Information Sciences 329 (2016) 800–818

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# Data prediction, compression, and recovery in clustered wireless sensor networks for environmental monitoring applications

Mou Wu [a], Liansheng Tan [a], Naixue Xiong [b,c,*]

[a] Department of Computer Science, Central China Normal University, Wuhan 430079, PR China
[b] School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China
[c] Department of Business and Computer Science, Southwestern Oklahoma State University, OK, USA

ARTICLE INFO

*Article history:*
Received 12 April 2014
Revised 15 September 2015
Accepted 2 October 2015
Available online 9 October 2015

*Keywords:*
Optimal step size
Least mean square (LMS)
Principal component analysis (PCA)
Data prediction
Environmental monitoring
Wireless sensor network (WSN)

ABSTRACT

Environmental monitoring is one of the most important applications of wireless sensor networks (WSNs), which usually requires a lifetime of several months, or even years. However, the inherent restriction of energy carried within the battery of sensor nodes brings an extreme difficulty to obtain a satisfactory network lifetime, which becomes a bottleneck in scale of such applications in WSNs. In this paper, we propose a novel framework with dedicated combination of data prediction, compression, and recovery to simultaneously achieve accuracy and efficiency of the data processing in clustered WSNs. The main aim of the framework is to reduce the communication cost while guaranteeing the data processing and data prediction accuracy. In this framework, data prediction is achieved by implementing the Least Mean Square (LMS) dual prediction algorithm with optimal step size by minimizing the mean-square derivation (MSD), in a way that the cluster heads (CHs) can obtain a good approximation of the real data from the sensor nodes. On this basis, a centralized Principal Component Analysis (PCA) technique is utilized to perform the compression and recovery for the predicted data on the CHs and the sink, separately in order to save the communication cost and to eliminate the spatial redundancy of the sensed data about environment. All errors generated in these processes are finally evaluated theoretically, which come out to be controllable. Based on the theoretical analysis, we design a number of algorithms for implementation. Simulation results by using the real world data demonstrate that our framework provides a cost-effective solution to such as environmental monitoring applications in cluster based WSNs.



## 1. Introduction

Recent climate change and natural disasters in the world have suggested the importance of the environmental monitoring, which is subsequently rapidly developing as a major application of wireless sensor networks (WSNs) [1–4]. For instance, within a WSN sensor nodes can be used in a harsh environment to periodically measure meteorological and hydrological parameters around its surroundings, such as light, temperature, humidity, wind speed and direction. Especially by using the advanced

* Corresponding author. Tel.: +1 4046454067.
*E-mail addresses:* mou.wu@163.com (M. Wu), l.tan@mail.ccnu.edu.cn (L. Tan), xiongnaixue@gmail.com, neal.xiong@swosu.edu (N. Xiong).

http://dx.doi.org/10.1016/j.ins.2015.10.004
0020-0255/

wireless communication and sensor technology, WSNs have advantages in applications over other traditional networks, on the aspects of such as withstanding ability, clustering for scalability, and self-organization properties [5–7].

Although WSNs provide several benefits in the field of environmental monitoring, energy conservation should always be taken into account in almost all application areas. The main reason is that sensors in such environment are impossible to be recharged or replaced, that means the energy is a limited amount. Therefore, energy saving becomes one of the major design concerns in WSNs, and some energy-efficient schemes are proposed to reach the goal of energy saving from all aspects [8–13]. A sensor node is typically deployed with sensing, computing and wireless communicating modules, in which the communication module consumes the most electricity [5,14]. Moreover, in the context of continuous monitoring, the most of data changes at a slow speed, which results in a large amount of data redundancy in space or time, subsequently frequent communications between sensor nodes will be a waste of limited energy. Basically, the increase of network lifetime will be proportional to the reduction in the number of transmitted data packets. Following this principle, data reduction has become one of the most enhanced solutions that is aimed to reduce the amount of data transmissions[15–18].

The most efficient way to obtain data reduction in WSN is data prediction that uses the prediction values instead of the real ones, thereby avoiding the data transmission. In a real-world scenario, it is often unnecessary and yet costly to obtain the precise measurements for each sample period. Data prediction techniques focus on minimizing the number of transmitted measurements from the sensor nodes during continuous monitoring process. However, one key concern is to ensure the accuracy of the prediction within a user-given error bound. For the periodical sensing applications especially environmental monitoring, each consecutive observation of a sensor node is temporally correlated to a certain degree. In our prediction model, the temporal correlation is exploited to perform the prediction of data for the monitoring application based on the user-defined error tolerance. The result of using this correlation-based approach is a dual prediction protocol that has a remarkable effect on reducing the frequency of data transmissions in a way that guarantees the prediction accuracy.

One alternative approach to realize data reduction is using compressing techniques [19,20] that lead a reduction in the amount of transmitted data because the size of data is reduced. In general, we can classify the data compression schemes into two categories: lossless and lossy compression. Lossless data compression demands the original data to be perfectly reconstructed from the compressed data. By contrast, lossy data compression allows some features of the original data that may be lost after the decompression operation. For highly resource constrained WSN, lossless algorithms are usually not necessary despite the fact that they have better performance on data recoverability. To put it the other way, lossy compression is better able to reduce the amount of data to be sent over the WSN. In the case of lossy compression, the amount of compression and the reconstruction error are the important criterions to judge the quality of compression algorithms. Our work using the Principal Component Analysis (PCA) method to compress the original data is proved to be able to obtain satisfactory results in two ways. More importantly, the error generated by the PCA compression is negligible compared to the prediction error, which ensures the user's acceptable total error bound.

In order to obtain the energy-efficient scheme for continuous environmental monitoring, we develop in the present paper a novel framework with delicate combination of data prediction, compression, and recovery in cluster based WSNs. The main idea of the framework is to reduce the communication cost through data prediction and compression techniques whilst the accuracy is guaranteed. First, sensor nodes collecting environmental parameters are grouped into multiple clusters based on their physical locations. At the same time, a dual prediction mechanism using LMS prediction algorithm with optimal step size is implemented at sensor nodes and their respective CHs, which not only improves the prediction accuracy, but also achieves faster convergence speed during the initial stage of algorithm. Then the CHs extract the principal component of collected data by the PCA techniques after a sampling period, so redundant data can be prevented. Finally, data is successfully recovered at the base station (BS). Throughout the entire process, all errors are controllable and kept within the tolerable bound. After achieving data reduction, the size of recovered data at the BS is equal to that of raw sensory data collected by all nodes. It is advantageous for the BS to gain a more in-depth understanding of environment parameters. The simulation results also demonstrate that the combination of LMS prediction algorithm and PCA technique is energy-efficient for environmental monitoring applications in cluster based WSNs.

The remainder of the paper is organized as follows. In Section 2, we discuss the related work on data reduction techniques in WSNs. In Section 3, we describe the dual prediction mechanism between sensor nodes and the CH, where an optimal step size LMS (OSSLMS) prediction algorithm is presented. Section 4 proposes the approach of energy-efficient data compression by PCA technique with data redundancy being lifted. In Section 5, we evaluate the communication cost and analyze the mean square error during data prediction and compression. Simulation results are provided in Section 6 to validate the efficacy and efficiency of the proposed approach. Finally, Section 7 presents the conclusions of the whole paper and future work.

## 2. Related work

Many models have been proposed to perform data prediction in WSN. The auto regressive (AR) model uses the linear regression function embedded in the sink to calculate the estimation of future sensor readings. By regularly collecting local measurements, the sensor node can compute the coefficients of the linear regression based on past real values. These coefficients are then delivered to the sink to perform time series forecasting.

Within the context of AR model, the paper [21] proposed a general framework called PAQ (short for Probabilistic Adaptable Query system) to efficiently answer queries at the sink based on a simple AR model. An adaptive model selection algorithm