Control and Cybernetics

VOL. **21** (1992) No. 3/4

The experimental investigations of pattern recognition algorithms for second-order Markov chains

by

Andrzej Żołnierek

Technical University of Wrocław
Institute of Control and Systems Engineering

The paper deals with the classification problems in which the sequences of recognized patterns form second-order Markov chain. First the algorithm in the case of complete probabilistic information is presented and furthermore the respective algorithm with learning is proposed. Subsequently, the results of experimental investigation of both algorithms are shown.

1. Introduction

In many pattern recognition tasks there exist dependencies among the patterns to be recognized. For instance, this situation is typical in text recognition [7], or in recognition of state of technological processes [1], or in sequential medical diagnosis when we need to classify the sequence of medical tests for the same patient [2, 8, 9]. In such cases the Bayes approach with the assumption of Markov dependence among the patterns to be recognized is often made.

There is a great amount of papers dealing with recognition problems under assumption of first-order Markov dependence [1, 5, 7]. Furthermore, in [6], the pattern recognition algorithm for k–order ($k > 1$) Markov chains, but only in the case of complete statistical information, can be found. In this paper the pattern recognition algorithm with learning for second–order Markov chain is presented in detail. Subsequently, the results of experimental investigations (for algorithm with complete probabilistic information and with learning) are shown.

2. Statement of the problem

Let us consider the recognition problem of sequence of patterns in which there exist a second–order Markov dependence among the classes to which the patterns belong. Let $x_n = [x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(r)}]$, taking values in the r–dimensional space of observations, denote the vector of measured features of the n–th ($n = 1, 2, \ldots$) recognized pattern. Let us denote by j_n the class index to which the n–th pattern in question belong and let j_n take values in the set of class $M = 1, 2, \ldots, m$. We assume that couples (x_n, j_n) are realizations of the stochastic process $F_n(\overline{X}_n, \overline{J}_n)$ described by the probabilities:

$$P\{J_1 = j_1, J_2 = j_2, \ldots, J_n = j_n\} = p_n(\overline{j}_n) \tag{1}$$

and the joint conditional probability density functions:

$$f(x_1, x_2, \ldots, x_n / J_1 = j_1, J_2 = j_2, \ldots, J_n = j_n) = f_n(\overline{x}_n / \overline{j}_n) \tag{2}$$

where $\overline{x}_n = (x_1, x_2, \ldots, x_n)$ denotes the sequence of n patterns and $\overline{j}_n = (j_1, j_2, \ldots, j_n)$ denotes the sequence of their true identities. Additionally, we suppose stationarity and conditional independence in the sequence of random variables $\overline{X}_n$ i.e., for each n:

$$\overline{f}_n(\overline{x}_n / \overline{j}_n) = \prod_{\alpha=1}^{n} f(x_\alpha / j_\alpha) \tag{3}$$

where

$$f(x_n / j_n) = f_{j_n}(x_n), \qquad j_n \in M \tag{4}$$

denote the class density functions.
If the sequence $\overline{J}_n = (J_1, J_2, \ldots, J_n)$ forms a second–order Markov chain, i.e.:

$$P\{J_n = j_n / \overline{J}_{n-1} = \overline{j}_{n-1}\} = P\{J_n = j_n / J_{n-1} = j_{n-1}, J_{n-2} = j_{n-2}\} \tag{5}$$

for $n > 2, j_n, j_{n-1}, \ldots, j_1 \in M$ then the complete probabilistic information means that all density functions (4), and all transition probabilities of second-order Markov chain:

$$\begin{aligned} &P\{J_n = j_n / J_{n-1} = j_{n-1}, J_{n-2} = j_{n-2}\} = \overline{p}_n(j_n, j_{n-1}, j_{n-2}) \\ &n = 3, 4, \ldots, j_n, j_{n-1}, j_{n-2} \in M \end{aligned} \tag{6}$$

and its initial probabilities:

$$\begin{aligned} &P\{J_1 = j_1, J_2 = j_2\} = p_2(\overline{j}_2) \\ &j_1, j_2 \in M \end{aligned} \tag{7}$$

are known.

3. The pattern recognition algorithms

In [6] and [9] the Bayes pattern recognition algorithm for second-order Markov chains in the case of complete probabilistic information is presented. In the case of 0–1 loss function first the following functions:

$$\begin{aligned} &h_n(j_n, j_{n-1}, \overline{x}_n) = \\ &\qquad P\{J_n = j_n, J_{n-1} = j_{n-1}\} \cdot f(\overline{x}_n / J_n = j_n, J_{n-1} = j_{n-1}) \\ &\qquad n = 2, 3, \ldots, j_n, j_{n-1} \in M \end{aligned} \tag{8}$$

are defined. In [6, 9] it is also shown, that these functions can be calculated according to the recursive formula:

$$\begin{aligned} &h(j_n, j_{n-1}, \overline{x}_n) = \\ &\qquad f_{j_n}(x_n) \sum_{j_{n-2}=1}^{m} \overline{p}(j_n, j_{n-1}, j_{n-2}) h_{n-1}(j_{n-1}, j_{n-2}, \overline{x}_{n-1}) \\ &\qquad j_n, j_{n-1} \in M, \qquad n = 3, 4, \ldots \end{aligned} \tag{9}$$

with the initial condition:

$$\begin{aligned} &h_2(j_2, j_1, \overline{x}_2) = f_{j_2}(\bar{x}_1) p_2(\bar{x}2) \\ &j_1, j_2 \in M \end{aligned} \tag{10}$$

From the formulas (9),(10) we derive the decision functions

$$\begin{aligned} &g_n(j_n, \overline{x}_n) = \textstyle\sum_{j_{n-1}=1}^{m} h_n(j_n, j_{n-1}, \overline{x}_n) \\ &n = 2, 3, \ldots \quad j \in M. \end{aligned} \tag{11}$$

For $n = 1$ we use the Bayes decision rule for the sequence of independent patterns with following decision functions:

$$g_1(j_1, x_1) = f_{j_1}(x_1) \sum_{j_2=1}^{m} p_2(\overline{j}_2) \qquad j_1 \in M \tag{12}$$

Knowing (11) or (12) we classify the n–th recognized pattern to the class i_n, for which the value of decision function after observing $\overline{x}_n$ is the greatest one, i.e.:

$$\Psi(\overline{x}_n) = \Psi_n(x_n) = i_n \tag{13}$$

if:

$$g_n(i_n, \overline{x}_n) = \max g_n(j_n, \overline{x}_n) \qquad j_n \in M \tag{14}$$

The above described algorithm can be also extended for the higher–order Markov dependence and for the case of general loss function [6, 9]. When the class density functions (4) and the description of the Markov chain (6),(7) are unknown, we can use the information contained in the set of learning sequences [2, 9]: $[S]_N = S_1, S_2, \ldots, S_N$. Each sequence is the realization of the stochastic process (1),(2),(3) and contains k correctly classified patterns:

$$\begin{aligned} S_1 &= (x_{11}, j_{11}), (x_{12}, j_{12}), \ldots, (x_{1k}, j_{1k}) \\ S_2 &= (x_{21}, j_{21}), (x_{22}, j_{22}), \ldots, (x_{2k}, j_{2k}) \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ S_N &= (x_{N1}, j_{N1}), (x_{N2}, j_{N2}), \ldots, (x_{Nk}, j_{Nk}) \end{aligned} \tag{15}$$

The pattern recognition algorithm with learning in n–th moment of classification is as follows:

$$i_n = \Psi([S]_N; \overline{x}_n) = \Psi_{N,n}(x_n) \tag{16}$$

Now, let us consider the problem of using the set of learning sequences (15). In this case the transition probabilities of Markov chain $\overline{p}_{N,n}(i, j, 1)$ can be estimated by:

$$\overline{p}_{N,n}(i, j, l) = \frac{N_{n,ijl}}{N_{n,jl}} \quad i, j, l \in M, \quad n = 3, 4, \ldots \tag{17}$$

where

$N_{n,ijl}$ denotes the number of events such that $j_{\beta n} = i$, $j_{\beta n-1} = j$, $j_{\beta n-2} = l$, $\beta = 1, 2, \ldots, N$ in (15), $N_{n,jl}$ denotes the number of events such that $j_{\beta n-1} = j$, $j_{\beta n-2} = l$, $\beta = 1, 2, \ldots, N$ in (15).

The initial probabilities can be estimated by:

$$p_{N,2}(i, j) = \frac{N_{2,ji}}{N} \quad i, j \in M \tag{18}$$

When the second–order Markov chain is homogeneous:

$$P\{J_n = i/J_{n-1} = j, J_{n-2} = l =\}P\{J_{n-1} = i/J_{n-2} = j, J_{n-3} = l\} = \qquad (19)$$
$$\ldots = P\{J_3 = i/J_2 = j, J_1 = l\} = \overline{p}(i,j,l) \quad i,j,l \in M$$

and stationary:

$$P\{J_1 = i, J_2 = j\} = P\{J_2 = i, J_3 = j\} =$$
$$P\{J_{n-1} = i, J_n = j\} = p_2(i,j) \qquad (20)$$
$$i,j \in M.$$

for the estimation of probabilities [6, 7] we can use only one learning sequence (of course the sequence should be long enough):

$$S_N = (x_1, j_1), (x_2, j_2), \ldots, (x_N, j_N) \qquad (21)$$

In this situation the transition probabilities can be estimated by:

$$\overline{p}_N(i,j,l) = \frac{N_{ijl}}{N_{jl}} \frac{N-1}{N-2} \qquad (22)$$

where
N_{ijl} denotes the number of events such that $j_\alpha = i$, $j_{\alpha-1} = j$, $j_{\alpha-2} = 1$, $\alpha = 3, 4, \ldots, N$ in (21), and N_{ji} denotes the number of events such that $j_{\alpha-1} = j$, $j_{\alpha-2} = l$, $\alpha = 2, 3, \ldots, N$ in (21).
The initial probabilities can be estimated according to:

$$p_{N,2}(i,j) = \frac{N_{ij}}{N-1} \qquad (23)$$

where N_{ij} denotes the number of events that in (21) $j_\alpha = i, j_{\alpha-1} = j, \alpha = 2, 3, \ldots, N$.
Let us notice that according to the Bernoulli's theorem all these estimators are consistent. The value of class density functions can be estimated using nonparametric techniques [3, 4, 9] (for example using Parzen or Loftsgaarden estimators or using the least interval pattern recognition algorithm) and let:

$$f_{N,j_n}(x_n), \quad j_n \in M, \quad n = 1, 2, \ldots \qquad (24)$$

denote the value of density in class j_n in point x_n, which is obtained using the set of learning sequences (15).
Subsequently, we use estimators (22), (23), (24) to calculate functions (10), (10) and the decision functions (11), (12) as though they were correct. Finally, our pattern recognition algorithm with learning for second–order Markov chain is

as follows:

1. We determine all additional functions for $j_n, j_{n-1} \in M$ according to the following recursive formula ($n = 3, 4, \ldots$):

$$h_{N,n}(j_n, j_{n-1}, \overline{x}_n) = \qquad (25)$$
$$f_{N,j_n}(x_n) \sum_{j_{n-2}=1}^{m} \overline{p}_{N,n}(j_n, j_{n-1}, j_{n-2}) h_{N,n-1}(j_{n-1}, j_{n-2}, \overline{x}_{n-1})$$

with the initial condition for $n = 2$:

$$h_{N,n}(j_2, j_1, \overline{x}_2) = f_{N,j_1}(x_1) f_{N,j_2}(x_2) p_{N,2}(\overline{j}_2) \qquad (26)$$

2. We determine all the decision functions for $j_n \in M$ $(n = 2, 3 \ldots)$:

$$g_{N,n}(j_n, \overline{x}_n) = \sum_{j_{n-1}=1}^{m} h_{N,n}(j_n, j_{n-1}, \overline{x}_n) \qquad (27)$$

and for $n = 1$ we have similarly as in (12):

$$g_{N,1}(j_1, x_1) = f_{N,j_1}(x_1) \sum_{j_1=1}^{m} p_{N,2}(\overline{j}_2) \qquad (28)$$

3. We classify the n–th recognized pattern to the class i_n, if:

$$g_{N,n}(i_n, \overline{x}_n) = \max_{j_n \in M} g_{N,n}(j_n, \overline{x}_n) \qquad (29)$$

For the above presented algorithm its asymptotical optimality can be proved [9]. It means that if $N \to \infty$ then the distance between the rule with learning and corresponding Bayes rule goes in probability to zero and moreover the risk of the rule with learning in probability goes to the risk of Bayes rule in every moment of classification n.

4. Experimental results

In the experimental investigations of properties of the above described pattern recognition algorithms computer simulation was used. First, the values of couples (x_n, j_n) $n = 1, 2, \ldots$ of the stochastic process (3), (5) were generated. These numbers were treated as the patterns to be recognized or as the learning patterns. In both cases the true classes to which the patterns belong were also known. As a measure of the quality of recognition the frequency of misclassification (mf) was chosen:

$$mf = \frac{l}{L} \tag{30}$$

where l denotes the number of misclassified patterns in sequence of L recognized patterns. It is well–known, that in the case of $0-1$ loss function, the frequency of misclassification estimates the probability of error of the recognition rule and its average risk as well. In the case of complete probabilistic information it was important to check out whether taking into account the second–order Markov dependency could improve the quality of recognition, i.e. whether the pattern recognition algorithm for second–order Markov chains is better in quality than the pattern recognition algorithm for the first–order Markov chains of independent patterns. This comparison was possible because knowing the description of the second–order Markov chain we can calculate corresponding description of first–order Markov chain and class probabilities at every moment $n = 1, 2, \ldots$ needed in pattern recognition algorithm for independent patterns [9]. Furthermore, in the case of learning, it was important to investigate the asymptotical properties of proposed recognition algorithm. Let us notice, that experimental investigations were done for the case of two classes and for one–dimensional normal distributed patterns. The particular results were as follows:

A. The investigation of influence of changes of parameters of class distributions on the frequency of misclassification (mf).

In this test 30 sequences of length $n = 10$ of patterns, for which the sequence of classes forms a stationary and homogeneous second–order Markov chain, were generated. The transition probabilities of the chain were as follows:

$$\overline{p}(1,1,1) = 0.9 \quad \overline{p}(2,1,1) = 0.1 \quad \overline{p}(1,2,1) = 0.8 \quad \overline{p}(2,2,1) = 0.2$$

$$\overline{p}(1,1,2) = 0.1 \quad \overline{p}(2,1,2) = 0.9 \quad \overline{p}(1,2,2) = 0.2 \quad \overline{p}(2,2,2) = 0.8$$

First, the distance $||m_1 - m_2||$ between the mean values in the classes of distributions $f_1(x) = N(m_1, \sigma_1)$, $f_2(x) = N(m_2, \sigma_2)$ was changed, while $\sigma_1 = \sigma_2 = 3$. In the second test σ_2 was changed, while $\sigma_1 = 3, m_1 = 2, m_2 = 4$. The results are shown in Figures 1 and 2 respectively, where the frequencies of misclassification (mf) for particular pattern recognition algorithms were denoted by:

$mf3$ – for second–order Markov chains,

$mf2$ – for first–order Markov chains,

$mf1$ – for sequences of independent patterns.

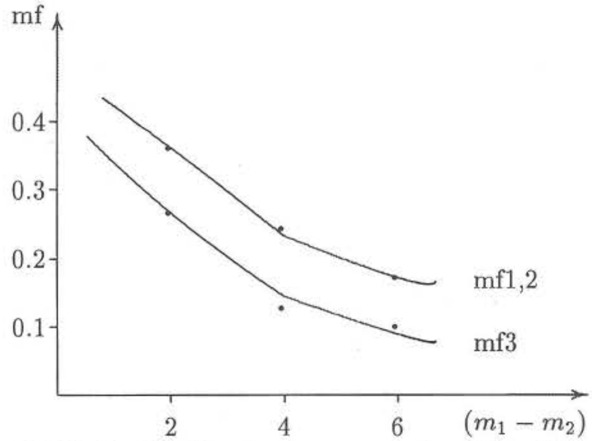

Figure 1. Misclassification frequency as a function of $\|m_1 - m_2\|$.

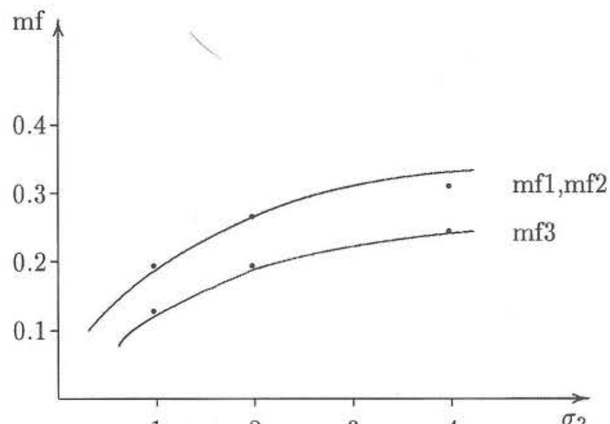

Figure 2. Misclassification frequency as a function of σ_2.

B. The investigation of influence of the transition probabilities of second-order Markov chain for "mf".

In this test only one sequence of 500 patterns from the classes with conditional density functions $f_1(x) = N_1(2;4)$, $f_2(x) = N_2(5;4)$ was generated. In the case of dichotomy for the evaluation of degree of second-order Markov dependency the special parameter γ:

$$\begin{aligned}\gamma &= |\overline{p}(1,1,1) - \overline{p}(1,1,2)| + |\overline{p}(1,1,1) - \overline{p}(1,2,1)| + \\ &+ |\overline{p}(1,1,1) - \overline{p}(1,2,2)| + |\overline{p}(1,1,2) - \overline{p}(1,2,1)| + \\ &+ |\overline{p}(1,1,2) - \overline{p}(1,2,2)| + |\overline{p}(1,2,1) - \overline{p}(1,2,2)|,\end{aligned} \tag{31}$$

was defined. Let us notice, that $\gamma = 0$ denotes that in the sequence of classes there does not exist dependency and that increasing of γ means increasing of degree of dependency ($\gamma \in [0,4]$). In these investigations, the parameter γ was changed and two cases of initial distribution of second-order Markov chain:

a) $p_2(1,1) = 0.4 \quad p_2(1,2) = 0.1 \quad p_2(2,1) = 0.1 \quad p_2(2,2) = 0.4$,

b) $p_2(1,1) = p_2(1,2) = p_2(2,1) = p_2(2,2) = 0.25$,

were taken into account. The results are shown in the Figures 3 and 4.

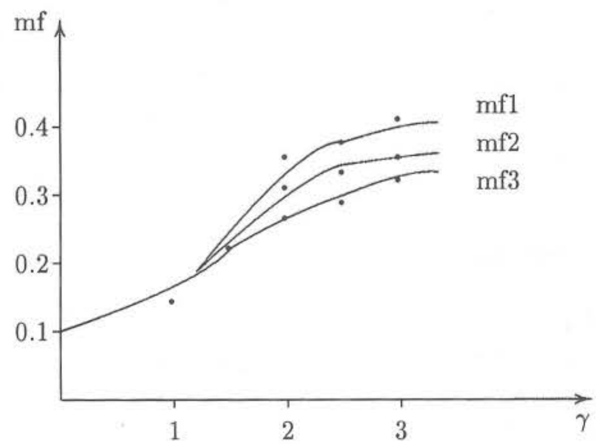

Fig.3. Misclassification frequency as a function of γ (case a).

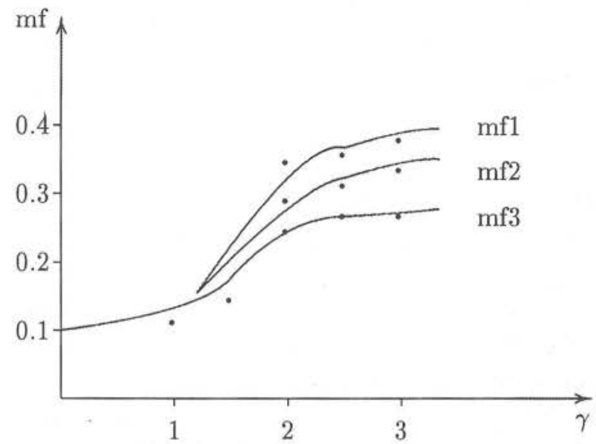

Fig.4. Misclassification frequence as a function of γ (case b).

C. The investigation of the influence of length of learning sequence for the misclassification frequency of the algorithm with learning. In this case, in the algorithm (16) the consistent Parzen estimator with the kernel function $K(x) = \exp(-x^2)/(2\pi)^{0.5}$ and sequence $h(n) = n^{-0.2}$ was used. For generation purpose, it was assumed that $f_1(x) = N(2;3)$ and $f_2(x) = N(8;3)$ and that the chain is homogeneous and stationary. In this investigation, first the test sequence of 50 patterns then the learning sequence of N patterns were generated. The results of investigation of influence of the length of learning sequence N for "mf" for different parameter γ are shown in the Figures 5 and 6, where by the horizontal lines the misclassification frequency of respective algorithms with the complete probabilistic information are shown.

5. Final remarks

In this paper the results of experimental investigation of pattern recognition algorithms for second–order Markov chains are presented. In the case of complete probabilistic information the comparison between the algorithms for second–order Markov chains, for first–order Markov chain and for the sequence of independent patterns (fig.1,2,3,4) was done. It shows us that taking into account the second–order Markov dependence (even in the case of not so strong degree of dependence i.e. for $\gamma > 1.5$), improves the quality of recognition in the cases when the sequences of patterns form such chain. Furthermore, in the case of the algorithm with learning, its asymptotical optimal properties were experimentally confirmed. Finally, one can say that the presented pattern recognition algorithms can be useful, and in the future they will be tested in some case of medical diagnosis.

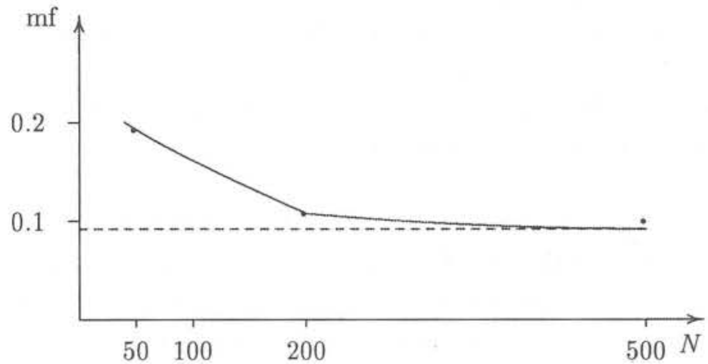

Fig.5. Missclassification frequency as a function of N ($\gamma = 2.7$).

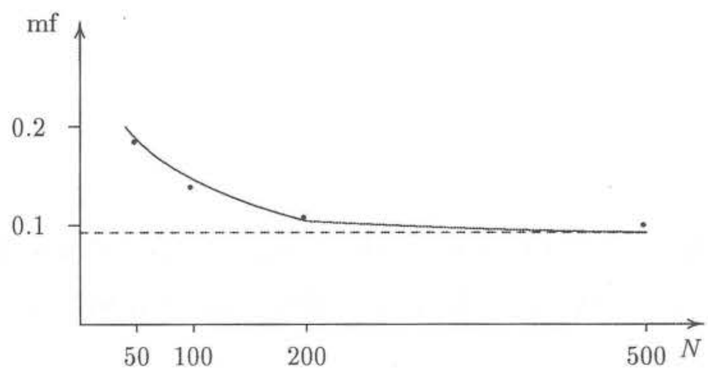

Fig.6. Missclassification frequency as a function of $N(\gamma = 3.2)$

References

[1] BUBNICKI Z., Algorytmy rozpoznawania dla prostych łańcuchów Markowa. Prace ICT Politechniki Wrocławskiej, Materiały i Studia, z.1, 1972.

[2] BUBNICKI Z., **Learning of Markov Chains Recognition and its Application to Biomedical Systems, Kooperative Systeme in Biologie and Technik,** *Oldenbourg Verlag,* München, 1980.

[3] BUBNICKI Z., Least interval pattern recognition algorithm and its application in control systems, IV IFAC Congress, no 21, 1969.

[4] GREBLICKI W., Pattern Recognition Procedures with Nonparametric Density Estimates, *IEEE Trans. on SMC,* vol. **8,** Nov.1978, pp.809–812.

[5] KURZYŃSKI M., Analiza porównawcza algorytmów rozpoznawania z uczeniem łańcuchów Markowa, Komunikat ICT Politechniki Wrocławskiej nr 110, 1974, (praca doktorska).

[6] KURZYŃSKI M., ŻOŁNIEREK A., A Recursive Classifying Decision Rule for Second-Order Markov Chains, *Control and Cybernetics,* vol. **9** pp.141–147, 1980.

[7] RAVIV J., Decision Making in Markov Chains Applied to the Problem of Pattern Recognition, *IEEE Trans. on IT,* vol. **IT–3,** Nov. 1967.

[8] ŻOŁNIEREK A., Pattern Recognition Algorithms for Controlled Markov Chains and Their Applications to Medical Diagnosis, *Pattern Recognition Letters,* **1,** pp. 299–303, 1983, North Holland.

[9] ŻOŁNIEREK A., Rozpoznawanie łańcuchów Markowa z zastosowaniem do diagnostyki medycznej, Raport ISiTS Politechniki Wrocławskiej, seria PRE nr 18, 1985, (Ph.D. dissertation).

Badania eksperymentalne algorytmów rozpoznania obrazów dla łańcuchów Markowa drugiego rzędu

Artykuł dotyczy zagadnień klasyfikacji, w których sekwencje rozpoznanych obrazów tworzą łańcuchy Markowa drugiego rzędu. Jako pierwszy przedstawiono algorytm dla przypadku pełnej informacji probabilistycznej, a następnie – odpowiedni algorytm z uczeniem. W dalszej części pokazano wyniki badań eksperymantalnych obu algorytmów.

Экспериментальные исследования алгоритмов распознавания образов для марковских цепей второго порядка

Статья касается вопросов классификации, в которых последовательности распознанных образов создают марковские цепи второго порядка. Первый алгоритм касается случая полной вероятностной информации, а в качестве второго представлен алгоритм с обучением. В следующей части работы рассмотрены результаты экспериментальных исследований обоих алгоритмов.