ELSEVIER

Signal Processing 54 (1996) 191-199

SIGNAL PROCESSING

Fast communication

A new time series classification approach

K.N. Plataniotis[a,*], D. Androutsos[a,1], A.N. Venetsanopoulos[a,2], D.G. Lainiotis[b]

[a] *Department of Electrical and Computer Engineering, Digital Signal & Image Processing Laboratory, University of Toronto, 10 King's College Rd., Toronto, Ontario M5S 3G4, Canada*

[b] *Intelligent Systems Technology, 3207 S. Hwy. A1A, Melbourne Beach, FL-32951, USA*

Received 7 June 1996

Abstract

A new, robust and computationally attractive approach to the problem of time series classification is discussed in this paper. Both the Bayesian as well as a new adaptive classification scheme for source selection are discussed. Simulation results are included to demonstrate the effectiveness of the new methodology.

Keywords: Time series classification; Neural networks; Logistic function

1. Introduction

The problem of time series classification arises in many real-world problems, such as speech identification, dynamic system identification, systems subject to failures/repairs, EEG diagnosis, piecewise linearization of nonlinear systems, target tracking and reconfigurable systems [9,4,6].

In such problems a set of source models is usually selected to represent the possible system behavior patterns. Thus, it is assumed that the observable time series $y(k), k = 1,2,\ldots,$ is generated by an unknown source model $S(\theta_i)$ where θ_i is a parameter which completely characterizes the model taking values in a finite set $\Theta = (\theta_1, \theta_2, \ldots, \theta_n)$.

Recently, a number of adaptive classification schemes have been devised. The objective of the different classification schemes is to recursively identify the source model which generates the time series by selecting the optimal value for the parameter θ [9,6]. Such schemes achieve their objective by utilizing a bank of neural network based predictors, each one trained off-line with labeled data from a particular source model $S(\theta_i)$. In the sequence, during the classification process, at time step $k = 1,2,\ldots$ each neural predictor is used to generate an estimate of the next available observation of the series $y(k)$ utilizing the past values $(y(k-1), y(k-2), \ldots, y(1))$. Based on the corresponding prediction errors the best model is selected on-line.

* Corresponding author. E-mail: kostas@dsp.toronto.edu.

[1] E-mail: zeus@dsp.toronto.edu.

[2] E-mail: anv@dsp.toronto.edu.

0165-1684/96/$15.00 Copyright © 1996 Elsevier Science B.V. All rights reserved.

PII S0165-1684(96)00155-7

In this paper we review the Bayesian approach to the problem and introduce a new, robust and computationally attractive classification scheme. The rest of the paper is organized as follows. In Section 2, we discuss the Bayesian approach and the new nearest neighbor classification module. Motivation and implementation issues are discussed in this section. The application of the proposed classification scheme to the problem of logistic source detection as well as comparison with existing schemes are discussed in Section 3. Finally, Section 4 summarizes our conclusions.

2. Adaptive time series classification

2.1. The Bayesian approach

The first approach to the problem discussed in the paper is the Bayesian one. Let us assume a random variable θ taking values in the set $\Theta = (\theta_1, \theta_2, \dots, \theta_n)$. In this context the time series $(y(k), y(k-1), y(k-2), \dots, y(2), y(1))$ is generated by the source model $S(\theta)$. At each time instant the Bayesian decision rule is utilized to generate an estimate of the random variable θ which generates the series.

The Bayesian classification scheme assumes that at time index k the sequence $(y(k), y(k-1), y(k-2), \dots, y(2), y(1))$ has been generated by the source model $S(\hat{\theta}(k))$ (as it is approximated by the corresponding neural predictor),with $\hat{\theta}(k) = \theta_j$ which has the maximum a posteriori probability

$$p_j(k) = \mathrm{Prob}(\theta = \theta_j \mid (y(k), y(k-1), \dots, y(1))), \tag{1}$$

with a priori probabilities $p_j(0) = \mathrm{Prob}(\theta = \theta_j \mid k = 0) = 1/n$, considering all models to be equiprobable.

The maximum a posteriori probability can be calculated recursively using Baye's rule, presented in [2,5,3], and applied initially within the context of neural networks in [9,4] and later used in [7,6].

Following the approach in [9] the posterior probability for the jth neural predictor, with $j = 1, 2, \dots, n$, is calculated recursively as follows:

$$p_j(k+1) = \frac{f(y(k+1) | ((y(k), y(k-1), \dots, y(1)), \theta_j)\, p_j(k)}{\sum_{i=1}^{n} f(y(k+1) | ((y(k), y(k-1), \dots, y(1)), \theta_i)\, p_i(k)}. \tag{2}$$

The calculation of the a posteriori probability $p_j(k)$ depends on the form of the neural network predictor $\hat{y}_j(k) = \hat{f}((y(k), y(k-1), \dots, y(1)), \theta_j)$ used to approximate the actual time series $y(k)$ under the assumption that the time series is generated by the source model $S(\theta_i)$ for $j = 1, 2, \dots, n$.

Multi-layer perceptrons with sigmoidal neurons trained with the backpropagation rule are used as predictors in this work. However, it must be emphasized that other predictors, such as linear predictors, polynomial networks or RBF (Radial Basis Function) neural nets can be used instead.

For each of the $j = 1, 2, \dots, n$ neural predictors, the prediction error $\hat{e}_j(k) = y(k) - \hat{y}_j(k)$ is calculated. Based on the error function used during the off-line training phase a different conditional probability function $f(\hat{e}_j(k) \mid (y(k-1), y(k-2), \dots, y(1)), \theta_j)$ is obtained.

If an L_p *Minkowski* error measure is used during training, the resulting error density has the form of the generalized Gaussian density [9,1] and thus, (2) transforms to

$$p_j(k+1) = \frac{f(\hat{e}_j(k) \mid \theta_j)\, p_j(k)}{\sum_{i=1}^{n} f(\hat{e}_i(k) \mid \theta_j)\, p_i(k)} = \frac{c_j \exp(-\beta_j (y(k+1) - \hat{y}_j(k+1))^p)\, p_j(k)}{\sum_{i=1}^{n} c_i \exp(-\beta_i (y(k+1) - \hat{y}_i(k+1))\, p_i(k)}, \tag{3}$$

where p is the shape parameter of the generalized Gaussian density, c_j a normalized constant and β a function of the variance. Based on the calculations, at time index $k+1$ the time series is classified to the source model (neural predictor) which maximizes the posterior probability $\hat{\theta}_j(k+1) = \max_{\Theta} p_j(k+1)$.

2.2. The nearest neighbor scheme

In this section we consider an improved and robust classification scheme. The new scheme is simple, robust to possible outliers, classifies as well as the Bayesian rule and has simpler implementation than the Bayesian scheme discussed in Section 2.1. As for the Bayesian approach, a bank of neural network-based predictors are trained off-line using labeled data. In the sequence, during the on-line phase we define the instantaneous prediction error $\hat{e}_j = \Phi(y(k) - \hat{y}(k))$, $k = 1,2,\ldots$, $j = 1,2,\ldots,n$, where $\Phi(\cdot)$ is a robust function of the prediction error selected to reduce the effects of any present outlier. In most cases the Bayesian approach of Section 2.1 is implemented using the square norm $\hat{e}_j = (y(k) - \hat{y}(k))^2$ to calculate the instantaneous prediction error [6]. However, if outliers are assumed present other error measures, such as the L_1 norm $\hat{e}_j = |(y(k) - \hat{y}(k))|$ or the the Huber minimax function

$$\Phi(e(k)) = e(k)^2, \qquad e(k) \leqslant T,$$
$$\Phi(e(k)) = T^2 + 2T(|e(k) - T|), \quad e(k) > T,$$

are most appropriate for the evaluation of the prediction error [8].

The instantaneous prediction error $\hat{e}_j(k)$ depends not only on the measurement $y(k)$ but also on past measurements which have been used to form each neural predictor. However, since the instantaneous prediction error cannot predict accurately the best source model, the aggregate error is utilized in the new scheme. For the jth neural predictor the aggregated prediction error during the on-line implementation phase is defined as follows:

$$d_j(k) = \gamma d_j(k-1) + \hat{e}_j(k), \tag{4}$$

where γ is a decay factor slightly less than 1 for discounting past data, and $j = 1,2,\ldots,n$, $k = 1,2,\ldots$.

Based on the corresponding aggregate prediction error a credibility weight w_j is assigned to the jth model. The credibility weight is defined as follows:

$$w_j(k) = \frac{d_{\max}(k) - d_j(k)}{d_{\max}(k) - d_{\min}(k)}, \tag{5}$$

with $j = 1,2,\ldots,n$ and $d_{\max}(k)$, $d_{\min}(k)$ are the corresponding maximum/minimum aggregated prediction errors.

The credibility weights defined above express the degree to which the jth predictor is close to the ideal, which best approximates the actual data model, and far away from the worst value, the outer rank. Both the minimum error rank position $d_{\min}$ and the worst rank $d_{\max}$ are occupied by at least one of the predictors under consideration. It is evident that the model with the minimum prediction error will be assigned the highest credibility weight.

The following comments should be made regarding our approach:

(i) The new decision mechanism is more versatile than the Bayesian approach discussed in Section 2.1. A number of robust functions Φ can be used to reduce the effects of any impulses present [8]. Thus, the new decision mechanism is robust to noise and prediction error.

(ii) As in the Bayesian scheme, the final decision is based not on the absolute predictive performance of a neural predictor, but on the relative performance of all neural predictors, resulting in a decision scheme immune to high levels of noise in the data.

(iii) The new scheme assigns the credibility weights in a recursive manner. Thus, the new scheme is appropriate for on-line classification of dynamic series without an increase in dimensionality. The new scheme can be implemented using only adders and ranking elements and is suitable for parallel implementation.

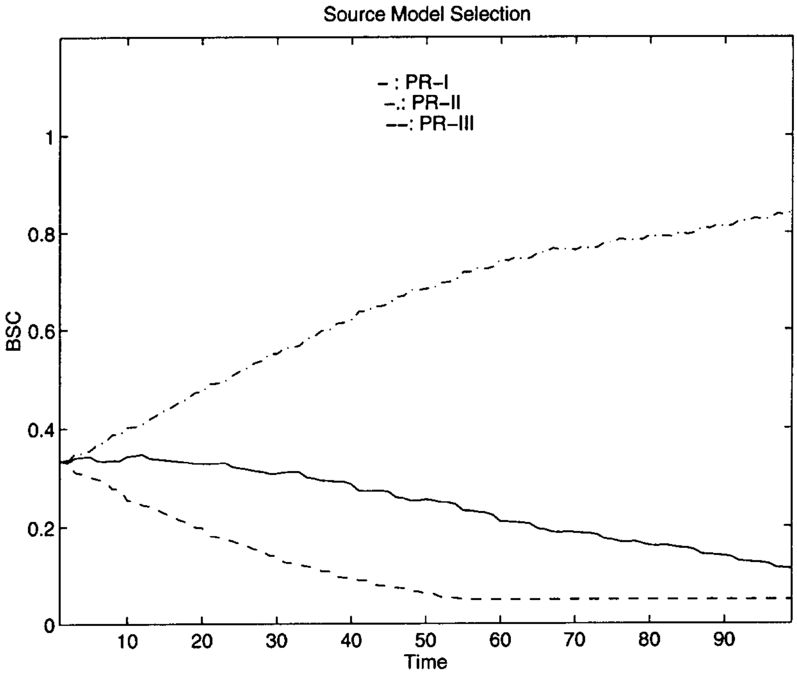

Fig. 1. Experiment I: Model selection, Bayesian approach.

(iv) The proposed new scheme is simpler and more robust than alternative suboptimal schemes, such as the so called ICRA (Incremental Credit Assignment) scheme proposed in [6]. Contrary to the approach in [6] no steepest descent procedure or ad-hoc defined parameters are used in our approach.

3. Application to logistic classification

To demonstrate the effectiveness of the proposed methodology, the problem of logistic time series detection is considered. A logistic time series is generated by the following difference equation [4,6]:

$$x(k) = a(x(k-1)(1-x(k-1)), \tag{6}$$

$$y(k) = x(k) + w(k), \tag{7}$$

with $k = 1,2,\dots$ and $w(k)$ is zero-mean white noise, uniformly distributed in the interval $[-0.25, 0.25]$.

In a first experiment we assume that the actual value of the parameter used to generate the time series is $a = 3.75$. The exact value is not assumed to be known. It is simply assumed that it lies in the interval $[3.5, 4.0]$. Hence, in the off-line phase a bank of three neural predictors has been trained on labeled data from the logistic time series of (6)–(7) with the following values:

Predictor I: $a = 3.5$, Predictor II: $a = 3.9$, Predictor III: $a = 4.0$.

The neural predictors are 2–4–1 sigmoidal feedforward neural networks trained via the backpropagation rule with learning rate of 0.05 and momentum term of 0.1. In all the simulation studies reported the proposed nearest neighbor scheme (NNSC) uses the L_1 norm to calculate the term $\hat{e}(k)$ in (4). The classification performance of the Bayesian scheme (BSC), that of the NNSC and finally that of the ICRA is plotted in Figs. 1–3. The

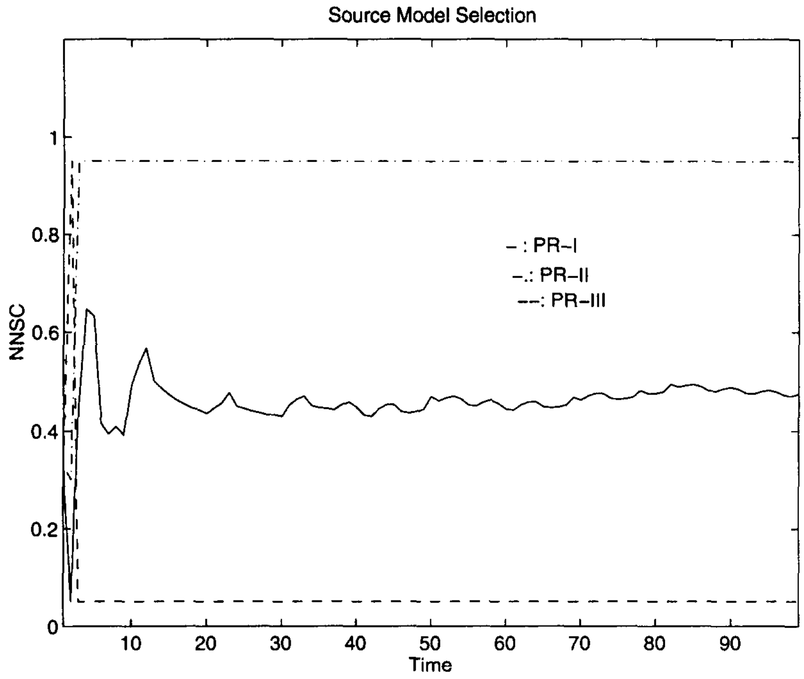

Fig. 2. Experiment I: Model selection, Nearest Neighbor approach.

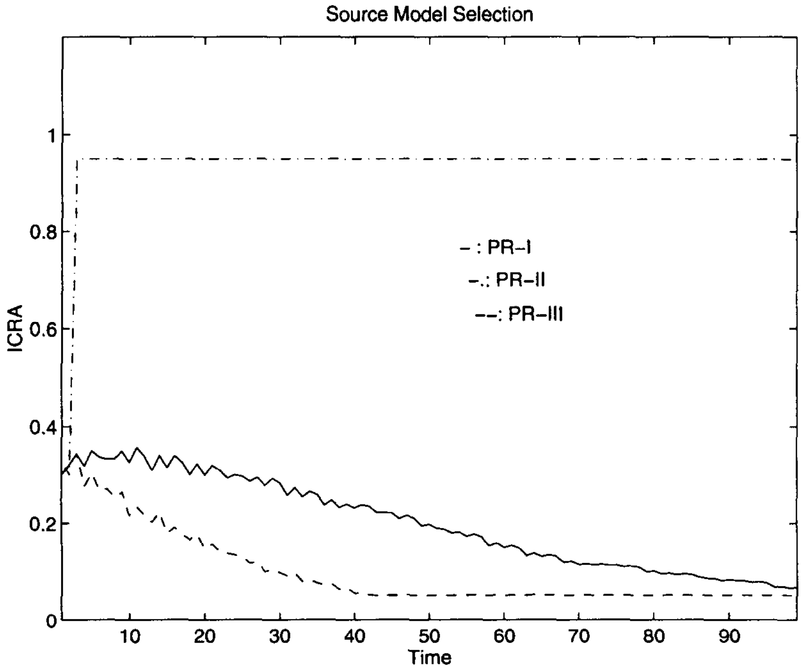

Fig. 3. Experiment I: Model selection, ICRA approach.

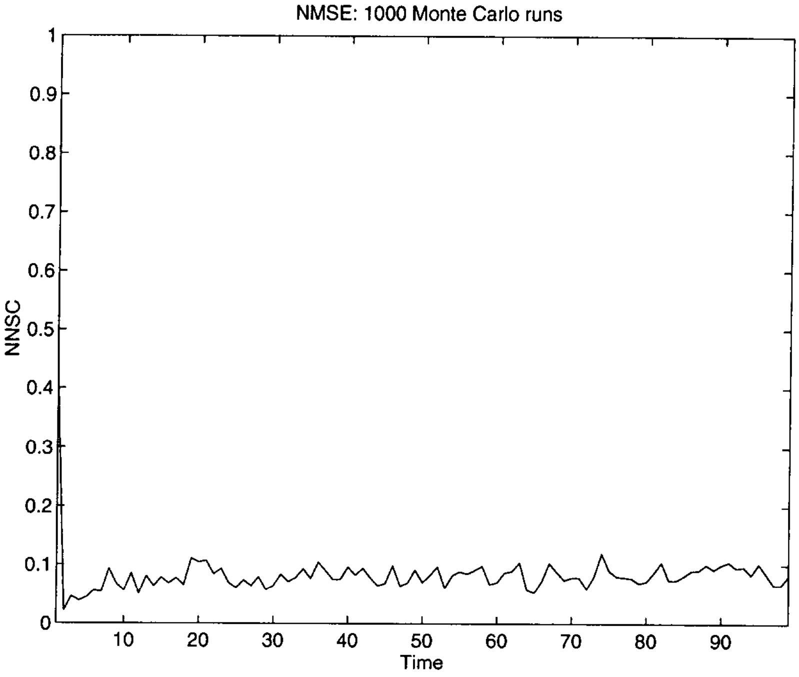

Fig. 4. Experiment I: Nearest Neighbor approach, NMSE, 1000 Monte Carlo runs.

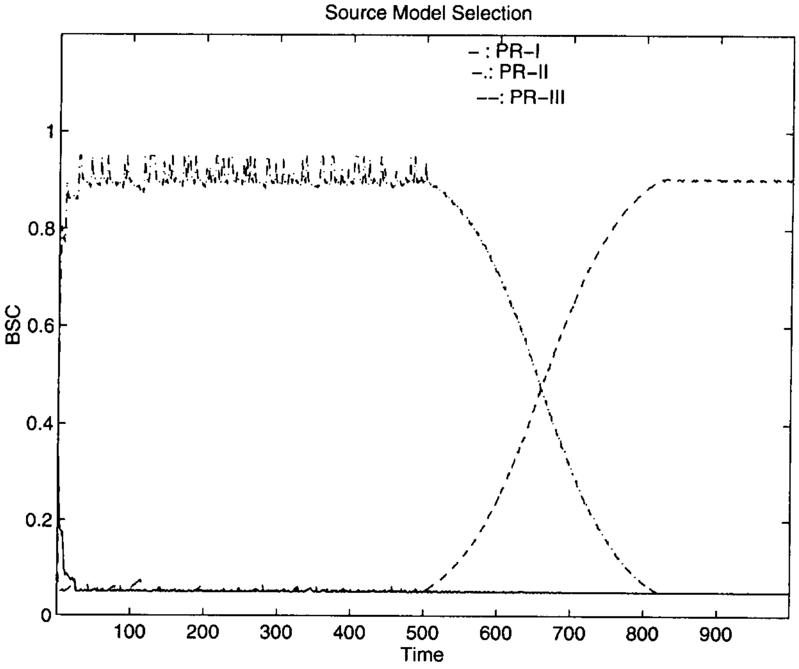

Fig. 5. Experiment II: Model selection, Bayesian approach.

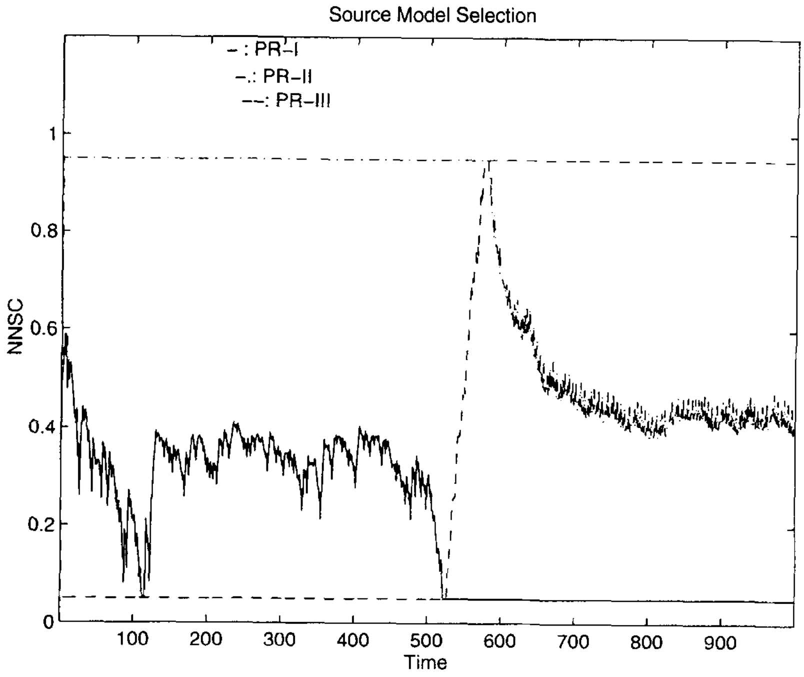

Fig. 6. Experiment II: Model selection, Nearest Neighbor approach.

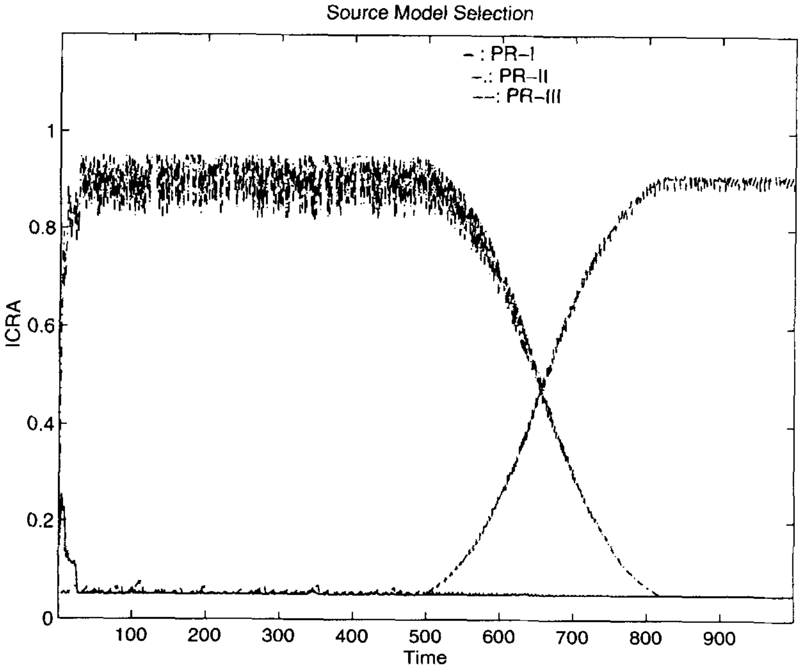

Fig. 7. Experiment II: Model selection, ICRA approach.

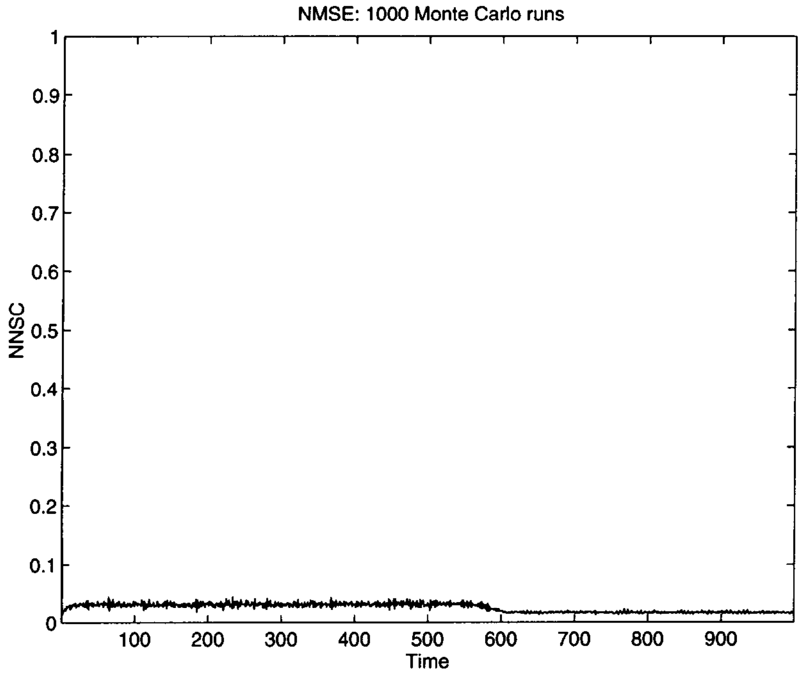

Fig. 8. Experiment II: Nearest Neighbor approach, NMSE, 1000 Monte Carlo runs.

prediction performance of the NNSC classification scheme in terms of the Normalized Mean Square Error (NMSE) for 100 steps averaged over 1000 Monte Carlo runs is depicted in Fig. 4.

In a second experiment, a source switching takes place. The actual time series is generated again using the system of (6)–(7). For the first half of the data (500 steps) the actual value of the parameter is $a = 4.0$. For the remaining 500 steps the parameter value used to generate the actual series is $a = 3.75$. During the off-line training phase three neural predictors have been trained using the following values:

Predictor I: $a = 3.5$, Predictor II: $a = 4.0$, Predictor III: $a = 3.75$.

The classification performance for the different schemes is plotted in Figs. 5–7 and the prediction performance of the NNSC in Fig. 8.

4. Conclusion

The following conclusions are supported from the experiments presented above:

- The new nearest neighbor decision scheme achieves high classification accuracy even for data corrupted with high noise, such as those used here. The nearest neighbor scheme is extremely robust even when the actual value is not included the search set used for training purposes. It should be emphasized here that the new scheme outperforms the ICRA which was devised to offer robustness in the design in such a situation.
- The new classification scheme retains the natural decoupled structure of the optimal Bayesian approach, since all the neural predictors needed can be independently realized. In addition, the NNSC converges to the source/model in the sample-space which is “closer” to the actual one if the later is not included in the design set.

- The nearest neighbor scheme is simple, easy to implement, requires no ad hoc parameters, and can adapt to switching source. It can be seen from Figs. 5–7 that the NNSC immediately identifies the switch in the active source model. On the other hand, both the Bayesian approach as well as the ICRA of [7] took more than 150 steps to identify the correct source/model.

In conclusion, the proposed nearest neighbor classification scheme outperforms other decision schemes used for the problem of time series classification, offers superior performance in the case of switching sources without complicate calculations and expensive implementation requirements.

References

[1] P. Burrascano, "A norm selection criterion for the generalized delta rule", *IEEE Trans. Neural Networks*, Vol. 2, No. 1, 1991, pp. 125–130.

[2] C.G. Hilborn and D.G. Lainiotis, "Optimal estimation in the presence of unknown parameters", *IEEE Trans. Syst. Sci. Cybernet.*, Vol. SSC-5, 1969, pp. 38–43.

[3] S.K. Katsikas, A.K. Leros and D.G. Lainiotis, "Passive tracking of a maneuvering target: An adaptive approach", *IEEE Trans. Signal Process.*, Vol. 42, No. 7, 1994, pp. 1820–1825.

[4] D.G. Lainiotis and K.N. Plataniotis, "Adaptive dynamic neural network estimation", *Proc. IEEE Internat. Conf. on Neural Networks, ICNN-94*, Vol. III, pp. 4710–4717.

[5] D. Lainiotis, S.K. Katsikas and S.D. Likothanasis, "Adaptive deconvolution of seismic signals: Performance, computational analysis and parallelism", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. AASP-36, No. 11, 1988, pp. 1715–1774.

[6] V. Petridis and A. Kehagias, "A recurrent network implementation of time series classification", *Neural Computation*, Vol. 8, 1996, pp. 357–372.

[7] V. Petridis and A. Kehagias, "Modular neural networks for MAP classification of time series and the partition algorithm", *IEEE Trans. Neural Networks*, Vol. 7, No. 1, 1996, pp. 73–85.

[8] I. Pitas and A.N. Venetsanopoulos, *Nonlinear Digital Filters: Principles and Applications* (Kluwer Academic Pub;ishers, Norwell, MA, 1990).

[9] K.N. Plataniotis, Distributed parallel processing state estimation algorithms, Ph.D. Dissertation Thesis, Florida Institute of Technology, Melbourne Beach, FL, 1994.

[10] K.N. Plataniotis, D. Androutsos, V. Sri and A.N. Venetsanopoulos, "A nearest neigbour multichannel filter", *Electronic Lett.*, 1995, pp. 1910–1912.