Control
and Cybernetics
VOL 9 (1980) No. 3

A recursive classifying decision rule for second--order Markov chains

by

MAREK KURZYŃSKI
ANDRZEJ ŻOŁNIEREK

Technical University of Wrocław
Institute of Technical Cybernetics
Wrocław, Poland

The paper deals with pattern recognition problems wherein there exist statistical dependences among the patterns to be recognized. As a mathematical model of this dependence a second-order Markov chain is adopted. Under assumption of complete statistical information and using Bayes' approach, the classifying decision rule which minimizes probability of misclassification is obtained. It is shown, that its discriminant functions can be recursively expressed and hence storage does not grow with the number of recognized patterns. Subsequently, the main result is extended to the higher-order Markov chains and to the case in which only learning sequence is available.

1. Introduction

In many pattern recognition problems there exist dependences among the patterns, to be recognized. For instance, this situation is typical for character recognition [7] recognition of state in technological processes [2], image classification [4], to name onlya few. Among the different concepts and methods of using „contextual" information in pattern recognition, an attractive from theoretical point of view and efficient approach is through Bayes' compound decision theory [10] in which a classifying decision is made on one pattern at a time, using additionally information from the entire past. Furthermore the assumption of Markov dependence among the patterns to be recognized is made. There is a great deal of available papers dealing with the recognition problems under assumption of a first-order Markov dependence. Based on this simplest model of statistical dependence, Raviv [8] derived decision rule, optimal with respect to a probability of misclassification. This result excellent was developed in [1], where recurrent form of recognition algorithm was presented and some decision rules with learning were proposed. Some next works in this area, in particular comparative analysis of different pattern recognition algorithms for first-order Markov chains and both theoretical and experimental studies of their properties can be found in [3, 5, 9].

In this paper the authors present maximum a posteriori probability decision rule under assumption of a second- order Markov dependence among the identities of recognized patterns. Subsequently it is shown, that its classifying functions can be recursively computed and hence storage does not grow with the number of recognized patterns.

2. Statement of the problem

Let us consider a problem of pattern recognition, that is concerned with the assignment of a given pattern to one of m known classes. Let x_n, taking values in the k-dimensional Euclidean space E^k, denotes the vector of measured features of n-th recognized pattern and j_n denotes the number of class to which the pattern in question belongs. Thus $\bar{x}_n \triangleq (x_1, x_2, \ldots, x_n)$ and $\bar{j}_n \triangleq (j_1, j_2, \ldots, j_n)$ state respectively feature vectors and true identities of a sequence of recognized patterns ($\triangleq$ signifies the defining equality).

Suppose, that x_n, and j_n are observed values of a couple of random variables (X_n, J_n) for $n=1, 2, \ldots$. Let j_n takes values in the set of consequtive integers $M \triangleq \triangleq \{1, 2, \ldots, m\}$. Subsequently suppose, that the sequence $J_1, J_2, \ldots, J_n, \ldots$ forms a second-order Markov chain [6], i.e.

$$P(J_n=j_n/J_{n-1}=j_{n-1}, J_{n-2}=j_{n-2}, \ldots, J_1=j_1)= \\ =P(J_n=j_n/J_{n-1}=j_{n-1}, J_{n-2}=j_{n-2}) \quad (1)$$

for all natural n and for every $j_1, j_2, \ldots, j_n \in M$. Notice, that for 2-nd order Markov dependence the initial probabilities:

$$p_{ij} \triangleq P(J_1=i,\ J_2=j), \quad i, j \in M, \quad (2)$$

and so-called trigram transition probabilities:

$$p^{(n)}_{i,j,k} \triangleq P(J_n=i/J_{n-1}=j,\ J_{n-2}=k), \quad i, j, k \in M\ n \geqslant 3 \quad (3)$$

determine any finite-dimensional distribution of random variables $\{J_n\}$, $n=1, 2, \ldots$ [6].

Let

$$f_i(x_n) \triangleq f(x_n/i) \quad (4)$$

be the conditional probability density function of X_n, given that $J_n=i$, $i \in M$, identical for all natural n. Suppose also, that probabilities (2), (3) and density functions (4) which determine the distribution of couples (X_n, J_n) $n=1, 2, \ldots$ are given. It states, that in this paper the case of complete probabilistic information is considered. For simplicity, suppose additionally conditional independence among the random variables X_n, $n=1, 2, \ldots$, which implies that:

$$f_n(\bar{x}_n/\bar{j}_n) = \prod_{i=1}^{n} f(x_i/j_i), \quad n=1, 2, \ldots, \quad (5)$$

where f_n denotes joint conditional density function of $\bar{x}_n$. This assumption states that, given the true identity of a pattern, the distribution of a measurement vector is independent of the features and true identities of previous and future patterns, but it is dependent only on the true identity of the pattern in question.

3. The pattern recognition algorithm

Let $L(i,j)$ be the loss incurred by the classifier, if a pattern from the class j is assigned to class i. It is well-known [8], that Bayes' decision rule:

$$i_n=\Psi^*(\bar{x}_{n-1}, x_n)\triangleq\Psi^*(x_n), \quad n=1, 2, \ldots \tag{6}$$

minimizes risk i.e. expected loss at the n-th stage of classification:

$$R[\Psi_n(\cdot)]\triangleq \underset{\bar{X}_n, J_n}{E}\{L(i_n, j_n)\}=\sum_{j_n=1}^{m}\int_{\chi_n} L(i_n, j_n)\, p(j_n)\, f(\bar{x}_n/j_n)\, d\bar{x}_n, \tag{7}$$

where $p(j_n)=P(J_n=j_n)$, $\chi_n\triangleq E^k\times E^k\times\ldots\times E^k$ (n times) and classifies n-th recognized pattern, given by measurement vector x_n to class i_n for which the conditional risk:

$$r(i_n, \bar{x}_n)\triangleq \underset{J_n/\bar{x}_n}{E}\{L(i_n, j_n)\}=\sum_{j_n=1}^{m} L(i_n, j_n)\, p(j_n/\bar{x}_n) \tag{8}$$

is the least one.

For the special case of $0-1$ loss function, i.e.

$$L(i,j)=\begin{cases}0 & \text{if } i=j\\ 1 & \text{otherwise,}\end{cases}$$

the rule (6) assigns the n-th pattern to the class with the highest a posteriori probability after observing $\bar{x}_n$ for all natural n, i.e.

$$\Psi_n^*(x_n)=i_n \text{ if} \tag{9}$$
$$p(i_n/\bar{x}_n)>p(s/\bar{x}_n)$$

for every $s\neq i_n$, $s, i_n\in M$.

The Bayes' risk (the minimum attainable risk) associated with above rule reduces to the minimum probability of error at the n-th stage of classification, which is given by

$$R_n^*=R[\Psi_n^*(x_n)]=1-\underset{\bar{x}_n}{E}\max_{i\in M} p(i/\bar{x}_n), \tag{10}$$

where expectation is taken with respect to the mixed (unconditional) distribution of $\bar{X}_n\triangleq(X_1, X_2, \ldots, X_n)$.

Now our objective is to calculate decision functions of (9) and to express them in terms of known quantities. From Bayes' formula we obtain:

$$p(j_n/\bar{x}_n)=\frac{f(\bar{x}_n/j_n)\, p(j_n)}{f(\bar{x}_n)}, \quad n=1, 2, \ldots, \; j_n\in M. \tag{11}$$

Since mixed density function $f(\bar{x}_n)$ is independent of j_n, we can maximize only the following discriminant functions:

$$g(j_n, \bar{x}_n) \triangleq p(j_n) f(\bar{x}_n/j_n) = f(\bar{x}_n)\, p(j_n/\bar{x}_n), \quad n=1, 2, \ldots, \quad j_n \in M. \tag{12}$$

Define functions:

$$g(j_n, j_{n-1}, \bar{x}_n) \triangleq p(j_n, j_{n-1}) \cdot f(\bar{x}_n/j_n, j_{n-1}), \quad n=2, 3, \ldots, \quad j_n, j_{n-1} \in M, \tag{13}$$

where $p(j_n, j_{n-1}) = P(J_n = j_n, J_{n-1} = j_{n-1})$ and notice, that they can be calculated recursively as follows. Using the fact, that transition probabilities (3) are independent of $\bar{x}_{n-1}$ and considering assumption (5) we have:

$$g(j_n, j_{n-1}, \bar{x}_n) = f(\bar{x}_n/j_n, j_{n-1}) \cdot p(j_n, j_{n-1}) = f(x_n, \bar{x}_{n-1}/j_n, j_{n-1}) \times p(j_n, j_{n-1}) =$$

$$= f(x_n/j_n, j_{n-1}, \bar{x}_{n-1}) \cdot f(\bar{x}_{n-1}/j_n, j_{n-1})\, p(j_n, j_{n-1}) = f(x_n/j_n)\, g(j_n, j_{n-1}, \bar{x}_{n-1}) =$$

$$= f_{j_n}(x_n) \sum_{j_{n-2}=1}^{m} p(j_n, j_{n-1}, j_{n-2}/\bar{x}_{n-1}) f(\bar{x}_{n-1}) = f_{j_n}(x_n) \sum_{j_{n-2}=1}^{m} p(j_n/j_{n-1}, j_{n-2}) \times$$

$$\times p(j_{n-1}, j_{n-2}/\bar{x}_{n-1}) f(\bar{x}_{n-1}) = f_{j_n}(x_n) \sum_{j_{n-2}=1}^{m} p^{(n)}_{j_n, j_{n-1}, j_{n-2}} \cdot g(j_{n-1}, j_{n-2}, \bar{x}_{n-1}), \tag{14}$$

for all natural $n \geqslant 3$ and for every $j_n, j_{n-1} \in M$, with initial condition:

$$g(j_2, j_1, \bar{x}_2) = f_{j_1}(x_1) \cdot f_{j_2}(x_2)\, p_{j_1, j_2}, \quad j_1, j_2 \in M. \tag{15}$$

From (12) and (13) we see that:

$$g(j_n, \bar{x}_n) = f(\bar{x}_n) \cdot p(j_n/\bar{x}_n) = \sum_{j_{n-1}=1}^{m} p(j_n, j_{n-1}/\bar{x}_n) f(\bar{x}_n) =$$

$$= \sum_{j_{n-1}=1}^{m} g(j_n, j_{n-1}, \bar{x}_n), \quad n \geqslant 2, \quad j_n, j_{n-1} \in M, \tag{16}$$

and therefore the knowledge of $g(j_n, j_{n-1}, \bar{x}_n)$ for $n \geqslant 2$, $j_n, j_{n-1} \in M$ allows to calculate (12) for all $n \geqslant 2$ and for every $j_n \in M$ and consequently suffices to construction of Bayes' decision rule (9).

At the first step of classification (for $n=1$) the discriminant functions can be calculated immediately, namely:

$$g(j_1, x_1) = f_{j_1}(x_1) \sum_{j_2=1}^{m} p_{j_1, j_2}, \quad j_1 \in M. \tag{17}$$

Thus, the formulas (14)÷(16) determine the recursive manner of calculation (12) with (17) as initial condition. For the general form of loss function $L(i, j)$, discriminant functions (8) of decision rule (6) are equal:

$$r(i_n, \bar{x}_n) = \frac{1}{f(\bar{x}_n)} \sum_{j_n=1}^{m} L(i_n, j_n) \cdot g(j_n, \bar{x}_n), \quad n=1, 2, \ldots, \quad i_n \in M \tag{18}$$

and they can be also calculated recursively as it was shown previously.

4. Extension to the k-th order Markov chains

The recursive technique of construction Bayes' decision rule can be straightforwardly extended to the higher-order Markov dependence. Generally, k-th order Markov chain is described by:

— initial probabilities

$$p_{j_1, j_2, \ldots, j_k} \triangleq P(J_1=j_1, J_2=j_2, \ldots, J_k=j_k), \quad j_1, j_2, \ldots, j_k \in M, \tag{19}$$

— transition probabilities

$$p^{(n)}_{j_n, j_{n-1}, \ldots, j_{n-k}} \triangleq P(J_n=j_n/J_{n-1}=j_{n-1}, \ldots, J_{n-k}=j_{n-k}), \quad j_n, j_{n-1}, \ldots, j_{n-k} \in M,\ n>k \tag{20}$$

In this instance, similary as in (14) and under the same assumptions we have:

$$g(j_n, j_{n-1}, \ldots, j_{n-k+1}, \bar{x}_n) = f(x_n/j_n) \sum_{j_{n-k}=1}^{m} p^{(n)}_{j_n, j_{n-1}, \ldots, j_{n-k}} \times \times g(j_{n-1}, j_{n-2}, \ldots, j_{n-k}, \bar{x}_{n-1}) \tag{21}$$

for all natural $n>k$ and for every $j_n, j_{n-1}, \ldots, j_{n-k} \in M$, where functions g are defined likewise as previously by:

$$g(j_n, j_{n-1}, \ldots, j_{n-k}, \bar{x}_n) \triangleq f(\bar{x}_n/j_n, j_{n-1}, \ldots, j_{n-k})\, p(j_n, \ldots, j_{n-k}). \tag{22}$$

For $n=k$ we have the following initial condition:

$$g(j_k, j_{k-1}, \ldots, j_1, \bar{x}_k) = \prod_{i=1}^{k} f_{j_1}(x_i)\, p_{j_1, j_2, \ldots, j_k} \quad j_1, j_2, \ldots, j_k \in M. \tag{23}$$

The discriminant functions (12) of decision rule (9) are related to (21) by:

$$g(j_n, \bar{x}_n) = \sum_{j_{n-1}=1}^{m} \sum_{j_{n-2}=1}^{m} \cdots \sum_{j_{n-k+1}=1}^{m} g(j_n, j_{n-1}, \ldots, j_{n-k+1}, \bar{x}_n). \tag{24}$$

For the k first steps of classification the discriminant functions (12) can be calculated immediately.

5. Final remarks

This paper deals with classification problem in which the sequence of recognized patterns is generated by a second-order Markov source, i.e. each pattern is a probabilistic function of its two immediate predecessors. For this case, under assumption of complete probabilistic information and using Bayes' approach, the pattern

recognition algorithm which minimizes probability of misclassification has been obtained. We have also shown that its discriminant functions at the n-th stage (for n-th recognized pattern in the sequence of patterns) can be expressed as functions of known data and the classifying functions at the $(n-1)$ th stage, i.e. they can be recursively computed. The recursive nature of the algorithm considerably reduces the memory space and the computation time. Subsequently, this main result has been extended to the higher-order Markov chains.

In the real world, there is often a lack of the exact knowledge of probabilities (2), (3) and density functions (4), whereas only partial information is available. For instance, there are situations in which only learning sequences, that is sets of correctly classified samples, are known. In this case one obvious and conceptually simple method is to estimate probabilities (2), (3) and densities (4) from the training sample set and then to use these estimators to calculate discriminant functions (12) according to (14) ÷ (17) as though they were correct. Similar procedure was employed in [5] for first-order Markov chains.

References

[1] BUBNICKI Z. Algorytmy rozpoznawania dla prostych łańcuchów Markowa, *Prace Instytutu Cybern. Techn. Politechniki Wrocławskiej, Seria: Studia i Materiały*, Nr 1, 1972 r.

[2] BUBNICKI Z. Least interval pattern recognition algorithm and its application to control systems. Materiały IV Kongresu IFAC, nr 21, Warszawa 1969 r.

[3] CHU J. T. Error bounds for contextual recognition procedure. *IEEE Trans. on Computers*, **20**, October 1971, 1203–1207.

[4] FU K. S. Syntactic methods in pattern recognition. New York, Academic Press, 1974.

[5] HASIEWICZ Z., KURZYŃSKI M. Algorytmy rozpoznawania z uczeniem dla prostych łańcuchów, Markowa. Prace VI KKA tom 1, Poznań 1974.

[6] Хеннекен П.А., Тортра А. Теория вероятностей и некоторые её приложения. Москва, Изд. Наука, 1974.

[7] NEUHOFF D. L., The Viterbi algorithm as an aid in text recognition. *IEEE Trans. on Information Theory*, **21**, April 1975, 222–226.

[8] RAVIV J. Decision making in Markov chain applied to the problem of pattern recognition. *IEEE Trans. on Information Theory*, **23**, October 1967, 536–551.

[9] REJTÖ L. Pattern recognition in Markovian case. *Problems on Control and Information Theory*, **5** (1976) 2, 135–147.

[10] TOUSSAINT G. T. The use of context in pattern recognition. *Pattern Recognition*, **10**, (1976) 3, 189–204.

Received, February 1980.

Rekurencyjny algorytm rozpoznawania łańcuchów Markowa II-rzędu

Praca dotyczy problemów rozpoznawania obrazów, w przypadku gdy występują statystyczne zależności pomiędzy ich klasyfikacjami. Jako matematyczny model tej zależności przyjęto łańcuch Markowa II-rzędu. Przy założeniu pełnej informacji probabilistycznej skonstruowano analitycznie

ciąg algorytmów rozpoznawania, z których i-ty minimalizuje prawdopodobieństwo błędnej klasyfikacji na i-tym etapie (dla i-tego rozpoznawanego obrazu). Pokazano, że funkcje klasyfikujące tych algorytmów mogą być wyznaczone rekurencyjnie, co znacznie upraszcza obliczenia i skraca ich czas. Rezultat ten uogólniono następnie na przypadek łańcuchów Markowa k-tego rzędu.

Рекуррентный алгоритм распознавания цепей Маркова второго порядка

Работа касается проблем распознавания образов в случае, когда между их классификациями выступает статистические зависимости. Как математическую модель этих зависимостей принято цепь Маркова второго порядка. При полной статистической информации построено ряд алгоритмов распознавания. $И$-тый алгоритм минимализирует вероятность ошибочных классификаций на $и$-тым шагу (для $и$-того распознаваемого образа). Показано, что дискриминационные функции могуть быть определены рекуррентно, что значительно уменьшает сложность вычислений и сокращает их время. В работе представлено обобщение этих результатов на случай Марковской цепи $к$-того порядка.