A Hybrid Neuro-Genetic Approach to Short-Term Traffic Volume Prediction

Shahriar Afandizadeh[1,*], Jalil Kianfar [2]

Received: January 2003, Revised: July 2008, Accepted: January 2009

Abstract: *This paper presents a hybrid approach to developing a short-term traffic flow prediction model. In this approach a primary model is synthesized based on Neural Networks and then the model structure is optimized through Genetic Algorithm. The proposed approach is applied to a rural highway, Ghazvin-Rasht Road in Iran. The obtained results are acceptable and indicate that the proposed approach can improve model accuracy while reducing model structure complexity. Minimum achieved prediction r2 is 0.73 and number of connection links at least reduced 20% as a result of optimization.*

Keywords*: Model; Traffic Prediction; Neural Network; Genetic Algorithm*

1. Introduction

Recent advances in the field of traffic data collection have provided transportation engineers with new features for traffic network management. These emerging technologies are utilized in TMC ; these centers can be designed for urban or rural highways. Traveler information and traffic management are primary functional areas of a TMC. In this paper, we focus on developing a model which covers both functional area of TMC.

Traffic detectors collect traffic parameters and transmit them through communication backbone to TMC. This data is utilized in order to improve traffic management strategies. Short-term Traffic Prediction Model is one of the TMC features, which puts traffic network data into practical use. The model receives real-time traffic condition on each link as input data and predicts link traffic condition for a few minutes ahead. Traffic condition can be in the form of volume, speed and travel time. TMC can take the advantage of this information for proactive traffic control. Proactive control targets near-term anticipated conditions, and the traffic network would operate (in theory, at least) under control strategies that are more relevant to the prevailing conditions. On the other hand, reactive control, as the name implies, reacts to already-observed conditions of the traffic stream.

* *Corresponding Author: Email: zargari@iust.ac.ir*

[1] *Associate Professor, Iran University of Science and Technology*

[2] *M.Sc. Transportation Engineering, Tehran, IRAN, + 98 21 73913138,*

It is worth mentioning that short-term traffic prediction model is quite distinct from conventional traffic prediction models. Conventional models predict traffic volume for the coming years or predict traffic volume in different scenarios while short-term traffic prediction models estimate traffic volume for only a few minutes ahead. Conventional models are based on a four-step method or direct models while short-term models are usually based on heuristic methods. Another difference is related to scope and usage of models. Conventional models are aimed at planning issues; on the other hand, short-term models insist on operation and management issues.

Neural networks and genetic algorithm are parts of Artificial Intelligence which have been applied in different areas successfully ?[1], ?[2]. The method proposed in this paper is a bi-level approach based on Neural Network and Genetic Algorithm. First, a primary model is developed using Network and developed Network is optimized by applying Genetic Algorithm.

2. Background

Successful implementation of ITS systems is highly dependent on the quality of information provided by systems[3]. Thus, continuing efforts have been conducted by various researchers in order to improve short-term traffic prediction models accuracy. During the last two decades, many different models have been presented based on dynamic traffic assignment, statistical methods and neural networks.

To model and forecast short-term traffic flow, statistical time series analyses, in the form of the

ARIMA family of models have been the most widely used approaches [4],[5]. But, although the temporal variability of traffic flow appears to be a crucial characteristic in short-term forecasting, the traditional time series methodologies frequently seem unable to capture the rapid variations of flow in urban areas[6].

Davis and Nihan attempted to replace the time series approach by a non-parametric regression approach, but they concluded that the performance of their k-nearest neighbor approach "performed comparably to, but not definitely better than, the time series approach" [7].

Several attempts have been made using Neural Networks as a substitute for the more traditional regression and time-series approaches [8], [9]. Common to all is a conclusion of potential superiority of Neural Networks and a recommendation for further in-depth investigation under different scenarios and using larger, real databases.

A recent approach in short-term traffic flow prediction is based on combination of Neural Networks and Genetic Algorithm, which is limited only to two cases. In 2002, Abdulhai et al. developed a model based on Time-Delay Neural Networks and GA. In this research, GA was used to determine general structure of TDNN, including number of inputs (look-back time), number of hidden neurons and learning rate. Steepest decent gradient method was selected as learning method [10].

In 2005, Karlaftis and others applied a neuro-genetic method for short-term traffic prediction. Their model was based on Multi-Layered Feed-forward Neural Networks. GA was employed in two levels in their research. First, the proper learning rate and momentum for back-propagation learning method were determined. Then proper number of hidden neurons was chosen using GA. Steepest decent gradient method was also selected as learning method in this research [11].

The approach presented in this paper is based on combination of neural networks and genetic algorithm, yet it differs from previous works in two different areas. In the proposed method, GA is used to optimize links which connect input cells to hidden cells thereby the final network would be partially connected. Furthermore, networks are trained using Levenberg-Marquardt method which is a second order training method, while other researchers used steepest gradient method for training network. In addition, the proposed approach is implemented in a rural highway for 68 day period, considering that rural trips pattern is more disorganized than urban trip patterns.

The remainder of this paper is organized as follows. The next section is a brief overview of neural networks and the fourth section briefly discusses genetic algorithm. The fifth section describes the proposed approach and sixth section presents the empirical results from a case study .Finally, the seventh section summarizes and discusses the findings of the paper.

3. Neural Networks

Neural Networks are part of Artificial Intelligence and inspired by human brain function. Up to now, several types of neural networks have been introduced and MLF neural network is one of these networks. A typical two-layer MLF (consists of a hidden layer and an output layer) can approximate any continuous nonlinear function to an arbitrary precision [12],[13]. Considering this characteristic, MLF neural networks are selected as core architecture for short-term prediction model.

3.1 Neural Network Principles

Neural networks consist of neurons or cells. Cells are grouped into layers. Each cell is connected to next layer cells through links or connection channels. Links transfer cells output to the next layer cells. Links have weights by which cell output is multiplied. The value of weighs is determined during training process. In fact, the knowledge of a neural network is stored in its weights.

3.2 Neural Network Training

A neural network can predict true outputs for corresponding inputs only when network parameters are selected properly. Link weight is

one of the network parameters. A process in which proper values for weights are selected is called network training. Backpropagation is a neural network training method. In this method, inputs are presented to neural network and then network output is calculated; the difference between desired outputs (target outputs) and network outputs (real outputs) shows network error. At the next step, value of weights is revised according to network error. Each interaction of revising weights values is called an epoch.

Neural Network training can be considered as a minimization problem which seeks for weights that minimize network error index ?[14]. Network error index is defined as the sum of squared error between target values and real values.

$$F(\mathbf{w}) = e^T e \quad (1)$$

where $F(\mathrm{w})$ is network error index, $\mathbf{w} = [w_1 \;\; w_2 \;\; \cdots \;\; w_N]$ is a vector containing all weights and e is a vector containing network error for training pairs. Levenberg-Marquardt is a training method for MLF neural Networks?[15]. In this method correction value for each weight (Δw) is calculated using (2):

$$\Delta w = \left[J^T J + \mu I\right]^{-1} J^T e \quad (2)$$

where J is Jacobean matrix and μ is learning rate. Learning rate is selected in each epoch based on results of correction in links weights. If weights correction brings an increase in $F(\mathrm{w})$, μ will be multiplied by decay rate $\beta (0 < \beta < 1)$, and if $F(\mathrm{w})$ decreased, μ will be divided by β. Levenberg-Marquardt can be illustrated in the following steps:

Step 1: Select an initial value for weights and μ (usually μ =0.1).

Step 2: Calculate $F(\mathrm{w})$.

Step 3: Compute Δw using (2).

Step 4: $w' = w + \Delta w$ and calculate $F(\mathrm{w}`)$.

If $F(\mathbf{w}) > F(\mathbf{w}')$

then $\mathbf{w} = \mathbf{w}'$, $\mu = \mu \cdot \beta (\beta = 0.1)$, go to step 2.

Else $\mu = \mu / \beta$, go to step 4.

This algorithm stops when the desired accuracy achieved or if the number of iterations exceeds defined number of iterations. We assume 100 for maximum number of iterations in our study.

Neural networks should be able to calculate true outputs for inputs they have not encountered before. This ability of neural networks is called generalization. In order to assess generalization ability of a neural network, available data is divided into three parts, training set, validation set and test set. Training set data is used to develop model. Validation set data is presented to all models developed, a model which produces best results will be selected as best model. Test set data are used to ensure generalization ability of best model ?[16].

4. Genetic Algorithm

Genetic Algorithm is an evolutionary method which is inspired from evolution of creatures in nature. In this method chromosomes and genes are utilized to solve problems. First, solutions are coded into chromosomes. Then an initial generation of chromosomes is produced. In the next steps, new generations of chromosomes are produced using genetic operators and positive characteristics of each generation are transferred to new generations ?[17]. Ultimately the last generation determines the answer to the problem. Reproduction, cross-over and mutation are most common genetic operators.

- Reproduction operator directly transfers some chromosomes to new generation.
- Cross-over operator selects two chromosomes as parents and then produces two offspring by combining the parents.
- Mutation operator randomly performs some changes in chromosomes in order to prevent search area from restricting to special part of space.

Chromosomes are compared based on their fitness. Chromosomes fitness is computed by fitness function, and then chromosomes are selected using selection mechanism. Ranking method and roulette-wheel are most common selection methods ?[18].

5. Proposed Method

Figure 1 depicts proposed architecture for short-term traffic flow prediction model. The

model receives recent temporal traffic profile in the form of i input variable, each variable represents traffic volume in a t-minute period. It is worth mentioning that in the presented approach, in addition to number of inputs and hidden cells, connection layout of hidden layer is also synthesized,. This means that some links will be omitted (dash lines in Figure 1).

The Proposed model is a MLF neural network including a hidden layer and an output layer. The output layer is only consisted of one cell; output of this cell represents the traffic flow for a t-minute period. The mathematical representation of the model is as follows:

$$\begin{cases} s_j = \sum_{j=0}^{i} w_{jk} V_k \\ z_j = \sigma(s_j) \\ V_t = \sum_{q=0}^{p} w_q z_q \\ k = 1,2,...,i \\ j = 1,2,...,p \end{cases} \quad (3)$$

where V_k refers to traffic volume in k^{th} t-minute interval, w_{jk} is weight of the link which connects k^{th} input to j^{th} hidden cell, s_j is net input of j^{th} hidden cell, σ is sigmoid function, w_q is weight of the link which connects q^{th} hidden cell to output cell and V_t is the traffic volume in t-minute period ahead.

The proposed approach for model development consisted of two major stages; first a primary model is developed using MLF neural networks and in the second stage, primary model is optimized using genetic algorithm. Figure 2 depicts these stages:

5.1 Setting up primary model

In the first step, a neural network is designed as a primary model. Number of input variables and hidden cells are determined in this step. Selection of neural networks inputs and hidden cells is a trail and error process. Trial and error process is started with a network containing minimum inputs and hidden cells, and then the number of inputs and hidden cells will be increased gradually. This process is finished when further addition of inputs and hidden cells do not have positive impact on model performance. The trial and error process is as follows:

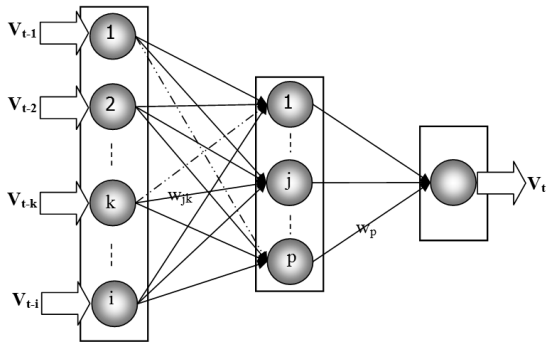

Fig. 1 Model Structure

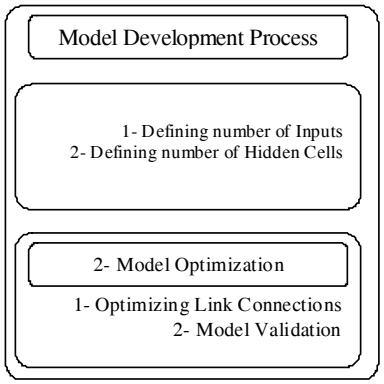

Fig. 2 Stages of model development

Step 1: Select p as number of input variables.

Step 2: Select q as number of hidden cells.

Step 3: Design a network with p inputs and q hidden cells and evaluate its performance.

Step 4: consider network performance

If it reached the desirable performance then go to step 7.

Else go to step 5.

Step 5: Considering last two networks, is network performance improved?

If yes, then go to step 2.

Else: $p=p+1$ and go to step 2.

Step 6: $q=q+1$ and go to step 3.

Step 7: end

Model mean squared error for validation data is a measure of network performance. When trail and error process finishes, the best network is selected and its parameters will be transferred to next step.

In common neural networks, all input variables are connected to hidden cells. These networks are called full connected networks. But full connection sometimes affects model

generalization ability. During training, the value of some link weights may become far greater than other weights. In this case, values carried by other links have no effect on net input of cells. This means that other links are omitted unintentionally.

5.2 Model Optimization

In order to decrease model complexity and also improve model generalization ability, the optimum layout of hidden layer connection links is designed using genetic algorithm. During optimization process, some links will be omitted. This could change model to a partially connected neural network. Figure 3 depicts the optimization process.

5.2.1 Chromosomes Encoding and First Population

Before applying genetic algorithm, answers to the problem should be introduced to algorithm. The aim of this step in developing short-term traffic prediction model is optimization of connection links; so connection layout of neural networks should be encoded in order to be understood by genetic algorithm. Binary encoding method is selected to encode connection links of model. Matrix C represents a typical encoding of connection links in a neural network:

$$C^1 = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pn} \end{bmatrix} \quad (4)$$

According to this encoding, if there exists a connection link from input variable j to hidden cell i, then c_{ij} would be equal to 1 and if else, c_{ij} would be equal to 0.

Considering the number of inputs and hidden cells which were determined in previous step, initial population should be generated. In this regard, several C matrices are generated randomly. Each matrix will be assigned to a neural network. These neural networks are genetic algorithm chromosomes which compromise initial population. Proper population size will be determined during case study.

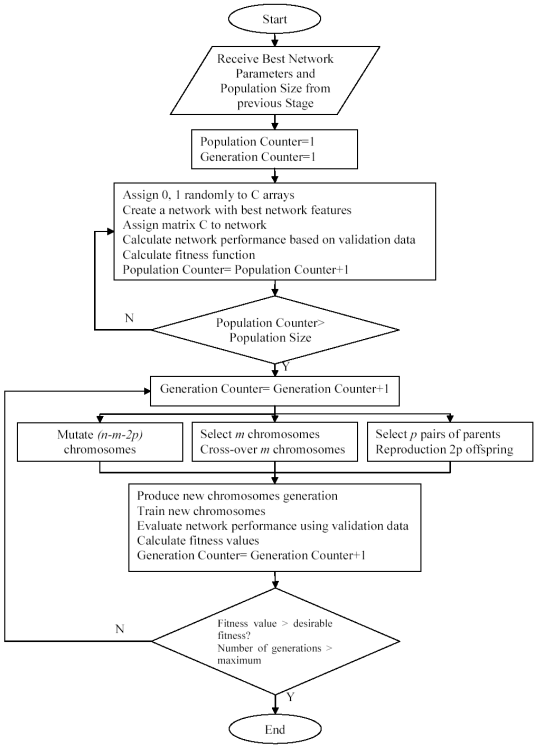

Fig. 3 Optimization process flowchart

5.2.2 Producing New Generations

Genetic algorithm is based on evolution of chromosomes in successive production of new generations. Genetic algorithm transfer superior characteristic of current generations to a new generation. Fitness function is used to evaluate chromosomes characteristics. While genetic algorithm is used to optimize link connections; yet the main purpose is to attain an acceptable model. So, the defined fitness function is representative of model error. Equation 7 shows fitness function:

$$Fitness(net_k) = \frac{1}{1+mse} \quad (5)$$

where *mse* is model mean squared error for validation data.

Genetic algorithm applies selection mechanism to choose chromosomes which are going to undergo genetic operators. Roulette-wheel is a selection mechanism in which chromosomes with high fitness are more likely to be selected.

$$p_r(net_k) = \frac{Fitness(net_k)}{\sum_{i=1}^{n} Fitness(net_i)} \quad (6)$$

where $p_r(net_k)$ is probability of network k to be selected, $Fitness(net_k)$ is fitness value of network k and n is population size. In this method, a random number is generated and the network corresponding to this number will be selected.

Genetic algorithm produces new generations using generic operators but an important question is the share of operators in production process. Proper share of genetic operators will be determined in the case study based on sensitivity analysis.

6. Case Study

Short-term traffic volume prediction models need a continuous temporal profile of traffic as their input. This data is usually obtained from a traffic sensor. The traffic data used in this research is obtained from an inductive loop detector in a rural highway, Ghazvin-Rasht Road in Iran [19]. In following sections, we try to use presented methodology to develop models for this highway.

6.1 Data Collection

As mentioned in previous section, data belongs to an Inductive Loop Detector. This data is recorded in PVR format which means there is a separate record for each vehicle passage in database. The original database contains 385,831 records; this data was recorded in winter 2005 form Jan 10 until March 19 2005. Data belongs to a 68 day period in winter and it is captured continuously 24 hour a day. A computer program in VBA was prepared for data accumulation and then MATLAB was selected as the main workplace for model development.

In order to assess generalization ability of a neural network, available data is divided into three parts, training set, validation set and test set. 50%, 25% and 25% is the share of sets in this study.

6.2 Population Size and Operators Share

Genetic algorithm parameters should be determined prior to model optimization. The parameters are population size, number of generations and operators shares. In order to determine these parameters, an initial model with 5 minute prediction horizon was designed based on section 5-1 guidelines. The model is composed of 6 input variables and 4 hidden cells. Then the proper share of operators was assumed to be reproduction 30%, cross over 50% and mutation 20%. Several optimizations were performed starting with different population sizes. Figure 5 depicts mean squared error of optimization results obtained from different population sizes. The population size of 25 provides the best mean squared error.

An investigation into the effects of number of generation on model performance shows that maximum model performance is achieved in 10th generation and keeping on producing further generations has no effect on model performance.

Operators share also affects the process outcomes. Various operator combinations were examined to find out most suitable operators combination. Figure 5 depicts the result obtained from some tested combinations. The best result occurs when the following combination is applied; reproduction 16%, cross-over 36% and mutation 48%.

6.3 Model Development

Following determination of genetic algorithm parameters in previous step, models with 5, 10 and 15 minutes prediction horizons are built according to the process depicted in Figure 3. Table 1 illustrates primary models for given horizons.

After designing primary models, optimum

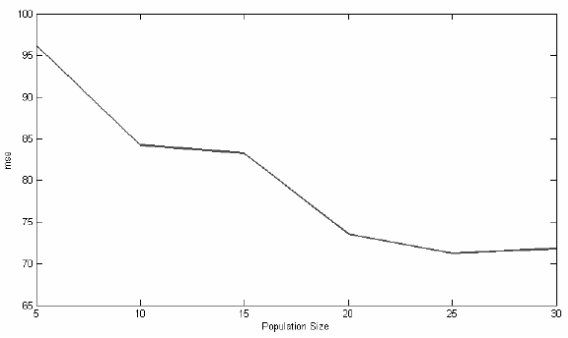

Fig. 4 Error vs Population size

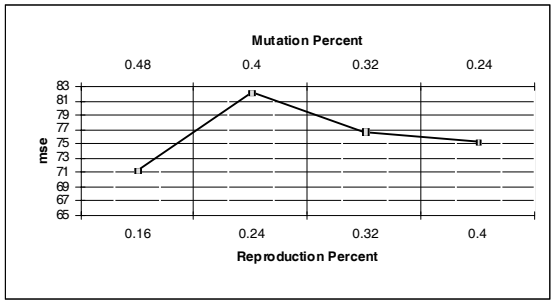

Fig. 5 Effect of operators share on model performance

connection arrangements are synthesized using genetic algorithm. Table 2 presents results obtained from optimization. These models are in fact final models. Table 3 presents models performance for whole data set. Mean squared error, average relative error, correlation coefficient between actual outputs and desired outputs, and r-squared are performance indices.

Figure 6 depicts 15-minute model predicted flow and observed flow in terms of passenger car unit for each data set. Model generalization ability is impressive and model has acquired traffic pattern well.

The results clearly show that although some of connection links are omitted, model performance has improved.

7- Conclusion

Table. 1 Primary Models Specification

Prediction Interval (Minute)	Number of Inputs	Number of Hidden Cells	Number of Links	r^2
5	6	4	24	0.722
10	3	3	9	0.831
15	6	4	24	0.856

Table. 2 Final Models Specification

Prediction Interval (Minute)	Number of Inputs	Number of Hidden Cells	Number of Links	r^2
5	6	4	15	0.734
10	3	3	6	0.846
15	6	4	19	0.867

Table. 3 Final Models Performance (for whole data set)

Prediction Interval (Minute)	Mean Squared Error	Average Relative Error	Correlation Coefficient	r^2
5	63.25	14.48	0.73	0.734
10	140.48	11.49	0.84	0.846
15	261.77	11.70	0.86	0.867

This paper has presented a short-term traffic flow prediction approach and produced a system based on an advanced Multi-Layer Feed forward Neural Network (MLF) model synthesized using Genetic Algorithms (GA).The model predicts flow values based on their recent temporal profile at a given highway site during the past few minutes. The model performance was validated using real traffic flow data obtained from the field.

Results obtained from training data, validation data and test data adduce that model generalization ability is satisfactory. For the case of 15-minute prediction horizon for instance, prediction r^2 indices are 0.85, 0.84 and 0.83 respectively which evidently shows model generalization ability.

Moreover, it was found that the longer the extent of prediction, the more the predicted values tend toward the mean of the actual for a given data resolution. In the case of optimization effectiveness, proposed approach performed well on reducing model complexity, meanwhile improving model generalization ability. As an

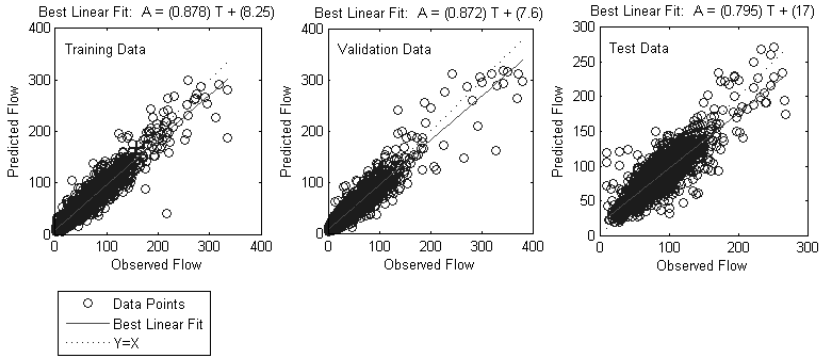

Fig. 6 Predicted flow vs observed flow for 15-minute horizon

example, number of links has been reduced 37% for 5-minute prediction model; furthermore, model r2 index increased from 0.72 to 0.73. The model performed acceptably using field data, bearing in mind that the real-data used is from a rural highway in winter, that is, the pattern of traffic is totally different from an urban highway.

References

[1] Wang, Lipo, Fu Xiuju, "Data Mining with Computational Intelligence", Germany, Springer, 2005.

[2] Kasabov, Nikola K., "Foundations of Neural Networks, Fuzzy Systems and Knowledge Engineering", England, MIT Press, 1996.

[3] Kan Chen; John C. Miles, "ITS Handbook 2000: Recommendations from the World Road Association (PIARC),"Artech House, October, 1999.

[4] Kirby, H., Dougherty, M., Watson, S.," Should we use neural networks or statistical models for short term motorway forecasting", International Journal of Forecasting 13, 1997.

[5] Williams, B.M. "Multivariate Vehicular Traffic Flow Prediction: An Evaluation of ARIMAX Modeling", Transportation Research Record 1776, 2001.

[6] Abdulhai, B., Porwal, H., Recker, W., „Short-term freeway traffic flow prediction using genetically-optimized time-delay-based neural networks", Transportation Research Board, 1999.

[7] Davis, G. and Nihan, N., "Nonparametric Regression and Short-Term Freeway Traffic Forecasting," Journal of Transportation Engineering, 117, 178–188, 1991.

[8] Smith, B. and Demetsky, M., "Short-Term Traffic Flow Prediction: Neural Network Approach," Transportation Research Record 1453, 1995.

[9] Dougherty,M. and Lechevallier,Y., "Short-Term Road Traffic Forecasting Using Neural Network," Recherche Transports Securite, English Edition 11, 1995.

[10] Abdulhai, Baher; Himanshu, Porwal; Will, Recker, "Short-Term Traffic Flow Prediction Using Neuro-Genetic Algorithms", ITS Journal, Vol. 7, pp.3–41, 2002

[11] Karlaftis, Matthew, G.; Eleni, I. Vlahogianni; John C. Golias, "Optimized and meta-optimized neural networks for short-term traffic flow prediction: A genetic approach", Transportation Research, Part C, Vol. 13, pp.211-234, 2005.

[12] Principe, J.C., Euliano, N.R., Lefebvre, C.W., "Neural And Adaptive Systems: Fundamentals Through Simulations", John Wiley and Sons, New York, 2000.

[13] Ham F. M., I. Kostanic, "Principles of Neurocomputing for Science and Engineering", McGraw Hill, New York, 2001.

[14] Haykin, S., "Neural Networks: A Comprehensive Foundation", McMillan, New York, 1994.

[15] Hagan M. T., M. B. Menhaj, "Training feed forward network with the Marquardt algorithm", IEEE Trans. on Neural Net., Vol. 5, No. 6, pp.989-993, 1994.

[16] Gupta, Madan, M. Liang Jin, Noriyasu Homma, "Static and Dynamic Neural Networks: From Fundamentals to Advanced Theory", John Wiley & Sons, 2003.

[17] Haupt, Randy L., Sue Ellen Haupt, "Practical Genetic Algorithms", New Jersey, Wiley Interscience, 2004.

[18] Melanie Mitchell, "An Introduction to Genetic Algorithms", MIT Press, Massachusetts, 1998.

[19] Iran Road Maintenance and Transportation Organization (RMTO) , www.rtmo.ir