DIXIEME COLLOQUE SUR LE TRAITEMENT DU
SIGNAL ET SES APPLICATIONS

NICE du 20 au 24 MAI 1985

A COMPARATIVE STUDY OF DECISION MAKING ALGORITHMS IN HIDDEN MARKOV CHAINS

Pierre A. Devijver

Philips Research Laboratory, Av. Em. Van Becelaere 2, B-1170 Brussels, Belgium

RESUME

Résumé: Cet article est consacré à une étude expérimentale comparative de trois méthodes de détection Bayesienne qui prennent en compte une information contextuelle modélisée par une chaine de Markov du premier ordre observée dans un bruit aléatoire. Le critère utilisé est celui de la minimisation de la probabilité d'erreur par symbole. Les méthodes envisagées ressortissent de la théorie statistique de la décision composite appliquée à des séquences finies d'observations.

De manière plus précise, les trois méthodes sont

i) la technique séquentielle composite (temps-réel) fondée sur la maximisation de la vraisemblance du signal compte tenu de la séquence des observations passées et présentes;

ii) la technique séquentielle composite à décision différée dans laquelle la décision est effectuée avec un retard fixe par rapport à l'observation;

iii) la technique composite globale fondée sur la maximisation de la vraisemblance du signal compte tenu de la séquence complète des observations.

Ces techniques ont été fréquemment utilisées dans le domaine de la reconnaissance de texte et de la parole et se sont montrées aussi d'une efficacité intéressante dans le domaine du traitement d'images. Toutefois, la littérature abonde en constatations discordantes quant à leurs performances et complexités respectives. C'est la question que nous nous proposons d'élucider.

Nous montrons que dans des conditions expérimentales normales, la méthode à décision différée offre sans conteste le meilleur compromis performance–complexité, et n'est supplantée par la méthode composite globale que dans des cas pathologiques peu susceptibles d'être rencontrés dans les applications.

Une explication intuitive de ce phénomène est proposée.

SUMMARY

Abstract: This paper reports and comments on experimental results obtained with three contextual decision rules for minimum probability of error per decision operating under the first-order Markov chain assumption for the sequence of class labels observed in memoryless noise. The rules belong in the realm of the statistical decision theory and are characterized by the range of their look-ahead capabilities.

Specifically, the rules we are interested in are:

i) the sequential compound decision rule based on the maximization of the joint likelihood of the signal and the sequence of past and present measurements;

ii) the sequential compound rule with fixed-lag look-ahead in which the decision is postponed until a fixed number of measurements have been acquired;

iii) the overall compound rule based on the maximization of the joint likelihood of the signal and the entire sequence of measurements.

These techniques have been extensively applied to text and speech recognition and are currently being explored with image processing applications in mind. However, the literature abounds with contoversial statements regarding their relative merits and drawbacks. It is the issue which is addressed here.

It is found, that under normal experimental conditions, one-step look-ahead offers the best tradeoff between performance and complexity. The forward-backward algorithm which constantly takes full look-ahead into account, is found superior only in rather pathological situations which are very unlikely to ever be encountered in practical applications.

An intuitive justification of these findings is proposed.

1. Introduction

The assumption of Markov dependence among pattern classes is one of the most effective ways of using contextual information in pattern recognition [1]. In addition to the information conveyed by the class-conditional probability distributions of feature vectors, contextual decision making algorithms exploit the information encoded in the probability distribution governing the temporal sequence of pattern-class identities. In this paper, we concentrate on algorithms which exploit both these sources of information for the purpose of achieving minimum probability of error *per decision*. (The Viterbi algorithm and its many variants [2] are excluded from consideration as they are designed to achieve minimum probability of error *per sequence* of decisions. Moreover, they are fairly well documented in the literature [3]).

Minimum probability of error under the Markov chain assumption can be achieved with three kinds of algorithms which can be characterized by their *look-ahead* capabilities. To be precise, let $\{X^1, \cdots, X^\tau, \cdots, X^T\}$ designate an ordered sequence of observed feature vectors, and $\{\omega^1, \cdots, \omega^\tau, \cdots, \omega^T\}$ designate the corresponding ordered sequence of unknown, pattern-class labels or interpretations.

Algorithms in the first category use no look-ahead. Feature measurement and decision making are synchronized. By this we mean that the decision on ω^τ is made using all the past feature vectors $X^1, \cdots, X^{\tau-1}$ and the current feature vector X^τ. A classical result in the statistical theory of compound decisions states that, for feature vectors observed in memoryless noise (see Sec. 2.1) minimum error probability per decision is achieved by the *sequential compound* decision rule of (1)

$$\begin{aligned} &\hat{\omega}^\tau = \omega_i \quad \text{if} \\ &\omega_i = \text{argmax}_{\omega_j} P\{\omega^\tau = \omega_j, X^1, \cdots, X^\tau\}, \qquad \tau = 1, \cdots, T \end{aligned} \tag{1}$$

where $\hat{\omega}^\tau$ designates the 'estimate' of ω at time τ and ω_i is one of the c possible class labels $\omega_1, \cdots, \omega_c$.

Algorithms in the second category use some fixed look-ahead. With n-step look-ahead, the decision on ω^τ is postponed until the $(\tau + n)$th feature vector has been acquired. With T fixed, the decision rule for minimum error probability is given by (2).

$$\hat{\omega}^\tau = \omega_i \quad \text{if}$$

$$\omega_i = \begin{cases} \text{argmax}_{\omega_j} P\{\omega^\tau = \omega_j, X^1, \cdots, X^{\tau+n}\}, & 1 \le \tau \le T - n, \\ \text{argmax}_{\omega_j} P\{\omega^\tau = \omega_j, X^1, \cdots, X^T\}, & \tau > T - n. \end{cases} \tag{2}$$

In what follows, we concentrate on $n = 1$ for reasons that will soon become apparent.

Algorithms in the third category use the largest possible look-ahead by postponing *any* decision until the entire sequence $X^1, \cdots, X^T$ has been acquired. The corresponding *compound* decision rule is given by (3).

$$\begin{aligned} &\hat{\omega}^\tau = \omega_i \quad \text{if} \\ &\omega_i = \text{argmax}_{\omega_j} P\{\omega^\tau = \omega_j, X^1, \cdots, X^T\}, \quad \tau = 1, \cdots, T. \end{aligned} \tag{3}$$

It is a remarkable result that, under the Markov chain assumption for the sequence of class labels, the probabilities in (1)–(3) can be computed in linear time.

The decision rule of (1) was first proposed independently by Raviv [4] and Abend [5]. Since then, it was used by too many researchers to be cited here. The rule of (2) was also proposed by Raviv [4]. In spite of Raviv's experimental results showing a considerable improvement over (1), the look-ahead technique does not seem to have been widely used (and goes unnoticed in Haralick's recent review [1]). Quite surprisingly, the decision rule of (3) does not seem to have appeared in the open pattern recognition literature until the brief outline by Haralick [1] of the unpublished BAMPS algorithm of Lehan. A different formulation of essentially the same algorithm can be found in Devijver [6]. However, (3) has been applied for more than a decade in information theory circles where it is used for optimal decoding purposes (see, *e.g.*, Bahl *et. al.* [6]). It is closely related to the work of Baum [7] which has led to quite successful learning methods [8], particularly in the field of speech recognition [9–12].

It is evident that the amount of information used by the decision rule increases with the look-ahead range. Intuitively, one should expect the performance to improve accordingly. It goes without saying that improved performance can be achieved only at the cost of increased time and space complexities. In the case of (3), the increase is quite substantial as we shall see in Sec. 2. However, our previous experience with these algorithms had indicated that, on a comparative basis, the performance of the compound decision rule did not level up to our expectation. In other words, the substantial effort involved in implementing (3) used not to return very high dividend. Therefore it was the purpose of the computer simulation reported here to elucidate the question of whether our deceiving results were to be attributed to a lack of accuracy of the assumed Markovian model or to the intrinsic behavior of the compound decision rule.

Our experiments provide a clear-cut answer. They give every indication that under "normal experimental conditions" the assumption of a first order Markov model does prevent us from using the information in the sequence of future measurements $X^{\tau+1}, \cdots, X^T$ beyond that which is encoded in the distribution of $X^{\tau+1}$. Significant improvement over the one-step look-ahead mode of decision occurred only in fairly pathological situations which are most unlikely to be encountered in practice.

A convenient feature of the Markov model is that it readily enables the computer simulation of stationary random sequences with prescribed temporal dependence. The Markov chain model is formalized in Sec. 2.1. The parametric family of Markov sources used in our experiments is specified in Sec. 2.3. A brief outline of the algorithms used to implement (1)–(3) can be found in Sec. 2.2. Experimental results are presented in Sec. 3 for various values of the Markov source entropy and the (non-contextual) signal-to-noise ratio. An intuitive justification of our experimental findings is attempted in Sec. 4.

2. Models and Algorithms

2.1. The theoretical model

We assume that the pattern-class generating mechanism can be modeled by a discrete parameter, discrete time, first order, homogeneous Markov chain with state space $\{\omega_1, \cdots, \omega_c\}$, [13]. We write $\omega^\tau = \omega_i$ to indicate that the process is in state ω_i at time τ. The Markov chain is specified in terms of an *initial*

state distribution $P_i = P\{\omega^1 = \omega_i\}$, $i = 1, \ldots, c$ and a matrix of stationary *state transition probabilities*

$$P_{ij} = P\{\omega^{\tau+1} = \omega_j | \omega^\tau = \omega_i\},$$

for $1 \leq i, j \leq c$,and $1 \leq \tau \leq T-1$. The Markov property yields the factorization

$$P\{\omega^1, \cdots, \omega^T\} = P\{\omega^1\} \prod_{\tau=1}^{T-1} P\{\omega^{\tau+1} | \omega^\tau\}, \tag{4}$$

The random process associated with the states is represented by c probability distributions $p_j(X) = p(X|\omega_j)$, $1 \leq j \leq c$. We make the assumption that $X^1, \cdots, X^T$ are state-conditionally independent, or that X's are observed in memoryless noise. This assumption yields a second factorization, vis.,

$$P(X^1, \cdots, X^T | \omega^1, \cdots, \omega^T) = \prod_{\tau=1}^{T} p(X^\tau | \omega^\tau). \tag{5}$$

In what follows, we assume that the initial and transition probabilities of the Markov chain as well as the state-conditional distributions are known to us.

2.2. The algorithms

In this section, we adopt the very elegant formulation of Baum [7]. The reader should be warned that it may not lead to the most efficient implementation. (See [10] for consideration of implementation details in the framework of the mixture identification problem.)

Under the assumptions just introduced, the likelihood L of a X-sequence of observations $X^1, \cdots, X^T$ is given by

$$\begin{aligned} L &= p(X^1, \cdots, X^T) \\ &= \sum_{\omega^1, \cdots, \omega^T = \omega_1}^{\omega_c} P(\omega^1) p(X^1 \mid \omega^1) \\ &\times \prod_{\tau=1}^{T-1} P(\omega^{\tau+1} \mid \omega^\tau) p(X^\tau \mid \omega^\tau) \\ &= \sum_{i_1, \cdots, i_T = 1}^{c} P_{i_1} p_{i_1}(X^1) \prod_{\tau=1}^{T-1} P_{i_\tau i_{\tau+1}} p_{i_{\tau+1}}(X^{\tau+1}). \end{aligned} \tag{6}$$

This follows readily from (4) and (5).

Let $\mathcal{F}_\tau(i) \doteq P(\omega^\tau = \omega_i, X^1, \cdots, X^\tau)$, $i = 1, \ldots, c$. Thus, $\mathcal{F}_1(i) = P_i p_i(X^1)$, and $\mathcal{F}_\tau(i)$ can be computed inductively forward by the recurrence

$$\mathcal{F}_\tau(i) = \begin{cases} P_i p_i(X^1) & \text{for } \tau = 1, \\ \sum_{j=1}^{c} \mathcal{F}_{\tau-1}(j) P_{ji} p_i(X^\tau) & \text{for } 2 \leq \tau \leq T. \end{cases} \tag{7}$$

Let $\mathcal{B}_\tau(i) \doteq P(X^{\tau+1}, \cdots, X^T \mid \omega^\tau = \omega_i)$, $i = 1, \ldots, c$ and $\mathcal{B}_T(i) = 1$, $\forall i$. Then, $\mathcal{B}_\tau(i)$ can be computed inductively backward by the recurrence

$$\mathcal{B}_\tau(i) = \begin{cases} 1 & \text{for } \tau = T, \\ \sum_{j=1}^{c} P_{ij} p_j(X^{\tau+1}) \mathcal{B}_{\tau+1}(j) & \text{for } T-1 \geq \tau \geq 1. \end{cases} \tag{8}$$

Now, the remarkable thing about these relationships is that [7]

$$L = \sum_{i=1}^{c} \mathcal{F}_\tau(i) \mathcal{B}_\tau(i) \tag{9}$$

identically in τ. The proof that (9) is equivalent to (6) involves nothing more than the distributive law.

Let us note that $\mathcal{F}_\tau(j)$ is the probability required for applying the decision rule of (1). Thus (7) describes one possible implementation of the sequential compound algorithm. On the other hand, it is readily seen that $\mathcal{F}_\tau(j)\mathcal{B}_\tau(j)$ is the probability required for applying (3). Thus, one computational scheme for implementing the compound decision rule amounts to:

i) a forward stage which consists in computing the $\mathcal{F}$ values using (7); these probabilities have to be stored for use during

ii) the backward stage which consists in computing the $\mathcal{B}$ values using (8) and forming the products $\mathcal{F}_\tau(j)\mathcal{B}_\tau(j)$ which are then used in (3).

This technique is often referred to as the "forward-backward" algorithm. From now on, we shall adhere to this convention. It is evident that the forward-backward algorithm is twice as costly as the sequential compound one. Moreover, it requires extra storage for $c(T-1)$ $\mathcal{F}$ values.

Let $\mathcal{G}_\tau(i) \doteq P\{\omega^\tau = \omega_i, X^1, \cdots, X^{\tau+1}\}$, $i = 1, \ldots, c$, $\tau = 1, \ldots, T-1$, as required by (2) for $n = 1$. Readily, the one-step look-ahead technique can be implemented by the recurrence

$$\mathcal{G}_\tau(i) = \begin{cases} \mathcal{F}_\tau(i) \sum_{j=1}^{c} P_{ij} p_j(X^{\tau+1}) & \text{for } \tau = 1, \ldots, T-1, \\ \mathcal{F}_\tau(i) & \text{for } \tau = T. \end{cases} \tag{10}$$

The one-step look-ahead technique is slightly more costly than the sequential compound algorithm and requires temporary storage for c $\mathcal{F}$ values. At the risk of belaboring the obvious, let us point out that—with or without look-ahead—the sequential compound algorithm can be implemented in real time while the forward-backward algorithm must be implemented off line.

2.3. The experimental model

In the experiments described hereafter, the Markov source was selected to be a simple, cyclic version of the Bakis model of speech production [14]. Specifically, we adopted a 6-state model with transition probabilities parameterized by $\mathbf{p}$, $0 \leq \mathbf{p} \leq 1$, according as

$$P_{i_\tau i_{\tau+1}} = \begin{cases} \frac{1-\mathbf{p}}{2} & \text{for } i_{\tau+1} = i_\tau \\ \mathbf{p} & \text{for } i_{\tau+1} = i_\tau + 1 \bmod 6 \\ \frac{1-\mathbf{p}}{2} & \text{for } i_{\tau+1} = i_\tau + 2 \bmod 6 \end{cases} \tag{11}$$

$\forall i_\tau \in \{1, \ldots, 6\}$. The stationary distribution of this source is an equiprobable distribution over all states. It was also taken as an initial distribution in order to avoid disruptive edge effects. Given $\mathbf{p}$, the entropy of the source is given by

$$H = (1-\mathbf{p}) - \mathbf{p} \log \mathbf{p} - (1-\mathbf{p}) \log(1-\mathbf{p}), \tag{12}$$

where logarithms are to the base of 2 and H is given in bits. The entropy function $H(\mathbf{p})$ is shown in Figure 1.

The generation of a pseudo-random Markov sequence was performed as follows:

i) An initial state was selected randomly with equal a priori probability for each possible state.

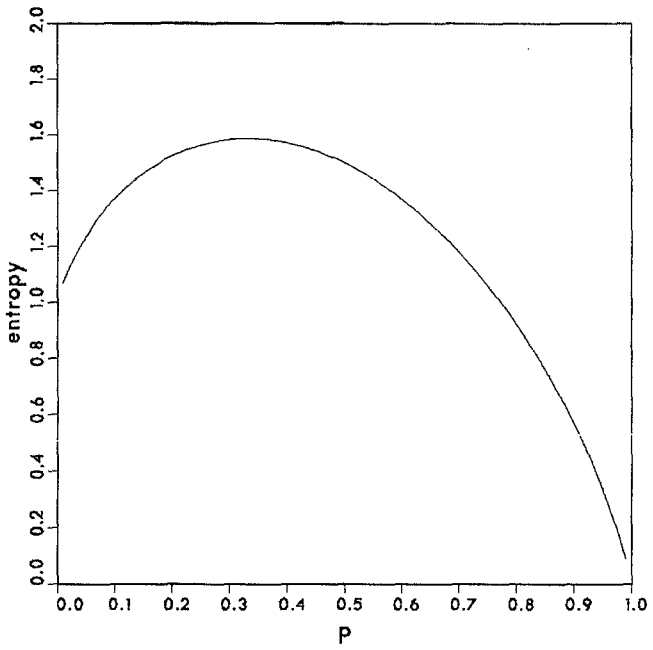

Figure 1: Entropy function for the experimental Markov source.

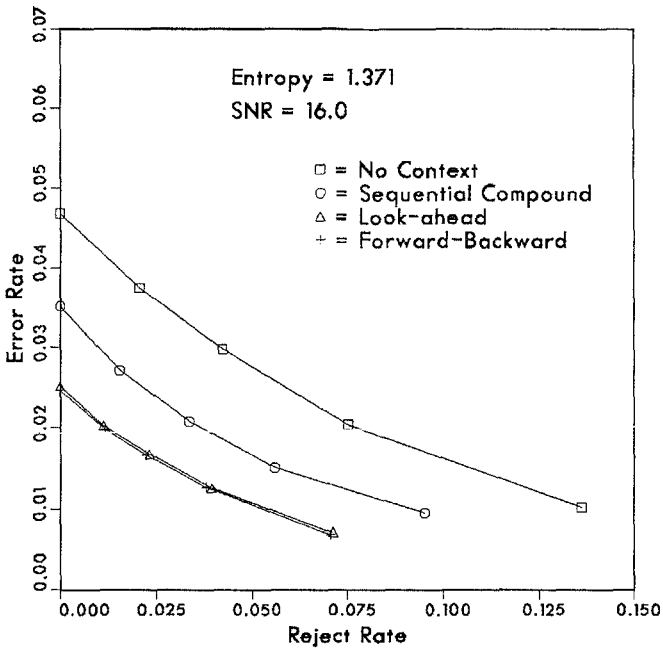

Figure 2: Error-reject curves, high entropy, high SNR.

ii) The selection of subsequent states was performed by partitioning the unit interval proportionally to the component of the row of the transition matrix indexed by the previous state, generating a randon number, and chosing the new state according to the subinterval in which the number felt.

The state-conditional distributions were chosen to be 2-dimensional normal, *i.e.*, $p_i(X) \sim N(\mu_i, I)$. I is the 2×2 identity matrix. Mean vectors μ_i's were located at the vertices of a regular hexagon inscribed within a circle of radius R. Ignoring context, the signal to noise ratio (SNR) was measured by the ratio of the traces of the between- and within-class covariance matrices. In our configuration, SNR $= R^2$. Generation of pseudo-random bivariate normal data was performed in the standard way.

From this choice for an experimental model, it turns out that experimental results can be characterized by only two parameters, namely $\mathbf{p}$ and R, or equivalently, $H(\mathbf{p})$ and SNR.

3. Experimental Results

Each experiment consisted in generating 10^3 random sequences of length $T = 10$ according to the methods described in Sec. 2.3. [Experiments with longer sequences did not produce noticeably different results. It should be noted that all three algorithms perform in exactly the same way in classifying the last element in each sequence. As the results reported are obtained by averaging the errors over all elements, short sequences induce a sligth comparative bias which is detrimental to the best algorithm, namely, forward-backward.] Classifications were performed using the decision rules of (1)–(3) combined with the algorithms of Sec. 2.2. For comparison purposes, classification was also performed with a non-contextual Bayes rule. Classification results are illustrated in the form of error-reject curves [15].

A typical example of the results obtained is illustrated by Fig. 2. In this example, the SNR was moderately high, SNR = 16, which corresponds to a minimal error probability less that 0.05 when the Markov source information is not used and the reject option disabled. The entropy of 1.371 is also moderately high, ($\max_{\mathbf{p}} H(\mathbf{p}) = 1.585$), though the source is already quite skewed with a probability $\mathbf{p} = 0.6$ of direct transition to the next state in the cyclically ordered sequence of states.

The curves shown in Figure 2 illustrate the effectiveness of contextual decision rules—something Raviv [4] had already done long ago—but above all, they show the definite failure of the forward-backward algorithm to improve in any significant way over the performance achieved under the one-step look-ahead mode for the particular model selected in this experiment. However, as stated above, the results in Figure 2 are a typical example of the behavior of the three contextual rules over a wide range of values for the parameters $\mathbf{p}$ and R.

The effectiveness of the forward-backward technique becomes apparent only at low signal to noise ratios. Figure 3 displays the results of an experiment in which the entropy was the same as in the previous example while the SNR was lowered down to 5.76 (for an error-rate of the order of 0.23 for the non-contextual rule with no reject). It is plain, however, that the improvement achieved by the forward-backward algorithm still belongs in the realm of "second-order" effects.

It is worth emphasizing that, however small, differences between performance figures are statistically significant due to high correlation. In actual fact, minute examination of individual errors revealed that these differences arise from the more powerful algorithm being able to correct some of the errors incurred by the less powerful one. In this sense, sets of individual errors are nested.

Finally, we were able to lay bare the potentiality of the forward-backward algorithm by retaining the low value of 5.76 for the SNR as in the previous example and turning the Markov

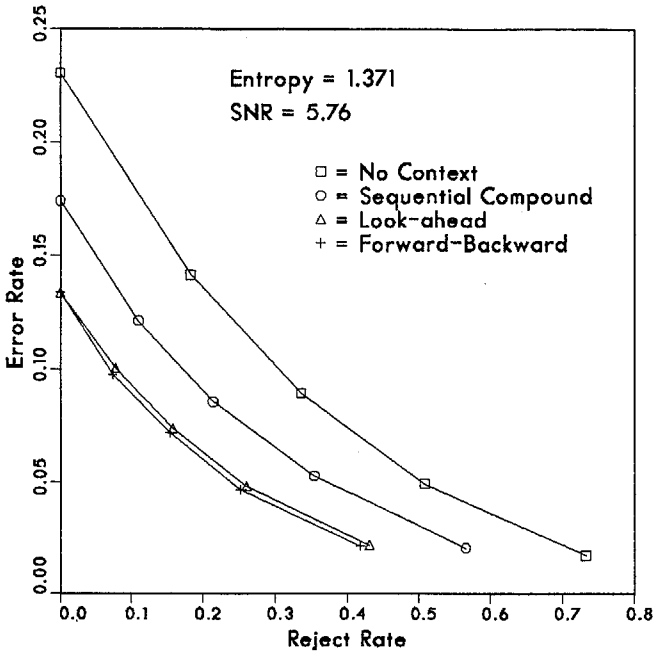

Figure 3: Error-reject curves, high entropy, low SNR.

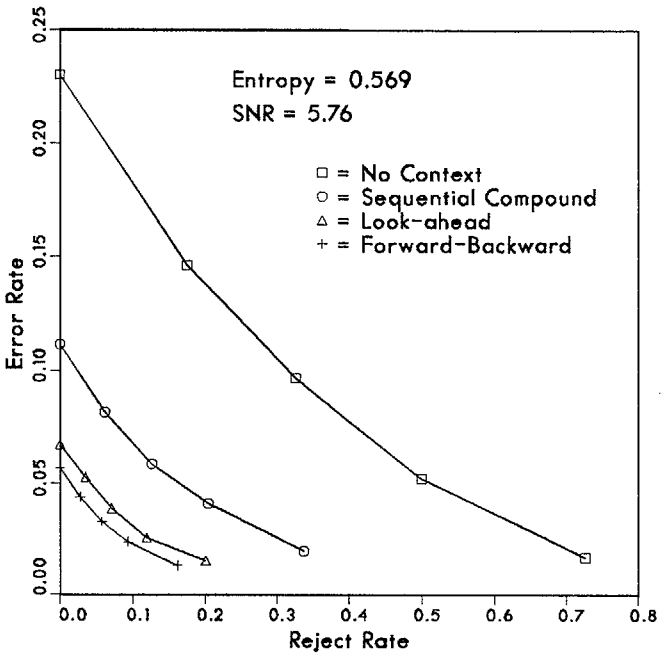

Figure 4: Error-reject curves, low entropy, low SNR.

source into a fairly skewed one with a low entropy of 0.569 bits (corresponding to p=0.9). This is illustrated by Figure 4. It is evident that these experimental conditions are quite unnatural and are very unlikely to be encountered in practical applications.

Figure 5 displays performance ratios [non-contextual (nc), sequential compound (sc), and one-step look-ahead (la) versus forward-backward (fb)] versus $H(\mathbf{p})$ for $\mathbf{p}$ in the range [0.4, 0.95] and an SNR of 5.76. Readily, these curves lead us to three conclusions:

i) The lowest the entropy of the Markov source, the most effective the contextual decision rules.

ii) Using one-step look-ahead yields significant improvement at fairly low computational costs. [In this respect, it seems worthwhile to recall that in Raviv's experiments [4], one-step look-ahead under a first order assumption appeared superior to the sequential compound rule (with no look-ahead) under a second order Markov chain assumption.]

iii) In terms of classification performance and under normal experimental conditions, the forward-backward algorithm can hardly do better than the one-step look-ahead method.

4. Comments

Our experiments have shown that, under "normal experimental conditions", the forward-backward algorithm does not improve in any significant way over the one-step look-ahead mode of decision. By the virtue of the computer simulation, this counter-intuitive observation may no longer be attributed to a lack of accuracy of the assumed Markovian model. The justification must therefore be searched among the properties of the model. A precise analysis of the problem appears to be extremely difficult and will not be attempted here. Though, it is our hope that the following considerations may help clarify the issue somewhat.

It is a well known property that a first order Markov chain is also a first order Markov process—not necessarily a chain—in the reverse direction [16]. This property implies that for $\tau < T$,

$$P\{\omega^\tau|\omega^1,\cdots,\omega^{\tau-1},\omega^{\tau+1},\cdots,\omega^T\} = P\{\omega^\tau|\omega^1,\cdots,\omega^{\tau-1},\omega^{\tau+1}\}. \tag{13}$$

In plain words, Eq. (13) shows that conditioning on the one-step look-ahead ω-sequence is equivalent to conditioning on the entire ω-sequence. [Equation (13) has a direct extension in terms of n-th order Markov chains and we have every reason to believe that the right thing to do under an nth order assumption is to use the n-step look-ahead decision rule prescribed by (2).]

It is equally well known that (13) applies only when the conditioning sequence is known exactly. When, as in our case, all that is known—or estimated—are the probabilities of the possible conditioning sequences, the actual outcome of the "terminal" sequence $\omega^{\tau+2},\cdots,\omega^T$, does affect our estimation of ω^τ.

The situation is further aggravated by the fact that what we do observe is not the ω-process but the X-process where X's are probabilistic functions of ω's through their class-conditional probability distributions. Hence, in general, the X-process is in no way Markov and there is no analogue to (13) when conditioning is in terms of X variables. The analysis of the conditions under which a function of a Markov chain is again Markov is known as the *lumpability* problem [16] (see also [17] for a recent reference). Unfortunately, the theory of lumpability has not reached the stage where it can handle such complex problems as the one considered here.

These theoretical hurdles not accounted for, let us briefly consider the case of a high signal to noise ratio. The knowledge of $X^{\tau+1}$ is most of the time highly indicative of the class to which it belongs. Hence, it looks as if $\omega^{\tau+1}$ were almost exactly

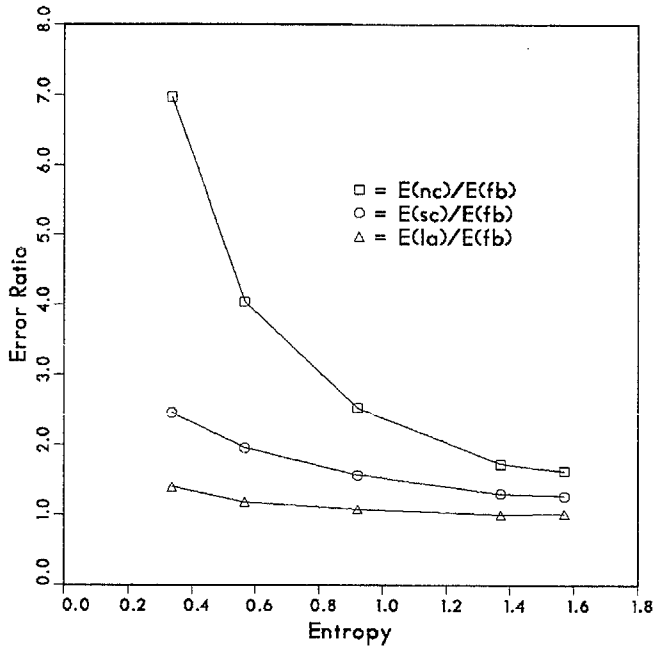

Figure 5: Performance ratios versus Markov source entropy (SNR = 5.76).

known and (13) approximately applies in most cases thereby disabling the capabilities of forward-backward to improve over one-step look-ahead. Very doubtful cases occur rather rarely. So the full potentiality of forward-backward is exploited equally rarely. Moreover, only a fraction of the "would be incorrect decisions" are turned into correct decisions. The net result is that the improvement reduces to second order effects. At low signal to noise ratio, the situation is somehow reversed and the forward-backward algorithm becomes more efficient.

In comparison, the role of the source entropy is more apparent. A source with high entropy assigns low probabilities to a large number of possible sequences of class labels, while a low-entropy source assigns high probabilities to a small number of possible sequences. When this information can be taken into account—that is, at low SNR—it is quite natural that low entropy should lead to more significant improvement. Although quite intuitive, these considerations are born out by our experimental results.

REFERENCES

[1] R.M. Haralick, "Decision making in context," *IEEE Trans. Pattern Anal., Machine Intell.*, PAMI-5, pp. 417–428, July 1983.

[2] G. Forney, "The Viterbi algorithm," *Proc. IEEE*, Vol. 6, pp. 268–278, 1973.

[3] G.T. Toussaint, "The use of context in pattern recognition," *Pattern Recognition*, Vol. 10, no. 3, pp. 189–204, 1978.

[4] J. Raviv, "Decision making in Markov chains applied to the problem of pattern recognition," *IEEE Trans. Inform. Theory*, IT-3, pp. 536–551, Oct. 1962.

[5] K. Abend, "Compound decision procedures for unknown distributions and for dependent states of nature," in *Pattern Recognition*, L.N. Kanal ed., Washington D.C. : Thompson Book Cy, 1968, pp. 207–249.

[6] P.A. Devijver, "Classification in Markov chains for minimum symbol error rate," *Proc. 7th Intern. Conf. Pattern Recognition*, Montreal, Aug. 1984, pp. 1334–1336

[7] L.E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process," *Inequalities*, Vol. 3, pp. 1–8, 1972.

[8] P.A. Devijver, "Cluster analysis by mixture identification," to appear in *Data Analysis in Astronomy*, V. Di Gesù and L. Scarci eds., New York: Plenum, 1984.

[9] F. Jelinek, and R.L. Mercer, "Interpolated estimation of Markov source parameters from sparse data", in *Pattern Recognition in Practice*, E.S. Gelsema ad L.N. Kanal eds., Amsterdam, The Netherlands: North Holland, 1980, pp. 381–397.

[10] F. Jelinek, R.L. Mercer, and L.R. Bahl, "Continuous speech recognition: Statistical methods," in *Handbook of Statistics*, Vol. 2, P.R. Krishnaiah and L.N. Kanal eds., Amsterdam, The Netherlands: North Holland, 1982, pp. 549–573.

[11] S.E. Levinson, L.R. Rabiner, and M.M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process in automatic speech recognition," *B.S.T.J.*, Vol. 62, no. 4, pp. 1035–1074, April 1983.

[12] L.R. Rabiner, S.E. Levinson, and M.M. Sondhi, "One the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition," *B.S.T.J.*, Vol. 62, no. 4, pp. 1075–1105, April 1983.

[13] K.L. Chung, *Markov Chains With Stationary Transition Probabilities*, Berlin: Springer-Verlag, 1967.

[14] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, Vol. 64, pp. 532–556, April 1976.

[15] P.A. Devijver, and J. Kittler, *Pattern Recognition: A Statistical Approach*, Englewood Cliffs: Prentice Hall, 1982.

[16] J.G. Kemeny, and J.L. Snell, *Finite Markov Chains*, New York: Springer-Verlag, 1976.

[17] A.M. Abdel-Moneim, and F.W. Leysiffer, "Lumpability for non-irreducible finite markov chains," *J. Appl. Prob.* Vol. 21, pp. 567–574, Sept. 1984.