Cluster-Based Federated Learning for Automotive Customer Profile Analysis

Qixuan Cui[1,a,*]

*[1]Southwest Jiaotong University, No. 111, Section 1, North Second Ring Road, Jinniu District,
Chengdu, Sichuan Province, China*
a. 736776835@qq.com
**corresponding author*

Abstract: In the context of increasingly fierce competition in the automotive market, car dealerships are shifting from a "product-oriented" approach to a "customer-oriented" one. This paper establishes a customer profile tagging system based on car purchasing behaviors and customer attributes, and constructs a customer profile model through cluster analysis. However, traditional machine learning algorithms face limitations such as insufficient data, single sources, and incomplete tags. Additionally, data cannot be directly shared due to data security concerns. Therefore, this paper applies a FederatedKMeans algorithm, which combines federated learning and K-means clustering, to construct profiles. The superiority of this algorithm in terms of performance is demonstrated through empirical tests, and it is used for cluster analysis to categorize customers into different groups, providing corresponding marketing strategies.

Keywords: Customer profiling, Federated learning, K-means

1. Introduction

In recent years, as the macro economy has continued to recover and improve, the automotive industry has experienced stable growth. As of 2023, China's vehicle ownership reached 336 million vehicles, with 486 million drivers[1]. The total sales volume of the automotive market in China reached 30.094 million vehicles, a significant increase of 12% compared to 2022[2]. In the context of the automotive market's demand and ownership nearing saturation and the intense competitive marketing environment, the differences in function and quality among automotive products are diminishing. Companies are shifting their focus from competition based on the product itself to competition within the customer market. Therefore, identifying customer needs in automotive marketing has become a focal point for major car dealerships. Customer profiling, as a model that characterizes customer needs, can use information about a customer's attributes, behaviors, preferences, and consumption habits to apply personalized tags, precisely targeting different types of customer groups.

Machine learning technology provides technical support for the implementation of customer profiling. However, traditional machine learning algorithms can only perform feature extraction and data analysis based on the data from a company's own database, which often leads to issues such as insufficient customer data, a single source of customers, and incomplete feature tags. From this perspective, the model needs data expansion. However, because of competitive relationships between companies, direct integration of original data can easily lead to privacy breaches, thus jeopardizing a company’s information security. From this aspect, dealers cannot share data directly. Federated learning, as an emerging distributed machine learning model, allows the joint local data of all participants to collaboratively train a model. During the training process, original data is not uploaded; instead, processed or encrypted intermediate parameters are uploaded, thus balancing the contradiction between data needs and privacy protection.

Based on the above analysis, this paper proposes to construct a customer profile model using a cluster-based Federated K-means learning framework, analyzing car buyer groups while protecting data privacy.

2. Related Work

2.1. User Profiling

The concept of user profiling is commonly believed to have been first proposed by the "Father of Interaction Design," Alan Cooper[3], who considered it a virtual representation of real users, based on a model constructed from a large amount of real data.

The analysis and construction of user profiles have also received widespread attention and in-depth study by researchers. Many scholars have successfully addressed real-world problems in different scenarios by integrating user profile analysis methods with various business contexts. For instance, Xue Haitao[4] and others researched automotive user profiles, providing guidance for precise marketing strategies in the automobile industry. Huang Xuejin[5] constructed user profiles based on vehicle charging behavior data, providing a basis for guiding orderly charging and electric grid planning expansion. Li Zonghua[6] and others proposed a user profiling model based on big data from connected vehicles, offering important references for optimizing travel experiences and related infrastructure development.

Currently, both domestically and internationally, cluster analysis and its improved algorithms are most commonly used in user profile research. Nie Xiaowei[7] and others used the K-prototype algorithm to cluster mixed-feature data and construct user profiles. Li Wei[8] and others introduced a multi-view bi-partition K-means algorithm based on Mahalanobis distance, solving the issues posed by the influence of attribute dimensions in multi-view scenarios using Euclidean distance. Han Lu[9] and others designed a K-means clustering algorithm based on Sugeno set order, reconstructing the dynamic clustering algorithm of K-means.

2.2. Federated Learning

The concept of federated learning first emerged in 2016, proposed by Google for mobile devices[10]. Their main idea was to construct machine learning models based on datasets distributed across multiple devices.

In recent years, the popularity of research in federated learning has continuously increased, and significant research achievements have been made in the area combining federated learning with user profiling. For example, Gao Sheng[11] and others proposed a personalized model training method and system that combines federated learning with user profiling. Chen Sien[12] and others introduced a user profiling and recommendation model based on federated learning. Dou Zhicheng[13] and others implemented a personalized search system that enhances privacy protection through federated learning.

3. Constructing Profiles

A complete customer profiling model involves three steps: data collection, label system design, and customer profile model construction.

3.1. Label Design

This paper designs a labeling system for car-buying customer profiles, which encompasses basic customer attributes, behavioral attributes, and vehicle attributes, as illustrated in Figure 1.

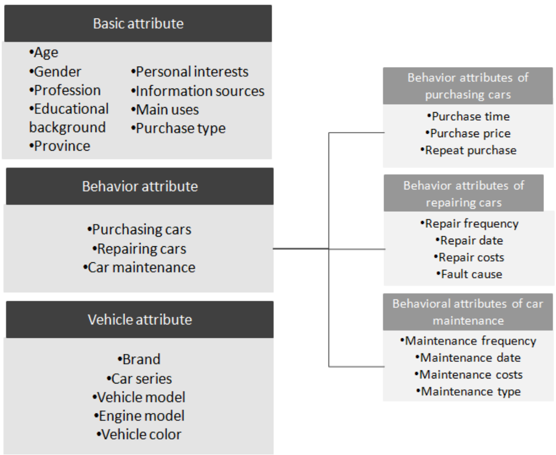

Figure 1: Automotive Customer Profiling Label System

3.2. Model Construction

This paper constructs a customer profiling model based on a clustered federated learning architecture. In the clustered federated learning algorithm, there are two key roles: the central server and the participant clients. In this study, participant clients refer to dealership companies at the same hierarchical level, each possessing its own customer dataset locally. The central server in this study refers to a third-party platform that has no competitive relationship with the dealership companies, responsible for aggregating model parameters and constructing a global customer profile model.

3.2.1. K-means Clustering Algorithm

The K-means clustering algorithm, as an unsupervised learning method, groups similar customer data points into clusters without the need for explicit labels or categories, thereby dividing customers into different groups.

Given a dataset X, and n samples within the dataset, these samples are divided into k clusters, with each cluster having a corresponding cluster center. The clustering objective is to minimize the sum of distances between all data samples and their respective cluster centers within the clustering groups. The K-means clustering process includes the following steps:

1. Initialize Centers: Randomly select k data points from the dataset as the initial clustering centers.

2. Calculate Distances: Calculate the distance from each sample to the clustering centers and assign data points to the nearest cluster center.

3. Recalculate Centers: After all data points are assigned, recalculate the means of the k clusters to serve as the new cluster centers.

4. Iterate: If the cluster centers have changed compared to the previously calculated k cluster centers, return to step 2; otherwise, proceed to step 5.

5. Termination: The iteration ends and the clustering results are output when the cluster centers no longer change.

3.2.2. Cluster-Based Federated K-means Learning Algorithm

The clustered federated learning algorithm combines the advantages of federated learning in protecting data security with the characteristics of unsupervised learning through clustering analysis. This paper uses the FederatedKMeans algorithm, a cluster-based federated learning algorithm derived from K-means, whose architecture is shown in Figure 2:

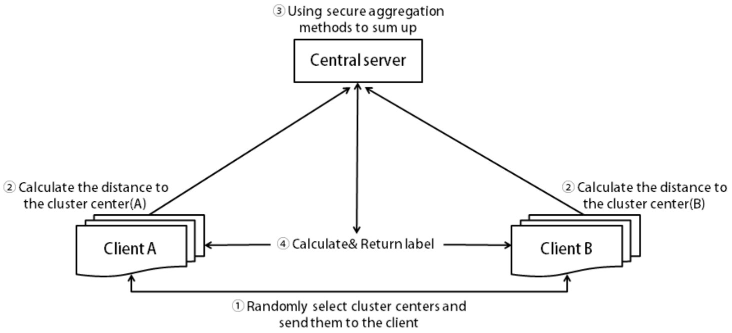
③ Using secure aggregation methods to sum up
Central server
② Calculate the distance to the cluster center(A)
② Calculate the distance to the cluster center(B)
Client A
④ Calculate& Return label
Client B
① Randomly select cluster centers and send them to the client

Figure 2: FederatedKMeans Algorithm Framework

The FederatedKMeans algorithm is a federated learning method that aggregates distributed data into k clusters. The implementation process is described from the perspectives of the central server and the clients:

The central server first initializes K cluster centers, then selects a proportion P of clients from the total M to form the participant set G_t for global model updates. Initial model parameters are distributed to G_t; clients perform local model training and subsequently upload the model parameters to the server. The server calculates the size of the global model training cluster set $(\lambda_1,\ldots,\lambda_k \in R)$ Λ, and with the learning rateα, computes the weighted global model training cluster centroids $(\gamma_1,\ldots,\ \gamma_k \in R^m)$ Γ set. The next round of model training begins until the set number of federated learning global training rounds is reached. The process is as shown in Algorithm 1.

```
Algorithm 1: FederatedKMeans Central Server Algorithm
Server executes:
  initialize cluster centroids $\Gamma_{t=0}$randomly.
  m ← max(P·M,1)
  for each round t =1,...,T do
      $G_t$ ← (random set of m clients)
      for each client i $\in G_t$ in parallel do
          $(\Lambda_t^{(i)}, \Gamma_t^{(i)})$ ← ClientKMeans(i, $\Gamma_{t-1}$)
      $\Lambda_t = \sum_{i=1}^{m} \Lambda_t^{(i)}$
      $\Gamma_t^* = \frac{1}{\Lambda_t} \sum_{i=1}^{m} \Lambda_t^{(i)} \Gamma_t^{(i)}$
      $\Gamma_t = \Gamma_{t-1} + \alpha \cdot \left(\Gamma_t^* - \Gamma_{t-1}\right)$
```

Clients M_i receive global model parameters and update their local training models. They divide their private dataset D_i into small batch-size datasets B and shuffle them randomly. In each batch b $\in$B, clients randomly select K cluster centroids $(\gamma_1,\ldots,\gamma_k \in R^m)$ to form K clusters c_k. Each data point x_j calculates its distance to centroid γ_k, with x_j belonging to the cluster c_j that has the closest centroid γ_k. The cluster size set ($\lambda_1,\ldots,\lambda_k \in R$) Λ and the cluster centroids ($\gamma_1,\ldots,\gamma_k \in R^m$) Γ set are computed based on the local learning rate β. After local training, model parameters are uploaded to the server for global model updating, awaiting the next model training. The process is as shown in Algorithm 2.

```
Algorithm 2: FederatedKMeans Client Algorithm
Client executes:
  ClientKMeans(i,Γ)
      B←(split D_i in batches of size B)
      for each epoch e=1,...,E do
          initialize cluster sizes Λ to zero.
          randomly shuffle order of batches in B
          for each batch b∈B do
              for each data x_j∈b in parallel do
                  $c_j = \arg\min_k \left\| x_j - \gamma_k \right\|^2$
              $\Lambda^b = \sum_{j=1}^{B} l_{\{c_j=k\}}$
              $\Gamma^b = \frac{1}{\Lambda^b} \sum_{j=1}^{B} l_{\{c_j=k\}} \cdot x_j$
              $\Lambda = \Lambda + \Lambda^b$
              $\Gamma = \Gamma + \beta \cdot \frac{\Lambda^b}{\Lambda} \cdot \left(\Gamma^b - \Gamma\right)$
      return Λ,Γ
```

4. Experiments and Results

4.1. Dataset and Preprocessing

The Multi-Value Chain Collaborative Services Cloud Platform is a professional platform providing information services for the automotive industry chain. Over more than ten years since its establishment, it has supported thousands of companies, generating millions of business data entries and providing reliable data services for these companies' production, procurement, sales, claims, and maintenance operations. Based on this platform, this paper takes a dealership in the automotive marketing chain as an example. As of January 1, 2024, 2500 historical sales records, along with 5000 maintenance and 5000 repair records from upstream and downstream service stations, were extracted as sample data, quantified as shown in Table 1.

Table 1: Examples of Sample Feature Data

ID	Purchase time	Purchase price	Repeat purchase	Repair frequency	Repair time	Repair costs	Maintenance frequency	Maintenance time	Maintenance costs
1	342	53000	1	1	75	483.7	1	30	727.2
2	459	42000	1	5	388	421.4	1	362	583.5
3	450	45500	2	3	261	632.2	3	349	630.3
…	…	…	…	…	…	…	…	…	…

To eliminate the impact of excessive numerical differences, this paper uses the Z-score standardization method for normalization. For a given matrix element X_i, the standardization formula is as follows: $X^* = \frac{X_i - mean}{standard\ deviation}$.

4.2. Experimental Metrics

In addressing multi-classification tasks, we commonly rely on metrics such as precision, recall, and the F1 score to comprehensively evaluate the performance of the model. The formulas for precision (P) and recall (R) are as follows:

$$P = \frac{TP}{TP+FP} \tag{1}$$

$$R = \frac{TP}{TP+FN} \tag{2}$$

The paper also introduces the Macro-average as a metric to evaluate the classification effect of the model. The Macro-average is a composite score derived by calculating the F1 score for each category and then averaging these scores arithmetically. The formula for the Macro-average is provided below.

$$Macro - average = \frac{2 \times \frac{1}{K}\sum_{k=1}^{K} P_k \times \frac{1}{K}\sum_{k=1}^{K} R_k}{\frac{1}{K}\sum_{k=1}^{K} P_k + \frac{1}{K}\sum_{k=1}^{K} R_k} \tag{3}$$

4.3. Model Comparison

This paper used the Python language and the Pytorch open-source framework to build a federated learning simulation platform. In the experiment, multi-threading technology was utilized to simulate the training and federated communication processes of 5 participants, setting the number of iterations for training data to 10, a learning rate of 0.01, a local training batch size of 120, local training rounds to 3, and global training rounds to 10000.

The experiment compared the performance of the FederatedKMeans algorithm based on clustered federated learning and a single participant's K-means algorithm, selecting results from iterations 1, 15, 25, 50, 100, and 1000, with the precision and macro-metrics results shown in Figures 3 and 4. Here, K1-K5 represent the results of K-means clustering algorithms performed by 5 individual participants on their respective data.

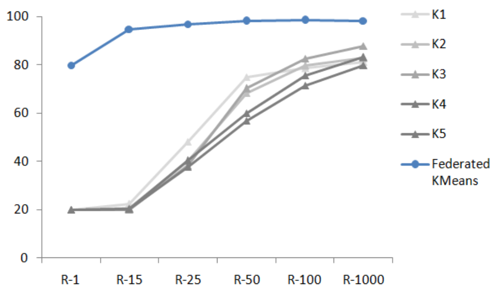

Figure 3: Comparison of precision between FederatedKMeans and single-party K-means

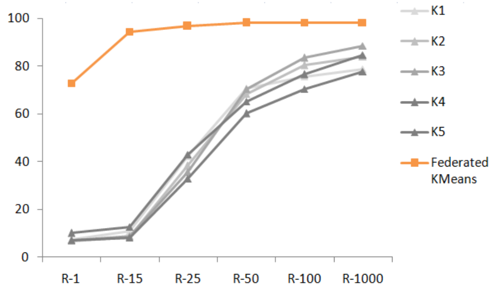

Figure 4: Comparison of Macro-average between FederatedKMeans and single-party K-means

Averages were calculated for the precision and macro-metrics for K1-K5, providing average values for different metrics of the K-means algorithm. A comparison between K-means and FederatedKMeans algorithms is shown in Table 2.

Table 2: Accuracy and Macro-average Results of Clustered Federated Learning versus Single-Party Learning

	Index	R-1	R-15	R-25	R-50	R-100	R-1000
KMeans	Precision	20.00%	20.74%	40.98%	66.12%	77.63%	83.03%
	Macro-average	7.74%	9.79%	38.42%	67.03%	77.31%	82.80%
FederatedKMeans	Precision	79.88%	94.85%	96.92%	98.1%	**98.45%**	98.31%
	Macro-average	72.85%	94.38%	96.7%	98.12%	**98.38%**	98.24%

The comparative results demonstrate that FederatedKMeans significantly outperforms the single K-means algorithm on both precision and macro-average metrics. This advantage mainly reduces the issues of scarce data samples or insufficient sample diversity encountered by individual participants during training, and increases the number of training data samples, thus enhancing the model's precision and macro-average scores in the test set.

4.4. Experimental Results and Analysis

The silhouette coefficient is an indicator used to evaluate the effectiveness of clustering; the closer the value approaches 1, the better the clustering performance. The silhouette coefficient for this experiment is shown in Figure 5.

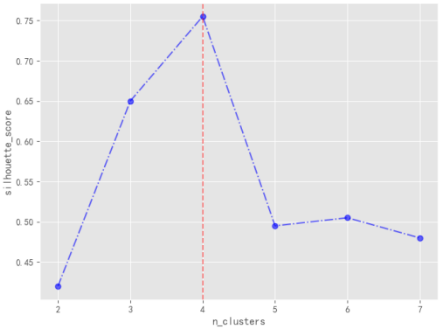

Figure 5: Clustering Silhouette Coefficient Diagram

As can be seen in the diagram, the best clustering result occurs at cluster=4. The distribution of customer types is shown in Table 3.

Table 3: Distribution of Customer Types

cluster				valid	missing
1	2	3	4		
184	880	654	327	2045	0

The analysis of these four data types reveals four distinct customer groups. The first type, comprising approximately 9% of customers, makes high-value car purchases and frequents after-sales services, indicating high customer value and brand loyalty. Dealers should pay special attention to the dynamics of this customer group, recommending high-end luxury models to them. The second type, making up 43%, spends a moderate amount on car purchases and frequents maintenance services, indicating that these customers prioritize car maintenance and safety. Dealers could recommend higher-grade models with better performance. The third type, accounting for 32%, spends moderately on car purchases with long intervals between purchases and low frequency of after-sales consumption, suggesting potential car buying needs. Dealers could guide consumption appropriately and recommend the latest models according to customer preferences. The fourth type, making up 16%, spends less on car purchases and infrequently seeks maintenance services, indicating a lower level of attention to cars. Dealers could schedule regular visits to prevent customer attrition.

5. Conclusion

This paper analyzes car purchasing customer information and historical sales records from dealerships to establish a car purchasing customer profile model using the FederatedKMeans clustering federated learning algorithm. The experimental results demonstrate that this algorithm surpasses the single participant K-means algorithm in both precision and macro-average metrics. Based on this algorithm, car purchasing customers were categorized into four distinct customer groups, with corresponding marketing recommendations provided. Future steps could involve recommending vehicles (model, color, etc.) to customers based on their historical purchasing preferences and similarities among customers.

References

[1] Zhang, T.P. (No date provided, assume current year if recent). The number of motor vehicles in China reaches 435 million. Retrieved fromhttps://www.gov.cn/lianbo/bumen/202401/content_6925362.htm

[2] Gao, K., & Wang, Y.Y. (2023). China's automobile production and sales exceed 30 million for the first time. Retrieved from https://www.gov.cn/yaowen/liebiao/202401/content_6925448.htm

[3] Cooper A,Robert Reimann R,Cronin D.About Face 3:The Essentials of Interaction Design[M].New Jersey:Wiley Publishing Inc.,2007:19-22.

[4] Xue, H.T., He, H.Y., Chen, Y.Z., et al. (2023). Precision marketing strategies for Wuling new energy vehicles. Times Automotive, (1), 184-187.

[5] Huang, X.J., Zhong, J.X., Lu, J.Y., et al. (2023). Prediction method of electric vehicle charging load based on user profiles. Journal of Jilin University: Engineering Edition, 53(8), 2193-2200.

[6] Li, Z.H., Zhai, J., Diao, G.T., et al. (2022). Research on the application of electric vehicle driver behavior profiles based on vehicle network data. Science and Technology Trend, (18), 61-64.

[7] Nie, X.W., Deng, K., & Tang, Y. (2023). Research on the construction path and application value of college student user profiles based on K-prototype. Chinese Journal of Science and Technology (Full Text Edition) Education Science, (2), 150-153.

[8] Li, W., Hu, Y.F., & Li, P.L. (2020). Research on university library user profiles based on multi-viewpoint binary k-means. Journal of Zhejiang University of Technology, 48(2), 141-147.

[9] Han, L., Su, Z., & Li, A.H. (2019). Research on credit user clustering under Sugeno measure. Systems Engineering Theory and Practice, 39(11), 2750-2759.

[10] McMahan H B,Moore E,Ramage D,et al.Communication-Efficient Learning of Deep Networks from Decentralized Data[C]//International Conference on artificial intelligence and statistics.Seattle:[s.n.],2017:1273-1282.

[11] Gao, S., Zhou, X.Y., Zhang, B.S., et al. (2021). Personalized model training method and system combining federated learning and user profiling. China Patent.

[12] Chen, S.E. (2021). A method for analyzing urban traffic travel data based on federated learning. China Patent.

[13] Dou, Z.C., Yao, J., & Wen, J.R. (2021). A personalized search system with enhanced privacy protection based on federated learning. China Patent.