Power user Data Feature Matching Verification Model based on TSVM Semi-supervised Learning Algorithm

Yakui Zhu*, Rui Zhang, Xiaoxiao Lu
Marketing Service Center, State Grid Hebei Electric Power Co. Ltd, Shijiazhuang, China

***Abstract*—The existing model for identifying user data features based on smart meter data adopts a supervised learning method. Although the model has good identification performance under the condition of sufficient index samples, matching data are difficult to obtain and the marking cost is high in real life. The identification accuracy is significantly reduced when the matching data are insufficient or unavailable in the supervised learning method. In view of the above problems, based on the smart meter data, this paper proposes a feature recognition method for residential user data based on semi-supervised learning, which uses three indicators to evaluate the recognition performance of the proposed semi-supervised learning method for residential user data features and to find the appropriate feature selection method and data acquisition resolution. Then, explore the role of this method in real life when there is insufficient or unavailable matching data. Experimental results show that the performance of the proposed semi-supervised learning algorithm is better than that of the supervised learning algorithm, and the accuracy of the proposed algorithm is better than or close to that of the supervised learning algorithm.**

***Keywords*—Power system; data matching; data characteristics; semi-supervised learning algorithm; load model**

I. INTRODUCTION

With the continuous development of smart electricity business such as demand response and energy efficiency management, scholars at home and abroad have conducted in-depth exploration and research on the relationship between smart meter data and residential user data characteristics. The characteristic information of residential user data affects the electricity consumption behavior of residents, and conversely, the portrait information of residents can also be identified from their electricity consumption behavior and data [1]. Therefore, the topic of exploring the potential correlation between smart meters and residential user data characteristics is mainly divided into two categories: one is to explore the residential electricity load pattern and analyze how the user data characteristics affect the electricity load type; the other is to identify the residential user data characteristics through the residential electricity load characteristics [2].

The residential load has strong randomness, and the characteristic data of residential users can help power companies to better understand the characteristics of residential peak load and the reasons for changes in electricity consumption behavior. Scholars in China and abroad have done a lot of research on the first type of topic [3]. Wang Yi et al. proposed a new dynamic clustering method for power consumption behavior. Firstly, the symbolic aggregation approximation measure is used for each user to reduce the size of the data set, and the Markov algorithm based on time series is used to establish a dynamic energy consumption model, which converts a large number of load data curves into matrix form. Secondly, the typical power consumption pattern is obtained through the fast clustering algorithm based on the density peak, and then the Kullback Liebler distance is used to evaluate the difference in power consumption behavior, and the users are clustered. The example in this paper verifies the effectiveness of this method [4]. Wang Fei et al. used the density-based clustering algorithm DBSCAN to extract the seasonal typical electricity consumption patterns of each user and used the K-means clustering algorithm to cluster the electricity consumption patterns, and finally used the association rule mining algorithm to explore the potential relationship between the residential electricity consumption pattern and its user data characteristic factors [5]. These works deeply explore the factors affecting the electricity load pattern of users, which can effectively promote the implementation of energy-saving projects. However, these studies need to integrate a large number of user data characteristics, classify and manage users according to specific user data characteristics rather than load patterns, and make some personalized service policies for different types of families, which are more easily accepted and understood by non-professional technicians [6]. Therefore, how to automatically, intelligently, and accurately identify the characteristic information of resident user data has become the core work of the association. In recent years, there are many methods applied to identify the characteristics of residential user data in smart meter data.

In the research of user portrait recognition based on the smart meter data, the key task is to extract and select the features of smart meter data, which directly affects the upper limit of the performance of the user portrait recognition model [7]. In the existing literature, the extraction of smart meter data features is focused on a single domain (single time domain or single frequency domain), which fails to comprehensively analyze the potential rules contained in smart meter data from multiple perspectives [8].

However, the existing models for identifying user data features based on smart meter data all adopt the supervised learning method. Although it achieves good recognition performance when the index sample is sufficient, the recognition accuracy is significantly reduced when the

*Corresponding Author.

matching data is insufficient or unavailable [9]. However, in real life, it is difficult to obtain matching data, the cost is high, and it is time-consuming and labor-intensive. How to save the cost of sample labeling while maintaining good user data feature recognition performance is an urgent problem to be solved [10].

To solve the problems in real life, such as difficulty in obtaining matching data, high cost, time-consuming, and labor-consuming, this paper proposes a feature recognition method of residential user data based on semi-supervised learning based on extracting the time domain and frequency domain features of smart meters. This method can make full use of the potential rules contained in a small number of matching data and a large number of non-index data to explore the relationship between smart meter data and residential user data characteristics and reduce the cost of index marking. In this paper, the effectiveness of semi-supervised learning is verified by the real CER data set, and two main factors affecting the performance of the semi-supervised learning recognition algorithm are analyzed.

The main innovations of this paper are:

1) It decomposes an original average daily power consumption curve by adopting a discrete wavelet transform and extracting frequency domain characteristics;

2) The method of combining time domain and frequency domain features is helpful to improve the accuracy of portrait recognition.

3) The applicability and expansibility of the feature extraction method. Based on the combination of time domain and frequency domain characteristics, the resolution of smart meter data acquisition will have an impact on the results of portrait recognition.

The main contents of the paper are as follows:

1) It introduces the research motivation, background and research status of this paper, and puts forward the solutions to the existing problems.

2) It explains the basic principle of the technical content.

3) The data feature recognition method based on semi-supervised learning is introduced.

4) Through the experimental analysis, the technical advancement and reliability of the research content of this paper are compared and tested.

5) In the conclusion part, the research results of this paper and the future work are summarized.

II. Basic Principles of Semi-supervised Classification

Semi-supervised generative classification algorithms assume that different classes of data are generated by potentially different "sources", and that the samples of each class follow a probability distribution $p(x \mid y;\theta)$, where θ is the parameter of the probability density distribution function [11]. If the samples without index and the samples with index come from the same probability distribution, the samples without index whose index values are inferred can be used as training samples to improve the classification accuracy of the model [12].

The semi-supervised expectation-maximization (EM) algorithm is a typical generative classification model, which assumes that its data distribution conforms to the Gaussian mixture model, and that the data of each class follows the normal distribution [13]. The joint probability density of the sample data and the index can be obtained from the conditional probability density function of the class, as shown in (1).

$$p(x, y \mid \theta) = p(y \mid \theta)\, p(x \mid y, \theta) \tag{1}$$

The parameter vector of the model can be determined according to the samples with and without indicators, which is transformed into solving the optimization problem as shown in (2) [14].

$$\max_{\theta}(\ln p(\{x_i, y_i\}_{i=1}^{l} \mid \theta) + \lambda \ln p(\{x_i, y_i\}_{i=l+1}^{l+u} \mid \theta)) \tag{2}$$

Where, λ is the artificially set parameter. $x_1, x_2, \cdots, x_l$ is the sample with index, and $x_{l+1}, x_{l+2}, \cdots, x_{l+u}$ is the sample without index. The EM algorithm is used to solve this problem. First, a set of parameters is estimated according to the sample with index, and then the expectation is calculated (step E) [15]. The sample without index is marked according to the estimated parameters, and the maximum likelihood estimate is calculated. Maximize the maximum likelihood estimate (M-step) obtained by E-step, re-estimate and update the parameters, and then start the next round of E-step calculation, repeat this EM process until the parameters converge [16].

III. Feature Recognition Method of Residential user Data based on Semi-supervised Learning

A. Overall Methodology Framework

The process of identifying the features of residential user data based on semi-supervised learning algorithm can be divided into three parts, and the overall method framework is shown in Fig. 1.

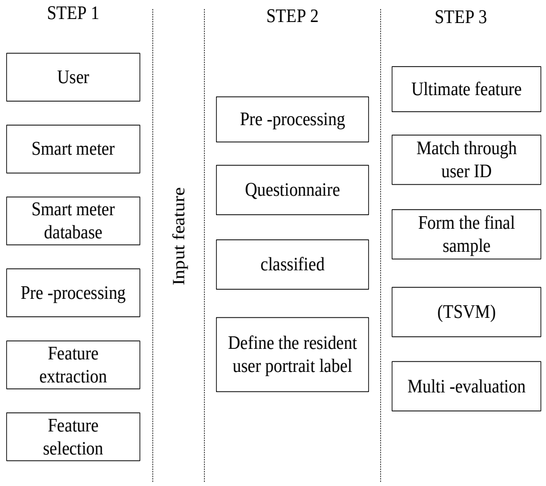

Fig. 1. Overall Method Framework of Resident user Data Feature Recognition Model based on Semi-supervised Learning.

1) First, remove the abnormal values and missing values of the smart meter data, then extract 54 time domain features and 24 frequency domain features. Normalize the features, and finally use the feature selection algorithm to screen the extracted load features, and select the input features of the subsequent identification model according to the importance ranking [17].

2) Sort out the questionnaire data related to user data characteristics, mainly analyze the user data characteristics that have a greater impact on power consumption, and classify and define the classification indicators for each type of user data characteristics according to the answer results of the questionnaire. The classification indicators can be used as real user indicators for the calculation of model performance indicators [18].

3) Match the finally selected smart meter data characteristics with the resident user data characteristic indexes through the user ID to form a final sample set. The semi-supervised learning method is used to train a small number of samples with indicators and a large number of samples without indicators to obtain the user data feature recognition model based on semi-supervised learning. Finally, the performance evaluation index is used to verify the effectiveness of the model [19].

B. Feature Selection

In this paper, three methods are used to select features respectively to facilitate the subsequent verification of the impact of different feature selection methods on semi-supervised learning to identify the features of residential user data.

1) Filtration: For the 78 data features extracted from the CER data set, the filtering method first uses the variance discrimination method, sets the variance threshold, and eliminates the features whose variance is less than the threshold (that is, there is no discrimination) [20]. Then the Pearson correlation coefficient method is used to calculate the Pearson correlation coefficient of the remaining features and the real user data feature identification classification index, and the R most important (i.e., most relevant) features are selected according to the coefficient value ranking [21].

2) Packaging method: For the packaging method, the Recursive Feature Elimination (RFE) method based on Logistic Regression (LR) algorithm is used in this section to select the features extracted from the CER data set. Firstly, a weight value is assigned to each original feature in the initial training. Secondly, the LR model is used to predict the classification index. Then the predicted classification index is compared with the real index to calculate the recognition error. The weight of each feature is updated according to the error, and the feature with the smallest absolute value of the weight is proposed in each round [22]. Repeat this step until the required number of features is reached. Finally, these features with larger weights are used as the input of the subsequent classification model. A schematic diagram of feature selection based on the LR-RFE algorithm is shown in Fig. 2.

3) Embedding method: The Random forest measures its ability to identify user data features by calculating the PI value of each feature. When calculating the importance of the extracted feature N, a decision tree i is created. The OOBErrori is first calculated. Then, the values of the out-of-bag data feature N are randomly rearranged, and the rest of the features remain unchanged to form a new out-of-bag data set OOBi. According to the new OOB, the OOBErrori is recalculated. The PI value of the feature N in the ith tree can be obtained by subtracting the results of the two calculations, and the calculation formula is shown in (3).

$$PI_i(N) = OBError_i^{'} - OOBError_i \tag{3}$$

The calculation process is repeated for each tree of the random forest, and the final PI value of the feature N can be obtained by summing and averaging the PI values of the feature N of each tree, as shown in (4).

$$PI(N) = \frac{1}{c}\sum_{i=1}^{c} PI_i(N) \tag{4}$$

C = ntree represents the number of decision trees used in the random forest. If the importance of a feature ranks high, it means that its value has discrimination between different samples. After the eigenvalues are randomly reordered on the out-of-bag data set, their discrimination for different user samples is reduced, thus improving the OOBErrori. Therefore, the higher the PI value, the higher the importance of the feature. The process of using the random forest algorithm to rank the importance of features is shown in Fig. 3.

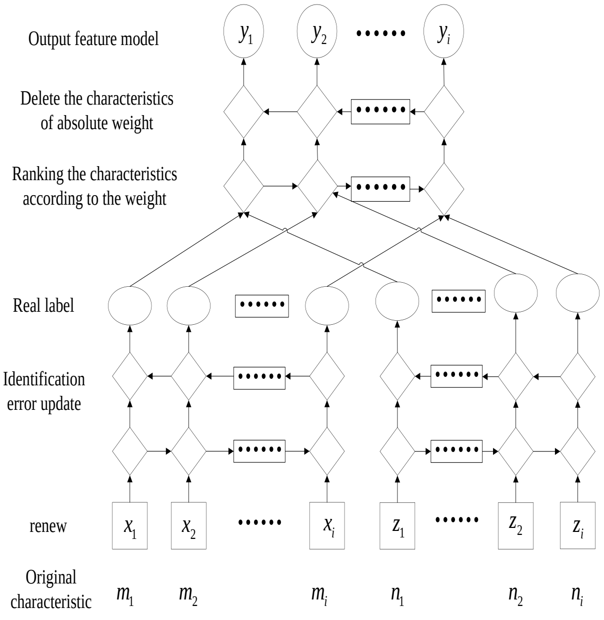

Fig. 2. Schematic Diagram of Feature Selection based on LR-RFE Algorithm.

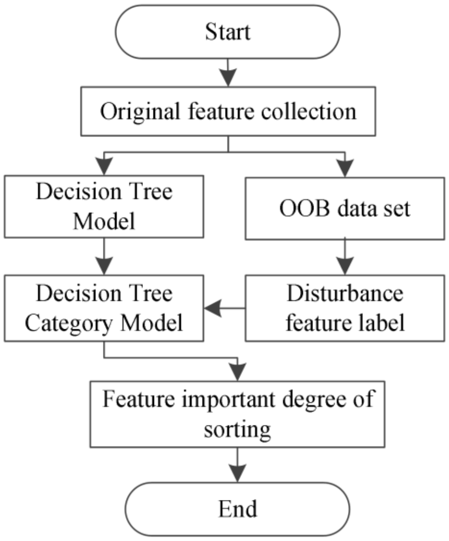
Start
Original feature collection
Decision Tree Model
OOB data set
Decision Tree Category Model
Disturbance feature label
Feature important degree of sorting
End

Fig. 3. Ranking Process of Feature Importance Calculated by Random Forest Algorithm.

C. Performance Evaluation Index

In this paper, the TSVM semi-supervised learning method is used to identify the characteristics of residential user data. Three performance evaluation indicators, namely ACC, F1-Score and AUC, are used to evaluate and analyze the recognition model.

When training the SVM classifier, five-fold cross validation is used to train the samples to verify the recognition performance of the data features of residential users more reliably. To evaluate the classification performance of the classifier from multiple perspectives, several evaluation indexes are given here.

- Accuracy

For user data features with M category indexes, the confusion matrix C of $M \times M$ can be calculated. J represents the number of features with the category index of m that are misclassified into the category index of n. If m = n, then Cm, n represent the number of correct classifications, and vice versa. Accuracy (ACC) can be expressed by formula (5).

$$Accuracy = \frac{\sum_{m=1}^{M} C_{m,m}}{\sum_{m=1}^{M}\sum_{n=1}^{M} C_{m,m}} \tag{5}$$

For the binary classification problem, the confusion matrix shown in Table I can be obtained by comparing the sample indicators identified by the classification model with the real sample indicators.

TABLE I. BINARY CLASSIFICATION CONFUSION MATRIX

	Positive	Negative
The prediction is positive	TP	FP
The prediction is negative	FN	TN

True Positives (TP): the number of samples that are actually positive and predicted to be positive; False Positives (FP): the number of samples that are actually negative and predicted to be positive.

False Negatives (FN): the number of samples that are actually negative and predicted to be positive.

True Negatives (TN): the number of samples that are actually negative and predicted to be negative. Therefore, ACC can also be expressed by the following formula (6):

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{6}$$

- F1-Score

According to the confusion matrix, several performance evaluation indexes can be obtained.

Precision: The proportion of samples that are actually positive (positive) among the samples predicted to be positive (positive), as shown in formula (7).

$$Presicion = \frac{TP}{TP+FP} \tag{7}$$

Recall: The proportion of samples that are correctly classified as positive among all samples that are really positive (positive examples), as shown in formula (8).

$$\mathrm{Re}call = \frac{TP}{TP+FN} \tag{8}$$

F1-Score is a comprehensive index reflecting precision and recall, and its value range is 0 to 1. The closer the value of F1-Score is to 1, the better the recognition performance of the model is, and its calculation formula is shown in (9).

$$F1-Score = \frac{2\times Presicion \times Recall}{Presicion + Recall} \tag{9}$$

IV. EXPERIMENTAL ANALYSIS

A. Experimental Environment Setting

This paper also uses the open CER data set to analyze the data characteristics of six residential users. See Table II for details.

To verify the effectiveness of the proposed method, the recognition performance of TSVM semi-supervised learning algorithm is compared with four classical KNN, RF, SVM and MLP supervised learning algorithms, and two examples are set to verify the results.

To verify whether the recognition performance of the resident user portraits can be improved by adding frequency domain features on the basis of time domain features, SVM is used to train the features of smart meters, and the grid search method is used to optimize the SVM parameters. 80% of the user sample is the training set and 20% is the test set. The selection of the main hyper-parameters of SVM in the three cases of only time domain feature (Model 1), only frequency domain feature (Model 2), and combination of time domain and frequency domain (Proposed Model) is shown in Table III.

TABLE II. USER DATA CHARACTERISTIC DESCRIPTION AND INDEX DEFINITION TABLE BASED ON CER DATA SET

Number	**User data Characteristics**	**User data characterization**	**Category**	**Indicators**	**Sample size**
1	Employment	Employment status of the family's main earner	Hire	1	1423
			Not hired	2	1026
2	Population	Number of family members	Little (no<2)	1	1321
			Many (no.N3)	2	1128
3	Housing types	Housing types	Freestyle	1	1299
			Connection type	2	1104
4	Occupancy rate	Is the house unused for more than 6 hours per day?	Yes	1	1619
		Is the house unused for more than 6 hours per day?	No	2	345
5	Cooking type	Type of cooking facility	Electricity	1	1712
			Non-electricity	2	737
6	With or without children at home	With or without children at home	Yes	1	1964
			No	2	485

TABLE III. MAIN HYPER-PARAMETER SETTINGS OF SVM CLASSIFIERS

SVM	Model 1			Model 2			Proposed model		
serial number	Kernel	C	gamma	Kernel	C	gamma	Kernel	C	gamma
1	RBF	99	0.02	RBF	23	0.2	RBF	5	0.02
2	RBF	97	0.02	RBF	85	0.2	RBF	59	0.002
3	RBF	33	0.02	RBF	3	0.2	RBF	67	0.02
4	RBF	2	20	RBF	2	200	RBF	2	20
5	RBF	22	0.02	RBF	43	20	RBF	32	0.02
6	RBF	2	0.2	RBF	77	0.2	RBF	2	0.2
7	RBF	2	20	RBF	2	200	RBF	2	20

First of all, for the basic results of TSVM, the accuracy values of population, housing occupancy, cooking type and whether there are children in the house are higher than 75%, and the accuracy values of employment and housing type are between 60% and 70%. The recognition accuracy of all user data features is higher than 60%, which shows the basic effectiveness of the semi-supervised learning method proposed in this paper.

Comparing the TSVM semi-supervised learning algorithm with other supervised learning algorithms, when the proportion of index samples is set to 5%, for any user data feature, the recognition accuracy, F1-Score and AUC of TSVM are higher than those of other supervised learning algorithms.

B. Example 2 Experimental Setup and Result Analysis

To further verify the performance of the proposed method, the proportion of samples with index for the semi-supervised learning algorithm is set to 5%, and the proportion of samples with index for the supervised learning algorithm is set to 10 times that of the semi-supervised learning model, that is, 50%. Here, the LR-RFE-based feature selection method is still used to select the top 20 features that are correlated with the classification target value. In this case, the ACC, F1-Score and AUC values of the TSVM semi-supervised learning algorithm and the four supervised learning algorithms of KNN, RF, SVM and MLP to identify the characteristic indicators of residential user data are shown in Fig. 4.

In Fig. 4, for the TSVM semi-supervised algorithm, except for the user-data feature of population number, the recognition ACC values of other user data features are higher than those of other supervised learning methods. For the population, the ACC and AUC values identified by TSVM with only 5% of the matching data are slightly lower than those identified by the supervised learning method with 50% of the matching data, but the difference is not significant.

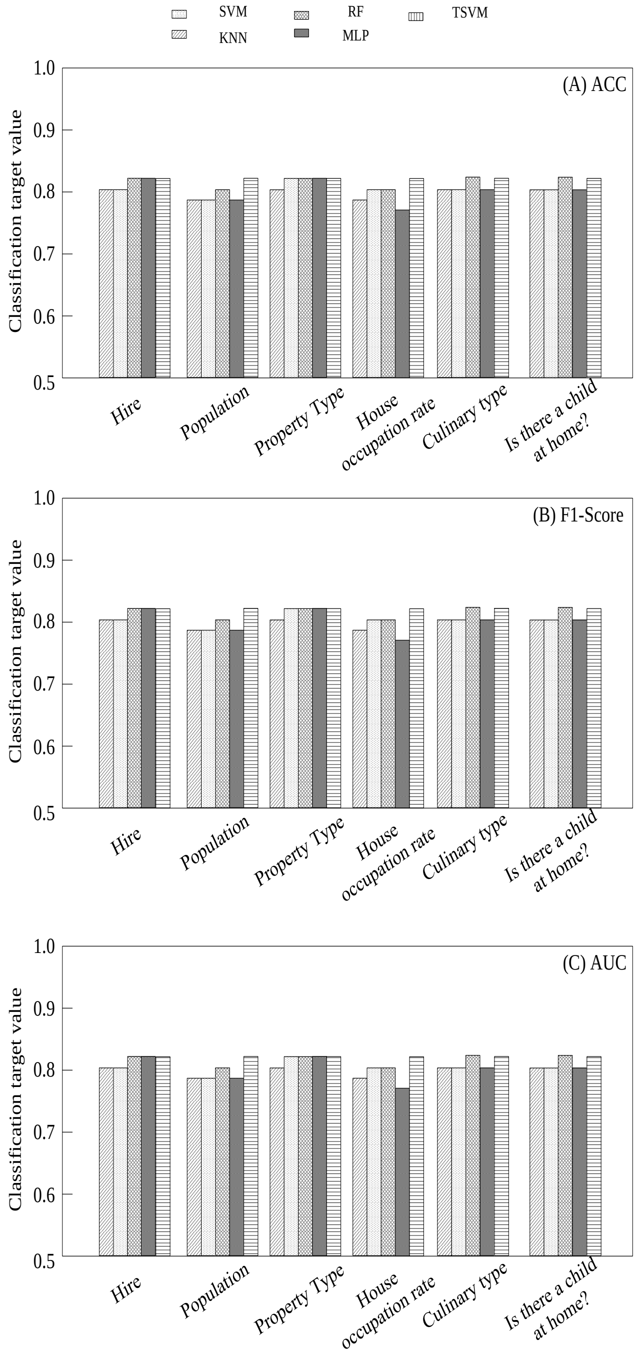

Fig. 4. Comparison of Semi-supervised and Supervised Recognition Algorithms.

V. Conclusion

In this paper, a feature recognition method for residential user data based on semi-supervised learning in the case of limited index samples is proposed. The main work includes:

1) Based on the extracted time domain and frequency domain features of smart meters, a feature recognition method for residential user data based on semi-supervised learning is proposed.

2) Then, the effectiveness of semi-supervised learning is verified by the real CER data set.

3) Two main factors affecting the performance of semi-supervised learning recognition algorithms are analyzed.

This method can make full use of the potential rules contained in a small number of matching data and a large number of non-index data to explore the relationship between smart meter data and residential user data characteristics and reduce the cost of index marking.

In future work, it is necessary to further explore the relationship between smart meter data and residential electricity consumption behavior habits, update the identification scenario, and integrate the user data characteristic information into residential load pattern forecasting, baseline load estimation or high demand response potential user identification from the user demand side.

References

[1] Meher S. Semi-supervised self-learning granular neural networks for remote sensing image classification. Applied Soft Computing, 2019, 83:105655.

[2] Muhammad F, Alberto S. Analyzing load profiles of energy consumption to infer household characteristics using smart meters. Energies, 2019, 12(5):773.

[3] Viegas J L, Vieira S M, Melício R, et al. Classification of new electricity customers based on surveys and smart metering data. Energy, 2016, 107:804-817.

[4] Zhong S, Tam K S. Hierarchical classification of load profiles based on their characteristic attributes in frequency domain. IEEE Transactions on Power Systems, 2015, 30(5):2434–2441.

[5] Gajowniczek K, Ząbkowski T, Sodenkamp M. Revealing household characteristics from electricity meter data with grade analysis and machine learning algorithms. Applied sciences, 2018, 8(9):1654.

[6] Hopf K, Sodenkamp M, Kozlovkiy I, et al. Feature extraction and filtering for household classification based on smart electricity meter data. Computer Science Research and Development, 2016, 31(3):141–148.

[7] Sun G, Cong Y, Hou D, et al. Joint household characteristic prediction via smart meter data. IEEE Transactions on Smart Grid, 2017, 10(2):1834-1844.

[8] Wang Y, Chen Q, Gan D, et al. Deep learning-based socio-demographic information identification from smart meter data. IEEE Transactions on Smart Grid, 2019, 10(3):2593-2602.

[9] Wang Y, Bennani I, Liu X, et al. Electricity consumer characteristics identification: a federated learning approach. IEEE Transactions on Smart Grid, 2021, Early Access.

[10] Wang F, Li K, Duić N, et al. Association rule mining based quantitative analysis approach of household characteristics impacts on residential electricity consumption patterns. Energy Conversion and Management, 2018, 171:839-854.

[11] Haben S, Singleton C, Grindrod P. Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data. IEEE Transactions on Smart Grid, 2017, 7(1):136-144.

[12] Kumar P, Banerjee R, Mishra T. A framework for analyzing trade-offs in cost and emissions in power sector. Energy, 2020, 195:116949.

[13] Li K, Cao X, Ge X, et al. Meta-heuristic optimization based two-stage residential load pattern clustering approach considering intra-cluster compactness and inter-cluster separation. IEEE Transactions on Industry Applications, 2020, 56(4):3375-3384.

[14] Wang F, Li K, Liu C, et al. Synchronous Pattern Matching Principle-Based Residential Demand Response Baseline Estimation: Mechanism Analysis and Approach Description. IEEE Transactions on Smart Grid, 2020, 9(6):6972-6985.

[15] Li K, Wang F, Mi Z, et al. Capacity and output power estimation approach of individual behind-the-meter distributed photovoltaic system for demand response baseline estimation. Applied Energy, 2019, 253:113595.

[16] Wang F, Xiang B, Li K, et al. Smart households' aggregated capacity forecasting for load aggregators under incentive-based demand response programs. IEEE Transactions on Industry Applications, 2020, 56(2):1086-1097.

[17] Wang F, Ge X, Yang P, et al. Day-ahead optimal bidding and scheduling strategies for DER aggregator considering responsive uncertainty under real-time pricing. Energy, 2020, 213:118765.

[18] Lu Q, Lü S, Leng Y, et al. Optimal household energy management based on smart residential energy hub considering uncertain behaviors. Energy, 2020, 195:117052.

[19] Finck C, Li R, Zeiler W. Economic model predictive control for demand flexibility of a residential building. Energy, 2019, 176:365-379.

[20] Li K, Mu Q, Wang F, et al. A business model incorporating harmonic control as a value-added service for utility-owned electricity retailers. IEEE Transactions on Industry Applications, 2019, 55(5):4441-4450.

[21] Li K, Liu L, Wang F, et al. Impact factors analysis on the probability characterized effects of time of use demand response tariffs using association rule mining method. Energy Conversion and Management, 2019, 197:111891.

[22] Lin J, Marshall K R, Kabaca S, et al. Energy affordability in practice: Oracle Utilities Opower's business Intelligence to meet low and moderate income need at Eversource. The Electricity Journal, 2020, 33(2):106687.