EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

ELSEVIER

European Journal of Operational Research 131 (2001) 253–261

www.elsevier.com/locate/dsw

An object-oriented neural network approach to short-term traffic forecasting

Hussein Dia

Department of Civil Engineering, The University of Queensland, Brisbane, Qld 4072, Australia

Abstract

This paper discusses an object-oriented neural network model that was developed for predicting short-term traffic conditions on a section of the Pacific Highway between Brisbane and the Gold Coast in Queensland, Australia. The feasibility of this approach is demonstrated through a time-lag recurrent network (TLRN) which was developed for predicting speed data up to 15 minutes into the future. The results obtained indicate that the TLRN is capable of predicting speed up to 5 minutes into the future with a high degree of accuracy (90–94%). Similar models, which were developed for predicting freeway travel times on the same facility, were successful in predicting travel times up to 15 minutes into the future with a similar degree of accuracy (93–95%). These results represent substantial improvements on conventional model performance and clearly demonstrate the feasibility of using the object-oriented approach for short-term traffic prediction.

Keywords: Neural networks; Traffic forecasting; Intelligent transportation systems; Advanced traffic management systems; Advanced traffic information systems

1. Introduction

The prediction of short-term traffic conditions is a vital component of advanced traffic management and information systems which aim to influence travel behaviour, reduce traffic congestion, improve mobility and enhance air quality. Traffic prediction models can be used to provide metropolitan traffic control centres with an automated tool for anticipating the congestion that may arise on road facilities and its expected duration. This information can then be provided to drivers in real-time to give them realistic estimates of travel times, expected delays and alternative routes to their destinations. Providing drivers with this information is believed to have the potential to alleviate traffic congestion and enhance the performance of the road network.

Traffic information provided to drivers may conceptually fall into one of three categories: *historical*, *current* and *predictive*. Historical information describes the state of the transportation system during previous time periods. Current information is the most up-to-date information about traffic conditions. A number of currently available intelligent transport systems (ITS) technologies allow for the provision of this information

E-mail address: h.dia@uq.edu.au (H. Dia).

0377-2217/01/$ - see front matter
PII: S 0 3 7 7 - 2 2 1 7 (0 0) 0 0 1 2 5 - 9

in real-time at intervals less than 20 seconds on a 24-hour basis. Predictive information, on the other hand, falls into two distinct categories: strategic and short-term. Strategic information is mainly needed for major decisions on road planning and includes prediction of traffic flows and conditions months or years into the future. In contrast, short-term predictive information often has a horizon of only a few minutes and is therefore more suited to implementation in traffic management and information systems.

Most of today's traffic control systems rely mainly on historical and current traffic data as a basis for traffic management actions. The performance of these systems is constrained because they lack the predictive capabilities. Ideally, traffic conditions should be anticipated and actions should be planned accordingly. Since drivers' decisions are affected by expected network conditions, it is also clear that the most useful type of information for a driver faced with travel choices is reliable predictive information. Drivers making travel decisions in the absence of predictive information are implicitly projecting future conditions from the historical and current information available to them. Therefore, predictions of what traffic conditions are likely to be in a few minutes' time (e.g., 5–10 minutes into the future) are needed for effective traffic management and information systems.

This paper describes the development and evaluation of an object-oriented neural network model that was developed to predict speed at a detector station up to 15 minutes into the future and shows how similar models were developed for freeway travel time estimation. The results reported in this paper clearly demonstrate the feasibility of implementing this approach for freeway short-term traffic forecasting and the potential for its implementation in other real-time applications such as arterial travel time estimation and automatic incident detection.

2. Previous research work

A number of traffic prediction algorithms have been developed or proposed over the last two decades. Their structure varies in the degree of sophistication, complexity and data requirements. Inductive loop detectors, embedded in the freeway pavement, are typically used to obtain the traffic data needed for these algorithms. Some of the most widely used traffic prediction models are those based on spectral analysis, ARIMA time-series models, Box–Jenkins analysis and Kalman filtering (Clark et al., 1993). However, these models have been shown to exhibit varying levels of performance especially during congested conditions (Okutani and Stephanedes, 1984). Some of the disadvantages of previous models also include the averaging (smoothing) of input data over long time intervals (e.g., 5 minutes). This may result in obscuring many of the time-space relationships in the data and could be a significant factor contributing to their poor predictive performance. It should be mentioned here that prediction errors in excess of 25% have been reported in a number of studies where such smoothing techniques have been used (Blue et al., 1994).

More recently, a number of studies have also investigated neural networks for predicting short-term traffic conditions (e.g., see Dougherty et al., 1994; Smith and Demetsky, 1994; Dougherty, 1995). The main limitation of these studies, however, is that they mainly implement static neural network architectures which may not fully capture the true dynamics of the underlying traffic data.

In contrast to previous studies, this paper discusses an object-oriented dynamic neural network model for predicting short-term traffic conditions. The object-oriented approach draws on recent advances in the theory of neural computation based on the principle of local rules of interaction among simple neural components (Principe et al., 1999). Using this approach, the neural network model is divided into functional blocks which communicate with each other in planes of activation which implement both the neural network and learning dynamics. A hierarchy of classes, which encapsulate the standard rules of interaction, allows complex network topologies to be constructed by simply interconnecting a small number of instantiated classes (Principe et al., 1999). This approach provides for modelling complex interactions such as mixing supervised and unsupervised learning rules in the same network or incorporating a recurrent processing element into the hidden layer of a feed-forward topology without the need for deriving new learning equations. Due to its object-oriented nature, this approach specifies what components do and how they interact with each other, rather than specifying rigid implementation of functions as in conventional programming (Principe et al., 1999).

The feasibility of the object-oriented approach is demonstrated in this paper through a time-lag recurrent network (TLRN) which was developed for predicting short-term speed data (up to 15 minutes into the future) on a section of the Pacific Highway between Brisbane and the Gold Coast in Queensland, Australia. The TLRN was trained via trajectory learning using a back-propagation through time (BPTT) algorithm. The paper also demonstrates how the object-oriented approach allows for modelling complex networks with a mixture of learning rules and processing element interactions which previously were difficult to model using conventional neural network paradigms.

3. Data for model development and evaluation

The neural network models presented in this paper were developed using field data collected from four inductive loop detector stations installed on a 1.5-km section of the Pacific Highway between Brisbane and the Gold Coast (Fig. 1). Inductive loop detectors were installed at approximately 500 m intervals to collect speed and flow data from each detector station (Lam et al., 1996). The raw data used in this study were collected over a 5-hour period on 2 days in April 1995. The 5-hour period comprised 2 hours of peak and 3 hours of non-peak conditions. The raw data from the three sections were then averaged over 20 seconds cycles resulting in a total of 5000 observations to be used in model development and evaluation.

The master data set (comprising 5000 observations) was divided into three data sets to be used for training, cross-validation and testing. The training set (comprising 60% of the data) was used for determining the network parameters while the cross-validation data set (10% of the data) was used to prevent the network from learning the idiosyncrasies in the training set and thus enable the model to generalise better. A third test set (comprising 30% of the data) was set aside for validating the performance of the trained models. This test set was independent of the data sets used for model training. Table 1 shows the breakdown of observations in the three data sets.

4. Neural network training

In order to develop a neural network model to perform traffic prediction, the network needs to be trained with historical examples of input–output

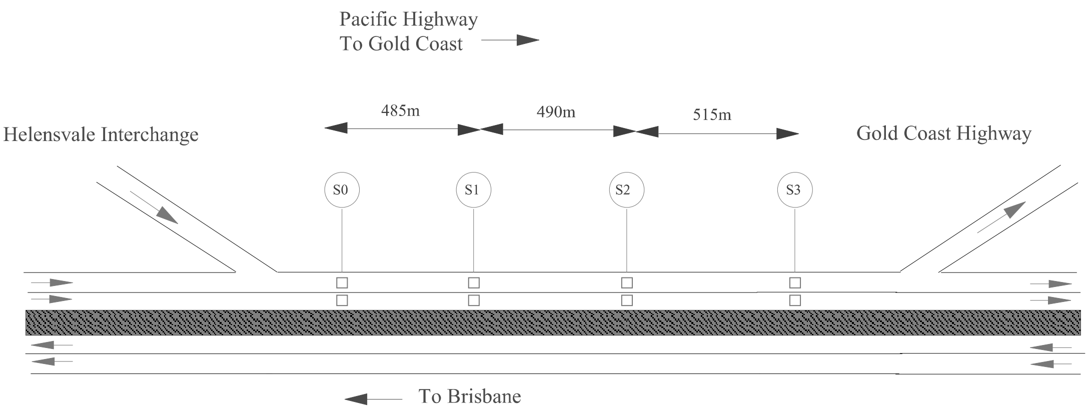

Fig. 1. Schematic of the Pacific Highway section used for data collection.

Table 1
Number of observations in the data sets used for model development

Data set	Percentage of master data set (%)	Number of observations
Training	60	3000
Cross-validation	10	500
Testing	30	1500
Total	100	5000

data. This paper will present the results for speed prediction only. Speed measurements from the current time interval (t_0) at a given detector station form the input to the neural network model. The output of the model comprises the speed measurements at the same station at some future time interval t_n.

4.1. Neural network architecture

As part of the model development process, decisions must be made about the architecture of the neural network. A number of experiments were conducted to determine the best neural network architecture. Four neural network models were developed to predict speed in the next time interval (i.e., 20 seconds into the future) based on current speed measurements. The models were then evaluated based on the prediction errors for the three data sets described earlier. A brief description of each of the models developed in this study is provided below (Principe et al., 1999).

4.1.1. Multilayer perceptrons (MLPs)

These layered feed-forward networks were trained with static back-propagation. Their main advantage is that they are easy to use, and that they can approximate any input–output map. The key disadvantages, however, are that they train slowly and that they implement static neural network architectures which may limit their predictive capabilities. These networks are presented here to serve as a base scenario against which other dynamic architectures can be compared.

4.1.2. Recurrent networks

These include both the fully and partially recurrent networks. Fully recurrent networks feed back the hidden layer to itself. Partially recurrent networks start with a fully recurrent network and add a feed-forward connection that bypasses the recurrency (Principe et al., 1999). These networks can have an infinite memory depth and thus find relationships through time as well as through the instantaneous input space.

4.1.3. TLRNs

Unlike most other neural network architectures which are purely static classifiers, these TLRNs are extended with short-term memory structures which makes them more suitable for non-linear time series prediction, system identification and temporal pattern classification.

4.1.4. Hybrid networks

The hybrid networks tested in this study comprised a combination of a principal component analysis (PCA) network and a TLRN. These networks combine both supervised and unsupervised learning in the same topology. Principal component analysis is an unsupervised linear procedure that finds a set of un-correlated features (principal components) from the input. A supervised procedure is then used to perform the non-linear classification from these components (Principe et al., 1999).

Each of the above models was trained on the training data set for a maximum of 1000 cycles. At the end of each cycle, the trained model was tested on the cross-validation set and the prediction error (i.e., mean squared error (MSE)) was computed. If the prediction error decreased, the model was saved and the training was continued. Training was stopped when the prediction error on the cross-validation data increased by a specified threshold. Table 2 below presents the results for each of the networks investigated in this study. These results clearly demonstrate the superior predictive performance of the dynamic neural network architectures when compared to the static MLP.

Table 2
MSE of predicted speed for a number of neural network architectures

Network type	MSE			
	Training set	Cross-validation set	Test set	Average of three sets
MLP	0.0253	0.0261	0.0098	0.0204
Hybrid	0.0087	0.0075	0.0096	0.0086
Recurrent	0.0088	0.0077	0.0075	0.0080
TLRN	0.0073	0.0076	0.0086	0.0078

5. Model performance evaluation

Based on the results reported in Table 2, the TLRN model was selected for further investigation. Table 3 below lists the performance results for nine TLRN models that were developed to provide speed forecasts for prediction horizons up to 15 minutes. The average percent errors (computed as the percentage difference between the actual and predicted speed) are based on the testing data set and are indicative of the generalisation performance of the model when applied in the field. These results clearly show that the neural network models are capable of predicting speed data up to 5 minutes into the future with a high degree of accuracy (90–94%). The accuracy of the models was found to decrease only to 88% and 84% for prediction horizons of 10 and 15 minutes, respectively.

Sample plots of actual and predicted speed measurements (20 and 60 seconds into the future) are presented in Figs. 2 and 3, respectively. These plots clearly demonstrate the high predictive accuracy of the models during both normal and incident conditions.

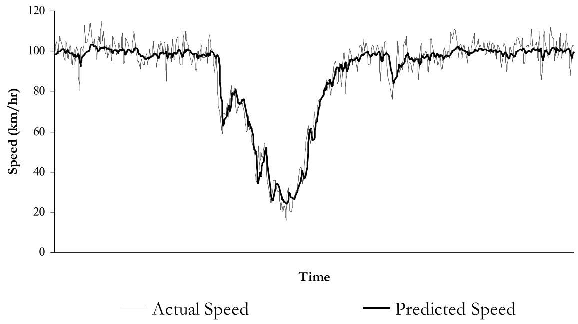

Fig. 2. Sample plot of actual and predicted speeds for a 20-seconds prediction horizon.

6. Freeway travel time estimation

The effectiveness of a wide range of ITS control strategies designed to alleviate traffic congestion depends heavily on the accuracy, credibility and reliability of travel time estimates. Dynamic estimates of travel time can be used to inform drivers of expected delays and travel times to their destinations and are important inputs into a number of real-time applications such as automatic incident detection and dynamic route guidance.

Table 3
Speed performance measures for different prediction horizons

Prediction horizon	Prediction error (MSE) (10^{-4})			Average percent error (%) (testing set)
	Training set	Testing set	Average of two sets	
20 seconds	73	86	80	6
40 seconds	85	99	88	7
1 minutes	95	111	103	7
2 minutes	126	129	127	8
3 minutes	156	151	154	8
4 minutes	183	172	178	9
5 minutes	215	199	207	10
10 minutes	337	358	348	12
15 minutes	443	522	483	16

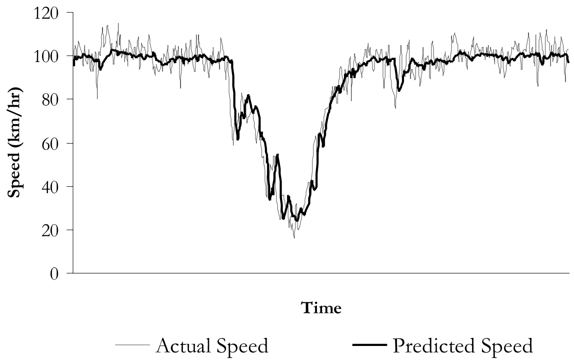

Fig. 3. Sample plot of actual and predicted speeds for a 60-seconds prediction horizon.

The next section describes briefly how the modelling framework presented previously has been implemented to estimate travel times on the same sections of the Pacific Highway.

6.1. Modelling framework

The neural network freeway travel time estimation modelling framework is shown in Fig. 4. This modelling framework is based on the hypothesis that knowledge of current traffic conditions and demands to use the facility are sufficient indicators for predicting the future state of traffic on a freeway section a few minutes into the future.

Consider the section of freeway shown in Fig. 4(a) which is defined by upstream and downstream detector locations. Traffic conditions at each station will be determined through two traffic variables. These are speed (km/h) and flow (number of vehicles passing the loop detectors in a unit time). The neural network to be used for prediction consists of an input layer comprising speed and flow data from the upstream and downstream stations as shown in Fig. 4(b). The output of the neural network is the travel time within the section at some future state (e.g., 20 seconds to 15 minutes into the future).

Fig. 4 illustrates how the neural network model works when performing travel time estimation within the section. The current data (at time interval t) from the upstream and downstream stations are used as input to the model. The neural network model then undergoes a training process during which the weights associated with the interconnections and thresholds associated with the nodes or processing elements are established. The output of the model represents the travel time within the same section at some future state ($t + n$, where $n = 20$ seconds to 15 minutes). These output estimates are then compared with the historical travel time values which were determined by video taping the sections under consideration and calculating travel times based on vehicle license plate matching. Prediction errors (degree to which the predicted travel times differ from the actual values) are computed. The training process continues until the prediction errors become very small and the network parameters converge to values that allow it to perform the desired mapping for each input–output example. Once the network is trained and its performance is shown to be satisfactory, it can then be used on-line to provide travel time predictions based on new real-time data.

6.2. Neural network model inputs

Table 4 shows the different input parameters tested in this study and their associated MSE results. These results clearly show that the upstream and downstream speeds are the most critical inputs to the model. The best performance model (an average MSE of 35×10^{-5} for the three data sets) is obtained by using speed and flow from both the upstream and downstream stations. Based on these results, it was decided to implement the model using speed and flow inputs from both the upstream and downstream stations.

6.3. Model performance evaluation

Table 5 lists the performance results for a number of models that were developed to predict travel times for prediction horizons up to 15 minutes. These results clearly show that the neural network models are capable of predicting travel times up to 15 minutes into the future with a high degree of accuracy (93–95%).

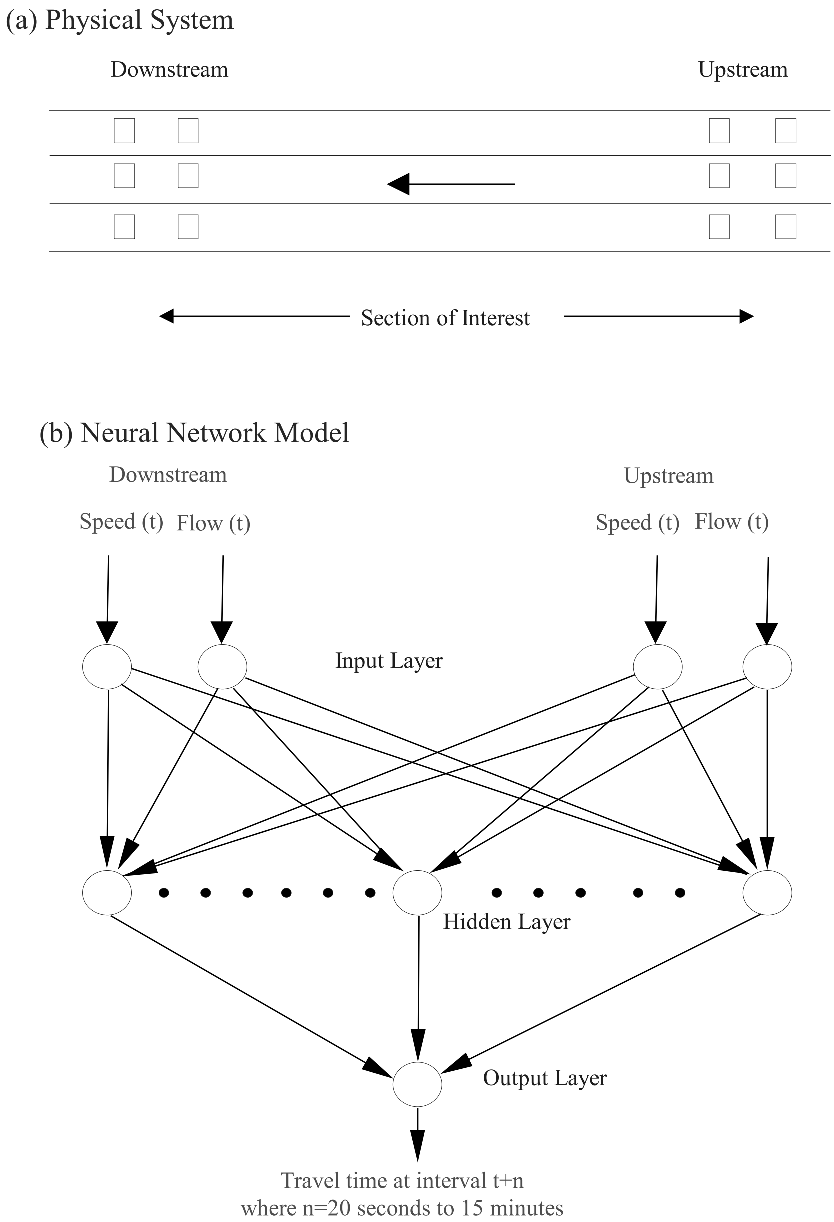

Fig. 4. Neural network freeway travel time estimation modelling framework.

7. Conclusions

The results reported in this paper clearly demonstrate the superior predictive performance of the dynamic neural network architectures (e.g., TLRN, recurrent and hybrid networks) compared to the static classifiers (e.g., MLPs). These results also demonstrate the feasibility of using the object-oriented neural network approach for short-term traffic forecasting. The models described in this paper were capable of predicting speed data up to 5 minutes into the future with a high degree of

Table 4
MSE of predicted travel times for different inputs[a]

Neural network input				Prediction error (MSE) (10^{-5})			
Upstream[b]		Downstream		Training	Cross-validation set	Testing set	Average of three sets
Speed	Flow	Speed	Flow				
✓	✓	✓	✓	95	1	9	35
✓	×	✓	✓	97	11	8	39
✓	×	✓	×	99	9	9	39
✓	✓	✓	×	99	9	11	40
×	✓	✓	✓	134	12	15	54
×	×	✓	✓	152	13	15	60
×	✓	✓	×	139	28	16	61
✓	✓	×	✓	318	47	20	128
✓	✓	×	×	340	47	19	135
✓	×	×	✓	343	47	18	136
×	✓	×	✓	1140	90	94	441

[a] ✓: input included; ×: input excluded.
[b] Average section length between upstream and downstream stations is 497 m (see Fig. 1).

Table 5
Travel time performance measures for different prediction horizons

Prediction horizon	Prediction error (MSE) (10^{-5})			Average percent error (%) (testing set)
	Training set	Testing set	Average of two sets	
20 seconds	97	12	55	5
1 minutes	167	14	91	6
2 minutes	285	24	155	6
4 minutes	647	24	346	7
5 minutes	945	24	485	7
10 minutes	1039	33	536	7
15 minutes	1299	37	668	7

accuracy (90–94%). Similar models that were developed for freeway travel time estimation were also successful in predicting travel times up to 15 minutes into the future with a similar degree of accuracy (93–95%). Further improvements in model performance (e.g., for prediction horizons greater than 5 minutes) may be obtained using additional traffic data inputs (e.g., flow and occupancy) and measurements from previous time intervals. Work is currently underway to collect larger data sets from a number of different locations to evaluate model performance under recurrent and non-recurrent congestion on both freeway and arterial networks. There is also scope in future research efforts to compare the performance of the neural network models with conventional statistical models based on the same data sets.

Acknowledgements

The author wishes to thank Mr. Richard Lam from the Department of Main Roads in Queensland and Mr. Ron Roper from ARRB Transport Research Ltd. for providing the data for this study. The author also wishes to thank the anonymous referees for their valuable comments and suggestions.

References

Blue, V., List, G., Embrechts, M., 1994. Neural net freeway travel time estimation. In: Proceedings of Intelligent Engineering Systems through Artificial Neural Networks, vol. 4. American Society of Mechanical Engineers, New York, pp. 11135–11140.

Clark, S.D., Dougherty, M.S., Kirby, H.R., 1993. The use of neural network and time series models for short term forecasting: A comparative study. In: Proceedings of PTRC meeting, Manchester.

Dougherty, M.S., 1995. A review of neural networks applied to transport. Transportation Research C 3 (4), 247–260.

Dougherty, M.S., Kirby, H.R., Boyle, R.D., 1994. Using neural networks to recognise, predict and model traffic. In: Ambrosino, G., Boero, M., Bielli, M. (Eds.), Artificial Intelligence in Traffic Engineering. VSP, Zeist, The Netherlands, pp. 233–250.

Lam, R., Ramsay, E., Luk, J., 1996. Congestion management measures for gold coast Pacific Highway. In: Proceedings of the Combined 18th ARRB Transport Research Conference and Transit New Zealand Land Transport Symposium (vol. 7, pp. 99–114), Christchurch, New Zealand.

Okutani, I., Stephanedes, Y.J., 1984. Dynamic prediction of traffic volume through Kalman Filtering Theory. Transportation Research Part B 18 (1), 1–11.

Principe, J.C., Euliano, N.R., Lefebvre, W.C., 1999. Neural and Adaptive Systems: Fundamentals Through Simulation. Wiley, New York.

Smith, B.L., Demetsky, M.J., 1994. Short-term traffic flow prediction: Neural network approach. Transportation Research Record 1453, 98–104.