
A Data Predication Model for Integrating Wireless Sensor Networks and Cloud Computing

Samer Samarah*

Department of Computer Information Systems, Yarmouk University, Irbid PO Box: 120, Jordan

Abstract

Cloud computing has been proved to be a promising solution for managing and processing big data by providing a data center-centric and efficient algorithms for managing and organizing the data. One of the cloud system’s data sources is Wireless Sensor Networks (WSNs). WSNs present a new way of data-stream sources in which data is received periodically from different sensors; resulting in a large amount of data accumulated over a short period. WSNs have limited resources in which a fine-detailed data streams lead to exhaustive energy consumption. In this paper, we propose a data prediction model that is built within the sensor nodes and used by the cloud system to generate the data. The purpose of the proposed model is to exempt the sensor nodes from sending a large amount of data and thus reduces the energy consumption of the sensor's battery. We manage to formulate the prediction model as a line equation through two n-dimensional vectors in n-space. Initial results showed that the proposed model will be capable to achieve a better error rate as compared to traditional data prediction techniques.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Peer-review under responsibility of the Conference Program Chairs

Keywords: Wireless Sensor Networks; Data Mining; Data Prediction; Cloud Computing Platform

1. Introduction

The growing usage of embedded devices and wireless technologies enabled Wireless Sensor Networks (WSNs) to contribute to variety of applications ranging from military applications, outdoor and indoor environmental

* Corresponding author. Tel.: +962-775-683-294
E-mail address: samers@yu.edu.jo
monitoring, crops monitoring, power monitoring, health monitoring and objects tracking. A wireless sensor node is a small device consists of one or more sensing devices, a microprocessor, a memory, a radio transceiver and one or more of batteries. The main functionality of the sensor node is to gather the data from the surrounding environment which has been measured by the sensing devices, make a local processing such as aggregation, then send the data to the base station through the radio transceiver.

WSNs present a new way for data-stream sources in which data is received periodically from different sensors; resulting in a large amount of data accumulated over a short period. In addition, many applications emerged recently to integrate data from different sources and domains; adding new dimensions for the collected data. With this big data, a new management and organization techniques become a necessity to cope with the complexity that is ever increasing.

Data transmission in WSN consumes a lot of energy. It considers the main factor of energy consumption comparing with processing and sensing. In order to reduce energy consumption, and prolong the life time of WSNs, one of the suitable solution is reducing the amount of the data transmitted. This issue needs suitable technique to deal with. In recent years, prediction models have been introduced with WSNs as a solution to reduce the amount of data needed to be transmitted to the Sink. These models differ mainly in the technique used, the amount of data needed to build the model, and in the prediction error produced.

Cloud computing has been proven to be a promising solution for managing and processing big data by providing a data center centric and efficient algorithms for managing and organizing the data. Many frameworks have proposed to integrate WSN and cloud computing in which the data is transmitted from the sensor network to the cloud.

Integrating WSN and cloud computing will add extra load on the sensor nodes. Data centric centers look for collecting fine-detailed data about the domain area, and thus the sensor nodes need to operate under a low sample rate that will affect negatively on the energy consumption of the sensor nodes. A solution is needed for such situation in which the data collected at the finest level while minimizing the energy consumption of sensor nodes. Data Mining, as a step in the knowledge discovery process, can be a useful technique to generalize a summarized data.

The main goal of the data mining techniques is to discover patterns among the data. Particularly, we will consider prediction technique, one of the data mining techniques that aims to construct a model to predict the future states based on the current reading. The prediction model is used to predict the future readings of sensor nodes and provide the information to the user without the need to communicate with sensor nodes and thus prolong the lifetime of the nodes. Several predication models have been proposed to improve the performance of the WSNs through reducing the amount of the data needed to be sent to the cloud system.

In this paper we propose a prediction model for n-dimensional time series data. The proposed model assumes a linear distribution for the data and capture the correlation among the data by the line equation trough two vectors in n-dimensional space.

This paper is organized as follows: section 2 discusses some related research. Section 3 provides a detailed and formal description of the proposed Prediction technique. Section 4 describes an initial analytical results. Finally, section 5 summarizes the conclusion of this research.

2. Literature Review

Prediction is among the earliest data mining techniques that has been used in wireless sensor networks. Prediction is concerning of building a model to predict the future data states of sensor nodes based on the current data. Two types Prediction can be defined, classification and regression. Classification is used to predict nominal or discrete values, while regression is used to predict continues or order values.

The main goal of regression based prediction algorithm is to build a regression model to predict the future reading of sensors nodes and exempt the nodes form sending their readings to the Sink. Two main paradigms have been proposed for maintaining regression models: 1) Dual predication scheme. In this paradigm, the same prediction model is built at the Sink node and the sensor node. During the operation time, the sensor node will update the Sink with the new value if the sensed value has a significant difference from the predicted value. 2) Single prediction
scheme\textsuperscript{8,9}. In this paradigm, a single model is built at the Sink node, or the sensor node, based on the reading of a subset of sensor nodes.

Olston et al.\textsuperscript{10}, proposed a dual prediction scheme that uses a constant prediction to provide the user with an answer that is guaranteed to be within a bounded interval [L, H]. The bounded interval is computed based on the user’s specified precision constrain \( \alpha \).

Dual Kalman Filter (DKF) is a regression-based data gathering framework proposed by Jan et al.\textsuperscript{11}. DKF uses "Kalman filter", a well known linear prediction technique\textsuperscript{12}, as the main technique to predict the future readings of the sensor nodes. In this framework, each sensor node and the Sink node run an identical version of the Kalman filter that has been built using the data sensed at sensor node and transmitted to the Sink, as well as, other parameters describing the environment under monitoring.

PREMON (PREdiction-based MONitoring) is a regression based prediction technique inspired by the encoding techniques used in video streaming\textsuperscript{13}. In this approach, sensor nodes and the Sink use TDMA based communication and it is assumed that sensors’ readings at each TDMA cycle form the intensity values of pixels in an image. This perspective allows the Sink to use the encoding technique used in MPEG2 to develop a prediction model to predict the future reading of sensor nodes.

The prediction subset is a different approach for regression based prediction proposed by Gianluce and Yan-Ale\textsuperscript{14}. In this approach, a subset of sensors is identified that is capable of predicting the readings of all the sensor nodes in the network. Different prediction subsets are defined to be used in turn to distribute the energy consumption.

ASAP (Adaptive Sampling Approach) is a regression based approach to energy-efficient periodic data collection in wireless sensor networks proposed by Gedik et al.\textsuperscript{15}. The main idea behind ASAP is to periodically identify a subset of nodes as sampler nodes such that all the reading of the sampler nodes are sent to the Sink, while the reading of non-sampler nodes are predicted using probabilistic model that is built locally in the network.

3. Methodology

In this section, we highlight the main steps required for building the prediction model. We start by providing a formal definition for the required terminologies through Section 3.1 followed by a detailed explanation for the steps taken for building the prediction model.

3.1. Problem Formulation

Let S be a sensor node in a particular WSN. Time is assumed to be divided into a set of equal-intervals windows-also called Time Periods (P). \( P = \{ p_1, p_2, \ldots, p_t \} \) represents the time periods over the sensor's data stream.

In each time period, the sensor collects a set of measures, depending on the sample rate of the sensor device, we refer to \( \mathbf{V}_i = (v_1, v_2, \ldots, v_n) \) as a vector of real numbers that represents the sensor readings over a period \( p_i \).

**Definition 3.1.1.**

\( \text{AVG}_V \) is defined to be the average vector among the Time Periods' vectors.

\[
\text{AVG}_V = \left( \frac{\sum_{i=1}^{t} V_i(1)}{t}, \frac{\sum_{i=1}^{t} V_i(2)}{t}, \ldots, \frac{\sum_{i=1}^{t} V_i(n)}{t} \right)
\]

**Definition 3.1.2.**

\( D(\mathbf{V}_i, \text{AVG}_V) \) is defined to be the Euclidean distance between vector \( \mathbf{V}_i \) and the vector \( \text{AVG}_V \).

\[
D(\mathbf{V}_i, \text{AVG}_V) = \sqrt{((V_i(1) - \text{AVG}_V(1))^2 + \cdots + (V_i(n) - \text{AVG}_V(n))^2}
\]

**Definition 3.1.3.**
Max_V is defined to be the vector among the Time Periods' vectors with the maximum Euclidean distance with respect to the vector AVG_V.

**Definition 3.1.4.**

Min_V is defined to be the vector among the Time Periods' vectors with the minimum Euclidean distance with respect to the vector AVG_V.

### 3.2. Model Building

The model building stage, we assume that the sensor node accumulate a set of readings over a given set of time periods P. Table 1 shows an example of the structure of the data collected.

<table>
<thead>
<tr>
<th></th>
<th>v1</th>
<th>v2</th>
<th>............</th>
<th>vn</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>R</td>
<td>R</td>
<td>............</td>
<td>R</td>
</tr>
<tr>
<td>P2</td>
<td>R</td>
<td>R</td>
<td>............</td>
<td>R</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>............</td>
<td>...</td>
</tr>
<tr>
<td>Pn</td>
<td>R</td>
<td>R</td>
<td>............</td>
<td>R</td>
</tr>
</tbody>
</table>

The model building process starts within the sensor by computing the AVG_V vector by averaging the values of all the time periods' vectors. After that, the Euclidean distance between time period's vector and AVG_V vector is computed. Then, the Max_V and Min_V vectors are determined.

### 3.3. Model Usage

The proposed model is based on the idea that all the values of the vectors will be distributed around a line in n dimensional space, n refers to the number of values in any given vector. Based on this assumption, we formulate the line equation between the time period vectors' boundaries (i.e., the Max_V and Min_V). The line equation is given by equation 1.

\[ Y = \text{Max}_V + t (\text{Min}_V - \text{Max}_V) \]  ..........Equ1

The model usage starts by the sensor nodes sending the Max_V and Min_V vectors to the cloud system through the base station. The cloud constructs the prediction model based on equation 1. In data transmission stage, and for each vector V_x, the sensor sends the scalar of the vector (t), which defined as the Euclidian distance between vector V_x and the vector AVG_V divided by the magnitude of the directional vector Min_V - Max_V.

\[ t = \frac{D(V_x, \text{Max}(V))}{|\text{Min}_V - \text{Max}_V|} \]  ..........Equ2
4. Analytical Results

In this section, we present an analytical results for the proposed prediction model. The data set used in the evaluation is data about irrigation collected from a wireless sensor network built at Yarmouk University, Jordan to control the irrigation process for the tomato plants. The data collected from three wireless sensor nodes connected to a base station and remotely accessed by a server like station called datalogger. The data contains 10000 recodes in which each record consists of a set of soil and sensor attributes captured by the sensor nodes (i.e., Volumetric Water Content VWC, Soil Temperature Ts, or Electrical Conductivity EC etc). Figure 1 shows the network architecture.

The proposed prediction model is evaluated with comparison to linear regression model. A subset of the data is used contains 30 readings over a period of 5 hours. The vector size varied from 3 to 5 values. Table 1 shows the results obtained. As shown, the proposed model outperform the linear regression in the error rate. The error rate computed based on Mean Squared Error formula.

Table 1: Mean Squared Error

<table>
<thead>
<tr>
<th>Vector Size</th>
<th>Linear Regression Mean Squared Error</th>
<th>Proposed Prediction Mean Squared Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2.636059</td>
<td>0.001649</td>
</tr>
<tr>
<td>4</td>
<td>0.823195</td>
<td>0.006245</td>
</tr>
<tr>
<td>5</td>
<td>0.23949</td>
<td>0.0116</td>
</tr>
</tbody>
</table>

Figure 1: Network Architecture

As for the message reduction gained using the proposed prediction, Table 2 compares the number of messages sent using the prediction model and the individual transmission of the data.

Table 2: Mean Squared Error

<table>
<thead>
<tr>
<th>Vector size</th>
<th>Individual transmission</th>
<th>Transmission using prediction model</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>21</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>28</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>35</td>
<td>15</td>
</tr>
</tbody>
</table>
5. Conclusion

This paper presented a prediction model for sensor's values. The technique assumed linear distribution of the data and model the readings as a vector in n-dimensional space. The proposed technique is aimed to be used to integrate wireless sensor networks and cloud computing by providing detailed data with minimum energy consumption. The evaluation results showed that the proposed prediction model was able to reduce the number of messages needed to transmit the sensor's readings to the cloud system with an acceptable error rate.

Acknowledgment

This work is supported in part by the USAID-NSF PEER Award (NAS Sub-Grant Award Letter Agreement No. PGA-2000003429 Sponsor Grant Award Number: AID-OAA-A-11-00012).

References