Regression-based Data Reduction Algorithm for Smart Grids

Bashar Chreim, Jad Nassar, Carol Habib

Regression-based Data Reduction Algorithm for Smart Grids

1st Bashar Chreim
CSM Department
JUNIA
Lille, France
Bashar.chreim@junia.com

2nd Jad Nassar
CSM Department
JUNIA
Lille, France
Jad.nassar@junia.com

3rd Carol Habib
CSM Department
JUNIA
Lille, France
Carol.habib@junia.com

***Abstract*—The evolution towards Smart Grids (SGs) represents an important opportunity for the energy industry. It is characterized by the integration of renewable and alternative energy resources into the existing power grids while ensuring a fine-grained control for the different measuring points. Therefore, this evolution requires the ability to send a maximum of data over the network in real time while controlling the grid. A Wireless Sensor Network (WSN) deployed across the grid is a potent solution to achieve this task. However, sensor nodes have limited energy and computation resources especially the battery powered ones. For that, reducing transmission is an essential priority in order to increase the lifetime of the network. Data prediction is a widely used, yet effective, solution in literature to accomplish this task. In this paper, we propose a Quality of Service (QoS) aware algorithm based on time series forecasting and linear regression for data prediction in WSN. Our algorithm takes into consideration the diversity of applications of SGs with different requirements while being energy efficient. We expect to reduce the number of transmission and energy consumption, while respecting the accuracy of the data.**

***Index Terms*—Smart Grids, Wireless Sensor Networks, Data Reduction, Data Prediction, Linear Regression, Linear Correlation, Quality of Service**

I. Introduction

The evolution towards Smart Grids (SGs) represents an important opportunity to shift the energy industry into a new era of reliability, availability and efficiency [1]. This evolution is manifested by the integration of renewable energy resources all over the grid and a two-way communication between the utility and the customers. Therefore, these changes require the ability to transmit in real time a maximum of data over the network, in order to monitor and control the different heterogeneous decentralized energy resources. A Wireless Sensor Network (WSN) deployed all over the grid on the different measuring and control points, is a potential and plausible solution to be used with SGs [1]. In a WSN, monitoring, processing and transmitting are the main tasks accomplished by sensor nodes. These sensors have limited energy and computation resources [2] since they are powered by batteries with a limited lifespan. Each time interval, source nodes perform data sampling and transmission to the destination via a set of sensor nodes distributed across the network. However, most of the time, sensed values do not change significantly between consecutive readings. Sending these samples periodically will cause exhaustion of the batteries of sensor nodes (knowing that wireless communication is considered the major energy consumer [2]) and information redundancy at the destination. For that, and in order to maximize the sensor nodes lifetime, data reduction has proven to be a potent solution. This is done by reducing the data transmission rate or aggregating data packets within the network. In this paper, we propose an energy efficient data reduction algorithm based on data prediction. It uses time series forecasting and linear regression models. More precisely, our algorithm exploits correlations among different data types collected from photo-voltaic cells and creates automatically a prediction model for each variable. To the best of our knowledge, our work is the first effort to apply linear regression in data reduction for WSNs combined with a time series forecasting model. The rest of the paper is organized as follows: Section II presents a summary of related work. Section III describes our proposed solution. Finally, section IV concludes the paper.

II. Related Work

In literature, numerous studies focused on developing data prediction techniques in WSNs. In [3], the authors used Auto Regressive Integrated Moving Average ($ARIMA$) model for multi-month forecasting of monthly mean daily global solar radiation. In [4], the authors proposed a prediction scheme using Seasonal $ARIMA$ ($SARIMA$) model for short term prediction that can predict using only limited dataset. This model was used to predict traffic flow. The main drawback with these methods is their requirement of high memory and computational overhead to initially build the model and to re-compute it when outdated. In [5], Least Mean Square (LMS) algorithm is used for data prediction in WSN. It consists of running two instances of the model at the sensor and the sink node, applying dual prediction scheme (DPS). The main complexity when using LMS is the task of choosing the best parameters to fit to a specific data type. In [6], the authors proposed a modification for the LMS algorithm by adding a phase of initialization and parameters determination. Moreover, other work focused on linear regression for data prediction. In [7], the authors used multiple linear regression to predict load from a set of weather and time metrics. In [8], the authors determined the most influential features for predicting

978-1-7281-9794-4/21/$31.00 ©2021 IEEE

photo-voltaic production, using linear regression. However, in these work, all input variables need to be available all the time in order to predict a single output. In other words, sensor nodes that are responsible of collecting the value of these variables need to execute the transmission task all the time, which may exhaust the batteries of the sensor nodes after a period of time.

III. Proposed Solution

The purpose of our contribution is to create an autonomous prediction algorithm for heterogeneous applications in WSN. It generates simultaneous prediction models for all variables by exploiting linear correlation among them. This is done using two types of models: *time series forecasting* and *linear regression*. The first selected variable is predicted using a *time series forecasting* model. After that, the remaining variables are predicted using *linear regression* models. In the rest of this section, we will detail the steps of our algorithm that is represented in algorithm 1.

- Our algorithm takes as *input* a dataset with a random number of variables (i.e., temperature, humidity, etc).
- In the first step, the correlation matrix (CM) is built (line 2 in Algorithm 1). It represents the correlation coefficient of each couple of variables.
- After that, the negative values are replaced by their absolute ones in the CM (since the positive and negative values have the same impact on the prediction).
- In the next step, the maximum value (Max) in the matrix must be identified (line 8 in Algorithm 1). It represents the correlation coefficient of the most correlated couple of variables.
- A *time series forecasting* model is created using $SARIMA$ [4] for one variable of the couple, randomly selected.
- The *output* of this time series model is used as *input* to create a *simple linear regression* (SLR) model (line 11 in Algorithm 1). It predicts the second variable previously identified.
- The next model should be created now by identifying the next maximum value in the matrix. It is a *multiple linear regression* (MLR) model, and takes as *input* the *outputs* of the previous created models (line 17 in Algorithm 1).
- This final step is repeated until prediction models are created for all variables. Our algorithm returns a matrix that represents the input/s and output variable/s of each created model.
- Once the execution of the algorithm finishes, all the prediction models are created. The prediction process can now take place.

In a nutshell, our algorithm works as follows: a *time series forecasting* model will predict the upcoming value of one of the variables of the application. Next, the predicted value is used as input by a *simple linear regression model* to predict the value of the second variable. Then, *multiple linear regression* models are executed simultaneously by taking the predicted values to predict the value of the next corresponding variable, and so on.

Algorithm 1 : Data prediction algorithm

Require: Dataset
Ensure: Prediction Models
1: $Count \leftarrow NbOfVariables(dataset)$ //Number of needed models
2: $CM \leftarrow BuildCorrelationMatrix()$
3: **for** each value in CM **do**
4: **if** $value < 0$ **then**
5: $value \leftarrow |value|$
6: **end if**
7: **end for**
8: $Max \leftarrow CM.MaximumValue()$
9: $X, Y \leftarrow Max.IndexInCM()$
10: $SARIMA(X)$
11: $SLR(X, Y)$
12: $Count \leftarrow Count - 2$
13: $Max \leftarrow 0$ //Replace maximum by zero in the matrix
14: **repeat**
15: $Max \leftarrow CM.MaximumValue()$
16: $X, Y \leftarrow Max.IndexInCM()$
17: $MLR(X\ and\ all\ previous\ outputs, Y)$
18: $Max \leftarrow 0$
19: $Count \leftarrow Count - 1$
20: **until** $Count == 0$ //Each variable has its prediction model

IV. Conclusion and Future Work

In this paper, we presented a data prediction correlation based approach, that automatically generates prediction models for different heterogeneous variables. The main advantage of our approach is the ability to adapt to different applications with different requirements as per a SG environment while being energy efficient. The expected results aim to reduce the number of transmission and energy consumption in a SG controlled by a WSN, while respecting the QoS requirements. Several tests and investigations have to be performed (i.e., computer simulations) before the completion of this work. Later on, we will implement our algorithm in a real WSN scenario to validate our theoretical approach.

References

[1] J. Nassar, "Ubiquitous Networks for Smart Grids," Theses, Universite des Sciences et Technologies de Lille, Oct. 2018. [Online]. Available: https://hal.inria.fr/tel-01908825

[2] U. Raza, A. Camerra, A. L. Murphy, T. Palpanas, and G. P. Picco, "What does model-driven data acquisition really achieve in wireless sensor networks?" in *2012 IEEE International Conference on Pervasive Computing and Communications*. IEEE, 2012, pp. 85–94.

[3] B. Belmahdi, M. Louzazni, and A. El Bouardi, "One month-ahead forecasting of mean daily global solar radiation using time series models," *Optik*, vol. 219, p. 165207, 2020.

[4] S. V. Kumar and L. Vanajakshi, "Short-term traffic flow prediction using seasonal arima model with limited input data," *European Transport Research Review*, vol. 7, no. 3, p. 21, 2015.

[5] S. Santini and K. Romer, "An adaptive strategy for quality-based data reduction in wireless sensor networks," in *Proceedings of the 3rd international conference on networked sensing systems (INSS 2006)*. TRF Chicago, IL, 2006, pp. 29–36.

[6] J. Nassar, K. Miranda, N. Gouvy, and N. Mitton, "Heterogeneous data reduction in wsn: Application to smart grids," in *Proceedings of the 4th ACM MobiHoc Workshop on Experiences with the Design and Implementation of Smart Objects*, 2018, pp. 1–6.

[7] B. Dhaval and A. Deshpande, "Short-term load forecasting with using multiple linear regression," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 4, p. 3911, 2020.

[8] S. G. Kalvitkar, P. U. Shinde, A. V. Wagh, and G. Gunjal, "Solar energy prediction using machine learning."