Sensory Data Prediction Using Spatiotemporal Correlation and LSTM Recurrent Neural Network

Tongxin SHU

(*Department of Electrical and Computer Engineering*, *The University of British Columbia*, *Vancouver*, *Canada V6T 1Z4*)

Abstract: The Wireless Sensor Networks (WSNs) are widely utilized in various industrial and environmental monitoring applications. The process of data gathering within the WSN is significant in terms of reporting the environmental data. However, it might occur that certain sensor node malfunctions due to the energy draining out or unexpected damage. Therefore, the collected data may become inaccurate or incomplete. Focusing on the spatiotemporal correlation among sensor nodes, this paper proposes a novel algorithm to predict the value of the missing or inaccurate data and predict the future data in replacement of certain nonfunctional sensor nodes. The Long-Short-Term-Memory Recurrent Neural Network (LSTM RNN) helps to more accurately derive the time-series data corresponding to the sets of past collected data, making the prediction results more reliable. It is observed from the simulation results that the proposed algorithm provides an outstanding data gathering efficiency while ensuring the data accuracy.

Key words: Spatiotemporal correlation; LSTM Recurrent Neural Network; time-series prediction

1 Introduction

A reliable WSN for industrial process or environmental monitoring typically consists of a number of sensor nodes that are reasonably deployed in the region of interests (ROI)[1], seeking efficient data collection to report the latest changes within the environment. In this regard, a timely and accurate measuring process is of great significance since the deployer of the WSN needs to know the latest changes in the ROI in the application. However, constrained by limited energy and the characteristics of the WSN, some sensor nodes may deplete faster than others and consequently, the collected data might be inaccurate and even incomplete. Particularly, in sensor nodes that are deployed in remote or hazardous areas, it might be infeasible and human-labor consuming to interact. In such situations, a backup plan could be to use the neighboring sensor nodes to predict the missing data for the dysfunctional sensor nodes, relying on the spatiotemporal correlations among them.

The temporal correlation determines the connection among a set of time-series data, which may indicate that the value of certain data at some point is somewhat a continuity of the previous values. This is typically demonstrated when a set of seasonal data is being observed in environmental monitoring applications. For instance, the temperature in a water body may display a seasonal trend, where for any given point in the time domain, the corresponding value of the data could be approximated based on its adjacent time-series data. Some research that focuses on the temporal correlation has been carried out for time series prediction. The Least Mean Square (LMS) filter and the Auto Regressive Integrated Moving Average (ARIMA) model, for instance, are such convenient tools that require relatively minor computational memory and have the advantage of effectively capturing the characteristics of the data set in the time domain. Thanks to the emergence of those mathematical models, much energy could be saved by using time-series prediction without real sensing in some practical monitoring applications where a WSN is utilized[2]. Likewise, reduced communication among

different sensor nodes could be realized through the time-series prediction in pursuit of energy conservation[3]. Regardless, a suitable mathematical model with proper parameters would be energy-efficient for time-series prediction using temporal correlation.

The spatial correlationis the geographical relationship and connections among the sensor nodes with respect to their locations. This may mean that the measurement from certain sensor node could be predicted by its surrounding sensor nodes. In such context, if a sensor node fails to collect and provide a reliable measurement, the nearby sensor nodes, depending on their spatial correlation, could collaboratively predict a measurement with adequate confidence. This will also pave the path for energy conservation within a typical WSN for environmental monitoring applications. For example, the design of routing protocols in WSNs is a major research topic, which has drawn much attention. It is widely believed that an unbalanced deployment of sensor nodes in the ROI may cause some sensor nodes to deplete faster than others. In addition, a sensor node closer to the sink might take a heavier burden for node-to-node communication because most of the data packets might be sent to the sink via such a sensor node, which means a more frequent node-to-node communication and greater energy consumption. Fortunately, with the existence of spatial correlation, it is possible to turn off a "dying" sensor node and have its surrounding sensor nodes predict the future measurements. In this manner, it is even possible to further probe into some duty-cycling schemes[4] since some redundant sensor nodes could be selectively shut down and their corresponding sensor readings could be predicted by the neighbour sensor nodes.

Previously, Liu et al.[4] proposed to capture the temporal correlation by partitioning the time-series data into a piecewise segments and assume that the missing measurements have a linear relationship with the adjacent data. In addition, the spatial correlation is determined by clustering sensor nodes, which are a small group of sensor nodes geographically close to each other, forming a cluster. Within each cluster, the measurements of some sensor nodes could be collaboratively predicted by their surrounding sensor nodes using the spatial correlation. The number of sensor nodes within each cluster and the partitioned segments is also dynamically determined by how fast and dramatically the measured parameters are changed within the environment. Yoon et al.[5], proposed a scheme in a similar fashion, which first divides the WSN into a number of clusters, and then leverages the spatial correlation within each cluster to drastically reduce the number of in-network transmissions. However, in terms of the characteristics of the measured parameter, a different pre-defined threshold and a fine-tuning process have to be implemented to guarantee the performance of the algorithm. Pattem et al.[6] discovered that, by taking advantage of the spatial correlation among the sensor nodes, it is possible to realize a near-optimal routing scheme and data compression, which both contribute to handling the energy-efficiency challenges in practical WSN applications. Other than trying to decrease the number of transmissions and clustering the scattered sensor nodes, there are also schemes that aim to conduct compressive sensing within the sensor nodes by exploiting the spatiotemporal correlations[7-8]. In a nutshell, it is nontrivial to note that the spatiotemporal correlation is a key notion that should be studied to tackle the energy-efficiency issues in practical monitoring applications within a WSN.

Besides those lightweight mathematical models mentioned beforehand, the LSTM RNN is another efficient and promising tool that provides solid prediction results on time-series data. Different from the traditional time-series sequence predictors such as the ARIMA, LMS and Kalman filter, the LSTM RNN is much more favored today since it is not only able to carry characteristics of the time-series data throughout the time domain, but also capable of capturing both linear and nonlinear relationships in any given data segments, for future prediction.

Some recentwork such as [9-10] have seen success in implementing LSTM RNN for time-series prediction. In [9], the LSTM RNN has been adopted for forecasting the environmental data, where the data of air pollutants are predicted at high accuracy. Likewise, [11] have proposed to train an LSTM RNN model in order to predict the airline demand, which also displays time-series characteristics within the time domain.

Taking advantage of the effectiveness of the spatiotemporal correlations and the efficiency of the LSTM RNN, the present paper seeks to solve a practical issue in environmental monitoring. That is, when a specific sensor node is unable to conduct the sensing task in a normal manner, due to the energy constraint, the spatiotemporal correlation should be determined according to the neighboring sensor nodes. Moreover, in order to ensure the future measurements are sufficiently accurate, the LSTM RNN will be chosen for time-series data prediction.

The present paper is organized as follows: Section 2 presents the formulated problem along with the developed methodology. Section 3 demonstrates the simulation and the corresponding results. What follows are the Conclusion and Discussion, in Section 4.

2 ProblemFormulation and Methodology

2.1 Illustration for practical WSN applications and assumptions

The environmental monitoring applications that utilize WSN are present everywhere [12-13]. For instance, in a large aquatic field, where a number of sensor nodes might need to be deployed in order to cover the whole area as much as possible. As can be seen in Figure 1, in ideal conditions, the sensor nodes are required to constantly collect the water-related data such as the dissolved oxygen, temperature, conductivity, and oxidation-reduction potential. from the aquatic body. Meanwhile, as mentioned before, some sensor nodes closer to the sink might be ore active in transmitting data packets from node to node, thus consume more energy than other nodes that are farther to the sink. This unbalanced condition of energy consumption would result in the malfunction of the very sensor node. As illustrated in Figure 1, under a normal routing scheme, the sensor node denoted as N_1 is believed to consume more energy due to more frequent communication with its neighbor nodes. As a consequence, there exists a higher possibility that the sensor node N_1 may provide inaccurate measurement or malfunction. Another case would be an unexpected natural disaster or attack from the wildlife could cause the sensor nodes to malfunction. Hence, a backup plan to have other sensor nodes ready to work for the malfunctioned sensor nodes is always highly desirable.

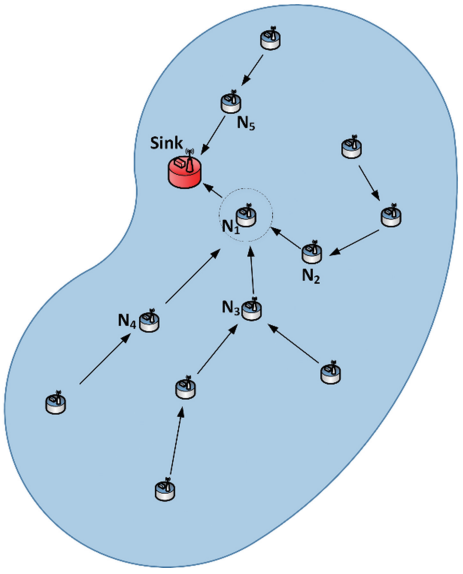

Fig. 1 An example of deploying the sensor nodes in a pool for aquatic monitoring.

Before further looking into the proposed scheme, a few assumptions have to be made to allow the formulated problem and the developed methodology more straightforward.

First, it is assumed that all the sensor nodes can reach at least another sensor node in its neighbor. That means, each sensor node has a sufficiently large communication range to cover at least one sensor node within its own range.

Second, the sensor nodes are not mobile, which means that they retain their initial position

once deployed. Considering the fact that the sensor nodes may be deployed on a water surface in aquatic monitoring applications, it is possible that some of the sensor nodes will slightly drift. However, there are many other environmental monitoring applications where the sensor nodes remain static[14]. Hence, for simplicity and scalability, it is assumed that all the sensor nodes in the WSN are non-mobile.

Last, each sensor node knows its own location as well as the others' via such techniques as Global Positioning Systems (GPS)[15].

2.2 Measurement estimation using spatial correlation

The selection of the surrounding sensor nodes is crucial. As there are many sensor nodes that simultaneously work in the field, there could be several sensor nodes surrounding a sensor node that is dysfunctional or running out of energy. For a sensor node that needs prediction using its surrounding sensor nodes, to properly select the candidate nodes, it is necessary to know their relative distance. Consider the case displayed in the Figure 2. If the sensor node N_1 starts to malfunction, it would not make much sense to use the data from N_7 and N_8 in the upper right since the distance is too far and the collected data samples in at the upper right and bottom left might vary drastically. Also, if the sensor node density is too high in a certain area, using all the surrounding sensor nodes that are quite close would also be a waste of energy. For instance, the prediction results of using node N_2 and N_3 together might lead to a similar prediction result as when only N_2 is used, since they are geographically close to each other. Hence, more preferably, the data collected from N_4, N_5 and N_6 might provide slight diversity, which could collaboratively reflect the overall condition of the monitored area. Thus it is more appropriate to have those nodes to be the potential candidate nodes.

In order to determine the proper candidate sensor nodes, a selection scheme has been developed as follows, where the candidate sensor nodes should meet the requirement as:

$$N_{i,j}=\{N_j \mid \alpha<\frac{d_{Ni,Nj}}{d_{AVG}}<\beta,\ j=1,\cdots,N,\ j\neq i\} \tag{1}$$

Here $N_{i,j}$ stands for the candidate sensor nodes chosen by sensor node N_i, α and β are the lower bound and upper bound that decide the range that N_i will choose the sensor nodes from, $d_{Ni,Nj}$ represents the distance between N_i and N_j, while d_{AVG} is the average distance between N_i and every other sensor node within the field.

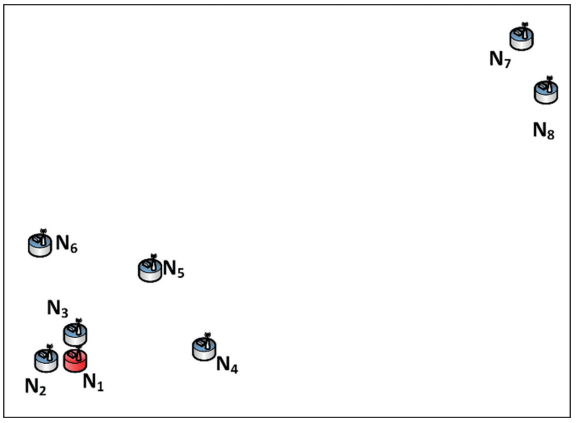

Fig. 2 An example for choosing suitable candidate sensor nodes.

In this regard, at any given point in the time domain, the estimated measurement at that moment forN_i may be expressed as:

$$x_i(t)=\frac{1}{n}\cdot\sum\left[\left(1-\frac{1}{\frac{d_{Ni,Nj}}{d_{AVG}}+1}\right)\cdot x_j(t)\right],\ \forall\ N_j\in N_{i,j} \tag{2}$$

where x_i is the estimated value for N_i at time t and x_j is the sensor reading of N_j at time t. Therefore, at any given time, the approximated measurement for any node could be estimated depending on its spatial correlation with its surround sensor nodes.

2.3 Time-series prediction using LSTM RNN

As for the time-series prediction, the LSTM RNN is essentially an extension of the traditional Re-

current Neural Network (RNN). An obvious advantage of the RNN over the traditional Neural Network (NN) is that, the time-series features among the data sequence can be carried throughout the time domain. This is owning to the fact that the outputs of each training iteration will be jointly fed as an input for next-round iteration. Nevertheless, one formidable challenge is that a gradient exploding or vanishing problem might occur and as a result, the impacts of past inputs on future predicted measurements might be gradually undermined. To overcome this problem, the LSTM RNN has been developed with a special component, which is called a memory cell. A memory cell consists of three components, which are input gate, output gate and forget gate. Compared to the traditional NN and RNN, it is the forget gate that makes the LSTM RNN superior because it can selectively choose how much information should be kept and carried out over any arbitrary time. Thus, the gradient exploding or vanishing problems could be avoided

However, the Recurrent Neural Network (RNN), compared to the traditional NNs, is capable of capturing the dependencies of long-range time-series sequences. That means, while processing the time-series sequences, the trained model acquires the correlation within the sequence, and the prediction is thereby influenced and determined by a long range of past inputs. Due to this superiority, RNNs have been widely utilized in various applications such as speech recognition, prescription systems, image processing, and natural language processing. Unfortunately, constrained by the inner structure of an RNN, the issue of gradient vanishing/exploding emerges, which necessitated the enhanced version LSTM RNN. A detailed description of LSTM is found in [16].

Figure 3 shows the structure of the LSTM cell.

As demonstrated in the figure, the LSTM RNN utilizes the memory cell to selectively store the information and carry them forward to the neurons. Respectively denote the forget gate, input gate, output gate, and cell state as f_t, i_t, o_t and C_t. Then the schemes for updating the prediction model and carrying the information forward are expressed as:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{3}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{4}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$\widetilde{C_t} = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{6}$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t} \tag{7}$$

$$h_t = o_t * \tanh(C_t) \tag{8}$$

where h_{t-1} stands for the output vector in the last iteration, σ denotes the sigmoid function, $\{W_f, W_i, W_o, W_c\} \in \mathbb{R}^{n \times 2n}$ represent the weight matrix and $\{b_f, b_i, b_o, b_c\} \in \mathbb{R}$ represent the bias matrix. In addition, (6) derives a new candidate value $\widetilde{C_t}$, deciding to what degree the information has to be retained and later on, added to a new state in (7). In the meantime, (5) and (8) determine the output vector h_t at the current time step.

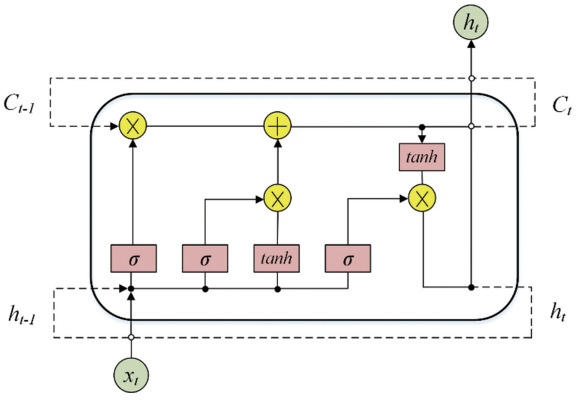

Fig. 3 Structure of the LSTM cell.

To summarize the proposed scheme, when the remaining energy of a sensor node is too low, a beacon signal should be sent to its surrounding sensor nodes for notification. Subsequently, the chosen sensor nodes start to conduct the time-series prediction based on its own past measurements. With each candidate nodes' prediction, the spatial correlation is then combined with weights according to equation (2). Finally, the estimated measurement at any given time for the node is computed.

3 Simulation study

3.1 Simulation setup

To evaluate the performance and validate the efficiency of the proposed scheme, the sensory data from Intel Berkley lab has been selected[17]. As can be seen from Figure 4, there are 56 sensor nodes in total, which recorded light, temperature, humidity and voltage data from Feb 28th to April 5th in 2004.

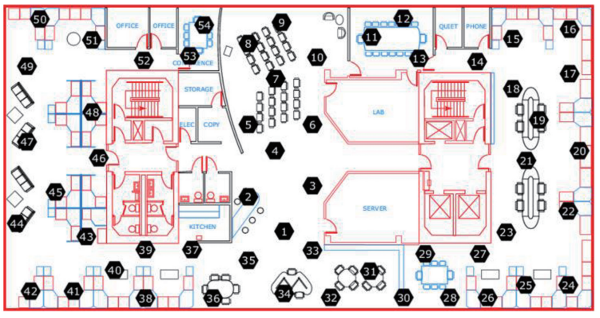

Fig. 4 54 Sensor nodes deployed in Intel Berkley research lab.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \quad (9)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{x}_i| \quad (10)$$

The node 12 has been randomly chosen and the measured temperature data from Feb 28 until March 3 are used for simulation, which consist of 3000 samples in total. The lower bound and upper bound are set as 0.1 and 0.3, respectively. In that sense, the sensor nodes N_{10} and N_{14} fall in the N_{12}'s range and serve as the candidate nodes for collaboratively predicting measurements using their spatial correlation. It is assumed that N_{12} starts to malfunction after collecting 3000 samples, then the remaining 1000 samples are predicted relying on N_{10} and N_{14}. After many rounds of trial and error, it is found that the optimal number of neurons in the first layer should be set as 40 while it should be 20 for the number of neurons in the second layer. To feed the LSTM RNN, the batch size is chosen to be 50 and the training epoch is 50 rounds.

3.2 Simulation results

Figure 5 and Figure 6 present the original data collected by N_{12} and the prediction results after using LSTM RNN and spatial correlations on N_{10} and N_{14}. It is believed that the accuracy of the predicted data is guaranteed while N_{12} stops collecting data from the environment.

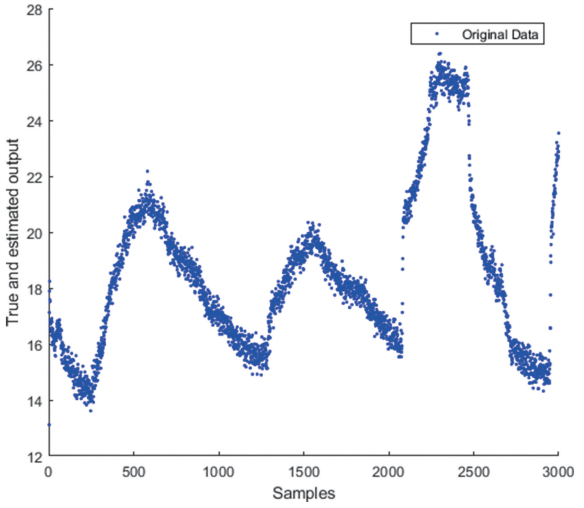

Fig. 5 Original temperature data collected by N_{12}

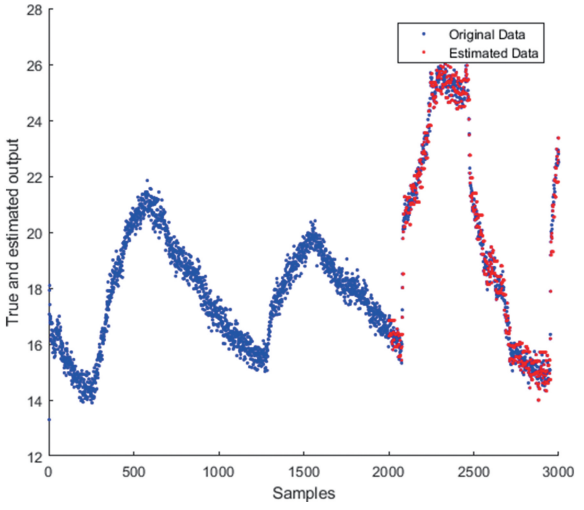

Fig. 6 Prediction results for N_{12} starting from the 2001th data.

To illustrate the accuracy of the developed scheme, the Mean Square Error (MSE) and Mean Absolute Error (MAE) are used to evaluate the prediction results. The MSE and MAE are defined as follows:

Here x_i represents the ground-truth value in the test data set and $\hat{x}_i$ stands for the predicted value. Table 1 indicates the corresponding performance after choosing different value for the upper bound.

Table 1 Prediction accuracy for different value of upper bound.

Upper bound	0.3	0.4	0.5	0.6	0.7
MSE	0.021	0.037	0.054	0.067	0.093
MAE	0.0342	0.0599	0.0689	0.0718	0.0864

4 Discussion and conclusion

From the presented simulation results, it is seen that when a larger upper bound is chosen, more sensor nodes will be involved for data prediction. However, the accuracy will be slightly undermined because of the diversity among the collected data. Despite that, the overall MSE and MAE are still numerically small, which shows the effectiveness and efficiency when a sensor node malfunctions or provides inaccurate measurements. Moreover, determination of the upper bound depends on the specific monitored area and the number of available sensor nodes deployed in the ROI. A trial-and-error process may be taken before choosing a suitable value for the upper bound.

This paper offered a solution when some sensor nodes were unable to provide reliable sensor readings. Focusing on the spatiotemporal correlation and the handy time-series prediction tool LSTM RNN, a scheme for selecting proper candidate nodes as well as collaboratively predicting future measurements were developed. It is believed that, when the whole WSN suffers from a heavy overload, there is a high chance that more sensor nodes need turning off or might go dysfunctional; hence, keeping only a few and most representative sensor nodes "alive" and having them simultaneously predict for other sensor nodes would be an energy-efficient strategy.

ACKNOWLEDGMENT

Funding for this research is provided by the Natural Sciences and Engineering Research Council of Canada.

References

[1] M. T. Lazarescu. (2013). Design of a WSN platform for long-term environmental monitoring for IoT applications, IEEE J. Emerg. Sel. Top. Circuits Syst., vol. 3, no. 1, pp. 45-54.

[2] S. K. Soni. (2012). Reducing the Data Transmission in WSNs using Time Series Prediction Model, 2012 IEEE Int. Conf. Signal Process. Comput. Control, pp. 1-5.

[3] L. Tan, M. Wu. (2016). Data Reduction in Wireless Sensor Networks: A Hierarchical LMS Prediction Approach. IEEE Sens. J., vol. 16, no. 6, pp. 1708-1715.

[4] C. Liu, S. Member, K. Wu, J. Pei, and S. Member. (2007). An Energy-Efficient Data Collection Framework for Wireless Sensor Networks by Exploiting Spatiotemporal Correlation. vol. 18, no. 7, pp. 1010-1023.

[5] S. Yoon and C. Shahabi. (2005). Exploiting Spatial Correlation Towards an Energy Efficient Clustered AGgregation Technique (CAG). IEEE Int. Conf. Commun. 2005. ICC 2005. 2005, vol. 5, no. C, p. 3307-3313 Vol. 5.

[6] S. Pattem, B. Krishnamachari, and R. Govindan. (2008). The Impact of Spatial Correlation on Routing with Compression in Wireless Sensor Networks. vol. 4, no. 4.

[7] B. Gong, P. Cheng, Z. Chen, N. Liu, L. Gui, and F. De Hoog. (2015). Spatiotemporal Compressive Network Coding for Energy-Efficient Distributed Data Storage in Wireless Sensor Networks. IEEE Commun. Lett., vol. 19, no. 5, pp. 803-806.

[8] L. Quan, S. Xiao, X. Xue, et al. (2016). Neighbor-Aided Spatial-Temporal Compressive Data Gathering in Wireless Sensor Networks. IEEE Commun. Lett., vol. 20, no. 3, pp. 578-581.

[9] V. Chaudhary, A. Deshbhratar, V. Kumar, and D. Paul. (2018) "Time Series Based LSTM Model to Predict Air Pollutant's Concentration for Prominent Cities in India."

[10] Karim, F., Majumdar, S., Darabi, H., & Chen, S. (2018). LSTM fully convolutional networks for time series classification. IEEE Access, 6, 1662-1669.

[11] Pan, B., Yuan, D., Sun, W., Liang, C., & Li, D.

(2018, June). A Novel LSTM-Based Daily Airline Demand Forecasting Method Using Vertical and Horizontal Time Series. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 168-173). Springer, Cham.

[12] Alippi, C., Camplani, R., Galperti, C., & Roveri, M. (2011). A robust, adaptive, solar-powered WSN framework for aquatic environmental monitoring. IEEE Sensors Journal, 11(1), 45-55.

[13] Chi, Q., Yan, H., Zhang, C., Pang, Z., & Da Xu, L. (2014). A reconfigurable smart sensor interface for industrial WSN in IoT environment. IEEE transactions on industrial informatics, 10(2), 1417-1425.

[14] Turner, J. S., Ramli, M. F., Kamarudin, L. M., Zakaria, A., Shakaff, A. Y. M., Ndzi, D. L., ... & Mamduh, S. M. (2013, December). The study of human movement effect on Signal Strength for indoor WSN deployment. In 2013 IEEE Conference on Wireless Sensor (ICWISE) (pp. 30-35). IEEE.

[15] Rahili, S., Lu, J., Ren, W., & Al-Saggaf, U. M. (2018). Distributed Coverage Control of Mobile Sensor Networks in Unknown Environment Using Game Theory: Algorithms and Experiments. IEEE Transactions on Mobile Computing, 17(6), 1303-1313.

[16] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems, 28(10), 2222-2232.

[17] http://db.csail.mit.edu/labdata/labdata.html

Authors' Biographies

Tongxin SHU received the B.S. degree in automation engineering from Xiamen University, Xiamen, China, in 2014, and the M.ASc. degree in mechanical engineering from the University of British Columbia, Vancouver, BC, Canada, in 2016. He recently obtained the Ph.D. degree in electrical and computer engineering, also from the University of British Columbia. His research interests lie in the areas of wireless sensor networks, deep neural networks, algorithm design, and power management issues in environmental monitoring applications.