GÜFBED/*GUSTIJ* (2018) CMES 2018 SEMPOZYUM EK SAYISI: 1-6
DOI: 10.17714/gumusfenbil.430137 Araştırma Makalesi / *Research Article*

Development of Internet Traffic Prediction Software Using Time-Series Multilayer Perceptron

Zaman Serili Çok Katmanlı Algılayıcı Kullanılarak Internet Trafik Tahmini Yazılımı Geliştirilmesi

Murat Can YÜKSEL[1,a], Mehmet Fatih AKAY[*1,b], Selami ÇİFTÇİ[2,c]
[1]Çukurova Üniversitesi, Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, 01250, Adana
[2]Türk Telekom Araştırma ve Geliştirme Merkezi, 34394, İstanbul

• Geliş tarihi / *Received*: 03.06.2018 • Düzeltilerek geliş tarihi / *Received in revised form*: 01.08.2018 • Kabul tarihi / *Accepted*: 12.09.2018

Abstract
Internet traffic prediction plays a fundamental role in network design, management, control and optimization. Although there exist several studies in litereture that focus on predicting Internet traffic using statistical and machine learning methods, to the best of our knowledge, a fully functional off-the-shelf software with different optimization capabilities has not been developed. The purpose of this study is to develop a new software for prediction of Internet traffic data based on time-series Multilayer Perceptron (MLP). The software includes features such as the optimization of the number of hidden layers and neurons in each layer and feedback delay optimization with respect to autocorrelations. The Internet traffic data from two different Internet Service Providers, varying by 1-hour and 5-minute time frequencies, have been used for testing the software. The datasets have been split into training and testing sets via 70-30% and 80-20% split ratios. The Mean Absolute Percentage Error (MAPE) has been utilized as the main error rate metric in order to evaluate the accuracy of the prediction models. It has been observed that the MAPE's of the Internet traffic prediction models change between 3.25 and 9.09. One can conclude that the developed software can be used for Internet traffic prediction within acceptable error rates.

Keywords: Multilayer perceptron, Traffic prediction, Time series

Öz
Internet trafik tahmini ağ tasarımı, yönetimi, kontrolü ve optimizasyonunda temel bir rol oynamaktadır. Literatürde Internet trafiğinin istatistiksel ve makine öğrenme yöntemleri kullanılarak tahmin edilmesi üzerine odaklanan çeşitli çalışmalar bulunmasına rağmen, bilgimiz dahilinde farklı optimizasyon seçeneklerine sahip, tam işlevsel bir yazılım geliştirilmemiştir. Bu çalışmanın temel amacı, Internet trafik tahmini için zaman serili Çok Katmanlı Algılayıcı (Multilayer Perceptron, MLP) üzerine kurulu yeni bir yazılımın geliştirilmesidir. Yazılımın içerdiği özellikler arasında gizli katmanların ve her bir katman içerisindeki nöron sayısının optimize değerlerinin bulunmasının yanı sıra en uygun zaman gecikmesi değerlerinin otokorelasyonlara bağlı olarak optimize edilmesi de bulunmaktadır. Yazılımın test edilmesi amacıyla iki farklı İnternet Servis Sağlayıcısı'ndan tedarik edilen, zaman frekansları 1 saat ve 5 dakika şeklinde değişen Internet trafiği verileri kullanılmıştır. Veri setleri %70-30 ve %80-20 oranlarında eğitim ve test verileri olmak üzere iki kısma bölünmüştür. Tahmin modellerine ait performansın değerlendirilmesi amacıyla, ana metrik olarak Ortalama Mutlak Yüzde Hata (Mean Absolute Percentage Error, MAPE) hesaplanmıştır. Internet trafik tahmini modellerine ait MAPE değerlerinin 3.25 ve 9.09 arasında değiştiği gözlemlenmiştir. Elde edilen sonuçlara göre, geliştirilen yazılım kabul edilebilir hata oranları ile Internet trafiğinin tahmin edilmesi amacıyla kullanılabilir.

Anahtar kelimeler: *Çok katmanlı algılayıcı, Trafik tahmini, Zaman serileri*

[*b] Mehmet Fatih AKAY; mfakay@cu.edu.tr; Tel: (0322) 338 71 01; orcid.org/0000-0003-0780-0679
[a] orcid.org/0000-0003-4846-849X [c]orcid.org/0000-0003-0780-0679

1. Introduction

Internet traffic can be described as a general term including the transmission of any type of Internet data among different devices and systems. The timely and accurate prediction of Internet data usage is a topic which has great importance in both literature and industry. One of the most popular methods used for this purpose is Time Series Analysis, a research concept about processing data collected in certain time periods.

There exist several studies in literature about the prediction of the Internet data traffic with different methods. Table 1 displays a summary of these studies.

Table 1. Literature studies on network traffic prediction

Study	Methods
Jiang et. al., 2009	ARMA,
Chabaa et. al., 2010	MLP, LM,
Liu et. al., 2011	CTSA, SVM
Cortez et. al., 2012	Naïve
Oliviera et. al., 2014	MLP, SAE
Katris et. al., 2015	FARIMA,
Daly, 2016	NAR
Bian et. al., 2017	LSTM
Sahrani et. al., 2017	NARMA
Narejo et. al., 2018	DBN

ANN, Artificial Neural Network; **ARMA,** Auto-Regressive Moving Average; **CTSA**, Chaotic Time Series Analysis; **DBN,** Deep Belief Network; **FARIMA**, Autoregressive Fractionally Integrated Moving Average; **LM**, Levenberg-Marquardt; **LSTM,** Long-Short Term Memory, **MLP,** Multilayer Perceptron; **NAR,** Non-Linear Autoregressive Neural Networks; **NARMA,** Non-Linear Autoregressive Moving Average; **RBP**, Resilient Back Propagation; **SAE**, Stacked Autoencoder; **SVM**, Support Vector Machine;

Although there exist several academic studies on Internet traffic prediction, the efforts are limited to developing prediction models with various methods and doing comparisons. The main purpose of this study is to develop a standalone desktop application which can perform time-series prediction using MLP, a very common machine learning method that is often used for neural network-based studies in literature. It is also aimed to develop the software with a variety of options including customization of train/test ratio, optimization of the number of hidden layers, the amount of neurons in each layer, feedback delays and training functions.

Section 2 gives brief information about the datasets and software development. Section 3 presents the details of the software features and sample results. Section 4 concludes the paper.

2. Datasets and Software Development

2.1. Dataset Generation

In this study, four data sets that have been obtained from two different Internet Service Providers have been used in order to create and evaluate prediction models. Two out of the four data sets (being referred to as A1H and B1H, respectively) consist of time-periodical data being measured in 1-hour time intervals while the time interval is 5 minutes for the other two data sets (being referred to as A5M and B5M, respectively). Table 2 shows the statistics of the data sets used for testing the prediction models generated by the software.

Table 2. Dataset statistics

Name	Number of Samples	Time Interval	Data Format	Data Type
A1H	1231	1 hour	Bit	Download
A5M	14772	5 minutes	Bit	Download
B1H	1657	1 hour	Bit	Download
B5M	19888	5 minutes	Bit	Download

2.2. Software Development

The development of the software has been carried out using MATLAB programming language, MATLAB R2017a Compiler and Neural Network Toolbox. The design of the user interface has also been done in the same environment. The performance of the prediction models has been evaluated by calculating MAPE, the formula of which is given in (1).

$$MAPE = \frac{100}{N}\sum_{t=1}^{n}\left|\frac{A_t - P_t}{A_t}\right| \quad (1)$$

In (1), A_t is the actual value, P_t is the predicted value and N is the number of samples in the test set.

3. Features of the Software

3.1. Software Components

The main features which have been developed are listed below.

3.1.1. Train-Test ratio adjustment

The amount of rows in the data set used for training and testing the network can be adjusted either with the exact amount or the percentage of the data set rows which are intented to be used as the training data whilst building the prediction models.

3.1.2. The number of neurons in each hidden layer

The amount of neurons in the hidden layers, which are the structures between the input and the output layers in an MLP network, can be adjusted either manually or by the optimization algorithms which are applied the training part of the data set when selected. Two main optimization options which have been provided by the software are 80-20% sequential data division and 10-fold cross validation. The software finds the optimum number of neurons in every hidden layer by creating sub-models and comparing the error rates produced by those sub-models. Afterwards, the optimized hidden layer size is used in order to build the main prediction model with the purpose of obtaining the best possible prediction accuracy.

3.1.3. The number of hidden layers

The option of using 1, 2 or 3 hidden layers in the neural network can be chosen by the user.

3.1.4. The feedback delay values

The feedback delays, being one of commonly-used criteria in time series analysis, can be decided with several options. The feedback delay optimization options are based on making use of dataset autocorrelation. The autocorrelation value of each row in the training data can be calculated by the software and displayed by clicking on the “Data Set AC” button. The main options for optimizing the feedback delays are using the indices which have a higher autocorrelation than a certain threshold, or sorting all autocorrelations in a descending order and using an array of best indices with the highest autocorrelations. These options also create an opportunity to test whether the autocorrelation has an effect on finding the optimum feedback delays.

3.1.5. Network training function

The activation function used for the neural network, which is also decided by the user.

3.1.6. Save-Load option

The user is able to create a brand-new prediction model and save the network settings after the prediction, then use the same settings in order to produce the same prediction results.

3.2. Prediction Steps

The prediction steps followed by the software can be seen below.

- The neural network is created with respect to the initial parameters which are decided manually or automatically by the user’s preference.
    - In the manual mode, the dataset is divided as training and testing sections. The neural network is created with respect the values entered manually on the user interface.
    - In the optimization mode, the training and testing data division is similar to the manual mode. In order to apply optimization, the training data is also divided into sections depending on the selected optimization option. The feedback delays are generated with respect to the training data indices with the best autocorrelations. Subsequent to the data preparation process and feedback delay selection, the software begins creating sub-networks, each of them built by a unique hidden layer size. Finally, the main network is created with the number of neurons that yielded the best performance in the sub-networks.
- The software opens a new window for displaying the comparison between the predicted values and the actual test data both in a table and a line graph, along with the calculated *MAPE*.
- The option to save the network settings is presented to the user in the results screen.

In order to test the software, eight different prediction models have been created using the given data sets with different train/test split ratios. Table 3 shows the testing results for each data set.

Table 3. *MAPE*'s for each data set

Data Set	Train-Test Ratio	*MAPE*
A1H	70-30	9.09
A1H	80-20	6.63
A5M	70-30	5.53
A5M	80-20	3.24
B1H	70-30	6.20
B1H	80-20	4.30
B5M	70-30	8.06
B5M	80-20	6.90

3.3. Screen Shots

Figure 1 through Figure 5 show some screen shots of the developed software.

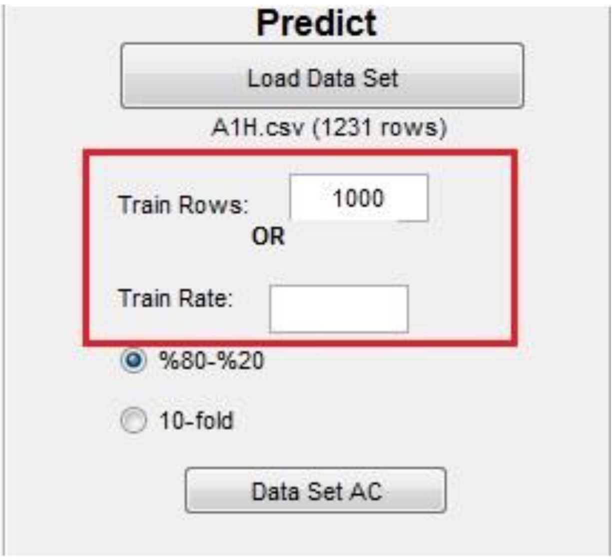

Figure 1. Adjustment of train-test ratio

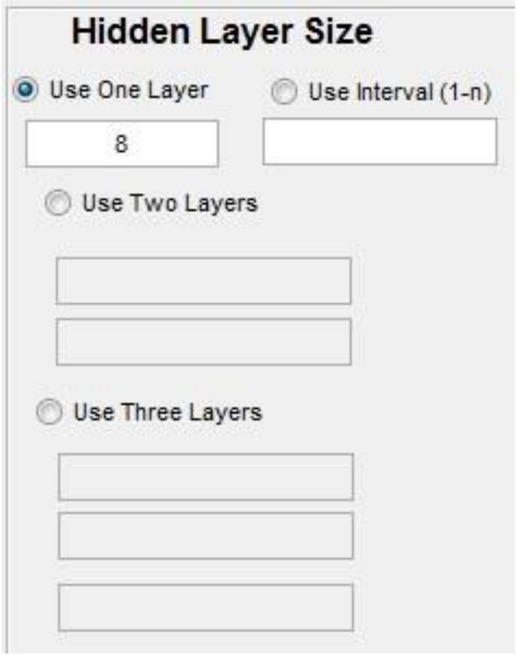

Figure 2. Number of hidden layers and neurons

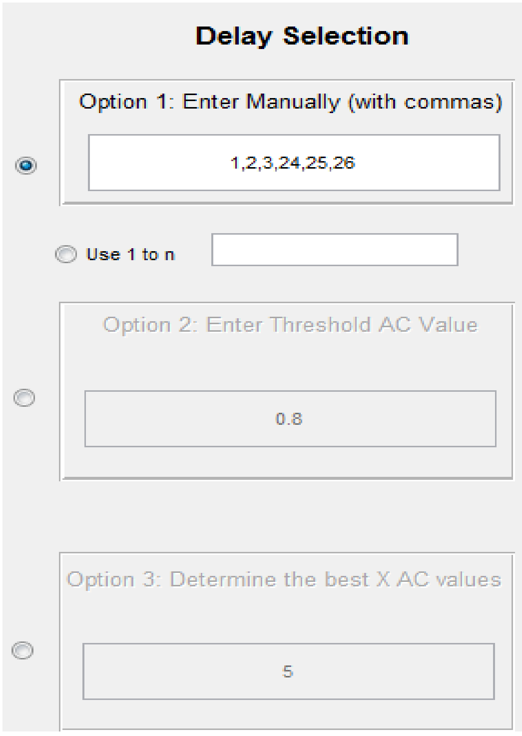

Figure 3. Feedback delay selection

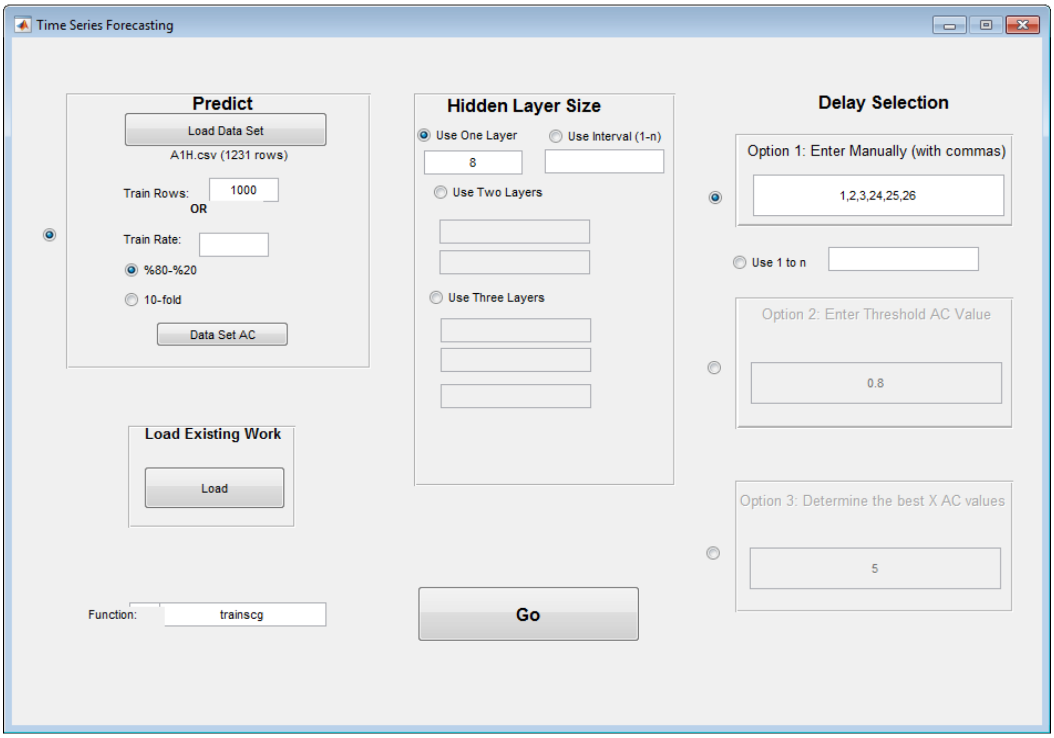

Figure 4. User interface - 1

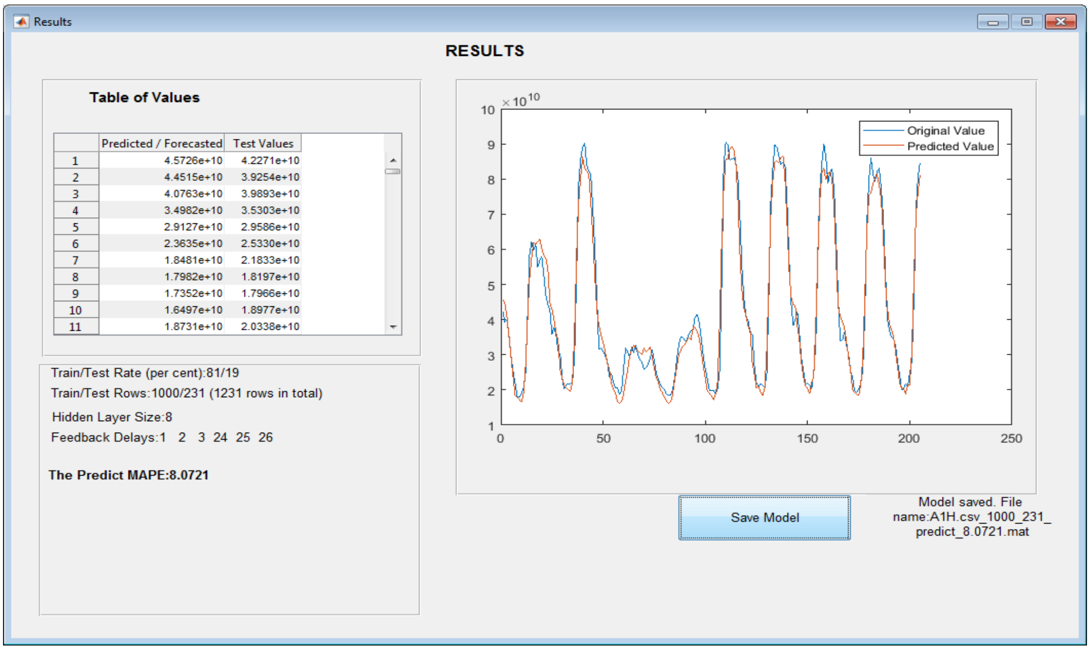

Figure 5. User interface - 2

4. Conclusion

The purpose of this study is the development of a new software based on machine learning and time series analysis, which can predict the Internet data traffic. The project has been developed using MATLAB programming language and the Neural Network Toolbox. The main options of the software include the optimization of hidden layer size, the number of hidden layers and the feedback delays. The performance of each prediction is calculated by using *MAPE*, which has been used as the main error metric. Using four different data sets consisting of time-series data, the models produce prediction *MAPE* values between 3.25 and 9.09. It can be concluded that the software can be used for the network traffic prediction within the acceptable error rates.

Acknowledgment

The authors would like to thank Cukurova University Scientific Research Projects Center for supporting this work under grant no. FYL-2018-10374.

References

Bian, G., Liu, J., ve Lin, W., 2017. Internet Traffic Forecasting using Boosting LSTM Method. DEStech Transactions on Computer Science and Engineering.

Chabaa, S., Zeroual, A. ve Antari, J., 2010. Identification and Prediction of Internet Traffic Using Artificial Neural Networks. Journal of Intelligent Learning Systems and Applications. 2(3), 147-155.

Cortez, P., Rio, M., Rocha, M. ve Sousa, P., 2012. Multi-scale Internet Traffic Forecasting Using Neural Networks and Time Series Methods. Expert Systems. 29(2), 143-155.

Daly, C., 2016. H. 265 Video Traffic Prediction Using Neural Networks. Georgia Southern University Research Symposium. s127.

Jiang, M., Wu, C. M., Zhang, M. ve Hu, D. M., 2009. Research on the Comparison of Time Series Models for Network Traffic Prediction. Acta Electronic Sinica. 37(1), 2353-2358.

Katris, C., ve Daskalaki, S., 2015. Comparing forecasting approaches for Internet traffic. Expert Systems with Applications, 42(21), 8172-8183.

Liu, X., Fang, X., Qin, Z., Ye, C., ve Xie, M., 2011. A Short-Term Forecasting Algorithm for Network Traffic Based on Chaos Theory and SVM. Journal of Network and Systems Management. 19(4), 427-447.

Narejo, S., ve Pasero, E., 2018. An Application of Internet Traffic Prediction with Deep Neural Network. In Multidisciplinary Approaches to Neural Computing. pp. 139-149. Springer, Cham.

Oliveira, T. P., Barbar, J. S. ve Soares, A. S., 2014. Multilayer Perceptron and Stacked Autoencoder for Internet Traffic Prediction. In Network and Parallel Computing. pp. 61-71.

Sahrani, M. N., Zan, M. M. M., Yassin, I. M., Zabidi, A., ve Ali, M. S. A. M., 2017. Artificial Neural Network Non-linear Auto Regressive Moving Average (NARMA) Model for Internet Traffic Prediction. Journal of Telecommunication, Electronic and Computer Engineering. 9(1-3), 145-149.