Data reduction in sensor networks: performance evaluation in a real environment

Abdallah Makhoul and Hassan Harb

***Abstract*—Data reduction is an effective technique for energy saving in wireless sensor networks. It consists on reducing sensing and transmitting data while conserving a high quality of collected information. In this paper we propose an online data reduction model based on Kruskal-Wallis test that allows sensor nodes to adapt their sensing rates based on the data variance. Then, we propose a local aggregation algorithm to reduce further the data set size before sending to the sink. Experimentation on real telosB sensor network testbed shows the effectiveness of our approach in reducing the size of data transmitted over the network and thus saving energy.**

***Keywords*—Wireless sensor networks, data reduction, adaptation and aggregation, Kruskal-Wallis test, telosB mote, real experiments.**

I. Introduction

Wireless sensor networks (WSN) present a low cost solution to enhance our lives. Their main advantages are fast, easy deployment and low maintenance cost [1]. Indeed, data reduction is one of the most efficient ways to reduce energy consumption in WSNs. It consists on reducing the amount of data sensed and transmitted to the sink.

In the literature, one can find various data reduction approaches based on in-network processing, data compression or data prediction methods. Adaptive filtering techniques were proposed in [2], [3], [4], [5]. They are based mainly on algorithms like least mean square [3], [5] and Kalman Filter [2], [4]. These works focus on coordinating adaptive filters at the sensor node and prediction techniques at the sink. Each node stops send data when the error between the filter input and the filter output is within a specified threshold. Then, the filters are ready to predict the future data. Other data prediction approaches are also studied to conserve energy in WSN [6], [7], [8], [9], [10]. It means to predict future information with the use of various algorithms and prevent transmitting the raw data. Stochastic approaches [6], time-Series forecasting [8] and heuristics and algorithms [9], [10] are used. Moreover data compression can be applied in sensor networks. It reduces the size of data transmitted in the network by involving encoding at nodes and decoding at the sink [11], [12], [13]. Although these approaches predict sensed values and allow efficient data reduction, however they present several disadvantages. They are almost complex, sometimes they generate communication overhead, and the sink may need some transmissions to detect failures.

In this paper, we present two phases data reduction method. It is less complex and suitable for limited resources sensor nodes. The first phase uses Kruskal-Wallis test, instead of Bartlett test used in [14], and aims to reduce the data acquisition on sensors by adapting their sensing rates to the varying nature of the sensed data. Indeed, Bartlett test has a serious weakness if the data normality assumption is not met. Consequently, in this paper we study the Kruskal-Wallis test which does not impose the data normality assumption. The second phase of our approach consists on aggregating collected data online based on some similarity properties before sending them to the sink. To evaluate our technique, we conducted several experiments on a real environment sensor networks based on telosB nodes. In our experiments, we compared our results to those obtained with Bartlett test [14] and a data compression method, called S-LEC, proposed recently in [11].

II. Data acquisition reduction

In this section, we introduce the first phase, i.e. data acquisition phase, which is based on the Kruskal-Wallis test. The idea here is to apply the Kruskal-Wallis test in order to verify if there is high variation in the collected measurements. In the affirmative case the sensing rate must be at its maximum else we adapt the sampling rate according to the variation and to the situation risk as explained in [14].

Subsequently, the Kruskal-Wallis H test is a rank-based nonparametric test that can be used to determine if there are statistically significant differences between two or more groups of data [15]. We propose to present the Kruskal-Wallis test via the following example.

A. Illustrative Example

Consider the readings taken in three periods p_1, p_2 and p_3 as shown in the following table.

	p_1	r_1	p_2	r_2	p_3	r_3	Total
	8.2 10.3 9.1 12.6 11.4 13.2	1 8 $3 \rightarrow 4$ 10 9 11	10.2 9.1 13.9 14.5 9.1	7 $4 \rightarrow 4$ 14 15 $5 \rightarrow 4$	13.5 8.4 9.6 13.8	12 2 6 13	
n_i	6		5		4		$N = 15$
$Total$		43		44		33	

TABLE I. Readings example

In order to apply the Kruskal-Wallis test, we must first order the readings in all periods by increasing order of their values. The order of each measure represents its rank, denoted r where

A. Makhoul is with FEMTO-ST Institute/CNRS, the DISC department, Univ. Bourgogne Franche-Comté, Belfort, France, e-mail: abdallah.makhoul@univ-fcomte.fr.

H. Harb is with the department of Computer Science, American University of Culture and Education (AUCE), Nabatiyeh/Tyre, Lebanon, e-mail: hassanharb@auce.edu.lb.

$r \in [1, N]$. Then, we assign for each measure its rank as shown in Table I. Subsequently, in case where a measure is repeated several times, called "tied", the mean value of their rank is assigned to the measure. Then, consider the following two hypothesis:

- H_0: the three probabilities distributions are the same.
- H_1: the three probabilities distributions are not the same with some desired false-rejection probability (risk α).

Then, the Kruskal-Wallis test statistic is given by:

$$H = \frac{12}{N \times (N+1)} \sum \frac{r_i^2}{n_i} - 3 \times (N+1) \qquad (1)$$

where:

- N is the total number of measures in all periods.
- n_i is the number of measures during a period.
- r_i is the rank of each period.

[1]

Thus, based on Equation 1, H is calculated as follows:

$$\begin{aligned} H &= \frac{12}{15 \times (15+1)} \left(\frac{43^2}{6} + \frac{44^2}{5} + \frac{33^2}{4} \right) - 3 \times (15+1) \\ &= 0.3805 \end{aligned}$$

Hence, for $\alpha = 0.05$ we have $H_t = 5.991$, we have $H < H_t$. Thus, the null hypothesis (H_0) is accepted and the sensor sampling rate should be adapted.

Algorithm 1 describes the adaptive sampling rate algorithm at the sensor node based on the Kruskal-Wallis test. For each round, every node decides to increase or decrease its sampling rate according to the difference between its collected measures and the application risk level. First, the node computes the rank for each measure (lines 7-8). Then, it uses the Behavior function [14] to adapt its sensing rate only if the calculated difference between measures is less than the Kruskal-Wallis threshold (lines 10-14). As explained in [14], we define the application risk level and we express this level by a quantitative variable R which can take values between 0 and 1 representing the low and the high risk level respectively.

III. Sensors data aggregation

A. Definitions and Notations

In periodic applications, each period p is divided into $\mathcal{T}$ equal time slots where, at each slot, a sensor S_i captures a new reading r_{i_j}, then, it forms a vector of readings during the period p as follows: R_i=$\left[r_{i_1}, r_{i_2}, \ldots, r_{i_\mathcal{T}}\right]$.

Mostly, R_i may contain redundant (or very similar) readings, especially when the monitored condition varies slowly or when the slots are short. In order to eliminate similar values from the vector R_i, we define $Similar$ function as follows:

Definition 1: Similar function. We define the $Similar$ function between two readings as:

$$Similar(r_{i_j}, r_{i_k}) = \left\{ \begin{array}{lll} 1 & \text{if} & |r_{i_j} - r_{i_k}| = 0, \\ 1 & \text{if} & |r_{i_j} - r_{i_k}| \leq \delta, \\ 0 & \text{otherwise.} & \end{array} \right\}$$

Algorithm 1 Adaptive Sensing Rate Algorithm.

```
Require: $p$ (1 round = $p$ periods), $\tau$: period size, $R$: application criticality, $\alpha$: false-rejection probability.
Ensure: $S_t$ (instantaneous sampling speed).
1:  $S_t \leftarrow \tau \ measures/period$
2:  while $Energy > 0$ do
3:      for $i = 1 \rightarrow p$ do
4:          takes measures at $S_t$ speed
5:      end for
6:      for each round do
7:          sort measures by increasing order of their values.
8:          compute the rank of each measure
9:          find $H_t$
10:         if $H \leq H_t$ then
11:             $S_t \leftarrow BV(H, H_t, R, \tau)$ (BV behavior function).
12:         else
13:             $S_t \leftarrow \tau \ measures/period$
14:         end if
15:     end for
16:     $M_i' \leftarrow local\_Aggregation(M_i, \delta)$ // $M_i$ is the set of measures collected at the current period
17:     $send\ (M_i')$
18: end while
```

where r_{i_j} and $r_{i_k} \in R_i$ and δ is a threshold determined by the application. Furthermore, two readings are considered similar if and only if their $Similar$ function is equal to 1. This means that two measures are considered redundant if they are equals or similar.

Then, we define the weight of a reading as follows:

Definition 2: Reading's weight, $wgt(r_{i_j})$. The weight of a reading r_{i_j} is defined as the number of similar readings (according to $Similar$ function) to r_{i_j} in the same vector R_i.

B. Aggregation Phase Algorithm

Algorithm 2 presents the aggregation process which is running by the sensors themselves at each period. For each new captured reading, S_i searches for similarities of the new taken reading. If a similar reading is found, the new one is deleted and the corresponding weight is incremented by 1, else the sensor adds the new reading to the set and initializes its weight to 1. Consequently, S_i will possess a set of readings/weights, e.g. R_i'=$\{(r_{i_1}', wgt(r_{i_1}')), (r_{i_2}', wgt(r_{i_2}')), \ldots, (r_{i_k}', wgt(r_{i_k}'))\}$ where $k \leq \mathcal{T}$, which will be sent to the sink.

IV. Experimental Results

To evaluate our proposal, we deployed four Crossbow telosB motes, collecting temperature and humidity measures in our laboratory for about three days (Fig. 1). The proposed approach (Kruskal-Wallis test), the Bartlett test [14], the S-LEC data compression technique [11] and the naïve approach[2], e.g. without adapting data collection, are implemented on motes

[1] the numbers'12' and '3' are fixed by the test and do not depend on the number of periods.

[2] naïve approach has been implemented in order to calculate the data loss measures in the two compared tests.

Algorithm 2 Local Aggregation Algorithm.

```
Require: new reading $r_i$, $\delta$: similarity threshold.
Ensure: set of readings with their weights: $R'_i$.
for each existing reading $r_j \in R'_i$ do
    if $Similar(r_i, r_j) = 1$ then
        $wgt(r_j) \leftarrow wgt(r_j) + 1$
        disregard $r_i$
    else
        $wgt(r_i) \leftarrow 1$
        $R'_i \leftarrow R'_i \cup \{(r_i, wgt(r_i))\}$
    end if
end for
```

Ids=1, 2, 3 and 4 respectively. Every 30 seconds, each mote collects new measure and sends to a fifth mote, called *collector* with Id=0, connected to a laptop machine. The *collector* only relays the data received from the motes to the laptop machine which acts as the sink node. Then, we implemented a Java application on the laptop machine in order to perform daily statistics over the data sent from the motes. In Table II, we show the metrics used in our experiments.

Parameter	Description	Value
$round$	round size	2 periods
τ	period size	50 measures
δ	similarity threshold	0.05 for temperature 0.01 for humidity
R	application criticality	0.6
α	false-rejection probability	0.01 in day 1 0.025 in day 2 0.05 in day 3

TABLE II. EXPERIMENTAL ENVIRONMENT.

A. Data collected reduction size study

Fig. 2 shows a comparison between the number of temperature and humidity measures sent daily by the motes. The obtained results show that Kruskal-Wallis test can reduce up to 19% and 26% of temperature and humidity measures sent with S-LEC technique. Whilst, comparison between Kruskal-Wallis and Bartlett tests shows that for a small value of α, e.g. in days 1 and 2, the mote sends fewer number of measures (temperature and humidity) using Bartlett test, otherwise, data collected are more reduced using Kruskal-Wallis test. This means that Kruskal-Wallis is more flexible than Bartlett to reject the null hypothesis when the false-rejection probability increases.

B. Adapting sensor sensing rate

Fig. 3 shows how the temperature and humidity sensors are able to adapt their sampling rate using the Kruskal-Wallis test. The obtained results show that the sampling rate of each sensor is dynamically adapted after each round. This confirms the decrease in the volume of the collected data compared to the naïve approach, e.g. $S_t = 50$ measures/period. We can also notice that the sampling rate of the temperature sensor has been more adapted compared to the humidity sensor.

Fig. 5 shows an illustrative example for which data are collected by the temperature sensor using Kruskal-Wallis test. Compared to naïve collection, the results show that our adapting algorithm allows each sensor to collect most important measures in the round after adapting its sampling rate.

C. Energy consumption study

In nesC programming, there is no model provided by tinyOS in order to measure the energy consumed in telosB [16]. In our experiments, we used the same radio model used in [17] as the most used model to evaluate the energy consumption in WSNs. Since the energy consumption is highly related to the amount of data collected/sent (see Fig. 2), the obtained results show that: **a)** Kruskal-Wallis test conserves the energy of the mote up to 44% compared to S-LEC. **b)** for α equals to 0.01 and 0.025, the Bartlett test consumes about 12% less energy compared to the Kruskal-Wallis test in the first two days. **c)** For $\alpha = 0.05$ in the third day, Kruskal-Wallis test reduces the energy consumption in the mote about 6% compared to Bartlett test.

D. Data accuracy/integrity

Fig. 6 shows the percentage of data loss using the three approaches. The percentage of data loss is calculated by searching the measures collected in the naïve approach (mote Id=1) and do not collected (nor similar values) by the other motes. The obtained results are dependent on the number of data sent to the sink (see Fig. 2); thus, more data are sent less measures are lost. Indeed, we observe that the percentage of data loss using kruskal-Wallis test does not accessed 3.1% whereas it arrives to 6.4% and 4.3% using S-LEC and Bartlett test. Therefore, we consider that Kruskal-Wallis test is more efficient in terms of saving data integrity compared to other approaches.

E. Further Discussion

In this section, we make more discussion about the three compared approaches then, we give some directions as to which approach is more suitable for a desired application. Regarding the energy preserving viewpoint, Kruskal-Wallis and Bartlett tests significantly reduce the energy consumption in motes, compared to S-LEC approach. Furthermore, comparison both tests shows that Bartlett conserves more energy when α is small (e.g. ≤ 0.025) whereas, the energy is more preserved using Kruskal-Wallis for a higher value of α (e.g. > 0.025). Therefore, for applications where we need to conserve the energy of the network, the decision makers should be used the suitable test depending on the chosen value of α. Regarding the data accuracy viewpoint, Kruskal-Wallis test can save the integrity of the collected data more than those in other approaches in almost cases. Hence, it is suitable to use the Kruskal-Wallis test for the application that does not permit flexibility regarding data accuracy.

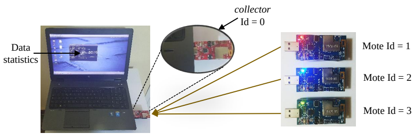

Fig. 1. Sample of TelosB nodes used in the experimentation.

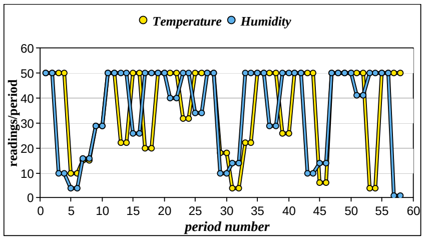

Fig. 3. Instantaneous sampling rate.

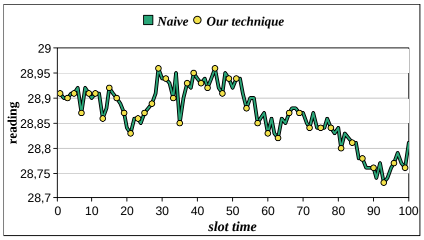

Fig. 5. Example of data collected during a period.

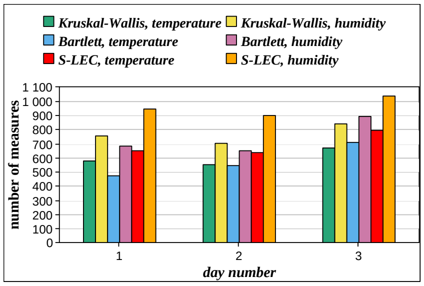

Fig. 2. Number of measures received by the sink.

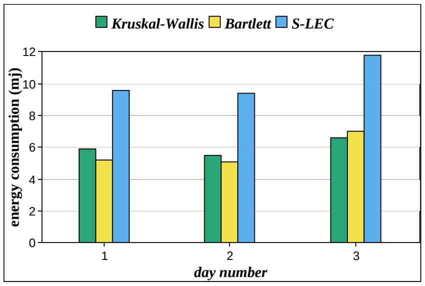

Fig. 4. Energy consumed in each mote.

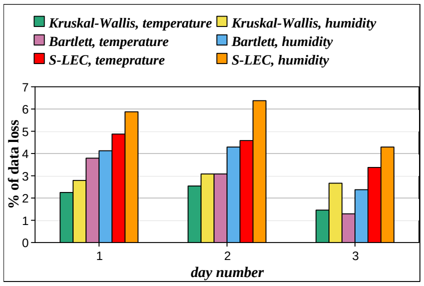

Fig. 6. Percentage of data loss using each technique.

V. Conclusions

In this paper, we proposed energy-efficient data reduction and aggregation techniques dedicated to wireless sensor networks. They were targeted to minimize the amount of data retrieved/communicated by the network without loss in fidelity. We studied a data collection model based on a Kruskal-Wallis test that allows each sensor node to adapt its sensing rate to the changing of the monitored condition. Furthermore, we added a second level of local data aggregation in order to reduce data transmitted in the network and save energy. We showed via real experimentation and telosB testbed that our approach can be effectively used to increase the sensor network lifetime, while still keeping the quality of the collected data high.

As a future work, we seek to extend our adaptive sensing technique in order to take into account the correlation between neighboring sensor nodes.

References

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci., "Wireless sensor networks: a survey." *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–114, August 2002.

[2] B. Q. Ali, N. Pissinou, and K. Makki, "Approximate replication of data using adaptive filters in wireless sensor networks," *2008 3rd International Symposium on Wireless Pervasive Computing*, pp. 365–369, 2008.

[3] A. Jain, E. Y. Chang, and Y.-F. Wang, "Adaptive stream resource management using kalman filters," *Proceedings of the 2004 ACM SIGMOD Int. Conf. on Management of Data*, pp. 11–22, 2004.

[4] R. Abdolee and B. Champagne, "Diffusion lms algorithms for sensor networks over non-ideal inter-sensor wireless channels," *2011 International Conference on Distributed Computing in Sensor Systems and Workshops (DCOSS)*, pp. 1–6, 2011.

[5] M. Stern, K. Böhm, and E. Buchmann, "Processing continuous join queries in sensor networks: A filtering approach," *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, pp. 267–278, 2010.

[6] D. Chu, A. Deshpande, J. M. Hellerstein, and W. Hong, "Approximate data collection in sensor networks using probabilistic models," *Proceedings of the 22Nd International Conference on Data Engineering*, pp. 48–, 2006.

[7] S. Samarah, "Vector-based data prediction model for wireless sensor networks," *Int. J. High Perform. Comput. Netw.*, vol. 9, no. 4, pp. 310–315, Jan. 2016.

[8] D. Tulone and S. Madden, "Paq: Time series forecasting for approximate query answering in sensor networks," *Proceedings of the Third European Conference on Wireless Sensor Networks*, pp. 21–37, 2006.

[9] J. Wang, K. Damevski, and H. Chen, "Sensor data modeling and validating for wireless soil sensor network," *Comput. Electron. Agric.*, vol. 112, no. C, pp. 75–82, Mar. 2015.

[10] A. Sinha and D. K. Lobiyal, "Prediction models for energy efficient data aggregation in wireless sensor network," *Wirel. Pers. Commun.*, vol. 84, no. 2, pp. 1325–1343, Sep. 2015.

[11] Y. Liang and Y. Li, "An efficient and robust data compression algorithm in wireless sensor networks," *IEEE Communications Letters*, vol. 18, no. 3, pp. 439–442, 2014.

[12] E. Zimos, D. Toumpakaris, A. Munteanu, and N. Deligiannis, "Multiterminal source coding with copula regression for wireless sensor networks gathering diverse data," *IEEE Sensors Journal*, vol. 17, no. 1, pp. 139–150, 2017.

[13] J. He, G. Sun, Z. Li, and Y. Zhang, "Compressive data gathering with low-rank constraints for wireless sensor networks," *Signal Processing*, vol. 131, pp. 73–76, 2017.

[14] A. Makhoul, H. Harb, and D. Laiymani, "Residual energy-based adaptive data collection approach for periodic sensor networks," *Ad Hoc Netw.*, vol. 35, no. C, pp. 149–160, Dec. 2015.

[15] P. E. McKight and J. Najab, "Kruskal-wallis test," *Corsini Encyclopedia of Psychology*, 2010.

[16] J. Griessen, "http://tinyos-help.10906.n7.nabble.com/energy-consumption-on-telosb-td22083.html," 2012.

[17] S. Madden, "Intel berkeley research lab," in *http://db.csail.mit.edu/labdata/labdata.html*, 2004.