Noname manuscript No.
(will be inserted by the editor)

On the Predictability of Next Generation Mobile Network Traffic using Artificial Neural Networks

I. Loumiotis · E. Adamopoulou · K. Demestichas · T. Stamatiadi · M. Theologou

Received: date / Accepted: date

Abstract Though the introduction of the new 4th Generation (4G) mobile access technologies promises to satisfy the increasing bandwidth demand of the end-users, it poses in parallel the need for novel resource management approaches at the side of the base station (BS). To this end, schemes that try to predict the forthcoming bandwidth demand using supervised learning methods have been proposed in the literature. However, there are still open issues concerning the training phase of such methods. In the current work, the authors propose a novel scheme that dynamically selects a proper training set for artificial neural network prediction models, based on the statistical characteristics of the collected data. It is demonstrated that an initial statistical processing of the collected data and the subsequent selection of the training set can efficiently improve the performance of the prediction model. Finally, the proposed scheme is validated using network traffic collected by real, fully operational BSs.

Keywords Backhaul resource management · artificial neural networks · relative standard deviation

I. Loumiotis · E. Adamopoulou · K. Demestichas · T. Stamatiadi · M. Theologou
School of Electrical and Computer Engineering, National Technical University of Athens, Greece, GR15780
E-mail: i_loumiotis@cn.ntua.gr

E. Adamopoulou
E-mail: eadam@cn.ntua.gr

K. Demestichas
E-mail: cdemest@cn.ntua.gr

T. Stamatiadi
E-mail: tstamatiadi@cn.ntua.gr

M. Theologou
E-mail: theolog@cs.ntua.gr

1 Introduction

During the last few years, there has been an increasing demand for wireless broadband services [1] which necessitates for an access network capable of satisfying high end-user data rates. The introduction of 4G technologies promises to meet these expectations by offering high capacity, low latency and seamless mobility [2]. Towards this direction, the convergence of the wireless access network with a passive optical network (PON) at the backhaul has been proposed in order to avoid the creation of bottlenecks and to facilitate the quality of service provisioning [3]. However, such an approach raises new issues concerning the resource management of the backhaul network.

Traditionally, the commitment of resources has been flat and its main scope has been to satisfy a worst-case scenario. Unfortunately, this approach, which is essentially based on overdimensioning, cannot be applicable in 4G networks characterized by high peak rates, since this may lead to an unnecessary commitment of valuable and costly resources. Thus, novel approaches are required for the efficient resource management of the backhaul network. These approaches should take advantage of the special characteristics of the new 4G networks and also comply with the new trends in communication networks, such as self-organization [4]. Self-organized networks consist of elements that are able to monitor their environment, interact with each other and react to the changes of the environment. In this light, the base stations (BSs) should be able to monitor their environment, analyze the transferred network traffic, predict their own needs for resources and proactively request their commitment.

In the literature there are already significant research attempts concerning the prediction of network traffic [5]-[8]. In [5], the appropriateness of artificial neural networks (ANNs) for forecasting the bandwidth traffic demand is shown. In [6], the authors study the prediction of Internet backbone traffic using wavelet multiresolution analysis (MRA) and an autoregressive integrated moving average model. In [7], a Pi-Sigma time delay neural network (PSN-TDNN) is used in order to predict scene changes of a real-time variable bit rate video in Asynchronous Transfer Mode (ATM) networks. Though the above works make successful predictions of network traffic, there are no guidelines about the size of the dataset that should be used for the training phase of the prediction model, something which implies that all the available collected data are assumed to be necessary for this process. An attempt towards this latter direction has been made in [8] where the authors study the network traffic collected by the National Science Foundation (NSF) and the National Laboratory for Advanced Network Research (NLANR) Measurement and Network Analysis Group. They use training sets of different size and argue that a small training set can provide more accurate and computationally efficient results for their short-term prediction model. However, their approach is problem-specific and lacks of a unified framework for selecting the proper training set.

In this paper, the authors study the problem of backhaul resource allocation in 4G networks using ANNs. The proposed scheme consists of an Intelligent

Agent located at the side of the BS that is responsible for the management of the BS's backhaul network resources. The Intelligent Agent monitors the current network traffic, stores all the necessary data that are later used in order to predict the forthcoming bandwidth demand, and proactively requests the commitment of the necessary resources based on these predictions. To enable this advanced prediction functionality, an appropriate training set is required. A small training set can result into a prediction model that cannot extend over new unseen data (underfitting). On the other hand, a large training set can result into a more accurate prediction model at the expense of a prohibitively high computational cost. Hence, there is a tradeoff concerning the selection of the training set for the prediction model. As a result, it becomes apparent that the proper selection of the training set is of great importance for the efficient performance of the prediction model. To this end, the authors elaborate on the impact of the statistical characteristics of the training set on the performance of an ANN-based prediction model. Specifically, by exploiting the statistical properties of the collected data, a dynamic training set selection scheme is proposed which can optimize the performance of traffic prediction models that are based on supervised learning techniques, such as ANNs. It is demonstrated that an initial statistical processing of the collected data can efficiently improve the network traffic prediction model. Finally, data collected from real-life fully-operational BSs are used in order to validate the proposed scheme. According to the authors' best knowledge, there exists no prior work in the literature providing the necessary guidelines for the efficient selection of training sets in network traffic prediction models.

The rest of the paper is organized as follows. Section 2 describes the proposed model and the collected measurements. In Section 3, the proposed scheme is implemented and the experimental results are presented. Finally, Section 4 concludes the paper.

2 System Model

2.1 Model Description

Consider an Intelligent Agent located at the side of the BS, as can be seen in Fig.(1), that is responsible for monitoring the BS's environment and storing all the necessary data (timestamped network traffic measurements), which will be subsequently used for the training process of the ANN. Then, based on this prediction model, the Intelligent Agent estimates the forthcoming needs for resources and proactively requests their commitment from the backhaul network. Thus, the significance of the training phase on the performance of the prediction model becomes apparent.

The Intelligent Agent must select an appropriate set out of the collected data in order to train the ANN-based prediction model. If the entire set of data is used for the training process, the resulting computational cost might be prohibitively high. On the other hand, if a small amount of the most recent

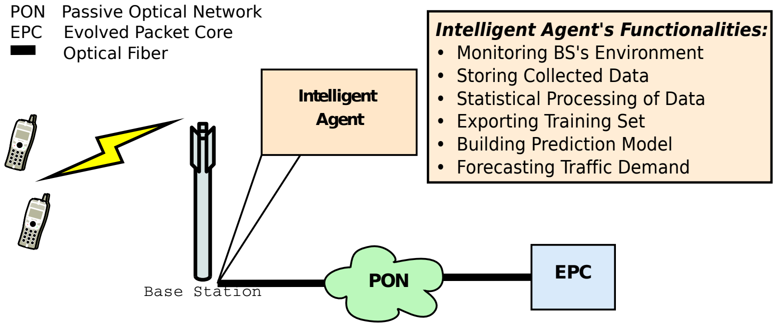
PON Passive Optical Network
EPC Evolved Packet Core
▬ Optical Fiber
Intelligent Agent
Intelligent Agent's Functionalities:
• Monitoring BS's Environment
• Storing Collected Data
• Statistical Processing of Data
• Exporting Training Set
• Building Prediction Model
• Forecasting Traffic Demand
Base Station
PON
EPC

Fig. 1 The Intelligent Agent at the side of the BS.

data is used as a training set, then the prediction model will not be able to extend over new unseen data. Hence, there is a tradeoff concerning the size of the training set.

As a result, the Intelligent Agent should efficiently select the proper training set. For this reason, a statistical processing of the collected data should initially take place, and then, based on these results a proper training set, capable of accurately predicting the forthcoming bandwidth demand, should be exported. The statistical processing of the collected data will allow the Intelligent Agent to identify the basic characteristics of the network traffic demand and dynamically select the appropriate training set that provides accurate prediction results.

In the current work, the authors use the notion of the relative standard deviation (RSD) so as to depict and exploit the statistical properties of the collected data. The RSD is a measure of precision regarding the collected data and is defined as

$$RSD(\%) = \frac{\sigma}{\mu} \times 100 \tag{1}$$

where σ is the standard deviation and μ is the average value of the data set. It is reasonable to expect that, for a network traffic demand pattern, a set of training data that is characterized by a small RSD experiences a more consistent behavior and can result into a more accurate prediction model than a training set with a higher RSD.

Practically, a small RSD for a set of collected data implies that the measurements are averaged around their mean value, while a high RSD refers to collected data with great variations. The former may correspond, for example, to busy times in a crowded cell where the bandwidth is shared among all the subscribers, while the latter may refer to quiet times where a small amount of subscribers exploits a large portion of the available bandwidth [9].

2.2 Measurements

For the validation of the proposed scheme, a set of data collected by two fully operational BSs located in Athens, Greece, are used. The BSs support High

Speed Packet Access (HSPA) connectivity and belong to the largest mobile operator in Greece. The data consist of 5456 hourly averaged measurements of downlink and uplink throughput[1] and concern partly sparse data expanding for a period of over one year starting from February 2012.

Intuitively, it is expected that the traffic pattern would experience certain periodicities, as the bandwidth demand is increased during the rush hours in the morning and in the evening, while during the night hours it is significantly decreased. In order to verify the existence of such periodicities, the fast Fourier transformation (FFT) has been employed. Because of the sparsely collected data, only a consecutive portion of them has been used for the FFT. Any missing values of the chosen sample for the Fourier transformation were interpolated by averaging either the preceding and the following measurement in case of only one missing value, or the measurements one week before and one week after, for the case of a series of missing values. These values were used only for the Fourier transformation analysis in order to detect periodicities, and in the subsequent phase of the ANN training, these interpolated values are omitted.

The collected data and the FFT results are depicted in Fig.(2) and in Fig.(3) for the downlink and the uplink case of BS1, respectively. Similar results are also derived for BS2. From Fig.(2) and Fig.(3), it becomes apparent that there exists a dominant period of 24-hours in the set of collected data. As a result, the demand pattern of the resources features a daily cycle, which implies that the behavior of the traffic demand from one day to the following day can be predicted sufficiently. Apart from the dominant period of 24-hours, another noticeable period is that of 12-hours and for the downlink traffic there is also a smaller period of 8-hours.

3 Experimental Results

3.1 Artificial Neural Network Parameters

In the present section, the experimental results regarding the performance of three different types of ANNs in the network traffic prediction problem and their dependence on the statistical properties of the collected data are presented. For comparison purposes, three different kinds of ANNs have been investigated, namely a multilayer perceptron (MLP) neural network [10], a general regression neural network (GRNN) [11] and a group method for data handling (GMDH) neural network [12]. Concerning the configuration of the ANNs, for the MLP the optimal number of hidden layers was computed and used, whereas for the GRNN the optimal value for the smoothing parame-

[1] The collected data correspond to the aggregated throughput derived by a mixture of services that varies over time, as requested by the subscribers.

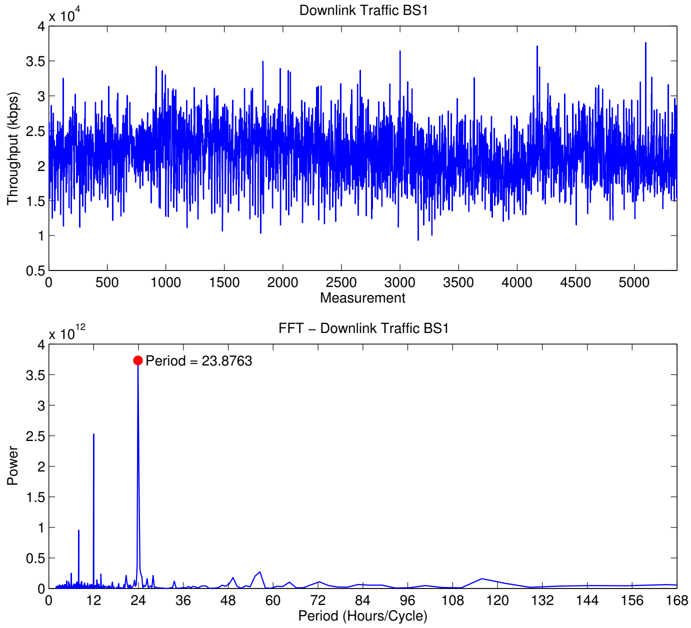

Fig. 2 Collected data and the fast Fourier transformation for the downlink traffic of BS1.

ter was detected and selected[2]. For the case of GMDH networks, a 64-order polynomial was used.

Due to the periodicities experienced by the network traffic demand pattern, the input data of the ANN-based prediction model have to be time associated. Thus, the set of input variables (predictors), denoted by $\mathbf{x}$, for the training set consists mainly of the day and time that the data were collected, while the output, denoted by y, is the measured hourly-based average bandwidth demand. Furthermore, special days in the calendar, such as holidays, that may influence the traffic demand are also taken into consideration. As a result, the input variable $\mathbf{x}$ and the output variable y can be expressed as

$$\mathbf{x} = (D, M, H, DT, Y, SE)$$

$$y = BW$$

where D is the name of the day (e.g. Monday), M is the name of the month, H is the time of the hour of the day, DT is the sequence number of the day of the month (e.g. 25), Y is the number of the year, SE is a binary variable used to designate whether a special occasion exists in the corresponding day that could possibly influence the demand, and BW is the hourly-averaged bandwidth demand.

[2] In order to calculate the optimal parameters for the MLP case, multiple networks were built and evaluated using 4-fold cross validation [13], while for the case of GRNN the conjugate gradient algorithm was used.

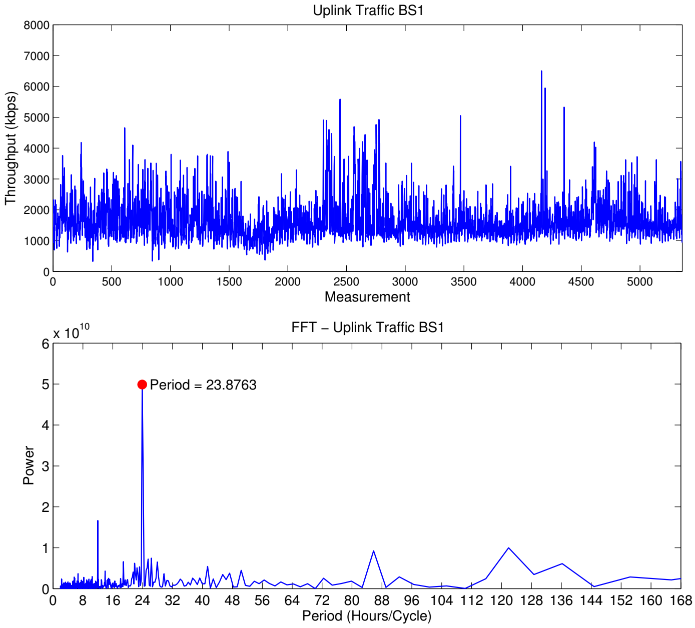

Fig. 3 Collected data and the fast Fourier transformation for the uplink traffic of BS1.

Finally, for the validation process of the prediction model, the 10-fold cross validation [13] technique is employed, whilst the mean absolute percentage error (MAPE) of the validation process is used as a figure of merit in order to compare the performance of the different types of ANNs.

3.2 Statistical Analysis of the Training Set

In order to investigate the impact of the statistical characteristics of the training set on the performance of the prediction model, the authors separated the collected data into 5 different and equal classes. Each class contains 1091 collected measurements, and the RSD is derived for each one in order to investigate its impact on the performance of the prediction model. These classes were used to build the prediction model and the results of the validation process are depicted in Fig.(4) and in Fig.(5) for BS1 and BS2, respectively, for both the downlink and the uplink traffic.

From these figures, it becomes clear that there is an evident dependence of the MAPE of the prediction model on the statistical properties of the training set, i.e. the RSD. Specifically, it can be seen that, when the collected data are more dispersed around the mean value, i.e. a high RSD is experienced, the performance of the prediction model degrades, which corresponds to a high MAPE. On the contrary, when the data are more gathered around the mean value, i.e. RSD is low, the MAPE of the ANN-based prediction model is small.

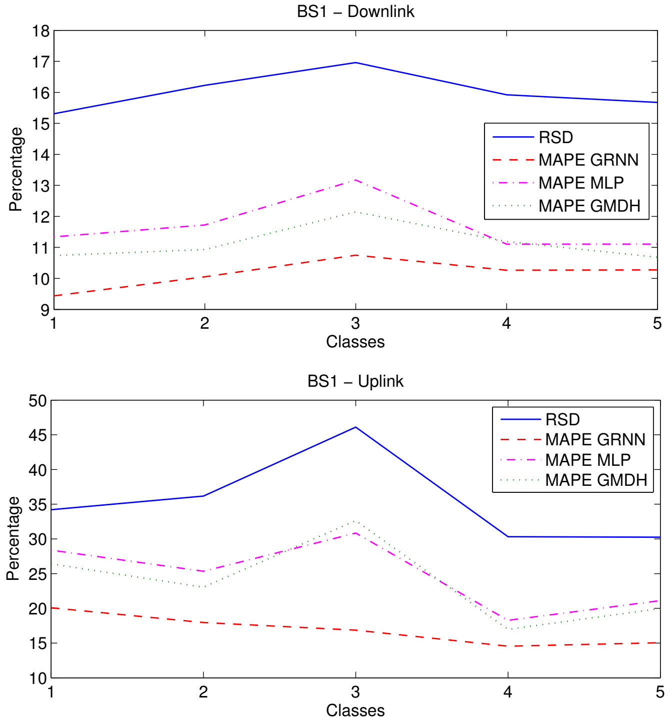

Fig. 4 Mean absolute percentage error (MAPE) for three types of ANNs and the corresponding relative standard deviation (%RSD) for the downlink and the uplink traffic of BS1, for different training sets.

Furthermore, it becomes apparent that all of the three types of ANNs follow the trend of the RSD. Moreover, it is evident that the GRNN model is more robust to the changes of the statistical properties of the training set, and yields more accurate results than the other two ANNs. Hence, these properties, in conjunction with the general characteristics of the GRNN [11], render it more suitable for the implementation of the prediction model.

3.3 Dynamic Training Set Selection

In the previous subsection, it was demonstrated that there is a direct relationship of the RSD with the performance of the ANN-based prediction model. Thus, the Intelligent Agent should initially process the collected measurements in order to dynamically export, based on the RSD value, a proper dataset for the training phase of the prediction model. In order to elaborate more on the implementation of the dynamic selection phase of the training set, the authors separated the collected data into 8 classes. Each class i contains the $682 \cdot i$ most recent data, and for each training set the RSD is calculated. Hence, class 1

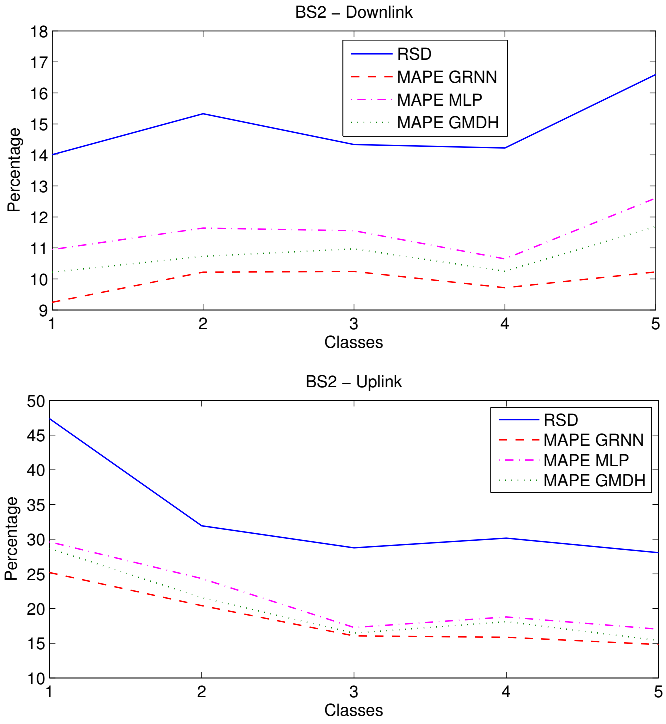

Fig. 5 Mean absolute percentage error (MAPE) for three types of ANNs and the corresponding relative standard deviation (%RSD) for the downlink and the uplink traffic of BS2, for different training sets.

contains the 682 most recent data, whereas class 8 contains the total amount of the collected data.

The results for BS1 and BS2 are depicted in Fig.(6) and in Fig.(7), respectively, for both the downlink and the uplink traffic. It can easily be observed that there is a direct connection of the RSD with the performance of the prediction model as it was proven in the previous subsection.

From Fig.(6) it is clear that, in general, a smaller training set results in a lower RSD and a corresponding lower MAPE. Thus, a small training set suffices to enable the efficient functionality of the Intelligent Agent of BS1. The same case also holds for the uplink traffic of BS2 as can be seen in Fig.(7). However, for the downlink case of BS2, a large training set experiences a smaller RSD and a corresponding lower MAPE. Hence, a larger training set is more appropriate for this case. As a result, it is not the size of the training set but rather the value of the RSD that provides a good approximation (indication) for the expected performance of the prediction model. Thus, the value of the RSD can be used as a guideline in order to dynamically select the proper dataset for the training phase.

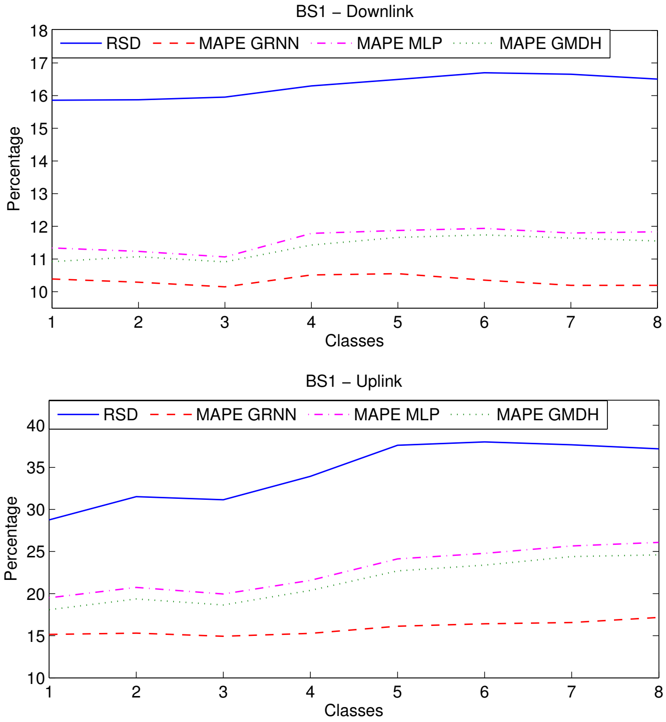

Fig. 6 Mean absolute percentage error (MAPE) for three types of ANNs and the corresponding relative standard deviation (RSD) for the downlink and the uplink traffic of BS1, with respect to the size of the training set.

4 Conclusion

In the current work, the impact of the statistical properties of network traffic on the performance of an ANN-based prediction model was investigated. To this end, the notion of RSD has been employed to identify the statistical characteristics of the collected data used as a training set for the ANN. The importance of an initial statistical processing of the collected data was identified and a novel scheme that dynamically selects a proper training set based on the special characteristics of the dataset was proposed. For the validation of the proposed scheme, real data coming from two fully operational BSs was used. Finally, it was concluded that the RSD metric provides an excellent guideline for dynamically selecting the optimal training dataset, and it was demonstrated that a training set consisting of collected data with significant fluctuations around their mean value can degrade the performance of the prediction model. On the contrary, training sets with measurements averaged around their mean value can provide more accurate results.

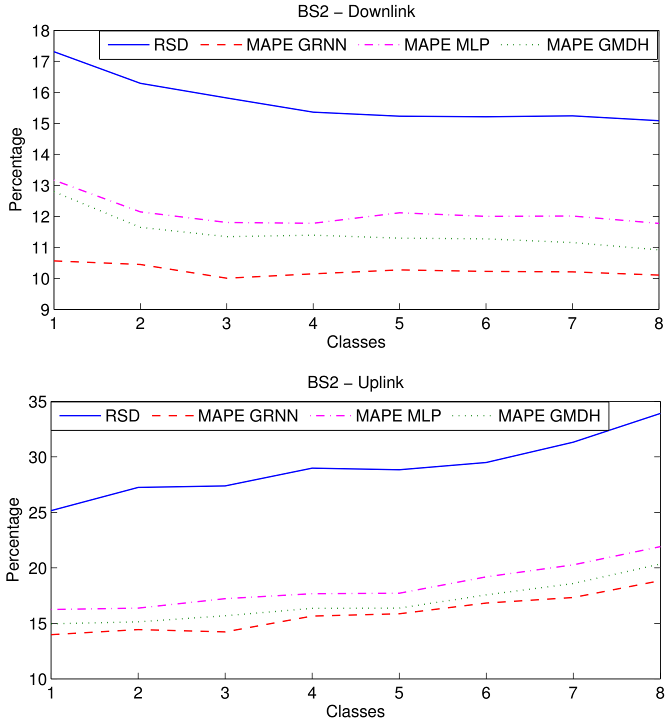

Fig. 7 Mean absolute percentage error (MAPE) for three types of ANNs and the corresponding relative standard deviation (RSD) for the downlink and the uplink traffic of BS2, with respect to the size of the training set.

References

1. *Cisco Visual Networking Index: Forecast and Methodology 2011-2016*, White paper, 2012.
2. Rinne, M., Tirkkonen, O. (2010). LTE, the radio technology path towards 4G, *Computer Communications*, 33(16), 1894-1906.
3. Orphanoudakis, T., Kosmatos, E., Angelopoulos, J., Stavdas, A. (2013). Exploiting PONs for mobile backhaul, *IEEE Communications Magazine*, 51(2), S27-S34.
4. Prehofer, C., Bettstetter, C. (2005). Self-Organization in Communication Networks: Principles and Design Paradigms, *IEEE Communications Magazine*, 43(7), 78-85.
5. Loumiotis, I., Stamatiadi, T., Adamopoulou, E., Demestichas, K., Sykas, E. (2013). Dynamic Backhaul Resource Allocation in Wireless Networks using Artificial Neural Networks,*IET Electronic Letters*, 49(8), 539-541.
6. Papagiannaki, K., Taft, N., Zhang, Z.L., Diot, C. (2005). Long-Term Forecasting of Internet Backbone Traffic, *IEEE Transactions on Neural Networks*, 16(5), 1110-1124.
7. Chong, S., Li, S.Q., Ghosh, J. (1995). Predictive Dynamic Bandwidth Allocation for Efficient Transport of Real-Time VBR Video over ATM, *IEEE Journal in Selected Areas in Communications*, 13(1), 12-23.
8. Zhani, MF, Elbiaze, H. (2009). Analysis and Prediction of Real Network Traffic, *Journal of Networks*, 4(9), 855-865.
9. NGMN Alliance (2011). *Guidelines for LTE Backhaul Traffic Estimation*, white paper.
10. Haykin, S. (1999). *Neural Networks, A Comprehensive Foundation*, 2nd edition, Prentice Hall.

11. Specht D.F. (1991). A General Regression Neural Network, *IEEE Transactions on Neural Networks*. 2(6), 568-576.
12. Farlow, S.J. (1981). The GMDH Algorithm of Ivakhnenko, *The American Statistician*, 35(40), 210-215.
13. Liu, L., Özsu M. T. (2009). *Encyclopedia of Database Systems*, Springer.