Forecasting the Unknown Dynamics in NN3 Database Using a Nonlinear Autoregressive Recurrent Neural Network

E. Safavieh, S. Andalib, and A. Andalib

Abstract—**In this paper, a nonlinear autoregressive (NAR) recurrent neural network is used for the prediction of the next 18 data samples of each time series in a set of 11 unknown dynamics in NN3 Database. The models are built on the reconstructed state spaces of data and no other domain knowledge is available to be used. Here, we clarify that the employed method is in part similar to a superior subclass of recurrent neural network, namely the nonlinear autoregressive model with exogenous inputs (NARX). Using the extensive available research about NARX networks, we briefly explain that our model is preferred to the both non-recursive and even other recurrent predictors, because of its intrinsic ability for learning long term dependencies in time series. As the desired values of the predicted time series are not available yet, no analysis have been performed on the presented results.**

I. Introduction

IN time series prediction, we wish to build a model that is responsible for generation of a given time series. However, when the underlying dynamics is affected by a set of explanatory variables, which are not known, the model that captures the dynamics cannot be identified directly. In these cases, a proper mapping of the observable output of the unknown dynamical system may be helpful for the prediction of the time series at hand. Identification of this mapping is discussed in the theory of dynamic reconstruction [1]. Using the result of this theory the analysis, e.g. prediction, of the observable time series is possible. This is valuable if the evolution of the points in the reconstructed state space tracks that of the unknown dynamics in the original state space. Under some conditions, Takens delay-embedding theorem [2] introduces a diffeomorphic map (one to one deferential mapping) as the result of dynamic reconstruction theory. According to this theory, with estimating two parameters of embedding dimension D_E and normalized embedding delay τ, and constructing a predictive mapping, dynamic reconstruction is achieved. The reconstruction consists of two steps: first, a delay line should be designed to latch the information for D_E points, and after that, a predictor that identifies the unknown mapping must be trained. This step is the heart of dynamic modeling.

Various nonlinear predictors are employed for the time series prediction task, including regression- and neural network-based algorithms. A review on some common methods for prediction of a chaotic time series, namely electricity price is presented in our recent work [3]. As it is described in [3], in many time series there is a very important characteristics, which is known as long-term dependencies, meaning that the temporal contingencies presented in the input/out sequences span long intervals. We have shown in[3], [4] that this is the case for financial time series and electricity price historical data.

It is mentioned in [5] that recurrent neural networks (RNNs) may be considered as a good choice for mapping input sequences to output sequences. In contrast, static systems, i.e. those with no recurrent connection, even if they include some lagged values of input data, have a finite impulse response and have difficulties with respect to RNN to store information for an indefinite time [5]. However, even simple recurrent networks, e.g. Elman RNN, are not ideal approaches for learning long-term dependencies because of the problem of *vanishing gradient*, meaning that under special condition given in section II, the fraction of error gradient due to information n time steps in the past exponentially decreases as n increases.

To tackle the problem of vanishing gradient, a class of recurrent neural networks, called nonlinear autoregressive model with exogenous inputs (NARX) is proposed [6], which has various advantages over simple recurrent networks. Not only has the NARX model less sensitivity to long-term dependencies [6], but also it has a very good learning capability and generalization performance [7].

It should be mentioned that NARX is different from a popular Auto Regressive (AR) model [8] which performs a simple linear transformation of the visited values of the time series. As it is mentioned in section II, the NARX is a nonlinear model which estimates the next values of the time series based on its last outputs instead of the actual measurements as is used in ARIMA models. Furthermore, the NARX uses a nonlinear structure, e.g. a neural network, for estimating the model's parameters. In contrast, the coefficients of a simple AR model are estimated using simple statistical methods like *least squares estimation*. Therefore, a NARX model enjoys a better generalization capability.

In this paper, we introduce the NARX model to use the research history that proves the advantages of this model over static and even simple recurrent structures. However,

Manuscript received January 31, 2007. This work is supported in part by Sepanta Robotics & AI Research Foundation (SRRF).

E. Safavieh, S. Andalib, and A. Andalib are with the SRRF, Tehran, 15757-18616, Iran (e-mail: e.safavieh@srrf.net, s.andalib@srrf.net, a.andalib@ srrf.net).

the time series that we are going to challenge in this paper are not multidimensional, in the sense that for each time series there is not any other explanatory data with values corresponding to the original one. Therefore, we are obligated to employ a simplified NARX network known as NAR. However, as the advantages of NARX are due to the feedback structure of this model, we are still interested in using a nonlinear autoregressive model without any exogenous inputs.

In this paper, we have employed aforementioned NAR model to forecast a set of 11 time series from the dataset of NN3 forecasting competition. As a preprocessing step, we first calculate some characteristics of these time series based on *Takens' embedding theorem* and the *theory of dynamic reconstruction*. The reconstructed state space is then used to generate 18-step ahead forecasts.

This paper is organized as follows. In the next section, we briefly review the architecture of NAR in the framework of NARX recurrent neural network and enumerate its advantages over other recurrent structures. In section III, we review the Takens' embedding theory. In section IV, we present the reconstructed dynamics of time series in NN3 database. In section V, we implement the methods explained in section II regarding the reconstructed state space addressed in section IV. Finally, the numerical results are presented in section VI.

II. NARX Recurrent Neural Network

A NARX model is a class of discrete-time nonlinear autoregressive systems, which has endogenous inputs as well as exogenous inputs, and can be stated as:

$$\hat{y}(t+1) = f(y(t),\cdots,y(t-D_y), u(t),\cdots,u(t-D_u)) + \varepsilon_t \quad (1)$$

here $\{y(t)\}$ is the time series of interest that should be predicted and $\{u(t)\}$ is another time series with terms associated with that of $\{y(t)\}$. The terms $u(t),\cdots,u(t-D_u)$ are the exogenous inputs and may be produced with an input delay line with memory of order D_u. Similarly $y(t),\cdots,y(t-D_y)$ are the endogenous inputs and may be produced with a delay line memory of order D_y. f is some nonlinear function, e.g. a multi layer perceptron (MLP), that estimates the next value of $\{y(t)\}$, $\hat{y}(t+1)$ and ε_t denotes the additive noise of the estimation. The architecture of a NARX recurrent neural network is shown in Fig. 1. The first layer of the MLP network consists of $D_u + D_y$ buffer neurons, corresponding to the outputs of the two delay lines.

Considering the above architecture, we define the cost function E at the time t as:

$$E_t = 1/2(\hat{y}(t) - y(t))^2 \quad (2)$$

We may mathematically explain the effect of vanishing

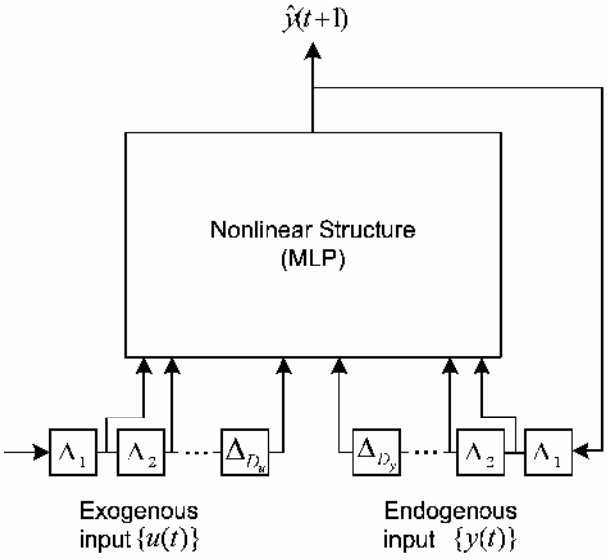

Fig. 1. A NARX recurrent neural network.

gradient on the derivatives of E_t with respect to the weight vector $\mathbf{w}$:

$$\frac{\partial E_t}{\partial \mathbf{w}} = (\hat{y}(t) - y(t))\frac{\partial \hat{y}(t)}{\partial \mathbf{w}} = (\hat{y}(t) - y(t))\sum_{\upsilon \prec t}\frac{\partial \hat{y}(t)}{\partial \hat{y}(\upsilon)}\frac{\partial \hat{y}(\upsilon)}{\partial \mathbf{w}} \quad (3)$$

It has been shown that if a network is to latch the information robustly, i.e. if it is to store information for a long period of time in the presence of noise, then for a term with $\upsilon << t$, $\left|\partial \hat{y}(t)/\partial \hat{y}(\upsilon)\right| \to 0$ [5]. In this condition, the gradient decays exponentially [6], meaning that there is not any chance for the terms that are far from t to change the weights in such a way that allow the network's state to jump to a better basin of attraction.

The above scenario comes true for all recurrent structures. However, one can postpone vanishing of gradient in NARX recurrent neural networks with increasing the number of delays in the output delay line of this architecture [6]. As it may be seen in Fig. 1, the output delay line of NARX networks, as jump-ahead connections, provides shortcuts for propagating gradient information more efficiently when the network is unfolded in time. In an unfolded recurrent neural network, the hidden units from the pervious states are considered as an additional set of inputs. This property makes NARX a proper tool for modeling dynamics that exhibit long-term dependencies.

As it is mentioned in section I, the NARX which is used in this paper does not have any exogenous inputs.

III. Embedding Theory and Dynamic Reconstruction

By the results of Takens' embedding theorem [9] and the theory of dynamic reconstruction [1], it is able to reconstruct the dynamics of interest, which were initially unobservable due to various, and maybe unknown parameters that effect on them.

Takens' theorem implies that by the means of two parameters of embedding dimension D_E and normalized embedding delay τ, it is guaranteed that the evolution of the points in the new state space follows that of the original state space. In this way the analysis, e.g. prediction of the new state space is fruitful.

Suppose the time series $\{y(t)\}$ as the observable output of the unknown dynamical system $\{x(t)\}$; $\{y(t)\}$ may be defined as follows:

$$y(t) = g(x(t)) \tag{4}$$

where $g(\cdot)$ is a scalar-valued function.

According to Takens' theorem, we can define D-dimensional reconstructed dynamics $\{\mathbf{y}_R(t)\}$ by the following equation:

$$\mathbf{y}_R(t) = [y(t), y(t-\tau), ..., y(t-(D-1)\tau], \quad D \geq 2d+1 \tag{5}$$

where d is the state space dimension of the unknown dynamics.

As it is mentioned above, the evolution of the points $\mathbf{y}_R(t) \to \mathbf{y}_R(t+1)$ tracks that of the unknown dynamics $x(t) \to x(t+1)$. So to challenge a problem concerned with the prediction of the time series $\{x(t)\}$, it is fruitful to predict the time series $\{y(t)\}$. This can be satisfied by a nonlinear model of $f: \Re^D \to \Re^1$, which performs the following mapping:

$$\hat{y}(t+1) = f(\mathbf{y}_R(t)) \tag{6}$$

where $\hat{y}(t+1)$ is the one-step ahead predicted value of the time series $\{y(t)\}$.

Equation (6) may be extended to τ-step mapping with a different model of f:

$$\hat{y}(t+\tau) = f(\mathbf{y}_R(t)) \tag{7}$$

There are many methods for estimating the embedding parameters. One of the methods for choosing the embedding delay is to choose the first point of autocorrelation function that goes bellow *zero zone* [10]. In [11] it is selected as the reciprocal of the highest relevant frequency of the time series. It is proposed in [12] to use the *information dimension*, which is a kind of fractal dimensions, as the time delay. However, one of the most reliable methods is presented by Fraser [13]. Due to this method known as average mutual information (AMI), the normalized embedding delay τ is heuristically set to the value in which the mutual information between $y(t)$ and $y(t-\tau)$ attains its first minimum. This way, the values $y(t)$ and $y(t-\tau)$ are essentially independent of each other in the sense that they may serve as two coordinates of the reconstructed space. However, they are not so independent as to have no correlation with each other.

There are also some methods for choosing embedding dimension D_E such as correlation dimension and the *Integral Local Deformation* (ILD) algorithm [14]. One of the best and simplest methods is presented in [15]. According to this method, the minimum acceptable value of D, denoted by D_E, is determined by looking at the first local minimum of the *false nearest neighbors* (FNN) under changes in the embedding dimension from $D \to D+1$.

In the next section, we calculate the embedding parameters of NN3 time series.

IV. Reconstructed Dynamics of the time Series in NN3 Database

We have calculated the parameters of τ and D_E for the set of 11 time series, including NN3-101 through NN3-111. It should be mentioned that due to the high nonstationarity of four time series, namely NN3-101, NN3-105, NN3-108 and NN3-109, first order differencing is used for detrending the raw data. By detrending, we mean eliminating the long-term trends of the signal, leaving only short-term oscillations. In fact, for these time series the reconstruction is performed on this secondary signal. Here, the normalized embedding delay is calculated by the autocorrelation function of each time series and the first point bellow zero zone is considered as the value of τ. Although, the method of average mutual information (AMI) is more reliable; however, it requires more historical data that is not available in our experiments. The embedding dimension is also calculated by the method of false nearest neighbors (FNN). We have depicted the autocorrelation and[15] FNN function for the time series NN3-106 in Fig. 2 and Fig. 3 respectively. The complete results for all of the time series in the database are reported in Table I.

V. Algorithm Implementation

To generate the prediction of the next 18 observation, NAR networks are used to model the functions f defined in (7). The reconstructed state space of each data set is used here as the endogenous inputs. The reconstruction vectors $\mathbf{y}_R(t)$ defined by (5) is of dimension D_E. Furthermore, the size of output delay line memory required to perform the embedding, is $\tau \times D_E$. However, the delay line memory is only required to provide D_E outputs. Therefore, we use τ equally spaced taps, representing spars connections to the nonlinear structure of NAR. The embedding parameters as well as the nonlinear structures, which are used for different time series in NN3 database are mentioned in Table I.

As the training sets of the given time series are extremely

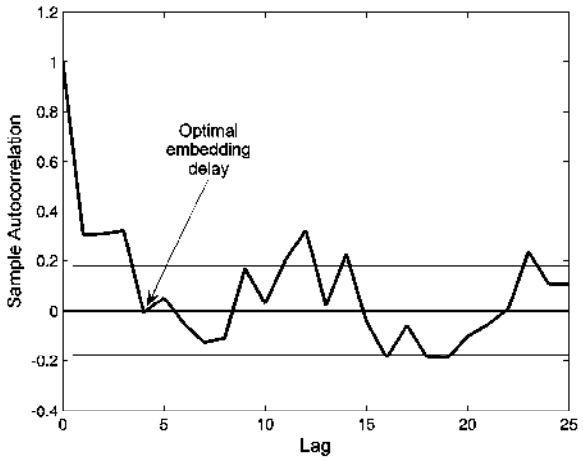

Fig. 2. The autocorrelation function of time series NN3-106. The first point bellow zero zone is considered as the value of normalized embedding delay.

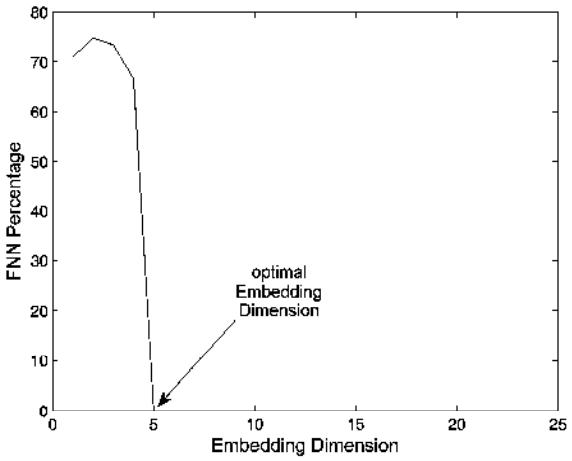

Fig. 3. The percentage of false nearest neighbors for data points of NN3-106 time series. The location of the first minimum is reported as the minimum acceptable value of embedding dimension.

TABLE I
THE EMBEDDING VARIABLES, EMPLOYED ARCHITECTURES, AND PREDICTION RESULTS FOR THE 11 TIME SERIES

Time Series	τ	D_E	Network	Train NMSE	Validation NMSE
NN3-101	3	4	[De 3 5 1]	0.049531	0.077242
NN3-102	6	4	[De 5 1]	0.042137	0.119053
NN3-103	6	8	[De 6 2 1]	0.058877	0.94272
NN3-104	6	4	[De 5 3 1]	0.002938	0.025725
NN3-105	3	3	[De 5 1]	0.130188	1.686444
NN3-106	4	5	[De 5 3 1]	0.016508	0.299395
NN3-107	4	4	[De 2 6 1]	0.015089	0.851522
NN3-108	3	3	[De 6 2 1]	0.426625	1.732382
NN3-109	4	6	[De 5 1]	0.181523	1.388405
NN3-110	5	3	[De 5 1]	0.052671	0.055022
NN3-111	3	3	[De 3 1]	0.114562	10.92656

TABLE II
FORECASTS RESULTS FOR THE NEXT 18 DATA SAMPLES OF EACH TIME SERIES IN NN3 DATASET

Time Series#	101	102	103	104	105	106	107	108	109	110	111
t+1	5236	4074	3953	7441	4630	4765	3782	2237	3590	2773	3184
t+2	5246	9107	39163	7400	4731	5311	3757	2837	3734	2613	3513
t+3	5040	9099	39350	7287	4827	4777	3703	3025	3702	2546	2404
t+4	5219	9042	32050	6925	4809	4431	3640	3423	3816	2494	3114
t+5	5007	8945	13489	6910	4861	4368	3708	3066	3894	2539	3138
t+6	5185	8744	4586	3438	5015	4673	3804	3195	3901	2121	2521
t+7	5067	8440	4127	3079	4974	4305	3763	4285	3804	2341	2496
t+8	4847	7967	3928	3630	4962	4476	3804	3343	3968	2650	2333
t+9	5261	7580	3921	6571	5117	5284	3777	3358	3914	2711	3163
t+10	5037	7219	4261	7176	5092	4190	3760	3710	3877	2667	2507
t+11	5109	6848	4552	7231	5044	5379	3795	3398	3927	3038	2591
t+12	5036	6282	4111	7408	5125	5172	3821	3419	3965	2839	2854
t+13	5146	5732	4099	7400	5133	4771	3786	3146	3945	2575	3076
t+14	5116	9078	36065	7410	5084	5403	3759	3562	3988	2541	3180
t+15	4923	9114	36168	7315	5082	4771	3759	3637	4147	2563	2427
t+16	5139	9118	31001	7146	5129	4799	3772	3043	4108	2492	2859
t+17	4922	9099	7046	7028	5116	4443	3759	3783	4209	2501	2817
t+18	5076	9041	4161	3702	5048	4747	3775	3827	4322	2615	2776

small, we have just considered 10% of each data set as validation data. The validation data in these experiments play a very important role. As the desired values of the time series are not available, to achieve the best results, we have trained many NAR networks in each case. Then, the network with the best performance over validation data is used to generate the forecasts for the next 18 data samples.

VI. Forecasting Results

The normalized mean square error (NMSE) values of the predictions are reported in Table I. Furthermore, the forecasted time series are presented in Table II. As the desired outputs of each network are not available yet, we have not been able to do any analysis on the results.

References

[1] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Englewood Cliffs, NJ: Prentice Hall, 2nd ed., 1999, ch. 14.

[2] F. Takens, "On the numerical determination of the dimension of an attractor," in D. Rand, and L. S. Young, eds., *Dynamical Systems and Turbulence,* Annual Notes in Mathematics, vol. 898, pp.366-381, Berlin: Springer-Verlag.

[3] A. Andalib, S. E. Safavieh, and F. Atry "τ-Step Ahead Forecasts for the Hourly Ontario Energy Price: A Comparative Study, A Critical Analysis of Current One-Step Ahead Forecasts," *IEEE Trans. Power Syst*., , submitted for publication

[4] A. Andalib, A. R. Sharafat, M. R. Zakeri Nasab, and S. E. Safavieh, "Optimal embedding of NARX networks for learning long-term dependencies in noisy and nonstationary time series," *IEEE Trans. Neural Netw.,* submitted for publication.

[5] Y. Bengio, P. Simard, P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 10, pp. 157-166, Mar. 1994.

[6] T. N. Lin, B. G. Horne, P. Tino, and C. L. Giles, "Learning long-term dependencies in NARX recurrent neural networks," *IEEE Trans. Neural Netw.*, vol.7, no. 6, pp.1329-1337, Nov. 1996.

[7] T. N. Lin, B. G. Horne, and C. L. Giles, "How memory order effect the performance of NARX networks," Inst. Adv. Comput. studies, Univ. Maryland, Collage Park, Tech. Rep. UMIACS-TR-96-76 and CS-TR-3706, 1996.

[8] G. Box, G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, second edition. Oakland, CA: Holden-Day, 1976.

[9] F. Takens, "Detecting strange attractors in fluid turbulence" in D. Rand, and L. S. Young, eds., *Dynamical Systems and Turbulence,* Berlin: Springer-Verlag.

[10] A. M. Albano, J. Muench, C.Schwartz, A. I. Mees, P.E. Rapp. "Singular value decomposition and the Grassberger-Procaccia algorithm". Phys. Rev. A, 38:3017-3026, 1988.

[11] D. S. Broomhead, G. P. King, "On the qualitative analysis of experimental dynamical systems", In S. Sarker, editor, *Nonlinear phenomena and mental chaos*, page113. Adam Hilger, Bristol, 1986.

[12] W. Liebert, H. Schuster, "Proper choice of the time delay for the analysis of chaotic time series" Physics Letters A, Volume 142, Issue 2-3, p. 107-111.

[13] A. M. Fraser, H. L. Swinny. "Independent coordinates for strange attractors from mutual information". Phys. Rev. A, 33:1134-1140, 1986.

[14] T. Buzug, G. Pfister, "Optimal delay time and embedding dimension for delay-time coordinates by analysis of the global and local behavior of strange attractors", Phys. Rev. A. 45:7073-7084, 1992

[15] M. B. Kennel, R. Brown, and H. D. I. Abarbanet, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Phys. Rev. A 45*, pp. 3403-3411, 1992.