Time Series forecasting with Recurrent Neural Networks NN3 Competition

Mahmoud Abou-Nasr
Research & Advanced Engineering
Ford Motor Company
Email: mabounas@ford.com

Abstract--- **In the NN3 competition, 111 time series of different number of points are given to the participants. The objective is to predict the values of the eighteen points which follow the given points in each series. This paper presents the approach used to predict the required points. For each of the provided series, eighteen recurrent neural networks were trained. The respective networks from one to eighteen were trained to predict one to eighteen steps ahead. All the networks used were comprised of one input node, one recurrent layer of nodes of sigmoidal activation and one unit of delay between them, one none recurrent layer of nodes with sigmoidal activation and one output linear node.**

I. INTRODUCTION

In the NN3 forecasting competition, samples of points of different time series are provided, and the participants were asked to forecast the next eighteen points in the respective series.

Forecasting has practical applications in many fields like forecasting consumption demand of a certain product or commodity, forecasting the weather conditions in a certain area or forecasting the performance of a stocks to name a few. Not surprisingly, researchers have studied this problem extensively. Some of the challenges of time series prediction are:

- Most of the time no underlying model is known in advance.
- The time series could be highly nonlinear or chaotic.
- The time series could be non-stationary.
- The data representing the time series could be very noisy.
- The sampling rate of the data could be inadequate.
- Scarcity of data necessary to represent the underlying dynamics.
- Erroneous or missing data in the sample representing the time series.

There is a plethora of techniques in the literature that tackle the prediction of points in a time series e.g. feedforward neural networks, recurrent neural networks, radial basis functions neural networks and support vector machines. All of these methods have been shown to be universal function approximators.

Nonlinear estimation techniques have to deal with the following challenges:

- The over fitting problem: the function F estimates the given sample data adequately but does not estimate new data as well.
- Nonlinear optimization techniques seeking global minima for the error functions can find local minima instead.
- They require intensive computation.

The rest of this paper is organized as follows. Section II is a brief overview of the recurrent neural network structures, the type of recurrent networks used in this competition, and some of the training heuristics used in this work. Section III describes the approach followed to forecast the eighteen points in the 111 time series of the NN3 competition. Section IV concludes the paper with a summary.

II. RECURRENT NEURAL NETWORKS

Multilayer perceptrons (MLP) allow only feedforward connections between each neurons and the neurons in the following layer. Recurrent neural networks in contrast allow arbitrary connections between neurons, both forward and recurrent (feedback). A nonlinear mapping obtained by a recurrent neural network is not only dependent on the current input, but also is dependent on the previous inputs through the feed back connections to the input.

II.1. Structures of Recurrent Neural Networks

- One of the first proposed recurrent neural network structures is the Elman Jordan network. This is an MLP with a single hidden layer. It also contains a layer called context layer which contains the previous activations of hidden and output neurons respectively. Networks of this structure were successful in finding temporal patterns while learning formal grammars [1, 2].

- Another structure is the diagonal recurrent neural network where recurrent connections are only used in the in the diagonal of the weight matrix.

The recurrent neural network used in this work is a special class of recurrent networks where recurrent connections are only allowed within the nodes of the hidden layers. A simplified diagram with two hidden layers is shown in Fig. 1.

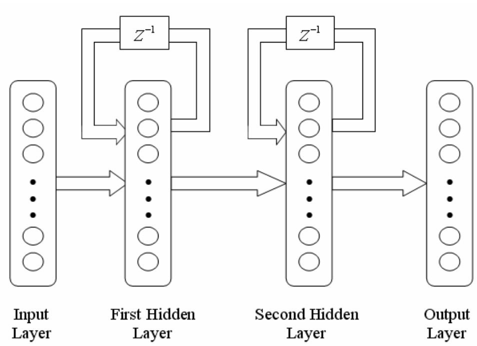

Figure 1: A Recurrent Network

The training algorithm used for training the recurrent network in this work is based on the Extended Kalman Filtering algorithm (EKF). Details on the EKF training can be found in [3]. The neural networks simulations and their training were performed using the neural network software tools developed by Dr. Lee Feldkamp and the neural network group in Ford Scientific Research Labs. Following are some of the heuristics that we followed while training the neural networks for the NN3 forecasting:

- The use of multistream training:

Some of the training algorithms may cause the network to perform well in the neighborhood of the most recent pattern it was trained on, and perform poorly on patterns it learned earlier in the training process. This is called the recency effect which in essence is saying that the network is not as good in modeling long time dependencies between the sequence elements. To overcome this problem multistream EKF training is used which is a combination of sequence scrambling and batch mode training [3]. The weight updates are performed after presenting the network with streams from different regions of the input data. The starting point of each stream is selected at random at each presentation.

- Data scaling:

Training data was scaled to the range of [-1, 1]. This is to avoid saturating the nonlinearities of the nodes activation functions.

- Training and Testing Data

One tenth of the data at the end of each series was set aside for testing the respective neural network predictors.

- Early stopping of the training:

Too long of a training may result in over fitting. It is suggested that the training is stopped as soon as the performance on the testing data is not improving.

III. APPROACH

A recurrent neural network with the following general architecture was used in this work:

- One input node
- Two hidden layers
    i. The first is comprised of recurrent nonlinear nodes with sigmoid activation functions and one step delay between nodes.

    ii. The second is comprised of nonlinear nodes with sigmoid activation functions and no feedback.

- One linear output node

The number of the nodes in each of the hidden layers was determined experimentally for each time series.
Around 10 percent at the end of each time series were set aside for testing purposes. For each time series, neural networks with different number of nodes in the hidden layers were trained on the training set and tested on the testing set. The smallest size neural network that performed well in the training as well as in the testing was chosen for the forecasting task. An ensemble of ten networks of this same architecture and the same was formed. The number of the networks in the ensemble is an arbitrary decision made for the purpose of managing the computational workload and is provided in this paper as a documentation of the exact procedure followed by the author. Whether a smaller or a bigger number of networks has any significance on the predicted results is a valid question one might ask, but could not be answered at this point without knowing how close the current predictions are to the actual points. Five of the networks in the ensemble were trained on all the sample points of the time series (training data and testing data). The other five was trained only on the training data. The final prediction of the ensemble is the average of all the predictions of the individual networks, with all the predictions of the individual networks having equal weights.

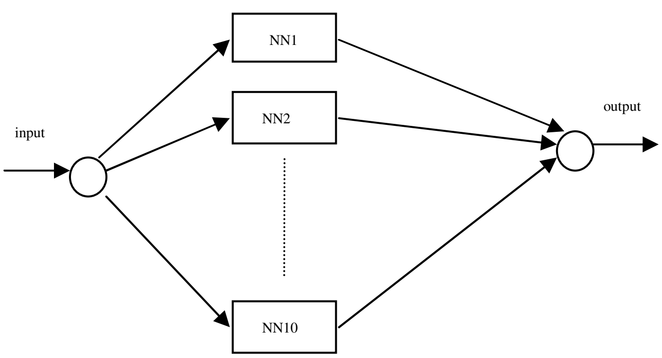

Figure 2: Ensemble of 10 Recurrent Neural networks. The output represents the prediction of the ensemble and is the average of the predictions of the individual networks.

Eighteen ensembles were formed following the above procedure to predict the required eighteen points for each of the 111 series of the competition. Ensemble one predicts one step ahead and is used to predict the first missing point. Ensemble two predicts two steps ahead and is used to predict the second missing point, and so on with ensemble eighteen predicting eighteen steps ahead and is used to predict the eighteenth missing point.

IV. CONCLUSION

An approach to forecast eighteen points at the end of 111 time series of the NN3 competition is provided in this paper. The approach is based on recurrent neural networks trained with EKF multistream training. Recurrent neural networks have been used successfully in many real world problems. One of the advantages of the technique used in this paper is that it requires minimal preprocessing of the series data. Also it requires no pre-assumption of the model underlying the data. In the absence of strong general nonlinear theory offering suggestions on which technique is better in what situation, experimentation and heuristics are relied upon to guide the choice of the methodology, architecture and size of the neural networks in this work.

REFERENCES

[1] J. L Elman. Finding structure in time. *Cognitive Science,* 14(2): 179-211, 1990.

[2] S. Haykin. *NeuralNetworks, a Comprehensive Foundation.* Prentice Hall, 1999.

[3] L.A. Feldkamp and G.V Puskorious. A signal processing framework based on dynamical neural networks with application to problems in adaptation, filtering and classification. In *Proceedings of the IEEE,* volume 86, pages 2259-2277, 1998.