Applied Energy 163 (2016) 134–141

Contents lists available at ScienceDirect

Applied Energy

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

Forecasting the EV charging load based on customer profile or station measurement?

Mostafa Majidpour [a,*], Charlie Qiu [a], Peter Chu [a], Hemanshu R. Pota [b], Rajit Gadh [a]

[a] Smart Grid Energy Research Center, UCLA, Los Angeles, CA, USA
[b] School of Engineering & Information Technology, The University of NSW, Canberra, ACT 2610, Australia

HIGHLIGHTS

- We compare the forecasting of the EV charging load based on two different datasets.
- Customer profile dependent dataset is prone to privacy invasion.
- Station measurement dataset is directly measured from charging outlets.
- Results show customer profile based prediction is faster due to less preprocessing.
- We found that both datasets yield comparable forecasting error.

ARTICLE INFO

Article history:
Received 18 April 2015
Received in revised form 26 October 2015
Accepted 31 October 2015

Keywords:
Electric Vehicle
Privacy
Time series
Load forecasting

ABSTRACT

In this paper, forecasting of the Electric Vehicle (EV) charging load has been based on two different datasets: data from the customer profile (referred to as charging record) and data from outlet measurements (referred to as station record). Four different prediction algorithms namely Time Weighted Dot Product based Nearest Neighbor (TWDP-NN), Modified Pattern Sequence Forecasting (MPSF), Support Vector Regression (SVR), and Random Forest (RF) are applied to both datasets. The corresponding speed, accuracy, and privacy concerns are compared between the use of the charging records and station records. Real world data compiled at the outlet level from the UCLA campus parking lots are used. The results show that charging records provide relatively faster prediction while putting customer privacy in jeopardy. Station records provide relatively slower prediction while respecting the customer privacy. In general, we found that both datasets generate comparable prediction error.

1. Introduction

The most distinct feature of the smart grid is its extensive use of information and communication technologies to improve the efficiency and reliability of the generation and distribution of electricity. A large volume of information is gathered from different meters that might be sufficient to reveal the behavior of different players such as suppliers and consumers. This calls for a privacy concern as is pointed out [1].

Electric Vehicle (EV) charging related data is no exception to privacy issues and has its own problems. One such problem is the large battery size in today's EVs, which may require a relatively large amount of charging time depending on the charging station capabilities. The long charging time may obligate EV owners to charge their EVs in places other than their household, including public charging stations or charging stations at their work place. This implies that not only utilities have access to charging data through household chargers, but also charging station administrators in work places and public stations have access to them. These data, when used for different analysis in utility or public charging station operation or planning, might expose information such as the pattern of entrance and exit times of the customers from charging lots or their home, hence risking their privacy. To exemplify the battery size, consider Chevrolet Volt, Nissan Leaf, and Tesla, with a battery size of 16.5 kW h, 24 kW h, and up to 85 kW h, respectively [2]. With Level 1 household chargers (16A at 230VAC which delivers 3.3 kW) it will take the Nissan Leaf around 8 h, and Tesla (Model S85) around 25 h to charge completely.

In this paper, our application is forecasting (predicting) EV charging load based on historical charging data. We have two available datasets: charging record that comes from anonymous

* Corresponding author.
E-mail address: mostafam@ucla.edu (M. Majidpour).

http://dx.doi.org/10.1016/j.apenergy.2015.10.184
0306-2619/

customer profiles and station records that come from measurements (voltage, current, etc.). Either one of them can be used for building a load time series and hence forecasting at the outlet level.

One might wonder since the charging records are anonymized, there is no threat to customer's privacy. However, anonymizing might not be enough, as in a famous incident, the medical record for the then governor of Massachusetts was extracted easily from anonymous medical records when combined with voter registration rolls [3]. The medical records were anonymous but they had sex, ZIP code, and birth-date of patients. This incident shows that even anonymity is not enough and anonymous data might still be revealing when combined with other datasets.

Therefore, we deal with two datasets: The charging record comes from customer profiles, and, as pointed out earlier, it is prone to privacy issues. On the other hand, the station record does not have any information about specific customers and hence protects customer privacy. We compare the accuracy and speed of the prediction process using these two types of records. Specifically, for EV charging data, we investigate the potential increase in prediction accuracy and speed, as a tradeoff of endangering customer privacy. Interestingly, we found that prediction accuracy is not significantly increased while using the privacy-jeopardizing dataset (charging records). To our knowledge, this type of comparison has not been done in this context.

The rest of this paper is organized as follows: Section 2 provides a brief review of the existing literature, Section 3 formulates the problem, Section 4 reviews the prediction algorithms applied on time series based on both of the datasets. Section 5 discusses the structure of each dataset from the charging stations at the University of California, Los Angeles (UCLA) parking lots and the preprocessing stages to convert each of them into a time series. Section 6 reports the result of applying the prediction algorithms on each of the time series and then analyses the results with nonparametric statistical tests to investigate statistically meaningful differences. Section 7 provides the conclusion and future work.

2. Literature review

Privacy has been an important issue in smart grids and it is one of the factors holding people back from participating in the use of these new technologies [4].

There are various levels of invasion of privacy in smart grid context. For example, at the smart household level, the different ways that household privacy might be invaded are: access to power consumptions records, presence of different players with potential access to data such as service provider and distribution operator in the economic smart grid, using wireless communication technology between devices that might make communications vulnerable, accessing energy devices from the internet, and third parties that are not involved in any part of power generation and distribution but monitor the customer usage with customers' approval [5].

According to [6], the current resolution of smart meter data (usually between 15 min and 1 h) invades customer privacy and the data might not be necessary for most of the smart grid planning and distribution functions. Some other research still relies on anonymous data to protect the privacy where the pattern of the EV customer driving times is used for designing an optimal charging algorithm [7].

There have been various suggestions on how to preserve privacy. Some of them are based on the idea of aggregating the data instead of using individual data. For example, in [8], building energy usage is investigated and only the aggregated data, instead of individual data, is used for analysis. Similarly, instead of individual EV charging data, the aggregated data of EV loads has been used for coordinating the EV charging operation in [9].

On the other hand, centralized and distributed algorithms for the routing of the information flows which preserve the privacy based on cryptographic methods is proposed in [10]. Cryptographic methods are also used in [11] to perform privacy preserving bill calculations. All these methods fall under Secure Signal Processing (SSP) methods which protect the sensitive data by encryption and provide tools to analyze the data under the applied encryption [12].

According to [13], privacy is exposed even when relatively infrequent measurements are acquired, and on the other hand, the energy management system with battery can protect the customer privacy. The role of the battery and its use in more effectively protecting privacy is also discussed in [14].

Lastly, the above articles focus mostly on protecting privacy at the measurement level. Another way to protect the privacy is at the data mining algorithm level. Privacy preserving machine learning algorithms started to become more important in early 2000s [15,16]. In subsequent years, privacy preserving versions of different machine learning algorithms such as nearest neighbor [17], Bayes classifiers [18], Support Vector Machines [19], and logistic regression [20] were introduced in literature to address privacy-preserving in algorithmic level.

There is a rich literature for time series forecasting in various disciplines. Ref. [21] provides a comprehensive review of different models. Machine Learning algorithms have also been successfully employed in the forecasting realm [22]. In the current work, we compare four machine learning based prediction algorithms on two time series built from different measurements of one phenomenon, one of which contains privacy-insensitive information whereas the other does not. We show that the dataset without privacy-sensitive information allows us to make equally accurate charging load prediction compared to the other dataset, thus precluding the need for privacy-preserving data mining techniques.

3. Problem formulation

The objective is to predict the energy consumption in the next 24 h at each charging outlet based on two datasets namely charging record and station record, and comparing the two approaches. Formally, we assume there is a function relating the predicted available energy and the past consumed energy:

$$\widehat{E}(t) = f(E(t-1), E(t-2), \ldots) \tag{1}$$

where $E(t)$ is the actual energy consumption at time t, $\widehat{E}(t)$ is the prediction of the energy consumption at time t, and $(E(t-i))$ indicates the past energy consumption records.

As is the usual practice in forecasting, we are interested in finding an estimation of $E(t)$ according to a particular performance (or error) criterion. To this end, we have chosen Symmetric Mean Absolute Percentage Error (SMAPE). For day i, the SMAPE is defined as:

$$SMAPE(i) = \frac{1}{H}\sum_{t\in day i}\frac{|E(t)-\widehat{E}(t)|}{E(t)+\widehat{E}(t)} \times 100, \tag{2}$$

where H is the horizon of prediction in a given day ($H = 24$ in this paper).

In this paper the most recent ten percent of the data is used to evaluate the performance of the algorithm (test dataset). Note that the test dataset is not used in either the parameter selection or training phase.

We use the notation $\hat{\boldsymbol{E}}(t)$ as the vector of prediction for the next 24 h ending at t (Fig. 1a). Also, let $t_r = \{1, 2, \ldots, N_{tr}\}$ and $t_s = \{N_{tr}+1, \ldots, N\}$ be two sets of indices for the training and test sets, respectively. Later on, in the parameter selection phase, parts