A Modeling Study of Sensor Data

Feng Liu[1, a] , Jianyong Wang[1,b,*], Ming Liu[2,c]

[1]College of Science, Huazhong Agricultural University

No.1, Shizishan Street, Hongshan District, Wuhan, Hubei Province, 430070, P.R.China

[2]Unit 95025, Chinese People’s Liberation Army, China

[a]liufeng@mail.hzau.edu.cn, [b]wjy01@mail.hzau.edu.cn, [c]liumingwht@163.com

* Corresponding author

Keywords: Wireless Sensor Network, Sensor Data Analysis, Data modeling

Abstract. Nowadays, Internet of Things (IoT) has been becoming a hot research topic. Being an important part of Internet of Things, the wireless sensor networks collect various types of environmental data and construct the fundamental structure of the IoT applications. In order to find out the characteristics of the environmental data, in this paper, we focus on four types of these sensor data: temperature, humidity, light and voltage, and employ statistical methods to analyze and model these sensor data. The results of our research can be used to solve the missing sensor data estimation problem which is inevitable in the wireless sensor networks.

Introduction

Sensors and actuators construct the fundamental structure of Internet of things. The environmental data collected by these devices are applied to various kinds of IoT applications. So, these sensor data are the key elemental factor for an application. However, because of the harsh outdoor environmental condition, the failure of sensors is very common and the phenomenon of data missing is inevitable in the wireless sensor network, which affects the stability and feasibility of the applications which is based on these wireless sensor networks. In order to estimate the missing data, different methods have been proposed. But, they seldom took the features of these sensor data into consideration and then design different estimation methods according to different kinds of environmental data. Therefore, in this paper, we will pay more attention to the analysis of these sensor data and try to find out the features and characteristics of these sensor data.

To estimate the missing sensor data, Verma et al. [1] propose the latent variable based data estimation method which is evaluated on a real life sensor network consisting of 122 environmental monitoring stations and the results show that this method can effectively reconstruct the missing data. Lu et al. [2] propose an adaptive inverse-distance weighting spatial interpolation technique. In [3], a KNN imputation procedure using a feature-weighted distance metric based on mutual information is proposed.

Data measurement and modeling method have been used in many research areas such as the Internet data processing and simulation. Basher et al. [4] employ flow-level distributional models to analyze P2P and Web traffic and present an extensive characterization of Web and P2P traffic. He et al. [5] perform a detailed analysis and modeling of different types of P2P traffic in the flow level. In [5] they find that the traffic volume, connection duration and connection inter-arrival times of different types of P2P applications have different characteristics and can not be characterized uniformly with a single model.

In [6], Bin et al. propose four data mining models for the IoT. These four models include multi-layer data mining model, distributed data mining model, grid based data mining model and the data mining model which is from multi-technology integration perspective.

In this paper we want to introduce these Internet traffic analysis methods into the sensor data analysis research area.

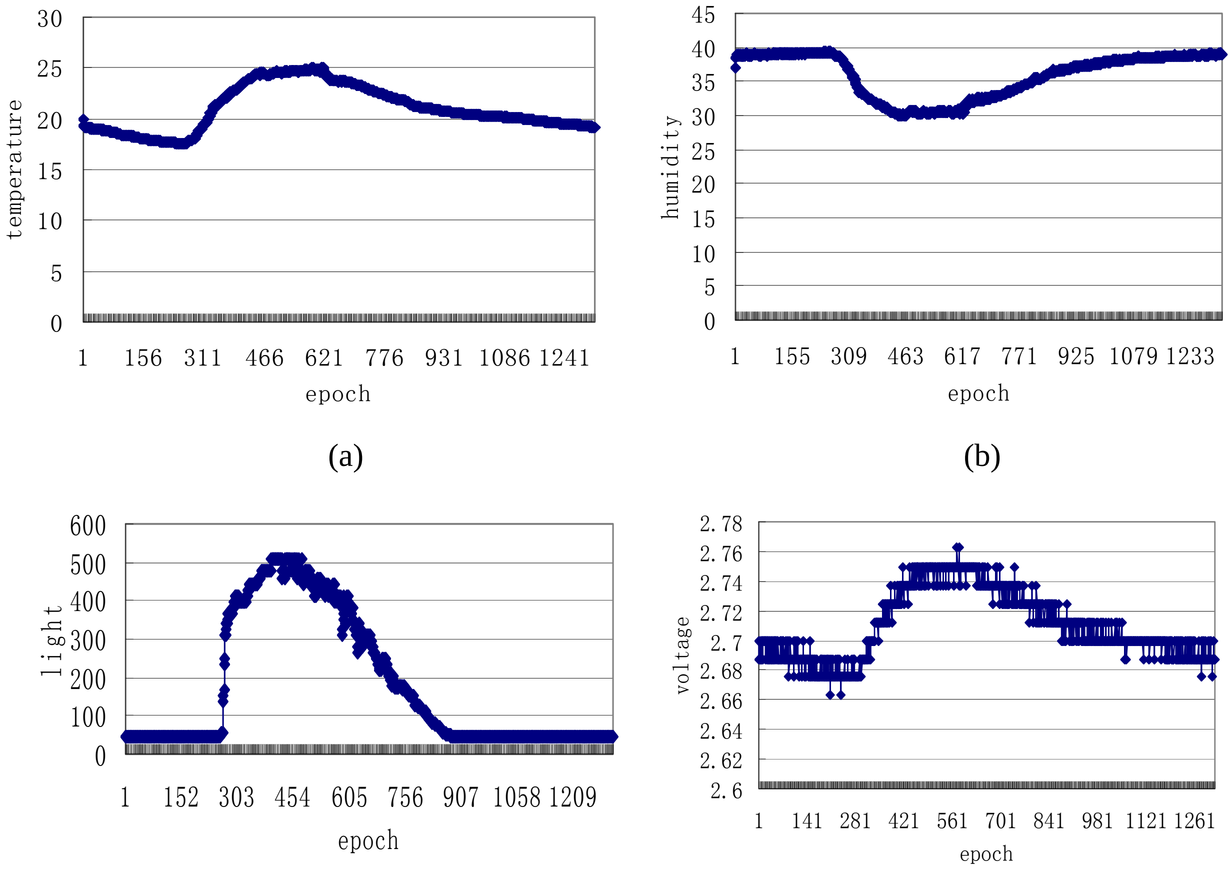
Published by Atlantis Press, Paris, France.

The remainder of this paper is organized as follows. In the section of "Sensor data analysis", we will analyze the characteristics of different types of environmental data. In the section of "Sensor data modeling", we will employ adequate mathematical model to descript these features and characteristics. At last, we will conclude and expand future our work.

Sensor data analysis

In this section, we will analyze the features of these sensor data. We perform our task by using the data set [7] which is collected by Intel Berkeley Research Lab. These data collected from 54 sensors which are deployed in the lab. These sensor data are collected every 31seconds including four types: temperature, humidity, light and voltage. Epoch is a monotonically increasing sequence number from each mote and two readings from the same epoch number were produced from different motes at the same time.

We take sensor node 1 for example and analyze the data values of temperature, humidity, light and voltage in a whole day of 28 February, 2004.



(a)

(b)

(c)

(d)

Figure 1. Four types of environmental data in one day: (a) temperature (b) humidity (c) light (d): voltage

Fig.1 illustrates the four types of environmental data in one day, including temperature, humidity, light and voltage. These changing characteristics are very important for missing data estimation and we can choose adequate estimation algorithm for different types of environmental data in different situation.

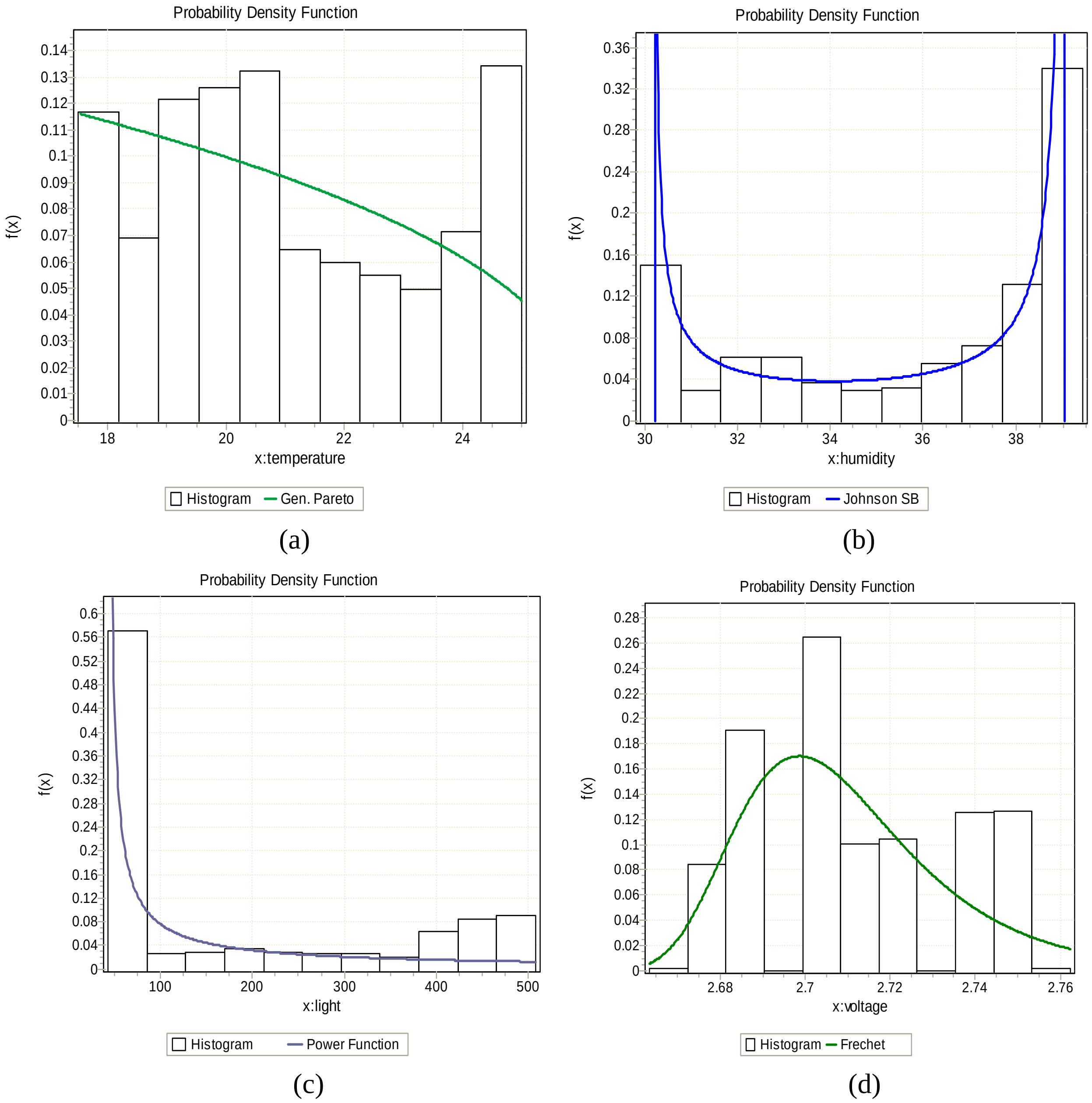
From fig.1 (a) we can find that during the morning and the night, the temperature is low and is stably changing, while during the day time, the temperature is much higher and after the sun rise, the temperature will increase dramatically. Fig 2(b) shows the humidity values, and from fig2(b), we can see that around the 12 o'clock am, the humidity value reach the lowest point. From (a) and (b) we can draw the conclusion that the temperature value and the humidity value have strong relationship. So,

Published by Atlantis Press, Paris, France.

we can estimate the missing temperature data not only based on these temperature data but also based on the humidity values. Fig 1 (c) shows the light value of a data. During the night, the light value is near 0 and it reaches the highest point around the 12 o'clock am. From fig.1 (d) we can see that the change of voltage is not as stable as the other three types of environmental data. So, it is much harder to estimate the missing voltage sensor data.

Sensor data modeling

In this section, we will present the distribution models of the four types of environmental data. We perform a statistical fitting to find the analytical distribution and we use the Kolmogorov-Smirnov (K-S) goodness of fit test to choose the best fitting analytical distribution.



(a) (b)

(c) (d)

Figure 2. The modeling results: (a) temperature (b) humidity (c) light (d) voltage

Fig.2 shows the modeling results of the four types of environmental data and the best fitting analytical distributions of them are Gen.Pareto, Johnson SB, PowerFunction and Frechet respectively. Particularly, for the voltage data it is very hard for us to find a distribution that fits these data well enough. That is because that this variable does not change stable during a day as show in fig 1(d). So for this kind of unstable variables, when we estimated the missing sensor data, we should employ special methods.

Published by Atlantis Press, Paris, France.

Summary

Because sensor data processing is a critical work in Internet of Things, in this paper, we focus on the sensor data's analysis and try to find the most fitted distributions for the four types of environmental data. The results of our work show that some environmental variables change stable and have strong correlation with other environmental variables, while some variables are not stable. Consequently, when we estimate the missing sensor data, we should choose different methods according to the data's characteristics and the relationship of different environmental variables should also be taken into consideration. For the unstable variables, it is much harder to estimate the missing data. In our future work, we will employ these data characteristics to construct the missing data estimation algorithms and pay more attention to the estimation work of the unstable environmental variables.

Acknowledgment

This work is supported by “Specialized Research Fund for the Doctoral Program of Higher Education” under Grant No. 20120146120002, and by Natural Science Foundation of Hubei Province of China under Grant No. 2012FFB02801, and by “the Fundamental Research Funds for the Central Universities” under Grant No. 2011QC019.

References

[1] N. Verma, P. Zappi, and T. Rosing, "Latent variables based data estimation for sensing applications." in 2011 Seventh International Conference on Intelligent Sensors, Sensor Networks and Information, 2011, pp. 335-340

[2] G. Y. Lu, and D. W. Wong, “An adaptive inverse-distance weighting spatial interpolation technique,” Computers & Geosciences, vol. 34, no. 9, pp. 1044-1055, 2008

[3] P. J. Garcia-Laencina, J.-L. Sancho-Gomez, A. R. Figueiras-Vidal et al., “K nearest neighbours with mutual information for simultaneous classification and missing data imputation,” Neurocomputing, vol. 72, no. 7-9, pp. 1483-1493, 2009

[4] N. Basher, A. Mahanti, A. Mahanti et al., "A comparative analysis of web and Peer-to-Peer traffic," in Proceeding of the 17th International Conference on World Wide Web, 2008, pp. 287-296.

[5] G. He, J. Hou, W.-P. Chen et al., "One Size Does Not Fit All: A Detailed Analysis and Modeling of P2P Traffic." in conference of GLOBECOM '07, 2007, pp. 393-398

[6] B. Shen, Y. Liu, and X. Wang, "Research on data mining models for the internet of things." in conference on Image Analysis and Signal Processing, 2010, pp. 127-132

[7] S. Madden. "Intel Berkeley research lab data,” on http://berkeley.intel-research.net/labdata