PROCEEDINGS OF SCIENCE

Missing Data Reconstruction Using Adaptively Updated Dictionary in Wireless Sensor Networks

Liang Zhao[12]
College of Informatics, Huazhong Agricultural University
Wuhan, 430074, China
E-mail: zhaoliang323@mail.hzau.edu.cn

Fang Zheng[34]
College of Informatics, Huazhong Agricultural University
Wuhan,430074, China
E-mail: zhengfang@mail.hzau.edu.cn

Due to external interference or fault, the collected sensor data is often missed or abnormal. It's significant to reconstruct the missing data, especially the large-scale missing data. In this paper, a missing sensor data reconstruction method based on the adaptively updated dictionary is presented. The K-SVD algorithm is used to train the historical data frames which are collected at different time to generate the original dictionary atoms. Moreover, in order to meet the real-time, continuous characteristics of sensor data, an adaptive dictionary update algorithm is studiedwhich. It calculates the correlation between the current reconstructed data frame and the largest weight frame in the training dictionary to update the dictionary. The experimental results are fully analyzed by the open data set. The results show that the proposed method has higher reconstructed precision especially the interval of data frames which is more than 60 minutes compared with other commonly used methods .

CENet2017
22-23 July 2017
Shanghai, China

[1]Speaker
[2]This study is supported by the Natural Science Foundation of Hubei Province of China (program No. 2016CKB705) and the Fundamental Research Funds for Central Universities (Program No.2014QC008).
[3]This study is supported by the Natural Science Foundation of Hubei Province of China (program No. 2015CFB524) and the Fundamental Research Funds for the Central Universities (Program No. 2015BQ023).
[4]Corresponding Author

PoS(CENet2017)040

1.Introduction

Due to interference or fault, the collected sensor data is often missed or abnormal, but some applications need complete data set without missing or abnormal data, otherwise it will impact the prediction accuracy. In this sense, it's significant to reconstruct the missing data by using their inherent characters, especially the large-scale missing data.

There are many methods commonly used to reconstruct the missing sensor data based on temporal correlation, spatial correlation, interpolation method or sparse theory. In general, the linear or the nonlinear reconstruction methods are selected according to the stably changing character of sensor data in a short time period, using the non-missing data adjacent to the missing node[1][2]. Considering the collected sensor data as a time series [3], the linear regression method is used to fill or predict the missing variables [4][5][6]. There are also many nonlinear methods such as least squares [7], support vector regression [8], or neural network model[9] are used. Generally speaking, most of these reconstruction methods only consider the time character of the missing data, the reconstruction accuracies are not high when the data acquisition interval is larger or the data is non-stably changed.

The Kriging interpolation [10][11], IDW[12] [13], KNN[14], Gaussian Mixture Model[15], or RNN[16] are often used to process the geographic information. These methods are also used to estimate the missing data in WSNs. The core of these methods is to find the most appropriate neighbor nodes of the reconstructed node and the appropriate coefficient for each neighbor[17]. But the accuracies of these methods are poor when the sensor data has large fluctuation or big lost rate. Additionly, the compressive sensing theory has been used to collect the sensor data or to reconstruct the missing data combined with the sparse or low-rank character [18][19][20].

Furthermore the reconstruction method based on sparse dictionary is the common method, the key work of which is to select an appropriate sparse dictionary to reduce the estimation error [21]. Training the historical data to get the sparse or over-complete dictionary is a feasible method.The current data frame is sparsely represented by the data atoms of the dictionary. Usually, the discrete cosine transform is used to transform the sensor data into the sparse representation of high-frequency and smooth low frequency signals, then to compute the value of the l_1 norm minimization or singular value decomposition to reconstruct the missing data[22]. At the same time, some methods aim to improve the prediction accuracy and reduce the required sampling data with the temporal correlation and the spatial correlation of sensor data [23]. The basic idea of these methods is to calculate the missing data with high computational complexity in an iterative way.

In this work, we present a new reconstruction method based on adaptively updated dictionary. The K-SVD algorithm is used to train the historical data frames at different time to generate the original dictionary atoms, then, in order to improve the reconstruction accuracy, an updating algorithm is applied to update the dictionary atoms with the current reconstructed frame adaptively.

2.Problem Description

Assuming that all sensor nodes are deployed in a two-dimensional space, where each sensor node generates a real-time, continuous data stream, and meets: (1) there are m sensor nodes and only one sink node in the monitoring area. All the sensor data is transferred to the

sink by a single step to reconstruct the missing data; (2) all sensor nodes are homogeneous and stationary; (3) each sensor node has an identical and a constant sampling rate.

At the $(t-1)$ th, the t th and the $(t+1)$ th time, the sensor data is collected with missing data during transmission. In the following figure, the hollow points represent the sensor data which is lost, and the solid points represent those are not.

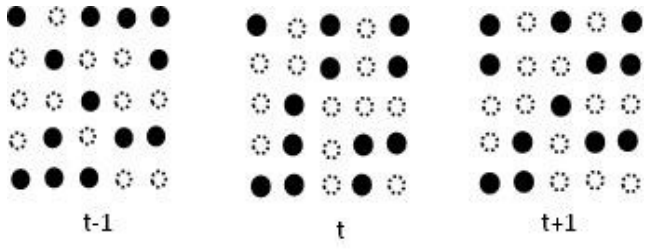

Figure 1: Sensor Data Acquisition

3.Temporal correlations of data frames

Supposing that there are m sensor nodes deployed in the monitoring area, the sense data of the node i at the time j is recorded as y_{ij} , where $i=1,2,\cdots,m$, $j=1,2,\cdots,n$. The continuous data frames are obtained if the data of each node is collected continuously. In order to describe conveniently, the sense data of all nodes at the same time will be organized into a frame as follows:

$$y_j=[y_{1j}, y_{2j},\cdots, y_{mj}]^T \tag{3.1}$$

y_j represents the j^{th} data frame, then n data frames are collected to form a matrix as follows:

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & & & y_{2n} \\ \vdots & & & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mn} \end{bmatrix} \tag{3.2}$$

the temporal correlation between two data frames is calculated as follows[20]:

$$R(\tau)=\frac{1}{\Delta n+1}\sum_{n=n_0}^{n_o+\Delta n} y_n^T\, y_{n+\tau} \tag{3.3}$$

Where y_n is the data frame of the nth time, n_0 is the initial time of the data frames, and $\Delta n+1$ is the total number of the data frames in the time interval τ . The temperature and humidity data frames are selected from 3:00 pm, March 7, 2004 to 5:00 am, March 7, 2004 of the data set [24] to calculate the time correlation. For example description, set $n_0=1$, $\tau=1$ and $\Delta n=900$, there are approximately 900 data frames from one minute interval to 900 minute intervals.

The correlations of different intervals are shown inFigure 2.

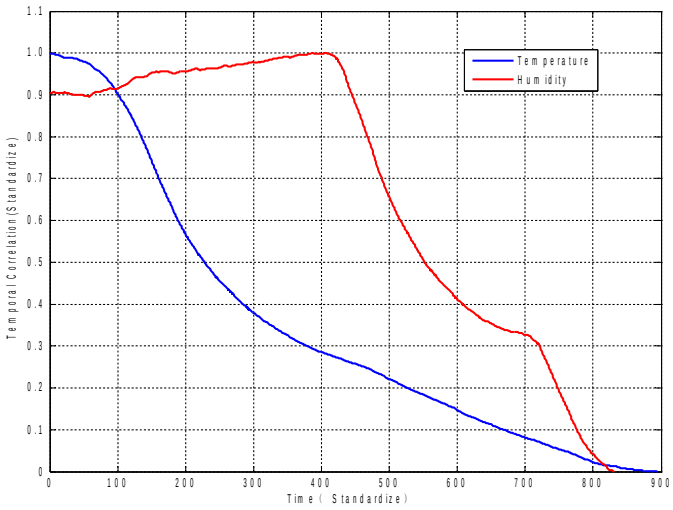

Figure 2: Time Correlation of Data Frames

In Figure 2, when the interval between two frames is within the range of 100 minutes, the correlation is over 0.9. The correlation of humidity data frames is very high even within 400 minutes. It shows that the current frame has strong correlation with its historical frames. But when the interval is sufficiently large, there is a sharp decline in the correlation figure. Consequently, selecting the proper historical frames is important for ensuring a higher correlation within the current frame.

4.Sparse Representation Theory based on Dictionary Learning

4.1 Theory Model of Sparse Representation

In the sparse representation theory, the target signal can be approximatly represented or represented as a linear combination of finite atoms, which constitute a $m \times n$ matrix, denoted as a dictionary D , any signal y can be represented by a linear combination of p elements in the dictionary D if the atoms are orthogonal, namely:

$$y \approx Da = a_1 d_1 + a_2 d_2 + \cdots + a_p d_p \tag{4.1}$$

Where $a \in R^n$ is the linear coefficient, d_i is the atom of D .

4.2Generating K-SVD-based Dictionary

Due to the smoothly changed environmental parameters, the collected sensor data frames have high temporal correlation. The shorter the interval of two frames is , the stronger the correlation is . Combining the temporal correlation andto generate the sparse dictionary, the K-SVD-based dictionary learning algorithm is used to train the historical frames within a certain time interval to obtain the basis function dictionary, then to reconstruct the current sensor frame according to the dictionary atoms and the corresponding sparse coefficients by sparse coding.

Supposing there be m sensor nodes, n data frames are selected as the training data set, which can be expressed as $y = [y_1, y_2, \cdots, y_n]$, where $y_i \in R^n$, $i = 1,2 \cdots, n$. For a given training data y , the dictionary learning algorithm is described as follows:

The inputs : a training data set $y_{m \times n}$ of m sensor nodes' historical data frames.

The outputs : the dictionary D .

The algorithm steps are:

① Input the training data set $y_{m\times n}$, where $m = n$;

② Set the initial size of the dictionary as K ;

③ Call the K-SVD algorithm to produce the dictionary atoms iteratively;

④ Output the dictionary.

4.3Update the Sparse Dictionary Adaptively

In order to reduce the current reconstruction error impact on the next reconstruction frame, it's particularly important to select the dictionary atoms having the maximum correlation with the current data frame. When a fixed dictionary is used to reconstruct the data frames, the reconstruction error will become bigger as the sampling interval increasing. Thus an adaptive dictionary update method is presented to reduce the reconstruction error, which is feasible to selectively update the dictionary atoms. The dictionary updating algorithm proposed in this study is as follows:

The input is : the latest reconstructed data frame $\hat{y}_n$.

The output is : the updated dictionary $D^{'}$.

The algorithm steps are:

① Calculating the weight $w_i = \|d_i\|_2$ for each data frame of the corresponding dictionary atom;

② Updating the weight as $w_i = w_i * e^{\alpha_i}$ according to the sparse coefficient of each data frame;

③ Getting the index of the data frame having the minimum weight w_i , and assigning to i_0 ;

④ Replacing the data frame d_{i_0} with $\hat{y}_n$, that is $d_{i_0} \leftarrow \hat{y}_n$;

⑤ Calculating the average weight of all the dictionary atoms, and assigning to $w_{i_0} \leftarrow mean(w)$;

⑥ outputting the updated dictionary $D^{'}$.

5.Implementation of the Reconstruction Algorithm

The reconstruction algorithm based on adaptively updated dictionary consists of two steps: the first step is to build the sparse dictionary based on K-SVD andthe second step is to update the dictionary atom according to the updating condition. The flow of the algorithm is as below:

a) The input : the training data $y=[y_{1,}y_{2,}\cdots,y_n]$, and set N as the total number of data frames to be reconstructed.

b) The output : the reconstruction error.

c) The algorithm steps are:

① Obtain the historical data frames $y=[y_{1,}y_{2,}\cdots,y_n]$;

② Produce the initial dictionary D ;

③ Call the l_1 norm minimization algorithm to get the sparse coefficient α_i of each frame d_i ;

④ Reconstruct the data frame $\hat{y}_n \approx \sum_{i=1}^{k} a_i d_i$ with the dictionary atom d_i and the coefficient α_i;

⑤ Analyze the dictionary update condition, the judging method is as follows:

If the $sim(\hat{y}_n, d_m) < \varepsilon$ (d_m is the corresponding data frame of the maximum coefficient α_m, $\varepsilon(0 \sim 1)$ is the update threshold value). If the similarity of $\hat{y}_n$ and d_m are less than ε, then go to ⑥, otherwise go to ⑦;

⑥ Call the update algorithm to update D with $\hat{y}_n$ to get D';

⑦ Calculate the reconstruction error as $r_n = \sqrt{\sum_{i=1}^{m} (y_n - \hat{y}_n)^2 / m}$;

⑧ If data frames finish , go to ⑨;otherwise go to ③;

⑨ Exit.

6.Experimental Results and Analysis

In this paper, the monitoring data sets of (Intel Berkeley Research Lab Data) Intel Berkeley Lab's are used. There are 54 sensor nodes deployed in the Intel Berkeley Lab which generated 32 million data approximately every 30 seconds in 36 days, including temperature, humidity, light and the voltage value of all the nodes. The training data of 30 minutes interval are chosen to generate the basis function dictionary for reconstructing the temperature data of 52 nodes (because there are too much missing data of the 5th and the 15th node, so the two nodes' data are not taken into account) in the 89th time by using the discrete cosine transform (DCT), K-SVD-based method (K-SVD) and the updated K-SVD-based method (MK-SVD).

In order to show all the 52 nodes' data clearly, 5 nodes are selected out of every 11 nodes. The compared curves of the reconstruction data based on different methods and the original data are shown in Figure 3.

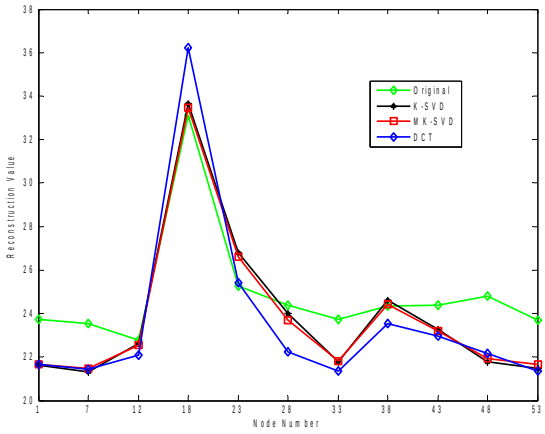

Figure 3: Reconstruction Data of Different Algorithms Compared with the Original Data

There is a great fluctuation of the 18th node's original data. In Figure 3, the K-SVD, MK-SVD and DCT methods are used to reconstruct the missing data, which have some fluctuations. The data reconstructed by the MK-SVD is most similar to the original data. The data frame of the next 10 moments is reconstructed based on the sparse dictionary generated by DCT, K-SVD

and MK-SVD respectively. The reconstruction errors based on different dictionaries of 30 minutes interval are shown in Figure 4.

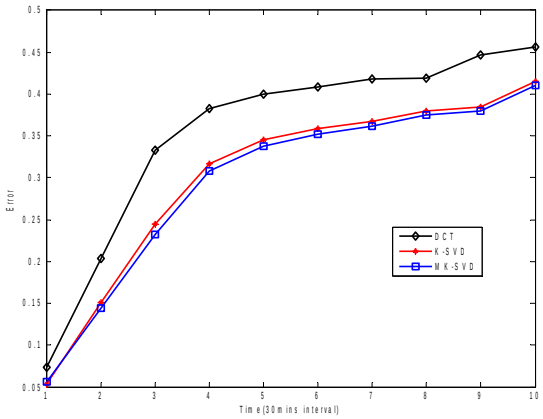

Figure 4: Reconstruction Error Comparison based-on Different Dictionaries

The experimental parameters are: the time interval is 30 minutes, the sparsity k is 2, the dictionary size is 52 * 80, the Lagrange coefficient λ is 0.1, and the ε is 0.6. From the figure, the reconstruction errors are bigger when the interval is bigger, because the temporal correlation of data frame is smaller when the interval is bigger, which will affact the reconstruction error. But the reconstruction error based on the DCT dictionary which fits for non-stationary change data is significantly larger than the other two methods, and the error of the method based on the updated dictionary is smaller than that of the method based on K-SVD. From the experiment, the algorithm based on adaptively updated dictionary can effectively reduce the reconstruction error.

The reconstruction errors of different time intervals before and after the dictionary updated are shown in Figure 5.

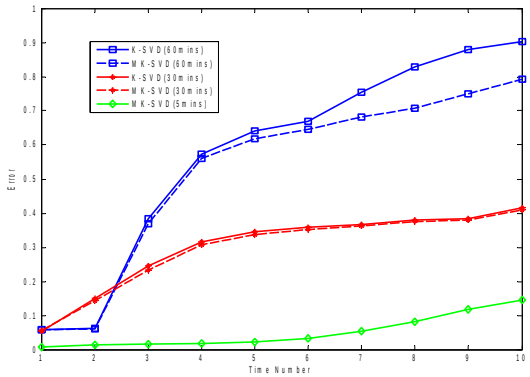

Figure 5: Reconstruction Error Comparison of Different Time Intervals before and after the Dictionary Updated

In Figure 5, the solid line represents the reconstruction error based on the not updated dictionary, and the dotted line represents the reconstruction error based on the updated dictionary. From the Figure, the bigger the time interval between the training data frames is, the more effective the proposed method is . When the training data is at 5 minute intervals, the reconstruction errors of the two methods are coincide with each other. The main reason is that the larger the time interval is, the smaller the correlation between data frames is. The least relevant training can be replaced by updating the dictionary to reduce the reconstruction error.

Conversely, the smaller the time interval is, the bigger the correlation of the data frames is andthe less chance the dictionary is updated.

7.Conclusion

In this paper, the missing sensor data reconstruction algorithm based on adaptively updated dictionary learning is proposed. Upon analyzing the correlations between the data frames at a certain time interval of all the sensor nodes deployed in monitoring area, the main contents are as follows:

The K-SVD algorithm is used to produce the basic sparse dictionary. The historical data frames is used as training data to get dictionary atoms based on K-SVD algorithm for reconstruction, which constitutes the dictionary D;

An adaptive dictionary update algorithm is presented. The original dictionary generated by training the historical data frames will be updated if the new reconstructed data frame satisfies the update conditions.

Experimental results show that the algorithm proposed in this paper has smaller reconstruction error especially when the time interval of the sensor data frames is bigger, more than 60 minutes in particular. In our next study, we will improve the dictionary updating algorithm to make it more efficient and less computational complex.

References

[1] Pan Liqiang, Li Jianzhong. *A Multiple Regression Model Based Missing Values Imputation Algorithm in Wireless Sensor Network*[J]. Journal of Computer Research and Development, 2009, 46(12):2101-2110

[2] PAN Li-Qiang LI Jian-Zhong LUO Ji-Zhou. *A Temporal and Spatial Correlation Based Missing Values Imputation Algorithm in Wireless Sensor Networks*[J]. Chinese Journal of Computers, 2010, 33(1):1-11

[3] Chen guangping, *Missing value estimating algorithm based on time series data properties*[J]. Computer Engineering and Applications, 2012, 48(12):135-138.

[4] Guestrin C, Bodik P, Thibaux R, et al. *Distributed regression: an efficient framework for modeling sensor network data*[C]. Third International Symposium on Information Processing in Sensor Networks, 2004, Berkeley, California,USA

[5] Jain A, Chang E Y, Wang Y-F. *Adaptive stream resource management using Kalman Filters*[C]. Proceedings of the ACM SIGMOD International Conference on Management of Data, 2004, Paris, France

[6] Stojkoska B, Solev D, Davcev D. *Data prediction in WSN using variable step size LMS algorithm*[C]. The fifth international conference on sensor technologies and applications, 2011, Nice/Saint Laurent du Var, France

[7] Kamal A R M, Hamid M A. *Reliable data approximation in wireless sensor network*[J]. Ad Hoc Networks, 2013, 11(8):2470-2483

[8] Liu X, Fang X, Qin Z, et al. *A Short-Term Forecasting Algorithm for Network Traffic Based on Chaos Theory and SVM*[J]. J. Network Syst. Manage., 2011, 19(4):427-447

[9] Murgante B, Borruso G,Lapucci A, et al. *Nonlinear Black-Box Models for Short-Term Forecasting of Air Temperature in the Town of Palermo*[J]. Springer Berlin Heidelberg,2011,183-204

[10] Mendez D, Labrador M, Ramachandran K. *Data interpolation for participatory sensing systems*[J]. Pervasive and Mobile Computing, 2013, 9(1):132-148

[11]Agarwal B T A. *A New Approach to Spatio-Temporal Kriging and Its' Applications* [Ph.D thesis]. Ohio State: Ohio State University, 2011

[12] Villas L A, Boukerche A, Guidoni D L, et al. *An energy-aware spatio-temporal correlation mechanism to perform efficient data collection in wireless sensor networks*[J]. Computer Communications, 2013, 36(9):1054-1066

[13] Umer M, Kulik L, Tanin E. Kriging for Localized Spatial Interpolation in Sensor Networks [C]. 20th International Conference, SSDBM 2008, Hong Kong, China

[14] Li Y Y, Parker L E. *Nearest neighbor imputation using spatial-temporal correlations in wireless sensor networks*[J]. Special Issue: Resource Constrained Networks, 2014, 15(1):64-79

[15] Xiaobo Yan, Weiqing Xiong, Liang Hu,et al. *Missing Value Imputation Based on Gaussian Mixture Model for the Internet of Things*[J]. Mathematical Problems in Engineering, 2015, 2015(3):1-8

[16] Moustapha A I, Selmic R R. Wireless Sensor Network Modeling Using Modified Recurrent Neural Networks: Application to Fault Detection[J]. IEEE Transactions on Instrumentation and Measurement,2008, 57(5):981-988

[17] Park I, Choi J, Jin Lee M, et al. *Application of an adaptive neuro-fuzzy inference system to ground subsidence hazard mapping*[J]. Computers & Geosciences, 2012, 48(0):228-238

[18] Quer G, Masiero R, Pillonetto G, et al. *Sensing, Compression, and Recovery for WSNs: Sparse Signal Modeling and Monitoring Framework*[J]. IEEE Transactions On Wireless Communications, 2012, 11(10):3447-3461

[19] Yan W J, Wang Q, Shen Y, et al. *Deterministic Measurement Matrix Generation for Compressive Data Gathering and Reconstruction in WSNs*[J]. Journal of Internet Technology, 2013, 14(6):973-984

[20] Di G, Zicheng L, Xiaobo Q, et al. *Sparsity-Based Online Missing Data Recovery Using Overcomplete Dictionary*[J]. IEEE Sensors Journal, 2012, 12(7):2485-2495

[21] Guo D, Qu X, Huang L, et al. *Sparsity-Based Spatial Interpolation in Wireless Sensor Networks*[J]. Sensors, 2011, 11(3):2385-2407

[22] Linghe K, Mingyuan X, Xiaoyang L. *Data loss and reconstruction in sensor networks*[J]. IEEE Infocom, 2013, 1654-1662

[23] Jie C, Qiang Y, Hongbo J, et al. *STCDG: An Efficient Data Gathering Algorithm Based on Matrix Completion for Wireless Sensor Networks*[J]. IEEE Transactions on Wireless Communications, 2013,12(2):850-861

[24] S.Madden, Intel Berkeley Research Lab Data [OL], 2006, http://db.csail.mit.edu/labdata/data.txt.gz.