Wireless Sensor Missing Value Estimation Algorithm Based On Multi-Attribute

Xingliang Zhang¹, Tao Fang¹, Chunlei Yang¹, Zhengzheng Huang¹, and Xiaodie Zhang¹
¹College of Information and Engineering, Hebei GEO University, Shijiazhuang 050000, China

Abstract. Because the wireless sensor is arranged in the environment of unmanned management and complex, in the process of collecting data and transmitting data, it often leads to data loss due to the influence of itself or external environment. The best way to reduce the impact of missing value is to estimate the missing value. In this paper, we propose a missing value estimation algorithm based on time attribute and trust mechanism. We use Arima to predict the time attribute, and use the relationship between current value, historical value and error to predict the future data. We use subjective logic to convert the interaction information between nodes into trust value, and then Linear Regression prediction value by selecting the number of trust nodes. Finally, according to the optimal fit degree, weight distribution is carried out to form the final prediction value. Because the algorithm not only considers the node data of the trusted neighbor, but also predicts the future data changes through the changes of its own historical data, it has higher accuracy and lower error when compared with other algorithms.

1 Introduction

Wireless sensor network is a self-organizing network composed of sensor nodes with communication and computing capabilities. Due to the low cost and flexible deployment of nodes, it has been widely used in geological detection, data acquisition, military reconnaissance and many other related fields. Wireless sensor network is often deployed in an unmanaged or harsh environment. It may be interfered with other signals or be limited by its own power, which often lead to data loss in the process of transmission. Therefore, the integrity and reliability of data should be considered when deploying WSN. Therefore, the integrity and reliability of data becomes an issue that needs to be considered when deploying wireless sensor network.

In recent years, some experts and scholars have applied the missing value of wireless sensor network to the related fields such as temperature and humidity data collection, road leveling state data collection and so on. Zheng Li¹ et al. used Bayesian vector autoregression method to impute the missing data of irregular traffic collision data, and compared with other algorithms, it has better performance of imputation phase. Liu [2] proved that the convolutional neural network is not only applied to image processing but also has a good effect on missing data complement. Zhang Wangjuan [3] used Convolutional Neural Network to predict the missing values of time and space dimensions respectively, and compared the multi-dimensional and single dimensional prediction values through
experiments. Xu Ke et al. used BP neural network to make up the missing values of temperature and humidity, and then fused the data by batch estimation and weighted average. Wu Jiawen and others proposed to use K-Means and Fuzzy Neural Network to reconstruct the missing data after clustering.

Based on the interaction information between wireless sensor nodes, this paper predicts the missing value by calculating the direct trust value and indirect trust value of the surrounding nodes. Firstly, considering the existence of interaction information between each wireless sensor node, through the discount operator and fusion operator in the subjective logic, we can transfer and fuse the interaction information of directly or indirectly adjacent nodes. Then we can find the most trusted node and take its data value as the prediction value of the missing node.

2 Supplementary value of trust attribute

The data changes in this wireless sensor network are not isolated, but change regularly with time. When there is a missing value in a node, the node with missing data can effectively estimate the current missing value through the previous time change rule. In this paper, we use the ARIMA (Autoregressive Integrated Moving Average Model) to estimate the missing value of the existing time attribute. ARIMA model is a time series prediction method, which uses its own historical data to connect the next data for prediction. ARIMA model consists of two parts, one is AR (Auto Regressive) and the other is MA (moving average).

2.1 Data preprocessing

When using ARIMA to predict data, it is necessary to ensure that the data is stable, because unstable data cannot capture the rules, and thus cannot predict. Therefore, before using ARIMA model, data preprocessing should be carried out to ensure that the data is stable. If there is non-stationary state in the data, the data will be processed for stationarity first. In this paper, the difference method will be used for stationarity processing. The first order difference is expressed as:

\[ y_t = Y_t - Y_{t-1} \]  

The second order difference is expressed as:

\[ y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \]
\[ = Y_t - 2Y_{t-1} + Y_{t-2} \]  

where \( y_t \) represents the data after the difference, \( t \) represents the data at the time, \( y_{t-1} \) represents the data at the time of \( t-1 \), and \( y_{t-2} \) represents the data at the time of \( t-2 \).

2.2 ARIMA model

2.2.1 AR model

AR describes the relationship between current data and historical data, and forecasts its own data through the historical data of its own variables. When AR forecasts its own data, it needs to ensure that the data is stable, so as to find out its own rules, and then carry out effective prediction. The expression of p-order autoregressive model is:

\[ y_t = \mu + \sum_{i=1}^{p} r_i y_{t-i} + \epsilon_t \]
Among them, \( y_t \) is the predicted value of the current value, \( u \) is the constant term, \( p \) is the order, \( y_{t-i} \) is the data of the previous \( i \) period, \( r_i \) is the correlation coefficient, and \( \varepsilon \) is the error.

### 2.2.2 MA Model

Ma is that the current predicted value has nothing to do with the measured value of previous periods, but has a certain relationship with the previous error value. MA is to estimate the current data value through the linear combination of prediction errors in previous time periods. The expression of q-order moving average model is:

\[
y_t = \mu + \sum_{i=1}^{q} \varphi_i \varepsilon_{t-i} + \varepsilon_t
\]

Among them, \( y_t \) is the predicted value of the current value, \( u \) is the constant term, \( q \) is the order, \( \varepsilon_{t-i} \) is the error of the previous time period, \( \varphi_i \) is the correlation coefficient, \( \varepsilon_t \) is the white noise with the mean value of 0 and the variance of \( \sigma^2 \). MA model can eliminate the random fluctuation in the prediction data.

### 2.2.3 ARIMA model

ARMA combines AR model with MA model, Model expression of ARMA:

\[
y_t = \mu + \sum_{i=1}^{p} r_i y_{t-i} + \sum_{i=1}^{q} \varphi_i \varepsilon_{t-i} + \varepsilon_t
\]

where \( p \) is the autoregressive term and \( q \) is the moving average number of items, and can be written as: ARMA \((p,q)\). Compared with ARMA model, ARIMA has more data preprocessing parts. In the data preprocessing part, the data is first differentiated before processing to make the data conform to the situation of stable distribution. Other data processing parts are the same as ARMA model. ARIMA model can be expressed as ARIAM \((p,d,q)\). Where \( p \) is the number of autoregressive terms, \( d \) is the number of differential times, and \( q \) is the number of moving average terms.

### 3 Supplementary value of trust attribute

Trust is based on knowledge and experience to judge the unknown party, and it is a subjective judgment to the entity. According to the judgment, it can be divided into three levels of trust: disbelief, partial trust and complete trust.

### 3.1 Subjective logic

In subjective logic, evidence space is composed of a series of observable events, which are generated by entities, including positive events, negative events and uncertain events. In subjective logic, the trust relationship between nodes is \( w=(b,d,u) \), \( b \) is the probability of trust, \( d \) is the probability of distrust, \( u \) is the probability of uncertainty, where \( b+d+u=1 \) and \( b,d,u \in [0,1] \). The subjective logic is represented by trust mapping function:

\[
w_{ij} = (b^i_j, d^i_j, u^i_j)
\]

\( b^i_j, d^i_j, u^i_j \) are the probability of trust, the probability of distrust, the probability of uncertainty.
3.1.1 Discount operator

The trust between nodes is transitive, and the discount operator is used to calculate the value of transitive trust decaying between nodes, that is, the calculation method of transitivity is as follows:

\[ w_{kj}^i = w_j^i \Theta w_k^j \]

where:

\[ b_{kj}^i = b_j^i \cdot b_k^i \]
\[ d_{kj}^i = d_j^i \cdot d_k^i \]
\[ u_{kj}^i = d_j^i + u_j^i + b_j^i \cdot u_k^j \]

3.1.2 Fusion operator

When multiple nodes have trust to the same node, the trust value of these nodes can be fused, that is, the fusion operator is:

\[ w_{kj}^{ij} = w_k^i \Theta w_k^j \]

where:

\[ b_{kj}^{ij} = (b_j^i \cdot u_k^j + b_j^j \cdot u_k^i) / U \]
\[ d_k^{ij} = (d_k^j \cdot u_k^i + d_k^i \cdot u_k^j) / U \]
\[ u_k^{ij} = (u_k^j \cdot u_k^i) / U \]

\[ U = u_k^i + u_k^j + u_k^i \cdot u_k^j \quad \text{且} \quad U \neq 0 \]

3.2 Trust mechanism

Trust evaluation mechanism in wireless sensor networks is composed of direct trust and indirect trust. Direct trust is the trust value obtained by the direct transmission and interaction of adjacent nodes, while indirect trust is the indirect transmission of data to cluster head by the way of node hop transmission by non-adjacent nodes. When a node evaluates other nodes, it makes a comprehensive evaluation by evaluating the direct and indirect trust of the nodes.

3.2.1 Direct trust

In wireless sensor network, each sensor node will record and save the interaction information with other nodes, such as the number of successful or unsuccessful packet forwarding, the number of undetermined packets and so on. Record value \( R_{nm} = (\alpha, \beta, \gamma) \) of any node \( n \) to adjacent contact \( m \). Where \( \alpha \) represents the number of successful packet forwarding, \( \beta \) represents the number of unsuccessful packet forwarding, and \( \gamma \) represents the undetermined packet.

The record value of node \( n \) to adjacent contact \( m \) reaches direct trust value \( D_{nm}(t_0 + \Delta t) = (\alpha, \beta, \gamma) \), Expressed as:
\[ w_m^n = (b_m^n, d_m^n, u_m^n) \]  \hspace{1cm} (19)

where:

\[ b_m^n = \frac{\alpha_1}{\alpha_1 + \beta_1 + \gamma_1} \]  \hspace{1cm} (20)

\[ d_m^n = \frac{\beta_1}{\alpha_1 + \beta_1 + \gamma_1} \]  \hspace{1cm} (21)

\[ u_m^n = \frac{\gamma_1}{\alpha_1 + \beta_1 + \gamma_1} \]  \hspace{1cm} (22)

The record value of node \( n \) to adjacent contact \( m \) reaches direct trust value.

3.2.2 Indirect trust

When there are \( k \) intermediate nodes, the evaluation subject node \( p_n \) cannot evaluate the object \( p_m \) through direct trust, but can indirectly evaluate \( p_m \) through the intermediate node \( p_{ij} \).

The interaction information of each node is counted, which includes positive event number \( \alpha \), negative event number \( \beta \), uncertain event number \( \gamma \). The trust value \( w_{ij}^n (j \in [1, k]) \) between \( p_n \) and its adjacent contact is calculated by recording information.

The indirect trust value it of a node in the middle layer is calculated by using the discount operator:

\[ IT_{it} = w_{im}^n \otimes w_{im}^n \]  \hspace{1cm} (23)

where:

\[ b_{im}^n = b_{i1}^n \cdot b_{m1}^n \]  \hspace{1cm} (24)

\[ d_{im}^n = d_{i1}^n \cdot d_{m1}^n \]  \hspace{1cm} (25)

\[ u_{im}^n = d_{i1}^n + u_{i1}^n + b_{i1}^n \cdot u_{m1}^n \]  \hspace{1cm} (26)

\( b_{i1}^n, d_{i1}^n, u_{i1}^n \) are the number of records of node \( p_n \) to object node \( p_{i1} \) and node \( p_n \) to node \( p_{ij} \). \( b_{m1}^n, d_{m1}^n, u_{m1}^n \) is the number of records of node \( p_{i1} \) to object node \( p_m \) and node \( p_{i1} \) to node \( p_m \). When there are multiple intermediate points with \( p_{ij} (j \in [1, k]) \), as shown in Figure 1, the indirect trust value of node \( p_n \) to \( p_m \) can be calculated, that is:

\[ IT_{nm} = w_{im}^{n1} \otimes w_{im}^{n1} \otimes ... \otimes w_{im}^{nk} \]  \hspace{1cm} (27)

4 Test and simulation

In order to verify the effectiveness of the algorithm, simulation is carried out in the existing agricultural environment. The 40m * 200m square area of the greenhouse is selected as the simulation experimental scene. In the square area, 100 nodes are randomly distributed to collect the temperature data and the converging node is in the center. The communication radius of wireless sensor node is about 20 m. The sampling period is 10 min, and
MATLAB is used for experimental simulation. In order to reflect the reliability of the algorithm, the wireless sensor nodes are optionally designed with data loss.

In order to ensure the accuracy of the algorithm data prediction, this paper will predict the non-missing data and compare with the real measured data. Because there is no missing value in the original measurement data, there should be no missing value in the selected measurement data during the simulation experiment. After selecting the experimental data, a number of measured data values are randomly selected as the missing values.

In order to ensure the objectivity of the experiment, the root mean square error (RMSE) is used to calculate the error.

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}
\]  

where \( n \) is the number of missing values, \( y_i \) is the real measurement data, and \( \hat{y}_i \) is the data predicted by the algorithm.

In order to prove the effectiveness of the algorithm, this paper will compare it with two computational memory experiments.

LIN: This algorithm uses time attribute to make up linear value, constructs linear function to predict missing value, builds linear function according to the relationship between adjacent time data and time, and then forecasts missing value.

KMRA: The algorithm predicts the missing value by time attribute and space attribute, establishes the correlation between the space attribute and the adjacent nodes, and establishes the linear relationship between the time attribute and the data of its own time, then synthesizes the predicted value of the two attributes by weight distribution.

![Figure 2(a). Sampling interval/min.](image)

![Figure 2(b). Number of missing.](image)

It can be seen from Figure 2 (a) that the error of the three algorithms will increase with the acquisition time interval. LIN algorithm only makes linear prediction from the time attribute, so the error of LIN algorithm will be higher than that of KMRA and the algorithm in this paper. Although the KMRA algorithm uses the time attribute and the space attribute...
to predict the complement value, the space attribute's complement value also uses the distance relationship with other nodes to establish the linear relationship. The distance attribute of KMRA algorithm only selects the nodes through the distance.

It can be seen from (b) in the figure that when the number of consecutive missing values increases, the error will also increase. When there are continuous missing values, the difference between the two known values of LIN algorithm will be larger, and the time span will be larger, which will cause the error to increase with the number of missing values. The reason why this algorithm is better than KMRA algorithm is that this algorithm first estimates the change trend of its own data through Arima algorithm, rather than through the data of its own adjacent time like KMRA algorithm.

5 Summary

When wireless sensor is collecting data and transmitting data, it is inevitable to encounter unexpected circumstances that lead the data cannot be obtained by the target node. When there is a continuous lack of value, it will have an impact on the analysis and application of the subsequent data. In this paper, we propose the prediction value of ARIMA's time attribute and the prediction value of multiple linear regression based on trust mechanism, and integrate them through weight distribution. This algorithm can effectively predict and supplement the missing value of wireless sensor network. By comparing the real temperature data collected, the experiment shows that the algorithm can effectively predict and estimate the sensor data.

The authors would like to acknowledge the science and technology planning project in Hebei Province (Grant no. 19450904D; Grant no. 20550301D; Grant no. 20557638D).

References

1. Zhenning Li, Hao Yu, Guohui Zhang, Jun Wang. A Bayesian vector autoregression-based data analytics approach to enable irregularly-spaced mixed-frequency traffic collision data imputation with missing values[J]. Transportation Research Part C, 2019, 108.
2. Liu, Feng. A New Missing Values Estimation Algorithm in Wireless Sensor Networks Based on Convolution[J]. Sensors & Transducers, 2013, 20(4).
3. Wangjuan Zhang, Guoyan Xu, Minjia Li, Shuai Zhu. Missing data filling method based on convolutional neural network[J]. Microelectronics & Computer, 2019, 36(03):48-52+57.
4. Ke Xu, Dang Feng, Feng Wang, Xiuyuan Peng. Sensor layout of solar greenhouse based on data fusion[J]. Journal of Shenyang University of Technology, 2019, 41(01):91-97.
5. Jiawen Wu, Guanghui Li. Missing data reconstruction algorithm of wireless sensor network based on KM-FNN[J]. Chinese Journal of Sensors and Actuators, 2019, 32(08):1245-1252.