Server Load Prediction Based on Dynamic Neural Networks

Ghannam Aljabari
Supervisor: Hashem Tamimi
Master of Informatics
Deanship of Gradate Studies and Scientific Research
Palestine Polytechnic University

Introduction

Predicting server load is involved in distributed system applications such as load balancing and sharing. Load balancing has been widely adopted in enterprise data centers for distributing workload across multiple servers or nodes in order to provide IT services for a large number of online clients. Many machine learning methods have been applied for load prediction. However, some researches show that applying Neural Networks (NN) technique is more efficient in predicting the load in future time. This paper is to investigate and compare two dynamic NN models in server load prediction including Time-Delay NN (TDNN) and Nonlinear Autoregressive Network with eXogenous inputs (NARX).

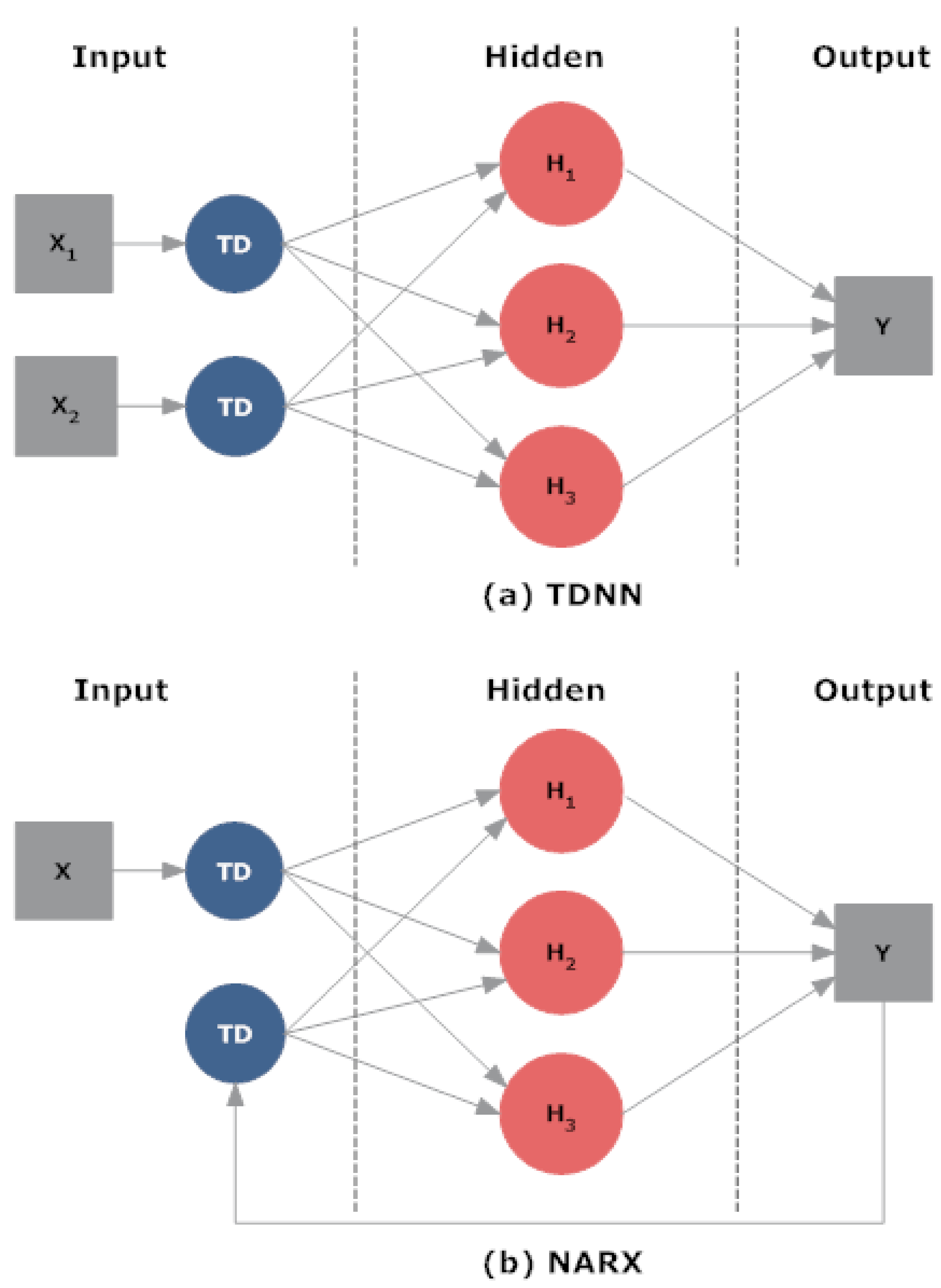

Figure 1: Dynamic Neural Networks

Results

Two set of experiments was conducted to investigate forecasting with NN. In TDNN set of experiments, the load average was used only to predict the next value. While in NARX model, another time sequences were used in addition to load average , in our case total number of processes and memory usage.

Figure 2 shows the predicted values of the server load in the last day with the actual values based on TDNN. The figure show that NN is able to accurately forecast the next value of the server load. The best result was obtained with maximum time delay of 6 and 0.0257 prediction error.

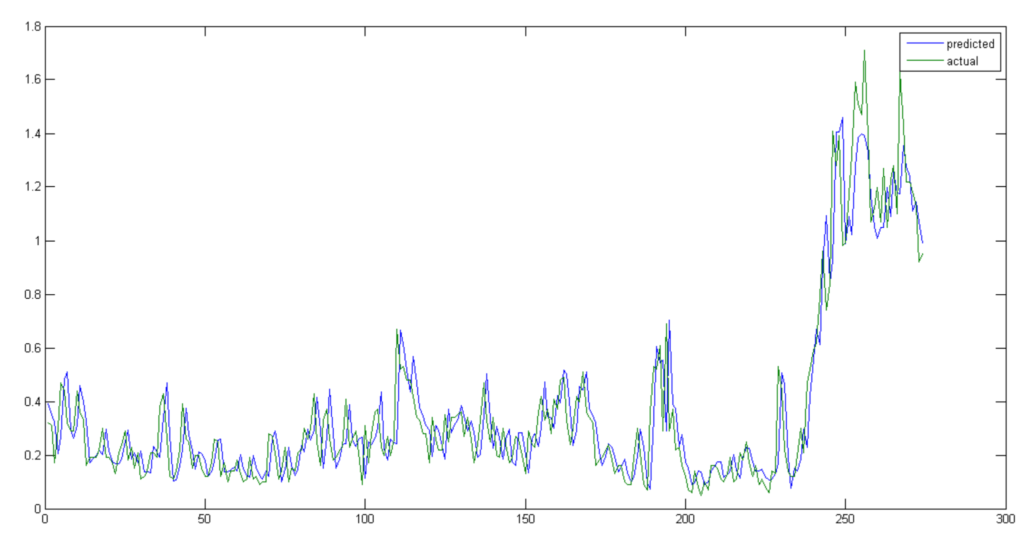

Figure 2: TDNN predicted and actual data

Figure 3 shows the predicted values of the server load in the last day with the actual values based on NARX. The best result was obtained with maximum time delay of 6 and 0.0244 prediction error .

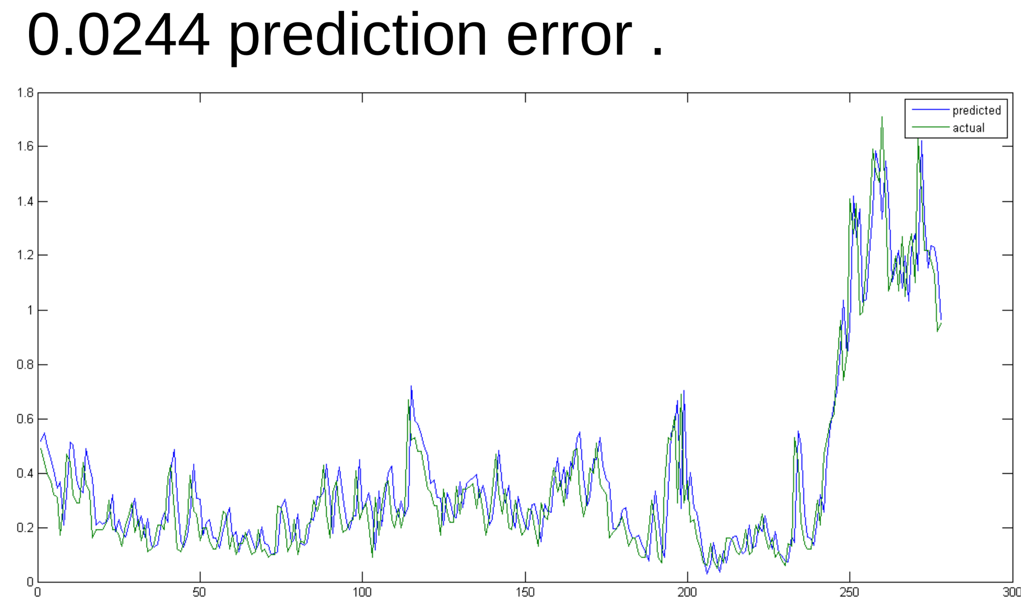

Figure 3: NARX predicted and actual data

Server Load Prediction

The goal of server load prediction is to estimate some future load values based on current and past load data. Mathematically:

$$y'(t) = f(y(t-1), y(t-2), y(t-3), \ldots)$$

Where *t* represents elapsed time , *y(t)* represents the time sequence of the load, and *y'(t)* is the predict value of the load at future time point.

Proposed Solution

Dynamic Neural Networks have been proposed for server load prediction. The reason of selecting dynamic NN is the ability of this approach to accurately forecast the non-linear pattern in load data.

In addition, dynamic NN can be more compact and hence faster to evaluate than other approaches such as Support Vector Machine (SVM) .

Two dynamic NN models have been applied for server load prediction: the TDNN and NARX, as shown in Figure 1.

Methodology

1. The load data was obtained from Webmail server of Palestine Polytechnic University for 7 days and with interval of 5 minutes (2016 sample).
2. The data includes load average, memory usage and total number of processes running on the server.
3. The samples of 6 days have been taken as training data and the 7th day as testing data.
4. The load data was normalized to take values between zero and one and then converted to a time sequence.
5. Root Mean Squared Error (RMSE) was used to compute the prediction error. The prediction error represents the difference between the predicted value and the actual value.
6. In all NN models, 10 neurons was used in the hidden layer, and the default Levenberg-Marquardt (LM) function was used for training.
7. The average prediction error was generated with 7-fold cross validation.

Conclusion

Two dynamic NN models have been investigated for time-varying data. NARX model provide better results in comparison to TDNN model in server load prediction.