j, $i = 0, \ldots, c, j = 1, \ldots, c$

$(0, 1, \lambda_r)$ ... $(0, 1, \lambda_r)$ loss function means that $L_{ii} = 0$, $L_{ij} = 1$, $i \neq j$, $L_{0i} = \lambda_r$, $i, j = 1, \ldots, c$

$r_i(\mathbf{x})$ ... conditional risk or expected loss of making decision $d(\mathbf{x}) = d_i$

$R(d)$ ... average risk associated with the decision rule d

$E(d)$ ... error rate or overall probability of error or probability of misclassification

$R_r(d)$ ... reject rate or reject probability

$r^*(d)$ ... minimum conditional risk

$R^*(d)$ ... minimum average risk

1.2.4 Notation for sample-based decision rules

$\mathbf{x}_{ij}$ ... jth vector sample from class ω_i

$\boldsymbol{T}_i$ ... ith training set; collection of N_i independent vector in Ω identically distributed

$\boldsymbol{N}_i$ ... number of training samples from class ω_i; size of training set $\mathbb{T}_i$

$\boldsymbol{T}_{(N)}$ ... collection of c training sets; $\mathbb{T}_{(N)} = \{\mathbb{T}_1, \ldots, \mathbb{T}_c\}$

$\boldsymbol{N}$ ... $N = \sum_{i=1}^{c} N_i$; total number of vectors from all c classes; size of training set $\mathbb{T}_{(N)}$

$\hat{\boldsymbol{d}}(\mathbf{x})$ ... sample-based decision rule;

$\hat{\boldsymbol{\Omega}}_i$ ... the decision rule $\hat{d}(\mathbf{x})$ induces $c + 1$ sets $\{\hat{\Omega}_1, \hat{\Omega}_2, \ldots, \hat{\Omega}_c\}$, where $\mathbf{x} \in \hat{\Omega}_i$ iff $\hat{d}(\mathbf{x}) = d_i$

$\boldsymbol{E}(\hat{\boldsymbol{d}})$ ... actual error rate (error rate of the sample-based decision rule $\hat{d}$)

$\hat{E}(\hat{d})$ ... apparent error rate of rule $\hat{d}$

$R(\hat{d})$ ... actual average risk of rule $\hat{d}$

$\hat{R}(\hat{d})$ ... apparent average risk of rule $\hat{d}$

1.2.5 Notation for dimensionality reduction

Observation space or measurement space: (M-dimensional)	Space for a finite number of observations $x_1, \ldots, x_M$
Feature space: (m-dimensional) $m \leqq M$	Space with points $\mathbf{y} = (y_1, \ldots, y_m)^T$, where y_i is the ith feature obtained as a function of the observations $x_1, \ldots, x_M$
$\lambda(\mathbf{y})$	Measure of discriminant power (MDP) of the vector $\mathbf{y}$. $\lambda: \mathbb{R}^m \to \mathbb{R}$.
Decision space: (one-dimensional)	Set of real numbers indexing the classes which can be decided.

2. STATISTICAL PATTERN CLASSIFICATION

In this chapter there are presented only those theoretical facts from the area of statistical pattern classification, which are relevant to the conception of PREDITAS system.

2.1 Basic elements of statistical classification

The basic idea of statistical pattern classification can be summarized as follows:

Suppose that several classes (or categories $\omega_1, \omega_2, \ldots, \omega_c$ are mixed in a large population. Denote the mixing proportions (a priori class probabilities) by $P_1, P_2, \ldots, P_c$, where each $P_i \geqq 0$ and $\sum_{i=1}^{c} P_i = 1$. *Individuals* (*objects, patterns*) can be examined with respect to each of m-characteristics or *features*, $x_1, \ldots, x_m$, quantitative or qualitative. An individual selected at random from the mixed population may be regarded as a random vector $(i, \mathbf{x})$ in which i indexes the individual's class and $\mathbf{x}$ denotes the m-dimensional *feature vector* or *classification vector*. To *classify* (or *allocate* or *identify*) an individual is to guess his class, given the observed $\mathbf{x}$. Probability structure is specified by a priori class probability P_i and by multivariate class – conditional probability density function $p_i(\mathbf{x}) \triangleq p(\mathbf{x}) \mid \omega_i)$ for $i = 1, 2, \ldots, c$ and $\mathbf{x} \in \Omega$ where Ω denotes the space of all such $\mathbf{x}$ – so called *feature space*; $\Omega \subset \mathbb{R}^m$. The unconditional density of $\mathbf{x}$ is given by $p(\mathbf{x}) = \sum_{i=1}^{c} P_i p_i(\mathbf{x})$ and it is assumed that $p(\mathbf{x})$ is nonzero over the entire space Ω.

In the sequel we use the notation "pattern $\mathbf{x}$" for the individual (pattern, object) described by the feature vector $\mathbf{x}$. Further the difference in writing between a random vector and its realization is omitted.

The problem of classifying a new pattern $\mathbf{x}$ can now be formulated as a statictical decision problem (testing of hypothesis) by defining a *decision rule* (or *classification rule* o *discrimination rule*) $d(\mathbf{x})$ where $d(\mathbf{x}) = d_i$ means that the hypothesis "$\mathbf{x}$

is from class ω_i" is accepted (cf. e.g. Anderson [1], Chow [10]). In other words the decision rule $d(\mathbf{x})$ is a function of $\mathbf{x}$ that tells us which decision to make for every possible pattern $\mathbf{x}$. So we have c possible decisions. When very ambiguous situations arise as to the decision to be made, it may be advantageous to allow the classification system to withhold or reject the decision and to mark the pattern for a later special processing. We designate the *reject option* by writing $d(\mathbf{x}) = d_0$. In such a case we have $c + 1$ decisions in a c class problem. Since the number of possible decisions is finite a decision rule d may be defined as an ordered partition $\{\Omega_0, \Omega_1, \ldots, \Omega_c\}$ of the feature space Ω. The rule d assigns a pattern $\mathbf{x}$ to class ω_j (i.e. $d(\mathbf{x}) = d_j$) if and only if the observed $\mathbf{x} \in \Omega_j$, $j = 1, \ldots, c$ (and the classification is correct if and only if the true class of the pattern $\mathbf{x}$ is ω_j) and rejects a pattern $\mathbf{x}$ (i.e. $d(\mathbf{x}) = d_0$) if and only if $\mathbf{x} \in \Omega_0$. The collection of all decision rules is denoted by $\mathscr{D}$.

The boundaries separating the sets Ω_i and Ω_j, $i, j = 0, \ldots, c$, $i \neq j$ are called the *decision boundaries*. In regular cases these boundaries may be determined by forming $c + 1$ functions $D_0(\mathbf{x}), \ldots, D_c(\mathbf{x})$, $D_i(\mathbf{x})$: $\Omega \to \mathbb{R}$ (i.e. $D_i(\mathbf{x})$ is a mapping from the feature space $\Omega \subset \mathbb{R}^m$ to the real numbers) called *decision* or *discriminant functions*, which are chosen in such way that for all $\mathbf{x} \in \Omega_i$ holds

$$D_i(\mathbf{x}) > D_j(\mathbf{x}), \quad j = 0, \ldots, c, \quad i \neq j .$$

Then the equation of the decision boundary separating the set Ω_i from the set Ω_j is

$$D_i(\mathbf{x}) - D_j(\mathbf{x}) = 0 .$$

Classification of a pattern $\mathbf{x}$ is performed as follows:

assign $\mathbf{x}$ to the class ω_i iff

$$D_i(\mathbf{x}) > D_j(\mathbf{x}), \quad j = 1, \ldots, c, \quad i \neq j ;$$

reject $\mathbf{x}$ iff

$$D_0(\mathbf{x}) > D_j(\mathbf{x}), \quad j = 1, \ldots, c .$$

Consider now the problem of classifying an arbitrary pattern $\mathbf{x}$ of unknown class. Assume that densities $p_i(\mathbf{x})$, $i = 1. 2, \ldots, c$ are known along with a priori probabilities P_i, $i = 1, 2, \ldots, c$.

First, associated with a decision $d(\mathbf{x}) = d_j$ is an *observation-conditional non-error rate*; the conditional probability of correct classification, given an observation $\mathbf{x} \in \Omega_j$, is the class a posteriori probability $P(\omega_j \mid \mathbf{x})$ which can be computed by the Bayes rule

$$P(\omega_j \mid \mathbf{x}) = P_i p_i(\mathbf{x}) / p(\mathbf{x}) . \tag{1}$$

The *observation-conditional error rate* or the conditional probability of classification error or misclassification is then

$$e_j(\mathbf{x}) = 1 - P(\omega_j \mid \mathbf{x}) . \tag{2}$$

Second, the *class-conditional non-error rate* given ω_i or the probability of correct classification of a pattern $\mathbf{x}$ from class ω_i is

$$G_{ii} \triangleq P\{\mathbf{x} \in \Omega_i \mid \text{class} = \omega_i\} = \int_{\Omega_i} p_i(\mathbf{x})\, d\mathbf{x}\,, \tag{3}$$

and

$$G_{ij} \triangleq P\{\mathbf{x} \in \Omega_i \mid \text{class} = \omega_j\} = \int_{\Omega_i} p_j(\mathbf{x})\, d\mathbf{x}\,, \quad j \neq i \tag{4}$$

is the *class-conditional error rate given* ω_j $(i \neq j)$ i.e. the probability of incorrect classification of a pattern $\mathbf{x}$ from class ω_j to class ω_i.

The *class-conditional reject rate given* ω_j i.e. the probability that the pattern $\mathbf{x}$ from class ω_j is rejected to be classified is

$$G_{0j} \triangleq P\{\mathbf{x} \in \Omega_0 \mid \text{class} = \omega_j\} = \int_{\Omega_0} p_j(\mathbf{x})\, d\mathbf{x}\,. \tag{5}$$

Third, the (unconditional) *non-error rate* or the overall probability of correct classification is

$$C(d) = \sum_{i=1}^{c} P_i G_{ii}\,, \tag{6}$$

the (unconditional) *error rate* or the overall probability of error or probability of misclassification is

$$E(d) = \sum_{i=1}^{c} \sum_{j=1}^{c} P_j G_{ij}\,, \quad j \neq i\,, \tag{7}$$

the *acceptance rate* or the overall acceptance probability is

$$A(d) = \sum_{i=1}^{c} \sum_{j=1}^{c} P_j G_{ij}\,,$$

and the *reject rate* or reject probability is

$$R_r(d) = \sum_{i=1}^{c} P_i G_{0i}\,. \tag{8}$$

The above defined probabilities are related by the following equations:

$$\begin{aligned} C + E &= A \\ A + R_r &= 1\,. \end{aligned} \tag{9}$$

The error rate and the reject rate are commonly used to describe the performance level of pattern classification system. An error occurs when a pattern from one class is identified as that of a different class. A reject occurs when the classification system withholds its classification and the pattern is to be handled in a special way.

Furthermore, let $L_{ij}(\mathbf{x})$ be a measure of the *loss* or penalty incurred when the decision $d(\mathbf{x}) = d_i$ is made and the true pattern class is in fact ω_j, $i = 0, \ldots, c$, $j = 1, \ldots, c$. The costs L_{ij} are bounded functions such that for $j = 1, \ldots, c$ and all $\mathbf{x} \in \Omega$,

$$0 \leqq L_{jj}(\mathbf{x}) \leqq L_{ij}(\mathbf{x}) \leqq 1\,, \quad \text{for} \quad i = 0, \ldots, c\,.$$

The classification of a pattern $\mathbf{x}$ from class ω_j to class ω_i entails a loss $L_{ij}(\mathbf{x})$ with probability $P(\omega_j \mid \mathbf{x})$. Consequently, the expected loss – called the *observation-conditional risk* (or conditional risk) – of making decision $d(\mathbf{x}) = d_i$ is

$$r_i(\mathbf{x}) = \sum_{j=1}^{c} L_{ij}(\mathbf{x})\, P(\omega_j \mid \mathbf{x}) . \tag{10}$$

The *average risk* associated with the decision d partitioning the feature space into regions $\Omega_0, \Omega_1, \ldots, \Omega_c$ is then

$$R(d) = \sum_{j=1}^{c} L_{0j} P_j G_{0j} + \sum_{i=1}^{c} \sum_{j=1}^{c} L_{ij} P_j G_{ij} . \tag{11}$$

Relation between conditional risk and average risk is presented in the sequel.

Several designs for a classification rule can be compared according to their error rate (7) or (if the losses are known) according to their average risk (11).

2.2 Bayes decision rule for minimum risk

The goal of the statistical decision theory is to propose the decision rule d in such a way that the average loss per decision is as small as possible. In other words, the most desirable is the rule that minimizes over the domain $\mathscr{D}$ of all rules the average risk (11). That is, regard R as a function of domain $\mathscr{D}$, and define

$$R^* = \inf_{d \in \mathscr{D}} R(d) \tag{12}$$

then d^* is the *minimum-risk decision rule* if $R(d^*) = R^*$. The minimum-risk rules can always be constructed by minimizing conditional risk (10) for every $\mathbf{x}$:

Such decision rule d^* defined as

$$d^*(\mathbf{x}) = d_i \quad \text{if} \quad r_i(\mathbf{x}) \leqq r_j(\mathbf{x}) , \quad i, j = 0, 1, \ldots, c \tag{13}$$

is called the *Bayes decision rule.*

The Bayes rule (13) partitions the feature space into c *acceptance regions* (called the Bayes acceptance regions):

$$\Omega_j^* = \{\mathbf{x} \in \Omega : r_j(\mathbf{x}) = \min_{i=0,\ldots,c} r_i(\mathbf{x})\} , \quad j = 1, \ldots, c \tag{14}$$

and one *reject region* Ω_0^* (called the Bayes reject region):

$$\Omega_0^* = \{\mathbf{x} \in \Omega : r_0(\mathbf{x}) = \min_{i=0,\ldots,c} r_i(\mathbf{x})\} . \tag{15}$$

The overall acceptance region Ω_A^* is the union of $\Omega_1^*, \Omega_2^*, \ldots, \Omega_c^*$ and it holds $\Omega_A^* \cup \Omega_0^* = \Omega$.

From (13) we can see that the Bayes rule has the minimum conditional risk

$$r^*(\mathbf{x}) = \min_{i=0,1,\ldots,c} r_i(\mathbf{x}) \tag{16}$$

and the minimum average risk – also called the *Bayes risk*

$$R^* = \int_\Omega r^*(\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x} =$$

$$= \sum_{j=1}^{c} L_{0j} P_j \int_{\Omega_0^*} p_j(\mathbf{x})\, d\mathbf{x} + \sum_{i=1}^{c} \sum_{j=1}^{c} L_{ij} P_j \int_{\Omega_i^*} p_j(\mathbf{x})\, d\mathbf{x} ,$$

where Ω_i^* and Ω_0^*, $i = 1, \ldots, c$ are defined by (14) and (15), respectively.

Suppose now the case of zero loss function for correct classification, a unit loss for the classification error and a constant loss λ_r for rejection, i.e.

$$L_{ii} = 0 , \quad L_{ij} = 1 , \quad L_{0j} = \lambda_r , \quad i, j = 1, \ldots, c , \quad i \neq j . \tag{18}$$

We call λ_r the *rejection threshold*. Then the conditional risk $r_i(\mathbf{x})$ of (10) becomes

$$r_0(\mathbf{x}) = \sum_{j=1}^{c} \lambda_r P(\omega_j \mid \mathbf{x}) = \lambda_r \tag{19}$$

$$r_i(\mathbf{x}) = \sum_{\substack{j=1 \\ j \neq i}}^{c} P(\omega_j \mid \mathbf{x}) = 1 - P(\omega_i \mid \mathbf{x}) , \quad i = 1, \ldots, c . \tag{20}$$

Therefore the conditional risk reduces to the conditional probability of classification error $e_i(\mathbf{x})$ defined in (2). With respect to

$$e^*(\mathbf{x}) = \min_{i=1,\ldots,c} e_i(\mathbf{x}) = 1 - \max_{i=1,\ldots,c} P(\omega_i \mid \mathbf{x}) \tag{21}$$

$$d^*(\mathbf{x}) = d_i \quad \text{if} \quad e^*(\mathbf{x}) = e_i(\mathbf{x}) \leqq \lambda_r \tag{22}$$

$$= d_0 \quad \text{if} \quad \lambda_r < e^*(\mathbf{x}) ,$$

or in terms of a posteriori probabilities:

$$d^*(\mathbf{x}) = d_i \quad P(\omega_i \mid \mathbf{x}) = \max_{j=1,\ldots,c} P(\omega_j \mid \mathbf{x}) \geqq 1 - \lambda_r \tag{23a}$$

$$= d_0 \quad \text{if} \quad 1 - \lambda_r > \max_{j=1,\ldots,c} P(\omega_j \mid \mathbf{x}) . \tag{23b}$$

Clearly, for c classes, $0 \leqq 1 - \max_j P(\omega_j \mid \mathbf{x}) \leqq (c-1)/c$ with the equality in the case when all the classes have equal a posteriori probabilities. So, for the rejection option to be activated we must have $0 \leqq \lambda_r \leqq (c-1)/c$.

The decision rule (23) partitions the feature space into c acceptance regions $\Omega_i(\lambda_r)$, $i = 1, \ldots, c$:

$$\Omega_i(\lambda_r) = \{\mathbf{x} \in \Omega : P(\omega_i \mid \mathbf{x}) = \max_{j=1,\ldots,c} P(\omega_j \mid \mathbf{x}) \geqq 1 - \lambda_r\} ,$$

and one reject region $\Omega_0(\lambda_r)$:

$$\Omega_0(\lambda_r) = \{\mathbf{x} \in \Omega : 1 - \lambda_r > \max_{j=1,\ldots,c} P(\omega_j \mid \mathbf{x})\} .$$

The reject rate R_r^* is given by

$$R_r^* = \int_{\Omega_0(\lambda_r)} p(\mathbf{x})\, d\mathbf{x} , \tag{24}$$

and the error rate is

$$E^* = \sum_{i=1}^{c} \int_{\Omega_i(\lambda_r)} e^*(\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x} . \tag{25}$$

The probabilities R^* and E^* are functions of λ_r only, so we may write $R_r^* = R_r^*(\lambda_r)$, $E^* = E^*(\lambda_r)$. The Bayes risk (17) expressed in terms of these probabilities is

$$\begin{aligned} R^* = R^*(\lambda_r) &= \int_{\Omega_A(\lambda_r)} e^*(\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x} + \int_{\Omega_0(\lambda_r)} \lambda_r\, p(\mathbf{x})\, d\mathbf{x} \\ &= E^*(\lambda_r) + \lambda_r R_r^*(\lambda_r) , \end{aligned} \tag{26}$$

where

$$\Omega_A(\lambda_r) \triangleq \bigcup_{i=1}^{c} \Omega_i(\lambda_r) .$$

The optimal property of the Bayes rule may be formulated, using (26), as follows: for a given pattern classification problem (in other words, for a given underlying distribution) among all the rules with error rate equal to $E^*(\lambda_r)$ there exists no rule with reject rate less than $R^*(\lambda_r)$. Equivalently, to $R^*(\lambda_r)$, there exists no rule with error rate less than $E^*(\lambda_r)$. In other words, the Bayes decision rule is optimal in the sense that for given error rate it minimizes the reject rate.

It follows from equation (21)–(23) that the reject option is activated whenever the conditional error probability exceeds the rejection threshold λ_r. The option to reject is introduced to safeguard against excessive misclassification. However the tradeoff between the errors and rejects is seldom one to one. Whenever the reject option is exercised, some would be correct classifications are also converted into rejects. Analysis of the error-reject tradeoff is an important result of Chow [11].

The rejection threshold is an upper bound of both the error rate $E^*(\lambda_r)$ and the Bayes risk $R^*(\lambda_r)$ (see, e.g. Devijver and Kittler [14]).

2.3 Bayes decision rule for minimum error rate

Suppose that $L_{ii} = 0$, $L_{ij} = 1$ and $L_{0j} = \lambda_r$, $i, j = 1, 2, \ldots, c$, $i \neq j$ and assume that the loss λ_r exceeds the largest possible conditional probability of classification error, $(c - 1)/c$. Then it follows from (23b) that the reject option is never activated and the Bayes rule constantly makes decision according to the minimal conditional probability of error. We therefore have

$$d^*(\mathbf{x}) = d_i \quad \text{if} \quad P(\omega_i \mid \mathbf{x}) = \max_{j=1,\ldots,c} P(\omega_j \mid \mathbf{x}) . \tag{27}$$

Decision rule (27) partitions the feature space Ω into c acceptance regions Ω_i

$$\Omega_i = \{\mathbf{x} \in \Omega : P_i p_i(\mathbf{x}) > P_j p_j(\mathbf{x})\} , \quad i, j = 1, \ldots, c , \quad i \neq j . \tag{28}$$

The conditional Bayes risk is

$$e^*(\mathbf{x}) = 1 - \max_{j=1,\ldots,c} P(\omega_j \mid \mathbf{x}) \tag{29}$$

and the average Bayes risk is

$$E^* = \int_\Omega e^*(\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x} = \sum_{i=1}^{c} \int_{\Omega_i} P_j\, p_j(\mathbf{x})\, d\mathbf{x}\,, \quad j \neq i\,. \tag{30}$$

It is noted from (29) that the corresponding discriminant function implemented by a Bayes decision rule is essentially

$$D_i(\mathbf{x}) = P_i\, p_i(\mathbf{x})\,, \quad i = 1, \ldots, c \tag{31}$$

or equivalently

$$D_i(\mathbf{x}) = \log\left[P_i\, p_i(\mathbf{x})\right],$$

because for $\mathbf{x} \in \omega_i$, $D_i(\mathbf{x}) > D_j(\mathbf{x})$ for all $i, j = 1, \ldots, c$, $i \neq j$ (cf. Nilson [40]).

The equation

$$D_i(\mathbf{x}) - D_j(\mathbf{x}) = 0$$

or

$$\log \frac{D_i(\mathbf{x})}{D_j(\mathbf{x})} = 0$$

defines the *decision boundary* separating the acceptance region Ω_i from the acceptance region Ω_j, $i, j = 1, \ldots. c$, $i \neq j$.

2.4 Two-class classification

The general two class classification problem — also called the *dichotomy* is to test a pattern $\mathbf{x}$ with the decision function $D(\mathbf{x})$: $\Omega \to \mathbb{R}$. The value of this decision function is used to select class ω_i which the pattern $\mathbf{x}$ must likely belongs to or to reject its classification:

$$\begin{aligned} D(\mathbf{x}) &\leqq \alpha_1 && \Rightarrow \mathrm{d}(\mathbf{x}) = d_1 \\ D(\mathbf{x}) &\geqq \alpha_2 && \Rightarrow d(\mathbf{x}) = d_2 \\ \alpha_1 < D(\mathbf{x}) &< \alpha_2 && \Rightarrow d(\mathbf{x}) = d_0\,. \end{aligned}$$

The thresholds α_1, α_2 are chosen to minimize a classification criterion of optimality. Each criterion determines a decision function.

2.4.1 Bayes decision rule

It can be readily verified, that the Bayes rule (13) for two-class problem becomes

$$\begin{aligned} d^*(\mathbf{x}) = d_1 \quad &\text{if} \quad \frac{P(\omega_1 \mid \mathbf{x})}{P(\omega_2 \mid \mathbf{x})} \geqq C_1 \\ = d_2 \quad &\text{if} \quad \frac{P(\omega_1 \mid \mathbf{x})}{P(\omega_2 \mid \mathbf{x})} \leqq C_2 \\ = d_0 \quad &\text{if} \quad C_2 < \frac{P(\omega_1 \mid \mathbf{x})}{P(\omega_2 \mid \mathbf{x})} < C_1\,, \end{aligned} \tag{32}$$

where

$$C_1 \triangleq \frac{L_{12} - L_{02}}{L_{01} - L_{11}} \quad \text{and} \quad C_2 \triangleq \frac{L_{02} - L_{22}}{L_{21} - L_{01}} \tag{33}$$

if $L_{jj} < L_{ij}$ for $i = 0, 1, 2, j = 1, 2$.

Alternatively, the decision rule (32) can be expressed in terms of *likelihood ratio* $p_1(\mathbf{x})/p_2(\mathbf{x})$:

$$\begin{aligned} d^*(\mathbf{x}) &= d_1 \quad \text{if} \quad \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} \geqq \frac{P_2}{P_1} C_1 \\ &= d_2 \quad \text{if} \quad \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} \leqq \frac{P_2}{P_1} C_2 \\ &= d_0 \quad \text{otherwise} . \end{aligned} \tag{34}$$

It is easy to verify that boundaries in (34) are related to the class-conditional error rate given in (4) by the following expressions:

$$\frac{P_2}{P_1} C_1 \leqq \frac{1 - G_{21}}{G_{12}} , \quad \frac{P_2}{P_1} C_2 \geqq \frac{G_{21}}{1 - G_{12}} .$$

It follows from (34) that in the case of Bayes decision rule (34) the decision function is

$$D(\mathbf{x}) = \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} .$$

The equation $D(\mathbf{x}) = C_1 P_2/P_1$ defines the decision boundary separating the acceptance region $\Omega_1 = \{\mathbf{x} \in \Omega: p_1(\mathbf{x})/p_2(\mathbf{x}) \geqq C_1 P_2/P_1\}$ from the reject region $\Omega_0 = \{\mathbf{x} \in \Omega: C_2 P_2/P_1 < p_1(\mathbf{x})/p_2(\mathbf{x}) < C_1 P_2/P_1\}$. The decision boundary separating the acceptance region Ω_2 from Ω_0 is defined in a similar way.

For the two class case the condition for rejection, namely (23b), can never be satisfied when $\lambda_r > \frac{1}{2}$; hence the reject rate is always zero if the rejection threshold exceeds $\frac{1}{2}$. For two classes and $0 \leqq \lambda_r \leqq \frac{1}{2}$ the condition for rejection (23b) is equivalent to

$$\frac{\lambda_r}{1 - \lambda_r} < \frac{P_1 \, p_1(\mathbf{x})}{P_2 \, p_2(\mathbf{x})} < \frac{1 - \lambda_r}{\lambda_r} , \tag{35}$$

and therefore the decision rule (23) becomes in terms of likelihood ratio

$$\begin{aligned} d^*(\mathbf{x}) &= d_1 \quad \text{if} \quad \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} \geqq \frac{P_2}{P_1} \frac{1 - \lambda_r}{\lambda_r} \\ &= d_2 \quad \text{if} \quad \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} \leqq \frac{P_2}{P_1} \frac{\lambda_r}{1 - \lambda_r} \\ &= d_0 \quad \text{otherwise} . \end{aligned} \tag{36}$$

2.4.2 Decision rule with predetermined probability of error

In many practical situations it is advantageous and useful to predetermine the value e_i for the conditional probability of misclassification associated with the decision $d(\mathbf{x}) = d_i$, $i = 1, 2$. This restriction can be determined by the actual possibilities, e.g. of economical or technical type, which are at our disposal to settle the consequences of misclassifications.

We accept the pattern $\mathbf{x}$ for classification and identify it as from the ith class whenever its maximum of the a posteriori probabilities is equal or greater then $1 - e_i$. Likewise we reject the pattern $\mathbf{x}$ whenewer its maximum of the a posteriori probabilities is less than $1 - e_i$. It means that the *decision rule with predetermined conditional probabilities of errors* is:

$$\begin{aligned} d(\mathbf{x}) = d_i \quad &\text{if} \quad P(\omega_i \mid \mathbf{x}) = \max_{j=1,2} P(\omega_j \mid \mathbf{x}) \geqq 1 - e_i \\ = d_0 \quad &\text{if} \quad P(\omega_i \mid \mathbf{x}) = \max_{j=1,2} P(\omega_j \mid \mathbf{x}) < 1 - e_i \,. \end{aligned} \tag{37}$$

From (37) it follows that $P(\omega_j \mid \mathbf{x}) \leqq e_i$ must hold for acceptation the decision $d(\mathbf{x}) = d_i$, $i, j = 1, 2, j \neq i$. Therefore

$$d(\mathbf{x}) = d_1 \quad \text{if} \quad \frac{P(\omega_1 \mid \mathbf{x})}{P(\omega_2 \mid \mathbf{x})} \geqq \frac{1 - e_1}{e_1} \tag{38a}$$

$$= d_2 \quad \text{if} \quad \frac{P(\omega_1 \mid \mathbf{x})}{P(\omega_2 \mid \mathbf{x})} \leqq \frac{e_2}{1 - e_2} \tag{38b}$$

$$= d_0 \quad \text{if} \quad \frac{e_2}{1 - e_2} < \frac{P(\omega_1 \mid \mathbf{x})}{P(\omega_2 \mid \mathbf{x})} < \frac{1 - e_1}{e_1} \tag{38c}$$

and the reject option will be activated if the following inequalities hold

$$\frac{e_2}{1 - e_2} < 1 < \frac{1 - e_2}{e_1} \,.$$

Thus $0 \leqq e_i < \frac{1}{2}$, $i = 1, 2$, and $e_1 + e_2 < 1$ holds. The acceptance regions $\Omega_1(e_1)$ and $\Omega_2(e_2)$ depend now on the values e_1 and e_2 respectively:

$$\begin{aligned} \Omega_1(e_1) &= \{\mathbf{x} \in \Omega : \max_{j=1,2} P(\omega_j \mid \mathbf{x}) = P(\omega_1 \mid \mathbf{x}) \geqq 1 - e_1\} \\ \Omega_2(e_2) &= \{\mathbf{x} \in \Omega : \max_{j=1,2} P(\omega_j \mid \mathbf{x}) = P(\omega_2 \mid \mathbf{x}) \geqq 1 - e_2\} \end{aligned} \tag{39}$$

and the reject region depends on both values e_1 and e_2:

$$\Omega_0(e_1, e_2) = \left\{\mathbf{x} \in \Omega : \frac{e_2}{1 - e_2} < \frac{P(\omega_1 \mid \mathbf{x})}{P(\omega_2 \mid \mathbf{x})} < \frac{1 - e_1}{e_1}\right\} . \tag{40}$$

The error rate and the reject rate

$$E(e_1, e_2) = 1 - \sum_{i=1}^{2} \int_{\Omega_i(e_i)} p_i(\mathbf{x})\, d\mathbf{x} \tag{41}$$

$$R_r(e_1, e_2) = \int_{\Omega_0(e_1,e_2)} p(\mathbf{x})\, d\mathbf{x} \tag{42}$$

respectively, are now functions of both e_1 and e_2.

If e_i are related to the loss functions by

$$e_i = \frac{L_{0i} - L_{ii}}{L_{0i} - L_{ii} + L_{ij} - L_{0j}}, \quad i, j = 1, 2, \quad j \neq i$$

then

$$\frac{L_{12} - L_{22}}{L_{01} - L_{11}} = \frac{1 - e_1}{e_1} \quad \text{and} \quad \frac{L_{02} - L_{22}}{L_{21} - L_{01}} = \frac{e_2}{1 - e_2}.$$

Therefore from (38) and (32) it follows that the decision rule (38) with predetermined conditional probabilities of error is also minimum risk rule.

If $e_1 = e_2 = \lambda_r$, then the rule (38) coincides with the Bayes rule (36).

2.4.3 Bayes Gaussian rule

As an illustrative example, suppose that $p_i(\mathbf{x})$, $i = 1, 2$ is a multivariate Gaussian (normal) density function with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$; that is,

$$p_i(\mathbf{x}) = (2\pi)^{-m/2} |\boldsymbol{\Sigma}_i|^{-1/2} \exp\left[-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)\right], \quad i = 1, 2.$$

Then the Bayes rule (36) can be reformulated in the following way:

$$\begin{aligned} d^*(\mathbf{x}) &= d_1 \quad \text{if} \quad D(\mathbf{x}) \geqq T_1 \\ &= d_2 \quad \text{if} \quad D(\mathbf{x}) \leqq T_2 \\ &= d_0 \quad \text{if} \quad T_2 < D(\mathbf{x}) < T_1, \end{aligned}$$

where

$$D(\mathbf{x}) = \log \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} = -\tfrac{1}{2}\left[(\mathbf{x} - \boldsymbol{\mu}_1)^{\mathrm{T}} \boldsymbol{\Sigma}_1^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) - (\mathbf{x} - \boldsymbol{\mu}_2)^{\mathrm{T}} \boldsymbol{\Sigma}_2^{-1} (\mathbf{x} - \boldsymbol{\mu}_2)\right] \tag{43}$$

$$T_1 = \log \frac{P_2}{P_1} \frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} \frac{1 - \lambda_r}{\lambda_r} \quad \text{and} \quad T_2 = \log \frac{P_2}{P_1} \frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} \frac{\lambda_r}{1 - \lambda_r}. \tag{44}$$

The equation of the decision boundary separating the acceptance region $\Omega_1 = \{\mathbf{x} \in \Omega: D(\mathbf{x}) \geqq T_1\}$ from the reject region $\Omega_0 = \{\mathbf{x} \in \Omega: T_2 < D(\mathbf{x}) < T_1\}$ is

$$D(\mathbf{x}) = T_1. \tag{45}$$

It is clear that the equation of the decision boundary separating the acceptance region $\Omega_2 = \{\mathbf{x} \in \Omega: D(\mathbf{x}) \leqq T_2\}$ from region Ω_0 is

$$D(\mathbf{x}) = T_2. \tag{46}$$

The equations (45) and (46) are quadratic in $\mathbf{x}$.

If $\Sigma_1 = \Sigma_2 = \Sigma$, the equation (45) reduces to

$$\mathbf{x}^T \Sigma^{-1}(\mu_1 - \mu_2) = \tfrac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) + \log \frac{P_2}{P_1} \frac{1 - \lambda_r}{\lambda_r}, \qquad (47)$$

and the equation (46) reduces in a similar way too.

The decision boundary $D(\mathbf{x}) = T_i$, $i = 1, 2$, is now *linear* i.e. a *hyperplane* $\mathscr{H}_i$ in m-dimensional feature space. The hyperplanes $\mathscr{H}_1$ and $\mathscr{H}_2$ are parallel.

Let

$$\mathbf{w} \triangleq \Sigma^{-1}(\mu_1 - \mu_2) \quad \text{and} \quad w_i \triangleq \left[-\tfrac{1}{2}(\mu_1 + \mu_2)^T \mathbf{w} + T_i\right] \qquad (48)$$

for $i = 1, 2$ then the equation of hyperplane $\mathscr{H}_i$ is

$$\mathbf{x}^T\mathbf{w} + w_i = 0\,, \quad i = 1, 2\,.$$

In the important case when there are no rejects, the equation of the discriminant hyperplane $\mathscr{H}$ is obtained by taking $\lambda_r = \frac{1}{2}$ in (47). The hyperplane $\mathscr{H}$ is sometimes referred to as the *Anderson discriminant hyperplane* [1], specifically

$$\mathbf{x}^T \Sigma^{-1}(\mu_1 - \mu_2) = \tfrac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) + \log \frac{P_2}{P_1}\,. \qquad (49)$$

If a priori probabilities are equal ($P_1 = P_2$), the optimal regions Ω_1 and Ω_2 are complementary half-spaces. In this case the error rate of Bayes rule is $(1 - \phi(\frac{1}{2}\Delta))$, where ϕ denotes the standard univariate normal distribution function and Δ is the positive root of $\Delta^2 = (\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2)$, which is the so called *Mahalanobis distance measure* [1].

2.5 Linear discriminant functions

The classification rules discussed in Sections 2.2 and 2.3 relied heavily on the assumed knowledge of the underlying statistical distributions. In some simple cases we were able to obtain the explicit form of the Bayes rule as it was shown in Section 2.4.

A radically different approach is based on the assumption that the functional form of the decision boundary is selected a priori. In other words, the analytical expression of the decision boundary is given except for the values of the set of parameters.

For example, for a given functional form $D(\mathbf{x})$: $\Omega \to \mathbb{R}$ and the problem of c classes, one possibility is to have $c + 1$ discriminant function $D_i(\mathbf{x})$: $\Omega \to \mathbb{R}$; $i = 0, 1, \ldots, c$. Classification of a pattern $\mathbf{x}$ is performed by forming these c discriminant functions and assigning $\mathbf{x}$ to the class associated with the largest discriminant function.

The selection of a suitable functional form for a discriminant function is a problem of crucial importance.

Let us consider the family of linear discriminant functions of the form ([1], [34])

$$D(\mathbf{x}) = \sum_{i=1}^{m} w_i x_i + w_0 = \mathbf{x}^T\mathbf{w} + w_0 \,. \tag{50}$$

The m-dimensional vector $\mathbf{w} = (w_1, \ldots, w_m)^T$ is called the *weight vector* and w_0 is the *threshold weight*.

We introduce the necessary notation. Let $\boldsymbol{\mu}_i$ denote the class-conditional mean vector of $\mathbf{x}$ and

$$\boldsymbol{\mu} = \sum_{i=1}^{c} P_i \boldsymbol{\mu}_i \quad \text{with} \quad \boldsymbol{\mu}_i = \mathsf{E}_i\{\mathbf{x}\} \tag{51}$$

denote the mean vector of the mixture distribution. For c classes, let $\mathbf{S}_B$ designate the *between class covariance* (or *scatter*) *matrix*

$$\mathbf{S}_B = \sum_{i=1}^{c} P_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T = \sum_{i \leqq j}^{c} P_i P_j(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \,. \tag{52}$$

For the problem of c classes, $\mathbf{S}_B$ has rank at most equal to $(c - 1)$. So, in the two-class case, $\mathbf{S}_B$ has rank one, unless the two classes have identical mean vectors. Let $\boldsymbol{\Sigma}_i$ be the conditional covariance matrix

$$\boldsymbol{\Sigma}_i = \mathsf{E}_i\{(\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^T\} \,, \quad i = 1, \ldots, c \tag{53}$$

and $\mathbf{S}_w$ be the average class-conditional covariance matrix, frequently called the *within class scatter matrix*,

$$\mathbf{S}_w = \sum_{i=1}^{c} P_i \boldsymbol{\Sigma}_i \,. \tag{54}$$

If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}$, then $\mathbf{S}_w = \boldsymbol{\Sigma}$.

In what follows we will restrict ourselves to the two-class classification problem.

2.5.1 Linear classification without rejection

The decision rule corresponding to the discriminant function $D(\mathbf{x})$ of (50) is:

$$\begin{aligned} &\text{assign } \mathbf{x} \text{ to class } \omega_1 \text{ iff } D(\mathbf{x}) > 0 \\ &\text{assign } \mathbf{x} \text{ to class } \omega_2 \text{ iff } D(\mathbf{x}) < 0 \,. \end{aligned} \tag{55}$$

If $D(\mathbf{x}) = 0$, $\mathbf{x}$ can be assigned to either of the classes. The decision boundary is defined by the equation $D(\mathbf{x}) = 0$. As $D(\mathbf{x})$ is linear in $\mathbf{x}$ the decision boundary is a hyperplane in m-dimensional space.

The quality of classification according to (55) depends on the vector $\mathbf{w}$ and the weight w_0. In this section we shall present the least-mean-squared-error (cf. e.g. Wee [49]) and Fisher criteria (cf. Fisher [20]) for the determination of $\mathbf{w}$ and w_0.

(i) The *least-mean-squared-error* (MSE) criterion of optimality of a linear

discriminant function $D(\mathbf{x})$ is taken to be

$$J(\mathbf{w}, w_0) = \sum_{i=1}^{2} P_i \, \mathsf{E}_i\{(D(\mathbf{x}) - \delta_i)^2\}$$

where δ_i is the desired value of $D(\mathbf{x})$ if ω_i is true, $i = 1,2$. $J(\mathbf{w}, w_0)$ is a measure of the mean-square-distance between $D(\mathbf{x})$ and δ_i, $i = 1, 2$. The closer $J(\mathbf{w}, w_0)$ is to zero, the better $D(\mathbf{x})$ is in the MSE sense. With the following arbitrary choice

$$\mathsf{E}\{D(\mathbf{x}) \mid \mathbf{x} \in \omega_1\} = 1 \quad \text{and} \quad \mathsf{E}\{D(\mathbf{x}) \mid \mathbf{x} \in \omega_2\} = -1$$

and with S_W non-singular, the following values can be determined (see e.g. [14])

$$\mathbf{w}^* = \frac{2P_1P_2 \, \mathbf{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{1 + \operatorname{tr}(\mathbf{S}_W^{-1}\mathbf{S}_B)}$$

$$w_0^* = (P_1 - P_2) - \frac{2P_1P_2}{1 + P_1P_2 \operatorname{tr}(\mathbf{S}_W^{-1}\mathbf{S}_B)} \, \boldsymbol{\mu}^{\mathrm{T}} \mathbf{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) .$$

The minimum of $J(\mathbf{w}, w_0)$ is then

$$J(\mathbf{w}^*, w_0^*) = \frac{4P_1P_2}{1 + \operatorname{tr}(\mathbf{S}_W^{-1}\mathbf{S}_B)} .$$

The equation of the decision boundary $\mathbf{x}^{\mathrm{T}}\mathbf{w}^* + w_0^* = 0$ in terms of the parameters of the underlying distributions is after the analytical modifications

$$\mathbf{x}^{\mathrm{T}}\mathbf{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \boldsymbol{\mu}^{\mathrm{T}}\mathbf{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{P_1 - P_2}{2P_1P_2}\left[1 + P_1P_2 \operatorname{tr}(\mathbf{S}_W^{-1}\mathbf{S}_B)\right] . \tag{56}$$

Remark. If the classes have equal apriori probabilities ($P_1 = P_2$) and the class-conditional probability density functions have identical covariance matrices ($\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$), then the equation of the linear decision boundary in (56) coincide with the equation of the Bayes optimal Anderson discriminant plane (49), because (56) and (49) both reduce to

$$(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 0 . \tag{57}$$

If we further assume that the classes are normally distributed, then MSE criterion yields the Bayes-optimal decision boundary. Consequently, the error rate can be calculated as in Section 2.3.4. In more general cases at the best we will be able to derive an upper-bound on the error rate E_{MSE} associated with the discriminant function in (56):

$$E_{\mathrm{MSE}} \leqq \frac{J(\mathbf{w}^*, w_0^*)}{1 - J(\mathbf{w}^*, w_0^*)} = \frac{4P_1P_2}{(1 - 4P_1P_2) + \operatorname{tr}(\mathbf{S}_W^{-1}\mathbf{S}_B)} . \tag{58}$$

Therefore a small mean-squared error $J(\mathbf{w}^*, w_0^*)$ or a large $\operatorname{tr}(\mathbf{S}_W^{-1}\mathbf{S}_B)$ is synonymous with a small expected error rate (see e.g. [14]).

(ii) In classical discriminant analysis which finds its origin in the classical paper by R. A. Fisher [19], the optimal vector $\mathbf{w}$ is the one for which the *Fisher discriminant ratio* ([20], [14])

$$F(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}, \tag{59}$$

with $\mathbf{S}_B$ and non-singular $\mathbf{S}_W$ as in (52) and (54), respectively, reaches maximum. The maximum of the right-hand side of (59) with respect to $\mathbf{w}$ occurs, when $\mathbf{w}$ is given by

$$\mathbf{w}^* = \mathbf{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{60}$$

Then

$$F(\mathbf{w}^*) = \max_{\mathbf{w} \neq 0} F(\mathbf{w}) = P_1 P_2 (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{S}_W^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \operatorname{tr}(\mathbf{S}_W^{-1} \mathbf{S}_B). \tag{61}$$

Remark. The vector $\mathbf{w}^*$ of (60) corresponds identically to the weight vector of the MSE linear discriminant function in (56).

The maximalization of the Fisher discriminant ratio does not involve the threshold weight w_0. It is shown in [14] that there exists a threshold weight w_0^* such that $D(\mathbf{x}) = \mathbf{x}^T \mathbf{w}^* + w_0^*$ is Bayes optimal.

The determination of the linear discriminant function by using MSE and Fisher criterion does not involve very constraining assumptions regarding the underlying distributions. It is implicitly assumed that these distributions have first and second order moments. For $\mathbf{w}^*$ to be different from null vector we should have $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. If $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ then no decision regarding classification can be made because always $D(\mathbf{x}) = 0$. If the two-class distributions exhibit both separability of the means and the covariances and we wish to employ a (suboptimal) linear classifier then a minimum MSE approach is applied.

2.5.2 Linear classification with reject option

The allocation of a pattern to one class from two classes with possibility of reject option can be done by using linear discriminant function as follows:

$$\begin{aligned} &\text{assign the pattern } \mathbf{x} \text{ to class } \omega_1 \ (\omega_2) \\ &\text{if } \ \mathbf{x}^T \mathbf{w} \leqq -w_1 \ \ (\mathbf{x}^T \mathbf{w} \geqq -w_2); \\ &\text{reject the pattern } \mathbf{x} \text{ to classify} \\ &\text{if } \ -w_1 < \mathbf{x}^T \mathbf{w} < -w_2 . \end{aligned} \tag{62}$$

The problem is a proper choice of a vector $\mathbf{w}$ and two threshold weights w_1 and w_2. In the case of two multivariate normal distributions with mean vectors $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, equal covariance matrices $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ and $(0, 1, \lambda_r)$ loss function the choice for $\mathbf{w}$ and w_i, $i = 1, 2$ as in (48) leads to the Bayes optimal rule.

2.6 Linear transformation and Bayes classification

As mentioned above the most meaningful performance criteria for a classification are the average risk (11) (if the losses are known) or the error and reject rates (7) and (8), respectively. We can construct the Bayes optimal decision rule (13), however if the dimension of feature space is greater than one, then the Bayes risk or error rate are difficult to compute without additional class structure assumptions (e.g. normal distributions with equal covariance matrices).

In this connection the problem of the classification of any pattern $\mathbf{x}$ may be transformed from m-dimensional space $\mathbb{R}^m$ to one-dimensional space $\mathbb{R}$. All pattern vectors are projected on the line with the direction $\mathbf{w}$, i.e. the number

$$y = \mathbf{x}^T\mathbf{w} \tag{63}$$

is related to each vector $\mathbf{x}$. The transformed space we denote by $\mathcal{Y} \subset \mathbb{R}$. The transformed observations y for class ω_i are then distributed with the density $p_i(y, \mathbf{w})$, $i = 1, \ldots, c$; the common transformed density is $p(y, \mathbf{w})$. The Bayes risk (17) for the transformed data (63) may then be denoted by

$$R^*(\mathbf{w}) = \sum_{j=1}^{c} L_{0j} P_j \int_{\Omega_0(\mathbf{w})} P_j(y, \mathbf{w})\, \mathrm{d}y + \sum_{j=1}^{c} \sum_{i=1}^{c} L_{ij} P_j \int_{\Omega_i(\mathbf{w})} P_j(y, \mathbf{w})\, \mathrm{d}y\,, \tag{64}$$

where the transformed Bayes decision regions are given by

$$\Omega_0(\mathbf{w}) = \{y \in \mathcal{Y}\colon r_0(y) = \min_{i=0,\ldots,c} r_i(y)\} \tag{65}$$

and

$$\Omega_i(\mathbf{w}) = \{y \in \mathcal{Y}\colon r_i(y) = \min_{j=0,\ldots,c} r_j(y)\}\,, \quad i = 1, \ldots, c \tag{66}$$

with the conditional risk for transformed data

$$r_i(y) = \sum_{j=1}^{c} L_{ij} P_j p_j(y, \mathbf{w}) / p(y, \mathbf{w})\,, \quad i = 0, \ldots, c\,. \tag{67}$$

The Bayes optimal dicision rule is then defined as follows:

$$\begin{aligned} &\text{Assign a pattern } \mathbf{x} \text{ to class } \omega_i \text{ (i.e. } d(\mathbf{x}) = d_i) \\ &\text{if } \quad y = \mathbf{x}^T\mathbf{w} \in \Omega_i(\mathbf{w}), \quad i = 1, \ldots, c\,; \\ &\text{reject a pattern } \mathbf{x} \text{ (i.e. } d(\mathbf{x}) = d_0) \\ &\text{if } \quad y = \mathbf{x}^T\mathbf{w} \in \Omega_0(\mathbf{w})\,, \end{aligned} \tag{68}$$

where $\Omega_0(\mathbf{w})$ and $\Omega_i(\mathbf{w})$, $i = 1, \ldots, c$ are as in (65) and (66), respectively.

It is clear that $R^* \leqq R^*(\mathbf{w})$.

In two-class case for $(0, 1, \lambda_r)$ loss function, the decision rule (68) becomes

$$\begin{aligned} d(\mathbf{x}) = d_1 \quad &\text{if} \quad \frac{P_1 p_1(y, \mathbf{w})}{P_2 p_2(y, \mathbf{w})} \geqq \frac{1 - \lambda_r}{\lambda_r} \\ = d_2 \quad &\text{if} \quad \frac{P_1 p_1(y, \mathbf{w})}{P_2 p_2(y, \mathbf{w})} \leqq \frac{\lambda_r}{1 - \lambda_r} \\ = d_0 \quad &\text{otherwise}\,. \end{aligned} \tag{69}$$

Theoretical results related to finding vector **w** which minimizes $R(\mathbf{w})$ for two multivariate normal classes with equal a priori class-probabilities and (0, 1) loss function were initially presented by Guseman and Walker [27]. The associated computation procedure was presented by Guseman and Walker [28]. Results for the general case of c m-dimensional normal classes with arbitrary a priori class-probabilities and (0, 1) loss function appeared in Guseman, Peters, and Walker [29].

The decision procedure used in PREDITAS system is based on the decision rule (69), where the transforming vector **w** maximizes the Fisher discriminant ratio (59).

3. FEATURE SELECTION AND EXTRACTION PROBLEM

This chapter is devoted to the theoretical background of feature selection and extraction methods, as well as to the comparison of feature selection and extraction from the practical viewpoint. Finally, basic properties of so called search algorithms are discussed from the general point of view and some definitions concerning the effectivity of individual features in relation to the selected subsets are presented.

3.1 Problem formulation and basic approaches to its solution

One of the fundamental problems in statistical pattern recognition is to determine which features should be employed for the best classification result. The need to retain only a small number of "good" features arise due to computational reasons, cost considerations.

Two general strategies have been used for dimensionality reduction:

(i) *Feature selection*: Chose the "best" subset of size m from the given set of M features.

(ii) *Feature extraction*: Apply a transformation to the pattern vector and select the m "best" features in the transformed space. The each new feature is a combination of all the M features.

The pioneering work in the area of feature selection and extraction is associated with the names of Sebastyen [46], Lewis [36] and Marill and Green [38].

The goal of feature selection and extraction is to determine the features which are important for the discrimination between classes.

The main objective of feature selection methods is to accomplish dimensionality reduction by reducing the number of required measurements. This can be achieved by eliminating those measurements which are redundant or do not contain enough relevant information. Thus the problem of *feature selection* in the measurement space lies in selecting the best subset X of m features.

$$X = \{x_i : i = 1, 2, \ldots, m, x_i \in Y\}$$

from the set Y,

$$Y = \{y_j : j = 1, 2, \ldots, M\}$$

of M possible features representing the pattern, clearly $m < M$. By the best subset we understand the combination of m features which optimizes a criterion function $J(\cdot)$ over all possible subsets F of m features. Let sets X and F of features be represented by vectors $\mathbf{x}$ of $\mathbf{f}$ respectively, whose components are the elements of X and F. Then the problem of feature selection becomes one of finding vector $\mathbf{x}$ satisfying

$$J(\mathbf{x}) = \max_{\mathbf{f}} J(\mathbf{f}) .$$

In contrast, feature extraction is considered as a process of mapping original features into more effective features. Therefore feature extraction consists of extracting m features, each of which is a combination of all M initial features, $m \leqq M$. The feature extraction can be viewed as the transformation B

$$B = (b_1, \ldots, b_m) , \quad 1 \leqq m \leqq M$$

which maps pattern vector $\mathbf{y}$ from the M-dimensional feature space into m-dimensional space, i.e.

$$\mathbf{x} = B(\mathbf{y}) \quad \text{with} \quad x_i = b_i(y_1, \ldots, y_M) , \quad 1 \leqq i \leqq m$$

(in general $B(\mathbf{y})$ is any vector function of $\mathbf{y}$). The mapping $B(\mathbf{y})$ could be obtained by optimizing a criterion function $J(\cdot)$. More specifically, $B(\mathbf{y})$ is determined among all admissible transformations $\{\tilde{B}(\mathbf{y})\}$ as one that satisfies

$$J\{B(\mathbf{y})\} = \max_{\tilde{B}} J\{\tilde{B}(\mathbf{y})\} .$$

These mapping techniques can be either linear or nonlinear (cf. Biswas et al. [2]).

Remark. The class of linear mappings includes the feature selection since the selection of any m features can be accomplished by selecting the appropriate $m \times M$ matrix $\mathbf{B}$, consisting only of 0's and 1's.

From the theoretical viewpoint, an ideal criterion of feature set effectivity in a given space is the classification error. However, this criterion function cannot be used in practical applications because of its computational complexity and, therefore, in the literature there are proposed various concepts of class separability on which alternative feature selection and extraction criteria are based.

Decell and Guseman [13] have developed an extensive bibliography of feature selection procedures prior to 1978. Devijver and Kittler [14] provide perhaps the only comprehensive exposition of such measures through early 1982.

Criteria for effectivity of feature set may be divided in two basic categories:

(i) Criteria based on the notion of interclass distance.

(ii) Criteria based on probabilistic distance measures, on the measures of probabilistic dependence and on the information measures.

Criteria of the first group are based on the simplest concept of the class separability:

the greater is the distance between patterns of two classes, the better is the separability of the two classes.

A main drawback of the criteria of the first group lies in the fact that they are not directly related to the probability of misclassification. "Good" are such features which give the maximum of the interclass distance. These criteria can be formulated using only the parameters $\boldsymbol{\mu}$, $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$, $\mathbf{S}_W$, $\mathbf{S}_B$ and $\mathbf{S}_W + \mathbf{S}_B$ defined in (2.51), (2.53), (2.54) and (2.52), respectively, when restricting ourselves to a certain class of distances. A detailed information about the probabilistic distribution of the classes is not used in defining the interclass distance and thus the interclass separability measures are attractive from the computational point of view. As opposed to the first group, these criteria have a direct relationship to the probability of misclassification. However, when solving real problems from practice, the required complete information about the probabilistic structure is rarely known. Moreover, these measures are by far less attractive from the computational point of view, than those based on the interclass distance.

More details about these measures can be found in Devijver and Kittler [14] and their references on these measures.

3.2 Comparison of extraction and selection principle from practical viewpoint

Both the approaches to the dimensionality reduction – feature selection and a more general feature extraction have their justification. When deciding between them, one has to be aware of their respective priorities and drawbacks so as to be able to choose the right approach from the point of view of ultimate goals and requirements, concerning the diagnostic task to be solved.

Let us state briefly the priorities and drawbacks of both the approaches from a practical viewpoint.

Feature selection

i) Priorities:

1) Selected features do not loose their original physical interpretation. They have an exactly defined meaning which is in many application fields an essential fact.
2) Data which are not selected among the diagnostically significant features need not be measured or collected at all during the application phase. This fact may result in a considerable saving of time and costs in the phase of data collection where the ratio of unavoidable tedious human work is very often a restrictive "bottle-neck" factor.

ii) Drawbacks:

1) Preservation of the physical interpretability is unfortunately paid by impossibility to achieve generally the optimum in the framework of selected

criterion of feature significance as compared to general feature extraction (optimally extracted subset of m transformed features will have generally a better discriminative ability than the best subset of m original data components).

Feature extraction

i) Priorities:

1) Discriminatory power achievable with optimally transformed is generally higher than in the case of restricting ourselves only to selection without any transformation.

ii) Drawbacks:

1) Since new features are the functions of all the original data components (e.g. their linear combination), they lack a straightforward and clear physical meaning. They may be looked upon as a certain abstraction and cannot be practically reasonably interpreted.
2) None of M original data components can be saved in the application phase. Quite contrary, all of them must be measured and collected since the resultant m-dimensional feature vector is derived only from the complete original M-dimensional data vector by applying a suitable transformation.

From the outlined priorities and drawbacks of both the approaches we can see that their properties are to a certain extent rather contradictory.

Feature selection methods will be more suitable in cases when the potential user puts emphasis on preserving the interpretability of original data and prefers decision-making on the basis of meaningful features. Furthermore, they will be suitable when one of the goals is to reduce tediousness and costs of data collecting by finding the data components which can be completely excluded from further collecting process. Most problems of medical differential diagnostics belong to this class of tasks.

On the other hand there exist applications where the requirements are quite opposite. The emphasis is laid on optimum reduction of the problem dimensionality and neither any transformation of features nor the necessity of measuring all the original data components represent any problem at all. A typical example of such a diagnostic task is the VCG (vectorcardiogram) classification where the primary data vector is represented by a time series (sampled VCG signal), having several hundred members. Since the physical interpretation of respective sampled points of the VCG curve is discussable anyhow and, moreover, since the signal is recorded automatically and thus the possibility to reduce the number of measured components plays obviously a less important role, feature extraction methods are preferable in this case.

Since most diagnostic and pattern recognition problems we have encountered have belonged to the first discussed class of problems, the feature selection approach

has been given priority in the PREDITAS system. However, a feature extraction method based on Karhunen-Loève expansion [48] is also considered. It is used in respective cases, when justified by the specific character of a solved task, to perform a primary reduction of dimensionality before processing data sets by the PREDITAS system.

3.3 Feature selection

The feature selection problem itself consists of two different problems:

(i) to order single features or sets of features according to their discriminating power;

(ii) to determine an optimal number of features (i.e. the optimal dimensionality).

Since the second problem is closely connected with the overall classifier performance, we shall postpone its discussion to the respective part of this paper, where other related problems like the reliability of classification, optimal number of steps in the stepwise decision rule, etc. will be discussed.

The first particular problem can be viewed as an optimization problem requiring a criterion function $J(\cdot)$ and a so called *search procedure.*

Let

$$Y = \{y_j : j = 1, 2, \ldots, M\}$$

be the set of M possible features representing the pattern, $M > m$. As mentioned in Section 3.1, the best subset X of m features

$$X = \{x_i : i = 1, 2, \ldots, m, x_i \in Y\}$$

is the combination of m features which maximizes a criterion function $J(\cdot)$ with respect to any other possible combination $F = \{f_i : i = 1, 2, \ldots, m, f_i \in Y\}$ of m features taken from Y.

We will call the *individual significance* $S_0(y_i)$ of the feature y_i, $i = 1, 2, \ldots, M$ the value of the criterion function $J(\cdot)$ if only the ith feature is used.

Let us order features $y_i \in Y$ so that

$$J(y_1) \geqq J(y_2) \geqq \ldots J(y_m) \geqq \ldots \geqq J(y_M)$$

A low value of $J(y_i)$ implies a low discriminative power and thus the simplest method of selecting the best feature set seems to be to select the m individually best features in Y, i.e. the "best" set X of cardinality m would be defined as

$$X = \{y_i : \forall i \leqq m\}$$

However, it is well known that the set of m individually best features is not necessarily the best set of size m even for the case of independent features. Since it is highly unlikely that the features will be independent in practice, some more sophisticated methods of selecting the best feature subset must be employed.

Unfortunately, the only way to ensure that really the best subset of m features

from a set of M features is chosen is to explore all $\binom{M}{m}$ combinations (cf. Cover and Van Campenhout [12]). It is apparent that direct exhaustive searching will not be possible and that in practical situations computationally feasible procedures will have to be employed.

In most *search* procedures the resultant feature set is constructed either by including to or excluding from the current feature subset. More precisely; feature subset to be evaluated at the kth stage of a search algorithm is constructed by adding to or substracting from the appropriate feature subset, obtained at the $(k-1)$th stage of the algorithm a certain number of features until the required feature set of cardinality m is obtained.

The search procedures start either from an empty set where the resultant set is consecutively built up by adding respective features, or from the complete set of measurements Y where the resultant set is gradually built up by successive excluding of superfluous features. The former approach is known as the "bottom up" search is latter while referred to as the "top-down" search. Equivalent terms that may be found in the literature too and which are used in the PREDITAS system are "*forward*" or "*including*" for the "bottom up", and "*backward*" or "*excluding*" for the "top down" searching respectively.

Let F_k be a set of k features from the set of M available features Y, i.e.

$$F_k = \{f_i : i = 1, 2, \ldots, k, f_i \in Y\}$$

Further denote by $\bar{F}_k$ the set of $(M-k)$ features obtained by removing k features $f_1, \ldots, f_k$ from the complete set Y, i.e.

$$\bar{F}_k = \{y_i : y_i \in Y, 1 \leqq i \leqq M, y_i \neq f_j, \forall j\}$$

Let $J(\cdot)$ be a chosen criterion function (measure of class separability). Then in the "*bottom up*" approach the best feature set X_k at the kth stage of an algorithm satisfies

$$J(X_k) = \max_{\{F_k\}} J(F_k),$$

where $\{F_k\}$ is the set of all the candidate sets of k features and it is as a rule determined at the $(k-1)$th step of the search procedure. The initial feature set is an empty set, i.e.

$$X_0 \equiv F_0 \equiv \emptyset$$

and the final feature set X is given as

$$X \equiv X_m .$$

In the "*top down*" approach the initial feature set $\bar{X}_0$ is given as

$$\bar{X}_0 \equiv \bar{F}_0 \equiv Y.$$

Successively reduced feature sets $\bar{X}_k$, $k = 1, 2, \ldots, M-m$ are constructed so that

$$J(\bar{X}_k) = \max_{\{\bar{F}_k\}} J(\bar{F}_k),$$

where $\{\bar{F}_k\}$ is the set of all the candidate combinations of features, $\bar{F}_k$. The final feature set X is defined as

$$X \equiv \bar{X}_{M-m} .$$

Narendra and Fukunaga [39] show that if the feature separability criterion used satisfies the monotonicity property then the *branch and bound search strategy* guarantees the optimal feature subset without explicitly evaluating all possible feature subsets. It is basically a top down search procedure but with a backtracking facility which allows all the possible combinations of m features to be examinated. If the feature selection criterion functions are monotone then for nested sets of features $\bar{X}_i$, $i = 1, 2, \ldots, k$, i.e.

$$\bar{X}_1 \supset \bar{X}_2 \supset \ldots \supset \bar{X}_k ,$$

the criterion functions, $J(\bar{X}_i)$, satisfy

$$J(\bar{X}_1) \geqq J(\bar{X}_2) \geqq \ldots \geqq J(\bar{X}_k) .$$

By a straightforward application of this property, many combinations of features can be rejected from the set of candidate feature sets.

4. FEATURE SELECTION AND SEARCH METHODS IN PREDITAS SYSTEM

This chapter deals with the principle of feature selection used in the PREDITAS system, as well as with the employed search procedures. Their formalized description is presented and respective search procedures are compared both from the theoretical and practical point of view.

4.1 Principle and methods of feature selection employed in PREDITAS system

The feature selection procedure employed in the PREDITAS system is based on the criterion function called the *measure of discriminative power* (MDP), which is the measure of class separability belonging to the first group of measures discussed in Section 3.1. The MDP is defined for the m-dimensional pattern vector $\mathbf{y} = (y_1, \ldots, y_m)^T$ $(m \leqq M)$ by

$$\lambda(\mathbf{y}) = \mathrm{tr}\,(\mathbf{S}_W^{-1}\mathbf{S}_B) , \tag{1}$$

where $\mathbf{S}_W$ and $\mathbf{S}_B$ denote the average within class scatter matrix and the between class scatter matrix, respectively, defined by (2.54) and (2.52) for $c = 2$. This feature selection criterion is equal to the maximum value of Fisher's discriminant ratio (2.59). The advantage of the criterion (1) is a computational simplicity since it is formulated using only the parameters $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$. The main drawback is the fact that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ is the necessary and sufficient condition for $\lambda = 0$ and that does not take into consideration the difference of class covariance matrix (see Duda and Hart

[16]). The covariance matrices Σ_i have only the normalizing role. Several attempts have been made to extend the Fisher criterion (Malina [37]).

Before describing the search procedures used in the PREDITAS system, let us briefly introduce some basic concepts of search procedures and then the definitions of individual features significance in the process of searching the best resultant subset.

As stated in the previous chapter, feature selection is usually realized by means of search procedures. These search procedures may be either including or excluding and their basic forms are called "*straight methods*". The search for feature selection is independent of the criterion function (as opposed to feature extraction).

Straight including and excluding methods can be shown to be only step optimal, therefore, various heuristic procedures aimed at closer approaching the optimal set are utilized. Two such methods, which we have called "*floating methods*" as opposed to "*straight methods*", have been designed and implemented in the PREDITAS system.

Before describing the corresponding algorithms formally, the following definitions have to be introduced:

Let $X_k = \{x_i: 1 \leqq i \leqq k, x_i \in Y\}$ be the set of k features from the set $Y = \{y_i: 1 \leqq \leqq i \leqq M\}$ of M available features. Denote $\lambda(X_k) \triangleq \lambda(\mathbf{x}_k)$, where $\mathbf{x}_k = (x_1, \ldots, x_k)^T$.

The *significance* $S_{k-1}(x_j)$ *of the feature* x_j, $j = 1, 2, \ldots, k$ *in the set* X_k is defined by

$$S_{k-1}(x_j) = \lambda(X_k) - \lambda(X_k - x_j)\,. \tag{2}$$

The *significance* $S_{k+1}(f_j)$ *of the feature* f_j from the set $Y - X_k$

$$Y - X_k = \{f_i: i = 1, 2, \ldots, M - k, f_i \in Y, f_i \neq x_l\ \forall l\}$$

with respect to the set X_k is defined by

$$S_{k+1}(f_j) = \lambda(X_k + f_j) - \lambda(X_k)\,. \tag{3}$$

Remark. For $k = 1$ the term feature significance in the set coincides with the term of individual significance defined in Section 3.3.

We shall say that the feature x_j from the set X_k is

(a) the *most significant* (best) feature

(b) the *least significant* (worst) feature *in the set* X_k if

$$\text{(a)} \qquad S_{k-1}(x_j) = \max_{1 \leqq i \leqq k} S_{k-1}(x_i) \Rightarrow \lambda(X_k - x_j) = \min_{1 \leqq i \leqq k} \lambda(X_k - x_i) \tag{4a}$$

$$\text{(b)} \qquad S_{k-1}(x_j) = \min_{1 \leqq i \leqq k} S_{k-1}(x_i) \Rightarrow \lambda(X_k - x_j) = \max_{1 \leqq i \leqq k} \lambda(X_k - x_i)\,, \tag{4b}$$

respectively.

We shall say that the feature f_j from the set

$$Y - X_k = \{f_i: i = 1, 2, \ldots, M - k, f_i \in Y, f_i \neq x_l\ \forall l\} \quad \text{is}$$

(a): the *most significant* (best) feature

(b): the *least significant* (worst) feature

with respect to the set X_k if

$$\text{(a)} \qquad S_{k+1}(f_j) = \max_{1 \leqq i \leqq M-k} S_{k+1}(f_i) \Rightarrow \lambda(X_k + f_j) = \max_{1 \leqq i \leqq M-k} \lambda(X_k + f_i) \tag{5a}$$

$$\text{(b)} \qquad S_{k+1}(f_j) = \min_{1 \leqq i \leqq M-k} S_{k+1}(f_i) \Rightarrow \lambda(X_k + f_j) = \min_{1 \leqq i \leqq M-k} \lambda(X_k + f_i), \tag{5b}$$

respectively.

4.2 Algorithms of searching procedures in PREDITAS system

As stated before, there are two basic types of searching procedures in the PREDITAS system, namely *straight* and *floating* methods. In each type there are two algorithms – *including* (or *forward*) and *excluding* (or *backward*) named according to the basic direction of search. We shall describe first both the direct methods and afterwards the floating methods.

4.2.1 Straight methods

(i) *Sequential Including (Forward) Selection* (SIS)

The SIS method is a search procedure where one feature at a time is added to the current feature set. At each stage the feature to be included in the feature set is selected from among the remaining available features so that the new set of features gives the maximum value of MDP.

Let us suppose k features have already been selected to form the feature set X_k. If the feature f_j from the set $Y - X_k$ is the best feature with respect to the set X_k (i.e. (5a) is satisfied), then the new feature set X_{k+1} is given as

$$X_{k+1} = X_k + f_j$$

The algorithm is initialized by setting $X_0 \equiv \emptyset$.

The main disadvantage of the SIS method lies in the fact that once a feature is included in the feature set, the method provides no possibility for discarding it in later stages from the selected feature set though this could yield a higher value of MDP. (It is apparent that e.g. already the best pair of features does not necessarily contain the individually best feature selected in the first step of the algorithm).

(ii) *Sequential Excluding (Backward) Selection* (SES)

The SES method is a search procedure where one feature at a time is excluded from the current feature set. At each stage the feature to be excluded from the feature set is selected so that the new reduced set of features gives the maximum possible value of MDP (its decrease is minimized).

Let us suppose that at the kth stage of the algorithm k features have already been excluded from $\overline{X}_0 \equiv Y$ to form the feature set $\overline{X}_k$. If the feature x_j is the worst feature in the set $\overline{X}_k$ (i.e. (4b) is satisfied), then the reduced feature set $\overline{X}_{k+1}$ is given as

$$\overline{X}_{k+1} = \overline{X}_k - x_j$$

The algorithm is initialized by setting $X_0 \equiv Y$.

The main disadvantage of the SES method is analogous to that of the SIS method – once a feature is excluded from the feature set, the method provides no possibility for including it in later stages to the selected feature set though this could yield a higher value of MDP.

4.2.2 Floating methods

As stated earlier, two more sophisticated heuristic procedures aimed at overcoming the drawbacks of the straight methods have been designed for the PREDITAS system. They have been named "floating methods" since the resulting dimensionality in respective stages of the algorithm is not changing monotonously but is really "floating" up and down.

Though both these methods switch between including and excluding the features, we recognize two different algorithms according to the dominant direction of searching.

(i) *Sequential Including (Forward) Floating Selection* (SIFS)

Suppose that we are at an arbitrary node at level k. That means that for all preceding nodes at levels $i = 1, 2, \ldots, k-1$, we have value $\lambda_i \triangleq \lambda(X_i)$, where X_i is the set of i features. The current feature set after k features have been selected in $X_k = \{x_1, \ldots, x_k\}$ and $\lambda_k = \lambda(X_k)$.

The algorithm realizing the SIFS method can be then described as follows:

Step 1: (*Including*). Using straight SIS method include feature x_{k+1} from the set of available features, $Y - X_k$, to X_k, to form feature set X_{k+1} (i.e. the most significant feature x_{k+1} with respect to set X_k has been included in X_k). Therefore

$$X_{k+1} = X_k + x_{k+1}$$

Step 2: (*Conditional excluding*). Find the least significant feature in the set X_{k+1}. If x_{k+1} is the least significant feature in the set X_{k+1}, i.e.

$$\lambda_k = \lambda(X_{k+1} - x_{k+1}) \geqq \lambda(X_{k+1} - x_j), \quad \text{for} \quad \forall j = 1, \ldots, k,$$

then set $k := k + 1$ and return to Step 1.

If x_r, $1 \leqq r \leqq k$ is the least significant feature in the set X_{k+1}, i.e.

$$\lambda(X_{k+1} - x_r) > \lambda_k,$$

then exclude x_r from X_{k+1} to form a new feature set X'_k i.e.

$$X'_k = X_{k+1} - x_r.$$

Now if $k = 2$, set $X_k := X'_k$, $\lambda_k := \lambda'_k$ and return to Step 1 else go to Step 3.

Step 3: Find the least significant feature x_s in set X'_k. If

$$\lambda'_{k-1} \triangleq \lambda(X'_k - x_s) > \lambda_{k-1},$$

then exclude x_s from X'_k to form a new set X'_{k-1}, i.e.

$$X'_{k-1} = X'_k - x_s .$$

Set $k := k - 1$. Now if $k = 2$ then set $X_k := X'_k$, $\lambda_k := \lambda'_k$ and return to Step 1 else repeat Step 3.

The algorithm is initialized by setting $k = 0$ and $X_0 \equiv \emptyset$ and the SIS method is used until feature set of cardinality 2 has been obtained. Then the algorithm continues with Step 1.

(ii) *Sequential excluding (backward) floating selection* (SEFS)

Suppose that we are at an arbitrary node at level k. It means that for all the preceding nodes at levels $i = 0, 1, \ldots, k-1$ we have value $\lambda_{M-i} \triangleq \lambda(\overline{X}_i)$, where $\overline{X}_i$ is the set obtained by removing i atributes from the complete set of features $\overline{X}_0 = Y$. The current feature set after k features $x_1, \ldots, x_k$ have been discarded is $\overline{X}_k$ and $\lambda_{M-k} \triangleq \lambda(\overline{X}_k)$.

The algorithm realizing SEFS method can be then described as follows:

Step 1: (*Excluding*). Using straight SES method exclude from the set $\overline{X}_k$ feature x_{k+1} to form reduced set $\overline{X}_{k+1}$ (i.e. we exclude from the set $\overline{X}_k$ the least significant feature in the set $\overline{X}_k$). Therefore

$$\overline{X}_{k+1} = \overline{X}_k - x_{k+1} .$$

Step 2: (*Conditional including*). Find among the excluded features the most significant feature with respect to set $\overline{X}_{k+1}$. If x_{k+1} is the most significant feature with respect to $\overline{X}_{k+1}$, i.e.

$$\lambda_{M-k} = \lambda(\overline{X}_{k+1} + x_{k+1}) \geqq \lambda(\overline{X}_{k+1} + x_j) \quad \text{for} \quad \forall j = 1, \ldots, k ,$$

then set $k := k + 1$ and return to Step 1.

If x_r, $1 \leqq r \leqq k$ is the most significant feature with respect to set $\overline{X}_{k+1}$, i.e.

$$\lambda'_{M-k} \triangleq \lambda(\overline{X}_{k+1} + x_r) > \lambda_{M-k} ,$$

then include x_r to set $\overline{X}_{k+1}$ to form the new feature set $\overline{X}'_k$, i.e.

$$\overline{X}'_k = \overline{X}_{k+1} + x_r .$$

Now if $k = 2$, then set $\overline{X}_k := \overline{X}'_k$, $\lambda_{M-k} := \lambda'_{M-k}$ and return to Step 1 else go to Step 3.

Step 3: Find among the excluded features the most significant feature x_s with respect to set $\overline{X}'_k$.

If

$$\lambda'_{M-(k+1)} \triangleq \lambda(\overline{X}'_k + x_s) > \lambda_{M-(k+1)} ,$$

then include x_s to set $\overline{X}'_k$ to form the new set $\overline{X}'_{k+1}$, i.e.

$$\overline{X}'_{k+1} = \overline{X}'_k + x_s .$$

Set $k := k + 1$. Now if $k = 2$ then set $\overline{X}_k := \overline{X}'_k$, $\lambda_{M-k} := \lambda'_{M-k}$ and return to Step 1 else repeat Step 3.

The algorithm is initialized by setting $k = 0$, $\overline{X}_0 = Y$ and the SES method is used until 2 least significant features are excluded. Then the algorithm continues with Step 1.

4.3 Comparison and discussion of described search methods

The comparison of all four in Section 4.2 employed methods together is rather difficult. The differences between straight methods and floating ones lie mainly in computational and other practical aspects. On the other hand, the difference between forward and backward methods is in some respect quite fundamental, though the computational point of view plays its role, too. We shall, therefore, discuss and compare first the straight methods alone, then the floating methods and finally the straight methods with floating ones.

4.3.1 Comparison of straight methods

When deciding which of the two straight methods should be preferred in a particular task, one must be aware of their differences both from the theoretical and practical point of view.

The SIS method has the drawback inherent in all the "forward" sequential selection methods. The decision about including a particular feature at the first few steps of the algorithm is made on the basis of $\mathbf{S}_W$ and $\mathbf{S}_B$ matrices of a low order k ($k = 1, 2, \ldots$). From it follows that only low order statistical dependencies of original pattern components are taken into consideration. This drawback does not concern the "backward" sequential relation methods to which the SES method belongs. The decision on excluding a particular feature is made taking into consideration the statistical dependencies of original pattern components in full dimensionality.

More explicitly, the best features with respect to the set X_k in the SIS method are at earlier steps of the selection algorithm determined on the basis of the set X_k of low cardinality. Quite contrary, the worst features in the set X_k for the SES method are at earlier steps of the algorithm determined on the basis of the set X_k of high cardinality.

The just stated facts seem to prefer uniquely the SES method. However, the problem is somewhat more complicated. In practice we are restricted to training sets of limited size, which is often not adequate to the problem dimensionality. In this case obviously the training set is not sufficiently representative with respect to the basic data set. The corresponding $\mathbf{S}_W$ and $\mathbf{S}_B$ matrices of full dimensionality m are, therefore, not sufficiently representative either not mentioning numerical problems involved. From it follows a hidden drawback of the SES method, namely that though in comparison with the SIS method it takes better into consideration complex

mutual relations, these relations are not always determined quite reliably. On the other hand, though the SIS method utilizes at the beginning only less complex mutual relations, they are, however, determined more reliably even for training sets of a moderate size.

From the computational point of view there are some basic differences, too, since the numerical propertirs of the SIS and SES methods are different. Since the SES method starts at full dimension m, where inversion of the matrix S_W is computed, it may happen that due to dependencies of data components this matrix is singular, generally it means numerical instability. Thus its inversion cannot be computed and the process of feature selecting is terminated uncompleted with the warning that one of the features should be eliminated before proceeding further (this elimination is not made automatically on purpose since any of the list of dependent features can be eliminated and the reasons for selecting one depend on the user's scale of preferences).

On the other hand, the SIS method in such a case rejects a feature, the adding of which would lead to singularity, and the selection process will continue until its completion. Thus from purely numerical reasons, to begin with the SIS method is preferable in those cases when singularities in matrices can be expected. The SIS method ends in this case in a maximum dimensionality for which are matrices regular and then the backward SES method can be utilized.

4.3.2 Comparison of floating methods

The differences between the SIFS (*including*) and SEFS (*excluding*) methods are basically the same as those between the straight methods. However, the arguments in favour or against are not so strong in this case, since due to the "floating" principle both methods have the chance to correct decisions made wrongly in previous steps. In practice both the methods give quite often, though not always, the identical results.

Nevertheless, the numerical considerations discussed in the previous paragraph hold even in this case. When very little is known about the data set and some dependent components and thus the singularity of matrices is to be expected, the SEFS method should not be used prior to finding the maximum dimensionality for matrices regularity.

4.3.3 Comparison of straight and floating methods

The main difference between these two types of methods follows from the purpose with which have been the floating methods designed. They remove the fundamental drawback of straight methods – the total impossibility to make any desirable corrections in later stages of the algorithm. Though always not absolutely optimal, they approach the optimum subset of any dimension much closer than the only step-optimal straight methods.

However, as usually this advantage is at the expanse of priorities of straight methods. The floating methods need far more computer time so their use should be justified by special circumstances and only for the case of reliable representative training sets (there is no point in improving the selection criterion by a margin on the basis of complicated computations made on the basis of quite unreliable data).

As far as the form of final results is concerned, there is an important difference. The straight methods give a clear ordering of features significance (either in increasing or decreasing order). The form of results in floating methods is quite different. Since the optimal "compositions" of subsets of cardinalities differing by one (e.g. X_k, X_{k+1}) may easily differ by more than one feature, it is here in principle impossible to order features according to their significance. The result is in the form of a binary table denoting the presence or absence of respective features in optimal subsets of successive dimensions.

To summarize, we can state that at first the straight methods should be used to get a clear ordering of features significance and only afterwards the floating methods should be used when justified by special requirements and circumstances.

5. SOME PRACTICAL IMPLICATIONS

The design of the classifier in statistical pattern recognition generally depends on what kind of information is available about the class-conditional densities.

If the information about these densities is complete, then the Bayes decision theory may be used.

If it is assumed that the form of the class-conditional densities is known but some of the parameters of the densities are unknown then we have a *parametric decision* problem. In that case we replace the unknown parameters in the density functions by their estimated values.

If the form of the densities is not known then we operate in nonparametric mode. In that case we must either estimate the density function or use some *nonparametric decision* rule (e.g. k-nearest neighbor rule, see e.g. Devijver and Kittler [14].

Estimation of the expected performance of a classifier is an important, yet difficult problem.

In practice, the class-conditional densities, in some case also the a priori class probabilities, are seldom specified and only a finite number of *learning* or *training samples* is available, either labeled (*supervised learning*) or unlabeled (*unsupervised learning*). The label on each training pattern represents the class to which that pattern belongs. Unsupervised learning refers to situations, where is it difficult or expensive to label the training samples. This is a difficult problem and, in practice, often has to rely on techniques of cluster analysis [15] to identitfy natural grouping (classes) present in data. The results of cluster analysis are useful for forming hypotheses about the data, classifying new data, testing for the homogeneity of the data, and for compressing the data.

Pattern recognition research has considered various questions concerning the relationship between the limited size of training set, the number of features and the estimation of performance criteria. The designer must decide whether this sample size is adequate or not, and also decide how many samples should be used to design the classifier and how many should be used to test it.

In practice it is difficult to express or measure the loss incurred when the pattern is misclassified. In such cases the criterion of performance of the decision rule is the simple error rate rather than the average risk. It is very difficult to obtain an analytic expression of the error rate for a particular classification problem unless, of course, the class-conditional densities are Gaussian with common covariance matrices (see Section 2.4.3).

Various methods have been proposed to estimate the error rate. They depend on how the available samples are divided into training and test sets.

(i) *Resubstitution method*: In this method, the classifier is tested on the data set which has been used to design the classifier. But in this "testing on the training data" this method is optimistically biased: a sample based classification rule usually appears to be more successful with the data from which it was designed than in classifying independent observations [21].

(ii) *Holdout method*: In this method the available samples are randomly partitioned into two groups. One of these groups is used to design the classifier and the other one is used to test it. This provides an independent test set, but drastically reduces the size of the learning set. This method gives a pessimistically biased performance.

(iii) *Leave-One-Out method*: This method [35] is designed to alleviate these difficulties. It avoids drastically dividing the available sample set into the design set and the testing set. In this method each sample is used both for design and testing although not at the same time. The classification rule is designed using $(N - 1)$ samples and then tested on the remaining sample, where N is the total number of available samples. This is repeated N times with different design sets of size $(N - 1)$. Thus the procedure utilizes all available samples more efficiently.

By using these last two methods simultaneously we can obtain upper and lower bounds of the true performance of the classifier.

(iv) *Bootstrap method*: More recently, Efron [17], [18] proposed a resampling method, in which the artificial samples are generated from the existing samples, and the optimistic bias between the resubstitution error and the classifier error when tested on independent samples is eatimated from them.

Toussaint [47] catalogs these and other testing methods and gives an overview of some of the early associated work. More recent work is surveyed in Hand [30].

Fukunaga and Hayes investigated the effect of sample size on a family of functions, and found a manageable expression for the errors of classifiers [24], and applied these errors expression to the various methods of error estimation [25].

The current interest in paterrn recognition is to expand the domain of applications

of this methodology. This has resulted in more careful attention to practical issues such as reliable estimates of error rate, computational complexity, robustness of classifier, implementation of decision algorithms into hardware and utilizing contextual or expert information in the decision rules. These are difficult problems and we do not yet know all the answers.

6. SAMPLE-BASED PATTERN RECOGNITION

In the sequel, we shall suppose that information about the classes ω_i is available only in the form of the training sample sets $\mathcal{T}_i$ generated from classes ω_i, $i = 1, \ldots, c$. We shall assume that ith training set

$$\mathcal{T}_i = \{\mathbf{x}_{ij} \in \Omega: j = 1, 2, \ldots, N_i\}, \quad i = 1, \ldots, c \tag{1}$$

is a collection of N_i independent pattern vectors $\mathbf{x}_{ij}$ in Ω, identically distributed; N_i is the number of the pattern vectors in the set $\mathcal{T}_i$ called the *size* of the *training set* $\mathcal{T}_i$. The c training sets we shall collectively denote by $\mathcal{T}_{(N)} = \{\mathcal{T}_1, \ldots, \mathcal{T}_c\}$. The size of the training set $\mathcal{T}_{(N)}$ is $N = \sum_{i=1}^{c} N_i$.

The search for informative features and the design of an effective classifier, the essential steps in pattern recognition system design, will be now considered from the view of this kind of available information.

6.1 Sample-based classification procedures derived from density estimators

As stated earlier in Chapter 5, the a priori class probabilities and class-conditional densities are seldom specified. Then probability of error cannot be evaluated for an arbitrary classification rule; nor can the optimal rule be calculated. But labeled samples usually are available to determine estimators of class probabilities and class-conditional densities, and therefore construct decision rule based on this samples.

The decision rule based on the training samples we shall call *sample-based decision rule* and denote by $\hat{d}$.

Classification rules are always sample-based in practice. But from the perspective of statistical decision theory, sample-based rules are substitutes for unknown optimal rules.

Three issues concerned with sample-based classification rules are principal ones [26].

1. How to construct from the sample a "reasonable" rule to classify any future pattern $\mathbf{x}$ from a mixed population?
2. How good is "reasonable" rule relatively to the unknown optimal rule?
3. How good is "reasonable" rule in absolute sense? What is the estimate of its error rate?

The same sample data which determine a decision rule are often used to estimate its non-error rate, i.e. probability of correct classification. But in this "testing on the

training data" intuitive estimation methods are optimistically biased: a sample-based decision rule usually appears to be more successful with data from which it was designed than in classifying independent observations.

Ideally one would like to construct a less biased estimator which also has small variance, is robust when normality fails, and can be computed easily (e.g. Devijver and Kittler [14] reviewed estimators with these multiple criteria in mild).

6.1.1 Sample-based minimum error rate classification procedure

First, we shall consider the sample-based counterpart of the decision rule minimizing error rate or maximizing non-error rate defined in Section 2.3.

Most non-error rate estimators descend from one or the other of two concepts:

(i) *plug-in* methods substitute estimated parameters into theoretical expression for non-error rate;

(ii) *counting* a correct classifications of labeled samples and dividing by sample size gives sample non-error (success) proportion;

(iii) *posterior probability* estimators combine aspects of both approaches.

(i) *A general plug-in estimator of the error rate*

A general plug-in estimator of the error rate, is formed by substituting the sample-based estimates $\hat{P}_i$ and $\hat{p}_i$ of a priori probability P_i and class-conditional density p_i, $i = 1, \ldots, c$ into appropriate expression (2.7).

The plug-in estimator $\hat{E}$ of the error rate function E in (2.7) is

$$\hat{E}(d) = \sum_{j=1}^{c} \int_{\Omega_j} \hat{P}_i \hat{p}_i(\mathbf{x})\, \mathrm{d}\mathbf{x}\,, \quad i \neq j$$

for each rule $d \in \mathscr{D}$.

Making an analogy with the criterion for optimality, it seems reasonable to use a classification rule which minimizes $\hat{E}$. The argument which proves that the rule d^* defined by (2.27) is optimal, shows also that $\hat{E}$ is minimized by the density plug-in rule $\hat{d}$ whose partition sets are

$$\hat{\Omega}_i = \{\mathbf{x} \in \Omega\colon \hat{P}_i \hat{p}_i(\mathbf{x}) = \max_{j=1,\ldots,c} \hat{P}_j \hat{p}_j(\mathbf{x})\}\,, \quad i = 1, \ldots, c\,.$$

That is,

$$\inf_{d \in \mathscr{D}} \hat{E}(d) = \hat{E}(\hat{d}) = 1 - \int \max_{j=1,\ldots,c} \hat{P}_j \hat{p}_j(\mathbf{x})\, \mathrm{d}\mathbf{x}\,. \tag{2}$$

Therefore, the *plug-in decision rule* $\hat{d}$ which is Bayes optimal with respect to the estimated distributions is given by

$$\hat{d}(\mathbf{x}) = d_i \quad \text{if} \quad \mathbf{x} \in \hat{\Omega}_i\,, \quad i = 1, \ldots, c \tag{3}$$

Remark. If normal densities $p_i(\mathbf{x}) = p(\mathbf{x}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ $i = 1, 2$ with common covariance matrices are parametrically estimated by $\hat{p}_i(\mathbf{x}) = p(\mathbf{x}, \boldsymbol{\mu}_i\hat{}\,, \boldsymbol{\Sigma}\hat{}\,)$ with mean vector estimates $\boldsymbol{\mu}_i\hat{}$ and covariance matrix estimate $\boldsymbol{\Sigma}\hat{}$ and $P_i = \frac{1}{2}$, then region $\hat{\Omega}_1$ is the

half-space

$$\hat{\Omega}_1 = \{\mathbf{x} \in \Omega: \hat{D}(\mathbf{x}) < 0\}\,, \tag{4}$$

and

$$\hat{\Omega}_2 = \Omega - \hat{\Omega}_1\,, \tag{5}$$

where

$$\hat{D}(\mathbf{x}) = ((\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^{\mathrm{T}})\,\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \tfrac{1}{2}(\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2))\,.$$

Thus the plug-in rule $\hat{d}$ defined by (3) with $\hat{\Omega}_1$ and $\hat{\Omega}_2$ specified by (4) and (5), respectively, is closely related to Anderson linear discriminative hyperplane (2.49) with estimated parameters $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_2$ and $\hat{\boldsymbol{\Sigma}}$.

The statistic $\hat{E}(\hat{d})$ given by (2) may be called the *apparent error rate* [26]. The apparent error rate is an estimator of $E(d^*)$, the error rate of unknown optimal rule d^* or of $E(\hat{d})$, the true error rate of the sample-based rule $\hat{d}$ so called *actual error rate*. It is clear that the error rate of the optimal rule is not greater than the actual error rate:

$$E(d^*) = E^* = \inf_{d \in \mathscr{D}} E(d) \leqq E(\hat{d})\,.$$

But the error rate of optimum rule d^* often is greater than the expected value of the apparent error rate. A tendency of the apparent error rate toward optimistic bias arises from the simple fact that the expectation of the maximum several random variables exceeds the maximum of their expectations. Glick [26] stated that if the density estimators $\hat{p}_i$, satisfy the inequality $\mathsf{E}\{\hat{P}_i\,\hat{p}_i(\mathbf{x})\} \geqq P_i\,p_i(\mathbf{x})$ for $i = 1, \ldots, c$ and almost all $\mathbf{x} \in \Omega$ then $\mathsf{E}\{\hat{E}(\hat{d})\} \leqq E^* \leqq E(\hat{d})$. He derived that a sample-based density plug-in rule $\hat{d}$ is asymptotically optimal, subject to mild conditions on the density estimators; and also the apparent error rate $\hat{E}(\hat{d})$ converges to optimal error rate E^*.

(ii) *Counting: the sample success proportion.*

More intuitive than the plug-in method is the counting estimator or sample success proportion.

Let denote h_i the indicator function of the optimal partition region

$$\Omega_i^* = \{\mathbf{x} \in \Omega: P_i\,p_i(\mathbf{x}) \geqq P_j\,p_j(\mathbf{x})\,,\quad j = 1, \ldots, c\}\,,\quad i = 1, \ldots, c$$

$$h_i(\mathbf{x}) = \begin{cases} 1 & \text{if } P_i\,p_i(\mathbf{x}) = \max\limits_{j=1,\ldots,c} P_j\,p_j(\mathbf{x}) \\ 0 & \text{otherwise} \end{cases}$$

$$\sum_{i=1}^{c} h_i(\mathbf{x}) = 1 \quad \text{for all} \quad \mathbf{x} \in \Omega\,.$$

Decision rule d^* has class-conditional probability of correct classification, given ith class, which can be represented as

$$G_{ii} = \int_{\Omega_i^*} p_i(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int h_i(\mathbf{x})\,p_i(\mathbf{x})\,\mathrm{d}\mathbf{x} = \mathsf{E}_i\{h_i(\mathbf{x})\}\,. \tag{6}$$

A sample mean $\sum_j h_i(\mathbf{x}_{ij})/N_i$ is unbiased and, as sample size $N_i \to \infty$, converges to the theoretical $\mathsf{E}_i\{h_i(\mathbf{x})\}$.

Hence, it follows from (2.30) and (6) that for equal a priori probabilities $P_i = 1/c$, $i = 1, \ldots, c$ the non-error rate of the optimal decision rule can be represented as

$$C^* = \frac{1}{c} \sum_{i=1}^{c} \mathsf{E}_i\{h_i(\mathbf{x})\} = 1 - E^*$$

and for sample size $N_i = N/c$, the mean

$$\bar{C}(d^*) = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N_i} h_i(\mathbf{x}_{ij}) \tag{7}$$

is proportion of sample observations classified correctly by decision rule d^*. It is clear, that $\bar{C}(d^*)$ in (7) is unbiased and converging to the optimal rate C^*.

If densities p_i, $i = 1, \ldots, c$ are estimated by $\hat{p}_i$ and optimal decision rule d^* is estimated by $\hat{d}$ derived from the density estimators, then the corresponding indicators are

$$\hat{h}_i(\mathbf{x}) = \begin{cases} 1 & \text{if } \hat{P}_i \hat{p}_i(\mathbf{x}) = \max\limits_{j=1,\ldots,c} \hat{P}_j \hat{p}_j(\mathbf{x}), \quad i = 1, \ldots, c \\ 0 & \text{otherwise} \end{cases}$$

$$\sum_{i=1}^{c} \hat{h}_i(\mathbf{x}) = 1 \quad \text{for all} \quad \mathbf{x} \in \Omega .$$

The non-error rate estimator called *sample non-error rate* becomes

$$\bar{C}(\hat{d}) = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N_i} \hat{h}_i(\mathbf{x}_{ij}), \tag{8}$$

which is proportion of sample observations classified correctly by decision rule $\hat{d}$.

If the density estimators $\hat{p}_i$, $i = 1, \ldots, c$ were determined from data, say $\mathbf{x}'_{ij}$, which were independent of the $\mathbf{x}_{ij}$ explicit in the expression for $C(\hat{d})$, then $\mathsf{E}_i\{\hat{h}_i(\mathbf{x}_{ij})\}$ would be the sample-based decision's ith class conditional non-error rate and sample success proportion $\bar{C}(\hat{d})$ would be an unbiased estimator of $C(\hat{d})$, the true non-error rate of sample-based decision $\hat{d}$.

But in "testing on the training data" the sample success proportion is usually biased even more optimistically then the plug-in estimator. See studies cited in discussion of the plug-in method, e.g. [35].

(iii) *Posterior probability estimators*

Let optimal decision's posterior probability of correct classification, given observation $\mathbf{x}$, be denoted by

$$h(\mathbf{x}) = \frac{\max\limits_{j=1,\ldots,c} p_j(\mathbf{x})}{\sum\limits_{i=1}^{N_i} p_i(\mathbf{x})} \tag{9}$$

for equal a priori probabilities.

Then the non-error rate of the optimal decision rule defined by (2.27) can be

represented as

$$C^* = \frac{1}{c} \int \max \{p_1(\mathbf{x}), \ldots, p_c(\mathbf{x})\} \, d\mathbf{x} =$$

$$= \frac{1}{c} \int h(\mathbf{x}) \{p_1(\mathbf{x}) + \ldots + p_c(\mathbf{x})\} \, d\mathbf{x} = \frac{1}{c} \sum_{i=1}^{c} \mathsf{E}_i\{h(\mathbf{x})\} .$$

Corresponding sample mean $\sum_j h(\mathbf{x}_{ij})/N_i$ is unbiased and (by the strong law of large numbers) converges to $\mathsf{E}_i\{h(\mathbf{x})\}$ with probability one as sample numbers $N_i \to \infty$. Hence

$$\bar{C}(d^*) = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N_i} h(\mathbf{x}_{ij})$$

is unbiased and converges to $C^* = C(d^*)$ under either of the following sampling schemes:

- observations $\mathbf{x}_{ij}$, $i = 1, \ldots, c$, $j = 1, 2, \ldots, N_i$ are sampled separately from the c classes with equal sample sizes $N_i = N/c$; or
- observations are sampled from the mixture of classes, fixing only the total size $N = \sum_i N_i$ (in this sampling N_i would be binomial random variables and a priori probabilities P_i need not be known since they are implicitly estimated by proportions N_i/N, $i = 1, \ldots, c$).

If densities p_i are not specified, but estimated by some $\hat{p}_i$ (parametric or non-parametric) then the posterior probability (9) must be estimated by a corresponding $\hat{h}$. The posterior probability estimator of non-error rate becomes

$$\bar{C}(\hat{d}) = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N_i} \hat{h}(\mathbf{x}_{ij}) . \tag{10}$$

This expression, unlike $\bar{C}(d^*)$, is not unbiased for C^* even if, first, each $\hat{p}_i(\mathbf{x})$ is unbiased for $p_i(\mathbf{x})$ at all points $\mathbf{x}$; and, second, the density estimates are determined from sample data, say $\mathbf{x}'_{ij}$, which are independent of the $\mathbf{x}_{ij}$ explicit in the expression for $\bar{C}(\hat{d})$.

The estimator $\bar{C}(\hat{d})$ differs from $\tilde{C}(\hat{d})$, where independence implies unbiasedness. Moreover, parametric normal density estimation is not pointwise unbiased; and "testing on the training data" violates the independence condition. So $\bar{C}(\hat{d})$, considered as an estimator for the optimal rate C^*, in practice can have quite complicated bias.

6.1.2 Sample-based minimum risk classification procedure

Let us now concentrate our attention to the sample-based counterpart of the decision (2.13) in Section 2.2.

(i) *A general plug-in estimators of the average risk, the error rate and the reject rate*

A general plug-in estimators of the average risk, the error rate and the reject rate, are formed by substituting the sample-based estimates $\hat{P}_i$ and $\hat{p}_i$ of a priori probability P_i and a class conditional density p_i, respectively, $i = 1, \ldots, c$ into appropriate expressions.

The plug-in estimator $\hat{R}$ of the average risk R in (2.11), $\hat{E}$ of the error rate E in (2.7) and $\hat{R}_r$ of the reject rate R_r in (2.8) have, respectively, values

$$\hat{R}(d) = \sum_{j=1}^{c} L_{0j}\hat{P}_j \int_{\Omega_0} \hat{p}_j(\mathbf{x})\,\mathrm{d}\mathbf{x} + \sum_{i=1}^{c}\sum_{j=1}^{c} L_{ij}\hat{P}_j \int_{\Omega_i} \hat{p}_j(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$\hat{E}(d) = \sum_{j=1}^{c} \hat{P}_j \int_{\Omega_i} \hat{p}_j(\mathbf{x})\,\mathrm{d}\mathbf{x}, \quad j \neq i,$$

$$\hat{R}_r(d) = \sum_{j=1}^{c} \hat{P}_j \int_{\Omega_0} \hat{p}_j(\mathbf{x})\,\mathrm{d}\mathbf{x},$$

for each rule $d \in \mathscr{D}$.

Similarly as in the preceding section making an analogy with the criterion for optimality (2.12), we use a classification rule which minimizes $\hat{R}$. The average risk $\hat{R}$ is minimized by the density plug-in rule $\hat{d}$ whose partition sets are

$$\hat{\Omega}_i = \{\mathbf{x} \in \Omega: \hat{r}_i(\mathbf{x}) = \min_{j=0,\ldots,c} \hat{r}_j(\mathbf{x})\}, \quad i = 0, \ldots, c$$

where

$$\hat{r}_i(\mathbf{x}) = \sum_{j=1}^{c} L_{ij}\hat{P}_j\,\hat{p}_j(\mathbf{x})/\hat{p}(\mathbf{x}) \tag{11}$$

and

$$\hat{p}(\mathbf{x}) = \sum_{i=1}^{c} \hat{P}_i\,\hat{p}_i(\mathbf{x}).$$

Therefore

$$\hat{d}(\mathbf{x}) = d_i \quad \text{if} \quad \mathbf{x} \in \hat{\Omega}_i, \quad i = 0, \ldots, c. \tag{12}$$

That is

$$\inf_{d\in\mathscr{D}} \hat{R}(d) = \hat{R}(\hat{d}) = \sum_{j=1}^{c} L_{0j}\hat{P}_j \int_{\hat{\Omega}_0} \hat{p}_j(\mathbf{x})\,\mathrm{d}\mathbf{x} + \sum_{i=1}^{c}\sum_{j=1}^{c} L_{ij}\hat{P}_j \int_{\hat{\Omega}_i} \hat{p}_j(\mathbf{x}). \tag{13}$$

The statistic $\hat{R}(\hat{d})$ defined in (13) we may call the *apparent average risk*. $\hat{R}(\hat{d})$ is the estimator either of the average risk $R(d^*)$ of the unknown optimal decision rule d^* defined in (2.17) or of $R(\hat{d})$, so called the *actual average risk*, i.e. the value of the average risk function R at $d = \hat{d}$.

It is clear that the average risk of the minimum risk decision rule is not greater than the average risk of the sample-based rule $\hat{d}$ defined by (12)

$$R(d^*) = R^* = \inf_{d\in\mathscr{D}} R(d) \leqq R(\hat{d}).$$

But the average risk of the optimal rule d^* often is greater than the expected value of the apparent average risk.

(ii) *Counting: the sample success proportion.*
Let h_i denote the indicator function of the optimal partition regions

$$\Omega_i^* = \{\mathbf{x} \in \Omega: r_i(\mathbf{x}) = \min\{r_0(\mathbf{x}), \ldots, r_c(\mathbf{x})\}\}, \quad i = 0, \ldots, c$$

$$h_i(\mathbf{x}) = \begin{cases} 1 & \text{if } r_i(\mathbf{x}) = \min\limits_{j=0,\ldots,c} r_j(\mathbf{x}) \\ 0 & \text{otherwise} \end{cases}$$

$$\sum_{i=0}^{c} h_i(\mathbf{x}) = 1 \quad \text{for all} \quad \mathbf{x} \in \Omega .$$

Then the decision rule d^* defined in (2.13) has the average risk

$$R^* = \sum_{j=1}^{c} L_{0j} P_j \int h_0(\mathbf{x})\, p_j(\mathbf{x})\, d\mathbf{x} + \sum_{i=1}^{c} \sum_{j=1}^{c} L_{ij} P_j \int h_i(\mathbf{x})\, p_j(\mathbf{x})\, d\mathbf{x}$$

$$= \sum_{j=1}^{c} L_{0i} P_i \, \mathsf{E}_i\{h_0(\mathbf{x})\} + \sum_{i=1}^{c} \sum_{j=1}^{c} L_{ij} P_j \, \mathsf{E}_j\{h_i(\mathbf{x})\} ;$$

the error, reject and non-error rates are, respectively:

$$E^* = \sum_{i=1}^{c} P_i \int h_j(\mathbf{x})\, p_i(\mathbf{x})\, d\mathbf{x} = \sum_{i=1}^{c} P_i \, \mathsf{E}_i\{h_j(\mathbf{x})\}, \quad j \neq i ,$$

$$R_r^* = \sum_{i=1}^{c} P_i \int h_0(\mathbf{x})\, p_i(\mathbf{x})\, d(\mathbf{x}) = \sum_{i=1}^{c} P_i \, \mathsf{E}_i\{h_0(\mathbf{x})\} ,$$

$$C^* = \sum_{i=1}^{c} P_i \int h_i(\mathbf{x})\, p_i(\mathbf{x})\, d\mathbf{x} = \sum_{i=1}^{c} P_i \, \mathsf{E}_i\{h_i(\mathbf{x})\} .$$

Corresponding sample means $\sum_j h_0(\mathbf{x}_{ij})/N_i$ and $\sum_j h_i(\mathbf{x}_{ij})/N_i$ respectively, are unbiased and converge to $\mathsf{E}_i\{h_0(\mathbf{x})\}$ and $\mathsf{E}_i\{h_i(\mathbf{x})\}$, $i = 1, \ldots, c$, respectively, with probability one as sample size $N_i \to \infty$.

Hence, for equal prior probabilities, $P_i = 1/c$ and for sample sizes $N_i = N/c$ the mean

$$\tilde{C}(d^*) = \frac{1}{N} \sum_{i=1}^{c} \sum_{i=1}^{N_i} h_i(\mathbf{x}_{ij})$$

$$= \text{proportion of sample observations classified correctly by the rule } d^*$$

is unbiased and converging to optimal non-error rate $C(d^*)$. Similarly, then mean

$$\tilde{R}_r(d^*) = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N_i} h_0(\mathbf{x}_{ij})$$

$$= \text{proportion of sample observations rejected to classify by the rule } d^*.$$

If densities p_i are estimated by $\hat{p}_i$ and a priori probabilities P_i by $\hat{P}_i$, $i = 1, \ldots, c$ and the optimal decision rule d^* is estimated by $\hat{d}$ derived from the density and

a priori probability estimates, then the corresponding indicators are

$$\hat{h}_i(\mathbf{x}) = \begin{cases} 1 & \text{if } \hat{r}_i(\mathbf{x}) = \min\limits_{j=0,\ldots,c} \hat{r}_j(\mathbf{x}), \quad i = 0, \ldots, c \\ 0 & \text{otherwise}, \end{cases}$$

where $\hat{r}_i(\mathbf{x})$, $i = 0, \ldots, c$ is as in (11).

Hence, for $\hat{P}_i = N_i/N$, $i = 1, \ldots, c$ the non-error rate estimator called *sample non-error rate* or *sample success proportion* becomes

$$\tilde{C}(\hat{d}) = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N_i} \hat{h}_i(\mathbf{x}_{ij}) \tag{14}$$

= proportion of sample observations classified correctly by rule $\hat{d}$.

The reject rate estimators called *sample reject rate* becomes

$$\tilde{R}_r(\hat{d}) = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N_i} \hat{h}_0(\mathbf{x}_{ij}) \tag{15}$$

= proportion of sample observations rejected to be classified by the rule $\hat{d}$.

It is easy to verify that if the density estimates $\hat{p}_i$ were determined from data $\mathbf{x}'_{ij}$, which were independent of the $\mathbf{x}_{ij}$, $j = 1, \ldots, N_i$, $i = 1, \ldots, c$ explicit in the expression for $\tilde{R}_r(\hat{d})$ and $\tilde{C}(\hat{d})$, sample reject rate $\tilde{R}_r(\hat{d})$ would be an unbiased estimator of $R_r(\hat{d})$, the actual reject rate, and sample-error rate $\tilde{C}(\hat{d})$ would be an unbiased estimator of $C(\hat{d})$, the actual non-error rate.

6.1.3 Sample-based transformed classification procedure

Let us now derive the sample-based counterpart of the decision rule defined by (2.69) in Section 2.6.

The derivation starts from determination of the estimator $\mathbf{w}^\wedge$ of the discriminative direction $\mathbf{w}$ based on the training sets $\mathbb{T}_i$ defined by (1). The a priori probability P_i, is either known or can be estimated from the training set $\mathbb{T}_i$ by $\hat{P}_i = N_i/N$, $i = 1, \ldots, c$. Now all m-dimensional elements of the training set $\mathbb{T}_{(N)}$ are projected on the line with the direction $\mathbf{w}^\wedge$. The transformed training set we denote by $Y_{(N)} = \{Y_1, \ldots, Y_c\}$, where

$$Y_i = \{y_{ij} \in \mathbb{Y}: y_{ij} = \mathbf{x}_{ij}^{\mathrm{T}} \mathbf{w}^\wedge, j = 1, 2, \ldots, N_i\}, \quad i = 1, \ldots, c; \ \mathbb{Y} \subseteq \mathbb{R}.$$

On the basis of elements of the transformed training set Y_i, the density $p_i(y, \mathbf{w}^\wedge)$ of the transformed observation y is estimated by $\hat{p}_i(y, \mathbf{w}^\wedge)$ and the estimator of the common transformed density $p(y, \mathbf{w}^\wedge)$ is $\hat{p}(y, \mathbf{w}^\wedge) = \sum_{i=1}^{c} \hat{P}_i \hat{p}_i(y, \mathbf{w}^\wedge)$.

The sample-based decision rule $\hat{d}_w$ derived from density estimators based on the transformed training sets Y_i may be definite as follows:

Assign the pattern $\mathbf{x}$ to the class ω_i, i.e.

$$\hat{d}_w(\mathbf{x}) = d_i \quad \text{if} \quad y = \mathbf{x}^T\mathbf{w}^\wedge \in \Omega_i(\mathbf{w}^\wedge), \quad i = 1, \ldots, c; \tag{16a}$$

reject the pattern to classify, i.e.

$$\hat{d}_w(\mathbf{x}) = d_0 \quad \text{if} \quad y = \mathbf{x}^T\mathbf{w}^\wedge \in \Omega_0(\mathbf{w}^\wedge), \tag{16b}$$

where

$$\Omega_i(\mathbf{w}^\wedge) = \{y \in \mathbb{Y}: \hat{r}_i(y) = \min_{j=0,\ldots,c} \hat{r}_j(y)\}, \quad i = 0, \ldots, c$$

and

$$\hat{r}_i(y) = \sum_{j=1}^{c} L_{ij}\hat{P}_j \hat{p}_j(y, \mathbf{w}^\wedge)/\hat{p}(y, \mathbf{w}^\wedge)$$

is the estimator of the conditional risk for the transformed data $y = \mathbf{x}^T\mathbf{w}^\wedge$.

By using estimators presented in the preceding sections, the resulting apparent non-error rate and the resulting apparent reject rate are, respectively

$$\hat{C}(\hat{d}_w) = \sum_{i=1}^{c} \hat{P}_i \int_{\hat{\Omega}_i} \hat{p}_i(y, \mathbf{w}^\wedge)\, dy, \quad \hat{\Omega}_i \triangleq \hat{\Omega}_i(\mathbf{w}^\wedge) \tag{17}$$

$$\hat{R}_r(\hat{d}_w) = \sum_{i=1}^{c} \hat{P}_i \int_{\hat{\Omega}_0} \hat{p}_i(y, \mathbf{w}^\wedge)\, dy, \quad \hat{\Omega}_0 \triangleq \hat{\Omega}_0(\mathbf{w}^\wedge). \tag{18}$$

Let n_{ji} denotes the number of observations from the ith training set $\mathbb{T}_i$ correctly $(i = j)$ or incorrectly $(i \neq j)$ classified using decision rule $\hat{d}_w$ defined in (16) and n_{0i} the number of observations from $\mathbb{T}_i$ rejected to be classified by $\hat{d}_w$. Let P_i be estimated by $\hat{P}_i = N_i/N$. Then the resulting sample non-error rate, resulting sample error rate and resulting sample reject rate are, respectively

$$\tilde{C}(\hat{d}_w) = \sum_{i=1}^{c} \frac{N_i}{N}\frac{n_{ii}}{N_i} = \frac{1}{N}\sum_{i=1}^{c} n_{ii}, \tag{19}$$

$$\tilde{E}(\hat{d}_w) = \sum_{i=1}^{c} \frac{N_i}{N}\frac{n_{ji}}{N_i} = \frac{1}{N}\sum_{i=1}^{c} n_{ji}, \quad j \neq i \tag{20}$$

and

$$\tilde{R}_r(\hat{d}_w) = \sum_{i=1}^{c} \frac{N_i}{N}\frac{n_{0i}}{N_i} = \frac{1}{N}\sum_{i=1}^{c} n_{0i}. \tag{21}$$

6.2 Sample-based feature evaluation criteria

In Chapter 3 we have been only mentioned two basic categories of feature evaluation criteria functions and therefore here we do not intend to develop an exhaustive discussion about that how some of the concepts and methods encountered in preceding section in analysis of error estimation can be used to estimate these feature evaluation criterion functions (see e.g. Devijver and Kittler [14]).

The criteria of the first categories are based on the notion of interclass distance and can be formulated using only the parameters $\boldsymbol{\mu}$, $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$, $\mathbf{S}_W$, $\mathbf{S}_B$ defined in (2.51),

(2.53), (2.54) and (2.52), respectively. In practical applications of these criteria, the unknown parameters would replace by their estimates based on the training set (1).

A number of probabilistic distance criteria can be analytically simplified in the case when the classes are normally distributed [14]. In this case these criteria may be expressed in closed forms involving the mean vectors and covariance matrices of the class-conditional, multivariate normal distributions. Therefore, substitution of sample mean vectors and sample covariance matrices provides apparent estimates of these distance criteria. The nonparametric case be handled with very much same tools that we used to estimate the error probability in Section 6.1.

7. OPTIMIZATION OF FEATURE SELECTION AND CLASSIFIER DESIGN IN PREDITAS SYSTEM

7.1 Sample-based K-stepwise linear classification procedure used in PREDITAS system

The idea on which the sample-based classification procedure of the PREDITAS system is based, may be express in the following way.

Suppose we have the training set $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2\}$ of size N, where

$$\mathcal{T}_i = \{\mathbf{x}_{ij} \in \Omega : j = 1, 2, \ldots, N_i\}, \quad i = 1, 2, \quad \sum_{i=1}^{2} N_i = N. \tag{1}$$

Assume a priori probabilities P_i is known or have been estimated by $\hat{P}_i = N_i/N$ and the density $p_i(\mathbf{x})$ is unknown, $i = 1, 2$. Further assume that e_i, $0 < e_i < \frac{1}{2}$ is predetermined value for the observation-conditional error rate (2.2), i.e.

$$1 - P(\omega_i \mid \mathbf{x}) < e_i, \quad i = 1, 2.$$

The derivation of the sample-based classification procedure starts from determination of the estimator $\mathbf{w}^{\wedge}$ of discriminative direction $\mathbf{w}$. This estimator is computed by substitution of parameter estimates for the unknown parameters in (2.60), namely

$$\mathbf{w}^{\wedge} = \mathbf{S}_w^{\wedge -1}(\boldsymbol{\mu}_1^{\wedge} - \boldsymbol{\mu}_2^{\wedge}), \tag{2}$$

where

$$\mathbf{S}_w^{\wedge} = \sum_{i=1}^{2} \hat{P}_i \boldsymbol{\Sigma}_i^{\wedge} \tag{3}$$

and

$$\boldsymbol{\mu}_i^{\wedge} = \frac{1}{N_i} \sum_{i=1}^{N_i} \mathbf{x}_{ij} \tag{4}$$

$$\boldsymbol{\Sigma}_i^{\wedge} = \frac{1}{N_i - 1} \sum_{i=1}^{N_i} (\mathbf{x}_{ij} - \boldsymbol{\mu}_i^{\wedge})(\mathbf{x}_{ij} - \boldsymbol{\mu}_i^{\wedge})^{\mathrm{T}} \quad i = 1, 2. \tag{5}$$

Now all m-dimensional elements of the training set $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2\}$ are projected on the line with the direction $\mathbf{w}^{\wedge}$ given by (2). The transformed training set we denote

by $Y = \{Y_1, Y_2\}$, where

$$Y_i = \{y_{ij} \in \mathbb{Y}: y_{ij} = \mathbf{x}_{ij}^{\mathrm{T}}\mathbf{w}^{\wedge}\,,\ j = 1, 2, \ldots, N_i\}\,,\quad i = 1, 2\,;\quad \mathbb{Y} \subseteq \mathbb{R}\,.$$

On the basis of elements of the transformed training set Y_i, the density $p_i(y, \mathbf{w}^{\wedge})$ is estimated by $\hat{p}_i(y, \mathbf{w}^{\wedge})$ and the estimator of $p(y, \mathbf{w}^{\wedge})$ is $\hat{p}(y, \mathbf{w}^{\wedge}) = \sum_{i=1}^{2} \hat{P}_i\, \hat{p}_i(y, \mathbf{w}^{\wedge})$.

The sample-based decision rule $\hat{d}_w$ derived from density estimators based on the training sets transformed using the estimator $\mathbf{w}^{\wedge}$ given by (2) of the vector $\mathbf{w}$ may be defined as follows:

Assign the pattern $\mathbf{x}$ to the class ω_i, i.e.

$$\hat{d}_w(\mathbf{x}) = d_i \quad \text{if} \quad y = \mathbf{x}^{\mathrm{T}}\mathbf{w}^{\wedge} \in \Omega_i(\mathbf{w}^{\wedge})\,;\quad i = 1, 2\,; \tag{6a}$$

reject the pattern to classify, i.e.

$$\hat{d}_w(\mathbf{x}) = d_0 \quad \text{if} \quad y = \mathbf{x}^{\mathrm{T}}\mathbf{w}^{\wedge} \in \Omega_0(\mathbf{w}^{\wedge})\,, \tag{6b}$$

where

$$\Omega_1(\mathbf{w}^{\wedge}) = \{y \in \mathbb{Y}: \hat{l}(y, \mathbf{w}^{\wedge}) > c_1\} \tag{7a}$$

$$\Omega_2(\mathbf{w}^{\wedge}) = \{y \in \mathbb{Y}: \hat{l}(y, \mathbf{w}^{\wedge}) < c_2\} \tag{7b}$$

$$\Omega_0(\mathbf{w}^{\wedge}) = \{y \in \mathbb{Y}: c_2 \leqq \hat{l}(y, \mathbf{w}^{\wedge}) \leqq c_1\} \tag{7c}$$

with

$$\hat{l}(y, \mathbf{w}^{\wedge}) \triangleq \hat{P}_1\, \hat{p}_1(y, \mathbf{w}^{\wedge}) / \hat{P}_2\, \hat{p}_2(y, \mathbf{w}^{\wedge}) \tag{8}$$

and

$$c_1 = \frac{1 - e_1}{e_1}\,,\quad c_2 = \frac{e_2}{1 - e_2}\,. \tag{9}$$

The decision rule (6) is the transformed decision rule with predetermined probabilities considered in Section 2.4.2.

All the observations from the training set $\mathbb{T}$ we classify by using the rule $\hat{d}_w$ defined in (6) and determine the non-error rate, error rate and reject rate estimators, respectively, by using the same training set $\mathbb{T}$. Let n_{ji} denote the number of observations from the ith training set $\mathbb{T}_i$ correctly $(i = j)$ or incorrectly $(i \neq j)$ classified using the decision rule $\hat{d}_w$ and n_{0i} the number of observations from $\mathbb{T}_i$ rejected to be classified by the rule $\hat{d}_w$. Then from (6.19) to (6.21) it follows that the resulting sample non-error rate, resulting sample error rate and resulting sample reject rate are, respectively

$$\tilde{C}(\hat{d}_w^{(1)}) = \sum_{i=1}^{2} \hat{P}_i \frac{n_{ii}}{N_i}\,,$$

$$\tilde{E}(\hat{d}_w^{(1)}) = \sum_{i=1}^{2} \hat{P}_i \frac{n_{ji}}{N_i}\,,$$

$$\tilde{R}_r(\hat{d}_w^{(1)}) = \sum_{i=1}^{2} \hat{P}_i \frac{n_{0i}}{N_i}\,,$$

or, if P_i is estimated by $\hat{P}_i = N_i/N$, $i = 1, 2$

$$\tilde{C}(\hat{d}_w) = \sum_{i=1}^{2} \frac{N_i}{N} \frac{n_{ii}}{N_i} = \frac{1}{N} \sum_{i=1}^{2} n_{ii}, \tag{10}$$

$$\tilde{E}(\hat{d}_w) = \sum_{i=1}^{2} \frac{N_i}{N} \frac{n_{ji}}{N_i} = \frac{1}{N} \sum_{i=1}^{2} n_{ji}, \quad j \neq i \tag{11}$$

and

$$\tilde{R}_r(\hat{d}_w) = \sum_{i=1}^{2} \frac{N_i}{N} \frac{n_{0i}}{N_i} = \frac{1}{N} \sum_{i=1}^{2} n_{0i}. \tag{12}$$

Now, we discard from the training set $\mathbb{T}^{(1)} \triangleq \mathbb{T}$ all the samples that have been classified using the decision rule $\hat{d}_w^{(1)} \triangleq \hat{d}_w$. Let $\mathbb{T}^{(2)} = \{\mathbb{T}_1^{(2)}, \mathbb{T}_2^{(2)}\}$ denote the set of remaining observations. $\mathbb{T}_1^{(2)}$ and $\mathbb{T}_2^{(2)}$ are subsets of the set $\mathbb{T}^{(1)}$ of sizes $N_1^{(2)}$, $N_2^{(2)}$, $\sum_{i=1}^{2} N_i^{(2)} = N^{(2)}$. Note that $N^{(1)} = N - N^{(2)}$ is the number of observations from the set $\mathbb{T}$ classified by $\hat{d}_w^{(1)}$; $N^{(1)} = N_1^{(1)} + N_2^{(1)} = \sum_{i=1}^{2} (n_{1i} + n_{2i})$.

We derive the decision rule $\hat{d}_w^{(2)}$ on the basis of the training set $\mathbb{T}^{(2)}$:

$$\hat{d}_w^{(2)}(\mathbf{x}) = \begin{cases} d_1 & \text{if} \quad y_2 = \mathbf{x}^{\mathrm{T}}\mathbf{w}_2^{\wedge} \in \Omega_1(\mathbf{w}_2^{\wedge}) \\ d_2 & \text{if} \quad y_2 = \mathbf{x}^{\mathrm{T}}\mathbf{w}_2^{\wedge} \in \Omega_2(\mathbf{w}_2^{\wedge}) \\ d_0 & \text{if} \quad y_2 = \mathbf{x}^{\mathrm{T}}\mathbf{w}_2^{\wedge} \in \Omega_0(\mathbf{w}_2^{\wedge}), \end{cases} \tag{13}$$

where

$$\Omega_1(\mathbf{w}_2^{\wedge}) = \{y \in \mathbb{Y}: \hat{l}(y, \mathbf{w}_2^{\wedge}) > c_1\}$$

$$\Omega_2(\mathbf{w}_2^{\wedge}) = \{y \in \mathbb{Y}: \hat{l}(y, \mathbf{w}_2^{\wedge}) < c_2\}$$

$$\Omega_0(\mathbf{w}_2^{\wedge}) = \{y \in \mathbb{Y}: c_2 \leqq \hat{l}(y, \mathbf{w}_2^{\wedge}) \leqq c_1\}$$

and $\hat{l}(y, \mathbf{w}_2^{\wedge})$ is the value of $\hat{l}(y, \mathbf{w}^{\wedge})$ given in (8) for $\mathbf{w}^{\wedge} = \mathbf{w}_2^{\wedge}$; c_1 and c_2 are as in (9). Now we classify the samples from the set $\mathbb{T}^{(2)}$ by using the decision rule $\hat{d}_w^{(2)}$ defined in (13). Let $n_{ji}^{(2)}$ be the number of observations from $\mathbb{T}_i^{(2)}$ correctly $(i = j)$ or incorrectly $(i \neq j)$ classified by the rule $\hat{d}_w^{(2)}$ and $n_{0i}^{(2)}$ is the number of samples from $\mathbb{T}_i^{(2)}$ rejected to be classified. Then the non-error rate, error rate and reject rate estimators are, respectively,

$$\tilde{C}(\hat{d}_w^{(2)}) = \sum_{i=1}^{2} \frac{N^{(2)}}{N} \hat{P}_i \frac{n_{ii}^{(2)}}{N_i^{(2)}}, \tag{14}$$

$$\tilde{E}(\hat{d}_w^{(2)}) = \sum_{i=1}^{2} \frac{N^{(2)}}{N} \hat{P}_i \frac{n_{ji}^{(2)}}{N_i^{(2)}}, \tag{15}$$

$$\tilde{R}_r(\hat{d}_w^{(2)}) = \sum_{i=1}^{2} \frac{N^{(2)}}{N} \hat{P}_i \frac{n_{0i}^{(2)}}{N_i^{(2)}}. \tag{16}$$

The problem of classification the new pattern $\mathbf{x}$ is then solved by using the decision

rule $\hat{d}_w$ defined as

$$\hat{d}_w(\mathbf{x}) = \begin{cases} d_1 & \text{if } y_1 \in \Omega_1(\mathbf{w}_1^\wedge) \text{ or } y_1 \in \Omega_0(\mathbf{w}_1^\wedge) \wedge y_2 \in \Omega_1(\mathbf{w}_2^\wedge) \\ d_2 & \text{if } y_1 \in \Omega_2(\mathbf{w}_1^\wedge) \text{ or } y_1 \in \Omega_0(\mathbf{w}_1^\wedge) \wedge y_2 \in \Omega_2(\mathbf{w}_2^\wedge) \\ d_0 & \text{if } y_2 \in \Omega_0(\mathbf{w}_2^\wedge), \end{cases} \tag{17}$$

where $y_1 = \mathbf{x}^T \mathbf{w}_1^\wedge$.

The corresponding non-error rate and reject rate estimators are, respectively,

$$\hat{C}(\hat{d}_w) = \sum_{k=1}^{2} \frac{N^{(k)}}{N} \sum_{i=1}^{2} \hat{P}_i \frac{n_{ii}^{(k)}}{N_i^{(k)}}, \tag{18}$$

where $n_{ii}^{(1)} \triangleq n_{ii}$, $i = 1, 2$,

$$\hat{R}_r(\hat{d}_w) = \frac{N^{(2)}}{N} \sum_{i=1}^{2} \hat{P}_i \frac{n_{0i}^{(2)}}{N_i^{(2)}}. \tag{19}$$

If the relation

$$\hat{R}_r(\hat{d}_w) < \hat{R}_r(\hat{d}_w^{(1)}), \tag{20}$$

is satisfied, then we choice the decision rule $\hat{d}_w$ defined in (17) to solve the problem of classifying the new pattern to one of two classes.

The sample-based classification rule in the PREDITAS system is constructed by using a sequence of K sample-based decision rules $\{\hat{d}_w^{(k)}\}_{k=1}^K$, where each $\hat{d}_w^{(k)}$ is based on its own training set $\mathbb{T}^{(k)} = \{\mathbb{T}_1^{(k)}, \mathbb{T}_2^{(k)}\}$ of size $N^{(k)} = N_1^{(k)} + N_2^{(k)}$.

The description of the steps in our procedure for the design of the kth sample-based classification rule, $k = 1, 2, \ldots, K$ is as follows.

Step 1: Compute the vector $\mathbf{w}^{\wedge(k)} \in \mathbb{R}^m$ by (2), where $\boldsymbol{\mu}_i^\wedge$ and $\boldsymbol{\Sigma}_i^\wedge$ are sample mean vector and sample covariance matrix, namely

$$\boldsymbol{\mu}_i^\wedge = \frac{1}{N_i^{(k)}} \sum_{j=1}^{N_i^{(k)}} \mathbf{x}_{ij}$$

$$\boldsymbol{\Sigma}_i^\wedge = \frac{1}{N_i^{(k)} - 1} \sum_{j=1}^{N_i^{(k)}} (\mathbf{x}_{ij} - \boldsymbol{\mu}_i^\wedge)(\mathbf{x}_{ij} - \boldsymbol{\mu}_i^\wedge)^T.$$

Step 2: Project all observations from the training set $\mathbb{T}^{(k)}$ on the line with direction $\mathbf{w}^{\wedge(k)}$ according to

$$y_{ij}^{(k)} = \mathbf{x}_{ij}^T \mathbf{w}^{\wedge(k)}, \quad i = 1, 2, \quad j = 1, 2, \ldots, N_i^{(k)}.$$

Denote the transformed training set by $Y^{(k)}$.

Step 3: Obtain the estimate $\hat{p}_i(y, \mathbf{w}^{\wedge(k)})$ of the transformed conditional density function $p_i(y, \mathbf{w}^{\wedge(k)})$, $i = 1, 2$ by histogram technique

$$\hat{p}_i(y, \mathbf{w}^{\wedge(k)}) = \sum_{j=1}^{H} \Phi_h(y) \frac{n_i(h)}{N_i},$$

where $\Phi_{ih}(y)$ is the indicator function for the hth interval of a partition of

the range of y into intervals $I_1, \ldots, I_H$ (i.e. $\Phi_h(y) = 1$ if $y \in I_h$ and $\Phi_h(y) = 0$ otherwise, $h = 1, 2, \ldots, H$); $n_i(h)$ is the number of transformed observations from training set $Y^{(k)}$ that are in the interval I_h.

Step 4: Determine the points $w_1^{(k)}$ and $w_2^{(k)}$, $w_1^{(k)} < w_2^{(k)}$, $w_i^{(k)} \in \mathbb{R}$, $i = 1, 2$ on the line with the direction $\mathbf{w}^{\wedge(k)}$ in the following way:

for all $y \in Y^{(k)} \wedge y \leqq w_1^{(k)}$ it holds

$$\hat{l}(y, \mathbf{w}^{\wedge(k)}) > c_1 \tag{21}$$

or

for all $y \in Y^{(k)} \wedge y \leqq w_1^{(k)}$ it holds

$$\hat{l}(y, \mathbf{w}^{\wedge(k)}) < c_2 \tag{22}$$

and

for all $y \in Y^{(k)} \wedge y \geqq w_2^{(k)}$ it holds (21)

or

for all $y \in Y^{(k)} \wedge y \geqq w_2^{(k)}$ it holds (22),

where $\hat{l}(y, \mathbf{w}^{\wedge(k)})$ is defined by (8) for $\mathbf{w}^{\wedge} = \mathbf{w}^{\wedge(k)}$ and the constants c_1 and c_2 are determined by (9).

Step 5: Classify the samples from $\mathbb{T}^{(k)}$ using the following rule $\hat{d}_w^{(k)}$:

Assign $\mathbf{x}$ to the class $\alpha_1^{(k)}$, if $y^{(k)} = \mathbf{x}^{\mathrm{T}}\mathbf{w}^{\wedge(k)}$ belongs to $\mathbb{R}_{\alpha_1^{(k)}} = \{y \in \mathbb{Y}: y < w_1^{(k)}\}$;

assign $\mathbf{x}$ to the class $\alpha_2^{(k)}$, if $y^{(k)} = \mathbf{x}^{\mathrm{T}}\mathbf{w}^{\wedge(k)}$ belongs to $\mathbb{R}_{\alpha_2^{(k)}} = \{y \in \mathbb{Y}: y > w_2^{(k)}\}$, where

$$\alpha_i^{(k)} = \begin{cases} \omega_1 & \text{if (21) holds} \\ \omega_2 & \text{if (22) holds} \end{cases}, \quad i = 1, 2$$

reject to classify $\mathbf{x}$ if $y^{(k)} = \mathbf{x}^{\mathrm{T}}\mathbf{w}^{\wedge(k)}$ belongs to

$$\mathbb{R}_{\alpha_0^{(k)}} = \{y \in \mathbb{Y}: w_1^{(k)} \leqq y \leqq w_2^{(k)}\}.$$

Step 6: Discard from $\mathbb{T}^{(k)}$ all the samples that have been classified using the decision rule $\hat{d}_w^{(k)}$. Denote by $\mathbb{T}^{(k+1)}$ the set of remaining samples.

Step 7: Repeat the procedure for the training subset $\mathbb{T}^{(k+1)}$.

The above described process of generating the sequence of the decision rules starts from the training set $\mathbb{T}$ and continues until all the elements of the original training set $\mathbb{T}$ are classified or any of the terminating conditions are fulfilled.

The problem of classifying a new pattern $\mathbf{x}$ is solved by using the above defined sequence of the respective decision rules as follows.

Assign $\mathbf{x}$ to the class $\alpha_i^{(j)}$

$$\text{if} \quad y^{(k)} \in \mathbb{R}_{\alpha_0^{(k)}} \wedge y^{(j)} \in \mathbb{R}_{\alpha_i^{(j)}} \tag{23a}$$

for all $k, 1 \leqq k < j \leqq K$, $i = 1, 2$

Reject to classify $\mathbf{x}$ if $y^{(k)} \in \mathbb{R}_{x_0^{(k)}}$ (23b)

for all $1 \leqq k \leqq K$.

The non-error and reject rate estimators of the decision rule (23) are, respectively according to (18) and (19)

$$\tilde{C} = \sum_{k=1}^{K} \frac{N^{(k)}}{N} \sum_{i=1}^{2} \hat{P}_i \frac{n_{ii}^{(k)}}{N_i^{(k)}} \tag{24}$$

and

$$\tilde{R}_r = \frac{N^{(K)}}{N} \sum_{i=1}^{2} \hat{P}_i \frac{n_{0i}^{(K)}}{N_i^{(K)}}, \tag{25}$$

where $n_{ii}^{(k)}$ is the number of samples from training set $\mathcal{T}_i^{(k)}$ classified correctly by the rule $\hat{d}_w^{(k)}$ and $n_{0i}^{(K)}$ is the number of samples from $\mathcal{T}_i^{(K)}$ rejected to be classified.

7.2 Determination of optimal dimensionality

In practical applications when information about the classes ω_i, $i = 1, 2$, is available only in the form of training sets (1), the feature selection criterion (4.1) for m-dimensional vector $\mathbf{x}$ should be replace by its estimates based on the training set $\mathcal{T}$ which it will be called *sample-based measure of discriminative power*

$$\hat{\lambda}_m(\mathbf{x}) = \operatorname{tr}\left(\mathbf{S}_W^{\wedge -1} \mathbf{S}_B^{\wedge}\right), \quad m \leqq M, \tag{26}$$

where

$$\mathbf{S}_B^{\wedge} = \hat{P}_1 \hat{P}_2 (\boldsymbol{\mu}_1^{\wedge} - \boldsymbol{\mu}_2^{\wedge})(\boldsymbol{\mu}_1^{\wedge} - \boldsymbol{\mu}_2^{\wedge})^{\mathrm{T}}$$

and

$$\mathbf{S}_W^{\wedge} = \sum_{i=1}^{2} \hat{P}_i \boldsymbol{\Sigma}_i^{\wedge}$$

with $\boldsymbol{\mu}_i^{\wedge}$ and $\boldsymbol{\Sigma}_i^{\wedge}$ determined by (4) and (5), respectively.

Let us assume that the problem of feature selection has been already solved. It means that by means of a search procedure and the sample-based MDP (26) we have found for any dimensionality an optimal (or near optimal to be more exact) subset of features.

However, yet another difficult problem remains to be solved – that of determining an optimal dimensionality of the decision problem, it means an optimal number of features.

The difficulty of this problem results from several facts. First of all, even the concept of optimal dimensionality itself needs to be clarified since the optimality can be assessed from different viewpoints, like the probability of error or the reliability of given classification rule.

Another reason is that the feature selection criterion has no direct relationship to the probability of classification error. Only the relationship between MDP and dimensionality is monotonous. This could seem to be sufficient for choosing the right dimensionality however, another problem arises here.

Since feature selection is carried out on the training set, the size of which is limited, and the performance of the classifier should be optimized with respect to samples outside the training set, in practice even the above mentioned monotonicity is not fulfilled.

It has been proved both in practice and by theoretical results [21, 24] that for a given finite size training set an optimal dimensionality (optimal number of features) exists and that any further increase of the number of features deteriorates the overall performance on independent patterns, though the value of criterion function on the training set increases.

All these facts result in impossibility to find an exact and analytically justified procedure by means of which we could arrive to the optimal dimensionality.

However, a heuristic procedure can be utilized to get the required results. Let us assume that for a set of m features, $x_1, \ldots, x_m$, ordered according to their significance $S_{(1)}(x_i)$, defined by (4.2) for $k = 1$, resulting from the search selection algorithm, the following relations hold:

$$k_1 \hat{\lambda}_M \leqq \sum_{i=1}^{m} S_{(1)}(x_i) \leqq k_2 \hat{\lambda}_M \,, \tag{27}$$

where $\hat{\lambda}_M$ is the estimate of the MDP for the whole set of features, k_1, k_2 are constants from the interval $(0, 1\rangle$.

Determination of values k_1, k_2 will be discussed after a while, for the moment being let us assign e.g. the values $k_1 = 0{\cdot}3$ and $k_2 = 0{\cdot}95$, in order to provide an interpretation of the formula (27). Then denoting m_1 the minimum number of features satisfying the first inequality and m_2 the maximum number of features satisfying the second inequality, we can state that the relation (27) is satisfied for any $m \in \langle m_1, m_2\rangle$, where m_1 can be interpreted as a minimum possible dimensionality and m_2 as a maximum needed dimensionality.

In other words, increasing the dimensionality above m_2 can increase the sample-based MDP by not more than $(1 - k_2) \,.\, 100\%$ (i.e. in our case by not more than 5%). On the other hand decreasing the dimensionality below m_1 would result in the decrease of MDP by more than $(1 - k_1) \,.\, 100\%$ (i.e. in our case by more than 70%) compared with the value $\hat{\lambda}_M$ for the original dimensionality M.

All the dimensionalities $m \in \langle m_1, m_2\rangle$ are then to be investigated with respect to the performance of the corresponding decision rules and the one optimizing the performance on independent testing set is regarded as the *optimal dimensionality.*

Now the problem of determining the values of k_i, $i = 1, 2$ remains to be discussed. It would be undoubtedly very convenient to be able to introduce some fixed values for k_i, $i = 1, 2$ analogously e.g. to the levels of significance $\alpha = 0{\cdot}05$ etc.

Unfortunately, such an analogy cannot be made in this case. Determination of "reasonable" upper and lower limits for the sample-based MDP and from it following bounds for dimensionalities to be investigated is strongly dependent on data, more specifically on the form of the sample-based MDP as a function of dimensionality.

Furthermore, the absolute value of the sample-based MDP for the full dimensionality, equal $\hat{\lambda}_M$, plays an important role, too. If we have a problem where the original set of features carries a highly informative content from the viewpoint of discrimination, we can obviously afford to loose more from the value of $\hat{\lambda}_M$, i.e. the value of k_2 can be chosen lower than in the case of a problem where the features carry far less useful information and thus $\hat{\lambda}_M$ is lower than in the former case. In the latter case we cannot afford to loose much of $\hat{\lambda}_M$ and thus the value of k_2 closer to 1·0 would be more appropriate.

At the bottom end of the investigated dimensionalities range analogous considerations can be made.

The best "practical" way of selecting the optimal dimensionality is to make the decisions on the basis of graphical form of the sample-based MDP as a function of dimensionality. If this curve from certain dimensionality upwards inhibits a typical "plateau", the value of m_2 can be set equal to the dimensionality corresponding to the beginning of the plateau. At the bottom end we should look for a sudden beginning of a steep decline of the curve with deceasing dimensionality and to set the value of m_1 equal to the dimensionality corresponding to the beginning of the steep decline.

The final phase of the process of determining the optimal dimensionality is to test the performance of all the decision rules corresponding to dimensionalities $m \in \langle m_1, m_2 \rangle$ on independent testing set of patterns and to choose as the optimal dimensionality that one, for which the corresponding decision rule yields the minimum classification error.

Though the described procedure represents a combination of analysis of numerical results and heuristic approach, it has proved to be quite efficient and satisfactory in practice.

8. ARCHITECTURE AND USAGE OF PREDITAS SOFTWARE

The PREDITAS system consists of four problem independent programs used in succession, where the results of one program are utilized in the following ones:

(i) DATANAL

is a program which performs a basic statistical analysis of the input data set. The values found for each feature and separately for each class and the whole set are: minimum and maximum values, range, average value and standard deviation, distribution histograms for non binary features and for binary features x_j probabilities $P(\omega_i \mid x_j = 1)$ $i = 1, 2$, $j = 1, \ldots, m$ and $P(x_j = 1)$ which are estimated by the rates

$$\hat{P}(\omega_1 \mid x_j = 1) = \frac{N_{1j}^+}{N_{1j}^+ + rN_{2j}^+}, \quad r = \frac{N_1}{N_2} \tag{1}$$

$$\hat{P}(\omega_2 \mid x_j = 0) = \frac{N_{2j}^-}{N_{1j}^- + rN_{2j}^-}$$

$$\hat{P}(x_j = 1) = \frac{N_{1j}^+ + N_{2j}^+}{2N_1}$$

where m is dimensionality of the pattern vector $\mathbf{x}$, N_{ij}^+, $i = 1, 2$ are numbers of "true" values of jth feature in class ω_i and N_{ij}^- are numbers of "false" values of jth feature in classes ω_i. N_i are numbers of elements in classes ω_i and it holds evidently $N_i = N_{ij}^+ + N_{ij}^-$. Both estimations of probabilities are computed under assumption that a priori probabilities of classes are equal. The last computed value for each feature is its significance, considered independently on other features and evaluated by a number from 0 to 1:

$$S_j = \tfrac{1}{2} \int_{-\infty}^{+\infty} |p_{\omega_1}(x_j) - p_{\omega_2}(x_j)| \, \mathrm{d}x_j \triangleq \tfrac{1}{2} \sum_{k=1}^{K_i} |\hat{p}_{\omega_k}(x_j) - \hat{p}_{\omega_k}(x_j)| \tag{2}$$

where $p_{\omega_i}(x_j)$, $i = 1, 2$ is probability density in class ω_i and $\hat{p}_{\omega_i}(x_j)$ its estimate by histogram for non binary features. For binary features the estimate

$$\hat{p}_{\omega_i}(x_j) = \frac{N_{ij}^+}{N_i} \tag{3}$$

is used.

The values of the feature ranges are used for elimination of those features the value of which is constant in the whole training set. Estimates $\hat{P}(\omega_1 \mid x_j = 1)$ and $\hat{P}(\omega_2 \mid x_j = 0)$ will be used for finding of such binary features, which alone can be considered as sufficient for classification because some of their two possible values, 1 or 0, occurs only in one of the classes. The value of at least one of both probable estimates will be in that case equal to 1 or 0.

(ii) **MEDIP**

is a program for the significance analysis of the features, using the measure of discriminative power (MDP) for its evaluation with the option of choosing from four selection methods. A detailed description of MDP criterion and four selection methods can be found in Section 4.1 together with comparison of their priorities and drawbacks. Briefly recalling, we can state that the straight methods either include or exclude at each step just one feature, which is the most or the least significant at this step, for the including and excluding method respectively. These methods are only step optimal since they do not have any possibility to go back and make corrections.

More sophisticated "floating" methods enable to return back and search even among already included or excluded features in the case of including or excluding method respectively. This results in possibility to correct the composition of subsets found in previous sets so as to approach closely the optimum subset. However, these methods are obviously more time-consuming and it is only fair to state that in most

practical problems, we have solved till now, only a little difference in significance analysis has been found. It does not mean at all that these methods should be neglected, they should be given priorities only in those cases where there usage is justified by the improved performance of the classifier.

As a result of the significance analysis we get a table describing the course of successive feature including or excluding, containing the values of MDP for the selected feature subsets, relative decrement to the maximum value of MDP in % and finally the significance of a given feature both in the subset and independently on other features. A graphical form of the significance analysis is available too. Another option enables to print out beside the table of significance analysis also the eigenvectors for all the steps of the selection process.

(iii) **SEDIF**

is a program for the stepwise decision rule construction using the most significant features selected according to the order of significance which was computed by MEDIP program. The decision rule construction is described in Section 7.1 as Step 1–6. For simplifying the program procedure, the decision rule $\hat{d}_w^{(k)}$ (described in the Step 5) is divided to two rules $\hat{d}_w^{(m)}$ and $\hat{d}_w^{(m+1)}$, where $m + 1 \leqq 2k$, in the following way:

$$\mathbf{w}^{\wedge(m)} = \mathbf{w}^{\wedge(k)}, \quad w_0^{(m)} = w_1^{(k)} \tag{4a}$$

$$\mathbf{w}^{\wedge(m+1)} = -\mathbf{w}^{\wedge(k)}, \quad w_0^{(m+1)} = -w_2^{(k)} \tag{4b}$$

If during the decision rule construction only one threshold value $w_1^{(k)}$ is determined, the equation (4b) is not applied. The rule specified in Step 5 is consequently modified as follows:

Assign $\mathbf{x}$ to the class $\alpha^{(m)}$ if $y^{(m)} = \mathbf{x}^{\mathrm{T}}\mathbf{w}^{\wedge(m)}$
belongs to $\mathbb{R}_{\alpha(m)} = \{y \in \mathbb{Y}: y^{(m)} > w_0^{(m)}\}$.

The results of this program are:
- histogram technique based estimate of ith class density projected on the line with optimal discriminative direction,
- estimate of error probability for each step,
- list of classified elements in each step.

Summary table specifies at each step the following characteristics:
- name of the assigned class (to which the elements are classified),
- estimate of error probability,
- updated estimate of a priori probability,
- error probability summarized over all previous steps,
- remaining part of the training set,
- risk index.

All these characteristics are specified both for each class separately and for the whole set.

(iv) **CLASSIF**

is a program for application of the stepwise decision rule produced by program SEDIF. This program exists in several variants both for main frame and personal computers and is adapted for classifying individual elements as well as the whole set e.g. in the case of testing decision rule using independent testing set. As a result of this program we get the decision about the class of classified element, the estimate of error probability for the element being classified its projection into the frequency histogram and the option to demonstrate the influence of any feature. The decision can be rejected if the probability of error is greater then the permitted value or if a feature value is out of the feature range.

9. EXAMPLE OF COMPLETE SOLUTION OF DIAGNOSTIC PROBLEM

One of the practical problems, which has been solved using the PREDITAS software package was the differential diagnosis of cholecystoalithiasis. Prior to a patient's operation it is very desirable to recognize the simple form of the disease from the form with complicating changes on the gull-duct which, if not removed during intervention, sooner or later enforce the reoperation. In the second case the intervention is especially in the small hospitals passed to more experienced surgeon and an instrumental preoperational examination of gull-duct is counted with. From the medical point of view this problem has been described in Keclík[1] and thus only methodical aspects are discussed here.

The problem is to classify patients before the operation into one of two classes: ω_1 – a complicated form of disease and ω_2 – a simple one. Each patient is characterized by 28 symptoms (= features), most of them are of the binary type (values 0 or 1) and the remaining of the real type. After the data analysis using DATANAL program, in several cases a group of binary features can be joined into one feature with ordinary type value. By means of natural numbers different quality levels of a new feature are designed, for example feature No. 3 in Table 1 describing different types of attack. The values and qualities cannot be joined arbitrarily but only in such a way that with the increased feature value the probability of one class increases too.

Among the features there was one (jaundice without attack) which was positive only at 4·9% of patients, but all of them belong into class ω_1 (= complicated form). Such a feature, if it is positive, can be considered as sufficient for the classification. Since for any element having this property the classification problem actually does not exist, none such have been included into the training set.

[1] M. Keclík, S. Bláha and A. Huslarová: Attempted preoperative differential diagnosis of simple and complicated cholecystolithiasis by means of a computer (in Czech). Čs. gastroenterologie a výživa *40* (1986), 5, 214–224.

Table 1. Symptoms for recognition of complicated and simple forms of cholecystolithiasis. The order is from the most significant symptom to the least significant one.

Order	Feature
1	Max. width of gull-duct (in mm)
2	Max. caught ALP (in the multiples of normal value)
3	Course of attacks (0 = without fever and shakes, 1 = sometimes fever, 2 = always fever, 3 = sometimes fever, and shake, 4 = always fever and sometimes shake, 5 = always fever and shake)
4	Stones (0, 1, 2 = undisplayed, suspection, displayed)
5	Pankreatitis (0, 1, 2 = never, once, repeatedly)
6	Jaundice after attact without fever (0, 1, 2 = never, once, repeatedly)
7	Cholangitis (0, 1, 2 = never, once, repeatedly)
8	Tangible drops (0, 1 = no, yes)
9	Number of attacks (1—6, 6 = six and more)
10	Age in the time of agreement (with operation)
11	Itching (0, 1 = no, yes)
12	Jaundice after attack with fever (0, 1, 2, 3 = never, once, moretimes, always)
13	Gull-bladder filling (0, 1 = no, yes)
14	Duration of obstructions (in years)
15	Sex (0, 1 = woman, man)
16	Jaundice lasted more than 14 days (0, 1 = no, yes)

It is necessary to pay attention to unascertained values, because for subsequent data processing the missing values are not allowed. The feature "maximum width of gull-duct in mm" can be measured on x-ray image. But the gull-duct is not always displayed and so this feature value may remain unknown. The training set had to be split into two parts with displayed and undisplayed gull-duct, respectively. The classification problem should be, therefore, solved separately for both parts. However, as the part with undisplayed gull-duct consisted only of 49 patients total, the problem has not been solved for this group.

After these arrangements the training set contains 108 pattern vectors of class ω_1 and 311 elements of class ω_2. Dimensionality of the problem is 16, 5 components are of the real type and 6 ones of the ordinary type. The list of features is in Table 1. (Features are ordered according to their significance.) Apriory probability of the class ω_1 – "complicated form" for the population has been estimated by the value $P_1 = 0{\cdot}20$, so $P_2 = 0{\cdot}80$. It has been requested the value 0·25 as permitted misclassification probability for the class ω_1 – "complicated form" and the value 0·10 for the class ω_2 – "simple form".

As a result of data processing by DATANAL program the basic statistical characteristics have been obtained, however, they provide no direct possibility for classification. Some from those characteristics are in Table 2.

Table 2. Basic statistical characteristics.

Feature	Relat. freq.		Complicated		Simple	
	cl. 1	cl. 2	average	st. dev.	average	st. dev.
Age			53·80	12·07	46·7	12·31
Duration of obstr.			8·91	8·94	6·07	7·26
Number of attacks			5·35	1·32	4·80	1·73
Max. ALP			1·16	1·02	0·67	0·31
Max. width of g.-d.			10·71	4·16	6·57	2·17
Sex	30·0	50·8	0·31	0·46	0·30	0·46
Course of attacks			0·79	1·22	0·42	0·66
Pankreatitis			0·34	0·63	0·15	0·45
Cholangitis			0·25	0·63	0·02	0·15
Jaundice without f.			0·29	0·66	0·12	0·44
Jaundice with fever			0·48	0·90	0·26	0·66
Jaundice long time	4·7	89·7	0·08	0·28	0·01	0·10
Gull-bladder filling	55·9	48·5	0·54	0·50	0·58	0·50
Stones			0·46	0·78	0·02	0·17
Itching	5·9	78·4	0·04	0·19	0·03	0·16
Tangible drops	7·7	24·4	0·04	0·19	0·12	0·32

Having used the second program of the system (MEDIP), we have obtained the significance analysis of the features. The course of feature excluding is in Table 3. The last columns of this table contain significances $S(x_i)$ of features. Column 5 contains the feature significances "inside" the whole subset of remaining features in that step, in which is the feature excluded. Column 6 contains the "individual" feature significances under hypothetical assumption that the feature is independent on the others.

Let us pay attention to the fact that the sooner excluded features have not the smallest values of individual independent significance. Until the dimensionality 9 has been reached the decrement of MDP was less then 5% of maximal MDP, thus all these features can be definitely excluded, because the significance of this whole group of features for the considered decision making is negligible.

From the significance analysis it follows that not more than 12 features, chosen according to the computed order, are sufficient for the decision rule construction. On the other hand, in the case of less than four features the loss of discriminative power is already essential. From these reasons the decision rules for all numbers of features in the range from 4 to 15 have been computed and evaluated from the

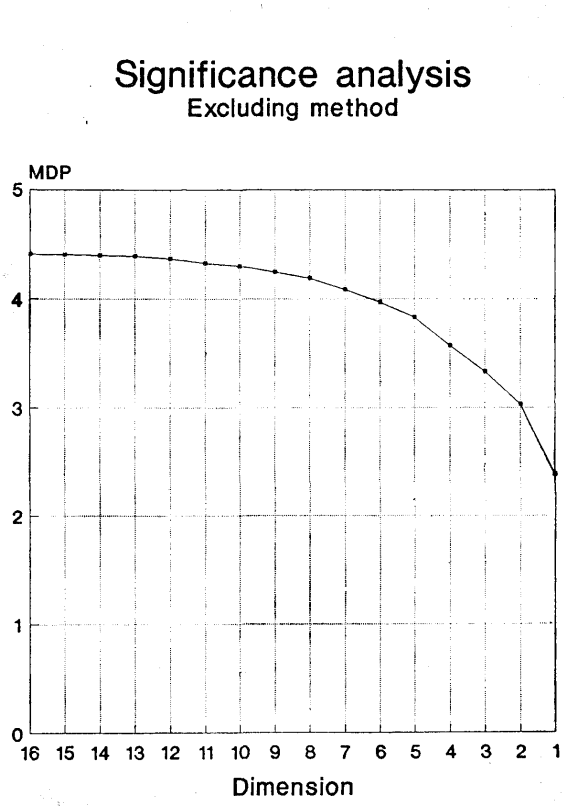

Fig. 1.

Table 3. Course of feature excluding.

Dimen-sionality	No of excluded feature	MDP	Decrement to Max in %	MDP of excluded feature	
				in the set	independ.
1	2	3	4	5	6
16		4·407	0·00		
15	16	4·406	0·02	0·0009	0·239
14	15	4·404	0·07	0·012	0·0002
13	14	4·390	**0·39**	0·014	0·140
12	13	4·372	**0·81**	0·019	0·005
11	12	4·334	**1·66**	0·037	0·078
10	11	4·296	2·52	0·038	0·124
9	10	4·246	3·67	0·051	0·321
8	9	4·190	**4·94**	0·056	0·076
7	8	4·090	7·20	0·099	0·068
6	7	3·966	10·02	0·125	0·573
5	6	3·826	13·20	0·140	0·118
4	5	3·569	19·01	0·256	0·139
3	4	3·332	24·39	0·237	0·847
2	3	3·025	31·35	0·307	0·474
1	2	2·384	**45·91**	0·642	0·840
	1	0·000	100·00	2·384	2·384

viewpoint of probability of error and reject probability. The results are presented in Table 4 and in Figure 2.

Probabilities G_{ij} $(i = 1, 2; j = 0, 1, 2)$ that patterns $\mathbf{x}$ from classes ω_i are mapped into acceptance and reject regions Ω_j are the results of classifications using decision

Table 4. Class conditional non-error (G_{ii}), error (G_{ij}) and reject (G_{0i}) rates.

M	G_{11}	G_{21}	G_{01}	G_{22}	G_{12}	G_{02}
5	0·355	0·075	0·570	0·712	0·010	0·278
6	0·486	0·075	0·439	0·721	0·013	0·266
7	0·411	0·056	0·533	0·744	0·013	0·243
8	0·393	**0.047**	0·560	0·753	**0·010**	0·237
9	0·467	**0·047**	0·486	0·724	0·013	0·263
10	0·477	0·056	0·467	0·750	0·016	0·234
11	**0·542**	0·065	**0·393**	0·782	0·022	0·196
12	0·477	0·084	0·439	**0·805**	0·022	**0·173**
13	0·383	0·065	0·552	0·776	0·042	0·182
14	0·374	0·056	0·570	0·776	**0·010**	0·214
15	0·383	0·047	0·570	0·737	**0·010**	0·253
CLIN.	0·598	0·196	0·206	0·907	0·016	0·077

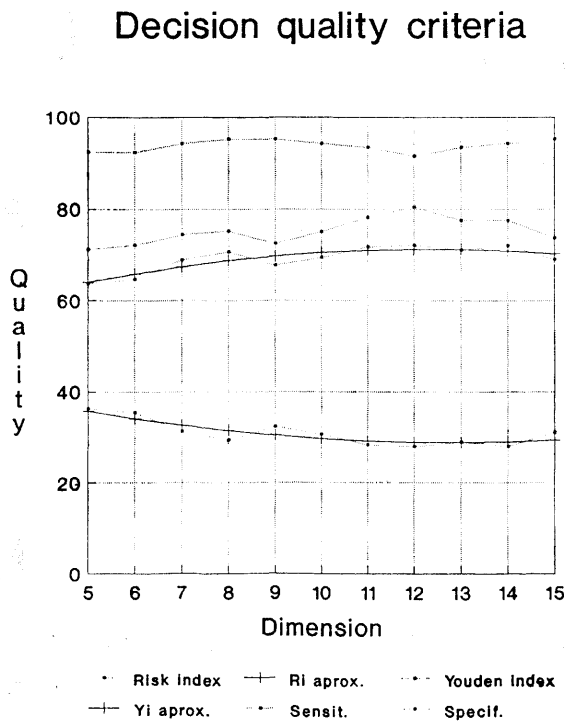

Fig. 2.

rules with the number of features from 5 to 15 and are recorded in Table 4. In the last row of this table there are presented results of pure clinical diagnostics of this problem for the same set of patients with the full number of investigated features. The classification quality has been evaluated by estimations of the classification errors and risk index. Most of all medical diagnostic problems cannot afford to have a case without decision: either is possible to made more informative examination or to abandon the error probability limitation at least one class. For this case the second possibility is the true one and therefore in the last step of the decision rule the patterns mapped into reject region Ω_{0i}, $i = 1, 2$, are classified into class ω_1, e.g. into class "complicated form". (For computation the value G_{02} must be added to the value G_{12} and value G_{01} to the G_{11}). On the medical field the test quality is evaluated by sensitivity and specificity of the test and by its joint in Youden's index. If $E_i(d) = P_j G_{ij}$ $(i = 1, 2;\ j = 1, 2;\ i \neq j))$ is the error probability of the rule d then specificity SP is

$$SP(d) = [1 - E_1(d)]\,100$$

and sensitivity SE is

$$SE(d) = [1 - E_2(d)]\,100$$

where the class ω_1 is the one which should be recognized in population, e.g. in our case the class "complicated form". Youden index YI is

$$YI(d) = SP(d) + SE(d) - 100\,.$$

For quality evaluation of the rule by average risk $R(d)$ (11) have been defined losses like this

$$L_{11} = L_{22} = L_{01} = 0\,, \quad L_{12} = L_{02} = 1\,, \quad L_{21} = L.$$

The actual value of the loss L_{21} is not known and can be supposed value from interval (0·25, 0·5). The variant for each of both values has been computed.

Each decision has two trivial solution dependent on assigning all elements into the one class. To each trivial solution corresponds an average risk and let $R_T(d)$ is the less of them. $R_T(d) = \min(P_1, LP_2)$. The average risk $R(d)$ must be less than $R_T(d)$ to be the decision rule usable. Risk index defined as

$$RI(d) = \frac{R(d)}{R_T(d)}\,100$$

gives relative value of remaining risk for the rule.

Values, which can be used as criteria for decision rule quality are in Table 5. In this table the purely clinical diagnosis is compared with computer diagnosis for different numbers of used symptoms. The clinical diagnosis has evidently high quality but computer diagnosis gives comparable results. Clinical diagnosis has outstandingly better the values of total error probability and error probability of the class ω_2, e.g. "simple form".

Table 5. Decision quality criteria expected conditional errors E_i, total error E, senzitivity SE, specificity SP, Younden's index YI and risk indexes RI for loss ratio $L = 0{\cdot}5$ and $L = 0{\cdot}25$.

M	E_1 %	E_2 %	E	SE	SP	YI	$RI_{0.5}$	$RI_{0.25}$
5	7·5	28·8	24·5	2·5	71·2	63·7	65·1	36·4
6	7·5	27·9	23·8	2·5	72·1	64·6	63·3	35·4
7	5·6	25·6	21·6	4·4	74·4	68·8	56·8	31·2
8	**4·7**	24·7	20·7	**5·3**	75·3	70·6	54·1	29·4
9	**4·7**	27·6	23·0	**5·3**	72·4	67·7	59·9	32·3
10	5·6	25·0	21·1	4·4	75·0	69·4	55·6	30·6
11	6·5	21·8	18·7	3·5	78·2	71·7	50·1	28·3
12	8·4	**19·5**	**17·3**	1·6	**80·5**	**72·1**	**47·4**	**27·9**
13	6·5	22·4	19·2	3·5	77·6	71·1	51·3	28·9
14	5·6	22·4	19·0	4·4	77·6	72·0	50·4	28·0
15	4·7	26·3	22·0	**5·3**	73·7	69·0	57·3	31·0
CLIN.	19·6	9·3	11·4	0·4	90·7	71·1	38·2	28·9

Computer diagnosis gives outstandingly better error probability for the class ω_1, e.g. "complicated form" in the whole investigated range of dimensionality and it was exactly the main reason for the solution of given problem. Comparison of quality evaluation by risk index depends on loss ratio $L = L_{21}/L_{12}$. Computer diagnostics, in this case, is better than clinical one, if loss ration is small, e.g. if a great difference between consequences of misclassifications for both classes exists.

Table 6. Weights for stepwise decision rule solving cholecystolithiasis problem.

Features	k	Feature weights w_{kj} 1	2	3	x_j min	max
Threshold weight W_k		2·3071	−3·3571	1·3903		
Max. width of gull-duct		−0·1515	+0·1515	−0·1117	0	30
Max ALP		−0·3965	+0·3965	−0·4779	0·06	5·6
Course of attacks		−0·1690	+0·1690	−0·0509	0	5
Stones		−0·4750	+0·4750	−0·4492	0	2
Pankreatitis		−0·3056	+0·3056	−0·1238	0	2
Jaundice without fever		−0·2459	+0·2459	−0·4041	0	2
Cholangitis		−0·4415	+0·4415	−0·5758	0	2
Tangible drops		+0·2951	−0·2951	+0·1518	0	1
Number of attacks		−0·0552	+0·0552	−0·0307	1	5
Age		−0·0061	+0·0061	−0·0007	18	59
Itching		0·3450	+0·3450	−0·0131	0	1
Jaundice with fever		0·0870	−0·0870	−0·1390	0	3
Assigned class α_k		ω_2	ω_1	ω_2		

Optimal criteria values are marked in Tables 4, 5 by different print and so evidently best solution given the group of 12 features. The decision rule has three steps. Weights $w_{k,j}$ for individual features x_j, $j = 1, \ldots, m$ and for each decision step $k = 1, 2, 3$ are in Table 6 and they are considered as weight vectors components $\mathbf{w}_k$. Stepwise decision rule (7.43) to the purpose of software simplicity was modified like this:

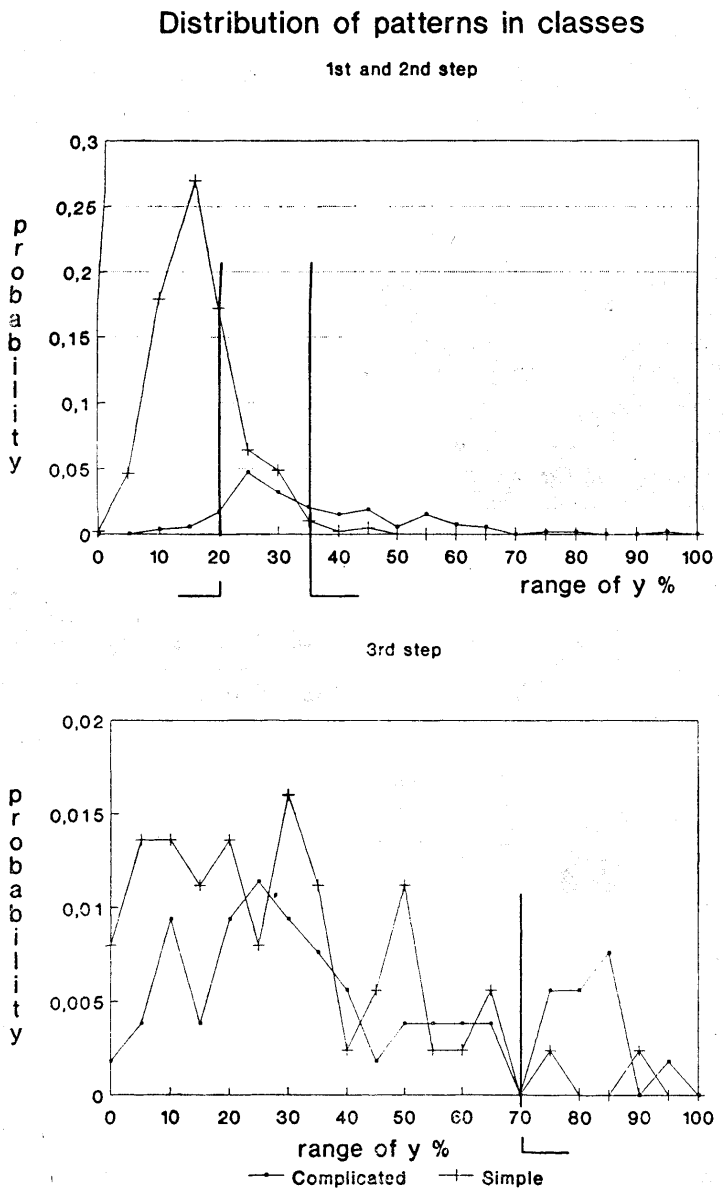

Fig. 3.

set $k = 1$ and

start: if $\mathbf{x}^T\mathbf{w}_k + w_k > 0$
then assign the pattern $\mathbf{x}$ to class α_k,
else set $k := k + 1$ and
if $k \leqq m$ then go to the start,
else reject to classify pattern $\mathbf{x}$.

Here $\alpha_k = \omega_1, \omega_2$ and sequence α_k, $k = 1, \ldots, m$ is the result of problem solution.

The features in Table 6 are ordered according theirs weights for the first decision step and this weights include range of feature values. As value $\mathbf{x}^T\mathbf{w}_k + w_k$ must be greater than 0 to be the patient $\mathbf{x}$ assigned to the class α_k, it is evident that most of all symptoms, if they have nonzero values, support classification of the patient $\mathbf{x}$ into class α_k with exception symptoms tangible drops and jaundice after attack with fever in the first two decision steps.

Distribution of the value $y = \mathbf{x}^T\mathbf{w}_k$, evaluated from the training set by histograms, are in Figure 3 which is valid for 1st and 2nd decision step. Acceptance region $\Omega_2(\mathbf{w}_1)$ for the 1st step and $\Omega_1(\mathbf{w}_2)$ for the 2nd step and reject region Ω_0 are marked out; see (7.7). All elements which are mapped into reject region Ω_0 are considered as new training set and the decision problem can be solved from the beginning. In the described problem the significance analysis was not applicated for the new training set; only a 3rd step of decision rule has been computed. Distribution of y is in Figure 3.

10. CONCISE CHARACTERISTICS OF OTHER APPLICATIONS

A number of various problems of diagnostic type from so different application fields as medicine, economics, remote serving, geology and power engineering has been solved by means of PREDITAS system.

Since one of the medical diagnostic problem has already been treated in more detail, we shall confine ourselves just to characterizing individual problems.

1. Diagnosis of risk pregnancy

The problem of finding the most important risk factors influencing the risk of abortion has been solved in cooperation with the hospital in Zlín.

A large training set consisting of various diagnostic data of 6166 women patients has been collected and labeled according to the fact whether abortion did or did not occur. A substantial reduction in the number of diagnostically significant features has been achieved by means of searching procedures in MEDIP program, resulting in cost and time saving of data preparation. The derived classification rule enables to classify a new women patient on the basis of particular values of several diagnostically important features and to make the decision about the risk of abortion.

The solution in the form of CLASSIF program, implemented on personal computer, is being at present tested directly in clinical use. The results can be found in more detail in Baran[1].

2. Diagnosis of mentally retarded children

This problem has been solved in cooperation with the Institute of Care of Mother and Child, Prague. Its essence is in the fact that the handicap can be cured quite often provided a special treatment starts at the first month of a newborn's life. However, the point is that medical specialists themselves are able to make the diagnosis not earlier than at the age of three years, when it is too late to begin with a successful treatment. So as a matter of fact, the diagnostic problem in question is a prediction problem, since in this case the decision making amounts to predicting the state of the child on the basis of data available at the first weeks of its life. The training set has been formed by the set of medical records of individual children which were stored until the child reached the age of three years when the diagnosis was made the sample record was labeled correspondingly.

A substantial reduction of original symptoms (more than one hundred symptoms) has been achieved by means of MEDIP program without loosing too much from the discriminative power. The results achieved in the first stage are very promising and on their basis a new training set of bigger size is being prepared at present.

3. Diagnosis of stress urinary incontinence

In a close cooperation with the Gynecological and obstetrical Clinic of the Charles University in Prague the problem of stress urinary incontinence diagnosis has been studied in Šindlář[2,3]. A special significance of individual uroflow-dynamic characteristics, taking into account their mutual dependencies.

4. Classification of enterprises with respect to typified software usability

Attempts to design typified software for solving specific tasks of automated control and to implement this software in Czechoslovak enterprises have been carried out. The enterprises were characterized by a set of economic indices and the training set consisted of feature vectors labeled according to found out usability of typified

[1] P. Baran, P. Mareš and S. Bláha: Influence of including social risk factors into risk frequency screening test on its quality (in Czech). Research Report OÚNZ Zlín 1990.

[2] M. Šindlář, P. Pudil and L. Papež: Use of discriminant analysis methods for determining the significance of UV profiles for diagnostics of stress urinary incontinence at women patients (in Czech). Čsl. gynekologie a porodnictví *45* (1981), 10, 700—703.

[3] M. Šindlář, P. Pudil and L. Papež: Use of measure of discriminative power for diagnostics of stress incontinence at women patients (in Czech). Proc. Conf. "Lékařská informatika", Praha 1981.

projects. The aim was to find those features which are the most significant features for differentiating between the classes of enterprises for which the typified projects are suitable or unsuitable, respectively.

The derived classification rule can be used for the decision whether an enterprise going to solve certain tasks of automated control should use the typified project or if it is not suitable. Moreover, the results of the significance analysis can be utilized with the aim to carry out such changes in typified software projects so as to extend their scope of applicability. More about it is in Říhová[4].

5. Recognition of data from remote sensing

Though in remote sensing problems predominantly contextual structural approaches to picture analysis are used, there are specific problems where the statistical feature approach can be used. A typical classification problem of this nature is recognition of forests based on the data from multispectral analysis recorded during the flight of aircraft. The results are presented in more detail in Bláha[5].

6. Recognition and classification of rock types in geology

The usability of PREDITAS system in geology has been demonstrated by the solution of two following problems. The first problem has been solved in cooperation with the Department of Geology of Mineral Deposits, Faculty of Science, Charles University, Prague. The aim was to carry out the multidimensional analysis of feature significance for recognition of two types of granite rocks from the Moldanubium of the Bohemian Massiff. The samples were characterized by their chemical components. The results are summarized in Pudil[6] and in more detail in Pudil[7].

The second problem solved in cooperation with the Institute of Geology and Geotechnics of the Czechoslovak Academy Sciences in Prague, was the problem of classifying tholeiitic and calc-alkaline volcanic rock associations based on their major elements. A correct classification of these two types is very important from the viewpoint of prospective mining of certain special metallic ones. The results are summarized in Bláha[8].

[4] Z. Říhová, P. Pudil and S. Bláha: Classification of enterprises with respect to usability of typified software for automated control. Proc. 7th Internat. Conf. on System Eng. "SI' 88", Mar. Lázně. Dům techniky ČSVTS Praha, 1988, pp. 210—211.

[5] S. Bláha and P. Pudil: Recognition of forest (in Czech). Research Report ÚTIA ČSAV, No. 1339, Praha 1985.

[6] P. Pudil and S. Bláha: Recognition of rock types according to geochemical data by means of pattern recognition methods (in Czech). Research Report ÚTIA ČSAV, No. 1333, Praha 1985.

[7] P. Pudil, S. Bláha and Z. Pertold: Significance analysis of geochemical data for rock type discrimination by means of PREDITAS program system. In: Proc. Internat. Symp. on Mat. Methods in Geology, Příbram 1989, pp. 119—125.

[8] S. Bláha, P. Pudil and F. Patočka: Program system PREDITAS and its possible application in geology. In: Proc. Internat. Symp. on Math. Methods in Geology, Příbram 1989, pp. 6—17.

7. Recognition of emergency states in power network

In large power network certain states may develop which, let just by themselves without operator's intervention, could lead to the network breakdown. Therefore, it is desirable, to be able to detect such risk or emergency states so that the operator may use some reserves of power or to intervene in another way with the aim to return the state of network back to normal.

Since acquiting data of emergency states and of possible breakdowns during real functioning of power network is hardly feasible, another approach has been adopted. A simulation model of the power network has been designed at the Faculty of Electrical Engineering, enabling to generate states characterized by multidimensional data vectors and to evaluate them according to the development of the network (state does or does not converge to a breakdown). The first results from PREDITAS system specifying the discriminatory plane offer the possibility to generate further training samples of the vicinity of the discriminatory plane and thus to specify its equation with more precision. So this problem solution is a perfect example of iterative approach to arriving to the final solution. This approach has been successfully verified on an artificial reference power network and the project is still under way.