IEEE Sensors Letters VOL. 1, NO. 6, DECEMBER 2017

Sensor networks

Performance Evaluation of Real-Time Multivariate Data Reduction Models for Adaptive-Threshold in Wireless Sensor Networks

N. A. M. Alduais[1*], Jiwa Abdullah[1**], Ansar Jamil[1], and Hadi Heidari[2†]

[1] Wireless and Radio Science Centre, Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Batu Pahat 86400, Malaysia

[2] School of Engineering, University of Glasgow, Glasgow G12 8QQ, U.K.

[*] Student Member, IEEE

[**] Member, IEEE

[†] Senior Member, IEEE

Manuscript received September 8, 2017; revised October 9, 2017; accepted October 25, 2017. Date of publication November 2, 2017; date of current version December 1, 2017.

Abstract—This article presents a new metric to assess the performance of different multivariate data reduction models in wireless sensor networks. The proposed metric is called updating frequency metric that is defined as the frequency of updating the model reference parameters during data collection. A method for estimating the error threshold value during the training phase is also suggested. The proposed threshold of error is used to update the model reference parameters when it is necessary. Numerical analysis and simulation results show that the proposed metric validates its effectiveness in the performance of multivariate data reduction models in terms of the sensor node energy consumption. The adaptive threshold improves the frequency of updating the parameters by 80% and 52%, in comparison to the nonadaptive threshold for multivariate data reduction models of MLR-B and PCA-B, respectively.

Index Terms—Sensor networks, internet of things, wireless sensor networks, multivariate data reduction, performance metric, threshold.

I. INTRODUCTION

In Wireless Sensor Networks (WSN)/Internet of Things (IoT), sensor data consist of either one attribute (univariate) or multiple attributes (multivariate) [1]. As the sensor board is aimed to collect merely one kind of data (light/temperature or humidity), this type of data is called univariate data [2]. Similarly, in some of IoT/WSN applications, each sensor board is equipped with multivariate sensors to support different requirements of applications. For example, IoT *Libelium Gases* sensor board supports multivariate sensors for measuring a few data such as humidity, temperature, and carbon dioxide at the same time [3].

Theoretically, energy efficiency of sensor board is influenced by the process of packet transmission from the sensor board to the gateway and its packet size. The energy consumed in sending one bit via sensor board is higher than running many microcontroller instructions [4]. Thus, Principal Component Analysis (PCA), Multiple/Simple Linear Regression (MLR), and other time series-based approaches are used as data reduction models for WSN to achieve low power consumption in sending the bit. For example, in recent work by Tan and Wu [5], a method to reduce the number of sensor node transmitted packets by applying the hierarchical Least-Mean-Square (HLMS) adaptive filter was presented. In prior works [6], the authors presented fast and efficient dual-forecasting method to reduce the number of sending messages by the sensor board. In [5] and [6], there is only univariate data with fixed threshold error investigated. In recent work [7], the authors proposed a new method based on forecasting to reduce the number of transmitted packets. The advantage of the proposed model is that the work could test the proposed model using vibration sensors datasets. However, it only addresses the univariate data. Therefore, this study focuses on the multivariate data which has high correlation.

Data reduction models with multidimensional sensors are presented in [8]–[10], where the authors applied MLR and PCA based models, respectively. It is noted that the original PCA approach is not suitable for real-time implementation on the sensor board level that has limited resource as the PCA has to learn new PCs for any change in the phenomenon by repeating complex matrix operations involved in singular value decomposition (SVD) operations [11]. Therefore, a lightweight version from PCA called as Candid Covariance-free Incremental PCA (CCIPCA) is proposed in [12]. In prior work [11], the authors used CCIPCA for reducing the multivariate data in WSN with fixed threshold and large size of training data. However, the accuracy of the data reduction models that is dependence on training decreases over time due to the increment in the approximation error. The retraining process aims to update the reference parameters to represent the new dynamic changes in the sensed data [11]. The increment in approximation error of the model during the real-time data collection is one of the significant challenges. The standard solution to this issue is accomplished by applying an adaptive model so that it is able to update its reference parameters during data collection. However, the act of increasing the frequency of the update of the model reference parameters will affect the efficiency of the sensor board energy.

Most of the current models have yet determine to determine the appropriate threshold for updating the common global model because of the dynamic nature of data variation [2]. The detailed explanation about the type of the threshold will be covered in the latter subsection. This challenge will increase when dealing with multivariate data type. The selection of threshold effects the model accuracy and frequency of the model

Corresponding authors: N. A. M. Alduais (e-mail: naifalduais@gmail.com).
Associate Editor: F. Falcone.
Digital Object Identifier 10.1109/LSENS.2017.2768218

1949-307X © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

updates, especially for the IoT-based WSN applications which have been developed to collect the sensed data in an unlimited period. Therefore, a new adaptive-threshold for data reduction models with multivariate data is proposed in this work.

II. MOTIVATIONS

The motivation to use UFM as new metric is the size of transmitted data after updating the model reference parameters which is larger or equal to the payload data size without reduction. It means that the sensor board requires more energy in updating stage than the reduction stage. This work is the first study that uses the UFM as metric to evaluate the real-time data reduction models in IoT/WSN.

This letter proposes the calculation of the model threshold during the training phase. The motivation for that is the minimum residual errors between the training data and approximated data occurred during the training phase. In univariate data, it is simple to use maximum absolute error or minimum least squares during the training phase as the threshold will later be used in the reduction phase. However, estimating the threshold value is a difficult in the multivariate data reduction models. Therefore, maximum relative approximation error in all attributes is suggested as a threshold to avoid the employment of different thresholds in the same sensor board. The advantages of the proposed threshold are *(1)* The threshold value will be estimated by the model itself without any human intervention at the sensor board. It reduces the human-dependency of the edge device and it is suitable for working in the smart environment. *(2)* The mechanism used to calculate the threshold is more accurate and suitable for the multivariate data. *(3)* The proposed threshold is adaptive such that the value of the threshold changes during data collection.

III. NUMBER OF UPDATE MODEL REFERENCE PARAMETERS

The number of updating models is affected by the type of mechanisms used to re-calculate the model reference parameters during data collection. In this article, the mechanism of updating models classifies into 3 categories: update model based on *(i)* window size; *(ii)* non-adaptive threshold; and *(iii)* adaptive threshold.

A. Update Model Based on Window Size

In this scenario, regardless of the approximation error, the model merely re-calculates its reference parameters when the number of sensed data samples is equal to the fixed window size. Window size w is entirely dependent on the application and it is selected by the sink. This study focuses on the update of the model when its approximation error increases. The numerical analysis is stated in the latter subsection to prove the effect of UFM on the energy consumption.

B. Update Model Based on Non-Adaptive and Adaptive Threshold

In this scenario, the model updates its reference parameter when the approximation error is larger than the specified threshold value. In this case, the threshold can be a fixed value selected by the sink. The threshold calculation during data collection may be adaptive or non-adaptive. The UFM values in the case of the non- adaptive threshold is larger than the adaptive one. The reason for that, the model based on non-adaptive threshold is entirely dependent on the value of threshold that has been calculated during the training phase and is used in reduction phase with no change in the value of that threshold. Furthermore, the probability that the value of the threshold to be small for the first time. In this case, the model will still be retrained as the dynamic data will change in most of the cases leading to the production of error that is larger than the threshold. Conversely, the adaptive threshold changes its value every time the reference parameters need updating.

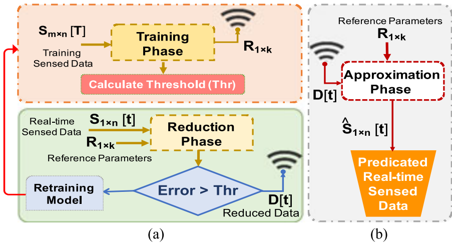

Fig. 1. General structure of multivariate data reduction model with adaptive threshold. (a) Sensor board level. (b) CH/BS level.

IV. MULTIVARIATE DATA REDUCTION MODELS WITH ADAPTIVE THRESHOLD

Fig. 1 shows general structure of the proposed adaptive threshold for multivariate data reduction models. It consists of three crucial phases including training phase, reduction phase at sensor board level and approximation stage at the sink level.

A. Estimate Reference Parameters/Approximation Data

In this article, the models based on PCA /MLR are mentioned because the proposed threshold has the potential to benefit different versions of PCA/MLR. However, for more clarity in this part, there are a few particulars versions of PCA and MLR will be discussed in this study. Due to limited resources of the sensor board, a lightweight version of PCA model called CCIPCA was used. It is explained in detail in [12], together with the steps of using CCIPCA in WSN as described in [11]. Furthermore, only 50 samples (training data) from 5000 samples used in this study for both models which is actually too small compared to training data have used in [11], where was about 700 samples (training data) from 1000 samples used.

B. Estimate the Threshold Value in Training Phase

Steps for calculating the proposed threshold are described in the following Pseudo Code.

C. Update Model During Reduction Phase

In this phase, the real-time sensed data is reduced by applying the multivariate data reduction model. The model should update its reference when the relative error is larger than the threshold value which has been estimated in the training phase. Additionally, the threshold value is adjusted during retraining/updating phase based on new reference parameters. In evaluating, the algorithm includes a counter C to account for the frequency of the model re-training. The following pseudo code describes the updating stage of the model.

Estimate approximation data/Reference parameters (PCA-based)

1) Standardises the training data $\bar{S}_{mXn}[T]$. **2)** Implements CCIPCA ($\bar{S}_{mXn}[T]$) and estimates the eigenvector matrix $R_{n\times n}$. Then, reduces the eigenvector matrix to $\boldsymbol{R_{PC\times n}}$. $R_{PC\times n}$ which is the reference parameters is produced. It is then saved at the sensor node and transmit one copy of the parameter to the sink. **3)** Standardises the new real-time sensed data $\bar{S}_{1\times n}[t]$, then reduces it before transmitting by applying (1).

$$D_{1\times pC}[t] = \bar{S}_{1\times n}[t] \times R_{PC\times n} \quad (1)$$

4) Sends the reduced data $\boldsymbol{D_{1\times pC}[t]}$ to the sink. **5)** Estimates the approximation data at the senor node/sink by applying (2).

$$\hat{S}_{1\times n}[t] = D_{1\times pC}[t] \times R_{PC\times n} \quad (2)$$

Estimate approximation data/Reference parameters (MLR-based)

1) After carefully studied the correlation between the multiple sensors on the same sensor board, the independence sensor s_i and dependence sensor s_h are selected. Ambient temperature is selected as dependence sensor s_h because it has the highest correlation with the surface temperature and relative humidity.
2) Calculates the reference parameters by applying (3) and (4). where $\bar{S}_i$, $\bar{S}_h$ are the average values for the variables S_i and S_h, respectively. In training phase, $\forall\, S_i,\ S_h \in R^{1\times m}, i \neq h,$ h *constant and* $i = 1, 2.., n$. $h = 1$ is the sensor index in the sensed data row $S_{1\times n}[t]$

$$\beta_{i,1} = \frac{\sum_{j=1}^{m}\left(S_{h,j}-\bar{S}_h\right)\left(S_{h,j}-\bar{S}_i\right)}{\sum_{j=1}^{m}\left(S_{h,j}-\bar{S}_h\right)^2} \quad (3)$$

$$\beta_{i,0} = \bar{S}_i - \left(\beta_{i1} \times \bar{S}_h\right) \quad (4)$$

Thus reference parameters are generated as $R_{(n-1)\times 2} = [\beta_{i,0}\ \beta_{i,1}]$, $i \neq h$, $i = 1, 2.., n$. **3)** Saves reference parameters in sensor boards and sends a copy of parameters to the sink.
4) Then, sends the reduced data $\boldsymbol{s_h}$ to the sink. **5)** Estimates the approximation data at senor node/sink by applying (5) $\forall\, s_i, \beta_{i,0}, \beta_{i,1}, s_h \in R^{1\times 1}$, $i \neq h$, h *constant*

$$s_i = \beta_{i,0} + \beta_{i,1} \times s_h \quad (5)$$

//Threshold estimation in Training Phase//

1 **Input**: Training data $S_{mXn}[T]$, where n is the number of sensors and m is the number of collected samples in a specific period of time [T].
2 **Estimates** the reference parameters where $\boldsymbol{R_{1\times k}}$ **//** k is the number of reference parameters.
3 **Calculates** the relative error between the training data $S_{m\times n}[T]$ and approximated data $\widehat{S}_{m\times n}[T]$ which is defined in (6).

$$\boldsymbol{E_{j\times i}[T]} = \frac{\left|\widehat{\mathbf{S}}_{\mathbf{j\times i}}[\mathbf{T}] - \mathbf{S}_{\mathbf{j\times i}}[\mathbf{T}]\right|}{\mathbf{S}_{\mathbf{j\times i}}[\mathbf{T}]} \quad \boldsymbol{i = 1, 2..n; j = 1, 2 \ldots, m} \quad (6)$$

4 **Estimates** Threshold (Thr) by selecting maximum relative error value for all sensors of the same board.
$\boldsymbol{Thr \leftarrow Max|E_{m\times n}[T]|}$ //

//Update model in Reduction Phase//

1) Reads new real-time sensed data $S_{1\times n}[t]$ at current time t
2) Calculates the approximated data $\widehat{S}_{1\times n}[t]$ by applying the reduction model with its reference parameters $R_{1\times k}$
3) Determines the model error at current time [t] as stated in (7)

$$\boldsymbol{E_{1\times i}[t]} = \frac{\left|\widehat{\mathbf{S}}_{1\times\mathbf{i}}[\mathbf{t}] - \mathbf{S}_{1\times\mathbf{i}}[\mathbf{t}]\right|}{\mathbf{S}_{1\times\mathbf{i}}[\mathbf{t}]} \quad \boldsymbol{i = 1, 2..n} \quad (7)$$

4 If $\boldsymbol{Max|E_{1\times n}[t]|} >$ **Thr,** then **Update** model; C = C + 1;
5 Calculates the Threshold (Thr) // Call *Training phase*
6 ELSE: Sends the reduced data; **End If** Go to Step 1

//Update model in Reduction Phase//

1) Receives new reduced data at current time t 2) **Estimates** the approximation data at sensor node/Sink by applying (2) for PCA***-based*** model **/** (5) for ***MLR-based*** model.
Note: 1) The approximation data at the sensor board is determined to calculate the model relative approximation error. 2) The approximation data at the sink is determined to reconstruct the original data.

D. Approximation Phase

V. PERFORMANCE EVALUATION

In this article, the multivariate data reduction models of PCA and MLR were applied to evaluate the proposed threshold and new performance metric. MATLAB software was used to simulate the data reduction models with adaptive and non-adaptive threshold using a real-time dataset called Lausanne Urban Canopy Experiment dataset (LUCE) [13]. LUCE is classified as a dynamic dataset, and it includes ambient temperature, surface temperature, and relative humidity.

A. Numerical Analysis for Different Multivariate Data Reduction Models With Fixed Buffer Size

A sensor board with multiple sensors transmitted N = 50000 samples during a time interval. The sensor board applied PCA and MLR models separately for each interval where the reduction ratio R% for PCA -1PC, PCA -2PC and MLR are 67%, 33% and 67% respectively. The model updated its reference parameters when the buffer size was set as W = {50 and 100}, $n = 3$, $E_{Byte} = 52.92\ \mu J$, $E_{bit} = E_{Byte}/8$, $S_d = n \times 32$ bits and S_R is 12 bytes, 24 bytes, and 16 bytes for PCA -1PC, PCA -2PC and MLR, respectively. Where the number of parameters for MLR is 4, and the number of reference parameters for PCA is 3 in case of 1PC and 6 in case 2PC. Table 1 shows list of symbols used. The total energy consumption during a specific

Table 1. List of Symbols Used.

Symbols	Description
n_R, n_T	the number of message transmissions in the reduction phase the number of UFM
S_d, S_R	the size of the original sensed data and of model reference parameters
R%	the model data reduction ratio
$E_{R,} E_T$	the cost of energy consumption for transmission of data in the reduction phase and retraining phase, respectively.
E	the total energy consumption during a specific period.
$\mathrm{E_{Byte}}, E_{\mathrm{bit}}$	the energy consumption per Byte and bits, respectively.
PC	the number of Principal Components (PC) for PCA

Table 2. Comparison Energy consumption.

W	Model	$E_R\,(uj) \times n_R$	$E_T(uj) \times n_T$	E(uj)
50	**PCA-1 PC**	23322432	1186368	**24508800**
	PCA-2 PC	11487168	1655232	**13142400**
	MLR	23322432	1423168	**24745600**
100	**PCA-1 PC**	23560416	593184	**24153600**
	PCA-2 PC	11604384	827616	**12432000**
	MLR	23560416	711584	**24272000**

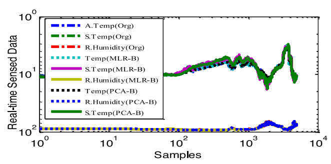

Fig. 2. Predicted data at sink vs. real-time sensed data.

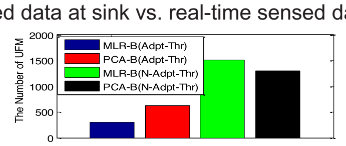

Fig. 3. Frequency of model updating with adaptive and non-adaptive thresholds for different models.

period is defined in (8):

$$\mathbf{E} = (E_R \times n_R) + (E_T \times n_T),\ n_R = N - n_T \tag{8}$$

$$E_R = (S_d \times \mathrm{R\%}) \times E_{bit} \tag{9}$$

$$E_T = ((S_d \times \mathrm{R\%}) + (S_R \times 8)) \times E_{\mathrm{bit}} \tag{10}$$

Table 2 shows the results of applying (8) for the above example. From the results, it is clear that the increment of the frequency of retraining/updating will negatively affect the energy consumption of the sensor board. It is because the size of the transmitted data after updating the model reference parameters is larger or equal to the payload data size without any reduction as defined in (10) and (9), which means that the edge device requires more energy in the transmission phase than the reduction stage. In this scenario (fixed window), UFM value is equal to (N/W). Thus, the UFM were 1000 and 500 for W = 50 and W = 100 respectively. The UFM can be reduced by selecting large value of W. But it is not a feasible solution due to the limited resource of the node.

B. Results and Discussion

Figs. 2–4 show the results of simulation for the PCA and MLR models with adaptive and non-adaptive threshold. It is clear that the adaptive threshold has managed to reduce the

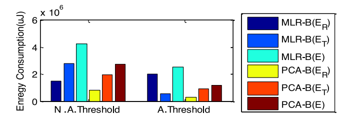

Fig. 4. Energy consumption of the sensor board (uJ).

frequency of model updating its reference parameters by 80% and 52%, which is better than the one with non- adaptive threshold for multivariate data reduction models MLR-B and PCA-B, respectively. The power consumption of the model by applying adaptive threshold is found to be less than the non-adaptive threshold. Based on the results, it is concluded that frequency of model updating is crucial in evaluating the multivariate data reduction models.

VI. CONCLUSION AND FUTURE WORK

Results show that the proposed metric validates its effectiveness in the performance of multivariate data reduction models in terms of the sensor node energy consumption. The adaptive threshold improves the frequency of updating the parameters by 80% and 52% in comparison to the non-adaptive threshold for multivariate data reduction models of MLR-B and PCA-B respectively. The proposed metric and threshold were tested using the environmental data. This study is recommended to be the future work test for the same model that employs multivariate vibration data.

ACKNOWLEDGMENT

This work was supported by the Fundamental Research Grant Scheme (FRGS) vote number 1532 from the Ministry of Education Malaysia.

REFERENCES

[1] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Comunm. Surveys Tut.*, vol. 12, no. 2, pp. 159–170, Apr.–Jun. 2010.

[2] M. A. Rassam, A. Zainal, and M. A. Maarof, "Advancements of data anomaly detection research in Wireless Sensor Networks: A survey and open issues," *Sensors*, vol. 13, no. 8, pp. 10087–10122, 2013.

[3] N. A. M. Alduais, J. Abdullah, A. Jamil, and L. Audah, "An efficient data collection and dissemination for IOT based WSN," in *Proc. IEEE Inf. Technol., Electron. Mobile Commun. Conf.*, 2016, pp. 1–6.

[4] K. A. Bispo *et al.*, "A semantic solution for saving energy in wireless sensor networks," in *Proc. IEEE Symp. Comput. Commun.*, 2012, pp. 000492–000499.

[5] L. Tan and M. Wu, "Data reduction in wireless sensor networks: A hierarchical LMS prediction approach," *IEEE Sensors J.*, vol. 16, no. 6, pp. 1708–1715, Mar. 2016.

[6] F. Strakosch and F. Derbel, "Fast and efficient dual-forecasting algorithm for wireless sensor networks," in *Proc. Sensor,* 2015, pp 859–863.

[7] I. B. Arbi, F. Derbel, and F. Strakosch, "Forecasting methods to reduce energy consumption in WSN," in *Proc. IEEE Instrum. Meas. Technol. Conf.*, 2017, pp. 1–6.

[8] C. Carvalho, D. G. Gomes, N. Agoulmine, and J. N. de Souza, "Improving prediction accuracy for WSN data reduction by applying multivariate spatio-temporal correlation," *Sensors*, vol. 11, no. 11, pp. 10010–10037, 2011.

[9] M. A. Rassam and A. Zainal, "Principal component analysis-based data reduction model for wireless sensor networks," *Int. J. Ad Hoc Ubiquitous Comput.*, vol. 18, no. 1/2, pp. 85–101, 2015.

[10] U. Jaimini, T. Banerjee, W. Romine, K. Thirunarayan, A. Sheth and M. Kalra, "Investigation of an indoor air quality sensor for asthma management in children," *IEEE Sensors Lett.*, vol. 1, no. 2, pp. 1–4, Apr. 2017.

[11] M. A. Rassam, A. Zainal, and M. A. Maarof, "An adaptive and efficient dimension reduction model for multivariate wireless sensor networks applications," *Appl. Soft Comput.*, vol. 13, no. 4, pp. 1978–1996, 2013.

[12] J. Weng, Y. Zhang, and W.-S. Hwang, "Candid covariance-free incremental principal component analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 8, pp. 1034–1040, Aug. 2003.

[13] "Lausanne urban canopy experiment," 2007. [Online]. Available: http://lcav.epfl.ch/Cms/lang/en/ pid/86035