



Schedulable capacity forecasting for electric vehicles based on big data analysis



Meiqin MAO¹ , Shengliang ZHANG¹, Liuchen CHANG^{1,2},
Nikos D. HATZIARGYRIOU³

Abstract Fast and accurate forecasting of schedulable capacity of electric vehicles (EVs) plays an important role in enabling the integration of EVs into future smart grids as distributed energy storage systems. Traditional methods are insufficient to deal with large-scale actual schedulable capacity data. This paper proposes forecasting models for schedulable capacity of EVs through the parallel gradient boosting decision tree algorithm and big data analysis for multi-time scales. The time scale of these data analysis comprises the real time of one minute, ultra-short-term of one hour and one-day-ahead scale of 24 hours. The predicted results for different time scales can be used for various ancillary services. The proposed algorithm is validated using operation data of 521 EVs in the field. The results show that compared with other machine learning

methods such as the parallel random forest algorithm and parallel k -nearest neighbor algorithm, the proposed algorithm requires less training time with better forecasting accuracy and analytical processing ability in big data environment.

Keywords Electric vehicle (EV), Schedulable capacity, Machine learning, Big data, Multi-time scale

1 Introduction

With increasing environmental concerns, electric vehicles (EVs) and renewable energy sources are receiving more and more attention all over the world [1]. According to the data from International Energy Agency [2], the number of global EV reached 3.1 million in 2017, increasing by 57% over the previous year, while the number of EV on the road is expected to reach 125 million by 2030. China will be the first country to start replacing traditional fuel vehicles with electric ones. At the same time, in 2030 renewable energy in China will account for 15% of the total energy consumption [3].

Large-scale integration of EVs and renewable energy into the grid poses great challenges in the operation of the power system due to their uncertainty and intermittent nature. Generally, large centralized energy storage systems (ESSs) can mitigate these problems, however, this would require expensive installations of large-capacity battery banks, pumped hydro and other large systems [4, 5].

Large-scale mobile and distributed ESS, composed of numerous on-board EV batteries can provide similar solutions, if their duality as ‘loads’ and ‘sources’ can be utilized and demand-side response technologies are applied [6, 7]. This allows to increase the penetration of renewable

CrossCheck date: 19 September 2019

Received: 28 January 2019 / Accepted: 19 September 2019 / Published online: 14 November 2019

© The Author(s) 2019

✉ Meiqin MAO
mmqmail@163.com

Shengliang ZHANG
slzhangmail@163.com

Liuchen CHANG
lchang@unb.ca

Nikos D. HATZIARGYRIOU
nh@power.ece.ntua.gr

¹ Research Center for Photovoltaic System Engineering, School of Electrical Engineering and Automation, Hefei University of Technology, Hefei 230009, China

² University of New Brunswick, Fredericton, NB E3B 5A3, Canada

³ National Technical University of Athens, 15780 Athens, Greece



power generation or improve the resiliency and stability by forming microgrids [8, 9].

An important prerequisite for EVs to provide ancillary services to utilities or efficient operation of microgrids is to forecast the EV schedulable capacity (EVSC) in a fast and accurate way. In this way, system operators can optimize the schedule for the participation of EVs in ancillary services.

In current literature, EVSC is generally obtained using probabilistic EV models [10–15], including plug-in time probability models based on binomial distributions [10], plug-in location probability models [11] and the aggregated queuing network model [12]. In [13–15], a Monte Carlo method is used to simulate the behavior of different types of EVs operating under realistic conditions, including start-stop time, charging rate, charging time, etc. In other non-probability models, the state of charge (SOC) of the EV batteries is used to obtain EVSC for individual and aggregated EVs [16, 17]. Several parameter hypotheses are needed in most of these models, also due to the scarcity of historical data.

With the development of communication and Internet of Things technologies, real-time operation data of individual EVs can be acquired from their battery management system (BMS). The large amounts of actual operation data such as SOC, times of EVs access to charging infrastructure, etc., enable to develop more accurate EVSC models. Nevertheless, dealing with the processing and analysis of a very large number of data poses great challenges. For example, if we assume that half of the 100 million EVs estimated on road in China by 2030 [2] will be involved in power system scheduling operation, and the collection interval of related information is one minute, the volume of data will reach 1–2 petabyte each year. Therefore, this paper treats the forecasting of EVSC based on real-time operation data of individual EVs as an essentially big data analysis problem.

Big data analysis and management are clear trends of future smart grids. This is challenging for traditional machine learning (ML) algorithms, since they are designed for a single machine and are not suitable to deal with big data [18]. Thus, more efficient ML algorithms for parallel computing or for big data are required.

The parallel processing methods proposed in literature can be divided into three groups. Group one refers to parallel processing of traditional algorithms by using Hadoop and Spark cluster technology [19–21]. Group two is the combination of clustering or optimization algorithms and traditional ML algorithms [22–24]. Group three is a combination of group one and group two [25–27]. In group one, the parallelization of the algorithm effectively improves the computing speed and accuracy of load forecasting in parallel computing framework MapReduce and

Spark. For example, [19] analyses the forecasting time and error for data sets with different sizes in different sizes of Hadoop clusters. In group two, clustering algorithms on large-scale data sets can be used to improve the performance markedly [22, 23]. For group three, [25] and [26] propose new hybrid algorithms, which combine the improved particle swarm optimization and extreme learning machine, fuzzy C clustering and support vector machines (SVM), respectively. The problems of over-fitting and long training time caused by the increase of data scale are faced by multi-distributed back propagation (BP) neural networks [27].

Different from the algorithms above, [19] and [28] propose ensemble learning algorithms of random forest (RF) and gradient boosting decision tree (GBDT), respectively. Ensemble learning algorithm integrates multiple base learners into a strong learner to improve the forecasting accuracy. It is considered as one of the important future research directions of ML [29]. Unlike traditional multi-linear regression (MLR) algorithm [30], GBDT can flexibly handle a certain number of different types of feature attributes, including continuous and discrete values [31]. Thus, it is widely used in traffic and load forecasting. However, since the output of the algorithm is the result of multiple iterations, there is a strong inter-dependence among regression trees, thus it is difficult to realize the parallelization of the GBDT algorithm.

The parallel GBDT (PGBDT) algorithm is derived from the GBDT algorithm and enables parallel computations. It requires less iteration time than GBDT by parallel processing of a large number of gradient and optimization computations without affecting the prediction accuracy of the model. Thus it is applicable to the big data scenario, although it has not been used so far for EVSC forecasting (EVSCF).

The application of big data analysis equipped with ML algorithms has been mainly found in the field of load forecasting, but rarely for EVSCF. In [15] and [32], the application of SVM, RF and decision tree (DT) algorithms are investigated for EVSCF. It is found that SVM and RF have an improved performance, when the forecasting curves of EVSCF fluctuate less, while RF is more effective than SVM, when there are large fluctuations [15]. DTs are heavily dependent on their input data, which means that even small variations in data may result in large changes in the structure of the optimal DT. In this paper, in order to address this problem, two new algorithms, suitable for big data analysis, i.e., PGBDT and parallel k-nearest neighbors (PKNN), are applied to the EVSCF problem and the results are analyzed.

With the rapid growth of EVs, un-controlled charging of a large number EVs may cause the phenomenon of “peak peaking”, i.e., increase the peak-to-valley difference of the utility and affect the stable operation of power grid.

EVSCF methods provide strong data support for load peak shifting, frequency regulation, economic dispatch and intelligent EV charging/discharging strategies. These different applications require the results of EVSCF for the scheduling of renewable energy or load at different time scales [33, 34]. For example, real-time load forecasting has a time horizon of several seconds to 10 minutes and is used for frequency/voltage regulation, in order to eliminate the effect of volatility of renewable energies [35]; ultra-short-term load forecasting has a time scale of one hour or short-term load forecasting of several hours to tens of hours for economic dispatch and peak shaving and valley filling [36, 37]. Forecasting of renewable generation is usually required for ultra-short-term 15 minutes to four hours ahead and for short-term 24–72 hours ahead [38].

So far, time scaling of EVs for power system operation has not been properly discussed. References [15] and [31] deal only with real time and one-day-ahead time scales. In this paper, ultra-short-term scaling of one hour is additionally incorporated for the first time in the EVSCF models. In this way, EVs can be used for more power system services such as real-time optimization, peak shaving and valley filling, economic dispatch, etc.

Overall, the main contribution of this paper is the development of EVSCF models for multi-time scales based on the PGBDT algorithm which is used to forecast EVSC faster and more accurately.

The results of real-time EVSCF based on a large amount of real-time operation data from BMS of individual EVs are used as historical data for training the EVSCF models for ultra-short-term scale of one hour and one-day-ahead scale of 24 hours. The PGBDT algorithm is initially proposed and tested on a big data platform for multi-time scale EVSCF models to prove its feasibility and effectiveness.

The rest of this paper is organized as follows. In Section 2, the PGBDT algorithm is described. Section 3 discusses EVSCF models for multi-time scales. In Section 4, the proposed models are validated and compared with parallel random forest (PRF) and PKNN algorithms on a big data platform. This is followed by conclusions in Section 5.

2 PGBDT algorithm

In the PGBDT algorithm [28], a training sample is composed of a 1-dimensional target vector y_i and a set of K -dimensional input vectors $\mathbf{x}_i = [x_i^1, x_i^2, \dots, x_i^K]$. The objective is to obtain $f^*(\mathbf{x})$ that is, mapping \mathbf{x} to y_i in training samples $S = \{(y_i, \mathbf{x}_i)\}_{i=1}^n$ of known value y_i and i^{th} set of feature attributes \mathbf{x}_i with the length K , while minimizing the loss function $L(y_i, f(\mathbf{x}_i))$ as shown in (1):

$$f^*(\mathbf{x}) = \arg \min_{f(\mathbf{x})} \sum_{i=1}^n L(y_i, f(\mathbf{x}_i)) \quad (1)$$

where $f(\mathbf{x}_i)$ is the i^{th} predicted value as the output of a mapping function. The loss function $L(y_i, f(\mathbf{x}_i))$ used here is the square error loss function of the regression problem as shown in (2):

$$L(y_i, f(\mathbf{x}_i)) = (y_i - f(\mathbf{x}_i))^2 \quad (2)$$

$f^*(\mathbf{x})$ can be approximated by the additive form of $f(\mathbf{x})$ as shown in (3):

$$f(\mathbf{x}) = \sum_{m=1}^M c_m b_m \quad (3)$$

where c_m is a scaling factor; and b_m is the least-squares coefficient of the base learner for the m^{th} iteration.

By defining the base learner as a J -terminal node regression tree, the specific steps of PGBDT algorithm are as follows.

Step 1: Initialize the model (3) by setting the initial and maximum number of iterations m as one and M , respectively, and the initial function $f_0(\mathbf{x})$ as shown in (4).

$$f_0(\mathbf{x}) = \arg \min_p \sum_{i=1}^n L(y_i, p) \quad (4)$$

where p is a constant value for minimizing the loss function; and $f_0(\mathbf{x})$ is a regression tree with only one node.

Step 2: Obtain the negative gradient of the loss function as shown in (5), and $f_{m-1}(\mathbf{x})$ is the model after $(m-1)^{\text{th}}$ iteration.

$$r_{mi} = - \left(\frac{\partial L(y_i, f(\mathbf{x}_i))}{\partial f(\mathbf{x}_i)} \right)_{f(\mathbf{x})=f_{m-1}(\mathbf{x})} \quad (5)$$

The m^{th} regression tree is constructed according to all samples and their negative gradients [39], and its splitting rule is to divide it into two regions according to the value s of the k^{th} feature attribute: $S_{\text{left}}(k, s) = \{x_i | x_i^k \leq s\}$ and $S_{\text{right}}(k, s) = \{x_i | x_i^k > s\}$. The minimization of the sum of regional variances after splitting is shown in (6):

$$\text{Gain}(k, s) = \min_{k, s} \left[\sum_{x_i \in S_{\text{left}}} \left(y_i - \frac{1}{n_{\text{left}}} \sum_{i=1}^{n_{\text{left}}} y_i \right)^2 + \sum_{x_i \in S_{\text{right}}} \left(y_i - \frac{1}{n_{\text{right}}} \sum_{i=1}^{n_{\text{right}}} y_i \right)^2 \right] \quad (6)$$

where the training sample S with size n is divided into left-dataset S_{left} and right-dataset S_{right} according to splitting s , the size of which are n_{left} and n_{right} , respectively.



Thus, its corresponding J -terminal node regions $R_{mj}, j = 1, 2, \dots, J$ are obtained.

Step 3: Obtain the corresponding least-squares coefficient b_{mj} of the m^{th} regression tree as in (7).

$$b_{mj} = \bar{r}_{mi} \varepsilon(\mathbf{x}_i) \quad \mathbf{x}_i \in R_{mj} \tag{7}$$

$$\varepsilon(\mathbf{x}_i) = \begin{cases} 1 & \mathbf{x}_i \in R_{mj} \\ 0 & \mathbf{x}_i \notin R_{mj} \end{cases} \tag{8}$$

where \bar{r}_{mi} is the average value of negative gradient of m^{th} regression tree; and $\varepsilon(\mathbf{x}_i)$ is an indicator function.

Step 4: Find the scaling factor c_m of the m^{th} regression tree for solving the ‘‘linear search’’ by (9).

$$c_m = \arg \min_c \sum_{\mathbf{x}_i \in R_{mj}} L\left(y_i, f_{m-1}(\mathbf{x}_i) + c \sum_{j=1}^J b_{mj}\right) \tag{9}$$

Step 5: Update the model $f_m(\mathbf{x})$ as (10).

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + c_m \sum_{j=1}^J b_{mj} \tag{10}$$

Step 6: If $m < M$, let $m = m + 1$ and repeat *Step 2* to *Step 5*, otherwise output the final $f_M(\mathbf{x})$.

After M iterations, the final model $f(\mathbf{x})$ is obtained as (11):

$$f(\mathbf{x}) = f_M(\mathbf{x}) = f_0(\mathbf{x}) + \sum_{m=1}^M \sum_{j=1}^J c_m b_{mj} \tag{11}$$

From (11), it can be seen that the PGBDT algorithm is a combined algorithm. It approximates the expected model by iterating a series of regression trees to improve the model accuracy and provides a strong predictive performance and generalization ability. Each regression tree can be parallelized by finding splits on each non-terminal node of the regression tree in parallel, whose splitting criteria depends on the minimization of variance after splitting. Therefore, the whole model of PGBDT can be parallelized by generating each regression tree in parallel during its generation process.

3 EVSCF models for multi-time scales

3.1 Time scales used in the proposed EVSCF models

In this paper, three time scales for EVSCF are proposed, i.e., real-time scale of one minute, ultra-short-term scale of one hour and one-day-ahead scale of 24 hours. This allows the provision of different ancillary services for power system, as shown in Fig. 1. For the one-day-ahead EVSCF, the forecasting is performed for every of the 24 hours in

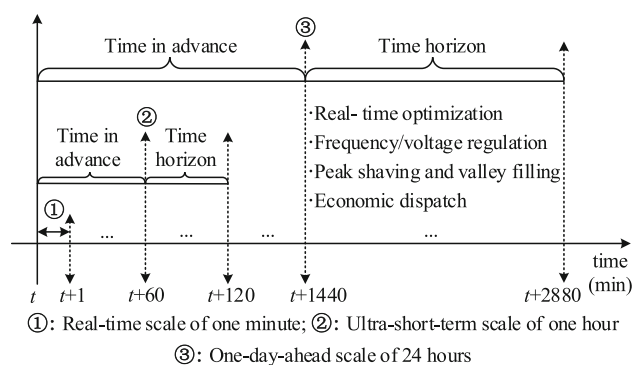


Fig. 1 Diagram of time scales in the proposed EVSCF

advance. Considering the uncertainty of the one-day-ahead scheduling, the schedule for the ultra-short-term scale of one hour is formulated to reduce the forecasting errors, which is performed one hour in advance. The real-time EVSCF is carried out one minute in advance. Its time scale is short and the precision is high and less affected by uncertain factors. It can be used for frequency and voltage regulation and further correction of scheduling errors. The real-time scale interval is one minute and the time scales for ultra-short-term and one-day-ahead are set to 60 minutes and 1440 minutes, respectively.

3.2 Real-time EVSCF model

3.2.1 Classification of EVs connected to grid

The proposed real-time EVSCF model is based on real-time data of individual EVs, which are acquired from the BMS of each EV.

In order to ensure the accuracy of real-time EVSCF model, the selected time scale for the prediction model is equal to the time interval of the real-time operation data acquisition, which is one minute in this paper. Since the SOC of the EV battery changes slightly within one minute, EVSC in real time is calculated dynamically through the real-time data acquisition and big data analysis method, and can be regarded as the forecasted value of EVSC for the next minute. To build the EVSCF model in real time, it is necessary firstly to classify the individual EVs accessing the utility network according to their levels of SOC so that the aggregated charging or discharging capacity of EVs can be obtained.

The participation of EVs in dispatch mainly depends on the scheduling period from the grid side t_s , the expected remaining time to disconnect from the grid $t_{d,l}$ and the minimum charging time to reach expected SOC $t_{d,c}$ from individual EV users. The rules to classify EVs are as follows [15].

- 1) If $t_{d,l} < t_s$ or $t_{d,l} < t_{d,c}$, EV_d is not allowed to participate in the scheduling plan.
- 2) If $t_{d,l} \geq t_s$ and $t_{d,l} \geq t_{d,c}$: ① if $SOC_d^t < SOC_d^{\min}$, EV_d is allowed to be charged; ② if $SOC_d^{\max} < SOC_d^t$, EV_d is allowed to be discharged; ③ if $SOC_d^{\min} < SOC_d^t < SOC_d^{\max}$, EV_d is allowed to be charged or discharged according to the scheduling plan.

Note that EV_d is the d^{th} EV; SOC_d^t is the SOC of EV_d at current time t ; and SOC_d^{\min} and SOC_d^{\max} are the minimum and maximum expected SOC for each EV_d , respectively.

3.2.2 Definition of charging/discharging rate

The charging/discharging rate $v_{d,t}$ is used to characterize the users' demands, as shown in (12). The value of $v_{d,t}$ is related to the initial SOC, ending SOC and $t_{d,l}$ of EV_d .

$$v_{d,t} = \frac{SOC_d^t - SOC_d^{t+t_{d,l}}}{t_{d,l}} \quad (12)$$

where $SOC_d^{t+t_{d,l}}$ is the SOC of EV_d at time $t + t_{d,l}$; $v_{d,t} < 0$ indicates that EV_d is charging; and $v_{d,t} > 0$ indicates that EV_d is discharging.

3.2.3 Real-time EVSCF model

The real-time EVSCF model includes real-time schedulable charge capacity (SCC) and schedulable discharge capacity (SDC) of EVSC based on real-time operation data of EVs from BMS and group classification above. In this paper, SCC and SDC of individual EVs are obtained from (13) and (14), respectively:

$$SCC_{d,t_s} = v_{d,t_s} C_d \quad v_{d,t} \in [\underline{v}_{d,t}, 0] \quad (13)$$

$$SDC_{d,t_s} = v_{d,t_s} C_d \quad v_{d,t} \in [0, \bar{v}_{d,t}] \quad (14)$$

where $\bar{v}_{d,t}$ and $\underline{v}_{d,t}$ are the limits of charging/discharging rate, respectively; SCC_{d,t_s} and SDC_{d,t_s} are SCC and SDC of EVSCF of EV_d at the scheduling period t_s , respectively; and C_d is the rated battery capacity of EV_d .

According to the real-time SCC and SDC of individual EVs as well as SCC and SDC of EVSCF of the cluster of EVs, $SCC_{t_s}^{all}$ and $SDC_{t_s}^{all}$ are derived as in (15) and (16), respectively:

$$SCC_{t_s}^{all} = \sum_{d=1}^N SCC_{d,t_s} \quad (15)$$

$$SDC_{t_s}^{all} = \sum_{d=1}^N SDC_{d,t_s} \quad (16)$$

where N is the total number of EVs connected to the grid.

3.3 Ultra-short-term and one-day-ahead EVSCF models

3.3.1 Construction of training dataset and testing dataset

Feature selection is required before establishing the training dataset and testing dataset for ultra-short-term and one-day-ahead EVSCF models. The following feature attributes of EVs are selected to train the model based historical data and time attributes of operations for individual EV.

- 1) The average values of SCC and SDC of EVSC at the same time t of the previous month are $\overline{SCC}_{t,mon}^{all}$ and $\overline{SDC}_{t,mon}^{all}$, which are calculated as in (17) and (18), respectively:

$$\overline{SCC}_{t,mon}^{all} = \frac{1}{l} \sum_{k=1}^l SCC_{t,mon}^k \quad (17)$$

$$\overline{SDC}_{t,mon}^{all} = \frac{1}{l} \sum_{k=1}^l SDC_{t,mon}^k \quad (18)$$

where l is the total number of days of the previous month; $SCC_{t,mon}^k$, $SDC_{t,mon}^k$ are the values of SCC and SDC of EVSC at the same time t on the k^{th} day of the previous month, respectively.

- 2) The average values of SCC and SDC of EVSC at the same time t last week are $\overline{SCC}_{t,week}^{all}$ and $\overline{SDC}_{t,week}^{all}$, which are calculated as in (19) and (20), respectively, where $SCC_{t,week}^k$, $SDC_{t,week}^k$ are the values of SCC and SDC of EVSC at time t on the k^{th} day of last week:

$$\overline{SCC}_{t,week}^{all} = \frac{1}{7} \sum_{k=1}^7 SCC_{t,week}^k \quad (19)$$

$$\overline{SDC}_{t,week}^{all} = \frac{1}{7} \sum_{k=1}^7 SDC_{t,week}^k \quad (20)$$

- 3) The values of SCC and SDC of EVSC at the same time t of the previous day are $SCC_{t,day}$ and $SDC_{t,day}$.

According to different time attributes, the following four feature attributes are selected as inputs at the training stage: current time t (a total number of 1440 time slots, represented by 0 to 1439), indication of rush hour, holiday or working time.

In summary, through the correlation analysis, the data set with length q is divided into two parts: training dataset



with length p and testing dataset with length $q-p$. The next step is to construct EVSCF models for ultra-short-term and one-day-ahead scales, as follows.

3.3.2 Ultra-short-term and one-day-ahead EVSCF models

Ultra-short-term and one-day-ahead EVSCF models with PGBDT proposed in this paper differentiate only in the time scales. Both are trained as follows:

- 1) Input training dataset A including the feature attributes and actual value of EVSC y_{t-p} , $A = \{(y_{t-p}, \mathbf{h}_{t-p}, \mathbf{w}_{t-p})\}_{t=1}^p$ comprises 10 feature attributes of the training dataset with length p ; $\mathbf{h}_{t-p} = [x_{t-p}^1, x_{t-p}^2, \dots, x_{t-p}^6]$ is a 6-dimensional vector with the historical data of EVSC; and $\mathbf{w}_{t-p} = [x_{t-p}^7, x_{t-p}^8, x_{t-p}^9, x_{t-p}^{10}]$ is a 4-dimensional vector with the time attributes of EVSC.
- 2) Set the parameters of the PGBDT algorithm including the number of iterations I and maximum depth d .
- 3) Train the model represented by (21) by the training dataset A :

$$y_{t-p} = f(\mathbf{h}_{t-p}, \mathbf{w}_{t-p}) \quad (21)$$

- 4) Substitute testing dataset B into the model, and obtain the predicted value of EVSCF y_t^e as (22):

$$y_t^e = f(\mathbf{h}_{t-p}, \mathbf{w}_{t-p}) \quad (22)$$

$B = \{(y_t, \mathbf{h}_t, \mathbf{w}_t)\}_{t=p+1}^q$ has 10 feature attributes of the testing dataset with length $q-p$.

3.3.3 Evaluation indexes

In order to evaluate the performance of the proposed PGBDT algorithm for the ultra-short-term and one-day-ahead EVSCF models, the mean absolute percentage error (MAPE) and root mean square error (RMSE) are chosen as evaluation indexes. The expressions are shown in (23) and (24), respectively:

$$MAPE = \frac{1}{n} \sum_{i=1}^n |(y_i - y_i^e)/y_i| \times 100\% \quad (23)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n [(y_i - y_i^e)/y_i]^2} \times 100\% \quad (24)$$

where y_i and y_i^e are the actual and forecasted EVSC values, respectively. If y_i is 0, it is replaced with the historical average of EVSC. The smaller the value of MAPE, the more accurate the predicted value is. RMSE is sensitive to outliers and can amplify the prediction errors. It can be used to evaluate the stability of the algorithm.

3.4 Implementation of EVSCF models with PGBDT algorithm and big data analysis

3.4.1 Real-time EVSCF framework based on big data

Equations (12)–(16) form the real-time EVSCF model. Although the proposed model looks simple, it is difficult to apply, since it needs to process the large amount of related data of EVs in real time. In this paper, Hadoop is used to solve the storage problem of big data by the Hadoop distributed file system (HDFS) [40]. Moreover, Spark designed for large-scale data processing is used. The Spark streaming can process stream data with a minimum interval of 500 ms. In this paper, the real-time processing interval is 60 s, which enables parallel computation to meet real-time requirements.

Parallel processing on Spark is shown in Fig. 2a. When EVs are connected to the grid, their operation information can be acquired and processed through the following functions. The Map function calculates the real-time EVSC of individual EVs. The ReduceByKey function combines the value (the output of Map function) of each key (the number of EVs). Real-time EVSCF values are obtained from the real-time EVSCF model. This is used as historical data for ultra-short-term and one-day-ahead EVSCF models, as discussed in the following sections.

3.4.2 Framework of EVSCF models based on PGBDT

The structure of EVSCF models based on PGBDT algorithm for multi-time scales is shown in Fig. 2. The real-time EVSCF model is built, as shown in Fig. 2a, and the historical data of the real-time EVSCF is combined with the time attributes to generate the training dataset and testing dataset. According to the different prediction periods, the training dataset and testing dataset are updated in order to apply rolling forecasting. Finally, one-day-ahead and ultra-short-term EVSCF models based on the PGBDT algorithm are trained, tested and evaluated, as shown in Fig. 2b.

4 Study cases

4.1 Big data platform configuration

Combining the advantages of Hadoop and Spark, a big data platform is constructed to test the proposed method. The hardware of configured big data platform consists of two IBM servers, which can communicate on the same network through Gigabit gateway. Based on Ubuntu 64-bit operation systems, a computer cluster containing four machines is set up, one of which is selected as the master

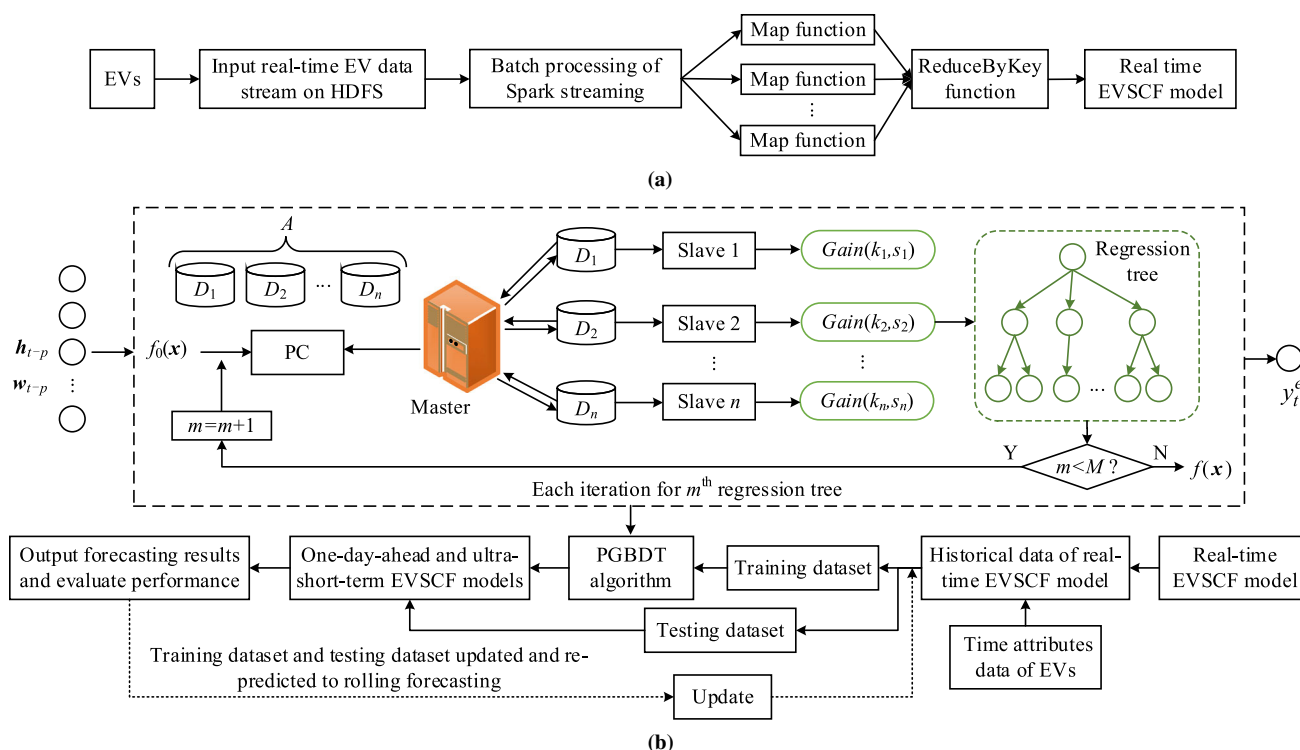


Fig. 2 Structure of EVSCF models for multi-time scales

Table 1 Big data platform configuration parameters

| Master/slave node | IP address | Software version |
|-------------------|----------------|----------------------------|
| Master | 192.168.16.135 | |
| Slave1 | 192.168.16.198 | Hadoop-2.7.0, Spark-1.6.0- |
| Slave2 | 192.168.16.199 | bin-hadoop-2.7.0 |
| Slave3 | 192.168.16.229 | |

Table 2 Processing time for different data size of real-time EVSCF

| Data scale (GB) | T_c (s) | T_s (s) | $S_{speedup}$ |
|-----------------|-----------|-----------|---------------|
| 0.5 | 22 | 246 | 11 |
| 1.0 | 37 | 643 | 17 |
| 5.0 | 109 | 2808 | 25 |
| 17.0 | 196 | 13030 | 66 |

node and the rest three as the slave nodes. The configuration parameters of the big data platform are shown in Table 1.

With the big data platform, the real-time data of 521 EVs are used to test the EVSCF models proposed in Section 3. These data are acquired from BMS of each EV with one-minute resolution, (17 GB in total) in the period from Nov. 1, 2015, 00:00 to Apr. 30, 2016, 23:59 [15].

4.2 Processing time analysis of different real-time EVSCF data scales

In this paper, the speed-up factor $S_{speedup}$ is defined as an evaluation index for measuring the parallelization degree of the big data platform, as shown in (25):

$$S_{speedup} = T_s/T_c \tag{25}$$

where T_s and T_c are the running time of single machine and cluster machines for processing big data, respectively.

With the proposed big data platform and the real-time operation data of EVs, real-time EVSCF is performed. Table 2 shows the values of $S_{speedup}$ for the processing methods by single machine and cluster machines. The achieved EVSCF increases from 0.5 GB to 17 GB.

It can be seen from Table 2 that with the increasing data size of real-time EVSCF, the speed-up factor increases from 11 to 66. The acceleration effect is obvious, reflecting the ability of the proposed method to process large-scale data.

4.3 Simulation results and discussions

4.3.1 Real-time EVSCF

The real-time EVSCF is performed using the operation data of EVs connected to the grid from Nov. 1, 2015, 00:00 to Apr. 30, 2016, 23:59 with one-minute resolution,



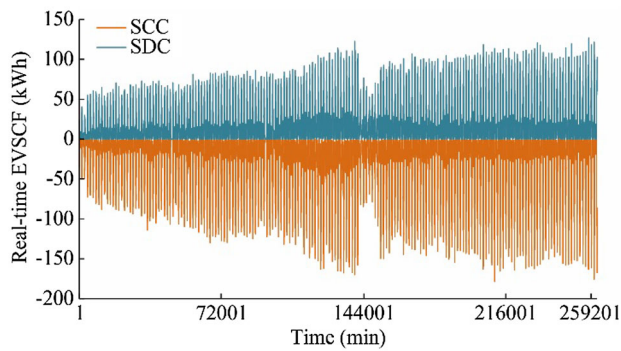


Fig. 3 Real-time EVSCF during Nov. 2015 to Apr. 2016

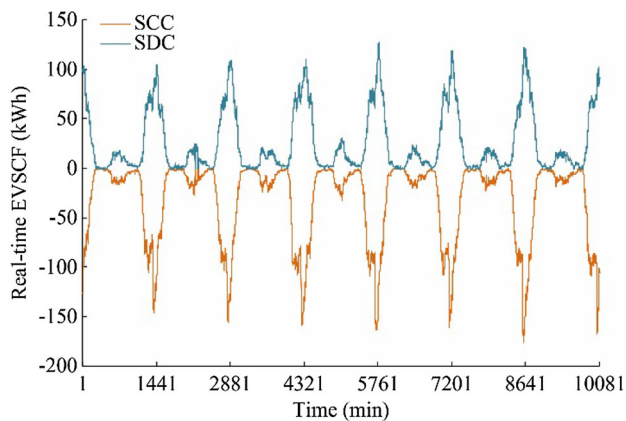


Fig. 4 Real-time EVSCF of a week during Apr. 24, 2016 to Apr. 30, 2016

including a total number of 262080 data points. The predicted results of both SCC and SDC are shown in Fig. 3. It can be seen that there is a drop in the curves of both SCC and SDC during the course Feb. 7, 2016 to Feb. 13, 2016 (time: 141120–151200 minutes), which is explained by the Chinese Spring Festival.

Figure 4 depicts SCC and SDC curve form Apr. 24, 2016 to Apr. 30, 2016. The daily trend of EVSCF is basically the same, because the travel time of EV buses is nearly the same every day. Since EVs will leave or access the grid at any time, the values of EVSCF always change and reflect the volatility of EVSC.

Figure 5 shows the results of real-time EVSCF for the specific day of Apr. 30, 2016. From Fig. 5, it is expected that the maximum values of SCC and SDC during this day are 167.557 kWh and 117.155 kWh, respectively, and the minimum values are zero. Figure 5 shows that the values of real-time EVSCF for the time periods from 05:00 to 08:59 (time: 301–540 minutes) and 16:00 to 18:59 (time: 961–1140 minutes) are close to zero, which reflects the intermittency of EVSC. For EVs, this happens during rush hours when EVs have completed charging and are

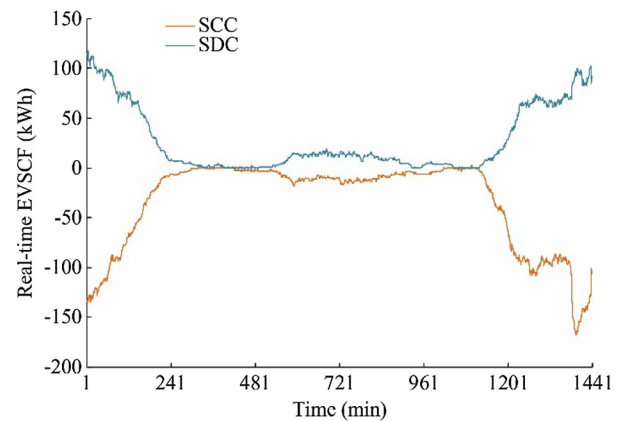


Fig. 5 Real-time EVSCF of Apr. 30, 2016

disconnected from the grid. Therefore, during these periods, there are few EVs participating in grid scheduling.

In summary, EVSC for EVs is lower during the daytime, close to 0 during rush hours and higher during the night. The time characteristics of EVSC are consistent with the operation frequency of buses. The probability of access to grid at night is much higher than at daytime, which results in higher EVSC for EV buses at night. Based on this characteristic, charging of EVs can be shifted, not only reducing the peak power, but also being charged at low electricity prices. Operation regularity also provides the basis for EVSC predictability. The analysis results of big data show the characteristics of EVSC, namely, volatility, intermittent and predictability.

4.3.2 Ultra-short-term EVSCF

For ultra-short-term and one-day-ahead EVSCF models, the real-time historical EVSC data from Nov. 1, 2015, 00:00 to Apr. 23, 2016, 23:59 are used for training datasets, while the historical EVSC data from Apr. 24, 2016, 00:00 to Apr. 30, 2016, 23:59 are used for testing datasets. Therefore, ultra-short-term EVSCF is set an hour in advance to forecast the next hour, rolling to the 168th hour (7×24 hours).

To demonstrate the effectiveness of the proposed method, the results from the PRF and PKNN algorithms [32, 41] are compared with the results of the PGBDT algorithm proposed in this paper. Taking into account the accuracy and processing time, the set of parameters for different algorithms are selected, as shown in Table 3.

The errors of SCC and SDC in MAPE and RMSE and the training time for ultra-short-term EVSCF obtained by the three ML algorithms are shown in Table 4. It can be seen that PGBDT has the best performance in both accuracy and training time, and the MAPE of SCC by PGBDT are 6.52% and 24.01% lower than those of PRF and PKNN,

Table 3 Set of parameters for different algorithms

| Algorithm | Set of parameters |
|-----------|---|
| PGBDT | Number of iterations $I = 4$; maximum depth $d = 8$ |
| PRF | Number of trees $T = 40$; number of bins $B = 64$; Maximum depth $d = 5$ |
| PKNN | Number of nearest neighbors $k = 100$ |

Table 4 Prediction errors and training time of ML algorithms for ultra-short-term EVSCF

| Algorithm | Training time (s) | MAPE (%) | RMSE (%) |
|-----------|-------------------|----------|----------|
| SCC-PGBDT | 6.59 | 3.79 | 4.79 |
| SCC-PRF | 9.82 | 10.31 | 16.09 |
| SCC-PKNN | 20.07 | 27.80 | 41.10 |
| SDC-PGBDT | 6.11 | 3.37 | 3.96 |
| SDC-PRF | 9.53 | 9.87 | 16.44 |
| SDC-PKNN | 20.78 | 27.89 | 37.15 |

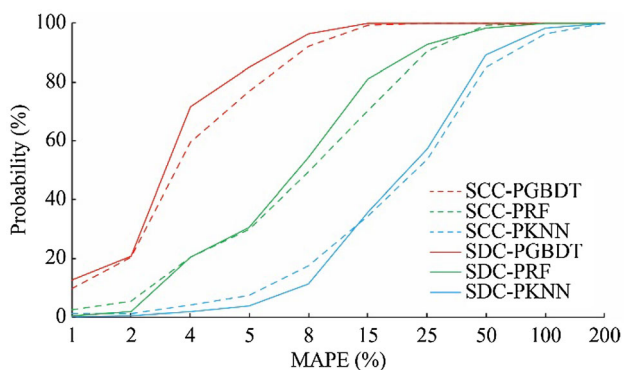


Fig. 6 Cumulative probability curves of ultra-short-term EVSCF

respectively. Similarly, MAPE of SDC by PGBDT are 6.50% and 24.52% lower than those of PRF and PKNN, respectively. The training times of SCC by PGBDT are 3.23 s and 13.48 s faster than those of PRF and PKNN,

respectively. Overall, the prediction accuracy and training time of PKNN are much inferior to those of PGBDT.

In order to quantitatively evaluate the reliability of PGBDT, PRF and PKNN algorithms, the cumulative probability curves of 168 hours of ultra-short-term EVSCF are obtained, as shown in Fig. 6. As can be seen, PGBDT has more than 92% of its results meeting the requirements of MAPE within 8%, which means that PGBDT has good generalization ability for 92% of new samples. At the same level of error, the data volume of PRF is about 50%, while that of PKNN is less than 20%. This reflects that the reliability of PGBDT is the highest among the three algorithms.

Figure 7 shows the actual value and predictive values of SDC for different algorithms for a typical day of Apr. 30 from 00:00 to 24:00. It can be seen that the curve of PGBDT values is consistent with the curve of actual values, and the other two curves have large deviations during the two selected periods. The amplitude of SDC changes more than 70% in two hours and within 10% in two hours, respectively.

To further evaluate the performance in different time periods, 24 hours are divided into three periods according to the operation practice of EVs, including peak hours of EVSC (00:00–02:59, 20:00–23:59), flat hours of EVSC (03:00–04:59, 09:00–15:59, 19:00–19:59), valley hours of EVSC (05:00–08:59, 16:00–18:59). The histograms of the evaluation indexes are shown in Figs. 8 and 9. As can be seen from Figs. 8 and 9, the errors between PGBDT and PRF are slightly different in peak hours. But in flat hours and valley hours, due to multiple iteration errors, the results from PGBDT are stable and much better than that of PRF and PKNN.

4.3.3 One-day-ahead EVSCF

One-day-ahead EVSCF is performed one day in advance for the next day. For example, in one-day-ahead 24-hour

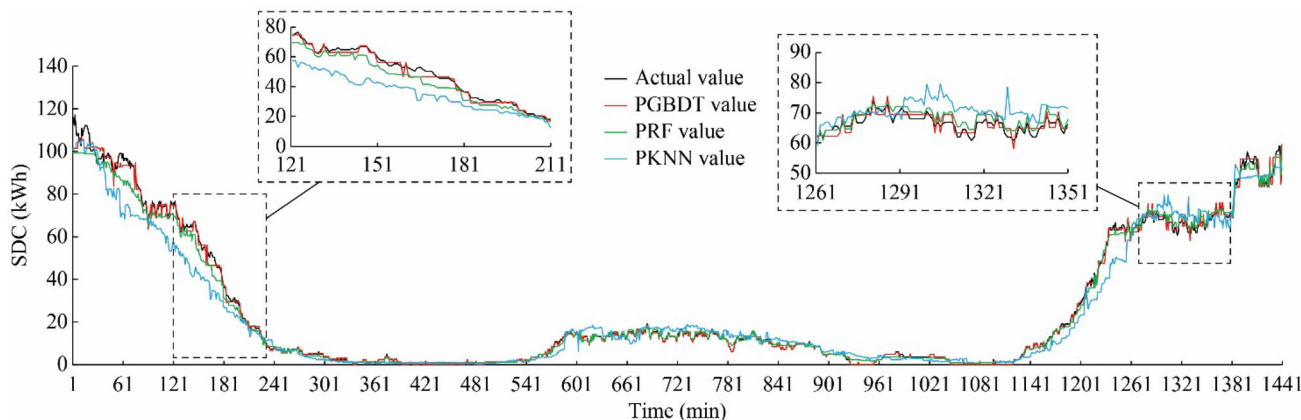


Fig. 7 Forecasting errors of SDC by three algorithms on Apr. 30, 2016, 00:00–24:00



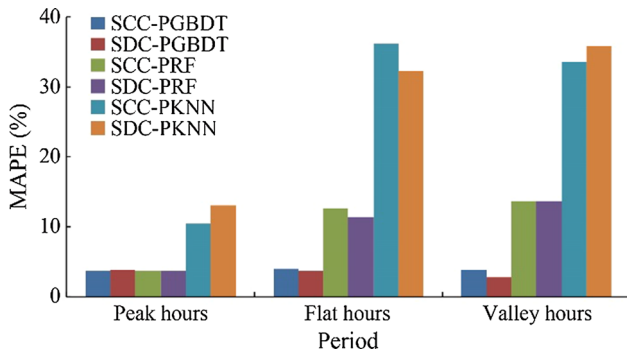


Fig. 8 Comparison of MAPE in different time periods

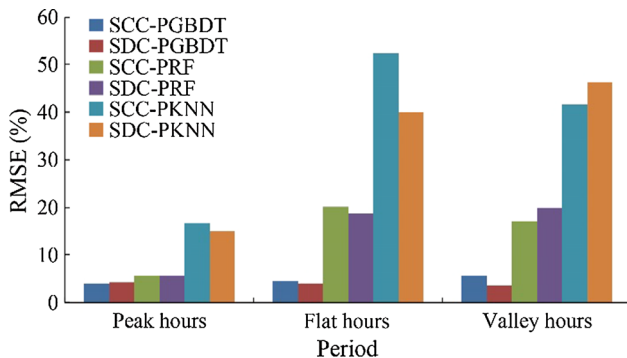


Fig. 9 Comparison of RMSE in different time periods

Table 5 Prediction errors and training time of ML algorithms for one-day-ahead EVSCF

| Algorithm | Training time (s) | MAPE (%) | RMSE (%) |
|-----------|-------------------|----------|----------|
| SCC-PGBDT | 8.19 | 4.11 | 4.15 |
| SCC-PRF | 11.63 | 11.00 | 19.45 |
| SCC-PKNN | 20.64 | 29.18 | 47.94 |
| SDC-PGBDT | 8.59 | 3.97 | 3.99 |
| SDC-PRF | 14.78 | 10.35 | 21.09 |
| SDC-PKNN | 21.25 | 28.02 | 62.44 |

EVSCF model, predicting that EVSC of Apr. 24, 2016 needs a training dataset from Nov. 1, 2015 to Apr. 23, 2016, predicting that EVSC of Apr. 25, 2016 needs a training dataset from Nov. 2, 2015 to Apr. 24, 2016, etc, in a rolling way, EVSC of Apr. 30, 2016 is predicted. Table 5 shows the forecasting errors of SCC and SDC in MAPE and RMSE and the training time based on PGBDT, PRF and PKNN algorithms for one-day-ahead EVSCF. Similar to the results of Table 4, PGBDT is the best among all the algorithms. Comparing with the results in Tables 4 and 5, it can be seen that the smaller the time scale of EVSCF is, the smaller the forecasting errors in MAPE are. The value of RMSE does not vary with the prediction time scale, and only depends on the complexity of the data and the amount of outlier data.

5 Conclusion

This paper investigates the EVSCF using big data analysis and ML algorithms. EVSCF models are established for multi-time scales based on actual operation data of EVs. Real-time EVSCF is achieved using the constructed big data platform, where the speed of Hadoop and Spark is 66 times faster than traditional methods. The proposed models are tested and compared with PRF and PKNN, exhibiting superior performance. The simulation results containing real operation data of EVs connected to the grid with one-minute resolution. It shows that for one-hour ultra-short-term EVSCF model, the PGBDT algorithm has the highest accuracy for SCC and SDC, with the forecasting errors in MAPE of 3.79% and 3.37%, and reduced training time by 30% and 60%, respectively, compared with those obtained by PRF and by PKNN. The performance of PGBDT-based EVSCF model for one-day-ahead 24 hours is much better than PRF and PKNN, proving its reliable forecasting performance and generalization ability. The simulation results also prove that the proposed PGBDT-based EVSCF models can take advantage of the analytical ability of ML under a big data environment and provide powerful support for EV participation in grid scheduling and ancillary services.

Acknowledgement This work was supported by National Natural Science Foundation of China (No. 51577047) and International Collaboration Project supported by Bureau of Science and Technology, Anhui Province (No. 1604b0602015).

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- [1] Mahmud K, Town GE, Morsalin S et al (2018) Integration of electric vehicles and management in the internet of energy. *Renew Sustain Energy Rev* 82(3):4179–4203
- [2] International Energy Agency (2018) Global EV outlook 2018. <https://webstore.iea.org/global-ev-outlook-2018>. Accessed 30 May 2018
- [3] Wang B, Wang Q, Wei Y et al (2018) Role of renewable energy in China's energy security and climate change mitigation: an index decomposition analysis. *Renew Sustain Energy Rev* 90:187–194
- [4] Abapour S, Nojavan S, Abapour M (2018) Multi-objective short-term scheduling of active distribution networks for benefit maximization of DisCos and DG owners considering demand response programs and energy storage system. *J Mod Power Syst Clean Energy* 6(1):95–106

- [5] Feng X, Gu J, Guan X (2018) Optimal allocation of hybrid energy storage for microgrids based on multi-attribute utility theory. *J Mod Power Syst Clean Energy* 6(1):107–117
- [6] Tushar MHK, Zeineddine AW, Assi C (2018) Demand-side management by regulating charging and discharging of the EV, ESS, and utilizing renewable energy. *IEEE Trans Ind Inf* 14(1):117–126
- [7] Zhu J, Gu W, Jiang P et al (2018) Integrated approach for optimal island partition and power dispatch. *J Mod Power Syst Clean Energy* 6(3):449–462
- [8] Poudel S, Dubey A (2018) Critical load restoration using distributed energy resources for resilient power distribution system. *IEEE Trans Power Syst* 34(1):52–63
- [9] Wang M, Mu YF, Jiang T et al (2018) Load curve smoothing strategy based on unified state model of different demand side resources. *J Mod Power Syst Clean Energy* 6(3):540–554
- [10] Han S, Han S, Sezaki K (2011) Estimation of achievable power capacity from plug-in electric vehicles for V2G frequency regulation: case studies for market participation. *IEEE Trans Smart Grid* 2(4):632–641
- [11] Agarwal L, Peng W, Goel L (2014) Probabilistic estimation of aggregated power capacity of EVs for vehicle-to-grid application. In: Proceedings of 2014 international conference on probabilistic methods applied to power systems (PMAPS), Durham, UK, 7–10 July 2014, pp 1–6
- [12] Lam AYS, Leung KC, Li VOK (2016) Capacity estimation for vehicle-to-grid frequency regulation services with smart charging mechanism. *IEEE Trans Smart Grid* 7(1):156–166
- [13] Guoqing W, Youbing Z, Jun Q et al (2014) Evaluation for V2G available capacity of battery groups of electric vehicles as energy storage elements in microgrid. *Trans China Electrotech Soc* 29(8):36–45
- [14] Leugoue E, Zhang J, Ndjansse SRD (2018) The research of V2G technology real-time charge and discharge capacity prediction. In: Proceedings of IOP conference series: earth and environmental science, vol 146. p 012066
- [15] Mao M, Yue Y, Chang L (2016) Multi-time scale forecast for schedulable capacity of electric vehicle fleets using big data analysis. In: Proceedings of 13th international symposium on power electronics for distributed generation systems, Vancouver, Canada, 27–30 June 2016, pp 1–7
- [16] Kumar KN, Sivaneasan B, Cheah PH et al (2014) V2G capacity estimation using dynamic EV scheduling. *IEEE Trans Smart Grid* 5(2):1051–1060
- [17] Zhang H, Hu Z, Xu Z et al (2017) Evaluation of achievable vehicle-to-grid capacity using aggregate PEV model. *IEEE Trans Power Syst* 32(1):784–794
- [18] Guo Y, Yang Z, Feng S et al (2018) Complex power system status monitoring and evaluation using big data platform and machine learning algorithms: a review and a case study. *Complexity*: 1–21
- [19] Wang D, Sun Z (2015) Big data analysis and parallel load forecasting of electric power user side. *Proc CSEE* 35(3):527–537
- [20] Xu Y, Cheng Q, Li Y et al (2017) Mid-long term load forecasting of power system based on big data clustering. *Proc CSU-EPSA* 29(8):43–48
- [21] Talavera-Llames RL, Pérez-Chacón R, Martínez-Ballesteros M et al (2016) A nearest neighbours-based algorithm for big time series data forecasting. In: Proceedings of international conference on hybrid artificial intelligence systems, Seville, Spain, 18–20 April 2016, pp 174–185
- [22] Dong X, Qian L, Huang L (2017) Short-term load forecasting in smart grid: a combined CNN and K-means clustering approach. In: Proceedings of 2017 IEEE international conference on big data and smart computing (BigComp), Jeju, Korea, 13–16 February 2017, pp 119–125
- [23] Prada J, Dorronsoro JR (2018) General noise support vector regression with non-constant uncertainty intervals for solar radiation prediction. *J Mod Power Syst Clean Energy* 6(2):244–254
- [24] Gan D, Wang Y, Yang S et al (2018) Embedding based quantile regression neural network for probabilistic load forecasting. *J Mod Power Syst Clean Energy* 6(2):244–254
- [25] Chen P, Li W, Chen Y et al (2017) A parallel evolutionary extreme learning machine scheme for electrical load prediction. In: Proceedings of IEEE computing conference, London, UK, 18–20 July 2017, pp 332–339
- [26] Zhao H, Tang Z, Shi W et al (2017) Study of short-term load forecasting in big data environment. In: Proceedings of 2017 29th Chinese control and decision conference (CCDC), Chongqing, China, 28–30 May 2017, pp 6673–6678
- [27] Su X, Liu T, Cao H et al (2017) A multiple distributed bp neural networks approach for short-term load forecasting based on hadoop framework. *Proc CSEE* 37(17):4966–4973
- [28] Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Ann Stat* 29(5):1189–1232
- [29] Guo FL, Zhou G (2017) Analysis of influencing factors on forecast accuracy of ensemble learning. In: Proceedings of 2017 10th international symposium on computational intelligence and design (ISCID), Hangzhou, China, 9–10 December 2017, pp 37–42
- [30] Wang WS, Ding J, Zhao YL et al (2003) Study on the long term prediction of annual electricity consumption using partial least square regressive model. *Proc CSEE* 23(10):17–21
- [31] Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. *J Anim Ecol* 77(4):802–813
- [32] Mao M, Wang Y, Yue Y et al (2017) Multi-time scale forecast for schedulable capacity of EVs based on big data and machine learning. In: Proceedings of 2017 IEEE energy conversion congress and exposition (ECCE), Cincinnati, USA, 1–5 October 2017, pp 1425–1431
- [33] Mayhorn E, Xie L, Butler-Purpy K (2017) Multi-time scale coordination of distributed energy resources in isolated power systems. *IEEE Trans Smart Grid* 8(2):998–1005
- [34] Chunyan L, Xiao C, Peng Z et al (2018) Multi-time-scale demand response dispatch considering wind power forecast error. *Power Syst Technol* 42(2):487–495
- [35] Liu H, Qi J, Wang J et al (2016) EV dispatch control for supplementary frequency regulation considering the expectation of EV owners. *IEEE Trans Smart Grid* 9(4):3763–3772
- [36] Mao T, Lau WH, Chong S et al (2016) A new schedule-controlled strategy for charging large number of EVs with load shifting and voltage regulation. In: Proceedings of 2015 IEEE PES Asia-Pacific power and energy engineering conference (APPEEC), Brisbane, Australia, 15–18 November 2015, pp 1–5
- [37] Karfopoulos EL, Panourgias KA, Hatzigiargyriou ND (2015) Distributed coordination of electric vehicles providing V2G regulation services. *IEEE Trans Power Syst* 31(4):1–13
- [38] State Grid (2010) Enterprise standards of state grid corporation: technical provisions for wind farms connecting to power grids, Beijing
- [39] Breiman L, Friedman JH, Olshen R et al (1984) Classification and regression trees. Chapman & Hall, New York
- [40] Fanibhare V, Dahake V (2016) SmartGrids: MapReduce framework using Hadoop. In: Proceedings of 2016 3rd international conference on signal processing and integrated networks (SPIN), Guangzhou, China, 12–13 December 2016, pp 400–405
- [41] Xu G, Shen C, Liu M et al (2017) A user behavior prediction model based on parallel neural network and k-nearest neighbor algorithms. *Clust Comput* 20(2):1703–1715



Meiqin MAO received the B.Sc., M.Sc. and Ph.D. degrees in electrical engineering from Hefei University of technology in 1983, 1988 and 2004, respectively. She is now a Professor with School of Electrical and Automation Engineering, Hefei University of Technology, Hefei, China. Her research interests include renewable energy generation technology, distributed power generation and microgrids, power electronics applied in power system. She serves as an Associate Editor for IEEE Journal of Emerging and Selected Topics in Power Electronics.

Shengliang ZHANG received the B.Sc. degree in electrical engineering and its automation from Northeast Electric Power University, Jilin, China, in 2017. He is currently pursuing the M.Sc. degree in power system and its automation at Hefei University of Technology, Hefei, China. His research interests include EV integration into smart grid and energy management system in microgrids.

Liuchen CHANG received the B.Sc. degree from Northern Jiaotong University, Beijing, China, in 1982, the M.Sc. degree from the China Academy of Railway Sciences, Beijing, China, in 1984, and the Ph.D. degree from Queen's University, Kingston, Canada, in 1991. Currently, he is Professor of electrical and computer engineering at

the University of New Brunswick, Fredericton, Canada. He is a fellow of Canadian Academy of Engineering. His principal research interests and experience include distributed power generation, renewable energy, and power converters.

Nikos D. HATZIARGYRIOU is professor at the Power Division of the Electrical and Computer Engineering Department of the National Technical University of Athens, Greece. He is currently the Chairman of the Hellenic Distribution Network Operator (HEDNO), the Distribution Utility of Greece. From February 2007 to September 2012, he was Deputy CEO of the Public Power Corporation (PPC) of Greece, responsible for Transmission and Distribution Networks, island DNO and the Center of Testing, Research and Prototyping. He is Fellow Member of IEEE, past Chair of the Power System Dynamic Performance Committee, Distinguished member of CIGRE and past Chair of CIGRE SC C6 "Distribution Systems and Distributed Generation". He is the Editor in Chief of the IEEE Trans on Power Systems. He is the chair of the Advisory Council of the EU Technology Platform on SmartGrids. He is a member of the Energy Committee of the Athens Academy of Science. His research interests include smartgrids, microgrids, distributed and renewable energy sources and power system security.