Regime Signaling Techniques for Non-stationary Time Series Forecasting

Radu Drossu Zoran Obradović
School of Electrical Engineering and Computer Science
Washington State University
rdrossu@eecs.wsu.edu zoran@eecs.wsu.edu

Abstract

An accuracy-based signaling technique is proposed as an alternative to a statistics-based signaling for detecting changes in a time series distribution. Three different forecasting scenarios are analyzed in order to decide whether to reuse historically successful neural network models or retrain new ones when a change in the distribution is signaled. The results obtained on low-noise and high-noise, non-stationary time series provide strong evidence in favor of the accuracy-based signaling technique.

1 Introduction

The theoretical work shows that, similar to traditional approximation techniques based on Taylor function expansion or Fourier series, neural networks (NN) are powerful computational structures able to approximate almost any arbitrary continuous function [2]. In addition, NNs can effectively construct approximations for unknown functions by learning from examples (known outcomes of the function), which makes them attractive in practical applications where traditional computational structures have performed poorly (e.g. ambiguous data or large contextual influence).

Many real-life time series are the result of complex and insufficiently understood interdependencies. Hence, forecasting models make use of incomplete information, while other factors not included in the models act as noise. In addition, real-life time series are sometime non-stationary, meaning that the data distribution is changing over time. Most often for non-stationary domains, a single model built on a certain data segment and used for all subsequent predictions is inadequate. A straightforward attempt is to stationarize the data by performing a de-trending preprocessing (e.g. a first or a second order discrete differentiation). More sophisticated methods provide solutions for certain types of non-stationarity (e.g. a reversible power transformation is successfully used to stabilize the variance of a series affected by a strong trend that cannot be removed by differencing [1]). However, not all non-stationary processes can be stationarized through data preprocessing. Forecasting such processes requires *on-line learning techniques*, where a given model is used for a limited time and a new model is constructed whenever a change of the underlying data distribution is detected. Although the on-line learning received recently considerable attention in the literature on computational learning theory [4], and is already applied to some specific classification problems [9], many important issues related to efficient forecasting of non-stationary time series are still unsolved. If time is not an issue, on-line learning can be accomplished by a *sliding window* technique, in which a new prediction model is built whenever a new data sample becomes available. However, in many real-life problems the data arrival rate is high, which makes this approach completely infeasible due to the computational complexity involved in repeatedly building NN prediction models. An alternative encountered in practice is the *uniform retraining* technique, in which an existing NN prediction model is used for a prespecified number of prediction steps (called reliable prediction interval), followed by the replacement of the existing model by one constructed using more recent data. A major disadvantage of uniform retraining is that it is often hard to determine an appropriate reliable prediction interval, as it might be changing over time.

Although theoretically possible, in practice it might be very difficult to efficiently learn a single global NN model for a non-stationary time series forecasting. An obvious difficulty of such a global approach is the selection of NN modeling parameters that are appropriate for all data segments. Additional serious problems include different noise levels in various

1060-3425/97 $10.00 © 1997 IEEE

data segments resulting in local overfitting and underfitting conflicts (it would be desired to stop training as not to overfit some data segments, while other data segments would still require additional training). An interesting multi-model attempt to forecasting *piecewise stationary* time series, where the process switches between different regimes, is by using a *gating network*, in which a number of NN experts having an identical structure are trained in parallel, and their responses are integrated by another NN also trained in parallel with the expert networks [7]. Briefly, due to an interesting combination of activation and error functions that encourages localization, in a gating network each expert network tends to learn only a subset of the training data, thus devoting itself solely to a sub-region of the input space. This competitive integration method showed quite promising results when forecasting a non-stationary time series having two regimes, but is not likely to extend well to more complex non-stationary processes due to overfitting problems of training a gating network system consisting of too many expert networks. In addition, the time required to train a complex gating network is likely to be prohibitively long for many real-life time series forecasting problems.

The multi-model forecasting approach proposed in this paper implies the use of a NN predictor until it is signaled that a new predictor is needed. This paper compares a statistics-based with an accuracy-based signaling technique for deciding whether to reuse "trusted" models or retrain new ones. The *statistics-based signaling* attempts to identify changes in the distribution by analyzing the statistical similarity of different data segments. An alternative signaling technique proposed in this paper for deciding when a new prediction model is needed is the *accuracy-based signaling*, in which changes in the distribution are identified based on prediction errors from previous time steps. Both for prediction accuracy and for computational efficiency it is desirable to make use of any previously successful prediction model.

Sections 2 and 3 provide the necessary statistics and neural network background concepts, while Section 4 explains the methodology employed. The experimental results obtained on relatively noise-free, as well as on extremely noisy, non-stationary time series with or without model reuse are presented in Section 5, followed by conclusions provided in Section 6.

2 Non-stationary Time Series

A *time series* $\{x_t\}$ can be defined as a function x of an independent variable t stemming from an unknown process. Its main characteristic is that its future behavior cannot be predicted exactly as in the case of a known deterministic function.

A *non-stationary* time series can be described as a time series whose characteristic parameters change over time. Different measures of stationarity can be employed to decide whether a process is stationary or not [5]. In practice, confirming that a given time series is stationary is a very difficult task, unless a closed-form expression of the underlying time series is known. Non-stationarity detection can be reduced to identifying two sufficiently long, distinct data segments that have significantly different statistics (distributions). In practice, common tests for comparing whether two distributions are different are [6]:

- Student's t-test;
- F-test;
- chi-square test;
- Kolmogorov-Smirnov test.

Student's t-test is applied to identify the statistical significance of a difference in means of two distributions assumed to have the same variance, whereas the *F-test* evaluates the statistical significance of a difference in variances. More commonly, if there aren't any assumptions regarding the means or variances of the distributions, a chi-square or a Kolmogorov-Smirnov test are performed.

In the *chi-square test*, the data range of the two data sets to be compared is divided into a number of intervals (bins). Assuming that R_i and S_i represent the number of data samples in bin i for the first and the second data set, respectively, the chi-square statistic computes

$$\chi^2 = \sum_i \frac{(R_i - S_i)^2}{R_i + S_i},$$

with the sum taken over all bins. The complement of the incomplete gamma function,

$$Q(\nu, \chi^2) = \frac{1}{\Gamma(\nu)} \int_{\chi^2}^{\infty} e^{-t} t^{a-1} dt,$$

where

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt,$$

is then evaluated and a small value of Q (close to 0) indicates that it is unlikely that the two distributions

are the same. Here, ν represents the number of degrees of freedom which in the case when the two sets have the same number of data samples ($\sum R_i = \sum S_i$), equals the number of bins minus one. If the previous restriction is not imposed, then ν equals the number of bins.

The *Kolmogorov-Smirnov (K-S) test* measures the absolute difference between two cumulative distribution functions S_{N_1} and S_{N_2} with N_1 and N_2 data points, respectively. The K-S statistic computes

$$D = \max_{-\infty<x<\infty} |S_{N_1}(x) - S_{N_2}(x)|.$$

The function Q_{KS} defined as

$$Q_{KS}(\lambda) = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2\lambda^2}$$

is computed for

$$\lambda = D(\sqrt{N_e} + 0.12 + 0.11/\sqrt{N_e}),$$

where N_e is the effective number of data points computed as

$$N_e = \frac{N_1 N_2}{N_1 + N_2}.$$

A small value of Q_{KS} (close to 0) indicates that it is unlikely that the two distributions are the same.

3 NN for Time Series Prediction

Neural network computational models consist of a number of relatively simple, interconnected processing units called *neurons*, working in parallel. Different neural network *architectures* can be used to solve a given classification or prediction problem. In an architecture, the neurons are interconnected through *synaptic links* (*weights*) and are grouped into *layers*, with synaptic links usually connecting neurons in adjacent layers. Typically three different layer types can be distinguished: *input* (the layer that external stimuli are applied to), *output* (the layer that outputs results to the external world) and *hidden* (the intermediate computational layers between input and output layers). The NNs used in this paper are of *feedforward* type, (in which the signal flow is from input layer towards output layer), with two hidden layers of neurons.

The most distinctive property of neural networks, as opposed to traditional computational structures, is called *learning*. Learning (training) represents the optimization of the neural network weights by using a set of examples (known outcomes of the problem for given conditions), so that the neural network computes a desired function. The final data modeling goal is not to memorize known patterns, but to *generalize* from prior examples (to be able to estimate the outcomes of the problem under unknown conditions).

When predicting a time series, a feedforward NN with k input units is trained on n consecutive process values $x_1, \ldots, x_n$, that are used to build a *training window* consisting of $n-k+1$ examples. Each example has the form $< x_{t-k}, \ldots, x_{t-1}, x_t >$, where $x_{t-k}, \ldots, x_{t-1}$ are used as inputs to the neural network and x_t is used for comparison to the actual process value at time step t. In this paper, the NN training is performed using the backpropagation algorithm, which is an iterative gradient descent method for minimizing the total squared prediction error on the training window [8]. The trained NN predicts the process value at time step m, where $m > n$, by using process values from k previous time steps as neural network input values. More formally,

$$\hat{x}_m = f(x_{m-1}, x_{m-2}, \ldots, x_{m-k}),$$

where f is an NN function obtained through learning on a given training window.

Although the prediction error plot represents a commonly encountered attempt to estimate a predictor's accuracy, it only provides a subjective means of visual evaluation. A more informative quantitative error measure for time series prediction evaluation is the *coefficient of determination*, which is a function of the mean squared error normalized by the variance of the actual data. It is computed as:

$$r^2 = 1 - \sum_{i=1}^{n}(x_i - \hat{x}_i)^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2,$$

where x_i and $\hat{x}_i$ denote actual and predicted process values, respectively, while $\bar{x}$ denotes the mean of the actual data. For a perfect predictor, the coefficient of determination should be one, whereas for a trivial mean predictor (one that always predicts the mean of the actual data) the coefficient of determination is zero.

4 Methodology

In this study we consider three different time series forecasting scenarios:

1. switching between two historically successful NN models (SWITCH);
2. reusing a historically successful NN model, or training a new one (REUSE);

3. retraining a NN model when signaled, without relying on any historically successful model (RETRAIN).

The SWITCH scenario assumes that two historically successful models were identified in the past. The objective is to detect in real-time which of the two models to use for prediction at any given time step. The REUSE scenario assumes an existing previously successful model. The objective is to decide in real-time whether to use the existing previously successful historic model for prediction, or to retrain a new NN on current data. Finally, the RETRAIN scenario is not assuming any previously successful model. The objective is to decide in real-time when to discard a NN predictor and retrain a new one on current data. The SWITCH and the REUSE scenarios are proposed in order to efficiently forecast piece-wise stationary processes with full or partial understanding of the number of different regimes, while the RETRAIN scenario is proposed for forecasting completely unknown higher order non-stationary processes.

The three scenarios were analyzed in the context of two different distribution-change signaling techniques, explained as follows.

4.1 Statistics-Based Signaling

This signaling technique attempts to identify changes in the data distribution by comparing the similarity of different data segments using either the chi-square or the K-S statistics.

For the SWITCH scenario, two historical data segments, D_{h1} and D_{h2}, both of length p, along with their successful NN models, M_{h1} and M_{h2}, trained on these segments are kept in a library. A current window, W, containing the p latest available data is compared distribution-wise (using either the chi-square or the K-S tests) to D_{h1} and D_{h2}, in order to decide which of the two historical data segments is more similar to it. The library model corresponding to the more appropriate historical data segment is then used for predicting the next time series value. This process is repeated when a new data sample becomes available (each time step).

For the REUSE scenario, a single historical data segment, D_h, used to build a previously successful NN model, M_h, as well as a temporary data segment, D_t, used to build a temporary NN model, M_t, both of length p, are kept in a library. The models M_h and M_t are also stored in the library. A current window, W, containing the p latest available data is compared distribution-wise (using either the chi-square or the K-S tests) to D_h and D_t, in order to decide whether to continue using one of the library models or to train a new model. For this purpose, a threshold has to be imposed on the confidence value obtained from the chi-square or K-S tests. If the test indicates more confidence in M_h, provided that the confidence value for M_h is larger than the specified threshold, then M_h is used for the current prediction. Similarly, if we are more confident in M_t and the confidence value is larger than the threshold, then M_t is used for the current prediction. Otherwise (none of the confidence values is larger than the imposed threshold), a new temporary NN model is trained on W and it replaces M_t, whereas W replaces D_t in the library. The new model is then used for the current prediction and the process is repeated when a new data sample becomes available.

In the case of the RETRAIN scenario, a data segment, D_t, of length p used to build a temporary NN model, M_t, is stored in a library. A current window, W, containing the p latest available data is compared distribution-wise (using either the chi-square or the K-S tests) to D_t, in order to decide whether to continue using M_t, or discard it and train a new NN model. For this purpose, a threshold imposed on the confidence value obtained from the chi-square or K-S tests is used to decide when the current model becomes inappropriate. If M_t is considered to be inadequate, W replaces D_t and a new NN model trained on W replaces M_t in the library. The new model is then used for the current prediction and the process is repeated when a new data sample becomes available.

4.2 Accuracy-Based Signaling

This signaling technique attempts to identify changes in the data distribution by measuring recent prediction accuracies of previously successful models.

For the SWITCH scenario, two historically successful NN models, M_{h1} and M_{h2}, are kept in a library. At each time step, the two models are compared based on their accuracy measured on a buffer containing b most recent process values, and the more accurate model is used for the current prediction.

For the REUSE scenario, a historically successful NN model, M_h, as well as a temporary NN model, M_t, are kept in a library. Similar to the SWITCH scenario, the accuracy of the two models is compared on the b most recent process values. The model having a better accuracy is used for predicting the current step, unless none of the models is a sufficiently good predictor on the b most recent process values. A model is considered to be sufficiently good if its accuracy on the b most recent process values is above $\alpha \min\{A_h, A_t\}$, where α is a prespecified threshold in

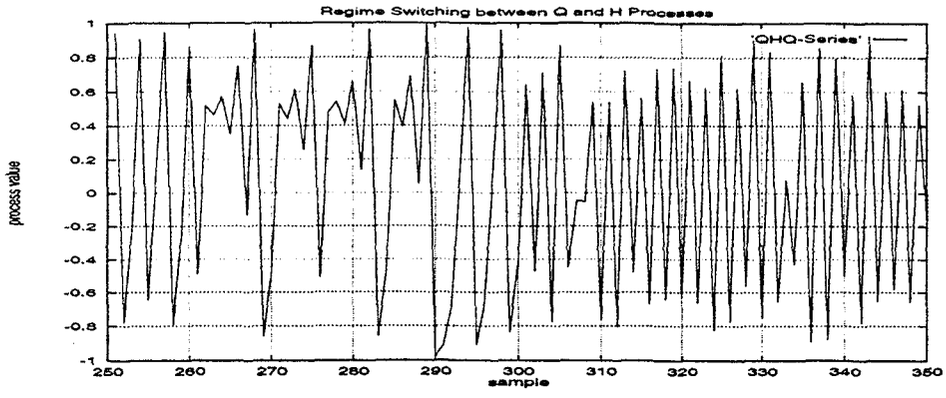

Figure 1: Regime Switching between Q and H Processes

the (0,1) range, while A_h and A_t are the training accuracies for the historical and the temporary model, respectively, computed on the process values used to build them. If none of the two existing models is satisfactory, a new NN model is trained, that replaces M_t in the library and is also used for the current prediction. This process is repeated whenever a new data sample becomes available.

In the case of the RETRAIN scenario, a temporary NN model, M_t, is stored in a library. Additionally, a corresponding training accuracy, A_t, is measured as for the REUSE scenario. If the accuracy M_t, measured on the b most recent process values is αA_t, model M_t is used for the current prediction. Otherwise, a new NN model is trained that replaces M_t in the library that is also used for the current prediction. This process is repeated whenever a new data sample becomes available.

Our experiments included two different accuracy measures. In the SWITCH scenario, the accuracy was computed as the mean of the absolute error values corresponding to the predicted values stored in the buffer. This quality measure can be used just in cases in which a decision has to be taken regarding which library model is the most adequate, irrespective of its actual prediction quality. For this reason, the REUSE and RETRAIN scenarios employ as an accuracy measure the coefficient of determination computed using the predicted values stored in the buffer.

5 Experimental Results

The experiments were performed on generic data, in contrast to the usual approach of using real-life time series. The main reasons for this decision were:

- a better understanding of the underlying phenomena;
- a rigorous control of regime switching between distributions;
- the possibility of computing the performance of an optimal predictor.

However, the time series were generated in such a way that they provide sufficient insight into real-life forecasting.

The time series used in our experiments were constructed by mixing data stemming from a deterministic chaotic process (Q) and a noisy, non-chaotic process (H), used earlier in [7]. The processes Q and H were generated according to the following rules:

$$x_{t+1} = 2(1 - x_t^2) - 1 \quad (Q)$$

$$x_{t+1} = \tanh(-1.2x_t + \epsilon_{t+1}) \quad (H),$$

where $\{\epsilon_t\}$ is a white noise process with mean 0 and standard deviation 0.32. For both processes, the initial values x_0 were taken from the [-1,1] range, and then the defining equations were applied repeatedly in order to generate additional process values.

Two different ways of mixing the Q and the H processes for building a time series were considered. A first time series (QHQ) was created by concatenating three data sections of lengths 300, 400, and 500 samples, respectively, in which the first and the last data segments stem from the Q process, whereas the second data segment stems from the H process. Similarly, a second time series (HQH) was created by concatenating data segments of lengths 300, 400 and 500 stemming from processes H, Q, and H, respectively. For conciseness, solely the results obtained on the QHQ series are presented here. For an in-detail

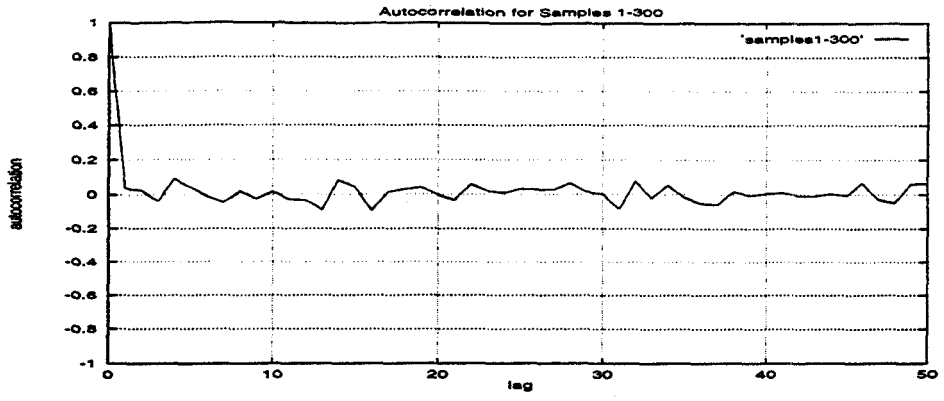

Figure 2: Autocorrelation for Data Samples 1-300

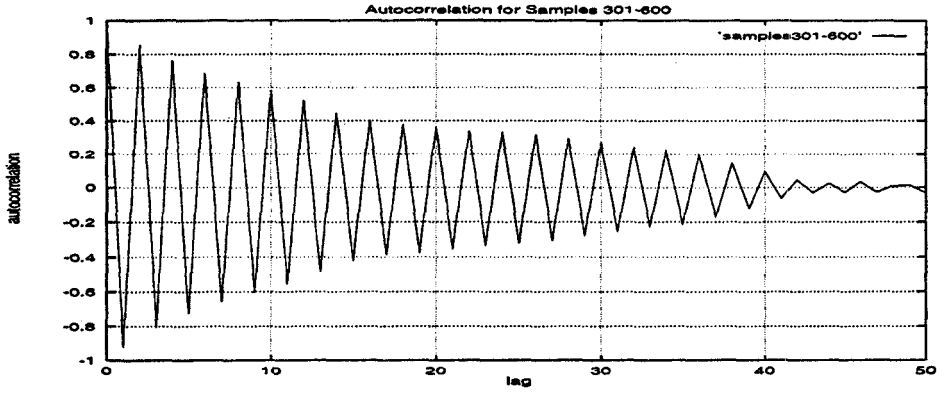

Figure 3: Autocorrelation for Data Samples 301-600

presentation of both the QHQ and the HQH series results, the reader should consult [3].

A segment of the QHQ time series comprising the first regime switch from process Q to process H (time series data samples 251-350) is presented in Fig. 1. Although the Q and H processes have basically the same means and variances, as well as data ranges, Fig. 1 illustrates the different time behavior of the two processes. Indeed, the autocorrelation plots for lags up to 50 on the first 300 and the next 300 time series data samples (shown in Figs. 2 and 3) indicate a dependence of autocorrelation on time origin, meaning that the underlying mixed time series is not wide-sense stationary [5].

To get insight into the robustness of our proposed methodology with respect to the data noise level, a high-noise time series was constructed by corrupting the QHQ time series with Gaussian additive noise of zero mean and standard deviation equal to half of the standard deviation of the uncorrupted data (the resulting time series, denoted as QHQ-N, is obviously not wide-sense stationary).

5.1 Low-Noise Experiments

Two feedforward neural networks M_1 and M_2, having 2 input units, two hidden layers of 4 units each and 1 output unit were trained (using the backpropagation algorithm) on two data segments of 200 samples each, stemming from the Q and the H processes, respectively. The statistics-based signaling technique compared windows of 200 samples using both chi-square and K-S tests. In the REUSE and RETRAIN scenarios, the threshold α employed for deciding when a new NN model must be trained was set to 0.8. An appropriate architecture, as well as an appropriate value for the parameter α were obtained through a reasonably short trial and error procedure and no claim is made that these are the optimal values.

The SWITCH Scenario

In our experiments, M_1 and M_2 represent the library models M_{h1} and M_{h2}, discussed in Section 4. The experiments compared the results obtained using

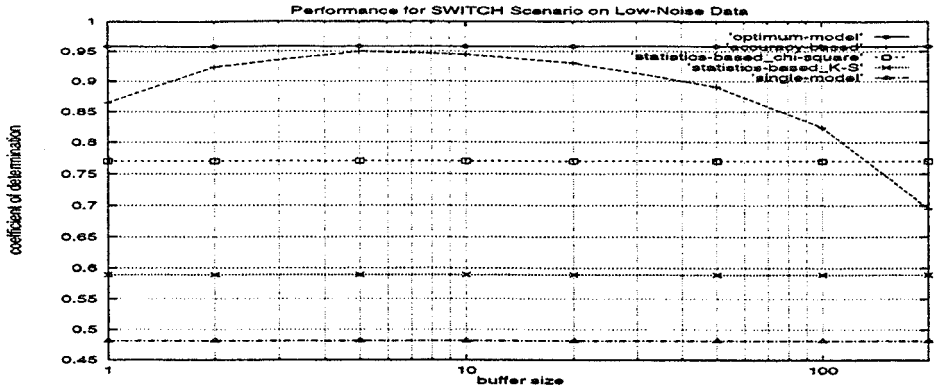

Figure 4: Performance for SWITCH Scenario on Low-Noise Series

statistics-based signaling and accuracy-based signaling to the results of both a single-model predictor as well as an optimal predictor. The single-model predictor represents a library model used for predicting the entire time series (M_1 in this case), whereas the optimal predictor is obtained by using both library models M_1 and M_2 and assuming that the switching points between distributions are detected without any delay (this is infeasible in practice unless the regime switching rules are entirely understood). The results obtained are shown in Fig. 4.

Although the statistics-based signaling technique yields a significantly better prediction accuracy as compared to using the single-model predictor, the results show that the accuracy-based signaling technique provides much better results as compared to the statistics-based one, using either the chi-square or the K-S tests. The accuracy-based signaling technique leads to excellent results for buffer sizes over a fairly wide range (2-30). It was also observed that small buffer sizes lead to performance that is comparable to that achieved when the regime switching points are completely known (optimal predictor curve).

The REUSE Scenario

In this case, M_1 corresponds to the library model M_h. The results obtained using the accuracy-based signaling technique for buffer sizes 50 and 100 are presented in Table 1.

The figures presented were obtained as averages over ten runs with different initial random weights for the retrained NNs. In all experiments included in Table 1 the standard deviations were very small. Consequently, by comparing Table 1 with Fig. 4, it can be concluded that even in the REUSE scenario the accuracy-based signaling technique yields significantly better results than the statistics-based signaling technique. For this reason, results obtained using the statistics-based signaling are not reported for the REUSE scenario. On the other hand, although the averaged value of the coefficient of determination (r^2) was larger when using a shorter buffer, a statistically significant difference cannot be claimed since the difference in r^2 mean values is smaller than the sum of the corresponding standard deviations. The number of NN retrainings observed in the experiments with buffer length 100 varied between 3 and 7, whereas it varied between 5 and 14 in the case of buffer length 50. These figures indicate that the experiments on longer buffers are computationally more efficient. However, even for the shorter buffer, the number of retrainings is very small as compared to the total number of predictions.

Buffer Size	r^2 Mean	r^2 Standard Deviation
50	0.866323	0.0117669
100	0.832089	0.0317474

Table 1: Performance for REUSE Scenario on Low-Noise Series

The RETRAIN Scenario

As in the REUSE scenario, the results presented in Table 2 were obtained using the accuracy-based signaling technique for buffer sizes 50 and 100.

The figures presented were obtained as averages over ten runs with different initial random weights for the retrained NNs. Again, the accuracy-based signaling results were better (with small standard deviation) than those obtained by the statistics-based sig-

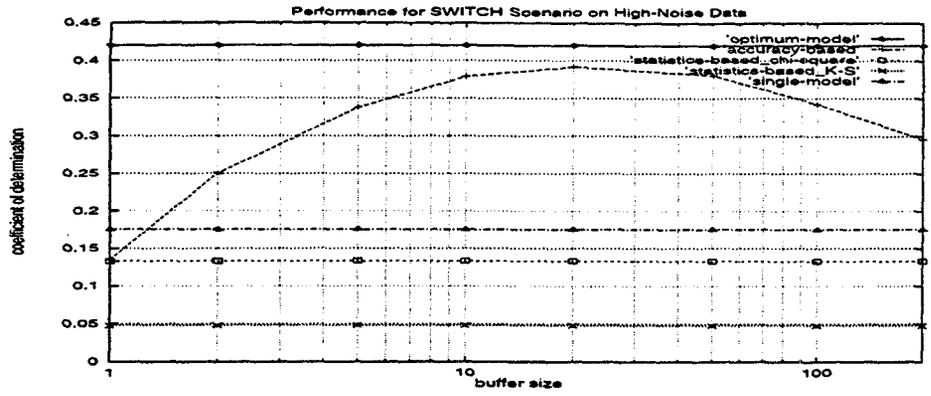

Figure 5: Performance for SWITCH Scenario on High-Noise Series

Buffer Size	r^2 Mean	r^2 Standard Deviation
50	0.801190	0.0285059
100	0.780692	0.0372022

Table 2: Performance for RETRAIN Scenario on Low-Noise Series

naling in the SWITCH scenario. Consequently, the statistics-based signaling was not considered in the REUSE scenario. The difference in performance for buffer sizes 50 and 100 was not statistically significant, while the experiments on longer buffer needed less computational resources (3 to 8 retrainings for buffer length 100, as compared to 6 to 16 retrainings for buffer length 50).

5.2 High-Noise Experiments

The NNs used in these experiments had the same architecture and parameters as those used in the low-noise experiments. However, the NNs M_1 and M_2 were now trained on two data segments of 200 samples each, stemming from the noise-corrupted Q and H processes. The window size used in the statistics-based signaling was again 200 samples, while the threshold α employed in the accuracy-based signaling was set to 0.6, smaller than in the low-noise experiments to account for the high noise level. Similar to the low-noise experiments, an appropriate architecture as well as an appropriate value for the parameter α were obtained through a reasonably short trial and error procedure and no claim is made that these are the optimal values.

In the SWITCH scenario, in spite of an extremely high noise level, the accuracy-based signaling technique lead once again to performance that was close to optimal. However, the statistics-based signaling technique was not only significantly less accurate, but not even consistently better than the single-model predictor that used a library model trained entirely on one distribution (see Fig. 5). As expected, due to a much larger amount of noise, the "optimal" buffer sizes for the accuracy-based signaling are larger as compared to the corresponding ones from the low-noise experiments (compare Fig. 5 with Fig. 4).

Buffer Size	r^2 Mean	r^2 Standard Deviation
50	0.271352	0.00888212
100	0.253184	0.0156959

Table 3: Performance for REUSE Scenario on High-Noise Series

Buffer Size	r^2 Mean	r^2 Standard Deviation
50	0.240188	0.0154151
100	0.222303	0.013072

Table 4: Performance for RETRAIN Scenario on High-Noise Series

Accuracy-based signaling results for the REUSE and the RETRAIN scenarios for high-noise experiments are again significantly better (with small standard deviation computed over 10 different runs) as compared to the statistics-based signaling results for the SWITCH scenario (compare Tables 3, 4 with Fig. 5). Consequently, without performing further experiments, it was possible to conclude that statistics-based signaling for the REUSE and the RETRAIN

scenarios were not appropriate. The number of NN retrainings in the high-noise experiments was consistently larger as compared to the low-noise ones (8 to 12 for buffer 100 and 15 to 23 for buffer 50, for the REUSE scenario, and 10 to 18 for buffer 100 and 23 to 33 for buffer 50 for the RETRAIN scenario). It can also be observed that the computational resources needed in the RETRAIN scenario were significantly larger than those needed in the REUSE scenario. However, once again the number of retrainings is reasonably small as compared to the length of the time series considered.

6 Conclusions and Further Research

This paper proposed an accuracy-based signaling technique for detecting changes in data distribution as an alternative to a statistics-based signaling technique. The obtained results provided strong support in favor of the accuracy-based signaling for non-stationary time series prediction. The method appeared to be applicable to both low-noise, as well as high-noise problems.

The proposed technique is using two domain-dependent parameters, buffer size b and threshold α. Appropriate parameter values can be obtained through a reasonably short trial and error procedure. Experiments indicated that the buffer size is fairly robust (satisfactory prediction results can be obtained with parameter values over a fairly large interval). Our hypothesis is that for non-malicious distributions the prediction accuracy is a convex function of the buffer length (see Figs. 2, 3, 6 and 7 for experimental support of this hypothesis). If so, the search for an optimal buffer length can be performed by starting with a small buffer length that is gradually increased until the predictor's accuracy stops increasing, as no larger buffer length is likely to yield any better prediction. This could be intuitively justified by observing that small buffers are unreliable due to imperfections in predictors and the existence of outliers, whereas large buffers result in long delays in distribution change detection. However, further research is needed in order to resolve this practically important hypothesis. A similar robustness analysis is required for a better understanding of the sensitivity threshold α.

Work on applying the proposed method to real-life time series prediction is currently in progress and results will be reported elsewhere. It is worth noting that in those problems different accuracy measures are more appropriate (e.g. in trading systems, the annualized rate of return should replace the r^2 accuracy measure used here). In this paper we investigated the possibility of reusing one or two historically successful models. However, it is worth mentioning that the approach proposed in this paper is easily extensible to a larger library of somewhat less successful, but promising, models that might be available in real-life time series predictions.

References

[1] S. M. Abecasis, and E. S. Lapenta, "Nonstationary Time-Series Forecasting Within a Neural Network Framework," *NeuroVe$t Journal,* Vol. 4, No. 4, 1996, pp. 9-16.

[2] G. Cybenko, "Approximation by Superpositions of a Sigmoidal Function," *Mathematics of Control, Signal, and Systems,* Vol. 2, 1989, pp. 303-314.

[3] R. Drossu and Z. Obradovic, "Regime Signaling Techniques for Non-Stationary Time-Series Forecasting," *NeuroVe$t Journal,* Vol. 4, No. 5, 1996, pp. 7-15.

[4] D. P. Helmbold, and P. M. Long, "Tracking Drifting Concepts by Minimizing Disagreements," *Machine Learning,* Vol. 14, No. 1, 1994, pp. 27-45.

[5] A. Papoulis, *Probability, Random Variables, and Stochastic Processes. Second Edition,* McGraw-Hill, 1984.

[6] W. H. Press et al, *Numerical Recipes in C. Second Edition,* Cambridge University Press, 1992.

[7] A. S. Weigend et al, "Nonlinear Gated Experts for Time Series: Discovering Regimes and Avoiding Overfitting," *International Journal of Neural Systems,* Vol. 6, 1995, pp. 373-399.

[8] P. Werbos, "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences," Doctoral Dissertation, Harvard Univ., Cambridge, Mass., 1974. Reprinted as *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting,* J. Wiley & Sons, New York, 1994.

[9] G. Widmer, and M. Kubat, "Learning in the Presence of Concept Drift and Hidden Contexts," *Machine Learning,* Vol. 23, 1996, pp. 69-101.