

Analysis of Electric Vehicle Charging Behavior under Differential Privacy Protection

Baoyi Wang^{a,*}, Ziyi Liu^b and Shaomin Zhang^c

School of Control and Computer Engineering, North China Electric Power University, HuaDian Road, Baoding, China

a. wangbaoyiqj@126.com, b. 2274890209@qq.com, c. zhangshaomin@126.com

Keywords: Electric Vehicles, Differential Privacy Protection, K-means Algorithm

Abstract: In order to reduce the impact of electric vehicles on the power grid, the charging behavior of electric vehicles is analyzed, which can improve the reliability of power grid operation. Understanding the charging behavior of electric vehicle users is helpful to improve charging service and guide charging behavior. Aiming at the problem of privacy disclosure in the process of data analysis, this paper proposes an electric vehicle charging data clustering algorithm under differential privacy protection. This method enhances the availability of clustering result by improving the selection of initial cluster center points and detecting outliers. The classification of electric vehicle users is realized through experiment, and the characteristics of charging mode corresponding to each user are analyzed.

1. Introduction

Electric Vehicles (EVs) have the advantages of energy saving and environmental protection, so they have been paid more and more attention in the world. The development of EVs is critical to achieving a comprehensive energy transformation in the field of transportation. With the popularity of EVs and the improvement of charging facilities, a large number of charging data is generated every year, which provides a data basis for the analysis of charging behavior. However, the charging behavior of EVs has the characteristics of randomness, which may cause many problems to the power grid, such as increasing the peak and valley load difference of grid, and affecting the stable operation of grid [1]. With the development of the competitive charging market, it has become an asset for power service providers to understand user charging behavior mode accurately [2].

Clustering algorithm, an important data mining algorithm, has been applied to analyze EV charging behavior. Literature [3] proposes a two-layer clustering model to determine the driving mode of EVs. Literature [4] uses British travel survey data to determine five typical traditional vehicle usage situations by k-means algorithm. In reference [5], k-means algorithm is used to cluster EV users and analyze all categories in detail. It can be seen that some achievements have been made in the analysis of EV charging behavior by using the clustering algorithm. However, the

above studies did not consider the influence of outliers and random initial center points on clustering result. Therefore, this paper improves the accuracy and availability of clustering result by detecting outliers and improving the selection of initial cluster center points.

The EV charging data may contain sensitive or private information from devices and users, and there is a risk of privacy leakage in the process of data clustering analysis [6]. Therefore, how to protect user privacy from being disclosed under the premise of accurately analyzing EV user behavior data is an urgent problem to be solved. Literature [7] considers that differential privacy protection technology is more suitable for data mining by comparing the scenarios applicable to privacy protection technology. Compared with other methods of privacy protection, differential privacy protection has the advantages of solid mathematical theory foundation and resistance to the maximum background knowledge attack. Literature [6] proposes a data clustering scheme based on k-means algorithm and differential privacy. Literature [8] proposes a method of power consumption data clustering analysis for mass users under differential privacy protection. Through the above studies, it's found that the key to privacy disclosure is cluster center. If the approximate value of each center point is provided in the process of clustering data, the sensitive data will be protected while ensuring the clustering accuracy. Therefore, in this paper, noise is added to the cluster center point, and differential privacy protection is used to reduce the risk of privacy leakage.

This paper improves the k-means algorithm by detecting outliers and selecting the initial cluster center point. In order to reduce the risk of privacy leakage, differential privacy protection is realized by adding Laplace noise to the cluster center. And this paper designs a method for the clustering analysis of EV charging data under differential privacy protection. The model proposed in this paper is helpful to guide the charging behavior and reduce the influence of EV charging on the power grid. It provides decision support for off-peak charging, and also helps to provide precise marketing services for different types of EV users.

2. Charging Behavior Analysis Model of Electric Vehicle

The model consists of the following parts: its input is preprocessed EV charging times data, the algorithm combines differential privacy protection with the improved k-means clustering algorithm, its output is the charging modes of EV users. The k-means algorithm is simple, efficient and easy to implement, Moreover, k-means algorithm and its improved algorithm have been widely used in the research of user behavior and classification, and there are successful cases in analyzing EV data. The model based on this algorithm has good practical application value.

2.1. Setting Number of Clusters

The number of clusters affects the clustering result, therefore the k-means algorithm needs to set the value of k in advance. The method used in this article to evaluate the clustering result is the Silhouette Coefficient, which is described as follows.

(1) Calculate a_i

a_i is the average distance between data d_i and other samples in the same cluster, and a_i is called intra-cluster dissimilarity of data d_i .

(2) Calculate b_i

$b_{i,m}$ is called the dissimilarity between data d_i and cluster $cluster_m$, which is obtained by

calculating the average distance between data d_i and all samples in other cluster $cluster_m$. Defined as: $b_i = \min\{b_{i,1}, b_{i,2}, \dots, b_{i,k}\}$, $m = 1, 2, \dots, k$.

(3) Calculate s_i

The Silhouette Coefficient of data d_i is as follows.

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \quad (1)$$

(4) Evaluate

In general, the average value of all s_i is the final measure used to evaluate clustering result.

2.2. Clustering Process of Electric Vehicle Charging Data

In this paper, the clustering process of EV charging data mainly includes:

(1) Selection of initial center point. First, calculate the density of each data point to determine the range of outliers, then sort the density and divide the data into k clusters in order, finally take the center of each cluster as the initial center point.

Set the dataset as $D = \{d_1, d_2, \dots, d_n\}$, n represents the number of data, the current data is represented by d_i , and other data are represented by d_l . The calculation formula of the distance from the current data point to other data point is shown in formula (2). The number of attribute value is 24, $j = 1, 2, \dots, 24$.

$$dist(d_i, d_l) = \sqrt{\sum_{j=1}^{24} (d_{i,j} - d_{l,j})^2} \quad (2)$$

The charging data of EVs usually contains data with few charging times. Removing the data when calculating the center point can improve the accuracy of clustering results. The density value of the data, as shown in formula (3), is obtained by the ratio of the number of data to the square of the distance [10].

$$density(d_i) = \frac{n}{\sum_{l=1}^n dist^2(d_i, d_l)} \quad (3)$$

The density values of all points are sorted in descending order. The parameter r of outliers is introduced. The number of data $n * (r - 1)$ are marked as outliers. In the later loop, the outliers are still part of the cluster partition, but they do not participate in the calculation of the center point. In this paper, $r = 0.9$.

The calculation formula of the initial center point is shown in formula (4).

$$centers_{m,j} = \frac{\sum_{l=(m-1)*p+1}^{m*p} d_{l,j}}{p} \quad (4)$$

Where p represents the number of data of each cluster except the outliers and $p * n \div k$.

(2) The k-means algorithm is used to cluster the dataset, and noise is added to the center point. Finally charging modes of EV users are obtained.

The calculation from the center point to the data point can be formulated as (5). Let the set of center point as $C = \{c_1, c_2, \dots, c_k\}$.

$$dist(d_i, c_m) = \sqrt{\sum_{j=1}^{24} (d_{i,j} - c_{m,j})^2} \quad (5)$$

Laplace mechanism is suitable for numerical data, and ϵ -difference privacy is achieved by adding random noise accords with Laplace distribution. In this paper, we use differential privacy protection to reduce the risk of privacy disclosure by adding noise to the cluster center. The calculation formula for updating the cluster center point is shown in (6).

$$newcenter_{m,j} = \frac{(sum_{m,j} + Lap(b))}{(count_m + Lap(b))} \quad (6)$$

Where $sum_{m,j}$ represents the sum of each dimension data in cluster m , $count_m$ represents the total number of data in cluster m , and $Lap(b)$ represents noise, $b = \Delta f / \epsilon$, Δf represents the sensitivity of the query function, ϵ is the privacy protection parameter, and it represents the amount of noise. According to formula (6), the sensitivity of this query sequence is $d + 1$, and d represents the dimension of dataset. And the dimension of dataset used in this paper is 24. This paper chooses to adjust the value of the parameter ϵ during the clustering process. The first budget consumption is $\epsilon/2$. And the budget consumption value is half of the previous iteration until end.

3. Algorithm Design

The purpose of data mining under privacy protection is to extract valuable information from a large amount of data while protecting the data privacy of users and devices. In order to better protect the privacy security of EV users, this paper adopts the improved k-means clustering algorithm under differential privacy protection. The input of the algorithm includes: the dataset $D = \{d_{1,1}, d_{1,2}, \dots, d_{1,24}; \dots; d_{n,1}, d_{n,2}, \dots, d_{n,24}\}$, the cluster number k , the outlier parameter r , the privacy protection parameter ϵ . The output of the algorithm is the clustering result, which is the charging mode of EV users.

The specific steps of the algorithm are as follows:

- (1) Traverse all data points, and calculate the distance from each data point to the other data points by using formula (2).
- (2) Calculate the density of each data point by using formula (3).
- (3) Sort the density values of all data points in the order of large to small.
- (4) Mark the $n \times (1 - r)$ data points at the end of the ordered queue as outliers.
- (5) All data points except the outliers are divided into k clusters according to the sorting order.
- (6) Calculate the initial clustering center point by formula (4).
- (7) Calculate the distance from each data point to each center point by formula (5).
- (8) Obtain the minimum Euclidean distance corresponding to each point, and each point is divided into the nearest center point.
- (9) Calculate $sum_{m,j}$ and $count_m$ in each cluster, and add the noise $Lap((d + 1)/\epsilon)$. Then the center point of each cluster is updated by equation 6.
- (10) If the convergence condition is not satisfied, repeat the step 7 to step 9; otherwise, output the clustering result.

4. Experimental Results

The data used in this paper comes from a 15-day charging dataset of an EV charging facility in China. The total number of samples is 31480. And the number of attribute features is 11, including user ID, charging start time, charging stop time, province, city, district and other attributes. Considering the different characteristics of charging behavior in different cities, in order to make the follow-up research more valuable, this paper select a city with a large amount of data for the experiment.

4.1. Setting Number of Clusters

For different values of k , the charging data of EV after data preprocessing is used for clustering analysis. The Silhouette Coefficient values corresponding to different k values are shown in Table 1. From the data in Table 1, it can be found that the Silhouette Coefficient value is the largest when k is taken as 5, that is to say, the optimal clustering result is achieved. So, choose $k = 5$.

Table 1: The Silhouette Coefficient with different k .

k	3	4	5	6	7	8
Silhouette Coefficient	0.5582	0.5687	0.5762	0.5516	0.5567	0.3217

4.2. Clustering result

According to the EV charging behavior analysis model mentioned in this paper, the pre-processed EV charging times data is taken as the input, and the clustering algorithm of EV charging data is used. The clustering result is shown in Figure 1. The different charging habits of each user can be analyzed from Figure 1.

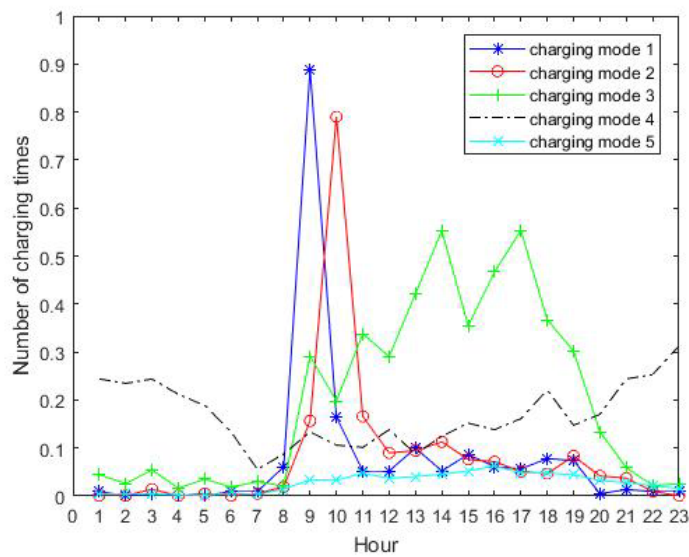


Figure 1: Charging mode of electric vehicle users.

According to the charging behavior characteristics of users, the electric power service providers can understand the personalized service requirements of each user, so as to adopt

differentiated service and marketing strategies for different types of users. For example, the users corresponding to charging mode 3 have a large number of cumulative charging times, which can provide more convenient and efficient charging services to improve the satisfaction of such users and maintain their loyalty. The charging time of the two types of users corresponding to charging mode 1 and 2 are all concentrated between 9 am and 10 am. The preferential policies such as time-sharing tariff can be used to encourage EVs to charge during off-peak periods and support the relevant decisions of dispatching management. Most of the users corresponding to charging mode 4 are charged at night, which is the low period of the grid load, so it can provide some differentiated services to reduce the loss of users. The charging mode 5 corresponds to less charging times of users, which belongs to potential user groups. These users should be maintained and motivated by active marketing programs [9].

5. Conclusion

In this paper, a clustering analysis method under differential privacy protection is proposed, which protects user privacy during the data analysis process. By improving the selection of initial clustering center and calculating the density of each data point to find outliers, the availability of data clustering results can be ensured under the premise of security. This paper determines the number of clusters k firstly, then use the clustering method to get the charging mode of EV users. Further work will focus on the analysis of charging load in different regions and periods, which have great significance for the protection of grid system.

References

- [1] Guo Jianlong, Wen Fushuan. *The influence and countermeasures of electric vehicle charging on electric power system [J]. Power automation equipment, VOL: 35, NO. 6, Pages: 1-9+30, 2015.*
- [2] L. Lin, Y. Li and L. Ma, "Research on Clustering Analysis of New Energy Charging User Behavior Based on Spark," *2018 International Conference on Sensor Networks and Signal Processing (SNSP), Xi'an, China, 2018, pp. 23-28.*
- [3] X.Li et al., "A data-driven two-level clustering model for driving pattern analysis of electric vehicles and a case study", *Journal of Cleaner Production, Vol.206, 2019, pp.827-837.*
- [4] C. Crozier, D. Apostolopoulou and M. McCulloch, "Clustering of Usage Profiles for Electric Vehicle Behaviour Analysis," *2018 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), Sarajevo, 2018, pp. 1-6.*
- [5] Y. Wang et al., "The Analysis of Electrical Vehicles Charging Behavior Based on Charging Big Data," *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China, 2019, pp. 63-67.*
- [6] J. Xiong et al., "Enhancing Privacy and Availability for Data Clustering in Intelligent Electrical Service of IoT," in *IEEE Internet of Things Journal, vol. 6, no. 2, pp. 1530-1540, April 2019.*
- [7] C. Clifton and T. Tassa, "On syntactic anonymity and differential privacy," *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW), Brisbane, QLD, 2013, pp. 88-93.*
- [8] Wang Baoyi, Hu Heng, Zhang Shaomin. *Power consumption data clustering analysis for mass users under differential privacy protection[J]. Power system automation, VOL: 42, NO. 2, Pages: 121-127, 2018.*
- [9] Zhang Lu, Li Guochang, Chen Yanxia, Sun Zhou, Wang Weixian, Tian Heping. *Electric vehicle user segmentation*

and value evaluation method based on Data Mining [J]. Power system protection and control, VOL: 46, NO. 22, Pages: 124-130, 2018.