energies

MDPI

Article

The Application of Improved Random Forest Algorithm on the Prediction of Electric Vehicle Charging Load

Yiqi Lu [1], Yongpan Li [2], Da Xie [1,*], Enwei Wei [3], Xianlu Bao [3], Huafeng Chen [2] and Xiancheng Zhong [3]

1 Shanghai Jiao Tong University, Minhang District, Shanghai 200240, China; luyiqi@sjtu.edu.cn
2 Shenzhen Power Supply Co. Ltd., Luohu District, Shenzhen 518001, China; yongpan_li@163.com (Y.L.); csgchf@126.com (H.C.)
3 Shenzhen Comtop Information Technology Co. Ltd., Shenzhen 518034, China; weienwei@szcomtop.com (E.W.); baoxianlu@szcomtop.com (X.B.); zhongxiancheng@szcomtop.com (X.Z.)
* Correspondence: profxzg@hotmail.com; Tel.: +86-21-3420-4298

Received: 9 October 2018; Accepted: 13 November 2018; Published: 19 November 2018

Abstract: To cope with the increasing charging demand of electric vehicle (EV), this paper presents a forecasting method of EV charging load based on random forest algorithm (RF) and the load data of a single charging station. This method is completed by the classification and regression tree (CART) algorithm to realize short-term forecast for the station. At the same time, the prediction algorithm of the daily charging capacity of charging stations with different scales and locations is proposed. By combining the regression and classification algorithms, the effective learning of a large amount of historical charging data is completed. The characteristic data is divided from different aspects, realizing the establishment of RF and the effective prediction of fluctuate charging load. By analyzing the data of each charging station in Shenzhen from the aspect of time and space, the algorithm is put into practice. The application form of current data in the algorithm is determined, and the accuracy of the prediction algorithm is verified to be reliable and practical. It can provide a reference for both power suppliers and users through the prediction of charging load.

Keywords: electric vehicle (EV); random forest; charging load; data analysis; load forecasting

1. Introduction

To solve the problems such as load balancing, capacity planning, and power quality caused by the access of large-scale electric vehicle (EV) [1], researchers have proposed many practical coordinated control schemes to guarantee the safety and reliability of the power system. For example, after the analyzing of the load demand in certain area, the EV chargers can be used to balance the unbalanced network without overloading the charger [2]. It has been proved that the load of EV can be converted into a tool to benefit the power system [3] by applying optimal charging schemes to arrange the charging and discharging through certain approaches such as demand side response [4,5]. However, these methods depend on the load prediction to a certain extent.

The variety of load forecasting is a large field of the researches on power system. As a new form of load, the exploration of EV load has already begun. At the beginning of these researches, the forecast of the EV load mainly based on the behavior of EV users or specific areas. By using the Markov Chain and Monte Carlo simulation, the seasonal and holiday characteristic of EV users can be analyzed, but the load prediction still has certain uncertainty [6]. Other studies focus on the space and time distribution of the EV load. Models can be established to simulate the fluctuation of EV load [7], and thus the load can be forecasted and thus it can be used as a feasible load to reduce the pressure of

the system [8]. The prediction can also be accomplished through the analyzing of specific location [9], or the spatial and temporal distribution [10,11]. The researches mainly emphasize on the modeling of EV or EV users [12]. Although such a method can perform part of the prediction function, it rarely involves the real-time fluctuation of a specific charging station, and it cannot be determined whether it can be applied to a charging system that has been changed after EV is widely popularized.

With the gradual popularization of EV, a lot of charging data has been accumulated. Meanwhile, the technique of machine learning is developing rapidly these years. In fact, a variety of load forecasting methods have been put into practice like wind power or solar energy [13]. Thus, associating the EV load prediction with machine learning is a practical way to realize EV load prediction. As the two most investigated field, support vector machine (SVM) is an effective method to forecast the daily load of relatively stable charging stations, and long short-term memory (LSTM) Neural Network is also an approach for time series forecast [14] as well as other methods of machine learning based on big data [15], however, the selection of features and data is still an unsolved problem [16]. The randomness of EV load is a large obstacle for most time series prediction algorithm such as LSTM, for the charging capacity for one day is not necessarily related to the capacity the day before. In fact, SVM can be used to forecast EV load. In recent researches, the prediction of electric bus charging stations is realized through SVM by the sampling of similar days [17]. The drawback of SVM is the lack of universality and data limitation. In addition, existing prediction method can hardly be used in location-based prediction method.

In contrast, the random forest algorithm (RF) used in this paper has been applied to power load forecasting from user-side, and the parallelized data processing mode implemented by random forest has the characteristic of high efficiency [18]. Actually, RF has already been applied in load prediction of the power system with higher precision and stability than ordinary SVM algorithm [19]. Bernoulli RF and quantile regression forest have also been applied in load forecasting and both receive high accuracy [20,21]. The RF is an integrated learning method of decision tree which has been proved to be an effective method to complete the prediction work [22].

The innovations of this paper are as follows: (1) all the characteristic parameters needed to improve the accuracy of the prediction algorithm are proposed; (2) the RF is applied in specific modality to realize the short-term prediction of the charging load; and (3) the load prediction method of charging station group is proposed by combining the advantages of classification tree and regression tree.

The structure of this paper is as follows. Firstly, the process and theoretical basis of RF are introduced. Then the application of RF in charging load forecasting is proposed according to the actual needs of charging load prediction. The design includes a single charging station and charging station group. The feature data selection method and a prediction algorithm for the charging station group are proposed in combination with the charging data. Then it analyzes the development status of EV in Shenzhen and shows the actual charging data of Shenzhen. Finally, through the analysis of the existing charging load data in Shenzhen. The important characteristic parameters are judged, and then the actual simulation is carried out in both the load prediction algorithm of single charging station and the charging station group to verify the effectiveness of the algorithm.

2. Random Forest Algorithm

Random forest is an integrated learning method that integrates multiple decision trees to eliminate the correlation between feature data. At the same time, the computational complexity of RF is only $O(n)$ (n stands for the number of samples) when the amount of data is large, furthermore, the algorithm can be run in parallel because of this integration to speed up.

RF reduce the correlation between decision trees by randomly selecting samples and features. Firstly, the same amount of data is selected randomly from the training sample in the original training data. Secondly, a part of the features is randomly selected to establish the decision tree. These two kinds of randomization make the correlation between each decision tree small, which reduces the error that may occur when the decision tree itself is over-fitting, and improves the accuracy of the model.

2.1. Gini Coefficient

During the generation of decision trees, the measure of the amount of information is defined as: the more the "uncertainty" of the data is reduced, the more information the partition can obtain. There are two common indicators for measuring this uncertainty: information entropy and *Gini* index.

Take K random variables, then the definition of the *Gini* coefficient is:

$$Gini(y) = 1 - \sum_{k=1}^{K} p_k^2 \tag{1}$$

where p_k indicates the different probabilities of taking Kth variable. It can be proved that when the Equation (2) is satisfied, the maximum $Gini(y)$ is obtained, and if $p_i = 1$ and $p_j = 0, (i \neq j)$, then $Gini(y) = 0$. This shows that more irregular y (y is the variable being discussed) is, the larger $Gini(y)$ is. Thus, the *Gini* coefficient can be used to measure uncertainty.

$$p_1 = p_2 = \cdots = p_K = \frac{1}{K} \tag{2}$$

2.2. Decision Tree

The algorithm used in this paper is the CART algorithm, which is the classification and regression tree (CART) using *Gini* gain in Equation (6) or least square as division criteria, for CART is more sophisticated, and can be used to solve both classification and regression problems [23].

2.2.1. Classification Decision Tree

- Load data set D on a node;
- If all the samples in D belong to the category c_k, the node will not continue to generate and mark it as c_k;
- If there is no optional feature, the category with the largest number of samples in D is taken as the category of the node;
- Otherwise, if feature $x^{(j)}$ has S_j different values $u_1^{(j)}, \cdots, u_{S_j}^{(j)}$ which satisfy $u_1^{(j)} < \cdots < u_{S_j}^{(j)}$ in the current data set, then:

  (i) If $x^{(j)}$ is discrete, $u_1^{(j)}, \cdots, u_{S_j}^{(j)}$ are selected as separation points a_p uccessively, then:

$$A_{jp} = \left\{ x^{(j)} = a_p, x^{(j)} \neq a_p \right\} \tag{3}$$

  (ii) If $x^{(j)}$ is continuous, $\frac{u_1^{(j)} + u_2^{(j)}}{2}, \cdots, \frac{u_{S_j-1}^{(j)} + u_{S_j}^{(j)}}{2}$ are selected as separation points a_p successively, then:

$$A_{jp} = \left\{ x^{(j)} < a_p, x^{(j)} \geq a_p \right\} \tag{4}$$

A_{jp} is the result of the division of $x^{(j)}$. According to the *Gini* gain, the feature $x^{(j^*)}$ with the greatest information gain of the jth feature. The corresponding dichotomy are calculated as the division criteria:

$$(j^*, p^*) = \underset{j,p}{\operatorname{argmax}}\, g_{Gini}(y, A_{jp}) \tag{5}$$

- If the stop condition is satisfied, take the category with the largest number of samples in D at this time as the output category;
- Otherwise, according to all possible values of $x^{(j^*)}$ (which is $\{a_1, \ldots, a_m\}$), divide D into $\{D_1, \ldots, D_m\}$:

$$(x_i, y_i) \in D_j, \left(x_i^{(j^*)} = a_j \right), \forall i = 1, \cdots, N \tag{6}$$

- Call the algorithm from Equation (1) for each D_j.

By looping through the above seven steps, a decision tree that meets the specific goal is generated.

2.2.2. Regression Decision Tree

The difference between the generation of regression tree and classification tree is the node partitioning criteria of nodes and the selection of output. The division criterion is the least squares method. For $x^{(j)}$, scan all its possible values, and select the separation point a_p, then $x^{(j)}$ will be divided into two parts R_1 and R_2. Find the value c_1 and c_2 in the output y, respectively, until the minimum value of Equation (11) is obtained. Then this a_p is the best separation point of $x^{(j)}$.

$$\min_{a_p}\left[\min_{c_1}\sum_{x_i\in R_1}(y_i-c_1)^2+\min_{c_1}\sum_{x_i\in R_2}(y_i-c_1)^2\right] \tag{7}$$

Similarly, the optimal partitioning features $x^{(j^*)}$ can be got by traversing j and the corresponding nodes.

The output value is determined by the average value of the corresponding range. Take R_1 as an example, the output value is:

$$c_o=\frac{1}{N_1}\sum_{x_i\in R_1}y_i \tag{8}$$

Among them, N_1 is the number of samples in R_1.

After the decision tree is generated, input the sample feature values that need to be processed, the corresponding output will be obtained.

3. Design of Random Forest Algorithm Application

Intuitively, RF can be thought of as generating a decision tree for each random sample from data set of the original data, and integrating the results of many decision tree outputs according to voting or averaging strategies as the final output.

This method of random sampling and the integrated output of the results is called Bagging. The specific algorithm process is as follows:

- Using Bootstrap, randomly extract n training samples from the original data set;
- k rounds of extraction are performed and k training sets are obtained;
- training k decision tree models for k training sets;
- For the problem of charging load prediction: the average of the prediction results of each model is used as the final prediction result.

RF can be intuitively understood through Figure 1. It is noted that the daily charging amount of different charging stations has a discrete characteristic, that is, the charging amount is much dispersed. Thus, step division of the original data is considered. According to the value range of the specific charging amount data, determine the intervals to cover the range. Then the small interference will be eliminated, and the effectiveness and accuracy of RF prediction algorithm will be improved. There are two main principles for the division of intervals:

(1) The amount of data in each interval is the same, which can ensure that each step after the division occupies the same proportion of the historical data. This principle is suitable for the small charge portion in the single station prediction;
(2) The length of the interval is the same. More intervals will be generated using this method, which requires a large amount of data. This principle is suitable for daily charging capacity prediction of station groups which is more uneven in data distribution.

After the bagging of pre-processed data samples, they are divided into k data packets. For each data packet, the regression decision tree is constructed separately: start from the starting node (root

node), the regression type is targeted to minimize the *Gini* coefficient (the uncertainty) through the CART algorithm, continue separating until the target or the maximum depth is reached.

The nodes that no longer bifurcate are called leaf nodes, and each leaf node is assigned an output value. This value is set differently from the classification decision tree algorithm. The average value of the corresponding value before preprocessing of this leaf node is the output. Applying the division process to each data packet, the learning process of the random forest model is realized.

When making predictions, the predicted data features will be input into the model. Each decision tree will generate independent prediction results, and the entire random forest will use the average of the results of all the decision trees as the final prediction result.

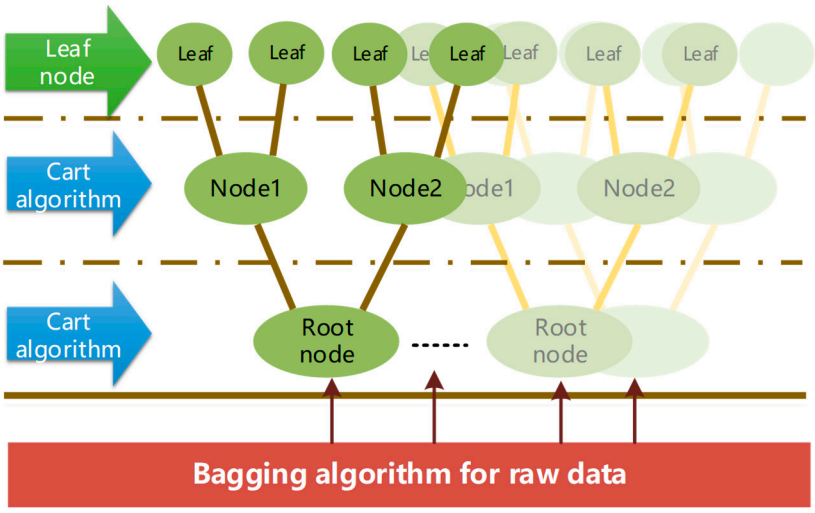

Figure 1. Schematic diagram of random forest algorithm (RF).

3.1. Charging Load Prediction Algorithm of a Single Station

For the charging load prediction of a single charging station. To meet the actual demand of the forecasting, the load is predicted by using RF regression tree in Section 2.2.2. The characteristic properties of the corresponding model are designed. The specific input and output data information is shown in Table 1. The characteristic attributes include the following categories:

- Date indicator (Year, Month and Day): an accurate judgment of the influence of climatic conditions such as temperature and humidity on the behavior of EV is difficult to make. Therefore, the attributes are directly integrated into the date indicator, and the impact of climate can be minimized with large amount of data;
- quantity: the importance is represented by a numerical value, which will be limited in 15 min;
- Activity indicator: the importance can be expressed by numerical values, which will be limited in 15 min. Important activities may cause a surge in regional charging load;
- Prosperity index: The infrastructure index in the prosperous index, which will fluctuate with the renovation of buildings and roads. This is an important indicator that affects the charging habits of EV users;
- Charging capacity: before the current time, the amount of power that has been given. The charging area and the charging capacity of many EV users in a period are relatively fixed, so the accumulated charging load of the daily charging station should also be recorded. The volume will have an impact on the remaining load prediction for that day.
- Previous day's charge: like the amount of charge, it can increase the temporality of the RF.

Table 1. Input and output table for RF.

Type	Data name	Description	Symbol
Common feature	Year	Indicates the year	X9
	Month	Indicates the month	x7
	Day	Specific date	x4
	Activity indicator	0–5 indicates the importance of activities	x2
	Prosperity index	Show the quality of area facilities	x14
Unique characteristics of single charging station	Weekend symbol	1/0 indicates whether it is a weekend	x13
	Holiday symbol	1/0 indicates whether it is a holiday	x12
	15-min quantity	Time passed that day	x11
	Charged amount	The amount of charge that day	X10
Unique characteristics of charging station group	Week symbol	From Monday to Sunday	x3
	Capacity indicator	Rated charging power of the station	x8
	Longitude	longitude of the station	x6
	Latitude	Latitude of the station	x5
	Previous day's charge	The amount of charge the day before	x1
Output	Charging capacity	Predict charge capacity every 15 min	y1
	Charging times	Predict charge times every 15 min	y2

3.2. Charging Load Prediction Algorithm of Station Group

Unlike normal loads, the charging load of EV tends to have group nature. In fact, predictions based on historical charging data from a single charging station is the way most current prediction algorithms use, and its accuracy can indeed meet the needs. Base on this, the charging load prediction algorithm of charging station group contains many stations by using the stepped daily charging capacity. The separation criterion of the classification tree is combined with the output selection of the regression tree. The input characteristics are shown in Table 1. Compared with the single station algorithm, the station group algorithm adds the following input characteristic data:

(1) Week symbol: indicates the position of the day in a week, the data contains the information of the weekend, and can reveal the characteristic attributes of different dates;
(2) Capacity indicator: indicates the rated capacity of each charging station. This value is obtained by summing the rated power of charging piles at each charging station. The capacity index reflects the prosperity of the location of the charging station to some extent.
(3) Longitude: the longitude of the location of each charging station;
(4) Latitude: the latitude of the location of each charging station is used to uniquely determine each charging station. The longitude and latitude indicators can effectively quantify the regional characteristics of different charging stations.

By integrating the 12 input characteristics belonging to the charging station group in Table 1, the charging load prediction for station group can be realized by RF. Since the charging station group considers the charging load variation characteristics of many charging stations of different sizes and regions, it is possible to simulate the short-term load changes of the respective charging stations.

The flow chart of the entire prediction process is shown in Figure 2.

After the original data is processed, k sample sets are obtained by bagging algorithm, and k decision trees are generated by the CART algorithm in Section 2.2 to form a random forest. Then, the forest can input and predict the charging load through the input within the predicted period.

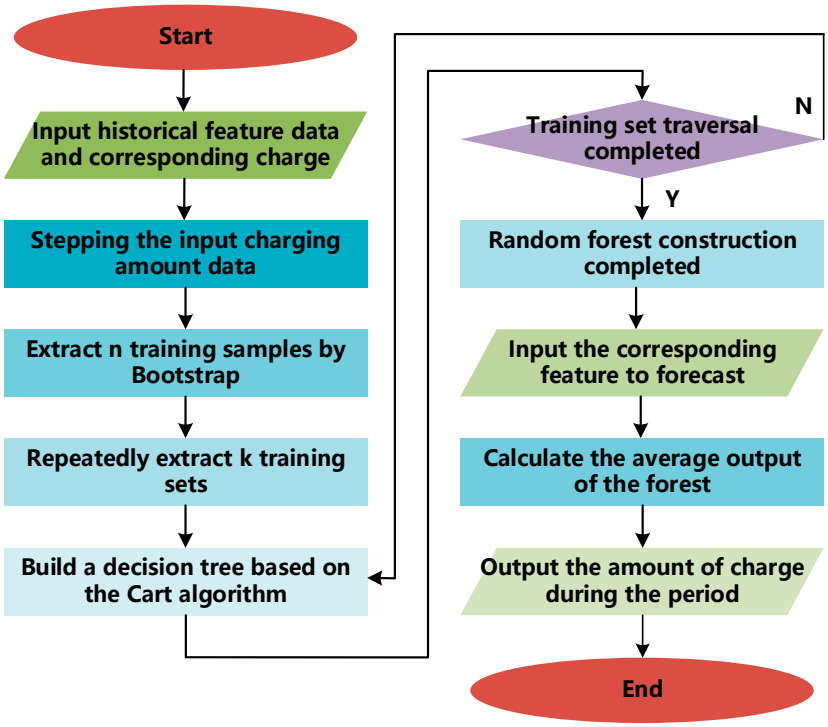

Figure 2. Flowchart of charging prediction algorithm.

3.3. Evaluation of Prediction Results

3.3.1. Error Analysis

For the results, the mean absolute percentage error (*MAPE*) and the root mean square error (*RMSE*) are used for evaluation. The error calculation formulas are shown in Equations (9) and (10), respectively.

$$\varepsilon_{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\hat{P}_N(i) - P_N(i)\right|}{P_N(i)} \times 100\% \tag{9}$$

$$\varepsilon_{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{P}_N(i) - P_N(i)\right)^2} \tag{10}$$

where $P_N(i)$ and $\hat{P}_N(i)$ (i = 1, 2, 3, ..., n) are the actual values and predicted values of the ith data point, respectively, and n represents the length of the data used for verification. ε_{MAPE} is regarded as the main judgment of error.

3.3.2. Analysis of Feature Importance

The importance of the input features is evaluated to verify the actual validity of the inputs. For each regression decision tree, the importance of a feature at a node refers to the variable of the *Gini* coefficient before and after the branch of the node, and its definition can be expressed as Equation (11).

$$W_{m,j} = Gini_m(y) - Gini_n(y) - Gini_p(y) \tag{11}$$

where n and p represent the two child nodes generated by node m, respectively. The characteristic importance of any decision tree i can be obtained by summing:

$$W_{i,j} = \sum_{m \in i} W_{m,j} \tag{12}$$

4. Case Analysis

In this section, the data of many charging stations in Shenzhen from 2016 to 2018 is analyzed, and charging load prediction for single station and station group is realized. The current situation and the effect of the application of RF are analyzed.

4.1. Analysis of the Construction of Charging Facilities and Charging Data in Shenzhen

Shenzhen City has jurisdiction over 10 districts including Luohu District, Futian District and Longgang District. The area of each district and the distribution of charging stations are shown in Figure 3a. Charging stations are most densely distributed in Nanshan District, Futian District and Luohu District. Baoan District and Longgang District are two districts with the largest number of charging stations. Nanshan, Futian, Longgang and Baoan are the most developed areas in Shenzhen. It is obvious that the distribution of charging stations is related to the economic strength of each district. Today, the total number of EV in the city has exceeded 80,000. According to the "2017 New Energy Vehicle Promotion and Application Financial Support Policy of Shenzhen", the government is now emphasizing on the construction of EV supporting facilities. This also indicates that the analysis and regulation research work of this new load of EV has entered the government's plan.

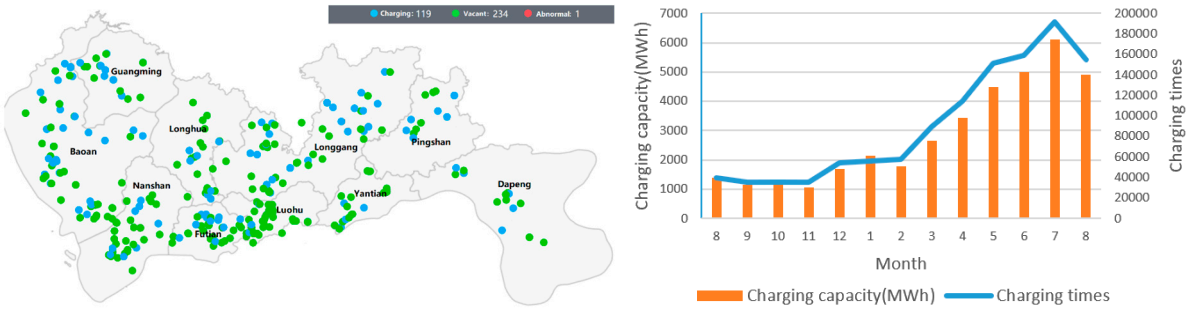

Figure 3. Charging facilities and data in Shenzhen: (**a**) the distribution of charging stations; and (**b**) charging history in Shengzhen.

At present, about 6000 charging piles have been built in Shenzhen. The charging history of Shenzhen is shown in Figure 3b, and the data of August is not complete. It is obvious that the recent increase in the charge capacity reflects the growing popularity of EV. This trend also increases the importance and urgency for the government's charging policy and related research.

The data used in the simulation are the information of charging station in Shenzhen and the data of charging capacity of different periods during two years. To fully display the spatial and temporal distribution characteristics of the selected data, the data is analyzed from temporal distribution and spatial distribution.

4.1.1. Temporal Analysis of Charging Data

In terms of time, the charging data includes the data from the second half of 2016 to the first half of 2018, which is enough to complete the prediction algorithm. From the monthly distribution, the distribution of the charging data in each month is shown in the violin diagram as shown in Figure 4a.

The monthly charging data is represented as a violin chart according to the distribution of charging capacity. The monthly distribution of the data is similar, consisting of a relatively centralized smaller charging capacity (wider part) in the lower part of the graph and a relatively larger charging capacity (slender part) in the upper part of the graph, so the monthly data form a needle. In the violin diagram, the area of the monthly figure is equal. Relatively speaking, the tip of the charge distribution is thicker and the bottom is narrower in summer, which indicates that the data distribution is narrower and the average charge is higher, while the average load in autumn and winter is smaller.

From the daily point of view, the charging data can also be plotted as shown in Figure 4b, and the daily charge distribution is very similar in shape to Figure 4a. Since the monthly date itself is not of practical significance, it needs to be matched with the month and week symbol to have the ability to express the meaning of time, so it is necessary to append the weekend symbol to regularize the daily charge load changes.

In fact, Figure 4a shows only the scattered charging station charging data with a daily charge capacity of less than 100 kWh, for the rest of the data distribution is like this situation. Using small charging data can make the graph clearer. Figure 4b shows almost all the data, showing the trend of peak charging capacity, and the specific trends need to be judged by the prediction algorithm.

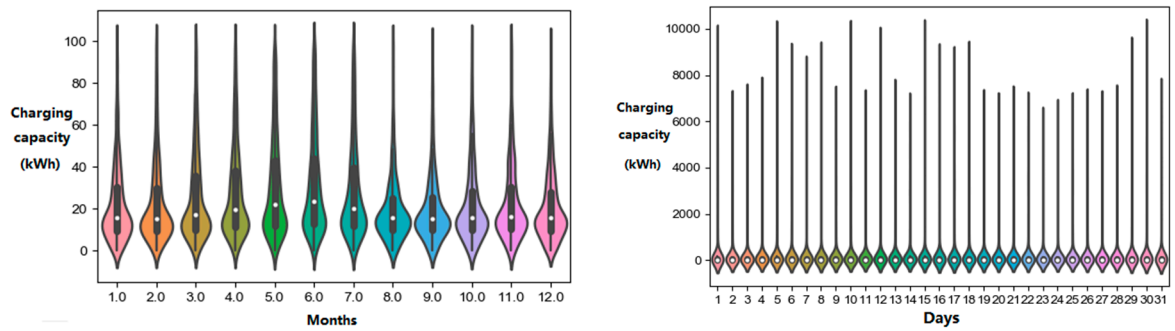

Figure 4. Violin chart of charging data distribution by month/day: (**a**) distribution by months; and (**b**) distribution by days.

4.1.2. Spatial Analysis of Charging Data

From the perspective of space, the latitude and longitude coordinates divide the geographic location of different charging stations. The relationship between the distribution of charging data and latitude and longitude coordinates is shown in Figure 5a. Since the distribution of the charging stations is discrete, Figure 5a is composed of a plurality of peaks on a plane. The horizontal coordinates respectively indicate the latitude and longitude of different stations, the ordinate and the color indicate the accumulation of the charging capacity of each station.

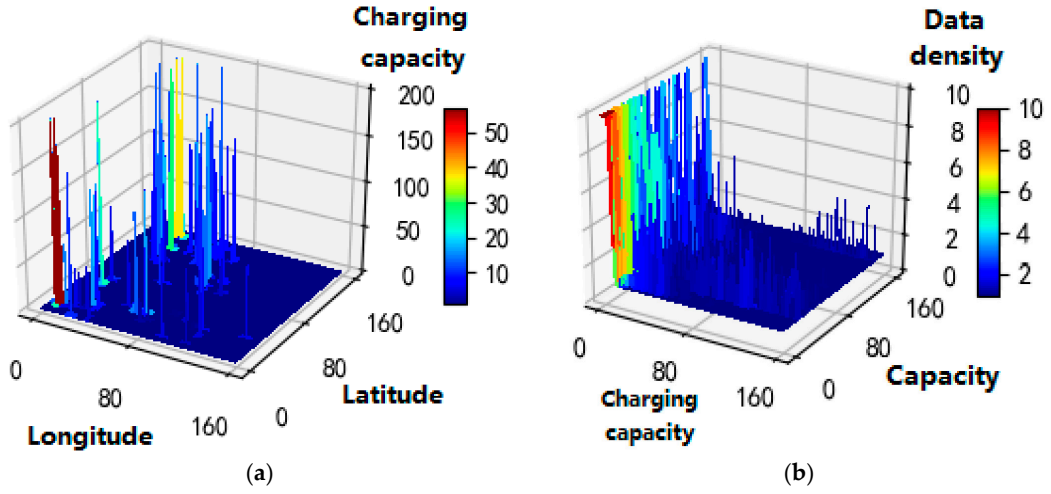

Figure 5. Relationship between charging data distribution and charging station character: (**a**) capacity and location; and (**b**) data density and stations.

It can be seen from Figure 5a that the charging data is clearly divided into certain concentrated areas, mainly two red areas. In fact, the total charging capacity in each area for a period will be

relatively stable. For example, the maximum deviation of the charging capacity between two months of the dark red part in the figure is only about 20%. This is an important foothold for the effectiveness of the charging station group prediction algorithm.

Figure 5b shows the relationship between the data distribution and the capacity of the charging stations. The horizontal coordinates are the capacity of each station and the converted value of charging capacity. The ordinate and color indicate the data density of the converted value and the corresponding station. Obviously, the data is generally concentrated near small charging capacity, especially the small-capacity station, they have a large amount of charging data (red part). This is because the main part of the EV charging is still small-capacity stations with small charging capacity. To facilitate direct observation, the data coordinates in the heat map have been quantified. Although the charging data is too small (less than 500 kWh), the large-capacity station still has a large charging capacity (as shown on the right side in Figure 5b, and charging changes with capacity. The distribution of the data also varies significantly, so it is also necessary to use capacity as an input feature.

4.2. Charging Load Prediction

4.2.1. Prediction of Single Station

To verify the effectiveness of the prediction algorithm mentioned in Section 3.1, the load data of a 524 kW charging station in Nanshan District, Shenzhen City was selected as a numerical example for simulation verification. The characteristic attributes of the training samples are selected as the year, month, day, 15-min quantity, weekend symbol, holiday symbol, activity indicator, and charged amount in Table 1.

(1) Training

Firstly, the daily charging capacity and the number of charging times are taken as the output, and the accuracy of RF is observed. 90% of the sample data is used as the training sample set, and the remaining 10% is used as the test sample set for the RF model. Select the sample characteristics of the year, month, day, weekend, holiday symbol, and activity indicator in Table 1. At the same time, specify 120 trees in the random forest. The depth of each tree is controlled within 80, and the average value is used as the output to obtain the load of the charging station. The test data is shown in Figure 6a. For the sake of brief observation, only some test data is shown in the figure.

The blue curve in the figure represents the prediction of RF, the green curve is the prediction of support vector regression (SVR) [24], and the orange curve is the actual value. SVR is selected to examine the actual effect of the RF prediction with the ε_{MAPE} of 9.03%, and the ε_{RMSE} of 457.21. Compared with ε_{MAPE} of 9.82%, and the ε_{RMSE} of 417.23 of the prediction of SVR.

As shown in Figure 6b, the prediction of charging times shown. The simulation shows that the prediction effect is also accurate, with ε_{MAPE} of 9.67%, and ε_{RMSE} of 16.46. Compared with ε_{MAPE} of 11.37%, and the ε_{RMSE} of 21.51 of the prediction of SVR.

Change the daily load data to the charging capacity every 15 min, add the 15-min sample feature in Table 1, and change the output to charging capacity every 15 min. The RF and SVR model thus trained by the data. The prediction of the test sample is shown in Figure 6c. RF Prediction result: ε_{MAPE}: 10.27%, ε_{RMSE}: 5.02. Compared with ε_{MAPE} of 11.53%, and the ε_{RMSE} of 7.82 of the prediction of SVR.

From the perspective of the prediction in the training set, RF model can achieve an average absolute error within 10% of the single charging station charging prediction. However, SVR model has similar results generally, but not as accurate. Then, the actual effect of the prediction process observation algorithm is simulated.

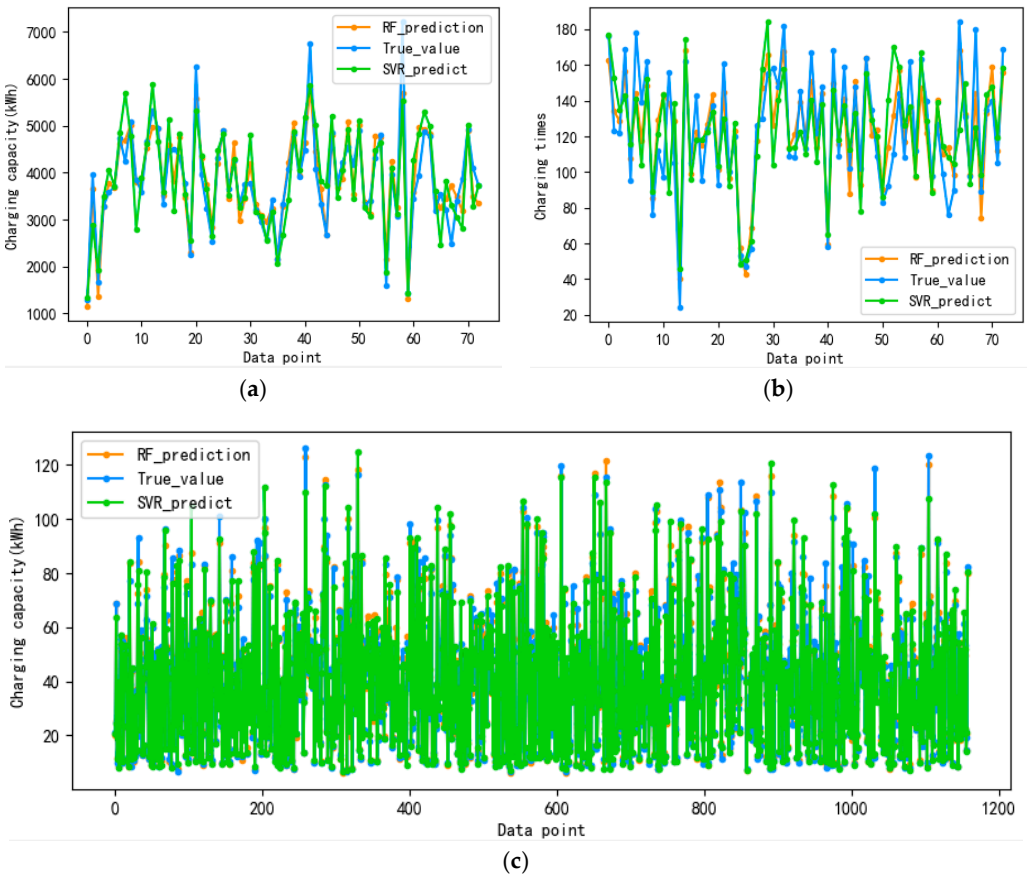

Figure 6. Load forecast curve of single charging station: (**a**) daily charging capacity; (**b**) daily charging times; and (**c**) charging capacity every 15 min.

(2) Prediction

After training the model, it can be used to realize the function of forecasting by the newly collected charging station load data in June 2018. For the charging data from 14 June to 26 June the outputs are shown in Table 2.

Table 2. Prediction data of single charging station every 15 min.

No.	Date	15-min Quantity	True Value	Prediction
1	6.26	33	62.79	64.20
2	6.25	17	42.67	40.22
3	6.25	59	33.14	33.15
4	6.24	10	29.36	27.45
5	6.24	80	23.01	22.82
6	6.22	56	28.79	28.53
7	6.18	34	60.03	57.38
8	6.17	64	25.25	25.94
9	6.14	4	11.36	12.70
10	6.14	71	8.84	14.36
		ε_{MAPE}		9.76%
		ε_{RMSE}		2.27

Since the characteristic data of the charged amount can only be acquired after the previous time elapsed, when using the algorithm for charging prediction, only the charging load of the next 15 min can be predicted in real time. As sown in Table 2, the prediction is very close to the actual value, with ε_{MAPE} of 9.76% and ε_{RMSE} of 2.27. Of course, the predicted value of the charging capacity can be used

as the charged amount portion to continue for the prediction for a longer period, but the accumulated error will gradually become not ignorable.

4.2.2. Charging Load Prediction of Station Group

After verifying the validity of the charging load prediction of the ordinary single charging station, the following is verified for the actual effect of the charging station group charging load prediction.

All small-capacity (less than 150 kW) charging stations that have been working normally in Shenzhen for two years or more are included in the station group, and their historical charging data is used as a sample for simulation. The characteristic attributes of the training samples are the year, month, day, week symbol, capacity mark, longitude, latitude, activity indicator and previous day charge in Table 1.

(1) Training

In the original sample data set, 10% of the charging data is extracted as the test sample, and the remaining 90% is the training sample, which is consistent with the training method of a single charging station.

To further improve the accuracy of the prediction algorithm and avoid the occurrence of over-fitting, the prediction of the test sample trained is observed by changing the structure of the random forest, thereby determining the best structure.

When the number of trees n and the tree depth m are 40 and 40, the performance of the RF algorithm on the test sample is shown in of Figure 7a; when n is 80 and m is 80, the performance is shown in Figure 7b; the performance when n is 120 and m is 120 is shown in Figure 7c; the performance when n is 140 and m is 140 is shown in Figure 7d. Although it can be roughly observed that the effect of prediction is different under different structures, it is difficult to determine the optimal forest structure. Therefore, n and m are traversed at intervals of 10, and the structure with the smallest value of ε_{MAPE} and ε_{RMSE} is extracted from it.

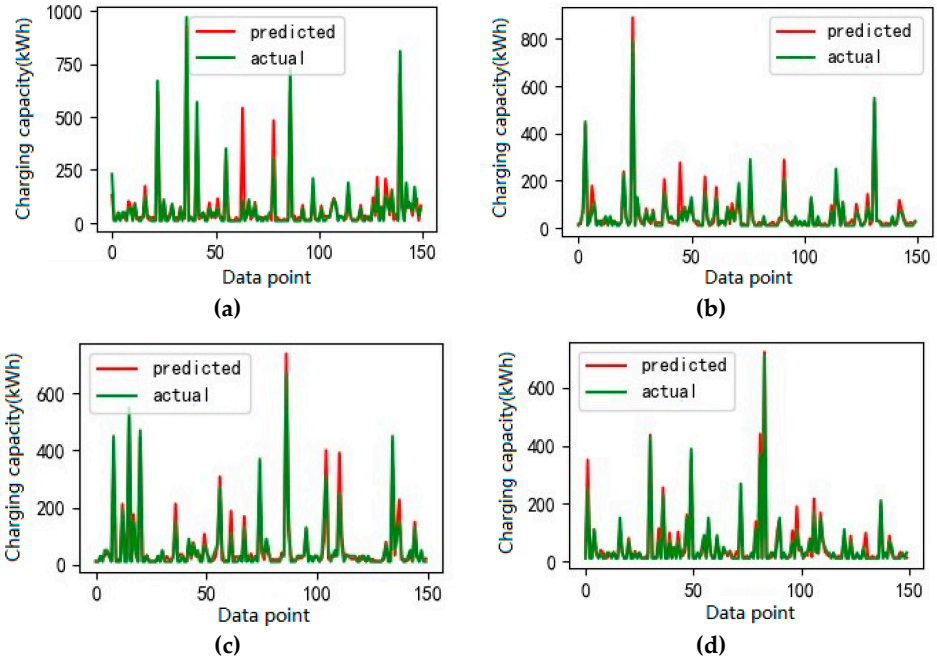

Figure 7. Performance of RF model on testing data set: (**a**) n = 40, m = 40; (**b**) n = 80, m = 80; (**c**) n = 120, m = 120; and (**d**) n = 140, m = 140.

As the structure of the forest changes, the changes of ε_{MAPE} and ε_{RMSE} are shown in Figure 8a,b. As the depth and number of trees increase, both ε_{MAPE} and ε_{RMSE} shows a trend of decreasing in

fluctuations. To avoid over-fitting, the upper limit of n and m is 200, and it is not necessary to continue to increase the upper limit because the effect of the forest will not continue to improve significantly, and the calculation time will be unnecessarily extended. In this range, the minimum ε_{MAPE} occurs when n is 140 and m is 160; the minimum ε_{RMSE} occurs when n is 120 and m is 180; and the minimum product of ε_{MAPE} and ε_{RMSE} occurs when n is 160 and m is 180. These three points are used as the actual application for charging prediction.

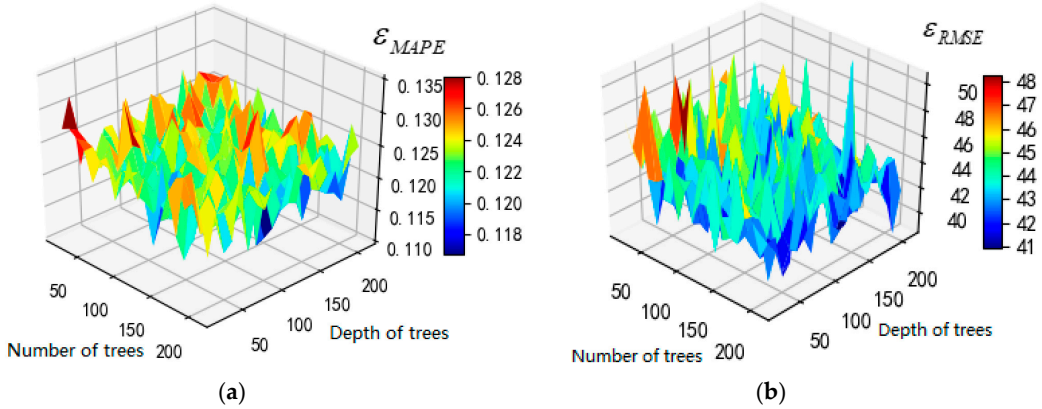

Figure 8. Relationship between forest structure and error by charging station group forecast: (**a**) ε_{MAPE}; and (**b**) ε_{RMSE}.

(2) Importance of input feature

According to the sample data and the forest structure with n of 140 and m of 160 in the previous section, after the forest is generated, the relative importance relationship of all the characteristic attributes in the model can be obtained. The distribution map is shown in Figure 9. The height of each column gives the average value of importance for each tree in the forest, and the black line segment above it represents the standard deviation. The features x1 to x9 represent the previous day's charge, activity indicator, week symbol, day, latitude, longitude, month, capacity indicator and year.

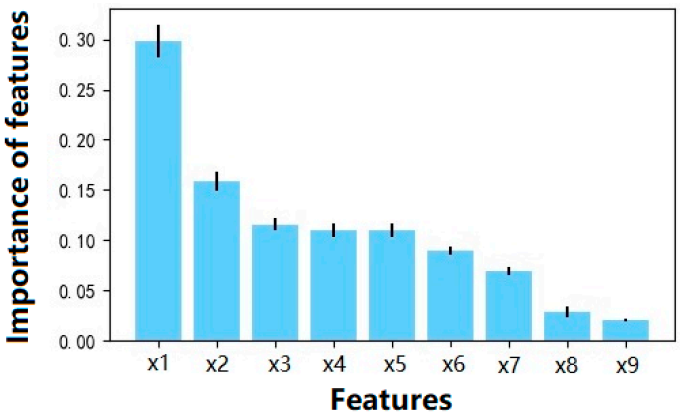

Figure 9. Importance of input characteristic of charging stations.

The most important one of the input characteristics is the previous day's charge. Its importance reaches 0.3, while the sum of all features is 1. This is because the indicator gives the algorithm timing characteristics and contains most of the information about the charging history. The second important factor is the activity indicator, which depicts the volatility of the charging load based on historical data, with an importance around 0.15. Next, come the time indicator, the latitude and longitude, their importance is around 0.1. The less important ones are capacity indicators and year. Indeed, capacity does not determine the direction of charging load at present, and the year does not have a significant

impact on the behavior of EV. However, these two indicators, especially the capacity indicators, also give a certain degree of improvement in the algorithm and the scalability of the data accumulation in the future years.

(3) Prediction

The prediction results obtained by the three forest structures are shown in Figure 10, wherein the predicted value 1 is given by the forest with n of 140 and m of 160. The predicted value 2 is given by a forest with n of 120 and m of 180; the predicted value 3 is given by a forest with n of 160 and m of 180. The red and purple curves are the output of SVR model and C4.5 algorithm (C4.5 uses ordinary information gain as the partition criterion compared with *Gini* entropy for CART). Since the amount of charging stations is relatively large, some typical prediction results are taken as an illustration.

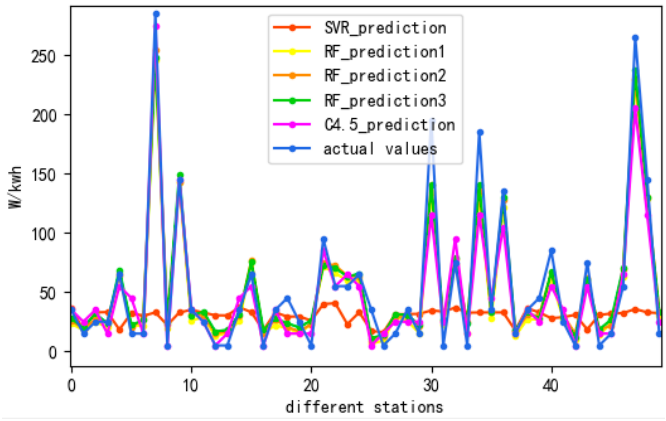

Figure 10. Prediction of charging capacity of station group.

First 10 results are shown in Table 3. It can be inferred that the RF model of station group is more accurate than SVR and C4.5 (decision tree using information gain as split criterion) [25], for it put less emphasis on temporality and output basing on different stations. SVR is not capable of the prediction of station group.

Table 3. 10 Prediction data for charging station group every day.

No.	True Value	RF_Prediction2	SVR_Prediction	C4.5_Prediction
1	36.5	28.8	34.8	36.9
2	19.8	20.1	18.3	25.6
3	26.8	27.2	32.5	34.7
4	27.4	24.8	33.0	15.9
5	66.0	67.5	20.1	54.2
6	11.9	18.3	34.7	15.2
7	18.4	21.2	30.2	14.8
8	288.9	249.9	33.3	276.0
9	149.9	147.2	32.7	143.5
10	31.8	29.3	34.8	25.6
	ε_{MAPE}	12.8%	55.5%	19.5%
	ε_{RMSE}	12.85	90.50	7.98

As shown in Figure 10, as the performance of the test sample, the three prediction curves of RF are highly accurate for single-day load prediction in the face of the actual prediction environment. Among them, the closest to the actual value is the predicted value of 2, with ε_{MAPE} of 10.83% and ε_{RMSE} of 39.59. Although the overall situation is good, it can be clearly seen from the figure that in some stations with small charging load, due to the large randomness of daily charging, it may be inaccurate for data

currently collected is not large enough. The prediction can still be further improved, as the charging prediction algorithm expands and new types of recorded data appear.

5. Conclusions

This paper proposes a method based on RF for EV charging load prediction and analysis, and apply it on Shenzhen actual charging data and application scenarios, and draws the following conclusions:

(1) The current EV industry in Shenzhen is still in a booming stage, and the charging load has a dispersion of small amount. After a large amount of charging data analysis, it can be observed that the charging load of EV also has temporal and spatial distribution characteristics. In terms of time, the charging load is higher in summer than in winter, and there are different distribution rules according to holidays. For space, the charging load has a distribution characteristic like that of the charging station group. Based on this, the data feature with the largest degree of discrimination is selected based on the existing data to provide the basis for the application of random forest in charging prediction.

(2) The proposed charging prediction algorithm of single station can effectively track the estimated charging capacity of the station every 15 min based on the actual recorded data. According to the simulation results, the prediction can reach a ε_{MAPE} of 9.76% and a ε_{RMSE} of 2.27. It can be used as a charging prediction method to provide reference for various EV charging load control strategies.

(3) The proposed charging prediction algorithm of station group can effectively track the estimated daily charging capacity of different charging stations based on the actual recorded data. According to the simulation results of the optimized forest structure, the prediction can reach a ε_{MAPE} of 10.83% and a ε_{RMSE} of 39.59. Also, it can be used for practical application.

Author Contributions: Conceptualization, D.X. and Y.L. (Yongpan Li); Methodology, Y.L. (Yiqi Lu); Validation, X.B., H.C. and X.Z.; Formal Analysis, Y.L. (Yiqi Lu); Investigation, Y.L. (Yongpan Li); Data Curation, E.W.; Writing-Original Draft Preparation, Y.L. (Yiqi Lu); Writing-Review & Editing, Y.L. (Yiqi Lu); Project Administration, D.X.; Funding Acquisition, E.W.

Funding: This research was funded by Shenzhen Comtop Information Technology Co. Ltd., under research "Electric Vehicle User Behavior Analysis and Power Grid Interaction Technology Based on Intelligent Terminal Technology" (090000KK52160041).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shahidinejad, S.; Filizadeh, S.; Bibeau, E. Profile of Charging Load on the Grid Due to Plug-In Vehicles. *IEEE Trans. Smart Grid.* **2012**, *3*, 135–141. [CrossRef]
2. Weckx, S.; Driesen, J. Load Balancing with EV Chargers and PV Inverters in Unbalanced Distribution Grids. *IEEE Trans. Sustain. Energy.* **2015**, *6*, 635–643. [CrossRef]
3. Wei, D.; Zhang, C.; Bo, S. A Time-of-Use Price Based Multi-Objective Optimal Dispatching for Charging and Discharging of Electric Vehicles. *Power Syst. Technol.* **2014**, *38*, 2972–2977.
4. Aziz, M.; Oda, T.; Mitani, T.; Watanabe, Y.; Kashiwagi, T. Utilization of Electric Vehicles and Their Used Batteries for Peak-Load Shifting. *Energies* **2015**, *8*, 3720–3738. [CrossRef]
5. Zhang, W.; Zhang, D.; Mu, B.; Wang, L.Y.; Bao, Y.; Jiang, J.; Morais, H. Decentralized Electric Vehicle Charging Strategies for Reduced Load Variation and Guaranteed Charge Completion in Regional Distribution Grids. *Energies* **2017**, *10*, 147. [CrossRef]
6. Chen, L.; Nie, Y.; Zhong, Q. A model for electric vehicle charging load forecasting based on trip chains. *Trans. China Electrotech. Soc.* **2015**, *30*, 216–225.
7. Qian, K.; Zhou, C.; Allan, M.; Yuan, Y. Modeling of Load Demand Due to EV Battery Charging in Distribution Systems. *IEEE Trans. Power Syst.* **2011**, *26*, 802–810. [CrossRef]
8. Madzharov, D.; Delarue, E.; D'Haeseleer, W. Integrating electric vehicles as flexible load in unit commitment modeling. *Energy* **2014**, *65*, 285–294. [CrossRef]

9. Omran, N.G.; Filizadeh, S. Location-Based Forecasting of Vehicular Charging Load on the Distribution System. *IEEE Trans. Smart Grid* **2017**, *5*, 632–641. [CrossRef]
10. Zhang, H.; Hu, Z.; Song, Y. A Prediction Method for Electric Vehicle Charging Load Considering Spatial and Temporal Distribution. *Autom. Electr. Power Syst.* **2014**, *38*, 13–20.
11. Shao, Y.; Mu, Y.F.; Yu, X.D.; Dong, X.H.; Jia, H.J.; Wu, J.Z. A Spatial-temporal Charging Load Forecast and Impact Analysis Method for Distribution Network Using EVs-Traffic-Distribution Model. *Proc. CSEE.* **2017**, *37*, 5207–5219.
12. Islam, M.S.; Mithulananthan, N.; Hung, D.Q. A Day-Ahead Forecasting Model for Probabilistic EV Charging Loads at Business Premises. *IEEE Trans. Sustain. Energy.* **2018**, *9*, 741–753. [CrossRef]
13. Liu, Y.; Shi, J.; Yang, Y.; Lee, W. Short-Term Wind-Power Prediction Based on Wavelet Transform—Support Vector Machine and Statistic-Characteristics Analysis. *IEEE Trans. Ind. Appl.* **2012**, *48*, 1136–1141. [CrossRef]
14. Jiao, R.; Zhang, T.; Jiang, Y.; He, H. Short-Term Non-Residential Load Forecasting Based on Multiple Sequences LSTM Recurrent Neural Network. *IEEE Access* **2018**, *6*, 59438–59448. [CrossRef]
15. Huang, X.; Jie, C.; Chen, Y. Load Forecasting Method for Electric Vehicle Charging Station Based on Big Data. *Autom. Electr. Power Syst.* **2016**, *40*, 69–74.
16. Guo, Q.; Wang, Y.; Sun, H.; Li, Z.; Xin, S.; Zhang, B. Factor Analysis of the Aggregated Electric Vehicle Load Based on Data Mining. *Energies* **2012**, *5*, 2053–2070. [CrossRef]
17. Liu, W.; Xu, X.; Zhou, X. Daily load forecasting based on SVM for electric bus charging station. *Electr. Power Autom. Equip.* **2014**, *34*, 41–47.
18. Wang, D.; Sun, Z. Big Data Analysis and Parallel Load Forecasting of Electric Power User Side. *Proc. CSEE* **2015**, *35*, 527–537.
19. Wu, X.; He, J.; Yip, T.; Lu, J.; Lu, N. Power System Short-term Load Forecasting Based on Improved Random Forest with Grey Relation Projection. *Autom. Electr. Power Syst.* **2015**, *39*, 50–55.
20. Wang, Y.; Xia, S.Q.; Wu, T.; Zhu, X. A Novel Consistent Random Forest Framework: Bernoulli Random Forests. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 3510–3523. [PubMed]
21. Zhang, W.; Quan, H. Dipti Srinivasan, Parallel and Reliable Probabilistic Load Forecasting via Quantile Regression Forest and Quantile Determination. *Energy* **2018**, *160*, 810–819. [CrossRef]
22. Herrera, F. On the Use of MapReduce for Imbalanced Big Data Using Random Forest. *Inf. Sci.* **2014**, *285*, 112–137.
23. Chen, F.; Deng, P.; Wan, J. Data Mining for the Internet of Things: Literature Review and Challenges. *Int. J. Distrib. Sens. Net.* **2015**. [CrossRef]
24. Che, J.; Wang, J. Short-term load forecasting using a kernel-based support vector regression combination model. *Appl. Energy* **2014**, *132*, 602–609. [CrossRef]
25. Yang, Y.; Chen, W. Taiga: Performance Optimization of the C4.5 Decision Tree Construction Algorithm. *Tsinghua Sci. Technol.* **2016**, *21*, 415–425. [CrossRef]