Multivariate Reduction in Wireless Sensor Networks

Orlando Silva Junior[1], Andre L. L. Aquino[2], Raquel A. F. Mini[1], Carlos M. S. Figueiredo[3]

[1] Department of Computer Science - Pontifical Catholic University of Minas Gerais
Belo Horizonte, MG, Brazil
[2] Department of Computer Science - Federal University of Ouro Preto
Ouro Preto, MG, Brazil
[3] FUCAPI – Research and Technological Innovation Center
Manaus, AM, Brazil
Email: `orlandosj@gmail.com`, `alla@iceb.ufop.br`,
`raquelmini@pucminas.br`, `mauricio@fucapi.com.br`

Abstract

In wireless sensor networks, energy consumption is generally associated with the amount of sent data once communication is the activity of the network that consumes more energy. This work proposes an algorithm based on "Principal Component Analysis" to perform multivariate data reduction. It is considered air quality monitoring scenario as case study. The results show that, using the proposed technique, we can reduce the data sent preserving its representativeness. Moreover, we show that the energy consumption and delay are reduced proportionally to the amount of reduced data.

1 Introduction

A wireless sensor network (WSN) [1,3] is a special kind of ad-hoc network with capacity to collect, process and send data to extern observers. The WSNs has some restrictions, such as energy and bandwidth. Thus, to send large amount of data can be problematic, causing excessive delay and diminishing the network lifetime. Due to these restrictions, it is necessary to adopt strategies to reduce the amount of data that are transmitted in the network.

Considering the phenomena characteristics, it is important to distinguish the sensor data as univariate or multivariate. Univariate data represents a sample of same phenomena variable. Therefore, multivariate data represents samples of different phenomena variables. These samples are originated from: different sensors of a specific node; or the same sensor of different nodes.

A usual technique to process multivariate data is the Principal Component Analysis (PCA) [5]. Thus, this work proposes an algorithm for multivariate data reduction in WSNs and it uses air quality monitoring scenario as case study. The technique uses PCA to classify data, so that only the most relevant samples are propagated to the sink. In addition, through the reduction, it is possible to diminish the energy consumption and the delay in the network, since communication is the task that consumes more energy.

Some related works consider univariate data reduction and they use techniques such as data aggregation [10], adaptive sampling [7], or sensor stream reduction [2]. Considering multivariate data reduction, there are some proposals that consider discrete wavelet transformation, hierarchical clustering, sampling and singular value decomposition techniques [8]. Specifically, the reduction based on PCA, we can find some contributions that apply PCA with prediction to improve the reduction [6]. Meanwhile, to provide the reduction in WSNs, it is necessary to evaluate some parameters in more details, such as the reduced data quality, energy consumption and delay. Differently of cited works, all of these aspects are focused and evaluated in this paper.

This paper is organized as follow. Section 2, we discuss the problem of multivariate reduction in WSNs. The proposed solution is presented in Section 3. The evaluation of the reduced data representativeness is shown in Section 4 and the network behavior is presented in Section 5. Finally, Section 6 presents the conclusions and future directions of this work.

978-1-4244-4671-1/09/$25.00 ©2009 IEEE

2 Problem Definition

The multivariate reduction problem in WSNs can be stated as follows: "*Considering an application for air quality monitoring generating multivariate data, is it possible to use a WSN infrastructure which reduces data based on PCA maintaining the data representativeness and reducing energy consumption and delay on the network?*"

To address this problem, the scope of this work consider the following assumptions:

- **Sensing moment reduction**: The application needs to reduce the data only in sensing moment. In this case, there are sensors array devices which collect, simultaneously, different organic compounds. Such device, considered as sensor node, can reduce the multivariate data after different environment samplings, avoiding unnecessary data traffic on the network.
- **Maximum reduction supported**: In this case, we need to identify what is the maximum level of reduction supported in the air quality monitoring application where the data representativeness is not compromised. To identify the maximum reduction, we set empirically values for the reduction in $n/2$ and $\log n$, where n is the number of data collected by each sensor. It is important to emphasize that when considering other applications different values should be evaluated.
- **Data representativeness**: Each application has its own requirements for quality, and so, for each application, different evaluation metrics can be employed. Considering application for the monitoring of air quality, we used hypotheses test and relative absolute error [4] since such test is suitable for this application.

Specifically, to data representativeness was used the hypothesis test Analysis of Variance (*ANOVA*), calculated through statistical program R[1]. The calculation is given by $F = D_B^2/D_W^2$, where D_B^2 represents spread between sets and D_W^2 the spread within the joints. Based on this calculation, the $p-value$ is used to determine if the null hypothesis H_0 must be accepted or rejected. In this case, to accept the null hypothesis indicates that there is no significant difference between the variances of the two sets. By convention, $\mathbf{\Phi}$ will be used to indicate the use of this test.

The absolute relative error considers a comparison between the averages of original and reduced data. This error is given by $\mathbf{\Upsilon} = 100\,Max\{|(\overline{X} - \overline{Y})/\overline{X}|\}$, where $\overline{X}$ and $\overline{Y}$ are the average of the original values and the reduced values, respectively. The $\mathbf{\Upsilon}$-error is calculated for each sensor and only the highest of them, situation where the technique is the worst, will be used.

3 Multivariate Reduction

This section presents the PCA-based algorithm used to reduce multivariate data in air quality monitoring applications. The main goal is generating a new data collection keeping the original data set characteristics with minimal loss of information.

In our sample algorithm, initially, a data classification is made based on the first principal component obtained through PCA. This classification groups the data considering the biggest, smallest, and intermediate values of the first principal component. Our sample algorithm uses only the biggest values classified because the pollution levels identified in air monitoring applications, generally, consider the components positively different to identify anomalous behaviors that indicate a higher pollutants concentration.

In order to illustrate multivariate generated data in this application, consider the matrix X_n^m the input data, where $n > 0$ represents the values monitored by each sensor and $m \geq 1$ represents the sensors responsible for obtaining environmental information. Thus, to reduce X, it is possible to consider three steps. The original set of sensory data X is used to calculate the principal components C in step 1. In step 2, first component C_1 is selected and sorted. The positions of biggest values in C_1, representing data positively different from X are used to determine positions of the lines in X that will compose the reduced data Y. Reduced data set Y containing the lines of X more representative for the application is obtained in step 3.

The pseudo-code is shown in Algorithm 1. In line 1, we have the calculation of the principal components. The complexity order of PCA calculation can be estimated in $O(m^2m'+m^2n)$, where m refers to original data dimension (number of sensors), m' is the reduced data dimension and n is the amount of generated data. According to [9], if $m > m'$ and $m > n$, the complexity order can be estimated in $O(m^2)$. As in this case $m = m'$ and $m < n$, we have $O(m^2n)$.

Algorithm 1 Multivariate reduction

```
Require: $X$ – input data, $r$ – reduction size
Ensure: $Y$ – reduced data

1: $C \leftarrow calculatePca(X)$
2: $I \leftarrow sort(C_1)$ /* $C_1$ is the first component of $C$ */
3: $I \leftarrow sort(I, r)$ /* $I$ are the more relevant values of $C_1$ */
4: for $i \leftarrow 1$ to $r$ do
5:     $Y_i \leftarrow X_{I_i}$
6: end for
```

In line 2, first component C_1 is sorted, where the index I of the more relevant values are obtained. Its complexity order is $O(n \log n)$, since $|C_1| = n$. Line 3 has the sort of vector I, considering only the r first index, to maintain the arrival order of the items chosen for Y. Complexity order

[1] *The R Project for Statistical Computing* – http://www.r-project.org/

of sort is $O(r\ \log r)$. Lines 4 – 6 built the reduced output data, whose complexity order is $O(r)$.

4 Data Representativeness Evaluation

Data quality evaluation considers air quality synthetic data sets to represent the situations in which the reduction is used in sensing phase. Simulations were performed through algorithms implemented in the statistical program *R*. We use for data representation the multivariate distributions normal and skew-normal. Thus, each scenario was executed with 93 different data sets. For each one, the results of $\mathbf{\Phi}$ and $\mathbf{\Upsilon}$, described in the Section 2, are analyzed.

The normal and skew-normal data generation was done from five media, chosen to simulate sensing of five different variables. The used values for the averages were $10, 30, 50, 70, 90$, with a standard deviation of 10%. Generated data size was varied in $n = \{256, 512,\ 1024, 2048\}$ and we applied the reductions $n/2$ and $\log n$.

The first analysis of data representativeness consider $\mathbf{\Phi}$, that represents the ANOVA test. The results indicate that there are no significant differences between the variances of the original data and the reduced data. As shown in Table 1, the the lowest value was equal to 0.69 and show a high level of significance, since values above 0.05 are satisfactory to accept the hypothesis.

Table 1. Analysis of variance (*p-value*)

Distribution	**($n = 256$)**		**($n = 512$)**		**($n = 1024$)**		**($n = 2048$)**	
	$n/2$	$\log n$	$n/2$	$\log n$	$n/2$	$\log n$	$n/2$	$\log n$
normal	0.89	0.85	0.83	0.87	0.76	0.86	0.72	0.84
skew-normal	0.88	0.86	0.84	0.85	0.80	0.84	0.69	0.85

The second analysis, $\mathbf{\Upsilon}$-error, represents the relative absolute error illustrated in Figure 1. In this case, we consider the monitoring realized by 5 sensors. Results show that the $\mathbf{\Upsilon}$-error, in the worst case (reduction $\log n$), the largest error was approximately 6% with the skew-normal distribution, due to the small number of details that were transmitted. These erros can be explained by the fact of only one sensor monitors each variable and thus there is no replication of data from different sensors. However, an important observation is that when the amount of generated data is increased, the technique presents practically the same performance, which demonstrates its scalability.

5 Network Behavior

It is known that communication consumes a lot of energy in WSNs. Thus, reducing the amount of transmitted data also reduces the energy consumption. The network behavior study shows the benefits of reducing the amount of data in terms of energy consumption and delay to deliver data to sink. It is important to emphasize that in the simulations we evaluated only the criteria identified as relevant to investigate the network behavior.

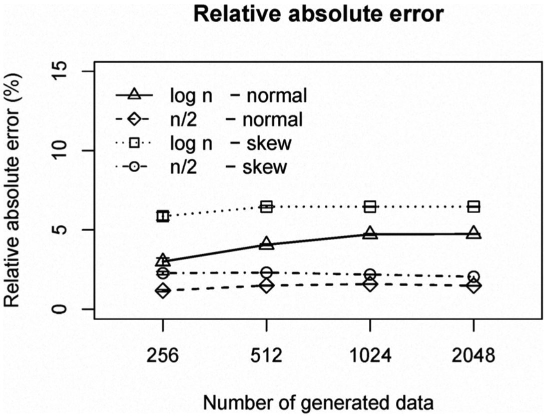

Figure 1. $\mathbf{\Upsilon}$-error with reduction in sensing

Regarding the simulation, the network behavior evaluation is performed through the Network Simulator 2 version 2.33. The simulation was executed with 33 random topologies and the results are presented with symmetrical asymptotic confidence interval of 95%.

Considering the network topology, we used a flat network that uses a routing algorithm based on tree of smaller way and all nodes have the same hardware configuration. The trees are built only once before the traffic begins, and the source nodes are randomly distributed in the air quality sensory region.

The network size varies with density and is obtained through $net_t = \sqrt{\pi\, a_r^2\, |S|/8,4791}$, where a_r is the radio range and S the sensors number. The size of the queue supported by each node varies with the amount of sensory data. The time for simulation is 1100s and the rates of traffic of 500s and 600s. The radio range is 50m, the bandwidth is 250 Kbps and the initial energy is 100J.

In this evaluation, we varied data size in $n = \{256, 512, 1024, 2048\}$, used a fixed number of sensors $m = 5$. The number of nodes on the network was set at 256 and only one source node is used. Again, to evaluate the performance of the algorithm it was used the reductions $n/2$ and $\log n$. Figure 2 shows the delay observed. As expected, delay diminishes significantly when we reduce the amount of transmitted data.

The same behavior was observed for energy consumption. As shown in Figure 3, when we reduce the amount of data, the energy consumption also diminishes considerably.

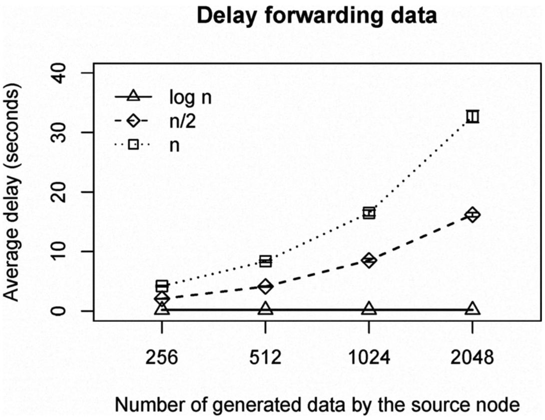

Figure 2. Evaluation of delay forwarding data

Moreover, the rates of the reduced data quality analyzed together with these characteristics emphasize still more the efficiency of the proposed solution.

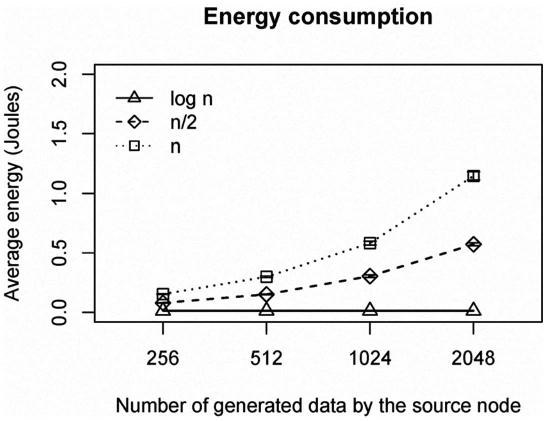

Figure 3. Evaluation of energy consumption

6 Conclusion and Future Work

In this work, we presented an algorithm that uses PCA to reduce the amount of data traffic in WSNs. Results show that such algorithm performs well when considering air quality monitoring applications. In all simulated scenarios, the observed errors were low, proving the feasibility of the solution to maintain data representativeness. Considering energy consumption and delay, the solution proposed also presented good results. As expected, by reducing the amount of traffic in the network, energy consumption and delay also diminish considerably.

As future work, we intend to apply the proposed method to process sensed data at leader node and through the routing task. Furthermore we intend to improve the evaluation of the reduction impact in the network behavior. We also plan to evaluate other solutions of multivariate data classification.

Acknowledges

This work is partially supported by the Brazilian National Council for Scientific and Technological Development (CNPq) under the grant number 477292/2008-9.

References

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci. A survey on sensor networks. *IEEE Communications Magazine*, 40(8):102–114, August 2002.

[2] A. L. L. Aquino, C. M. S. Figueiredo, E. F. Nakamura, L. S. Buriol, A. A. F. Loureiro, A. O. Fernandes, and C. N. C. Junior. Data stream based algorithms for wireless sensor network applications. In *21st IEEE International Conference on Advanced Information Networking and Applications (AINA'07)*, pages 869–876, Niagara Falls, Canada, May 2007. IEEE Computer Society.

[3] A. Boukerche. *Algorithms and Protocols for Wireless Sensor Networks*. Wiley-IEEE Press, 2008.

[4] A. C. Frery, H. Ramos, J. Alencar-Neto, and E. Nakamura. Error estimation in wireless sensor networks. In *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*, pages 1923–1928, New York, NY, USA, 2008. ACM.

[5] J. E. Jackson. *A User's Guide to Principal Components*. Wiley-Interscience, 1 edition, 2003.

[6] J. Li and Y. Zhang. Interactive sensor network data retrieval and management using principal components analysis transform. *Smart Materials and Structures*, 15(11):1747–1757, December 2006.

[7] S. Santini and K. Romer. An adaptive strategy for quality-based data reduction in wireless sensor networks. In *3rd International Conference on Networked Sensing Systems (INSS'06)*, volume 1 of *1*, pages 29–36, Chicago, IL, USA, 31 May – 2 June 2006. On-line.

[8] S. Seo, J. Kang, and K. H. Ryu. Multivariate stream data reduction in sensor network applications. In *2nd International Symposium on Ubiquitous Intelligence and Smart Worlds (UISW'05)*, volume 1 of *1*, pages 198–207, Nagasaki, Japan, December 2005. Springer.

[9] A. Sharma and K. K. Paliwal. Fast principal component analysis using fixed-point algorithm. *Pattern Recognition Letters*, 28(10):1151–1155, July 2007.

[10] J. Zhu and S. Papavassiliou. A resource adaptive information gathering approach in sensor networks. In *IEEE Sarnoff Symposium on Advances in Wired and Wireless Communication (SARNOFF'04)*, volume 1 of *1*, pages 115–118, Princeton, NJ, USA, April 2004. IEEE Computer Society.