Computers and Electronics in Agriculture 112 (2015) 75–82

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Sensor data modeling and validating for wireless soil sensor network

Ju Wang *, Kostadin Damevski, Hui Chen

Computer Science Department, Virginia State University, United States

ARTICLE INFO

Article history:
Received 6 December 2013
Received in revised form 25 November 2014
Accepted 21 December 2014
Available online 14 February 2015

Keywords:
Soil data
Modeling
Sensor network
High-tunnel greenhouse

ABSTRACT

Precise modeling of field sensor data is an important link in precision agriculture which uses a wireless network for data collecting and field management. A good sensor model allows accurate prediction of environmental variables even with incomplete sensor data and provides basis to assess the quality of sensor readings. We investigate a clustered sensor model using observations of nearby sensors. The proposed method uses a cluster of self-evolving sub-models to model the dynamic and correlation between the networked field sensors. Each cluster represents a set of closely-related sensor attributes. The model is shown to produce accurate sensor prediction when proper attributes are selected during model training. The clustered sensor model is evaluated using field data collected in a high tunnel greenhouse. Our experiment data indicate that correlation of sensor attributes can be identified from training data and significantly improve prediction accuracy with the presence of faulty sensor data.

1. Introduction

Wireless Actuating Networks (WAN) are increasingly being used in industry automation, precision agriculture and environment monitoring. They help us to acquire new knowledge of the physical world with unparalleled details, and to apply proper actions to control the underlying physical processes. To demonstrate the advantages and feasibility of WAN for precision agriculture, we developed an experimental smart high-tunnel greenhouse with a WAN to control environmental conditions in the high-tunnel greenhouses. Sensors and actuation devices (nodes) are installed in the field of the greenhouse to measure and control critical greenhouse parameters (see Fig. 1).

Our primary goal is to collect and control crucial parameters relevant to the growth of the plants in the high-tunnel greenhouses. The relevant parameters include ambient temperature, water potential, solar radiation level, soil pH level, and concentration of certain ion types. These sensor data will help agriculture researchers to better understand the growing, dormant, and flowering process of the plants.

One challenge encountered in our deployment is the lack of the means to assess the quality of sensor readings. Erroneous sensor readings are observed in many of our experiments. Our investigations show that faulty readings could be attributed to many possible sources including incorrectly calibrated sensor, low battery supply, bad connections, or improper sensor installation. Ni et al. (2009) shows that sensor readings can be erroneous due to flaws in the physical circuit board. In Vuran and Akyildiz (2006), data errors are caused by improper deployment and insufficient battery power to drive the sensors. Since accurate data is crucial to perform actuation in precision agriculture, such as turning on irrigation or fertilizer valves, the sensor data need to be carefully processed to prevent irreversible effects on the field/plants.

Much effort has been focused on statistical modeling of sensor data for prediction of future sensor readings Ni et al. (2009), Chu et al. (2006), and Le Borgne et al. (2007). The goal is to reduce the number of actual sensing and data transmissions and the overall battery consumption.[1] An important question in the model selection is: *How do we screen the past data to achieve the most accurate prediction for a given set of interested attributes?*

Our main contribution is a dynamic sensor data model based on clustering of sensor attributes. Our model provides both validation and prediction of sensor readings in one framework. In our model, the combined sensor observation, which consists of N attributes (parameters), is modeled by a time-varying k-mode random process

$$([X_1(t), X_2(t), \ldots, X_k(t))$$

Each mode $X_j(t)$ is defined as an multi-variable Gaussian process in n_j dimension $X_j(t) = \langle x_{j,1}(t), x_{j,2}(t), \ldots, x_{j,n_j}(t)\rangle$, such that

$$\sum_{j=1,K} n_j = N$$

* Corresponding author.
E-mail address: jwang@vsu.edu (J. Wang).

[1] Sensor device battery life is dominated by the duty cycle and networking activities.

http://dx.doi.org/10.1016/j.compag.2014.12.016
0168-1699/

K is the number of model clusters. The forming of cluster reflects the main difference between our model and prior data models such as Barbie-Q Query System (BBQ) Deshpande et al. (2004).

We implement our proposed method in a dual-model system. One Gaussian model runs on the sensor node which is directly driven by the local sensor readings. A matching joint Gaussian model runs on the monitoring computer at the base station which combine all sensor reading collected at the base. Both models are driven and updated by the same actual sensor readings, except that the one on the sensor node is more frequently updated than the one on the base station side. Our clustered model require less computing power than an un-clustered model. An un-clustered sensor model, albeit its mathematical elegance in describing the entire system, is likely noisier than one containing fewer selected parameters. Such claim is intuitively true: one might have more success to predict tomorrow's temperature based on today's temperature, rather than using the combined information of today's stock market and temperature. Agricultural sensors demonstrate varying degrees of correlations with other environment parameters.

In the rest of this paper, Section 2 discusses related studies; Sections 3 and 4 discuss the behavior of predictive data models in presence of abnormal sensor readings. Section 5 discusses the clustered data model, and Section 6 shows the performance evaluation results of the proposed scheme.

2. Motivation and related studies

Statistical modeling of wireless sensor data has been discussed by several authors Deshpande et al. (2004), Chu et al. (2006), and Ni et al. (2009). The objective of using statistical data models is to predict the value of a sensor without performing actual sensing. This has the apparent benefit of saving sensor battery and reducing network traffic, if the predicted reading is sufficiently accurate. In addition to the energy saving benefit, prediction is useful when the actual reading of a sensor cannot be obtained.

Two dominant wireless sensor network data models are: (1) Markov Gaussian models where the collection sensors are modeled as Gaussian, evolving through a Markov chain. An example of such work is in Deshpande et al. (2004). (2) Regressive model where typically a low order Auto-Regressive (AR) model is used for prediction Tulone and Madden (2006). A typical sensor network deployment would adopt a dual-model configuration Deshpande et al. (2004) and Le Borgne et al. (2007), where both the base station and sensor node run an updating process for model synchronization.

Existing work has established that a statistical model can indeed be very accurate and effective in providing valuable energy saving. However, a main drawback in previous work is an implicit assumption that the sensor data used to build the model consists of 100% good readings. If there is any abnormality or incorrect readings, they must be filtered out (usually manually). This can be difficult and overwhelming in practice. Tolle et al. (2005) reported that 49% of data cannot be used for meaningful interpretation in their experiment on monitoring the micro-climate of a redwood tree. Szewczyk et al. (2004) estimated that 3–60% of data from each sensor were faulty in their habitat monitoring experiment at Great Duck Island, Maine.

Data fault detection in software is performed by capturing significant departures of the observed value in one sensor node in relation to neighboring nodes or from the past history of the same node. These fault detection methods are based on the assumption that sensed data have spatial and temporal correlation.

This leads to a practical question: which part of the historical data should be used to assess a new observation? Thus far, the effect of erroneous sensor reading has not been systematically evaluated, which is part of our motivation for this work. Our work is one of the first to evaluate the confidence level of a reported measurement, and further determine the optimum usage of historical data.

3. Baseline observations

To evaluate the baseline correlation of sensor readings for our greenhouse sensor network, we installed a four-node, three-hop wireless sensor network to observe the sensor behavior, node battery level, and their correlations. All sensor nodes report their reading regularly to the base station node. To analyze all network activity, we use an external network sniffer that is capable of capturing all IEEE 802.11.4 traffic in the area. The sniffer is placed carefully to capture all sensor data packets in the system.

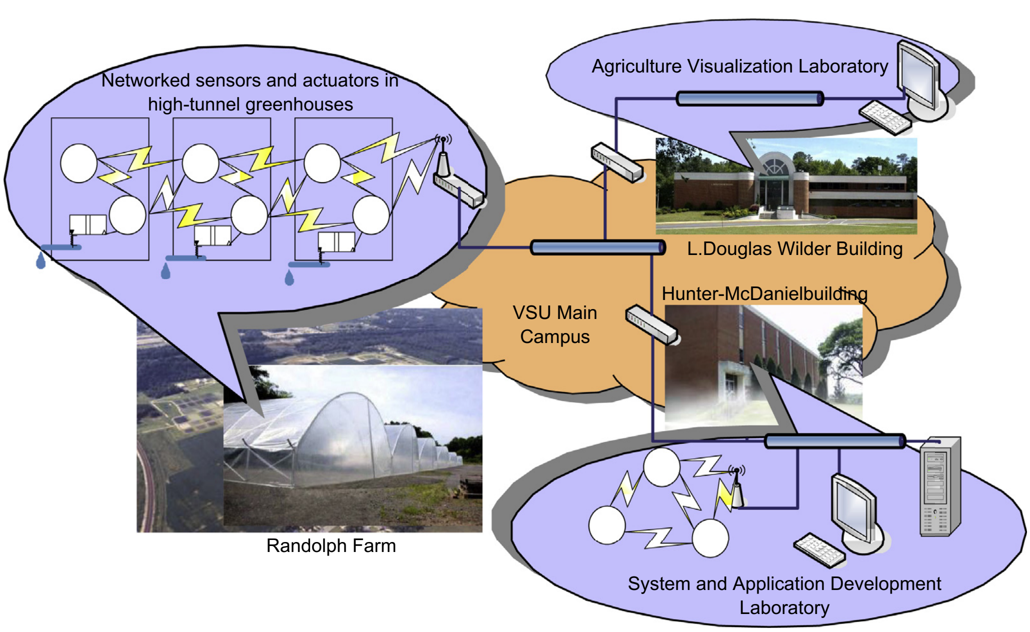

Fig. 1. Smart high-tunnel greenhouse.