Data Prediction Model in Wireless Sensor Networks: A Machine Learning Approach

Khushboo Jain[1(✉)], Manali Gupta[1], and Ajith Abraham[2,3]

[1] School of Computing, DIT University, Dehradun, UK, India
khushboojain2806@gmail.com
[2] Machine Intelligence Research Labs (MIR Labs), Auburn, Washington 98071, USA
ajith.abraham@ieee.org
[3] Center for Artificial Intelligence, Innopolis University, Innopolis, Russia

Abstract. In resource constraint wireless sensor networks (WSN), an important design concern is the optimization of the data transmission reduction of each sensor node (SN) to extent the overall network lifetime. Numerous cited works claim that the Data Prediction Method (DPM) is the most competent method for data transmission reduction among data aggregation, data regression, neural networks models, spatiotemporal correlation, clustering methods, adaptive sampling and data compression. The big data is generally communicated across the WSN which leads to packet collisions, packet drops and unnecessary energy consumption. Thus, we propose a machine learning model-based on Data Prediction Method (MLM-DPM) to solve these problems. The proposed work is simple yet efficient in terms of processing and needs a small memory footprint in SN. The proposed approach reduces the data transmission rates while maintaining data accuracy. The proposed work is estimated on real dataset attained from the Life Under Your Feet (LUYF) project and compared to two recent Data Prediction Methods. The simulation results were promising and justify the proposed claims.

Keywords: Data prediction method · Energy efficiency · Machine learning · Transmission suppression · Wireless sensor network

1 Introduction

In WSN, the continuous monitoring applications sense measurements at a very high rate. Continuous data transmissions by SNs at high frequency cause enormous energy investing and thus reduce the network's lifetime. Energy preservation becomes an essential topic in WSN applications as the SNs are generally battery-equipped [1]. As data transmissions at SNs deplete more energy than any other task [2], data transmission reduction advances further attentions for conserving the scare energy resources [3, 4]. Beside aggregation [5], data prediction methods (DPM) conserve scarce energy resources by avoiding unnecessary data transmissions. In DPM, each SN will train the models with the recent sensed measurement and sends the prediction model to the Base station (BS) or sink. Then, the SNs employ the same DPM as the sink employs to predict the future data.

A. Abraham et al. (Eds.): IBICA 2021, LNNS 419, pp. 130–140, 2022.
https://doi.org/10.1007/978-3-030-96299-9_13

In the literature, researches have been performed to perform data prediction in wireless networks to reduce the number of data transmissions and thus increase the network lifetime. It is observed in most of the research works that the numerous models and techniques proposed for data prediction are based on machine learning, deep learning, regression etc. or their combination with other techniques such as feature extraction, selection, filtering models etc. Zhao et al. [6] proposed a linear-regression prediction model based on the periodic nature of the sensor data which helps to obtain more accurate predicted values over popular data set of 54 sensor nodes deployed in the Intel Berkeley Research laboratory. Matos et al. [7] proposed a strategy based on linear regression to predict sensor data on the basis of data collected from other sensors. The proposed strategy also considered the outlier sensor data values. Raza et al. [8] proposed a novel derivate based prediction strategy which is simple and easy to use for predicting sensor data in wireless sensor networks over a 13 million data-points in four real world problems. Avinash et al. [9] proposed a model based on Kalman filter to predict the data of sensor nodes in WSN. Cheng et al. [10] proposed multiple nodes multiple features based bidirectional long short-term memory (LSTM) to extract features for data prediction using neural network approach. Further, Cheng et al. [11] focused on making the complete use of spatial as well as temporal correlation between sensor data to predict sensor data. They introduced one-dimensional Convolutional Neural Network and Bi-LSTM (Bidirectional Long and Short-Term Memory) to obtain features responsible for first prediction which is executed recursively to get final predicted value. Soleymani et al. [12] proposed a hybrid model using of decision tree, auto-regressive moving average and Kalman filter technique to predict sampling data of sensors for reducing transmissions in WSN.

Apart from predicting data at SNs, CHs and sink have also been used for data prediction in wireless networks. Wu et al. [13] applied least-mean-square dual prediction technique over cluster heads to obtain the next value of sensor nodes in WSN. They further tried to optimize the dual prediction technique by minimizing its number of steps required for data prediction. Further, Krishna et al. [14] proposed an autoregressive model of certain order p and applied it to the sensor nodes as well as base stations to predict the next sensor value. The autoregressive model works on the concept that the sensed values use to change slowly and also follow a pattern. Further, in this direction the author proposed an extended regression model in the research work [15] to predict data at sensor node and base station using popular data set of provided by Intel Berkeley Research lab. Further, to establish correlation between senor data values and thus reduce overhead of prediction large number of values, the author proposed a buffer-based linear filter model in [16] for data prediction. Moreover, to synchronize the predicted sensor data the author adopted a two-vector model using Extended-cosine regression approach [17]. The synchronization helps to reduce error due to continuous prediction which in turn reduces the number of transmissions within the network. To reduce the application overhead, the author proposed a light-weight novel approach called as Data Transmission Reduction Method (DTRM) which employ a dual prediction technique between CHs and sink to predict sensor values. The proposed light-weight DTRM is a model which is based on dual prediction and data aggregation [18].

The contributions of the proposed work are as follow:

- In-depth analysis of literature work of various data prediction techniques.
- Design and implementation of Machine Learning Model based Data Prediction Method (MLM-DPM) for WSNs.
- Simulation-based validation of the proposed work (MLM-DPM) on real application dataset along with the two related work to determine the performance attained by the proposed algorithm.

The remaining of this work is structured as follows. Section 2 presents the Machine Learning Model based Data Prediction Method (MLM-DPM). Section 3 discusses the implementation of the MLM-DPM for WSN based real applications. The simulation results and discussion are illustrated in Sect. 4. Finally, Sect. 5 concludes the work with future direction.

2 Machine Learning Model Based on Data Prediction Method (MLM-DPM)

In this section, we present a Machine Learning Model based Data Prediction Method (MLM-DPM) that examines the data collected from the SNs and determine the pattern from it, to predict the future data. An identical DPM is employed at both SNs and at the BS. The SNs and sink will regularly make the prediction of forthcoming observations on the basis of the same historical data. This method allows SN to avoid unnecessary data transmissions to the BS, while the predictions are within threshold or error limit.

The sink assumes that the predicted value of SN reflects the actual value of SN till its value gets updated by SN. The algorithm works in a streamlined manner: as soon as the sink receives SN value it will move in the training phase for detecting the pattern of the environmental attributes, which is examined. As long as variance of that attribute remains constant over time, a trend will be detected. For instance, if the feature i.e., ‘temperature’ declines from 32 °C to 31 °C to 30 °C, then it means that the decreasing trend is 1 °C per timeslot. The SN transmits this trend to the sink along with the “last value” of the learning process once a trend is known. Therefore, not all the sensed data is sent to the sink unless a decisive variation is found based on error limit. The SN and the sink will initiate to predict the next value based on the ‘last value’ sent and the trend found.

Meanwhile, the sensing process at the SN will remain unchanged. The SN sends data to the sink for each newly sensed measurement, which matches the actual data with the predicted data. When any variation is perceived, the SN transmits that value to the SINK as a sign that some variation is found. Then again, the sink will try to perceive the new trend and will repeat the training process.

Though the sink has more powerful memory and is capable of storing huge amount of data, the SNs uses a vector to store the ‘last reading’ of all SNs, and as soon as a shift occurs, it goes back to the vector to estimate the new trend. In any pattern is found it will send it directly to the sink to start the accurate forecasts. The SN will have to train the model again if any pattern is not observed.

3 Implementation of Machine learning Model based Data Prediction Method (MLM-DPM)

This section proposes the implementation of Machine learning Model based Data Prediction Method (MLM-DPM). We present a stepwise algorithm for SNs to learn and update the proposed model. For the sink we present an algorithm to rebuild and predict the data.

3.1 Training of MLM-DPM

In this proposed Machine learning Model based Data Prediction Method (MLM-DPM), only two sensed values are required to build the prediction model and only one value is needed for correction or for updating the model. Initially, the i^{th} SN sends the first two sensed readings r_i^1 and r_i^2 of time instance t_1 and t_2 to the BS. When the sink obtains a "new reading", it saves it with the corresponding time-stamp in its memory. *NR* signifies the "new reading", and t_{NR} is denoted when *NR* was received. Then, the sink will enter the learning phase for *n* consecutivedata to find thetrend *td* in the data. Both the SNsand the sink simultaneously calculates the "trend" by finding the variance between the two readings as follow in Eq. (1)

$$td_1 = r_i^2 - r_i^1 \tag{1}$$

After the variance is calculated, the trend denoted as *td* is used to predict the next data value in the time series. This is done concurrently at the SN and the sink based on of the measurement of the SNs. Because the sensor's data continues to move slowly and smoothly over time, the readings at neighboring time-ticks are generally very close to each-other. Hence, the following Eq. (2) will be used to predict the $\widehat{r_i^k}$ (data reading r_i at time k^{th} timeslot).

$$\widehat{r_i^k} = \widehat{r_i^{k-1}} + td_i \tag{2}$$

When the SN compares the predicted data $\widehat{r_i^k}$ with the actual data r_i^k to estimate the difference, if this difference lies between the threshold or prediction error (ε), then the actual data is flushed, and there will be no data transmission considering as a "successful prediction".

When the sink does not obtain data from any SN during that timeslot, it will assume that its predicted data lies within the threshold. Otherwise, if the predictedvalue surpasses the threshold error (ε), then the predicted value will be flushed and the real measurement is sent to the sink. The sink will update td_k as calculated in Eq. (3). Again, the previous *NR* reading is updated by the new received data r_i^k to retain track for possible future updates.

$$td_k = \frac{r_i^k - NR}{t_k - t_{NR}} \tag{3}$$

Algorithm 1: Model Training

Require: $\varepsilon, (t_1, t_2, \ldots t_n), (r_i^1, r_i^2, \ldots, r_i^n)$
Ensure: (ε, μ, t)
for each SN_i in WSN
    BS will broadcast ε to all SNs
    SN_i will sense r_i^1 and r_i^2
    Transmit r_i^1 and r_i^2 to BS
        $td_1 = r_i^2 - r_i^1$
        $NR = r_i^2$
        $t_{NR} = t_2$
        while $Energy_{SN} \neq 0$ do
            SN will do:
            read r_i^k at time t_k
            $\hat{r}_i^k = \hat{r}_i^{k-1} + \mu\, td_k$
        end while
end for

3.2 Data Prediction and Model Corrections

The sensed data from WSN time-series naturally upsurges, stand-still, and declines, particularly when the sensing attributes change over the time period. Since td_k is a linear rate of variation over a precise time interval. Thus, to synchronize and get the stability in the regression line over the data curve, td increased by the times of "correction rate" denoted as μ where $0 \geq \mu \geq 1$. Hereafter, the predicted values are calculated by Eq. (4) as follow instead of Eq. (2).

$$\widehat{r_i^k} = \widehat{r_i^{k-1}} + \mu td_k \tag{4}$$

The role of "correction factor" is to match the predicted data with the real sensed measurements which is achieved by introducing a penalty in the model, which is defined as correction factor (μ) based on a threshold and response on model accuracy while being unbiased. Unbiased means to find a balance between overestimation and underestimation. Thus, the idea is to reduce the bias in the model as td_k. Again, as discussed, the SN will not transmit the actual data to the sink until the predicted data is within the threshold or error budget (ε). Otherwise, when sink is likely to obtain one data reading, and it does not receive it, it decides that the data computed by the prediction model is correct. Though, if a predicted data $\widehat{r_i^n}$ at any time slip is not correct. In such scenario, the SN has to transmit the sensed measurements r_i^n to the sink to rectify the model instantly. This is achieved by Eq. (5) that supervised the "update phase" as presented below.

$$td_n = \frac{r_i^n - NR}{N} \tag{5}$$

To determine the model's accuracy for the next prediction cycle, the SN and the sink will concurrently estimate the "learning rate" (α) as done in Eq. (6).

$$\alpha = \frac{r_i^n - \widehat{r_i^n}}{F} \tag{6}$$

The prediction frame (F) produced is defined as the number of successful predicts that are estimated before the update. The smaller the value of α, the more exact will be the value of *NR*. If each estimation inside F is within the threshold, the α will be always less than ε. Assuming, α doesn't exceed ε, then both the SN and the sink will compute the percentage P in Eq. (7).

$$P = \frac{\alpha \times 100}{\varepsilon} \tag{7}$$

If $\alpha < 0$, the td must be increased by $P\%$ so that it better fit the data, thus the ε is increased by $P\%$. Otherwise, if $\alpha > 0$, td must be reduced by $P\%$ as stated below in Eq. (8)

$$\mu^{new} = \begin{cases} if\ (\alpha < 0),\ \mu^{new} = \mu + \frac{(P \times \mu)}{100} \\ else \qquad ,\ \mu^{new} = \mu - \frac{(P \times \mu)}{100} \end{cases} \tag{8}$$

Algorithm 2: Data Prediction and Model Corrections

Require: $\varepsilon, (t_1, t_2, \ldots t_n), (r_i^1, r_i^2, \ldots, r_i^n)$
Ensure: $(\varepsilon, \mu, F, \alpha)$
if $|\hat{r}_i^k - r_i^k \geq \varepsilon|$ **then**
Send r_i^k to the BS
$tr_n = \frac{r_i^k - NR}{N}$
$NR = y_i^k$
$\alpha = \frac{r_i^n - \hat{r}_i^n}{F}$
if $|\alpha \geq \varepsilon|$ **then**
$P = \frac{\alpha \times 100}{\varepsilon}$
if $(\alpha < 0)$, **then**
$\mu^{new} = \mu + \frac{(P \times \mu)}{100}$
else
$\mu^{new} = \mu - \frac{(P \times \mu)}{100}$
end if
end if
end if

4 Simulation Results

We conduct simulations on Network Simulator (NS2) to evaluate the performances of the proposed work (MLM-DPM) on real application dataset.

4.1 Simulation Setup

The work is experimented on real dataset obtained from the Life Under Your Feet (LUYF) project [19] which is collecting data from WSN measuring the environmental conditions of the soil. The LUYF project measure following data: soil temperature, soil water pressure, surface temperature, volumetric soil water content, surface moisture, CO_2 flux, light flux, battery voltage, total solar radiation and photo-active radiation. The visualization tools of these projects are grazor, silverKoala and sense Web. For our analysis we have simulated the LUYF project based on the of soil temperatures of four SNs and surface temperatures of ten SNs for the duration of 80 days from the LUYF project. Figure 1 demonstrates the amount of data collected by the LUYF project. The simulation parameters are presented below in Table 1.

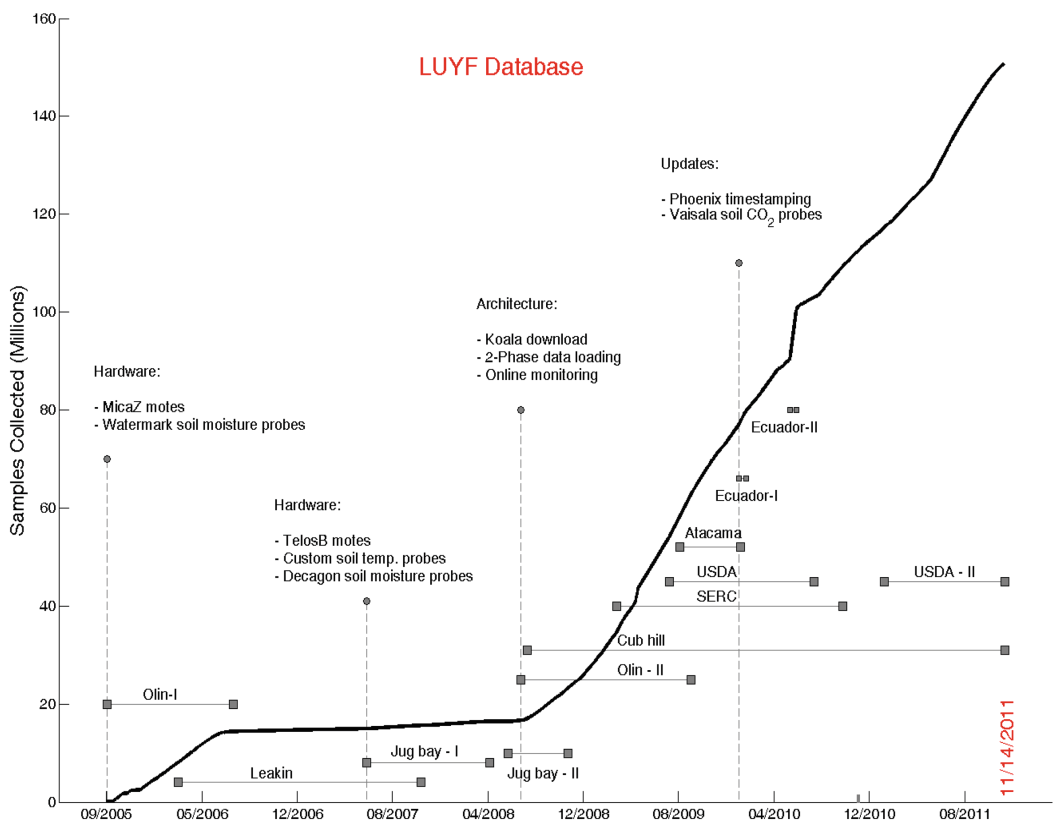

Fig. 1. The amount of data collected by the LUYF project [19] deployments

4.2 Transmission Suppression%

It can be determined by calculating the ratio of the transmitted data by using any prediction algorithm and original transmitted data without using any prediction algorithm.

$$Transmission\ Supression\% = \left(\frac{transmitted\ data\ by\ using\ prediction\ algorithm}{Original\ transmitted\ data\ without\ prediction\ algorithm} \right) \times 100 \tag{9}$$

Table 1. Simulation parameters

Representation	Parameter	Cost
E_0	Initial energy	100 J
ε	Prediction error	0.1 to 0.9 with the step of 0.1
-	Number of Soil Sensors	4
-	Number of Surface Sensors	10
-	Algorithms	MLM-DPM, ELR, P-PDA
-	Data packet	100 bytes
-	Control packet	48 bytes
$\{ro_1, ro_2, \ldots ro_{10}\}$	Data Collection Rounds	10
ε_{mp}	Multi-path fading amplifier energy	0.0013 $(pJ/bit)/m^4$
ε_{fs}	Free space amplifier energy	10 $(pJ/bit)/m^2$
E_{DA}	Aggregation energy	5 $(nJ/bit)/s$
E_{DP}	Prediction energy	5 $(nJ/bit)/s$
E_{RX}	Reception energy	50 $nJ/sfor\ 1 - bit$
E_{TX}	Transmission energy	150 $nJ/sfor\ 1 - bit$, 10 m

Figure 2 illustrates the results of transmission ratio in percentage by executing the MLM-DPM, ELR and P-PDA ten times by varying the threshold levels. The data transmission is reduced drastically to 22% at threshold level 0.1 and goes till 44% even at higher threshold of 0.9%. The data transmission suppression percentage of ELR and P-PDA is between 30% to 55% and 40% to 77% respectively.

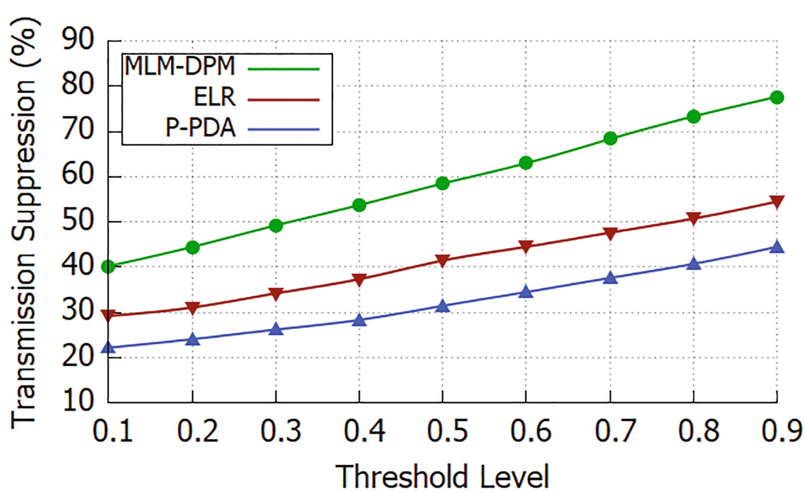

Fig. 2. Transmission Suppression (%) of MLM-DPM, ELR and P-PDA for the threshold levels

4.3 Energy Efficiency

Since the energy consumption is directly proportional to the number of data transmissions performed by the SNs. The transmission data reduction to the sink would suggestively increases the WSN lifetime. The higher the transmission suppression will be, the less data is communicated over the network and less energy will be depleted. The energy model of this work is based on the work done in research [15].

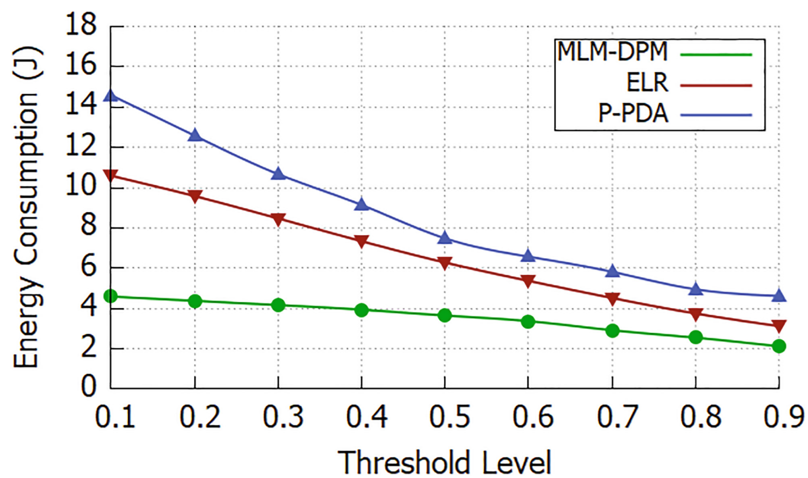

Fig. 3. Energy consumption of MLM-DPM, ELR and P-PDA for the threshold levels

We have compared the energy consumption of our proposed MLM-DPM technique with ELR and P-PDA for the various threshold levels over ten rounds of communication. The results are demonstrated in Fig. 3. The total energy consumption of MLM-DPM is less than 4 J even at the lower threshold level due to high data transmission and approximately 2 J at the higher threshold level. With ELR model the energy consumption is between 3 J to 10 J by varying the threshold level and with P-PDA the energy consumption is between 4.5 J to 14.3 J by varying the level.

4.4 Cost Factor

The accuracy of a good DPM should focus on minimizing the difference between the actual data and the predicted data, while being unbiased. Unbiased means to find a balance between overestimation and underestimation. Therefore, to estimate the performance in terms of the predicted time-series reliability at the BS, the root means square error (RMSE) metric is defined as follow:

$$RMSE = \sqrt{\frac{1}{N} \times \sum_{i=1}^{N} (r_i^n - \widehat{r_i^n})^2} \tag{10}$$

Figure 4 highlights the RMSE for the various threshold levels averaged over ten rounds of data communications. We have compared the RMSE of our MLM-DPM, ELR

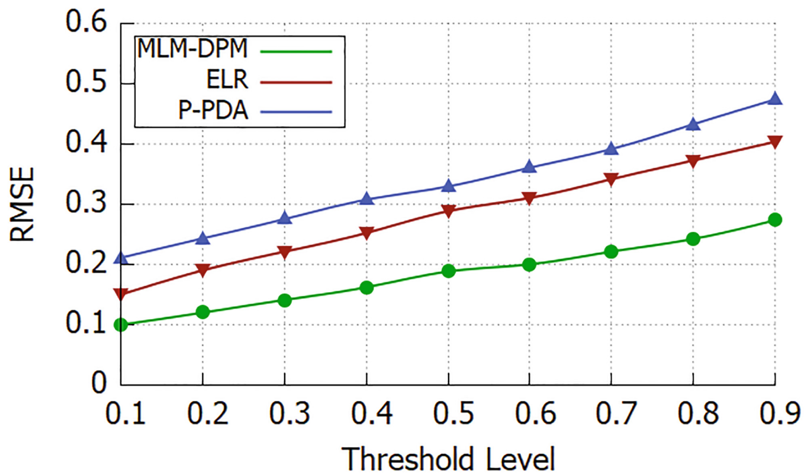

Fig. 4. RMSE of MLM-DPM, ELR and P-PDA for the threshold levels

and P-PDA compared with P-PDA and ELR, MLM-DPM has attained RMSE less than 0.27 which less than 1% even by varying the threshold levels. The MSE of ELR and P-PDA is less than 0.4 and 0.45 respectively, even at the higher threshold limits but it is comparatively higher than the proposed MLM-DPM. Altogether, MLM-DPM delivers much better performance metrics with reasonable error.

5 Conclusion

This paper puts forward a Machine learning based Data Prediction Method applied to a real-world WSNs. To improve energy efficiency by avoiding unnecessary data transmission and maintaining data quality by reducing the cost function, MLM-DPM algorithm is trained based on current data values and is employed to predict the future values as well as to regenerate the historical data. An algorithm based on machine learning based data prediction is proposed for model training in SNs to avoid complex computations and to improve the viability of the proposed work. Another algorithm for data prediction and model correction is proposed for the sink. The simulation results of this proposed MLM-DPM algorithm can (1) provide enhanced energy efficiency, (2) higher transmission suppression ratio and (3) reduced cost function for the real time applications of WSN.

References

1. Agarwal, A., Dev, A., Jain, K.: Prolonging sensor network lifetime by using energy-efficient cluster-based scheduling. Int. J. Sci. Technol. Res. ISSN 2277–8616, **9**(4) (2020)
2. Jain, K., Singh, A.: An empirical cluster head selection and data aggregation scheme for a fault-tolerant sensor network. Int. J. Distrib. Syst. Technol. (IJDST) **12**(3), 27–47 (2021)
3. Gupta, M., Varma, S.: Optimal placement of UAVs of an aerial mesh network in an emergency situation. J. Ambient. Intell. Humaniz. Comput. **12**(1), 343–358 (2020). https://doi.org/10.1007/s12652-020-01976-2

4. Gupta, M., Varma, S.: Optimal placement of UAVs forming aerial mesh networks to handle network issues. Adhoc & Sensor Wireless Networks. **48** (2020)
5. Jain, K., Kumar, A.: Energy-efficient data-aggregation technique for correlated spatial and temporal data in cluster-based sensor networks. Int. J. Bus. Data Commun. Netw. (IJBDCN) **16**(2), 53–68 (2020)
6. Zhao, J., Liu, H., Li, Z., Li, W.: Periodic data prediction algorithm in wireless sensor networks. In: China Conference on Wireless Sensor Networks, pp. 695–701. Springer, Berlin, Heidelberg (October 2012)
7. Matos, T.B., Brayner, A., Maia, J.E.B.: Towards in-network data prediction in wireless sensor networks. In: Proceedings of the 2010 ACM Symposium on Applied Computing, pp. 592–596 (March 2010)
8. Raza, U., Camerra, A., Murphy, A.L., Palpanas, T., Picco, G.P.: Practical data prediction for real-world wireless sensor networks. IEEE Trans. Knowl. Data Eng. **27**(8), 2231–2244 (2015)
9. Avinash, R.A., et al.: Data prediction in wireless sensor networks using Kalman filter. In: 2015 International Conference on Smart Sensors and Systems (IC-SSS), pp. 1–4. IEEE (December 2015)
10. Cheng, H., Xie, Z., Wu, L., Yu, Z., Li, R.: Data prediction model in wireless sensor networks based on bidirectional LSTM. EURASIP J. Wirel. Commun. Netw. **2019**(1), 1–12 (2019). https://doi.org/10.1186/s13638-019-1511-4
11. Cheng, H., Xie, Z., Shi, Y., Xiong, N.: Multi-step data prediction in wireless sensor networks based on one-dimensional CNN and bidirectional LSTM. IEEE Access **7**, 117883–117896 (2019)
12. Soleymani, S.A., et al.: A hybrid prediction model for energy-efficient data collection in wireless sensor networks. Symmetry **12**(12), 2024 (2020)
13. Wu, M., Tan, L., Xiong, N.: Data prediction, compression, and recovery in clustered wireless sensor networks for environmental monitoring applications. Inf. Sci. **329**, 800–818 (2016)
14. Krishna, G., Singh, S.K., Singh, J.P., Kumar, P.: Energy conservation through data prediction in wireless sensor networks. In: Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT), pp. 26–27 (May 2018)
15. Jain, K., Agarwal, A., Kumar, A.: A novel data prediction technique based on correlation for data reduction in sensor networks. In: Proceedings of International Conference on Artificial Intelligence and Applications, pp. 595–606. Springer, Singapore (2021)
16. Agarwal, A., Jain, K., Dev, A.: Modeling and analysis of data prediction technique based on linear regression model (DP-LRM) for cluster-based sensor networks. Int. J. Ambient Comput. Intell. (IJACI). **12**(4) (2021)
17. Jain, K., Kumar, A.: An energy-efficient prediction model for data aggregation in sensor network. J. Ambient. Intell. Humaniz. Comput. **11**(11), 5205–5216 (2020). https://doi.org/10.1007/s12652-020-01833-2
18. Jain, K., Kumar, A.: A lightweight data transmission reduction method based on a dual prediction technique for sensor networks. Trans. Emerg. Telecommun. Technol. e4345 (2021)
19. Dataset (Online): Life Under Your Feet (LUYF) project Retrieved March 11. From http://lifeunderyourfeet.org/en/src/ (2021)