http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (24): 8570-8575, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.8570.8575

Reasearch on Fitting Method for Defective-data-mending of Urban Traffic Flow Based on SARBF Neural Networks

[1]Chen Ning, [2]Wen Xiaoyue and [3]Qu Xiao
[1]School of Mechanical and Automotive Engineering, Zhejiang University of Science and Technology, Hangzhou, China
[2]Zhejiang Enjoyor Electronics Co. Ltd., Hangzhou, China
[3]School of Automation and Electrical Engineering, Zhejiang University of Science and Technology, Hangzhou, China

Abstract: It is inevitable that the defectiveness of urban traffic flow data always occur in collecting information due to the sensor's failure. In order to mend those defective data, a new fitting method based on SARBF neural networks for defective-data-mending of urban traffic flow is presented in this study. It is not only an approach to analyzing the traffic data based on spatial autocorrelation but also a method on mending the defective data based on the RBF neural network's fitting technique. Firstly, the effectiveness of the defective-data-mending for complete data is great improved by using the spatial autocorrelation according to the urban traffic grid. Secondly, not only the mending precision but also the limitation of regression analysis is developed because of using RBF neural network. The experiment was held to in Hangzhou city. It is shown by the experiment's results that the fitting method brought up in this study is quite practicable to mend the defective data of urban traffic flow.

Key words: Defective-data-mending, urban traffic flow, information collection, spatial autocorrelation, 5 RBF neural network

INTRODUCTION

Nowadays, the urban traffic flow in China is collected by the sensors buried at the node of traffic grid. However, defective data is always occurred due to the sensor failure of constructing, transmitting or processing, which is not benefit on data mining and analyzing deeply. So, the defective data need to be mended. The research on mending defective data at present includes the methods of cluster analysis (Chen *et al.*, 2012), principal component analysis (Zhang *et al.*, 2007), stepwise regression analysis (Zhou, 2006), EM (Han *et al.*, 2007) and DA (Sun *et al.*, 2013), predicting on grey system theory (Yang *et al.*, 2012) and artificial neural network (Zhang *et al.*, 2005) and so on.

On the methods of cluster analysis, principal component analysis, stepwise regression analysis, EM, DA, the defective data is mended by the relation between the historical data from the data-defective intersection and the complete traffic flow data in other intersections, without considering the spatial autocorrelation of the intersections in the traffic grid. On the methods of predicting on grey system theory and artificial neural network, the defective data is mended by predicting with the historical data from the data-defective intersection, without considering the data of other intersections= effect in the traffic grid, so the real-time influence of the traffic flow breakdown can=t be reflected on the defective data mending.

So, a new approach named SARBF neural network fitting is presented to mend the defective data, which combines analysis based on spatial autocorrelation and RBF neural network fitting method.

ALGORITHM DESIGN OF SARBF NEURAL NETWORKS FITTING METHOD

The subject of mending defective traffic flow data is the flow data from each traffic grid node. By analyzing the realistic flow data, its characteristics are known easily:

- Periodic in time, the traffic flow data of each intersection changed periodically. Figure 1 shows the variations of one week data at two intersections in Hangzhou City
- Correlative in space, the correlation of the traffic flow data in proximity intersections. Shown on Fig. 1, the change law of the two adjacent intersections is similar mostly

Corresponding Author: Chen Ning and Qu Xiao, Zhejiang University of Science and Technology, China

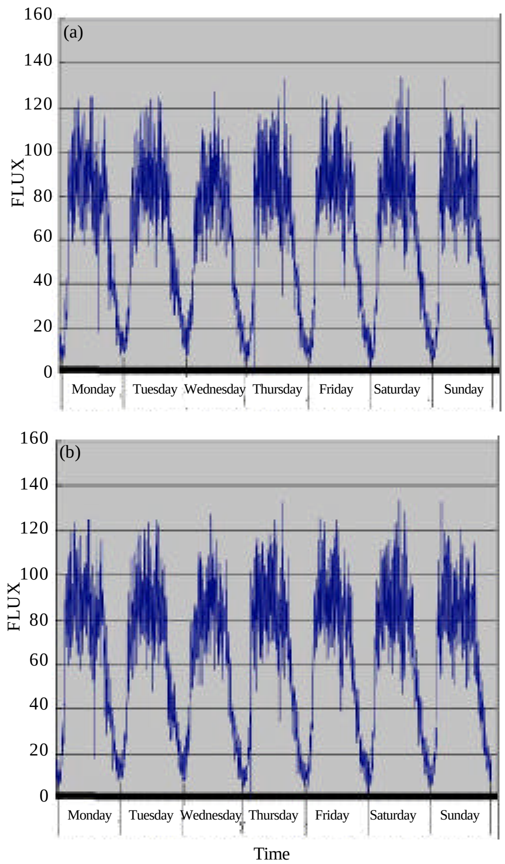

Fig. 1(a-b): Urban traffic flow data of two adjacent intersections, (a) Intersection A and (b) Intersection B

- On SARBF neural networks fitting method for mending defective traffic flow data, the defective data is mended by the complete data from other intersections. Before using neural network to mend the defective data, the data-complete intersections should be selected to improve the mending speed. As the characteristics shown in Fig. 1, the data-complete intersections can be selected by the spatial auto-correlation of the intersections in the traffic grid. Then, the defective data can be mended by the RBF neural network fitting on historical data of selected data-complete and data-defective intersections. The algorithm flow is shown in Fig. 2

Spatial auto-correlation analysis: The first step of the SARBF neural networks fitting method for mending defective traffic flow data is selecting the data-complete intersections, for which is the basis to study and establish a neural network model to solve the problem. The urban traffic grid is an organic whole with the intersections

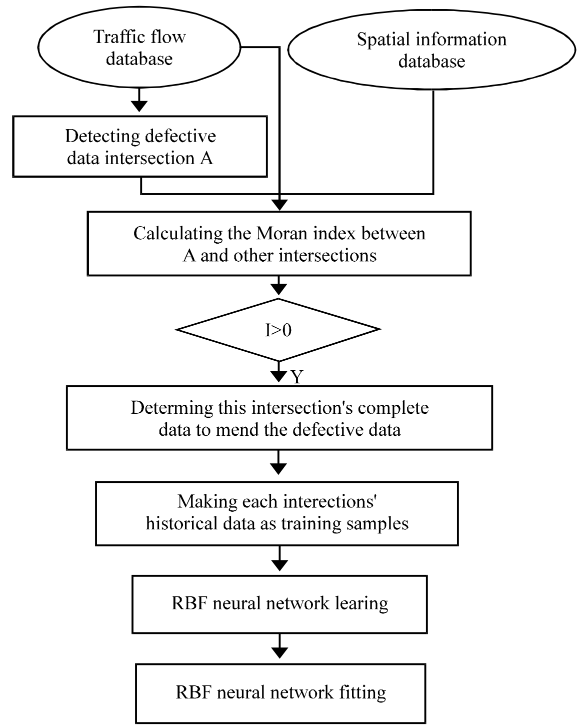

Fig. 2: SARBF neural networks fitting method

being connected by sections. So, the traffic flow of the most adjacent intersections is closely related, not completely identical, for the similarity about the regularity and style of resident's trip in one city. And the data-complete intersections which are strongly correlated to the data-defective intersection can be selected by Moran's I spatial autocorrelation analysis method, avoiding the redundancy of data.

The reflection of the spatial autocorrelation in urban traffic flow is the correlation degree of the traffic flow data from each intersection in the traffic grid. In spatial autocorrelation theory, the more close of two things, the more likeness. The procedure in judging by the Moran's I spatial autocorrelation as follows:

- Generating a data table of traffic grid node by the software of GIS, including the position relation of each node in the traffic grid. The intersections are coded from 1 to n
- Finding out the data-defective intersection (i) from the traffic flow database, then building a spatial adjacency matrix W_{ij}. If intersection j close to intersection i, $W_{ij} = 1$, else $W_{ij} = 0$. And $i \neq j$, $W_{ij} = 0$
- Calculating the decision index I for the spatial autocorrelation:

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij} \times (X_i - \overline{X})(X_j - \overline{X})}{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij} \times \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} \quad (1)$$

In which, W_{ij} is expressed as spatial adjacency matrix, X_i is the traffic flow data from intersection i and X_j is the traffic flow data from intersection j

- Determining the intersections with complete traffic data to mend the defective data from the I-value. The I-value calculated from the Moran's I formula is between ,S1 to 1. It is called positive correlation when I>0 and negative correlation when. The greater I-value, the higher spatial correlation is, as well as the lower I-value, the weaker spatial correlation is. And the I<0 spatial distribution would be random, when I-value tends to 0

RBF neural network fitting: The variation of the traffic flow is random, so the relation between each intersection is nonlinear, which can be founded easily and well by RBF neural network method.

The radical basis function neural network (RBFNN) is a feed-forward neural network model with the capacity of local approximation, so it can be used in mending the defective data in traffic flow. The weight from input layer to hidden layer is fixed to be 1, only the weight from hidden layer to output layer is adjustable. The response to input signal is occurred in local by the action function (radial basis function) of the hidden layer. And the great output would be produced when the input signal closed to the central range of the radial basis function.

The Gaussian kernel function is selected as the transform function in the hidden layer generally exerting its good properties as simplicity, radial symmetry and smoothness and accurate analyses, the form shown as:

$$G(\|x - x_c\|) = \exp\left\{-\frac{\|x - x_c\|^2}{2\sigma^2}\right\} \quad (2)$$

with which, x_c is the center of kernel function; ó is the width parameter of kernel function which control the radial action sphere of kernel function and could be determined by k-means clustering algorithm; ||·|| is norm(usually take Euclidean norm).

The RBF network of multi-input and 1 output is designed to mend the defective data. Meanwhile, the historical traffic data of the data-defective intersection and its correlated data-complete intersections are got to be training data set. The weight vector W of the hidden layer to output layer can be got by interpolation algorithm when the center of BRF is determined, because the input vector is mapped to the hidden layer.

Algorithm of time slice selection: The time slice can be acquired by analyzing the history data of every cross-road. The selection of threshold is very important. If the threshold becomes larger, the precision will be decreased. If it becomes smaller, then process time spending will increase greatly.

Considering two OD pairs of ($O_1 D_1$) and ($O_2 D_2$), their history matrixes of snapshot records are:

$$O_1 = \begin{pmatrix} x_{11}^{(1)} & t_{11}^{(1)} \\ \vdots & \vdots \\ x_{1n}^{(1)} & t_{1n}^{(1)} \end{pmatrix}, D_1 = \begin{pmatrix} x_{21}^{(1)} & t_{21}^{(1)} \\ \vdots & \vdots \\ x_{2n}^{(1)} & t_{2n}^{(1)} \end{pmatrix} \quad (3)$$

And:

$$O_2 = \begin{pmatrix} x_{11}^{(2)} & t_{11}^{(2)} \\ \vdots & \vdots \\ x_{1n}^{(2)} & t_{1n}^{(2)} \end{pmatrix}, D_2 = \begin{pmatrix} x_{21}^{(2)} & t_{21}^{(2)} \\ \vdots & \vdots \\ x_{2n}^{(2)} & t_{2n}^{(2)} \end{pmatrix} \quad (4)$$

In which, the threshold is p. The time slice algorithm is depicted as:

- **Step 1:** The redundant and invalid data can be removed by matching the pairs of keys and values in ($O_1 D_1$) and in ($O_2 D_2$), then new matrixes are obtained as:

$$\overline{O_1} = \begin{pmatrix} \overline{x_{11}}^{(1)} & \overline{t_{11}}^{(1)} \\ \vdots & \vdots \\ \overline{x_{1m}}^{(1)} & \overline{t_{1m}}^{(1)} \end{pmatrix}, \overline{D_1} = \begin{pmatrix} \overline{x_{21}}^{(1)} & \overline{t_{21}}^{(1)} \\ \vdots & \vdots \\ \overline{x_{2m}}^{(1)} & \overline{t_{2m}}^{(1)} \end{pmatrix} \quad (5)$$

And:

$$\overline{O_2} = \begin{pmatrix} \overline{x_{11}}^{(2)} & \overline{t_{11}}^{(2)} \\ \vdots & \vdots \\ \overline{x_{1h}}^{(2)} & \overline{t_{1h}}^{(2)} \end{pmatrix}, \overline{D_2} = \begin{pmatrix} \overline{x_{21}}^{(2)} & \overline{t_{21}}^{(2)} \\ \vdots & \vdots \\ \overline{x_{2h}}^{(2)} & \overline{t_{2h}}^{(2)} \end{pmatrix} \quad (6)$$

With:

(a): m n, h n

(b): $\overline{x_{1i}}^{(1)} = \overline{x_{2i}}^{(1)} \in \{x_{1i}^{(1)} \mid i = 1,...m\} \cap \{x_{2i}^{(1)} \mid i = 1,...m\}$, $(\overline{t_{1i}}^{(1)} < \overline{t_{2i}}^{(1)})$

(c): $\overline{x_{1i}}^{(2)} = \overline{x_{2i}}^{(2)} \in \{x_{1i}^{(2)} \mid i = 1,...h\} \cap \{x_{2i}^{(2)} \mid i = 1,...h\}$, $(\overline{t_{1i}}^{(2)} < \overline{t_{2i}}^{(2)})$

- **Step 2:** Assuming the threshold is p, m≤h, there are three conditions below:

- .If p>h, then the OD pair ($O_1 D_1$) and in ($O_2 D_2$) can't satisfy the threshold condition, the algorithm stop
- If $p \in (m, h]$,then only ($O_2 D_2$) satisfies the condition. So, p corresponding records can be chosen randomly from $\overline{O}_2$ and $\overline{D}_2$. The result is depicted as:

$$\overline{\overline{O_2}} = \begin{pmatrix} \overline{\overline{x}}_{11}^{(2)} & \overline{\overline{t}}_{11}^{(2)} \\ \vdots & \vdots \\ \overline{\overline{x}}_{1p}^{(2)} & \overline{\overline{t}}_{1p}^{(2)} \end{pmatrix}, \overline{\overline{D_2}} = \begin{pmatrix} \overline{\overline{x}}_{21}^{(2)} & \overline{\overline{t}}_{21}^{(2)} \\ \vdots & \vdots \\ \overline{\overline{x}}_{2p}^{(2)} & \overline{\overline{t}}_{2p}^{(2)} \end{pmatrix} \tag{7}$$

- If $p \geq m$, then the two pairs satisfy the condition. So, p corresponding records can be chosen randomly from O_1, $5D_1$, $5O_2$, and D_2. The result is depicted as:

$$\overline{\overline{O_1}} = \begin{pmatrix} \overline{\overline{x}}_{11}^{(1)} & \overline{\overline{t}}_{11}^{(1)} \\ \vdots & \vdots \\ \overline{\overline{x}}_{1p}^{(1)} & \overline{\overline{t}}_{1p}^{(1)} \end{pmatrix}, \overline{\overline{D_1}} = \begin{pmatrix} \overline{\overline{x}}_{21}^{(1)} & \overline{\overline{t}}_{21}^{(1)} \\ \vdots & \vdots \\ \overline{\overline{x}}_{2p}^{(1)} & \overline{\overline{t}}_{2p}^{(1)} \end{pmatrix} \tag{8}$$

And:

$$\overline{\overline{O_2}} = \begin{pmatrix} \overline{\overline{x}}_{11}^{(2)} & \overline{\overline{t}}_{11}^{(2)} \\ \vdots & \vdots \\ \overline{\overline{x}}_{1p}^{(2)} & \overline{\overline{t}}_{1p}^{(2)} \end{pmatrix}, \overline{\overline{D_2}} = \begin{pmatrix} \overline{\overline{x}}_{21}^{(2)} & \overline{\overline{t}}_{21}^{(2)} \\ \vdots & \vdots \\ \overline{\overline{x}}_{2p}^{(2)} & \overline{\overline{t}}_{2p}^{(2)} \end{pmatrix} \tag{9}$$

- **Step 3:** If $p \in (m, h]$, the time interval between the earliest time of $\overline{\overline{O_2}}$ and the latest time of $\overline{\overline{D_2}}$ can be obtained. For example, the earliest time of $\overline{\overline{O_2}}$ is:

$$t_{o_2} = \min(\overline{\overline{t}}_{1i}^{(2)})$$

and the latest time of $\overline{\overline{D_2}}$ is :

$$t_{D_2} = \max(\overline{\overline{t}}_{2i}^{(2)})(i=1,...,p)$$

the time interval is $t = t_{D2}\text{-}t_{o2}$. So, the tome slice is $\Delta t = t$. If $p \leq m$, then the two time interval between $\overline{\overline{O_2}}$ and $\overline{\overline{D_2}}$, $\overline{\overline{O_2}}$ and $\overline{\overline{D_2}}$ can be obtained. For example, the earliest time of $\overline{\overline{O_1}}$ is:

$$t_{o_1} = \min(\overline{\overline{t}}_{1i}^{(1)})$$

the latest time of $\overline{\overline{D_1}}$ is:

$$t_{D_1} = \max(\overline{\overline{t}}_{2i}^{(1)})$$

And the earliest time of $\overline{\overline{O_2}}$ is:

$$t_{o_2} = \min(\overline{\overline{t}}_{1i}^{(2)})$$

the latest time of $\overline{\overline{D_2}}$ is:

$$t_{D_2} = \max(\overline{\overline{t}}_{2i}^{(2)})(i=1,2,...,p)$$

the two time intervals are, respectively $t_1 = t_{D1}\text{-}t_{o1}$ and $t_2 = t_{D2}\text{-}t_{o2}$. To satisfy the threshold condition, $\Delta t = \max(t_1, t_2)$ is selected as the time slice.

EXPERIMENT IN HANGZHOU CITY

At present, the urban traffic flow in Hangzhou is collected by SCATS induced coils. The GIS application platform is generated by mapinfo, meanwhile, the traffic flow data is collected per 5 minutes. The defective traffic flow data is occurred for the failure of the induced coils, which could be mended by using the SABRF neural network fitting method.

Take the Wensan-Xueyuan east intersection (871) as example, in which the defective data was occurred. And the message of 871 in one day is shown on Table 1. The data in first row in Table 1 is the intersection number and "/" represents for defective data.

The intersections information table can be got by Mapinfo to construct the spatial weight matrix:

$$\begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 0 & 1 & 1 & 1 & 0 & \cdots \\ \cdots & 0 & 1 & \ddots & 1 & 0 & \cdots \\ \cdots & 0 & 1 & 1 & 1 & 0 & \cdots \\ \cdots & 0 & 0 & 0 & 0 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{10}$$

Then, determine the data-complete intersections to mend the defective data in Wensan-Xueyuan east intersection by calculating the Moran's I-value, such as Wen'er-Xueyuan intersection, Wensan-Jiaogong intersection and so on.

The training sample is the historical flux data which is selected from different intersections including the data-defective and data-complete intersections. The regularity of traffic flow is different in different periods by analyzing the historical data. So, in order to guarantee the precision of mending, the training sample should be gained from different periods to get different RBF neural networks. Meanwhile, distinguishing the time periods is important to the RBF neural networks mending model. The periods contain large periods and small periods. The large ones can be divided into several kinds, such as Spring Festival period (the Spring Festival holiday), Holiday timeperiod (The Labor Day, The National Day and New Year's Day), Weekends, Normal time (from Tuesday to Thursday), Unnormal time (especially Monday and Friday) and the small periods can be divided into morning peak, evening peak and other time.

Using RBF neural network fitting method to mend the defective data, the period when the defective data occurred should be judged firstly, then mend it according to the corresponding RBF neural network. For example, the defective data occurred in 871 intersection in December 1st in 2006 (Friday). Three different RBF neural networks should be established by the historical traffic data on Friday in different periods, which are morning peak, evening peak and other time separately.

MATLAB's control system toolbox function called NEWRB can be used to realize the function of mending the defective data by RBF neural network fitting method. In which, the flow vector of the data-complete intersection correlated to the data-defective intersection is taken as input vector, the flow vector of the data-defective intersection is taken as object vector, the GOAL of the mean square error take 0 and SPREAD take 1. The mean square error of network output would be decreasing continuously by increasing the radial nerve automatically and the training of the network end until the error attains to the GOAL.

After the neural network training to the historical data of the data-defective and data-complete intersections, the defective data can be mended by the simulation function:

Table 1: Part of traffic flow data in xihu district

Time	451	871	891	923	455	...
07:05	56	/	95	44	71	...
07:10	47	/	92	85	99	...
07:15	24	/	109	77	68	...
07:20	84	/	78	86	74	...
...	...	...	...	...	...	...

Table 2: Mending result, (2013.11.1)

		Mending result/cars		Relative error (%)	
Period	Real data/cars	RA	RBF	RA	RBF
07:00~07:05 (AM)	100	82	95	18	5
08:00~08:05 (AM)	168	192	159	14	5.4
08:50~08:55 (AM)	161	189	154	17.4	4.3
09:10~09:15 (AM)	193	219	183	13.5	5.2

$$y = \mathrm{sim}(net, P) \tag{11}$$

To verifying the mending precision of the RBF neural network fitting method, the mending results would be compared to the real data and the data from regression analysis(RA), the comparison result are shown in Table 2 and 3.

As shown in Table 2 and 3, the mending result's relative error by the RBF neural network fitting method is about 5%, while the result by regression analysis is above 10%. So, the mending precision of the RBF neural network fitting method is higher.

CONCLUSION

Using the SARBF neural networks fitting method to mend defective traffic flow data, the mending speed is improved by selecting part of the correlated data to be trained in neural network from spatial autocorrelation analysis. Meanwhile, the mending precision is also improved by RBF neural network method, avoiding the limitation of regression analysis. And the experiment to mend the defective traffic flow data in Hangzhou is shown that the mending precision is acceptable.

ACKNOWLEDGMENT

This study was supported in part by National Natural Science Foundation of China (Grant No. 61273240) and by Key Project of Chinese Ministry of Education (Grant No. 211070), as well as by Zhejiang Provincial Natural Science Foundation of China (Grant No. LY12F03024).

REFERENCES

Chen, X.Q., Z.X. Hou, Y.H. Wu and Z.W. Liu, 2012. A gray neural network model for traffic flow prediction at non-deterctor intersections. J. Syst. Simul., 24: 2655-2656.

Han, W.G., J.F. Wang, Y.G. Gao and J.J. Hu, 2007. Forecasting and analysis of regional traffic flow in space and time. J. Highway Transp. Res. Dev., 24: 92-96.

Table 3: Computation result of travelling time, (2013.11.1)

Samples	1	2	3	4	5	6	7	8	9	10	Extreme early/late time	Interval
Vehicle ID	A36B XX	A3T6 XX	A56 JXX	A56 UXX	B7E4 XX	AT79 XX	A70 XXX	AT78 XX	AT56 XX	B33I XX		
No. 1 roadcross	10:12:37	10:13:20	10:14:20	10:13:12	10:13:22	10:12:11	10:12:30	10:14:20	10:12:54	10:12:34	10:12:11	4min 13sec
No. 2 roadcross	10:14:39	10:16:07	10:16:20	10:16:03	10:16:24	10:15:55	10:15:55	10:16:10	10:14:59	10:15:51	10:16:24	

Yang, Z.S., H. Zhang and J. Li, 2012. Step-by-step regression used in traffic flow prediction of non-detector road intersections. J. Jilin Univ. (Eng. Technol. Edn.), 32: 16-19.

Zhang, H., W. Wang and H.Z. Gu, 2005. Application of cluster analysis and stepwise regression in predicting the traffic volume of lanes. J. Southeast Univ. (English Edn.), 28: 359-362.

Zhang, J., E.T. Yao, G.F. Ni and J. Ji, 2007. Application of EMD-RBF and image processing to weighing-in-motion of vehicles. J. Mech. Sci. Technol. Aerospace Eng., 26: 701-704.

Zhou, K., 2006. A new isolated point detecting algorithm based on statistical clustering RBF neural network. J. Comput. Sci., 33: 196-197.