Contextual and Isolated Algorithms for Multistage Pattern Recognition

MAREK KURZYNSKI
Chair of Systems and Computer Networks, Faculty of Electronics
Wroclaw University of Technology
Wyb.Wyspianskiego 27, 50-370 Wroclaw
POLAND

Abstract: - This paper deals with the recognition algorithms of a multistage classifier based on a decision tree scheme. For the given tree skeleton and features to be used, concepts of contextual and isolated decision rules (strategies) for performing the classification are discussed. These both strategies are compared in respect of classification accuracy and the upper bound of difference between their probabilities of misclassification is given. The empirical versions of contextual and isolated strategies were practically implemented in the computer-aided prognosis of sacroileitis development and results of classification accuracy on the real data are presented.

Key-Words: - Multistage recognition, contextual strategy, isolated strategy, medical diagnosis

1 Introduction

In many practical pattern recognition problems, the number of features and the number of pattern classes are both very large. In such cases it would be advantageous to use a multistage recognition system [1]. The procedure of multistage classification consists of the following sequence of activities.

At the first stage, some specified features x_0 chosen from among all accessible features x, describing the object being recognized are measured. These features constitute a basis for making a decision $d_i^{(1)}$. This decision, being the result of recognition at the first stage, defines a certain subset in the set of all classes and simultaneously indicates features $x_i^{(1)}$ (from among x) which should be measured in order to make a decision at the next stage. Now, at the second stage, features $x_i^{(1)}$ are measured, which together with $d_i^{(1)}$ are a basis for making the next decision $d_i^{(2)}$. This decision - like $d_i^{(1)}$ - indicates features $x_i^{(2)}$ necessary to make the next decision (at the third stage) and - again as at the previous stage - defines a certain subset of classes, not in the set of all classes, however, but in the subset indicated by the decision $d_i^{(1)}$. Generally, at the nth stage, the decision $d_i^{(n)}$, made on the basis of the measured features $x_i^{(n-1)}$ specified by the previous decision $d_i^{(n-1)}$, defines a subset in the set of classes indicated by the decision $d_i^{(n-1)}$ and specifies features $x_i^{(n)}$ necessary to make a decision at the (n+1)th stage. The whole procedure ends at the last, Nth stage, where the decision made ($d_i^{(N)}$) indicates a single class, which is the final result of multistage classification.

The action of the multistage classifier can be conveniently described by means of a decision tree, in which the terminal nodes represent pattern classes and every interior node is connected with an appropriate set of classes accessible from that node. In particular, the root-node represents the entire set of classes into which a pattern may be classified. In order to classify the unknown pattern into a class, one has to traverse a path of the tree starting at the root-node, and at each encountered nonterminal node one ought to take a decision on the further path in the tree, until a terminal node is reached. This terminal node represents the final classification and its label indicates to which class the unknown pattern is assigned.

The synthesis problem of a multistage classifier can be decomposed into the following three components [2]:

- The choice of decision logic of the classifier, i.e. the specification of the decision tree skeleton,
- Feature selection for every interior node,
- The specification of decision rules (recognition algorithms) at every interior node.

The present paper is devoted to the last problem only, i.e. we we shall deal with the determination of recognition algorithms, assuming that both the tree skeleton and features used at each nonterminal node are given. For the case of complete probabilistic information two concepts of strategies for performing

the classification at each nonterminal node are derived and discussed. First of them takes into account the context of decision making resulting from the fact, that multistage recognition is not a single activity, but constitutes a multistep decision process. The second one treats recognition tasks at each stage independently, as an isolated action.

This paper is a sequel to the author's earlier publications [3, 4, 5, 6, 7] and it yields a generalization and extension of the results included therein.

The contens of the work are as follows. Section 2 introduces necessary notations and provides the problem statement. In section 3 we present the concept of contextual and isolated strategies of multistage recognition and furthermore we discuss and compare both strategies with respect to the classification accuracy. In section 4 some remarks dealing with empirical versions of contextual and isolated strategies based on k-nearest neighbours procedure are presented. These empirical algorithms were practically implemented in the computer-aided prognosis of sacroileitis develpment and results of classification accuracy obtained on the real data are given in section 5. Finally, section 6 concludes the paper.

2 Preliminaries and Notation

Let us consider a pattern recognition task with m classes (m>2) which are organized into an (N+1)-level decision tree. This enables the introduction of the following notation:

$M^{(i)} = \left\{ d_1^{(i)}, d_2^{(i)}, \ldots, d_{m_i}^{(i)} \right\}$ - the set of node numbers (labels) at the i-th level of decision tree (i=0,1,2,...,N). In particular $M^{(N)} = \left\{ d_1^{(N)}, d_2^{(N)}, \ldots, d_m^{(N)} \right\}$ denotes the set of class numbers and $M^{(0)} = \left\{ d_1^{(0)} \right\}$ is a number of the root-node.

$M_j^{(i)}$ - the set of numbers of immediate descendant nodes of the $d_j^{(i)}$ node.

Let us denote by x the vector of observed values of all accessible features describing patterns being recognized; x is an element of an l-dimensional vector space $\mathsf{X} \subseteq \mathsf{R}^l$. We will continue to adopt the probabilistic model of the recognition problem, i.e. we will assume the the class number of the pattern being classified $d_i^{(N)} \in M^{(N)}$ and vector x are observed values of a couple of random variables $\boldsymbol{d}^{(N)}$ and $\boldsymbol{X}$, respectively. The complete probabilistic information requires the knowledge of *a priori* probabilities of classes

$$p_i^{(N)} = P(\boldsymbol{d}^{(N)} = d_i^{(N)}) \tag{1}$$

and class-conditional probability density functions (CPDFs)

$$f_i^{(N)}(x) = f(x / d_i^{(N)}), \quad x \in \mathsf{X},\ d_i^{(N)} \in M^{(N)}. \tag{2}$$

Every interior node of the constructed decision tree is connected with some features which observed values are a basis for making a decision at this node. Let $x_i^{(j)} \in \mathsf{X}_i^{(j)}$ denotes vector of features used at the $d_i^{(j)}$, which have been selected from the entire vector x. Thus $x_i^{(j)}$ is an observed value of the random variable $\boldsymbol{x}_i^{(j)}$ and its CPDFs can be determined from (2) by integration after complementary variables.

Let now

$$\psi_i^{(j)}: \mathsf{X}_i^{(j)} \to M_i^{(j)} \tag{3}$$

be a decision rule (recognition algorithm) used at the $d_i^{(j)}$ node, which maps observation subspace to the set of numbers of immediate descendant nodes of the $d_i^{(j)}$ node.

In the next section we consider and discuss different strategies π of a multistage classifier, i.e. the set of classifying rules (3), that is

$$\pi = \left\{ \psi_i^{(j)}: j = 0,1,\ldots,N-1, \quad i = 1,2,\ldots,m_j \right\}, \tag{4}$$

under assumption that complete probabilistic information is given, i.e. that both *a priori* probabilities (1) and CPDFs (2) are known.

3 Strategies of Multistage Classification

Procedure of multistage recognition is a compound multistep decision process in which final result, i.e. final classification depends on all intermediate decisions. Taking this fact into account in the procedure of construction of decision rules for particular interior nodes of decision tree, we have to formulate a global performance index. Next, minimizing (maximizing) it, we obtain recognition algorithms into which the context of decision making will be reflected. The set of such rules for all nodes forms strategy (4), called the contextual strategy of multistage classification.

In alternatively approach to construction of strategy π every recognition task is treated independently as an isolated (context-free) local classifier. Now, minimization undergoes a local criterion, evaluating an immediate effect of this decision. Such a procedure leads to the set of recognition algorithms, which together forms so-called isolated multistage recognition strategy.

Details of construction procedures for both strategies are presented in next subsections.

3.1 Contextual Strategy

Let us introduce as the total performance measure of multistage classifier the overall probability of error:

$$Pe(\pi) = 1 - P_\pi(A_1, A_2, ..., A_N) = P_\pi(B_1) + P_\pi(A_1, B_2) + ... + P_\pi(A_1, A_2, ..., A_{N-1}, B_N) \quad (5)$$

where A_i (B_i) denotes the event that at the i-th stage of classification a correct classification (an error) is made and $P_\pi(.)$ denotes the probability of respective events under strategy π.

Now, using the dynamic programming method [8] we can receive the optimal strategy π^*, which minimizes (5). Its decision rules are the following [3]:

$$\psi_i^{*(j)}(x_i^{(j)}) = d_k^{(j+1)} \text{ if} \quad (6)$$

$$p_k^{(j+1)} f_k^{(j+1)}(x_i^{(j)}) P_{\pi^*}(A_{j+2}, .., A_m / A_{j+1}, d_k^{(j+1)}) = \max_{l: d_l^{(j+1)} \in M_i^{(j)}} p_l^{(j+1)} f_l^{(j+1)}(x_i^{(j)}) P_{\pi^*}(A_{j+2}, .. / A_{j+1}, d_l^{(j+1)})$$

j=1,2,...,N-1, i=1,2,...,m_j , $P(.)$=1 for j=N-1 and $p_k^{(j+1)}$ is a sum of *a priori* probabilities of classes attainable from the node $d_k^{(j+1)}$.

What is interesting is the manner of operation of the above decision rule. Namely, its decision indicates this node for which *a posteriori* probability of set of classes attainable from it, multiplied by the respective probability of correct classification at the next stages of recognition procedure, is the greates one. In other words, the decision at any interior node of a tree depends on the future to which this decision leads.

3.2 Isolated Strategy

Another optimal multistage classifier strategy $\hat{\pi}$ may be considered, which does not take into regard the context and which decision rules are mutually independent. Formally, this isolated strategy can be derived minimizing the local criteria, which denote probabilities of misclassification for particular nodes of a tree, namely:

$$Pe_i^{(j)} = P(B_{j+1} / A_j, d_i^{(j)}), \quad j=1,...,N-1, \; i=1,..,m_j. \quad (7)$$

This is clear that algorithms of isolated strategy $\hat{\pi}$ reduces to the well known maximum *a posteriori* probability decision rules [9], viz.

$$\hat{\psi}_i^{(j)}(x_i^{(j)}) = d_k^{(j+1)} \text{ if} \quad (8)$$

$$p_k^{(j+1)} f_k^{(j+1)}(x_i^{(j)}) = \max_{l: d_l^{(j+1)} \in M_i^{(j)}} p_l^{(j+1)} f_l^{(j+1)}(x_i^{(j)}).$$

3.3 Comparison of Error Probabilities

Comparing (6) and (8) it is worth noting that using the isolated strategy we reduce the computational complexity at the sacrifice of the classification accuracy, because minimization of the error probability at each individual node of a decision tree does not necessarily lead to the globally optimal classifier.

It should be interesting to compare the contextual and isolated strategies with respect to the classification accuracy. The following lemma gives the upper bound of difference between the probabilities of misclassification for the both strategies to be considered.

Lemma

For N=2 the folowing inequality holds:

$$P_e(\bar{\pi}) - P_e(\pi^*) \le p_{\hat{\pi}}(B_1) \times \times \left[\max P_{\hat{\pi}}(B_2 / A_1, d_i^{(2)}) - \min P_{\hat{\pi}}(B_2 / A_1, d_i^{(2)}) \right]. \quad (9)$$

Proof. To simplify notation let $\hat{P}(.) = P_{\hat{\pi}}(.)$ and $P^*(.) = P_{\pi^*}(.)$. First notice that

$\hat{P}(B_2 / A_1, d_i^{(2)}) = P^*(B_2 / A_1, d_i^{(2)})$ and

$\hat{P}(B_1) \le P^*(B_1)$.

Now we have:

$$\hat{P}_e - P_e^* = \hat{P}(B_1) + \hat{P}(B_2, A_1) - P^*(B_1) - P^*(B_2, A_1) =$$

$$\sum_i \left(\hat{P}(B_1 / d_i^{(2)}) p_i^{(2)} + \hat{P}(B_2 / A_1, d_i^{(2)}) \hat{P}(A_1 / d_i^{(2)}) p_i^{(2)} \right) -$$

$$\sum_i \left(P^*(B_1 / d_i^{(2)}) p_i^{(2)} + P^*(B_2 / A_1, d_i^{(2)}) P^*(A_1 / d_i^{(2)}) p_i^{(2)} \right) =$$

$$\sum_i \{ \hat{P}(B_2 / A_1, d_i^{(2)}) p_i^{(2)} [\hat{P}(A_1 / d_i^{(2)}) + \hat{P}(B_1 / d_i^{(2)})] +$$

$$p_i^{(2)} \hat{P}(B_1 / d_i^{(2)}) [1 - \hat{P}(B_2 / A_1, d_i^{(2)})] \} -$$

$$\sum_i \{ P^*(B_2 / A_1, d_i^{(2)}) p_i^{(2)} [P^*(A_1 / d_i^{(2)}) + P^*(B_1 / d_i^{(2)})]$$

$$+ p_i^{(2)} P^*(B_1 / d_i^{(2)})) [1 - P^*(B_2 / A_1, d_i^{(2)})] \} =$$

$$\sum_i p_i^{(2)} \hat{P}(B_1 / d_i^{(2)}) [1 - \hat{P}(B_2 / A_1, d_i^{(2)})] -$$

$$p_i^{(2)} P^*(B_1 / d_i^{(2)}) [1 - \hat{P}(B_2 / A_1, d_i^{(2)})] \} \le$$

$$\hat{P}(B_1)\left\{\max_i[1-\hat{P}(B_2/A_1,d_i^{(2)})]-\min_i[1-\hat{P}(B_2/A_1,d_i^{(2)})]\right\}=$$

$$\hat{P}(B_1)[\max_i \hat{P}(B_2/A_1,d_i^{(2)})-\min_i \hat{P}(B_2/A_1,d_i^{(2)})].$$

Q.E.D.

4 Remarks on Multistage Recognition with Learning

In the real world there is often a lack of exact knowledge of a priori probabilities (1) and CPDFs (2), whereas only partial information is available. For instance, there are situations in which only a learning set, i.e. a set of correctly classified samples, is known. In these cases one obvious and conceptually simple method is to estimate appropriate probabilities and conditional densities from the training set and then to use these estimators to calculate discriminant functions of rules (6) or (8). As an example of such idea let us consider the k-nearest neighbour (*k*-NN) decision rule [9]. The *k*-NN procedure has been well investigated in literature and shown to be a powerful nonparametric technique for classification. In this approach one first finds the k nearest neighbours of the unknown pattern to be recognized among training patterns and next determines a decision according to a majority vote.

Application of this classification method to the multistage recognition leads to the following algorithms.

1. Contextual *k*-NN strategy (*k*-NN$_M$)

Its algorithms resulting from the decision rules (6) of contextual strategy π^* are the following:

$$\psi_i^{(j)}(x_i^{(j)})_{k-NN_M}=d_r^{(j+1)} \text{ if} \tag{10}$$

$$k_r^{(j+1)}(x_i^{(j)})\overline{P}(A_{j+2},..,A_m/A_{j+1},d_r^{(j+1)})=$$

$$\max_{l:d_l^{(j+1)}\in M_i^{(j)}} k_l^{(j+1)}(x_i^{(j)})\overline{P}(A_{j+2},..,A_m/A_{j+1},d_l^{(j+1)}),$$

where $k_r^{(j+1)}(x_i^{(j)})$ denotes the number of of neighbours to $x_i^{(j)}$ from the training set belonging to the group of classes available from the node $d_r^{(j+1)}$, contained in a minimum volume containing k neighbours to $x_i^{(j)}$. $\overline{P}(.)$ denotes empirical probability of appropriate event obtained from learning set using leave-one-out method [10]. Decision rules (10) and probabilities $\overline{P}(.)$ can be calculated alternately from the terminal level of a tree.

2. Isolated *k*-NN strategy (*k*-NN)

The same concept as previously, but now applied to the isolated strategy $\hat{\pi}$ of multistage classifier leads to the well known (for one stage recognition) form of k-NN rule, namely:

$$\psi_i^{(j)}(x_i^{(j)})_{k-NN}=d_r^{(j+1)} \text{ if} \tag{11}$$

$$k_r^{(j+1)}(x_i^{(j)})=\max_{l:d_l^{(j+1)}\in M_i^{(j)}} k_l^{(j+1)}(x_i^{(j)}).$$

5 An Practical Example: Prognosis of the Development of Sacroileitis

The aim of this study is to evaluate the usefulness of multistage recognition system and to compare classification accuracy of contextual and isolated strategies on the real data for the computer-aided prognosis of the development of sacroileitis (SI-itis), i.e. inflammatory changes of sacro-iliac joints of the spine. This area is a common dilemma in rheumatology and the decision must be made on the basis of patient's symptoms, characteristic for the initial phase of a disease [11]. Prognosis of SI-itis, as a pattern recognition task includes 4 classes organized into two-level decision tree depicted in Fig.1. The vector of features *x* contains the values of 37 items of clinical data presented in Table 1. In the Research Institute of Rheumatic Diseases in Piestany (Slovakia) 112 patients with SI-itis were examined and results of 37 clinical features were gathered. Each case record was provided with the firm diagnosis made after 5 years.

Table 1. Clinical features considered

GENERAL
Sex, Age
ANAMNESIS
Duration of disease, Recidive ischalgia, Ancylosing spondylitis in family history, Reiter's syndrome in history, Urinary infection
PAIN
Lumbar pain, Ischalgia irradiating to knee, Thoracic pain, Coxalgia, Wheel pain, Pain of sacro-iliac (S-I) joints (back), Pain of S-I joints (side), Pain of S-I joints (abdomen), Pain of S-I joints (hyperextension), Pain of S-I joints
PHYSICAL EXAMINATIONS
Iritis, Stiffness of back, Schober's distance, Thombayer's distance, Breathing excursion, Tenderness of sternoclavicular joints, Tenderness of wheels
X-RAY EXAMINATIONS
Unclearness of S-I joints, Erosions ofS-I joints, Sclerosis ofS-I joints, Partial ankylosis of S-I joints, Total ankylosis of S-I joints, Quadratization of vertebras, Anterior spondylitis, Syndesmitis, Changes of the symphysis
LABORATORY EXAMINATIONS
Sedimentation rate, α2 globulins, ASO level, HLA B27 level

In order to select the most „informative" symptoms for each nonterminal node (a total of 37 features were available for selection) the sequential forward selection procedure was used [12]. In this method, at first the best the single symptom is chosen, next, to the feature already selected, we add another one so as to create the best couple, then the best three features are chosen, including the first and second ones already selected, and so one. As the criterion of selection optimality the empirical probability of error was adopted. Repeating this procedure for each nonterminal node, we obtain the list of ranged features according to their importance in separating appropriate classes.

For the two-stage prognosis of the SI-itis development both k-NN and k-NN_M strategies were used for k=1,3,5 and their classification accuracy (frequency of misclassification) versus the number of features used at each node (according to the determined order) are presented in the Fig.2 a-c. It can be noted here that the use of contextual strategy k-NN_M in the multistage recognition leads, in general, to higher classification accuracy in comparision with the isolated form of the k-NN algorithm. For example, in the best case (k=3 and 11 features per node) the results are 14.2% and 20.9% for the 3-NN_M and 3-NN, respectively, i.e., contextual strategy represents an improvement of 6.7% over the isolated decision rules.

It should be interesting to compare the multistage approach with the one-stage classifier. For this purpose we used for computer-aided prognosis of the SI-itis development the one-stage recognition system, where distinction among all classes is made in one step. Results of classification accuracy for k-NN (k as previously) method are presented in Fig.2.d.

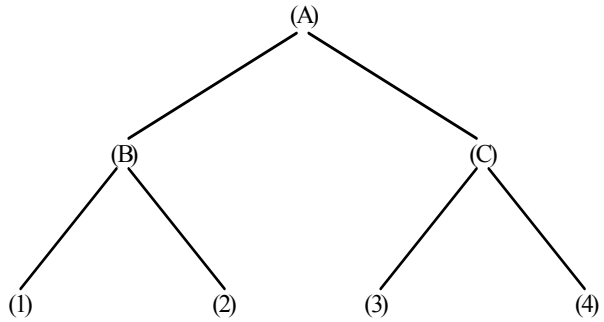

Fig.1. Decision tree for the prognosis of SI-itis:
Categories: (A) - SI-itis, (B) - Ankylosing spondylitis (Bechterew disease), (C) - Others
Final classifications: (1) - Ankylosing spondylitis (definitive), (2) - Ankylosing spondylitis (probable), (3) - SI-itis persistent, (4) - Miscellaneous/?

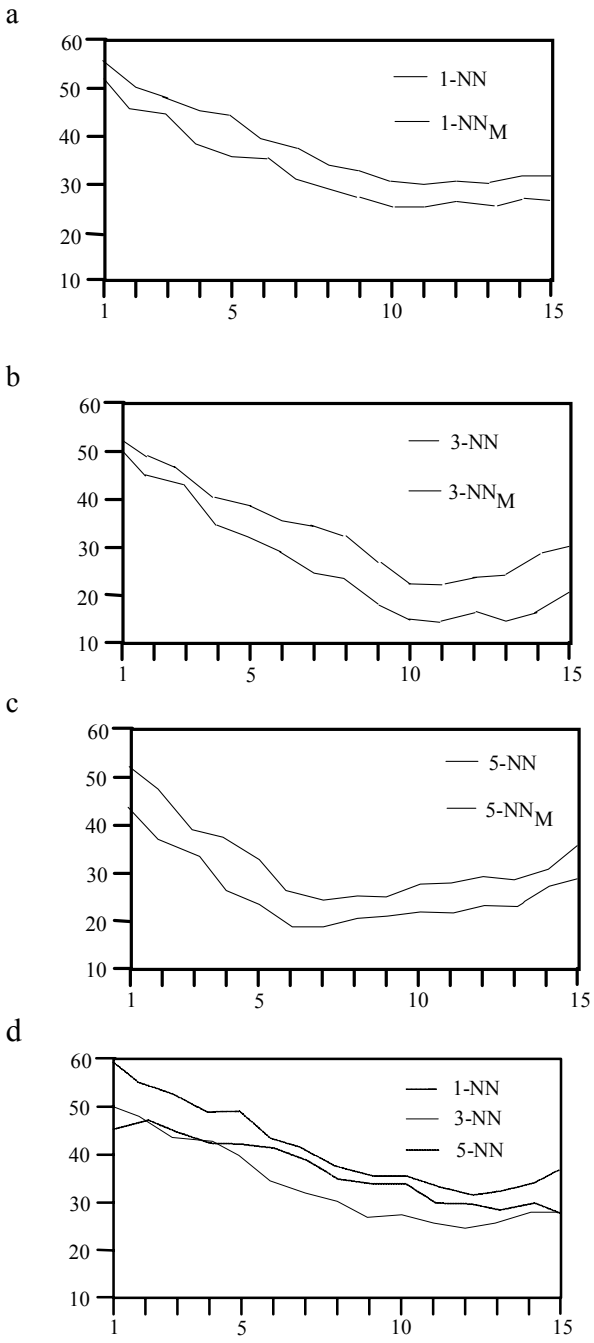

Fig.2. Frequency of misclassification (%) versus the number of features per node for two-stage (a,b,c) and one-stage classification (d).

6 Final Remarks

In this paper we have focused our attention on the decision rules for decision-tree classifier. For the case of complete probabilistic information two concepts of strategies for multistage decision making have been discused. First of them takes into account the context of decision making resulting from the fact, that multistage recognition is not a single activity, but constitutes a multistep decision process. The second one treats

recognition tasks at each stage independently, as an isolated action. The both contextual and isolated strategies were compared analitically with respect to the classification accuracy and furthermore, their empirical versions based on *k*-NN procedure were practically implemented in the computer-aided prognosis of SI-itis development.

The superiority of the presented empirical results for the contextual strategy over isolated one demonstrates the effectiveness of the proposed concept of using context in decision-tree classifier and yield some recommendation for a wide range of applications of multistage recognition system, not only in the medical domain.

Acknowledgement. This work was financed from the Polish Ministry of Science and Higher Education resources in 2007-2009 years as a research project No N518 019 32/1421.

References:

[1] Quinlan J., Decision trees and decision making, *IEEE Trans. on System, Man and Cybernetics*, 20, 339-346

[2] Safavian S., Landgrebe D., A survey of decision tree classifier methodology, *IEEE Trans. on System, Man and Cybernetics*, 21, 660-674

[3] Kurzynski M., On the Multistage Bayes Classifier, *Pattern Recognition*, 21, 355-365.

[4] Kurzynski M., On the Identity of Optimal Strategies for Multistage Classification, *Pattern Recognition Letters*, vol.10, no 2, pp.39-46.

[5] Kurzynski M., Probabilistic algorithms, neural networks, and fuzzy system applied to the multistage diagnsis of acute abdominal pain - a comparative study of methods, *Proc. 1st Int. Conference on Fuzzy Systems and Knowledge Discovery*, Singapore, 18-20 Nov. 2002

[6] Kurzynski M., Multistage empirical Bayes approach versus artificial neural network to the computer aided myocardial infraction diagnosis, *Proc. IEEE EMBS Conference*, Vienna, 4-7 Dec. 2002, 762-766

[7] Kurzynski M., Fuzzy Inference System for Multistage Diagnosis of Acute renal failure in Children, *Lecture Notes in Computer Science*, 2868, 99-108

[8] Bertsekas D., *Dynamic Programming and Stochastic Control*, Academic Press, New York 1998.

[9] Duda R., Hart P., Stork D., *Pattern Classification*., Wiley, New York 2001

[10] Fukunaga K., Hummels D., Leave-One-Out Procedure for Nonparametric Error Estimates, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11, 223-226

[11] Wood H., *Epidemiology of Rheumatic Disease*, [in:] Textbook of Rheumatic Disease, [ed.] Scott J., Churchill-Livingstone, London 1988.

[12] Kanal L., Patterns in Pattern Recognition, *IEEE Trans on Information Technology*, 20, 697-722.