International Journal of Information Engineering and Applications
2018; 1(1): 1-5
http://www.aascit.org/journal/information

AASCIT American Association for Science and Technology

Sensor Reduction Method for Intel Berekely Data Set Based on Machine Learning

Mohammad Abdulaziz Alwadi, Girija Chetty

Faculty of Information Sciences and Engineering, the University of Canberra, Canberra, Australia

Email address

wadi@uni.canberra.edu.au (M. A. Alwadi)

Citation

Mohammad Abdulaziz Alwadi, Girija Chetty. Sensor Reduction Method for Intel Berekely Data Set Based on Machine Learning. *International Journal of Information Engineering and Applications.* Vol. 1, No. 1, 2018, pp. 1-5.

Received: January 2, 2018; **Accepted**: January 20, 2018; **Published**: February 12, 2018

Abstract: In this paper a proposed energy efficient sensor reduction method for Intel Berkeley lab wireless sensor network data set based on machine learning. The experimental work in this paper using publicly available WSN dataset to show the possibility to reduce the number of sensors used in order to enhance the energy efficiency where the system resources and energy are always a key issue in Wireless sensor networks. The main concept is to perform certain experiments on the Intel Berkeley to come up with improved energy reduction method associated with the least number of sensors used to get better sensor life time and performance. Rest of the paper is organized as follows. Next sections describing the background and introduction, the details of the dataset used, and section 3 describes the sensor selection and routing approach, where the details of experimental results obtained are presented, and the paper concludes in Section 4.

Keywords: Machine Learning, Wireless Sensor Networks, Dataset, Sensor Life Time

1. Introduction

Wireless and wired sensor networks have become a focus of intensive research today, especially for monitoring large physical environments, and for tracking environmental or physical conditions such as temperature, pressure, wind and humidity. A wireless or a wired sensor network (WSN/SN) consists of a number of sensor nodes (few to thousands) storing, processing and rotating the data often to a base station for further computation [1], [2]. Sensor networks can be used in many applications, such as wildlife monitoring [3], military target tracking and surveillance [4], hazardous environment exploration [5] and natural disaster relief [6]. Many of these applications are expected to run unattended for months or years. Sensor nodes are however constrained my limited resources, particularly in terms of energy. Since communication is one order of magnitude more energy consuming than processing, the design of data collection schemes that limit the amount of transmitted data is therefore recognized as a central issue for wireless sensor networks. An efficient way to address this challenge is to approximate, by means of mathematical models, the evolution of the measurements taken by sensor over space and/or time. Indeed whenever a mathematical model may used in place of the true measurements, significant gains in communications may be obtained by only transmitting the parameters of the model instead of the set of real measurements. Since in most cases there is litter no a priori information about the variations taken by sensor measurements, the models must be identified in an automated manner. This calls for the use of machine learning and data mining techniques, which allow modeling the variations of future measurements on the bases of past measurements.

In this paper, a machine learning based formulation for energy efficient WSN (Wireless Sensor Network) Monitoring is proposed. The proposed approach involves an adaptive routing scheme to be used for energy efficiently and is based on selecting most signification sensors for the accurate modeling of the WSN environment. The experimental validation of the proposed scheme for publicly available Intel Berkeley lab wireless sensor network dataset shows it is indeed possible to achieve energy efficiently without degradation in accurate characterization and understanding of WSN environment. The propped machine learning and data mining formulation for achieving energy efficiency, provides better implementation mechanism in terms of tradeoff between accuracy and energy efficiency, due to an optimal combination of feature selection and classifier techniques

used in machine learning. By approaching the complexity of WSN/SN with a data mining formulation where each sensor is equivalent to an attribute of the data set, and all the sensor nodes together forming the WSN/SN set up equivalent to a multiple feature or attributes of the data set, it is possible to use powerful feature selection, dimensionality reduction and learning classifier algorithms from machine learning/data mining field, and come up with an energy efficient environment monitoring system [7]. In other word, by employing a good feature selection algorithm along with a good classification algorithm, for example it is possible to obtain an energy efficient solution with acceptable characterization or classification accuracy (where the WSN/SN set up is emulated with a data set acquired from physical environment). Here, minimizing the number of sensors for energy efficiently is very similar to minimizing the number of feature with an optimal feature selection scheme for the data mining problem. Further, the number of sensors chosen by the feature selection scheme leads to a routing scheme for collecting the data from sensors and transmitting them until reaches the base station node. As accuracy of data mining schemes rely on amount of previous data available for predicting the future state of the environment, it is possible to obtain an adaptive routing scheme, through the life of WSN, as more and more historical data becomes available, allowing trade- off between energy efficiency and prediction accuracy. It has been shown that it is possible to do this, with an experimental validation of our proposed scheme with a publicly available WSN dataset acquired from real physical environment The Intel Berkeley lab [8].

2. Intel Berkeley Data Set Description

The publicly available data set used for experimental validation consists of temperature measurements come from a deployment of 54 sensors in the Intel research laboratory at Berkeley [8]. The Deployment has been done between February 28th and April 5th, 2004. A table of the log file results of the deployment is provided in Table 1 below where sensor nodes are identified by numbers ranging from 1 to 54. The log file came out as a result of the sensor nodes distributed in the lab as per the following structure.

Table 1. *Intel lab log file.*

Intel Berkeley Lab log file							
Date	**Time**	**epoch**	**moteid**	**Temp.**	**Humidity**	**Light**	**voltage**
yyyy-mm-dd	hh:mm:ss.xxx	int	int	real	real	real	real

Many sensor readings from WSN test bed were missing, due to this being a simple prototype. A few subsets of measurements from this data set were selected. The readings were originally sampled every thirty one seconds. A pre processing stage where data was partitioned was applied to the dataset. After preprocessing, a several subsets of data were prepared. The approach used for this WSN test bed, involves an assumptions that data collection and transmission is done by some of the sensors (source nodes), that purely sense the environment and transmit their measurement to collector nodes (sink nodes/base station), and based on the relative distance between source nodes and sink nodes. The route or the path taken for sensor data to be transmitted from once node to other is predetermined at sink/base station node.

The nodes which actively participate in sensing the environment, and transmit the data, consume the power and those who do not participate in this activity do not consume any power. This is how the WSN can be made energy efficient. By involving only the optimum number of sensors to participate and leaving non- participating sensors in a sleep mode. However, if the number of sensors participating in not properly chosen this may impact on the accuracy. Therefore, it is essentials that a dynamic or adaptive routing scheme is used, where the machine learning/data mining can use larger training data from previous/historical data sets to predict the future environment accurately and continuously adapt the routing scheme for nodes based on threshold error measure for prediction accuracy and energy efficiency.

3. Sensor Selection and Routing Approach

Different sets of experiments were performed to examine the relative performance of sensor selection and routing approach proposed here. K-fold stratified cross validation technique for performing experiments was used, with k=2, 5 and 10, based on the training data available (using larger folds for larger training data). Further, to estimate the relative energy efficiently achieved, experiments have been performed with all sensors (without feature selection/sensor selection) algorithm, and with sensors selected by feature selection algorithm. As mentioned before, the feature selection algorithm allows selection of an optimal number of features or sensor nodes needed to characterize or to classify the environment (which in turn leads to energy efficient scheme). Further, time taken to build the model is also an important parameter, particularly for adaptive sensor routine scheme to be used for real time environment monitoring.

Table 2. *Results from Three experimental Scenarios.*

Exp #	Number of Sensors	Number of Training Samples	Feature/Sensor
1	27	35	1, 3, 4, 14, 19, 25, 27
2	25	2700	3, 13, 14, 19, 20, 27
3	25	5400	1, 2, 3, 13, 14, 18, 20, 24, 26, 27

Table 3. *Results of time taken from Three Experimental Scenarios.*

Exp #	Time (No feature selection)	Time with feature selection	RMS error (No feature selection)	RMS error (with feature selection)
1	0.61 sec	0.02 sec	19.5%	38.79%
2	12.96 sec	0.23 sec	12.07%	12.21%
3	25.98 sec	1.85 sec	8.6%	9.07%

For the first set of experiments, the first 27 sensors and a small set of training samples (35 temperature measurements) were used. As can be seen from the sensor participating shown in Table 2, sensor 27 is the sink node (emulating base station node), and sensors 1 to 26 participate in measuring and transmitting the environment around them to the sink node, where the machine learning prediction task is to estimate the measurement at sink node (sensor 27). The RMS error (root mean squared error) at the sink node (node 27) provides a measure of prediction for all source sensor nodes (1-26) in WSN participating in measuring the temperature in the environment and sending it to sink node, the RMS error is 19.5%, and with sensor selection scheme used with only 7 sensors participating in routing scheme, the RMS error is 38.79%. As can be seen in Table 3, with a moderate degradation in accuracy (19.5 to 38.79%), energy efficiency achieved is of the order of 3.7 (26/7). A new measure for energy efficiency is used, the life time extension factor (LTEF), which can be defined as:

$$lifeTimeExtensionFactor = \frac{totalnumberofsensors}{Sensorsparticipatingintheroutingscheme}$$

With 7 sensors out of 27 sensor nodes in active mode, the LTEF achieved is around 3 times, and 20 sensor nodes are in sleep mode. The trade off is a slight reduction in accuracy. This could be due to less training data used. Only 35 temperature samples for prediction scheme were used. With more data samples used in the prediction scheme, performance could be better. To test this hypothesis, the following next set of experiments will be performed.

For second set of experiments, 2700 training samples collected on different days were used. As can be seen in Table 2, With larger training data size, the participating sensors in the routing scheme are different, as the proposed feature selection algorithm chooses different set of sensors (3, 13, 14, 19, 20, 27). 25 sensors for this set of experiments were used, as two of the sensors (sensor 5 and sensor 15 did not have more than 35 measurements). With all 25 sensors in he routing scheme, the RMS errors is 12.07%, and with 6 sensor nodes (3, 13, 14, 19, 27), the error is 12.21%. This is a Significant improvement in prediction accuracy (from 38%%to 12.21%), with life time extension of 4.16 (25/6). As is evident here, by using larger training data (2700 temperature Measurements), it was possible to achieve an improvement in prediction accuracy and energy efficiency as well.

To examine the influence of increasing training data size, A third set of experiments with 5200 samples were performed. The performance achieved for this set of experiments is shown in Table 2. Here the adaptive routing scheme based on proposed feature selection technique selects 10 sensors (1, 2, 3, 13, 14, 18, 20, 24, 26, 27). For this set of experiments, the RMS error varies from 8.6% for all sensors participating in the scheme to 9.07% with LTEF of 2.5 (25/10). Though there is no degradation in prediction accuracy, there is not much improvement in energy efficiency, with doubling of training data size for the building the model this could be due to overtraining that has happened, with the network losing its generalization ability. So by increasing training data size, it may not be just possible to achieve performance improvement, for prediction accuracy (RMS error) and energy efficiency (LTEF), and a trade-off may be needed. An optimal combination of training data size, and number of sensors actively participating in routing scheme can result in energy efficient WSN, without compromising the prediction.

Accuracy. Figure 1 below shows explanation for the results from three set of experiments.

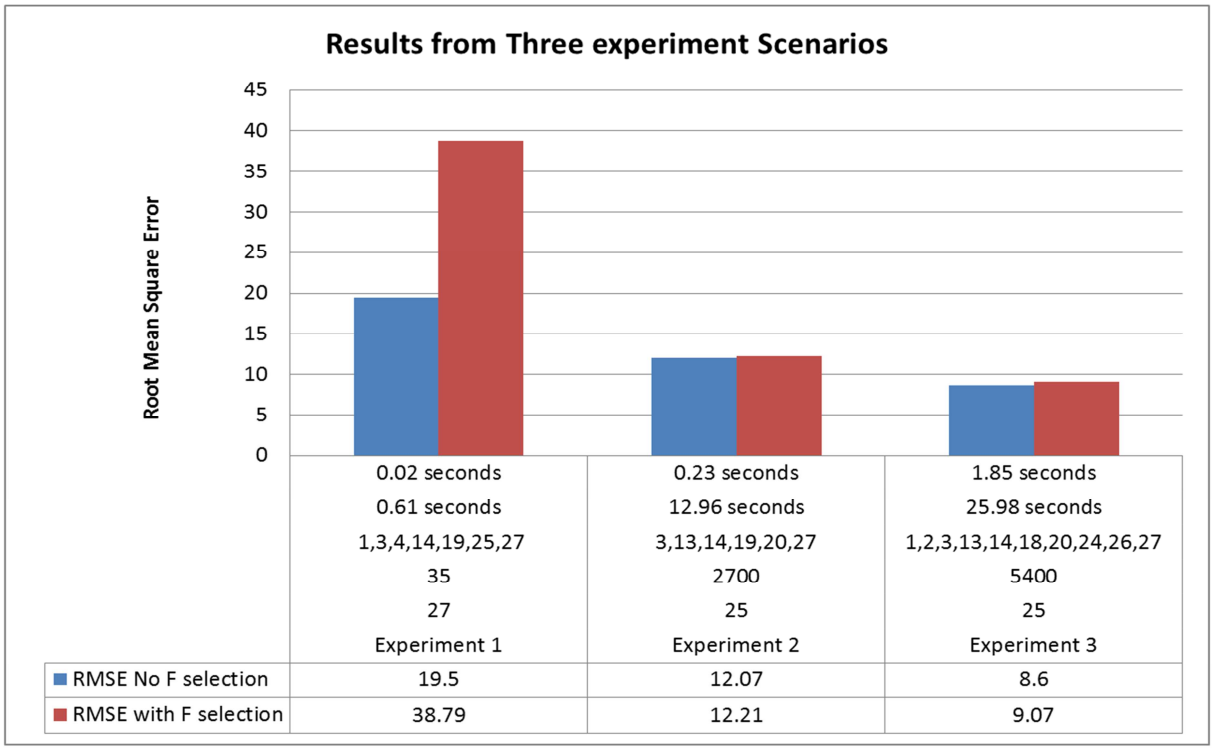

Figure 1. *Results from three experiments scenarios.*

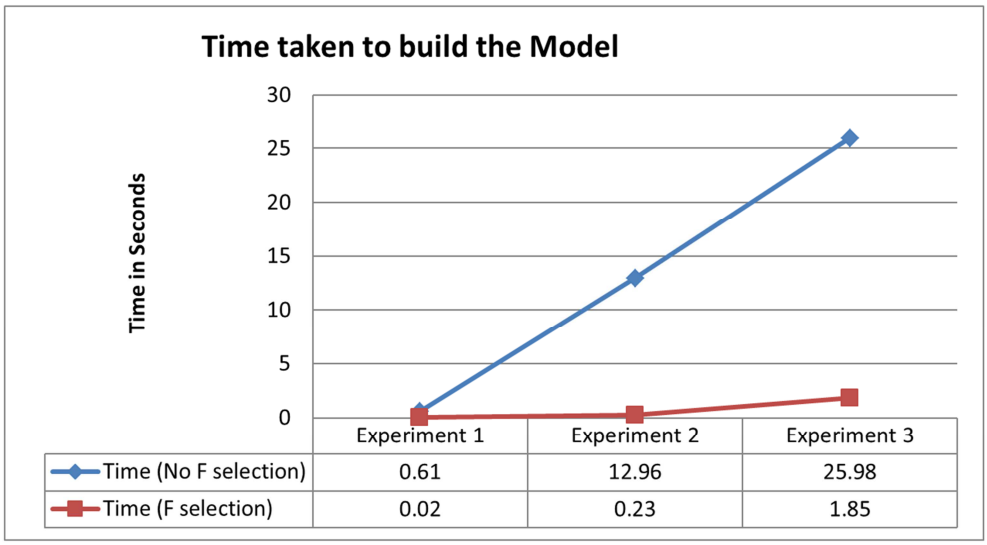

Figure 2. *Time taken to build the model.*

Further, another important parameter is model building time, as for adaptive sensor routing scheme to be implemented in real time WSN environment, routing scheme Journal of Advances in Computer Networks, Vol. 3, No. 4, December 2015 has to dynamically compute the sensors that are in active mode and in sleep mode. Out of 3 experimental scenarios considered here, as can be seen from Table 3, the model building time improves from 0.61 seconds to 0.02 seconds for experiment 1, from 12.96 seconds 0.23 seconds for experiment 2, and from 25.98 seconds to 1.85 seconds for experiment 3. So, the proposed adaptive routing scheme for sensor selection provides an added benefit of reduced model building times, suitable for real time deployment. The Figure 2 shows the times taken for the three experiments in comparison.

4. Conclusion

Energy sources are very limited in sensor networks, in particular wireless sensor networks. For monitoring large physical environments using SNs and WSNs, it is important that appropriate intelligent monitoring protocols and adaptive routing schemes are used to achieve energy efficiency and increase in lifetime of sensor nodes, without compromising the accuracy of characterizing the WSN environment. In this paper, we proposed an adaptive routing scheme for sensor

nodes in WSN, based on machine learning data mining formulation with a feature selection algorithm that selects few most significant sensors to be active at a time, and adapts them continuously as time evolves. The experimental validation for a real world publicly available WSN dataset, proves our hypothesis, and allows energy efficiency to be achieved without compromising the prediction accuracy, with an added benefit in terms of reduced model building times. Further work involves, developing new algorithms for sensor selection and environment characterization with WSNs and their experimental validation with other similar datasets that can lead to better energy efficiency. Also, the further research involves extending this work with adapting these classifiers for big data stream data mining schemes, for real time dynamic monitoring of complex and large physical environments in an energy efficient manner.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (references).

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp. 68-73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] Bodik, P., et al., Intel lab data. Online dataset, 2004.

[9] Y. Bengio, "Learning deep architectures for AI," Foundations and Trends in Machine Learning, vol. 2, no. 1, pp. 1–127, 2009.

[10] [MATLAB. 20/01/2012]; Available from: http://www.mathworks.com.au/.

[11] W. Branch, C. Giannella, B. Szymanski, R. Wolff, and H. Kargupta, "In-network outlier detection in wireless sensor networks," Knowledge and information systems, vol. 34, no. 1, pp. 23–54, 2013.