A traffic data imputing method based on multisource recurrent neural network

Zhongchang Ji[#], Wenxing Zhu*
School of Control Science and Engineering, Shandong University, Jinan, China

ABSTRACT

This paper proposes a method to fill in the missing traffic data by using multi-source data. Due to the regularity and specificity of traffic data, Gru network is used to capture missing patterns. The processed missing data, mask data and time interval data are input into Gru network for more in-depth information capture. The results of road speed matching for the floating vehicle data on the road in the corresponding period are further studied by Gru network, and the two results are fused to obtain the filling value of missing value.

Keywords: Traffic data, imputation, RNN-network, machine learning

1. INTRODUCTION

With the acceleration of economic development and infrastructure construction, the number of motor vehicles is increasing and the congestion phenomenon is becoming more and more serious. Therefore, the demand of intelligent traffic system research is also increasing. The development of cloud computing, big data and other technologies has promoted the rapid development of various transportation technologies, but also profoundly affected the decision-making of management departments. All kinds of traffic data contain rich traffic information and serve as the basis for various research and decision-making. However, in the real world, due to weather, communication network, equipment failure and so on, there are always missing traffic data collected by various sensors. Then, most of our traffic systems consider complete traffic data and extract important traffic information from it. Therefore, the problem of missing data is urgent.

In this paper, a traffic data filling algorithm based on multi-source data is proposed, which makes use of the feature collection ability of deep learning network, which is different from the traditional single data source filling algorithm and effectively improves the accuracy and effect of data filling.

Over the years, domestic and foreign scholars at have proposed many methods to solve data completion problem. There are some commonly used basic statistical methods in data-driven intelligent transportation system. Data is divided into numerical data and non-numerical data. When missing value is numerical data that can be added or subtracted, the average value is used to fill in the blank. And when the data is non-numerical, the mode that is used to fill in the missing spot. If the dataset conforms to a standard distribution rule, median interpolation can be used. It is a common method to deal with missing data by calculating the mean value, but this method is only suitable for relatively stable data. Reference[1] interpolated the missing data in the video detector by using the historical data and the average value of adjacent period and adjacent detector. Ku et al.[2] took advantage of the time-space correlation existing between the intersecting flow of multiple phase intersecting connected road segments, and adopted the k-mean clustering technique for the clustering of road segments with similar intersecting flow modes. The time space correlation is extracted and used for the interpolation of missing data points. Based on the study of common missing data repair methods, reference[3] proposed using autoregressive integrated moving average (ARIMA) to predict missing data, and achieved good results; Reference[4] proposed a temporal information enhancing long short term memory (t-LSTM) network to predict the traffic flow of a single road, and then applied it to the repair of abnormal data. This technology can well restore the characteristics of the original data and improve the repair accuracy. Reference[5] uses time label to enhance the input effect and improve the model performance effectively. Rodrigues et al.[6] adopted the multi-output Gaussian process (GP) for data imputation, which can model the complex spatial and temporal patterns in traffic data. Li et al.[7] proposed a hybrid spatiotemporal method to restore the missing values, which captures the temporal patterns and spatial residuals information by the prophet model and iterative random forest model. Wang et al. (2018)[8] utilized low-rank matrix

#201914390@mail.sdu.edu.cn; *zhuwenxing@sdu.edu.cn

International Conference on Computer Application and Information Security (ICCAIS 2021),
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12260, 122600C · © The Authors.
Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2637835

factorization to reconstruct missing traffic data; this process incorporates constraints on temporal evolution and spatial similarity. Chen et al. (2018)[9] proposed a tensor completion framework based on Tucker decomposition to accomplish the recovery task by discovering the spatiotemporal patterns and underlying structure from incomplete data. Zhang et al. (2020)[10] developed a geometric matrix completion model to estimate network-wide traffic flow; this approach integrates the traffic flow records and traffic speed information.

Feasible solutions to abnormal and missing data have been proposed in the above literature in different fields. Based on the spatio-temporal relationship between different types of traffic data, this paper proposes a missing data filling algorithm based on additional data, and takes into account the relevance of data before and after, using the self-learning ability of recurrent neural network for data recovery.

2. PRELIMINARIES

2.1 Pre-process

Data Missing Question is a commonly happened situation. Many scholars have researches on this subject. In this article, data-missing is about traffic data missing. There is a rather often pattern. Data is normally missing in missing in various position Due to weather, network, equipment, etc. Complete data can often be used directly, and missing data can be filled up with other information. For missing data, some mathematical processing is needed to represent missing information in the data.

Firstly, a multivariate time series is defined as $\mathbf{X}=\{\mathtt{x}_1,\mathtt{x}_2,\cdots,\mathtt{x}_T\}^T$,where R is the set of real numbers, $\mathtt{X}_t$ represents the data at the moment t， and Trepresents the length of time window. I Represents the length of all variables of a time slice data, i.e., $\mathtt{x}_t=\{\mathtt{x}_t^1,\mathtt{x}_t^2,\cdots,\ \mathtt{x}_t^I\}$, where x_t^i represents the value of the first variable at the time. To represent missing data, a corresponding $\mathtt{m}_t$ is defined for each variable $\mathtt{x}_t$, where the element is m_t^i, and its value is:

$$m_t^i=\begin{cases}1, & if\ x_t^i \text{ is not observed}\\ 0, & \text{if } x_t^i \text{ is observed}\end{cases} \tag{1}$$

In addition, we need to describe more missing information, so the time interval of missing data is defined as the time interval between the data and the data observed last time, namely

$$\delta_t^i=\begin{cases}l_t-l_{t-1}+\delta_{t-1}^i & \text{if t}>1, m_{t-1}^i=0\\ l_t-l_{t-1} & \text{if t}>1, m_{t-1}^i=1\\ 0 & \text{if t}=1\end{cases} \tag{2}$$

All the information obtained above is about the missing data that we processed in advance.

2.2 GRU

GRU Neural Network is a kind of Recurrent Neural Network (RNN). RNN is proposed to capture the input information before and after, which is essentially a memory function. The purpose of GRU is the same as that of LSTM, which is to solve the gradient problems of RNN in long-term memory and back propagation. GRU, on the other hand, has fewer parameters and is relatively easier to train, which can improve the training efficiency to a large extent.

The GRU network unit structure is shown in Figure 1. The forward propagation formula of GRU is as follows

$$r_t=\sigma(W_r\cdot[h_{t-1},x_t]) \tag{3}$$

$$z_t=\sigma(W_z\cdot[h_{t-1},x_t]) \tag{4}$$

$$\tilde{h}_t=tanh(W_{\tilde{h}}\cdot[r_t*h_{t-1},x_t]) \tag{5}$$

$$h_t=(1-z_t)*h_{t-1}+z_t*\tilde{h}_t \tag{6}$$

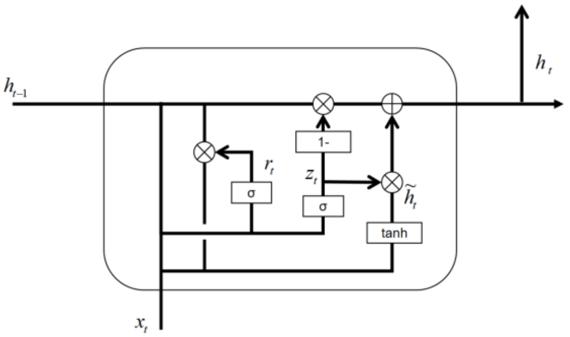

Figure 1. GRU structure.

Among them, the first three parameters W_r , W_t, r_t , $W_{\tilde{h}}$ are spliced, and the specific parameters are as follows.

$$W_r = W_{rx} + W_{rh} \tag{7}$$

$$W_Z = W_{ZX} + W_{zh} \tag{8}$$

$$W_{\tilde{h}} = W_{\tilde{h}x} + W_{\tilde{h}h} \tag{9}$$

where x_t is the data at the current moment, r is the reset gate, z is the forget gate, σ is the sigmoid function, and $*$ is the vector multiplied by elements. Is a candidate hidden layer that uses r_t to control how much of the previous memory needs to be retained. z_t controls how much information is forgotten from $\tilde{h}_t$ in the hidden layer of the previous moment, how much information of the hidden layer of the current moment $\tilde{h}_t$ needs to be added, and finally gets the output hidden layer information. The final output is

$$y_t = \sigma(W_o \cdot h_t) \tag{10}$$

3. MULTISOURCE TRAFFIC DATA IMPUTATION MODEL

Due to the inherent complexity of traffic flow, different traffic numbers provide different information, so different data processing modules are needed to solve the problem of obtaining its inherent information. We use RNN network to capture the data pattern of multi-source traffic data. Figure 2 shows the whole structure of the proposed model.

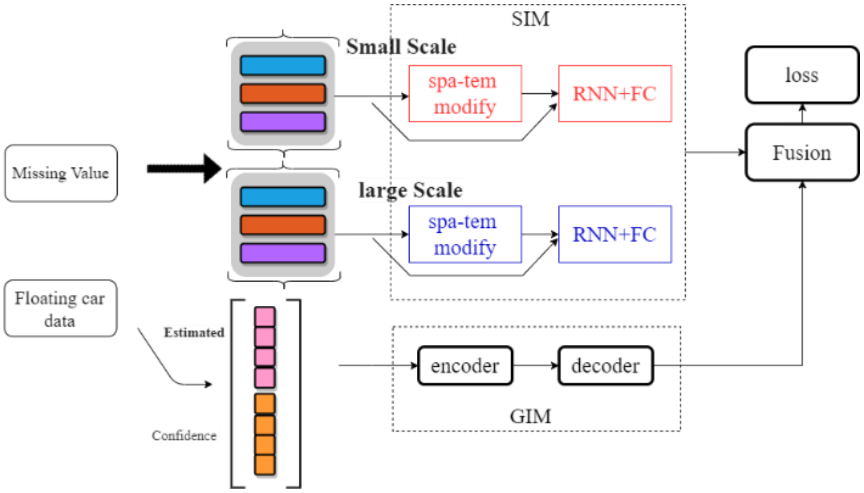

Figure 2. Multisource process model structure.

3.1 Check point data processing

GRU can process timing data, but for missing data, we need to process it before network processing, which cannot be entered into the network directly.

For the input at time T, we all have the network state at time $T-1$ and the previously preprocessed data $\{x, m, \delta\}$. h_{t-1}, mask data M and time interval data δ are used to further preprocess the missing data.

For missing data, because the data of different bayonets at the same time have spatial correlation, the data at the current time can be used to process the current data.

$$x_s = W_s x_t + b_s \tag{11}$$

The values of W_s on the diagonal are all 0, so x_s are the predicted values of data obtained from other dimensions. Correction was made using attenuation value λ_x associated with δ.

$$x_h = relu(W_h h_{t-1} + b_h) \tag{12}$$

$$x_h = relu(W_h h_{t-1} + b_h) \tag{13}$$

$$\lambda_x = 1 + (max(0, W_{\lambda x} h + b_{\lambda x})^2 + 1) \tag{14}$$

The mask vector M is used to judge whether the data is missing, and all the complete data are retained.

$$\tilde{x}_t = m_t x_t + (1-m)\hat{x}_t \tag{15}$$

h_{t-1} is also adjusted for recession

$$\tilde{h}_{t-1} = h_{t-1}\lambda_h \tag{16}$$

$$\lambda_h = exp\{-max(0, W_{\lambda h} h + b_{\lambda h}\} \tag{17}$$

After processing the data with m and δ, the model can better capture the missing information of data.

3.2 GPS data processing

GPS data includes latitude and longitude, UTC time, azimuth, current speed, and current vehicle status. According to the latitude, longitude and time of the data, the data is mapped to the corresponding section, and the RNN network is used to extract the information of the traffic flow at a certain time.

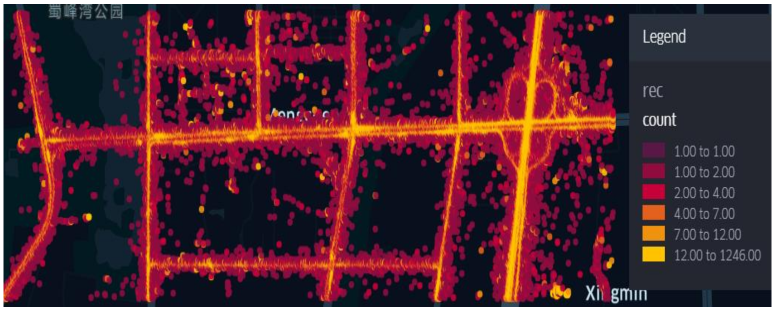

Figure 3. Heatmap of GPS dataset.

Figure 3 displays the amount of GPS dataset. It is showed that there are enough data to indicate the state of traffic flow. The road speed can be obtained by weighted processing of GPS data mapping, and RNN network is used for further feature extraction. Due to weather, equipment, network and other factors, GPS data may be biased, so it is necessary to deal with this special situation in advance. We need to map these to road and use a common direct estimation method to compute the road speed with confidence rate.

It is weighted with confidence according to the information source of the road section.

$$\delta_g = 1 - exp(c) \tag{18}$$

where c represents the number of GPS data mapped to the current section, and the more the number, the higher the confidence. The purpose of fusion module is to fuse GPS data with data features extracted from bayonet data. Data features extracted from different data are fused. Because the nodes of the data are spatially related and identical in time, it is possible to fuse tensors in the temporal dimension.

$$o = [o^m_{:jk}, o^g_{:jk}] \tag{19}$$

o^m is the feature extracted from main data, and o^g is the feature obtained from GPS data. The final results are output through the full connector layer.

4. EXPERIMENT

Here, we use Mae and RMSE as evaluation indexes. Then, the MAE is defined as:

$$MAE = \left(\frac{1}{N}\sum|y - \hat{y}|\right) \tag{20}$$

where y is the real value, $\hat{y}$ is the estimated value, and N is the total number of missing values.

This experiment uses microwave data and GPS data related to Huangke intersection in Hefei demonstration area. The data was collected from October 11 to 15, 2016 with an interval of 1 minute. Microwave data include road section number, acquisition time, time occupancy, speed, lane, etc. GPS data includes vehicle number, reporting time, latitude and longitude, speed, direction and other field information. By setting different miss rates, the completion effect of the algorithm is detected. We compare the experimental results of different algorithms under different missing rates (Table 1).

SVD method is singular value decomposition method, KNN and Haltrc is traditional methods to impute missing data. It can be seen that the performance of KNN is more than 50%, the effect is poor, and it is greatly affected by the missing items. The performance of Halrtc is acceptable, second only to the model proposed in this paper. In the case of different deletion rates, our model performed best, especially its performance was not sensitive to the change of deletion rate.

Table 1. MAE of the imputation under different missing rate.

Model	**Missing Rate**			
	25%	40%	55%	70%
SVD	11.27	12.43	14.63	15.47
KNN	7.20	7.6	10.48	10.64
Halrtc	7.34	7.79	8.33	8.74
Ours	6.15	6.29	6.79	7.12

5. CONCLUSION

This paper proposes a multi-source data filling algorithm based on improved GRU network to solve the problem of missing traffic data filling. Various kinds of data are used effectively, and the hidden information of missing data is considered to better obtain the information contained in the data. It improves the robustness of the algorithm and performs well in a variety of missing conditions.

ACKNOWLEDGEMENTS

This work is funded in part by the National Natural Science Foundation of China (Grant No.61773243), Major Technology Innovation Projects of Shandong Province (Grant No. 2019TSLH0203) and the National Key Research and Development Program of China (Grant No. 2020YFB1600501).

REFERENCES

[1] Nuovo, A. G. D., "Data analysis with fuzzy C-means: A study of its application in a psychological scenario," Expert Systems with Applications, 38(6), 6793-6797 (2011).

[2] Ku, W. C., Jagadeesh, G. R., Prakash, A., et al., "A clustering-based approach for data-driven imputation of missing traffic data," 2016 IEEE Forum on Integrated and Sustainable Transportation Systems, 1-6 (2016).

[3] Newsham, G. R. and Birt, B. J., Building-level occupancy data to improve Arima-based electricity use forecasts *Proc. of the 2nd ACM Work. on Embedded Sensing Systems for Energy-Efficiency in Building*, 13-18 (2010).

[4] Hamza, T., Amer, A. S. and Noor, A. E. R., "Dynamic L-RNN recovery of missing data in IoMT applications," Future Generation Computer Systems, (89), 575-583 (2018).

[5] Mou, L., Zhao, P., Xie, H., et al., "T-LSTM: A long short-term memory neural network enhanced by temporal information for traffic flow prediction," IEEE Access, (7), 98053-98060 (2019).

[6] Rodrigues, F., Henrickson, K. and Pereira, F. C., "Multi-output Gaussian processes for crowdsourced traffic data imputation," IEEE Trans. Intell. Transp. Syst., 20(2), 594-603 (2019).

[7] Li, H., Li, M., Lin, X., He, F. and Wang, Y., "A spatiotemporal approach for traffic data imputation with complicated missing patterns," Transp. Res. Part C: Emerg. Technol., 119, 102730 (2020) https://doi.org/10.1016/j.trc.2020.102730

[8] Wang, Y., Zhang, Y., Piao, X., Liu, H. and Zhang, K., "Traffic data reconstruction via adaptive spatial-temporal correlations," IEEE Trans. Intell. Transp. Syst., 20(4), 1531-1543 (2018).

[9] Chen, X., He, Z. and Wang, J., "Spatial-temporal traffic speed patterns discovery and incomplete data recovery via Svd-combined tensor decomposition," Transp. Res. Part C: Emerg. Technol., 86, 59-77 (2018).

[10] Zhang, Z., Li, M., Lin, X. and Wang, Y., "Network-wide traffic flow estimation with insufficient volume detection and crowdsourcing data," Transp. Res. Part C: Emerg. Technol., 121, 102870 (2020).