Application of Graph Neural Network to Traffic Information Big Data and Evaluation by Sites

R. Ogata [a], T. Miyazaki [a], Y. Kikuchi [a], Y. Murano [a], H. Sugawara [a]

[a] *Yachiyo Engineering Co., Ltd.*

In order to predict real-time traffic flow and control traffic dynamically to avoid traffic congestion, it is necessary to predict traffic speed with high accuracy. In this paper, short-time traffic speed prediction is conducted using traffic data from England, which is open and can obtain a wide range of traffic information, with an eye toward the aforementioned application. As in Ogata et al. (2023), in addition to machine learning methods such as LightGBM and GNN (Graph Neural Network), we also used LSTM, which is often used in time series forecasting. The comparison with their results for the entire 170 sites showed that the GNN was the most accurate, but LSTM may be superior depending on the value of partial autocorrelation. Finally, issues for improving the GNN are discussed.

Key Words: traffic prediction, machine learning, graph neural network, open data

1 Introduction

Since former President Barack Obama's efforts to promote open government in the United States, public organisations have been opening their data to the public in recent years. Traffic data is also becoming more open, and studies are being conducted to avoid traffic congestion.

In considering how to utilize traffic volume data, Miyazaki et al. (2023) focused on the publicly available traffic data of England and studied to predict traffic congestion using a decision tree model LightGBM. Long Short Term Memory (LSTM) is also often used for time series forecasting and was applied in traffic congestion forecasting (Fu et al., (2016); Zhao et al., (2017)). In addition, many methods based on Graph Neural Networks (GNN), which acquire spatial features through graph structures, have been developed in recent years and have demonstrated the highest accuracy on various benchmark datasets (Yu et al., (2018); Wu et al., (2019); Jiang et al., (2022); etc.). However, when considering decision tree models or RNN models such as LSTM, the number of target sites in many studies is limited to one or a few. In addition, the evaluation of the benchmark dataset only includes an overall evaluation of all sites, not individual sites. Ogata et al. (2023) conducted a site-by-site evaluation, but this was a comparison of only LightGBM and GNNs, and there was no study using RNN methods commonly used in time-series forecasting.

The purpose of this paper is to compare short-time speed predictions by sites using traffic data in England. We use methods including not only LightGBM and GNN but also LSTM and clarify the differences in the evaluations by model for each site.

2 Methodology

2.1 Data

The WebTRIS Traffic Flow API of National Highways (Highways England, 2023), which provides an overview of the traffic situation in England since 2015, allows users to download historical traffic volume data for various roads in England. In this study, time series traffic data at each observation site were downloaded from the WebTRIS Traffic Flow API, as in Miyazaki et al. (2022).

As in Ogata et al. (2023), a total of 170 sites were selected within a 30 km radius of Birmingham Town Hall. The data period was from January 1, 2016, to December 31, 2019, and used 2016 and 2017 as training data, 2018 as validation data, and 2019 as test data. Data were provided in 15-min intervals, and average speed (Avg mph) and total volume (Total Volume) were used. The average speed and total number of vehicles for 3 h (12 times) from the current situation to 2 h and 45 min ago were used as inputs, along with dummy variables (0 to 6) for the time of day (15-min time units, normalised to 0 to 1) and the day of the week. Because LightGBM and LSTM were used in training and inference for each of the 170 target sites individually, input data was also used for each site. Moreover, GNN uses the data of all sites as input data as training and inference were conducted for all sites at once. The output of each model is the average speed from 15 min to 180 min ahead for all 12 time periods.

2.2 Model

In this paper, the models were compared using machine learning methods. As in Ogata et al. (2023), the models used were LightGBM and Graph Wavenet, which is one of the GNNs that have been studied extensively in recent years. In addition to them, LSTM, which is commonly used for time series forecasting, was also used in the study.

The problem is that decision tree models such as LightGBM and RNN models, including LSTM, cannot account for spatial site-to-site dependencies. Moreover, GNNs are expected to improve accuracy by capturing the relationship between sites in a graph structure. Graph Convolutional Networks (GCN) are commonly used for GNNs in the field of traffic flow, but the Graph WaveNet used in this paper introduced the "self-adaptive adjacency matrix" method, which was able to acquire spatially independent relationships between sites and achieves high accuracy. For more information, please refer to the paper by Wu et al. (2019).

The Python library “optuna” (Akiba, 2019) was used for hyperparameter search in LightGBM, and the optimal parameters were adjusted for each site. As LSTM is computationally expensive, multiple attempts were conducted to set parameters. The number of epochs was set to 200 (it was stopped if there were no updates with an accuracy of 40 epochs) and the layers were set as the LSTM layer (32 neurons) + Dropout (0.3) layer + Dense layer. The hyperparameters of Graph WaveNet were set to default, and the number of epochs was set to 300.

3 Results

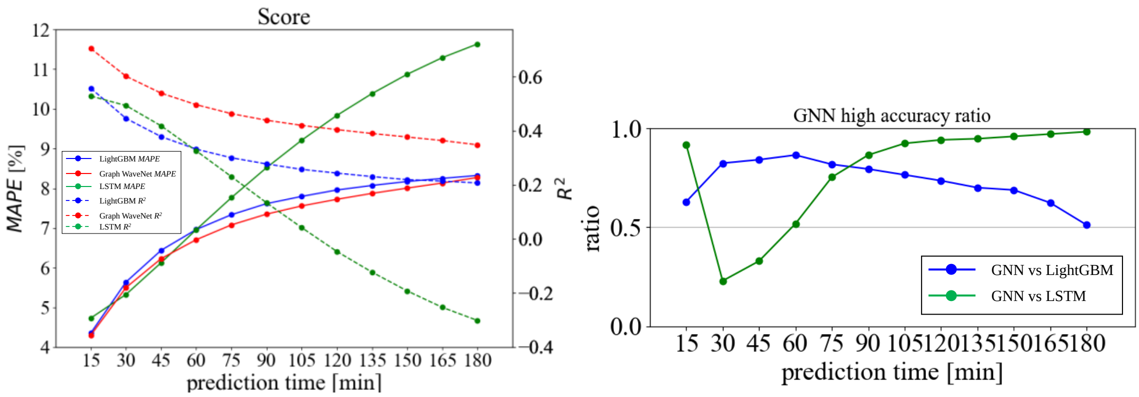
The Mean Average Percentage Error ($MAPE$) and the coefficient of determination (R^2) were used for a uniform evaluation of all sites. **Figure 1** (left) shows the average values of the calculated evaluation indices for all sites. The lower the value of $MAPE$, the better the accuracy, and the higher the value of the coefficient of determination, the better the accuracy. The results show that LSTM is more precise at making predictions 30 and 45 min in advance, whereas GNN is better at making predictions at other times. Because LSTM makes predictions based on past time series information, the accuracy of predictions made 30 and 45 min ahead of time was improved by capturing past features, however, predictions made beyond that point were judged to have benefited from features related to the time of day or other locations. As reported by Ogata et al. (2023), Graph WaveNet automatically acquires the spatio-temporal dependencies of the sites, which contributed to its high accuracy.

Next, **Figure 1** (right) shows the results of calculating the percentage of sites where the GNN is superior to the LightGBM and LSTM at each prediction time. Root Mean Square Error ($RMSE$) was used here to evaluate each site. The GNN was superior to most prediction times throughout, and there were many sites where the GNN had a consistent advantage in comparison to LightGBM (Ogata et al., (2023)). Moreover, GNN was inferior to LSTM in predicting 30 and 45 min ahead of time. As a result, LSTM was superior in the evaluation of all sites for both 30 and 45 min predictions, as described earlier (**Figure 1** (left)). Both LightGBM and LSTM were found to be more accurate than GNN in some cases and at some sites. $RMSE_i^{model1-model2}$ is defined in the following equation, with LightGBM and LSTM compared to GNN. Note that i is the predicted time index; e.g., i =3 represents 45 (=15x3) min ahead.

$$RMSE_i^{model1-model2} = RMSE_i^{model1} - RMSE_i^{model2}$$

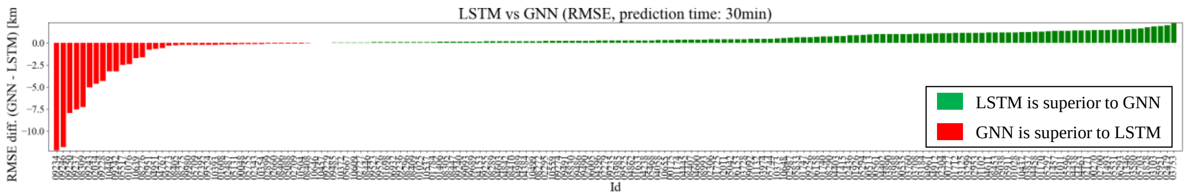
The results of the three models comparison for the entire prediction time, i = 1 to 12 averages for $RMSE_i^{model}$, confirmed that the GNN was superior at 70% of all sites. Since LSTM showed high accuracy at 30 and 45 min ahead, **Figure 2** shows the 30-min-ahead prediction results $RMSE_2^{GNN-LSTM}$ of LSTM and GNN, which shows overall high accuracy, sorted by $RMSE_2^{GNN-LSTM}$ values (ascending order from left to right). The results show LSTM was superior at most sites. Partial autocorrelations were calculated using the Python library "statsmodels" (Perktold, 2023) for the sites with the largest differences, Id = 3753 (LSTM advantage) and Id = 9234 (GNN advantage), and the results are shown in **Figure 3** (left and middle). The figure shows that the correlation at Id=3753 has a high value of 0.8 at lag = 1. Moreover, at Id = 9234, the lag decreases gently as the lag increases. The LSTM advantage in the 30 and 45 min ahead predictions was attributed to the high partial autocorrelation at lag = 1. A comparison of partial autocorrelation averages for LSTM-advantage and GNN-advantage sites confirmed a similar trend (**Figure 3** (right)). These results indicate that the high partial autocorrelation of the LSTM at lag = 1 is the reason for its high accuracy in predicting 30 and 45 min ahead of time.

Figure 1. (left) Accuracy comparison of LightGBM (blue line), LSTM (green line) and GNN (red line) (*MAPE* (solid line, left axis), R^2 (dotted line, right axis) (Right) Percentage of sites where GNN is superior (vs LightGBM: blue line, vs LSTM: green line)



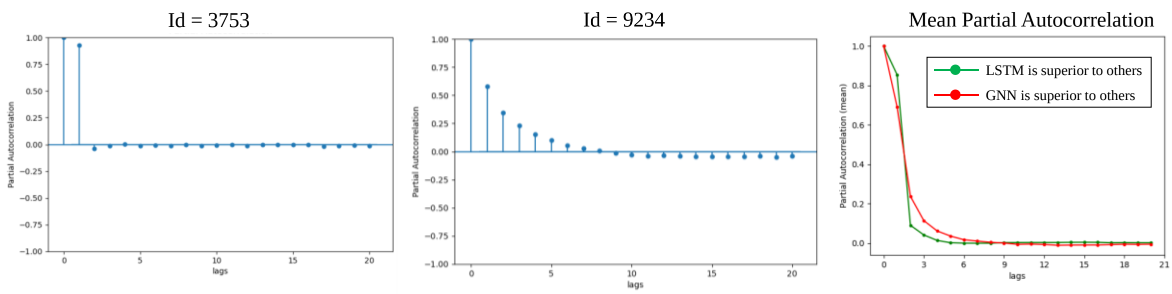
Source: own elaborations

Figure 2. Comparison of LSTM and GNN (30 min ahead forecast, $RMSE_2^{GNN} - RMSE_2^{LSTM}$)



Source: own elaborations

Figure 3. Partial Autocorrelation (left: Id=3753, middle: Id=9234, right; LSTM/GNN advantage sites mean)



Source: own elaborations

Conclusion

In this paper, we compare the results of average speed predictions from 15 to 180 min ahead of time by applying LightGBM, LSTM, and Graph WaveNet as GNN, to open data in England. In the evaluation of all 170 sites, the GNN was superior in the average of all predicted times, but the LSTM may be superior depending on the value of partial autocorrelation, such as 30 min ahead or 45 min ahead in this study. However, when predicting at all sites with large data sets, LSTM should be treated with caution because of the large computational cost during training.

As a later issue, this study has been discussed by $MAPE$, R^2, and $RMSE$, but has not been able to clarify the differences in characteristics between sites from a traffic engineering point of view. We understand that clarifying this is also an important issue to consider. In addition, because GNN handles all sites at once, the adjacent matrix calculation requires significant amount of memory when using large datasets. In this study, GNNs showed high accuracy in 2-3 h ahead-of-time predictions, and we believe that reducing the computational cost of GNNs will lead to more effective and efficient traffic management. Moreover, GNNs that take into account the relationship between data at different times of the day have been developed in recent years, and there is a possibility that they can provide accuracy in predicting 30 or 45 min in advance. Therefore, the study of improving the computational efficiency of the GNN and introducing relationships between data in time series is a subject for future work.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.， 'Optuna: A Next-generation Hyperparameter Optimization Framework', 2019, arXiv:1907.10902.

Fu, R., Zhang, Z., Li, L., 'Using LSTM and GRU Neural Network Methods for Traffic Flow Prediction', *Proc. 31st Youth Academic Annu.Conf. Chin. Assoc. Autom. (YAC)*, 2016, pp. 324-328.

Highways England, 'WebTRIS Traffic Flow API', https://webtris.highwaysengland.co.uk/api/swagger/ui/index. (Accessed 3rd February 2023)

Jiang, R., Wang, Z., Yong, J., Jeph, P., Chen, Q., Kobayashi, Y., Song, X., Fukushima, S., Suzumura, T., 'Spatio-Temporal Meta-Graph Learning for Traffic Forecasting', 2022, arXiv:2211.14701v3.

Miyazaki, T.， Oosawa, A., Kikucih, Y., Sugawara, H., 'Open traffic data of England and short-term traffic congestion prediction with AutoML', *Intelligence, Informatics and Infrastructure,* Vol. 3, No. J2, 2022, pp. 268-276.

Miyazaki, T. , Ogata, R. , Y., Murano, Kikuchi, Y., Sugawara, H., 'AI short-term speed prediction on an arterial road of england', *Intelligence, Informatics and Infrastructure,* 2023. (in press)

Ogata, R., Miyazaki, T. , Kikuchi, Y., Murano, Y., Sugawara, H., 'Comparison of Gradient Boosting Decision Tree and Graph Neural Networks for Short-Time Speed Prediction', *Intelligence, Informatics and Infrastructure,* 2023. (in press)

Perktold, J., Seabold, S., Taylor, J., statsmodels-developers, 'statsmodels', https://www.statsmodels.org/stable/index.html. (Accessed 3rd February 2023)

Yu, B., Yin, H., Zhu, Z., 'Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting', 2018, arXiv:1709.04875v4.

Wu, Z., Pan, S., Long, G., Jiang, J., Zhang, C., 'Graph WaveNet for Deep Spatial-Temporal Graph Modeling', 2019, arXiv:1906.00121v1.

Zhao, Z, Chen, W., Wu, X., Chen, P. C. Y. and Liu, J. 'LSTM network: A deep learning approach for short-term traffic forecast', *IET Intelligent Transport Systems*, Vol. 11, No. 2, 2017, pp. 68-75.