A Estimation Model of Missing Value Based on Wireless Sensor Network

Cai ZHANG¹, Yi ZHUANG¹,* and Jing LI¹

¹College of Computer Science and Technology
Nanjing University of Aeronautics and Astronautics
Nanjing, 210016, China
*Corresponding author

Keywords: Time series, Space correlation, ARIMA, Wavelet decomposition, Neural network.

Abstract. Wireless sensor network node affected by its own resources and environmental factors, exists the problem of the missing acquisition data. In view of the missing of WSN data, this paper uses ARIMA model and wavelet decomposition to forecasting the missing value, obtain the estimation of missing data in the time series. The estimation model based on time series named TS model. The adjacent node of missing data acquisition nodes as a reference, through the advantage of neural network model in non-linear relationship, se model is proposed. In this paper, from TC model and ST model, the time series are combined with the space correlation, STC model is proposed. Experiment shows STC model has a good effect on the estimation of missing data.

Introduction

Due to the small size of the Wireless Sensor Network nodes, the communication capacity, battery capacity, storage capacity and other hardware resources of WSN nodes are limited[1,2]. The components of the Wireless sensor network nodes are also relatively fragile. If WSN works in complex and volatile environment, the nodes are susceptible to environmental factors, their own factors and human factors, cause the breaking of communication link. Especially in the process of the high frequency acquisition and transmission, it can make the acquired data of WSN loss or abnormal. The usual method of processing loss and exception data is to throw away the data set about them. It will lose a lot of raw data, valuable information related to the missing data also tend to be ignored. It could be considerable effects on the data analysis results, sometimes will not only reduce the accuracy and reliability of data analysis, even directly affect the corresponding decision. The sensor nodes collect data sets except the loss and abnormal data are reliable. If filling missing data according to the reliable data for reasonable, we can improve the accuracy and reliability of the data analysis results. In wireless sensor network, data loss is inevitable, so it is necessary to fill in the missing data. The method of filling missing data is highly valued, and how to estimate the missing data is an urgent problem to be solved.

Related Research Work

There are a lot of researches for missing data of WSN at home and abroad. Literature [3] researches data missing problem based on data streams, WARA(Window Association Rule Mining)algorithm is proposed. When the data stream of a WSN node exist missing value, the algorithm first finds a node associated with it, and use the value of the associated node fill the missing value. In WSN, there may be a strong correlation between the two nodes, but there may be a linear function between the two nodes, the value of the sample is not necessarily approximate, so the accuracy of this algorithm is not high. Literature [4] proposes CARM(Closed Frequent Item sets Association Rule Mining)algorithm which is improved based on the algorithm. The algorithm is used to calculate the association rules of the data, find the frequent patterns of multiple data sources, and estimate the missing values by using frequent patterns. If there are no data tuple corresponding missing values in the frequent pattern, the algorithm will fail. So the WARA algorithm and CARM algorithm have great limitations, and can not
be widely used in the estimation of missing values. Literature [5] consider the characteristic of time space correlation, using linear interpolation algorithm for missing values in time domain, and using regression analysis method for missing values in space domain. In order to overcome the defects of linear interpolation in processing of missing data, this paper use wavelet decomposition to resolve the original time series, and use ARIMA model to estimate the missing data. For the correlation in the space, this paper use neural network to find the nonlinear relationships between sensors. Based on time series and space correlation, this paper proposes a time-space correlation model called TSC model.

**TSC Model**

**Problem Define**

The problem in this paper is how to accurately estimate the missing values in the WSN data set. First we give the definition of the problem.

Define; The perception data collected by a sensor node $N_i$ can be viewed as a time series $S_i=(<y_{i1}, T_1>, <y_{i2}, T_2>, ..., <y_{in}, T_n>)$. $y_{ij}$ is the value of the node $N_i$ in time $T_j$, Among them $j=\{1,2,...,n\}$. For any moment $T_v$, if the time of the observation value $y_{uv}$ missing, the problem of estimating the value of $\bar{y}_{uv}$ and make the $|\bar{y}_{uv} - y_{uv}|$ minimum is called the problem of missing value estimation.

**TS Model**

In many applications, the WSN measurement data is a continuous change of the physical quantity, so its monitoring value is usually a certain time correlation. If the value of $y_{uv}$ is missing, the TS model is used to estimate the missing value. TS model as shown in Fig.1 below. $S_{uv}$ represents the time series data of missing $y_{uv}$ data. $S_{uv}$ wavelet decomposition and reconstruction the two sequences $S_{uv}^h$ and $S_{uv}^l$, $S_{uv}^l$ is the shape of the original sequence, $S_{uv}^h$ is the details of the original sequence. Then the two sequences are estimated using ARIMA model, two sequences $\bar{S}_{uv}^h$ and $\bar{S}_{uv}^l$ were obtained after estimation. The two sequences were synthesized and the time series $\bar{S}_{uv}$ were obtained. $\bar{y}_{uv}$ in the sequence is estimated as the missing value.

![Figure 1. TS model.](image)

Wavelet decomposition and reconstruction is essentially through different band-pass filter will contain comprehensive information of a set of the original sequence is decomposed into multiple sets of different characteristics of time series. A set of signals reflect the intrinsic change of the original time series, that is, the approximation signal. Another set of signals reflects the impact of random disturbance, that is, the details of the signal. Prediction for different characteristics of the signal.

Wavelet decomposition can be achieved by means of Formula (1) Mallat algorithm.

In the formula, H and G are Low-pass filter and high-pass filter.

\[
\begin{align*}
C &= HS_{uv} \\
D &= GS_{uv}
\end{align*}
\]  

(1)

After wavelet decomposition, the sample quantities of the details sequence and the approximation sequence are reduced. Reducing the number of samples will affect the final result. So use the Mallat algorithm to decompose the sequence of each group should be reconstructed back to the original scale...
to increase the number of signals, the reconstruction algorithm described as shown below in Formula (2). $H^*$ and $G^*$ are dual operators of $G$ and $H$ respectively.

\[
\begin{aligned}
S_{uv}^p &= H^*C \\
S_{uv}^q &= G^*D
\end{aligned}
\] (2)

Random time series can be divided into stationary random sequence and non-stationary random sequence. If the average value of any time slice of the data sequence is constant, it is called stationary random sequence, otherwise it is non-stationary random sequence. Through the analysis, the WSN data is a non-stationary time series, which is suitable for the ARIMA model[6,7]. After self correlation and partial correlation function analysis of the monitoring data, can reflect the characteristics of WSN monitoring data by ARIMA model, the ARIMA model is shown in follows Formula (3).

\[
\phi(B)\nabla^d S_{uv} = \omega(B)\epsilon_i
\] (3)

In the formula, $\phi(B) = 1 - \phi_1B - \phi_2B^2 - \cdots - \phi_pB^p$, $\omega(B) = 1 - \omega_1B - \omega_2B^2 - \cdots - \omega_qB^q$, $S_{uv}$ is the time series with missing data, $B$ is time backward shift operator, $\nabla = 1 - B$ is difference operator, $p,q,d$ is Model order, $\phi_1,\phi_2,\ldots,\phi_p$ and $\omega_1,\omega_2,\ldots,\omega_q$ are Model parameters; $\epsilon$ is White noise with a certain variance of the mean value.

By the modeling process, ARIMA model is a process model with pre-built time-varying parameter, order and parameter of the model must use the latest data to identify and estimate. Thus realization of missing data estimate.

**SC Model**

It is well known that many sensor nodes are usually arranged in a specific monitoring area. The perception data of these nodes have spatial correlation, that is, the monitoring data collected by the sensor nodes in the physical location tend to be similar or there is a functional relationship. In SC model, in order to be able to estimate the missing values accurately $y_{uv}$, reduce the random error which is introduced by a single neighbor node to estimate the missing values, the node $N_i$ and its neighbors as a whole, using all of its neighbor nodes to estimate the missing values of the data. That is, for any time $v$, $y_{uv} = f(y_{1v}, y_{2v}, \ldots, y_{mv})$. So according to the correlation of neighbor nodes, $y_{uv}$ and other nodes of the perception of the data has a nonlinear relationship.

In order to obtain the complex nonlinear relationship, the SC model use neural network to describe the relationship between the Ni and its neighbor nodes. The data of the other correlative nodes are used as the input of the neural network, the data of the missing data node are the output data. Neural network are trained using the missing data near the time of the missing data, in order to reflect the relationship in the near time between each node.

**TSC Model**

In this paper, TSC model based on time-space is proposed to estimate the missing data with two models: the time series model and the spatial correlation model. The structure is as follows.
Figure 2. TSC model.

From Fig. 2, compared to the SC model, the TSC model adds a estimate of time series as input to the neural network, that is, the sequence $S_{uv}$ of missing data $y_{uv}$ calculated by TS model, obtain the time series $\hat{S}_{uv}$ including the estimate value $\hat{y}_{uv}$. There are three layers, m input nodes and J hidden layer nodes in neural network of the model. The weight between the $i$-th node of input layer and the $j$-th node of hidden layer is $w_{ij}$. The weight between the $j$-th node of hidden layer and output layer is $c_j$. The output is $\hat{S}_{uv}$ estimated by the synthesis of two models.

**Experiments**

In order to compare the accuracy of the TSC model proposed in this paper and other models, the experimental data of the dynamic test data of large deformation flexible body fabric are used. In this experiment, four WSN node data are used to estimate the data of one node marked missing data, compare the estimation results of three models. MSE is used to show the estimation error of each node, and the MSE is shown in the following Formula (4), where the mean is expressed as an average of the residuals between the true values and the estimated values.

$$MSE = \sqrt{mean[(y_{uv} - \hat{y}_{uv})]}$$  

(4)

Record 20 experiments, the experiment results are shown in the above Fig. 3. The TC model and the SC model are almost the same in the estimation precision and time. Compared to the TC model and the SC model, the STC model has a little improvement in operating time, but it can effectively reduce the error of the missing data estimation and improve the accuracy of the missing data estimation.
Figure 3. Experiment results.

Conclusions

The inherent characteristics of WSN make the problem of the missing data is unavoidable, and it also brings great difficulties to the various applications of WSN. The best way to solve the problem is to estimate the missing data. In this paper, a TS model is proposed, which is based on the time dependence of data. The model can estimate the missing data in a short time. We also propose a missing value estimation model named SC model based on the spatial correlation based for the non-stationary change data. Based on the two models, proposes a TSC model based on time and space is proposed, which can achieve better estimations of stable and non-stationary change. In the large deformation flexible body monitoring data set, the model is tested. The experiment results show that the TSC model has good accuracy and stability.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (General Program) under Grant No. 61572253, the National Natural Science Foundation for Youth of China under Grant No. 61202351 and the National Postdoctoral Fund under Grant No. 2011M500124.

References


