GMDH Modelling for Mobile User Throughput Forecasting

Isah A. Lawal
Faculty of Applied Computing and Technology (FACT) - Noroff University College
Elvegata 2A, 4608 Kristiansand - Norway
Isah.Lawal@noroff.no

ABSTRACT

This paper demonstrates the use of GMDH algorithm as an alternative approach for forecasting mobile user throughput in a cellular data network. We measure the hourly throughput per user for three weeks and use it to train a GMDH-based model for predicting the user throughput for the fourth week. We adopt a modelling strategy that employs a single next-day forecaster iteratively to estimate an entire week throughput. Our experimental results show that the GMDH-based forecaster performed very well with a mean percentage error as low as 1.87%. We also compare the performance of the GMDH-based forecaster with that developed using state-of-the-art LSTM method and show that it can achieve a comparable performance against the LSTM method. Moreover, the GMDH algorithmâs ability to select only effective input variables during model training reduces the dimensionality of the training data by 43% and allows the development of simpler and more interpretable throughput forecaster.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning approaches**;
• **Applied computing** → **Telecommunications**;

KEYWORDS

Throughput prediction, GMDH algorithm, LSTM method

ACM Reference Format:
Isah A. Lawal. 2020. GMDH Modelling for Mobile User Throughput Forecasting. In *The 35th ACM/SIGAPP Symposium on Applied Computing (SAC '20), March 30-April 3, 2020, Brno, Czech Republic.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3341105.3373843

1 INTRODUCTION

A major challenge for mobile network operators is how to plan and optimise network resources to guarantee the quality of service to their network subscribers during planned outages or peak periods [6, 8]. he throughput experienced by network users is an important indicator of the performance and quality of the network connection [3]. Thus, by predicting the expected throughput in advance, network operators can schedule and distribute network resources to cater to the need of their subscribers more efficiently [17]. Short-term (i.e. one hour to one week) throughput forecasting is crucial for network analysis to ensure performance benchmarks are being met and also for scheduling functions such as system update and base transceiver station maintenance [1].

Network throughput forecasting can be formulated as a time series prediction problem using historical network data and statistical analysis such as Autoregressive Integrated Moving Average (ARIMA) can be applied in solving the problem [11]. However, this approach relies heavily on user experience and may not appeal to mobile data network operators. The ARIMA approach, however, relies heavily on user experience and does not appeal to mobile network operators. Machine Learning (ML) methods such as Long Short-Term Memory (LSTM) [15] and Group Method of Data Handling (GMDH)[1] [4], on the other hand, allow the development of time series forecasting model without the prior knowledge of the characteristic of the time series data [2, 9]. In other words, a user does not need to explicitly specify the model relationship for any given time series data during model synthesis. Therefore, the forecasting model is free from bias and prior assumptions on the data distribution. LSTM has been used for prediction in many applications [5, 10, 16]. In spite of the extensive use of the method in forecasting problems, it still suffers from some limitations. These limitations include difficulties in determining critical network and training parameters, such as the size of the network layers, type of transfer function and learning rate. Moreover, with the LSTM method, it is difficult to understand the relative significance and contribution of the different input variables during the model synthesis. GMDH algorithm has emerged recently as a powerful tool for solving forecasting problems [13, 14]. The method offers the advantages of an automatic configuration of the model structure and efficient selection of only the most relevant input variables during model development, which provides a better insight into the model [13]. Inspired by the results achieved with the GMDH method in the past [7, 13, 14], we present the GMDH algorithm as an alternative method that alleviates many of the limitations of LSTM, for short-term network throughput forecasting in a cellular data network.

2 METHODOLOGY

The GMDH algorithm is a formalised paradigm for iterative and multi-layer polynomial regression that can be used to synthesise model for prediction [13]. The modelling process occurs in an evolutionary manner, where simple input-output polynomial regression relationships initialised in one iteration, is used to derive more accurate model representations in the next iteration [4]. After each iteration, the algorithm selects and retains only polynomial relationships and the input variable combinations that have good prediction

[1]Group Method of Data Handling [online]. http://www.gmdh.net/ [Last accessed 17.07.2019]

SAC '20, March 30-April 3, 2020, Brno, Czech Republic

ACM ISBN 978-1-4503-6866-7/20/03…$15.00
https://doi.org/10.1145/3341105.3373843

power for the next iteration [4]. The selective discarding of under-performing input variables has the advantage of limiting the model complexity by preventing the exponential growth of the polynomial expressions. The model synthesis is terminated automatically when the polynomial regression begins to have poorer prediction performance than those generated in the previous iterations.

2.1 GMDH-based one-step ahead model

To develop a GMDH-based one-step ahead model, we require a time-series data with N rows of data samples and $m+1$ columns for m independent variables $[F_0, F_1, F_2, \ldots, F_m]$ as inputs and one dependent variable $y = F_{m+1}$, as the targeted output. First, the GMDH algorithm split the data into training and selection sets. Then the modelling process begins by initialising the first iteration, whereby each pair of the training input variables (F_i, F_j ; i, j = 0,1,2,…,m) together with the corresponding output y are taking as the predictors to derive the following regression polynomial equation [4]:

$$\hat{y}_i = b_0 + b_1 F_i + b_2 F_j + b_3 F_i F_j + b_4 F_i^2 + b_5 F_j^2 \quad (1)$$

where $b_0, b_1, b_2, \ldots, b_5$, are the coefficient of the input variables. Each of the $m(m-1)/2$ regression equations produced is evaluated using the sample for the pair of input variables used to derive it, to generate new estimation $\hat{y}_i$, $\forall i$, $i \in [1, 2, \ldots, m(m-1)/2]$ known as the partial descriptor for the targeted output y. The Root Mean Square Error (RMSE) for each of the partial descriptors $\hat{y}_i$, over the selection set is computed to select only the $\hat{y}_i$ having good predicting power. The RMSE value r_i $\forall i$ is computed as [4]

$$r_i = \left[\frac{\sum_{k=1}^{n} (y_k - \hat{y}_{i_k})^2}{\sum_{k=1}^{n} y_k^2} \right]^{0.5} \quad (2)$$

where n is the size of the selection set. The polynomial equations and the corresponding $\hat{y}_i$ with r_i value below a preset error threshold are selected and kept while the under-performing ones are discarded. The minimum r_i value i.e. r_i^{min} recorded is also saved. The selected partial descriptors $\hat{y}_i$ are then used for repeating the model estimation process in the next iteration to derive the next set of higher-layer partial descriptors. At the end of next iteration, the r_i^{min} obtained is compared with the one saved in the previous iteration and the modelling process is terminated when the value of r_i^{min} begin to increase more than its previous value or when a desired model complexity is reached. The final GMDH-based one-step ahead forecasting model can be considered computationally, as a layered network of partial descriptors, each layer representing the results of an iteration. Algorithm 1, enumerates the steps involved in the GMDH modelling process, while Figure 1 shows an example of a GMDH-based one-step ahead forecaster developed in this work using the dataset described in Section 3.1. As illustrated in the figure, the model contains four-layered (i.e. layer 0-3) network and each layer consist of at least one partial descriptor (neuron) with linear or quadratic input combinations. Among the seven input variables ($F_0 - F_6$) provided per sample during modelling, the algorithm automatically selected only four most relevant ones (i.e. F_0, F_1, F_3, F_4), thus reducing the dimensionality of the input data by 43%.

Algorithm 1 GMDH modelling for one-step ahead forecast.

1: **Terms:** $\hat{y}$:-partial descriptor, r:-RMSE value, r^{min}:- lowest r
2: **Input:** Time series dataset
3: **Output:** Forecasting model
4: **Begin:**
  i. Split input data into training and selection sets.
  ii. Generate different pairs of the training input variables.
  iii. Set the error threshold and maximum number of layers.
5: **Step 1:**
  for *each pair of input variables* **do**
    Derive the polynomial equation in Eq.1.
    Compute $\hat{y}$ from Eq.1 with data for each pair of the input variables.
    Compute r for $\hat{y}$ over the selection set using Eq.2.
  end
  iv. Retain all polynomial equations for $\hat{y}$ with $r \leq$ threshold
  v. Represent selected polynomial equations as new layer of the model
  vi. Create new pairs of input combinations using all the $\hat{y}$
  vii. Record the value for r^{min}.
  while r^{min} *decreases & number of layers not reached* **do**
    Go to **Step 1.**
  end
  viii. Return model; a multi-layered polynomial network
6: **End**

2.2 Iterative forecast using GMDH-based model

We aim to forecast the network throughput for seven days ahead. However, the GMDH-based model by default can only be used for one-step-ahead or one-day forecasting and cannot be directly applied for seven-days ahead forecasting. Two possible strategies can be used to overcome the aforementioned limitation. This includes using the single GMDH-based one-step ahead model iteratively seven times, or the direct use of seven dedicated GMDH-based one-step ahead models, each independently trained to forecast the network throughput for each day of the week. For this paper, we adopt the former approach whereby the value of the network throughput forecasted in one iteration is provided as input to the same model in the next iteration until the network throughput for the entire seven days of the evaluation week forecasted. This approach is similar to the rolling-origin strategy used for the out-of-sample test to evaluate the accuracy of forecasting model [12]. There is a concern that the iterative method could lead to the accumulation of the forecasting error. However, in a practical setting, this approach is more computationally cost-effective, because only one GMDH-based model is developed and maintained over time, and not seven independent and dedicated models.

3 EXPERIMENTS AND DISCUSSION

3.1 Dataset

We used High-Speed Downlink Packet Access (HSDPA) dataset that was recorded from sixty different cell sites of a UMTS-based cellular network operator over a period of four weeks, in a densely populated urban area. The dataset consists of 683 data points representing the aggregated hourly average throughput per user in

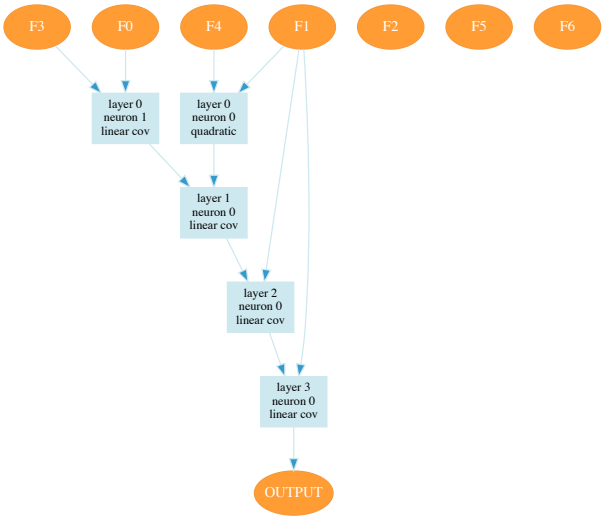


Figure 1: A four-layered GMDH-based next-day throughput forecaster. Only 4 (F0,F1,F3 and F4) input variables selected from the 7 (F0-F6) provided during training.

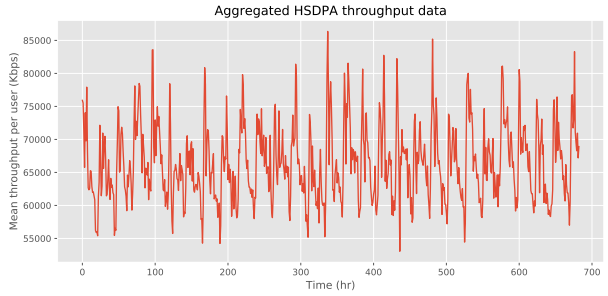


Figure 2: The plot of the HSDPA dataset showing the aggregated hourly average throughput per user in Kbps.

Kilobit per second (Kbps), for all the sixty cell sites combined. Figure 2 shows a plot of all the data points in the dataset. As illustrated in the figure, the dataset is characterised by a 24-hour seasonality and also exhibits high variability. The dataset is available online for public use[2].

3.2 Experimental Setup

We conducted two experiments as follows: In the first experiment, we aim to forecast the one-hour ahead average throughput per user. We use the first 3-week historical data for model development (training) and that for the fourth week for model evaluation. A sample in the training data have the following values $[F_0 = x_{i-6}, \ldots, F_4 = x_{i-2}, F_5 = x_{i-1}, F_6 = x_i]$ as input variables and x_{i+1} as targeted output. x_i is the measured average throughput per user for the i^{th} hour, where $i = 7$ is the time lag window. We set the maximum number of layers for the GMDH algorithm to infinite and the threshold for stopping the training to 10^{-3}. In the second experiment, we aim to forecast the average throughput per user for a whole one week ahead. Thus, we convert the HSDPA dataset to daily data by integrating the hourly throughput measurements over a period of 24 hours. Each datapoint in the new dataset now represents the average throughput per user per day. Like in the previous experiment, we use the daily average throughput data for the first 3-week for training and that for the fourth week for model evaluation. Figure 1 shows the model structure of a 4-layered GMDH-based next-day average throughput forecaster developed and used iteratively seven times (as discussed in Section 2.2) to predict the user throughput for the evaluation week. Each layer of the model consists of at least one neuron (i.e. partial descriptor with linear or quadratic input variables combination). The algorithm automatically selected only four input variables (F_0, F_1, F_3, F_4) from the seven input variables $F_i, i = 0, 1, \ldots, 6$ per sample provided to during the modelling. This is because the algorithm carves out under-performing inputs during the model synthesis. This automatic selection of relevant inputs accounts for 43% reduction in the number of input variables required to forecast the next-day average throughput. We evaluate the performance of the GMDH-based forecaster by comparing its predictions to the targeted values and computing the Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE) and RMSE of the predictions. The MAPE and MAE are computed as [12]

$$MAPE = \frac{\sum_i^K \frac{|y_i - \dot{y}_i|}{y_i}}{K} * 100$$

$$MAE = \frac{\sum_i^K |y_i - \dot{y}_i|}{K}$$

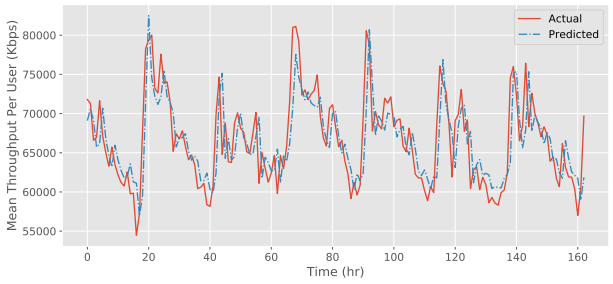
where y_i is the targeted value, $\dot{y}_i$ is the predicted value and K is the size of the evaluation set. We also compare the performance of the GMDH-based forecaster with state-of-the-art LSTM method. We implement an LSTM-based one-step ahead forecaster consisting of a single hidden layer of LSTM units [10]. We look for the optimum number of nodes for the hidden layer in the range [10, 100] by using 5 points equally spaced in a linear scale. We used Rectified Linear Unit (ReLU) as activation function, Adam optimiser and ran the model fitting process for 1500 epochs.

3.3 Discussion of Results

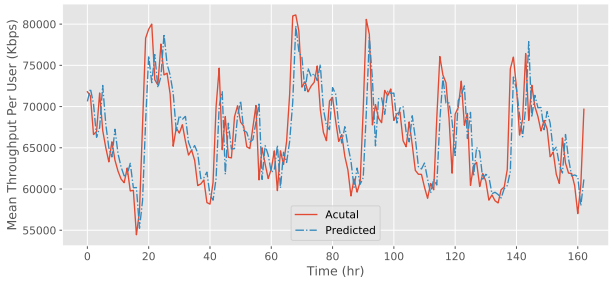
The results of the first experiment are shown in Table 1. The LSTM-based forecaster predicts the average throughput with a MAPE, MAE, and RMSE of 4.45%, 2993.57 Kbps and 3911.64 Kbps, respectively. Whereas the proposed GMDH-based forecaster performance better with MAPE, MAE, and RMSE of 4.06%, 2735.87 Kbps and 3575.25 Kbps, respectively. Figure 3 shows the plots of the actual and predicted hourly average throughput per user for both GMDH and LSTM based models.

The results of the second experiment are summarised in Table 2. The targeted average throughput per user for the evaluation week is 11,210,147.10 Kbps. The GMDH-based forecaster despite using only 57% of the input variables, predicted the throughput as 11,202,468.53 Kbps with MAPE of 1.87%, while the LSTM-based forecaster predicted a throughput of 11,304,785.50 Kbps with MAPE

[2]HSDPA Dataset: https://www.dropbox.com/s/fkt3gdlxogedf84/Dataset.xlsx?dl=0



(a) GMDH-based Model



(b) LSTM-based Model

Figure 3: The actual and predicted hourly average throughput per user. (a) GMDH-based model predicts with MAPE, RMSE, and MAE error of 4.06%, 3575.25 Kbps, 2735.87 Kbps, respectively. (b) LSTM-based model predicts the throughput with slightly worse MAPE, RMSE, and MAE error of 4.45%, 3911.64 Kbps, 2993.57 Kbps, respectively.

Table 1: Summary of the performances of the forecasting models in predicting the hourly average throughput.

Evaluation Criteria	GMDH-based Model	LSTM-based Model
RMSE (Kbps)	3,575.25	3,911.64
MAE (Kbps)	2,735.87	2,993.57
MAPE (%)	4.06	4.45

Table 2: Summary of the performance of the forecasting models in predicting the average throughput for the entire evaluation week. The value in bracket is the actual aggregated throughput for the evalution week.

Evaluation Criteria	GMDH-based Model	LSTM-based Model
Mean throughput per user (Kbps)	11,202,468.53 (11,210,147.10)	11,304,785.50 (11,210,147.10)
Input variables used (%)	57	100
RMSE (Kbps)	36,151.78	37,288.68
MAE (Kbps)	31,206.15	32,586.18
MAPE (%)	1.87	2.04

of 2.04%. Figure 4 shows the relative error committed by both methods on a daily basis for the entire week. As illustrated in the figure, the performance of the GMDH-based forecaster is comparable to

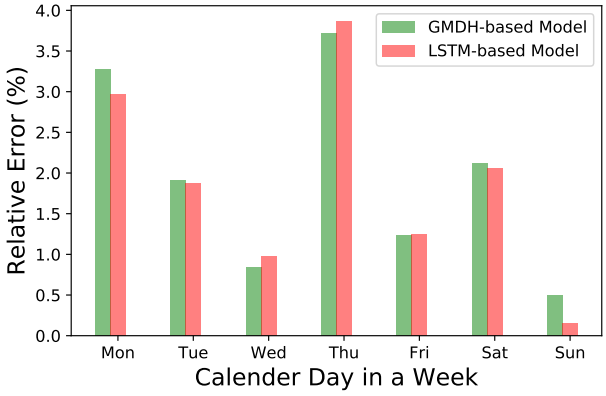


Figure 4: Relative error incurred by two the forecasters in predicting the daily throughput per user in a week.

the LSTM-based model, in addition to achieving a reduction of 43% in the input data dimensionality.

4 CONCLUSION

This paper demonstrates the use of the GMDH-based algorithm as an alternative approach for modelling and forecasting the average network throughput per user in UMTS-based cellular data network. We train a GMDH-based model on hourly mobile user throughput time-series data for three weeks and used it to forecast the throughput for the fourth week with MAPE as low as 1.87%. This result is better than the 2.04% MAPE obtained using state-of-the-art LSTM-based recurrent neural networks. Moreover, the GMDH algorithm's ability to automatically selects effective input variables accounts for a 43% reduction in the dimensionality of the training data required for the model and also gives better insight into the modelled throughput forecaster.

ACKNOWLEDGMENTS

This work was completed in part with the support and computing resources provided by Noroff School of Technology and Digital Media, Kristiansand, Norway.

REFERENCES

[1] S. A. Abdulkarim and I. A. Lawal. 2017. A cooperative neural network approach for enhancing data traffic prediction. *Turkish Journal of Electrical Engineering and Computer Sciences* 25 (2017), 4746–4756.
[2] G. Bontempi, S. Ben Taieb, and Y. Le Borgne. 2013. *Machine Learning Strategies for Time Series Forecasting.* Springer Berlin Heidelberg, 62–77.
[3] N. Bui, F. Michelinakis, and J. Widmer. 2014. A Model for Throughput Prediction for Mobile Users. In *Proc. of the 20th European Wireless Conference.* 1–6.
[4] S.J. Farlow. 1984. The GMDH algorithm. In *Self-Organising Methods in Modelling: GMDH Type Algorithms*, S.J Farlow (Ed.). Marcel-Dekker, New York.
[5] K. Kim, D. Kim, J. Noh, and M. Kim. 2018. Stable Forecasting of Environmental Time Series via Long Short Term Memory Recurrent Neural Network. *IEEE Access* 6 (2018), 75216–75228.
[6] I. A. Lawal, S. A. Abdulkarim, M. K. Hassan, and J. M. Sadiq. 2016. Improving HSDPA Traffic Forecasting Using Ensemble of Neural Networks. In *Proc. of the 15th IEEE International Conference on Machine Learning and Applications.* Los Alamitos, CA, USA, 308–313.
[7] RYM. Li, S. Fong, and KW. Sang Chong. 2017. Forecasting the REITs and stock indices: Group Method of Data Handling Neural Network approach. *Pacific Rim Property Research Journal* 23, 2 (2017), 123–160.

[8] Y. Liu and J. Y. B. Lee. 2015. An Empirical Study of Throughput Prediction in Mobile Data Networks. In *Proc. of the Conference on Global Communications*. 1–6.
[9] S. Makridakis, E. Spiliotis, and V. Assimakopoulos. 2018. Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLoS ONE* 13, 3 (2018).
[10] S. Muzaffar and A. Afshari. 2019. Short-Term Load Forecasts Using LSTM Networks. *Energy Procedia* 158 (2019), 2922 – 2927.
[11] FK. Oduro-Gyimah and KO. Boateng. 2019. Using autoregressive integrated moving average models in the analysis and forecasting of mobile network traffic data. *African Journal of Engineering Research* 7, 1 (2019), 1–9.
[12] LJ. Tashman. 2000. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting* 16, 4 (2000), 437 – 450.
[13] TM. Tsai and PH. Yen. 2017. GMDH algorithms applied to turbidity forecasting. *Applied Water Science* 7, 3 (2017), 1151–1160.
[14] L. Xie, J. Xiao, Y. Hu, H. Zhao, and Y. Xiao. 2017. China's energy consumption forecasting by GMDH based auto-regressive model. *Journal of Systems Science and Complexity* 30, 6 (2017), 1332–1349.
[15] B. Yang, K. Yin, S. Lacasse, and Z. Liu. 2019. Time series analysis and long short-term memory neural network to predict landslide displacement. *Landslides* 16, 4 (2019), 677–694.
[16] J. Yang and J. Kim. 2018. An accident diagnosis algorithm using long short-term memory. *Nuclear Engineering and Technology* 50, 4 (2018), 582 – 588.
[17] C. Zhang, P. Patras, and H. Haddadi. 2018. Deep learning in mobile and wireless networking: A survey. *arXiv preprint arXiv:1803.04311* (2018).