Household Profile Identification for Behavioral Demand Response: A Semi-supervised Learning Approach Using Smart Meter Data

Fei Wang[1,2,3], Xiaoxing Lu[1], Xiqiang Chang[4], Xin Cao[5], Siqing Yan[1], Kangping Li[6], Neven Duić[7], Miadreza Shafie-khah[8], João P.S. Catalão[9]

1. Department of Electrical Engineering, North China Electric Power University, Baoding 071003, China

2. State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources (North China Electric Power University), Beijing 102206, China

3. Hebei Key Laboratory of Distributed Energy Storage and Micro-grid, North China Electric Power University, Baoding 071003, China

4. State Grid Xinjiang Electric Power Co., Ltd, Urumqi 830018, China

5. China Suntien Green Energy Corporation Limited, Shijiazhuang 050022, China

6. Department of Electrical Engineering, Tsinghua University, 10084 Beijing, China

7. Department of Energy, Power and Environmental Engineering, Faculty of Mechanical Engineering and Naval Architecture, University of Zagreb, Ivana Lucica 5, HR-10000 Zagreb, Croatia

8. School of Technology and Innovations, University of Vaasa, 65200 Vaasa, Finland

9. Faculty of Engineering of University of Porto and INESC TEC, 4200-465 Porto, Portugal

Abstract—**Accurate household profiles (e.g., house type, number of occupants) identification is the key to the successful implementation of behavioral demand response. Currently, supervised learning methods are widely adopted to identify household profiles using smart meter data. Such methods could achieve promising performance in the case of sufficient labeled data but show low accuracy if labeled data is insufficient or even unavailable. However, the acquisition of accurately labeled data (usually obtained by survey) is very difficult, costly, and time-consuming in practice due to various reasons such as privacy concerns. To this end, a semi-supervised learning approach is proposed in this paper to address the above issues. Firstly, 78 preliminary features reflecting the household profiles information are extracted from both time and frequency domain. Secondly, feature selection methods are introduced to select more relevant ones as the input of the identification model from the preliminary features. Thirdly, a transductive support vector machine method is adopted to learn the mapping relation between the input features and the output household profile identification results. Case study on an Irish dataset indicates that the proposed approach outperforms supervised learning methods when only limited labeled data is available. Furthermore, the impacts of different feature selection methods (i.e., Filter, Wrapper and Embedding methods) are also investigated, among which the wrapper method performs best, and the identification accuracy improves with the increase of data resolution.**

Keywords—Behavioral demand response; Household profile; Smart meter data; Semi-supervised learning; Feature selection

Acronyms

ACC	Accuracy
ANN	Artificial neural networks
AUC	Area under the curve of ROC
BDR	Behavioral demand response
CNN	Convolutional neural network
DR	Demand response
DT	Decision tree
DWT	Discrete wavelet transform
FPR	False positive rate
KNN	K-nearest neighbor
LDA	Linear discriminant analysis
LR	Logistic regression
MLP	Multi-layer perception
RF	Random forest
RFE	Recursive feature elimination
ROC	Receiver operating characteristic
SVM	Support vector machine
TOU	Time of use
TPR	True positive rate
TSVM	Transductive support vector machine
WD	Wavelet decomposition

1. Introduction

1.1 Background and motivation

Demand response (DR) realizes the efficient utilization of massive flexible demand-side resources such as electric vehicles, distributed energy storages, elastic load [1,2], etc., which could achieve a similar regulation effect with the supply side in a more economical, fast and environmental-friendly way [3]. Therefore, DR has been widely recognized as an effective technique to maintain the reliability and improve the flexibility of the power system [4,5]. To meet the requirement of power grids, traditional DR programs encourage customers to change their normal load patterns utilizing dynamic pricing [6] or financial incentives [7]. In recent years, an emerging kind of DR named behavioral demand response (BDR) is receiving increasing attention. It takes advantages of social & behavioral science rather than economic signals as incentive to motivate customers' reduction in electricity usage during the peak event period. The practice of a U.S. company named Opower has proved BDR to be effective in peak load reduction [8]. Specifically, Opower provides energy-saving suggestions and sends personalized home energy reports to its customers, in which the energy consumption information of similar neighborhoods (with similar household profiles such as similar home types and number of occupants) are presented [9]. Such a report provides customers with a more intuitive and comprehensive perception of household energy usage compared to its similar neighborhoods, which would greatly stimulate their energy-saving action voluntarily. Relying on this technology, Opower has helped households save 25 billion kWh of electricity by June 22, 2020

[10], which equals the power to top off the charge for 2.2 billion smartphones and the energy gap of flipping 6.7 million classic lightbulbs to LEDs.

The information of household profiles is the basis for finding the "similar neighborhoods" in BDR. However, such household profile information is usually unavailable. How to accurately identify household profiles is very important for the implementation of BDR, which is the motivation of this paper. Formally, a household profile is composed of demographic, geographic, psychographic characteristics, purchase history as well as other personalized features of customers [11]. In this paper, electricity consumption behavior related household profiles are concentrated on, which mainly covers four categories, including *dwelling characteristics* (e.g., type of house, age of house), *socio-demographic* (e.g., employment situation of the householder, have children or not), *appliances and heating* (e.g., house heating, type of cooking) and *attitudes towards energy* (e.g., the willingness to reduce energy use, energy-saving efforts) [12]. The diagram of the household profiles is presented in Fig.1.

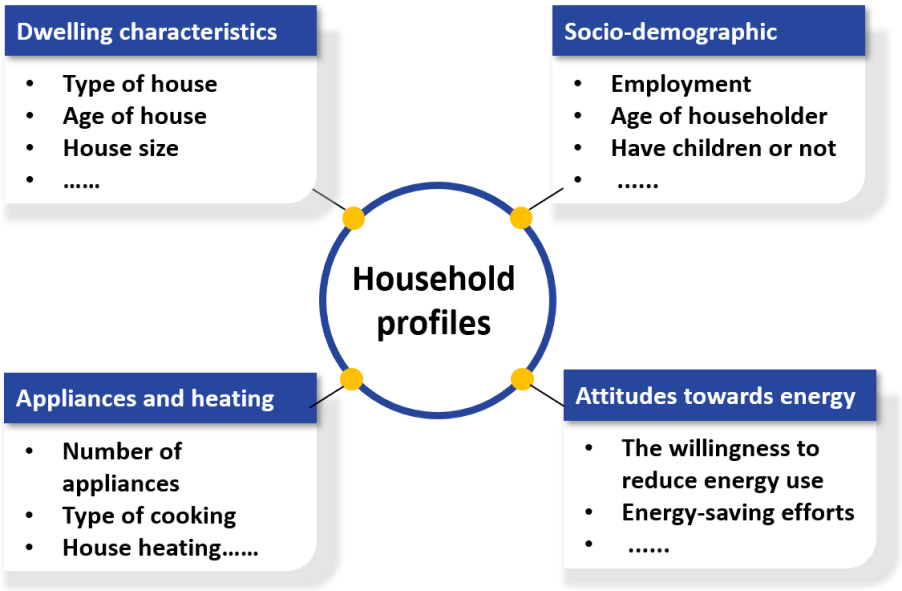


Fig. 1. The diagram of the household profiles

In the past years, the household profile information was collected through site-visiting, questionnaires, etc., which can hardly guarantee the authenticity of the collected information, and further causing the waste of massive manpower, material, and financial resources. Fortunately, smart meters have rapidly gained popularity around the world in recent years [13]. For example, it is reported by the U.S. Energy Information Administration (EIA), U.S. electric utilities had about 94.8 million Advanced Metering Infrastructure (AMI) installations by 2019, about 88% of the AMI installations were residential customer installations [14]. In the U.K., over 21 million smart meters were installed by the end of 2020 [15]. The widespread installation of smart meters enables the collection of fine-grained residential electricity consumption data [16], which contains abundant information of household profiles [17], thus making the identification of household profiles from smart meter data possible [18].

1.2 Literature review

In recent years, a variety of methods have been proposed to identify household profiles from smart meter data, which can be classified into three categories: single-, hybrid- and ensemble classification methods, as is summarized in Table1.

Single methods refer to the case where a single machine learning-based classification model is employed to identify the household profiles. For example, a Random Forest (RF) classification model is developed by Muhammad et al. [19] to identify

household profiles and is compared to other single methods including Decision Tree (DT), K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) classifiers. Viegas et al. [20] proposed a DT method to estimate the profile labels of new consumers. Zhong et al. [21] extracted the features from the frequency domain by using the Discrete Fourier Transform and used a Classification and Regression Tree method to classify the consumers into different groups. Gajowniczek et al. [22] used Artificial Neural Networks (ANN), KNN and SVM to identify household profiles. In our previous work [23], SVM was introduced to identify the specific household profiles. Then RF, KNN and Multi-layer perception (MLP) classifiers were adopted as benchmark methods to compare the identification performance with the SVM classifier.

Table 1 Summary of existing household profile identification methods

Category	Ref.	Identification methods
Single classification method	[19]	RF, DT KNN and SVM
	[20]	DT
	[21]	Classification and Regression Tree
	[22]	ANN KNN and SVM
	[23]	SVM, RF, KNN and MLP
Hybrid classification method	[24]	eCLASS (hybrid KNN and SVM)
	[25]	Hybrid multi-task supervised learning model
	[26]	Joint CNN-SVM model
	[27]	Federated ANN model
Ensemble classification method	[28], [29]	AdaBoost

Hybrid methods usually combine several single classification models together to identify household profiles. For example, Hopf et al. [24] extracted a set of features using an extended CLASS (hybrid KNN and SVM) system to infer the household profiles. The results indicate that the identification accuracy can be improved by the proposed hybrid model. Sun et al. [25] considered the joint analysis of different profiles to improve the generalization performance and predicted multiple household profiles simultaneously by a hybrid multi-task supervised learning model. Wang et al. [26] proposed a two-dimensional Convolutional Neural Network (CNN)-SVM model to automatically extract features from smart meter data and identify the household profiles. Recently, Wang et al. [27] proposed a new idea of using a federated learning approach for household profile identification.

Ensemble methods combine multiple weak classifiers into a strong classifier to improve the household profile identification performance. For example, Albert et al. [28] utilized the AdaBoost classifier to train the features extracted from the time series data smart meter to identify the specific appliances and household occupancy characteristics. Similarly, Beckel et al. [29] presented a taxonomy of 22 features for smart meter data and applied the AdaBoost classifier to further estimate more specific household profiles.

In summary, the above household profile identification methods are all supervised learning-based methods. Such methods can achieve promising performances when sufficient labeled training data is available, but show low accuracy if labeled samples are insufficient or even unavailable. However, the acquisition of labeled samples (i.e., usually obtained by survey) is usually difficult, costly and time-consuming in practice due to various reasons such as privacy concerns. How to reduce the labeled cost while maintaining the identification accuracy remains a critical technical issue to be addressed.

1.3 Contributions and paper structure

Semi-supervised learning method falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data) methods. It serves as a promising method to identify specific load profile when a small amount of labeled data and enormous unlabeled data is available. Among the semi-supervised learning technique, Semi-supervised Support Vector Machine method [30] is commonly adopted in literature and the most well-known one is the Transductive Support Vector Machine (TSVM) method [31]. It could not only classify the labeled samples well but also fully extracts useful information from the sample distribution covered by the substantial unlabeled samples, thus is expected to produce considerable improvement in identification performance in the case of insufficient labeled training data. In this paper, a semi-supervised learning method based on TSVM is applied to the household profile identification process to achieve performance improvement when the labeled data is insufficient. The main contributions can be summarized as follows:

(1) A TSVM-based semi-supervised learning approach is proposed to identify the households' profiles taking advantage of both labeled and unlabeled smart meter data. The proposed approach can effectively improve the identification accuracy when only limited labelled samples are available and significantly save the cost of sample labeling. It is very useful for the large-scale promotion of BDR.

(2) Numerous features reflecting households' power consumption characteristics are extracted from both time and frequency domain to comprehensively distinguish each customer, among which the more representative ones are chosen as the input of the TSVM identification model by feature selection methods to reduce data redundancy and improve calculation efficiency.

(3) Two case studies are conducted using a real-world dataset to illustrate the effectiveness of the proposed method. Furthermore, the impacts of feature selection methods and data resolution on feature identification performance are explored.

The rest of this paper is organized as follows. Section 2 introduces the proposed semi-supervised learning-based household profile identification approach. Section 3 evaluated the performance of the proposed method through simulation experiments, followed by several potential applications of the proposed approach given in Section 4. Finally, Section 5 highlights the conclusions and future works.

2. Methodology

2.1 Data preprocessing

The dataset is obtained from the Commission for Energy Regulation (CER) in Ireland [32] during the Smart Metering Electricity Customer Behavior Trials (SMECBTs) from July 14, 2009 to December 31, 2010. Over 4000 Irish residential customers participated in the trials with an electricity smart meter installed in their homes and they also completed comprehensively designed questionnaires, which mainly contain four aspects questions about the household's dwelling profile, socio-demographic, appliance and heating, attitudes toward energy.

2.1.1 Smart metering dataset

The smart metering dataset is constituted of electricity consumption data of 4232 residential customers with the interval of 30 min over one and a half year. The whole dataset is divided into two periods: benchmark and testing periods. The benchmark period is from 14th July to 31st December 2009, in which all customers are charged with a fixed electricity price. The testing period is from 1st January to 31st December 2010, in which some customers participate in time of use (TOU) DR program. To eliminate

the impact of TOU tariffs on customers' electricity consumption habits, the data from 14th July to 31st December 2009 is selected for analysis. The dataset is reduced to 2991 customers after removing 1241 customers with spurious and missing data.

2.1.2 Survey dataset

The surveys mainly contain four aspects of information including household's dwelling profile (e.g., dwelling type, year of construction), the socio-demographic data (e.g., employment or not of householder, number of occupants), the appliance and heating, attitudes toward energy. Every record of each customer in the survey dataset can be matched to the corresponding records in the smart meter dataset through the unique ID. In this research, six characteristics that can reflect the situation of the households are conducted in-depth analysis, since the residents' response rates to each question in the questionnaires are different, so the number of samples when predicting each household characteristic is different. The specific sample number will be shown in the case study.

2.2 Framework

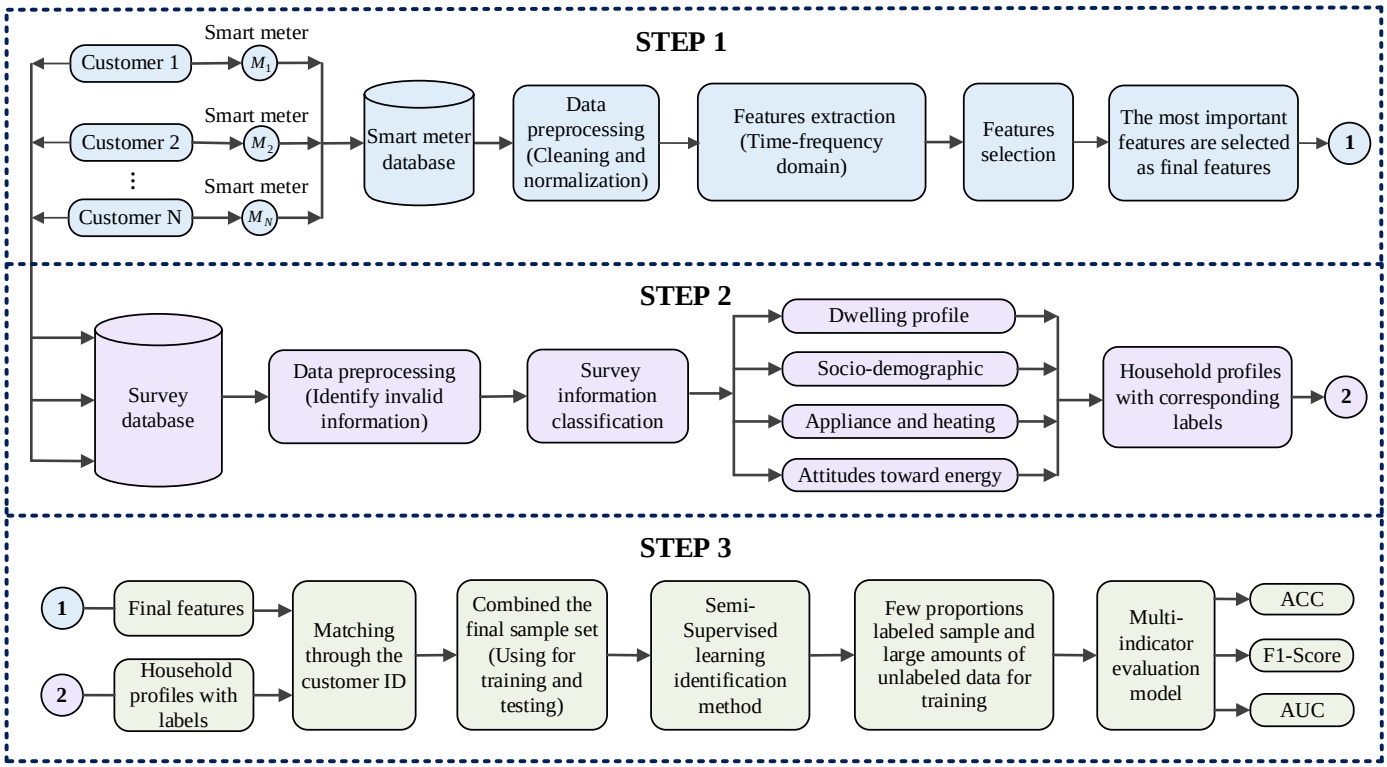


Fig. 2. The Framework of the proposed approach

The framework of the proposed approach is shown in Fig.2. The proposed approach is divided into three steps. Specifically, 54 time-domain features and 24 frequency-domain features are extracted based on the smart meter data. Three feature selection methods (Filter, wrapper and embedding) are used to select the significant features as the final input of the identification model. Meanwhile, the survey dataset is sorted out and effective information is extracted. In this paper, the final household profiles that are most significant to characterize the household situation of the residents are selected and the corresponding labels based on the answers to the questionnaire are calibrated to each household profile. Next, the selected features from smart meter data are matched with the six household profiles in the questionnaire through the consumer ID to form the final training samples. Finally, the

training samples are input into the semi-supervised learning method TSVM. The identification performance is evaluated using three indicators.

2.3 Feature engineering

2.3.1 Feature extraction

(1) Time-domain features extraction

Different households have different power consumption levels during weekdays, weekends and different time periods of one day. Based on these significant differences, 54 time-domain features are extracted to form the input vectors of the identification model. These features are computed using the smart meter data from July 14, 2009 to December 31, 2009. They could be divided into four categories:

1) 28 consumption features (e.g., the *c_total*, *c_max_total*, *c_min_weekend* represent the average, maximum and minimum load power values for all days over almost half-year);

2) 8 ratio features (e.g., *r_mean_max* represents the ratio of average load power values to the maximum load power values);

3) 9 temporal properties (e.g., *t_above_0.5kW_total* represents the proportion of load power records above 0. 5kW over almost half-year);

4) 9 statistical properties (e.g., *s_var_total* represent the variance values of load power for all days over almost half-year).

A large number of time-domain features are extracted to guarantee the integrity of the analysis and then are selected by three types of feature selection methods to reduce the redundancy.

Table 2 The extracted features

No.	Feature name	No.	Feature name	No.	Feature name
	Time-domain features				Frequency-domain features
1	c _total	28	c _min _night	55	CA1_cof_mean
2	c _weekday	29	r_ mean _ max	56	CA1_cof_max
3	c _weekend	30	r_ min _ mean	57	CA1_cof_min
4	c _day	31	r_ forenoon _ noon	58	CA1_cof_var
5	c _morning	32	r_ afternoon _noon	59	CA2_cof_mean
6	c _forenoon	33	r _evening _noon	60	CA2_cof_max
7	c _noon	34	r _noon _total	61	CA2_cof_min
8	c _afternoon	35	r _night _day	62	CA2_cof_var
9	c _evening	36	r _weekday _weekend	63	CA3_cof_mean
10	c _night	37	t_above_0.5kW_total	64	CA3_cof_max
11	c _max _total	38	t_above_0.5kW_weekday	65	CA3_cof_min
12	c _max _weekday	39	t_above_0.5kW_weekend	66	CA3_cof_var
13	c _max _weekend	40	t_above_1kW_total	67	CD1_cof_mean
14	c _max _morning	41	t_above_1kW_weekday	68	CD1_cof_max
15	c _max _forenoon	42	t_above_1kW_weekend	69	CD1_cof_min
16	c _max _noon	43	t_above_2kW_total	70	CD1_cof_var

17	c _max _afternoon	44	t_above_2kW_weekday	71	CD2_cof_mean
18	c _max _evening	45	t_above_2kW_weekend	72	CD2_cof_max
19	c _max _night	46	s _var _total	73	CD2_cof_min
20	c _min _total	47	s _var _weekday	74	CD2_cof_var
21	c _min _weekday	48	s _var _weekend	75	CD3_cof_mean
22	c _min _weekend	49	s _var _morning	76	CD3_cof_max
23	c _min _morning	50	s _var _forenoon	77	CD3_cof_min
24	c _min _forenoon	51	s _var _noon	78	CD3_cof_var
25	c _min _noon	52	s _var _afternoon		
26	c _min _afternoon	53	s _var _evening		
27	c _min _evening	54	s _var _night		

(2) Frequency-domain features extraction

Time-domain features could not depict the difference between fluctuation patterns of household electricity consumption, which presents close correlations with factors such as the number of occupants, the employment of chief income earner, etc. [33]. Hence, in order to capture the load profile's periodical pattern features, Discrete Wavelet Transform (DWT) is introduced to decompose the original smart meter data into several stationary parts (low-frequency signals) and fluctuation parts (high-frequency signals). The decomposed signal has better resolution performance than the original signal, which is conducive to improve the identification performance of the whole model.

DWT is an effective tool for analyzing complex data sequences. Firstly, the initial electricity consumption sequence S is decomposed into two parts: approximate subsequences CA1 and detailed subsequences CD1. Next, the approximate subsequence CA1 would be further decomposed into another two parts named CA2 and CD2 at Wavelet Decomposition (WD) level 2, similarly, the CA2 would be decomposed into CA3 and CD3 till to CAk and CDk at WD k-level. Practically, the three-level DWT is used to decompose the daily average electricity consumption curve of each household into three approximate components (CA1, CA2 and CA3) and three detailed components (CD1, CD2 and CD3). The average, maximum, minimum and variance features of each decomposed component are calculated by the analysis, a total of 24 features for each household are presented in the frequency-domain features of Table 2.

2.3.2 Feature selection

The purpose of feature selection is to find out the features with more discernibility to the classification results. In this paper, in order to explore the impact of different feature selection methods on the performance of household profile recognition, three types of feature selection methods (filter, wrapper, embedded) are adopted to select the most significant (i.e., most relevant) features respectively.

(1) Filter

Filter feature selection method directly selects the final feature subset according to the relationship between features and target labels, which is unconcerned with the final construction of the identification model. In this work, the variance discrimination method is adopted, and the variance threshold is set to K to exclude the features whose variance is lower than the threshold. Then

the Pearson correlation coefficients of the remaining features with respect to a given label are calculated and the *R* most significant (i.e., most relevant) features are selected.

(2) Wrapper

Different from the filter method, the feature subset selected by the wrapper method not only considers the classification model performance, but also greatly minimizes the redundancy between selected features. The wrapper method usually includes three steps: 1) the possibly best feature subsets are generated using the heuristic optimal search algorithm; 2) the feature subsets generated by the search algorithm are further evaluated through the performance of the classification model; 3) the ranking of the features in the top *R* by the performance evaluation of the classification model are selected as the final inputs of the subsequent model.

In this paper, the Recursive Feature Elimination (RFE) method is adopted using the Logistic Regression (LR)-based model. Firstly, a weight is assigned to each feature in the initial training. Secondly, the LR model is used to predict the classification labels. Then the predicted classification labels are compared with the true labels to get the identification error. The weights of features are updated according to errors constantly and the feature with the smallest absolute value weight is eliminated in each round. Repeat this step until the required number of features is reached. Finally, the features with the top *R* weights are selected as inputs of the classification model.

(3) Embedding

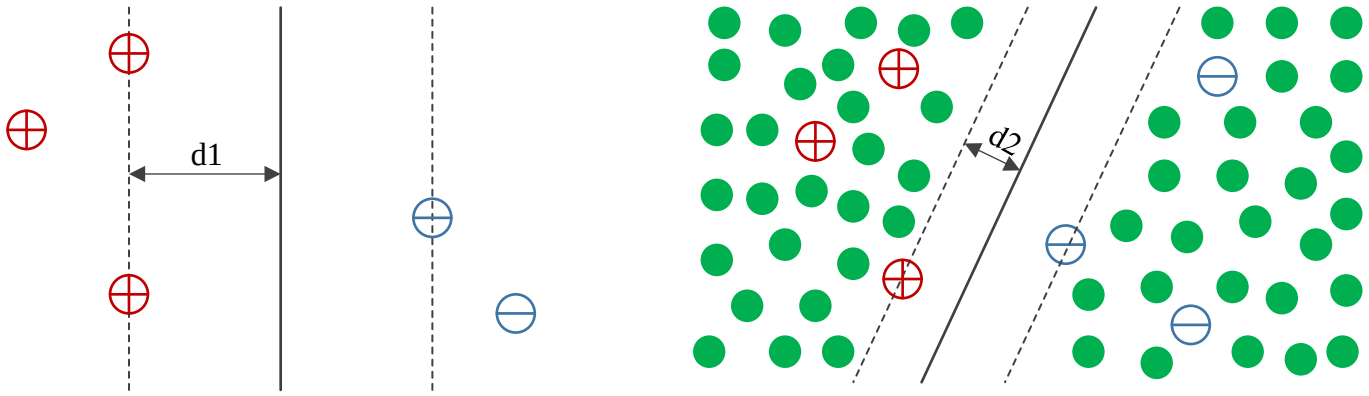
Embedding methods aim to reduce the computation time taken up for reclassifying different subsets which are done in wrapper methods. The main approach is to incorporate feature selection as a part of the training process of the classifier.

In this work, the RF algorithm is used to score the original features and the *R* most significant (i.e. most relevant) features are selected. The bootstrap technology is applied to the smart meter data set, *K* new random self-sample sets with the same size as the original sample are randomly selected and the *K* classification and regression trees are constructed. The unselected samples are marked as *K* out-of-bag samples (*OOB*) at each time. The importance of a feature is measured by the classification accuracy based on the *OOB* samples, which is defined as the average reduction of the classification accuracy before and after slightly perturbing the feature values of *OOB* dataset.

2.4 Transductive support vector machine method

Specifically, given a labeled sample set $D_l = \{(x_1, y_1), (x_2, y_2), ..., (x_l, y_l)\}$, where l represents the number of labeled samples, $y_i \in \{-1, +1\}$. The set of unlabeled samples is $D_u = \{(x_{l+1}), (x_{l+2}), ..., (x_{l+u})\}$, where u represents the number of unlabeled samples, $l \ll u$, $l+u=m$. The labeled and unlabeled data are represented using *L* and *U* respectively. In this paper, the transductive problem of finding the labels for *U* is explored.

The differences between SVM and TSVM are presented in Fig.3. Fig.3 (a) shows a completely labeled dataset. The linear decision boundary is established by SVM, the two dotted lines go through the nearest positive and negative instances. The distance *d1* is called the geometric margin and maximized by the learning of SVM and the labeled samples can be well classified. In Fig.3 (b), the green dots represent massive unlabeled samples. Though the margin *d2* is smaller than *d1*, the classification hyperplane not only classifies the labeled samples well but also makes full use of the sample distribution information covered by the substantial unlabeled samples. Compared to the classification established only by a few labeled samples, its identification performance has been improved.



(a) SVM decision boundary (b) TSVM decision boundary

Fig. 3. (a) With only labeled samples, the linear decision boundary that maximizes the distance to any labeled instance is shown in solid line. Its associated margin is shown in dashed lines. (b) The green dots represent the additional unlabeled samples, under the assumption that the classes are well-separated, the decision boundary seeks a gap in unlabeled sample.

The learning goal of TSVM is to make predictive labels for unlabeled samples in D_u, $\hat{y}=(\hat{y}_{l+1},\hat{y}_{l+2},...,\hat{y}_{l+u})$, $\hat{y}_i \in \{-1,+1\}$. The process of finding the optimal classification hyperplane can be formulated as the following optimization problem.

$$\min_{w,b,\hat{y},\varepsilon} \frac{1}{2}\|w\|_2^2 + C_l \sum_{i=1}^{l} \varepsilon_i + C_u \sum_{i=l+1}^{m} \varepsilon_i \tag{1}$$

$$y_i\left(w^T x_i + b\right) \geq 1-\varepsilon_i, i=1,2,...,l \tag{2}$$

$$\hat{y}_i\left(w^T x_i + b\right) \geq 1-\varepsilon_i, i=l+1,l+2,...,m \tag{3}$$

$$\varepsilon_i \geq 0, i=1,2,...m \tag{4}$$

where C_l and C_u represent the weight of labeled samples and unlabeled samples respectively. w is normal vector, b is displacement term. (w,b) represents the partition hyperplane. ε is tension vector. $\varepsilon_i (i=1,2,...,l)$ corresponds to the labeled sample, $\varepsilon_i (i=l+1,l+2,...,m)$ corresponds to the unlabeled sample.

Firstly, a SVM model is trained using labeled samples, ignoring the terms and constraints of D_u and $\hat{y}$ in the optimization function. The SVM model trained only using the labeled data is then adopted for temporary label assignment to unlabeled samples, and the predicted results of the SVM are given as pseudo-labels to the unlabeled sample. Currently $\hat{y}$ is the known data, and it is brought into the constraint function to obtain the standard SVM model, so the new classification hyperplane and slack vector can be obtained. Secondly, the labels of unlabeled samples at this time are inaccurate, so set the value C_u smaller than C_l that the labeled samples take up more weight. Next, if a pair of positive and negative samples near the classification boundary could be found such that an exchange of their temporary labels decreases the objective function value in Eq. (2)-(4), their classification labels will be exchanged and the classifier retraining. Repeat this step continually to adjust the label assignments. Gradually increase the impact of unlabeled samples on the optimization goal until $C_u = C_l$. Finally, the optimal classification hyperplane is solved and the identification model is obtained. The pseudo-code of the TSVM algorithm is shown in Appendix A.

3 Case study

In this section, the performance of the proposed method is tested and evaluated through simulation experiments, the simulation results are analyzed and the factors that may affect the performance of the proposed approach are discussed.

3.1 Experimental setup

According to the survey dataset and previous study, six household profiles (1#Employment, 2#Residents, 3#House-type, 4#Occupancy, 5#Cooking-type and 6# Children) are finally selected for identification. The No. of output, names, descriptions, label definitions and sample quantities are shown in Table 4.

Table 4 The household profiles description and classification definition

No. of profile	Profile name	Description	Classes	Label	No. of samples
1	Employment	Employment of the chief income earner	Employed	1	1423
			Not employed	2	1026
2	Residents	Number of residents	≤2	1	1321
			>2	2	1128
3	House-type	Type of house	detached or bungalow	1	1299
			semi-detached or terraced	2	1104
4	Occupancy	The Occupancy time more than 6h per day	Yes	1	1619
			No	2	345
5	Cooking-type	Type of cooking facility	Electrical	1	1712
			Not electrical	2	737
6	Children	Have children or not	Yes	1	1964
			No	2	485

3.2 Performance evaluation

3.2.1 Accuracy (ACC)

For a classification problem with Q classes, a $Q \times Q$ confusion matrix B can be obtained, where $B_{q,n}$ denotes the number of samples of class q classified into class n. If $q = n$, then $B_{q,n}$ denotes the number of samples that are correctly classified, and vice versa. Thus, the ACC can be calculated as Eq. (5).

$$ACC = \frac{\sum_{q=1}^{Q} B_{q,q}}{\sum_{q=1}^{Q}\sum_{n=1}^{Q} B_{q,n}} \tag{5}$$

3.2.2 F1-score

For a binary classification problem, the sample labels obtained by the classifiers are compared with the real sample labels, then a confusion matrix can be established, the structure is shown in Table 3.

Table 3 Binary classification confusion matrix

	True Positive	True Negative
Predicted Positive	*TP*	*FP*
Predicted Negative	*FN*	*TN*

TP, FN, FP, and TN represent the number of samples that are correctly predicted as positive, incorrectly predicted as negative, incorrectly predicted as positive, and correctly predicted as negative. Based on these four indices, the F1-score can be defined by Eq. (6) to evaluate the performance on the imbalanced label dataset.

$$F1\text{-}Score = \frac{2 \times Presicion \times Recall}{Presicion + Recall} \tag{6}$$

Among them:

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

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

F1-Score is a comprehensive evaluation metric that reflects the precision and recall of the identification models. The value of F1-Score is between 0 and 1. The closer F1-Score is to 1, the better the identification model is.

3.2.3 Area under the receiver operating characteristic (AUC)

The performance of the identification model can also be evaluated by the Receiver Operating Characteristic (ROC) curves, it is a comprehensive indicator to reflect sensitivity and specificity characteristics. The horizontal axis of the ROC curve is *FPR = FP / (FP + TN)*, the vertical axis of the ROC curve is *TPR = TP* / (*TP + FN*).

The area under the ROC curve is used as an indicator to evaluate the performance of the classifier. The meaning of AUC is:

1) AUC is a value between 0 and 1;

2) The larger the AUC value, the higher the correct rate and the better the classifier performance.

3.3 Results

In this section, two cases are performed to verify the effectiveness of the proposed approach. In each case, four well-known supervised learning methods (i.e., KNN, RF, SVM and MLP) are used as benchmarks for performance comparison. To make the result more reliable, 10-fold cross validation is conducted for each method.

3.3.1 Case1

Since the labeled data is difficult to obtain in practice, thus in case 1, the number of available labeled samples is assumed to be 5% of the total number of samples. That is, only 5% labeled samples could be utilized to train the semi-supervised learning model and supervised learning models respectively and the identification results of household profiles are compared. In this case, 20 most significant features are selected for feature selection using the wrapper method, after that, the values of ACC, F1-Score and AUC are calculated for each household profile to evaluate the performance. The obtained results are compared in Table 5.

Table 5 The comparison results between the semi-supervised learning and supervised learning method (The same number of labeled samples)

Household profiles	Evaluation indexes	SVM	KNN	RF	MLP	TSVM	Improvement
1#Employment	ACC	0.654	0.625	0.68	0.647	**0.700**[a]	**2.94%**[b]
	F1-Score	0.533	0.457	0.575	0.582	**0.664**	**14.09%**
	AUC	0.696	0.554	0.626	0.64	**0.735**	**5.60%**
2#Residents	ACC	0.745	0.72	0.745	0.748	**0.752**	**0.53%**
	F1-Score	0.759	0.764	0.766	0.682	**0.784**	**2.34%**
	AUC	0.805	0.724	0.714	0.659	**0.823**	**2.24%**
3#House-type	ACC	0.598	0.586	0.591	0.617	**0.625**	**1.30%**
	F1-Score	0.602	0.588	0.553	0.558	**0.671**	**11.46%**
	AUC	0.612	0.567	0.564	0.552	**0.668**	**9.15%**
4#Occupancy	ACC	0.826	0.822	0.815	0.82	**0.857**	**3.75%**
	F1-Score	0.905	0.885	0.824	0.855	**0.912**	**0.77%**
	AUC	0.579	0.543	0.537	0.513	**0.681**	**17.62%**
5#Cooking-type	ACC	0.739	0.700	0.721	0.689	**0.764**	**3.38%**
	F1-Score	0.822	0.804	0.779	0.751	**0.834**	**1.46%**
	AUC	0.686	0.689	0.641	0.671	**0.746**	**8.27%**
6#Children	ACC	0.790	0.778	0.805	0.754	**0.819**	**1.74%**
	F1-Score	0.787	0.777	0.786	0.766	**0.821**	**4.32%**
	AUC	0.848	0.782	0.783	0.742	**0.886**	**4.48%**

[a] The best indexes values

[b] The improvements of TSVM over the best performer among the other supervised learning methods

In terms of the basic results of the TSVM model presented in Table 5, among these six household profiles, the ACCs of 2#(Residents), 4#(Occupancy), 5#(Cooking-type) and 6#(Children) are higher than 75%. The ACCs of the remaining household profiles are between 60% and 70%. Note that the ACCs of all the household profiles predicted in this paper are higher than 60%, which verifies the effectiveness of the proposed approach. Clearly, the occupied time and whether having children or not has a great influence on the daily life of consumers and significantly affects the load profiles. The 5#(Cooking-type) directly determines the electricity consumption, it is easy to identify this factor from the smart meter data. Compared to other profiles, the 3#(House-type) has a relatively weak impact on the behavior of residential electricity consumption. The average ACCs, F1-scores and AUCs of these household profiles are 0.753, 0.781 and 0.757, respectively.

The improvement percentage of TSVM over the best performer among the other supervised learning methods are presented in bold. The TSVM has the most significant improvement over the best performer among the other supervised learning methods for the 4#(Occupancy) household profile in terms of the ACC and AUC values, and for the 3#(House-type) in terms of the F1-Scores. For each household profile, it can be clearly observed that the ACCs F1-Scores and AUCs of TSVM are improved obviously compared with other supervised learning methods which due to the TSVM has a stronger ability to capture potential information of a large number of unlabeled smart meter data. Semi-supervised learning method not only considers the class information

covered by the labeled samples, but also makes better use of the distribution regularity of massive unlabeled samples, which demonstrates competitive results of semi-supervised learning methods only with 5% labeled samples.

3.3.2 Case2

In order to further verify the performance of the proposed approach, we compare the identification performance of the semi-supervised learning model with a limited number of labeled samples and the supervised learning models with a large number of labeled samples. The rate of labeled samples used for training is set to 5% and 50% for the proposed approach and the benchmark methods, respectively. Similarly, the wrapper method is chosen for feature selection and the 20 most significant features are selected. The comparison results of ACC, F1-Score and AUC are presented in Table 6.

Table 6 The comparison of identification results between semi-supervised learning and supervised learning

Household profiles	Evaluation indexes	SVM	KNN	RF	MLP	TSVM
1#Employment	ACC	0.688	0.685	0.693	0.699	**0.700**[a]
	F1-Score	0.585	0.575	0.62	0.642	**0.664**
	AUC	0.731	0.667	0.676	0.713	**0.735**
2#Residents	ACC	0.750	0.740	0.755	**0.770**	0.752
	F1-Score	0.766	0.755	0.778	0.782	**0.784**
	AUC	**0.829**	0.749	0.751	0.748	0.823
3#House-type	ACC	0.603	0.601	0.616	0.620	**0.625**
	F1-Score	0.662	0.603	0.65	0.667	**0.671**
	AUC	0.642	0.561	0.599	0.650	**0.668**
4#Occupancy	ACC	0.841	0.833	0.839	0.851	**0.857**
	F1-Score	0.909	0.894	0.887	0.892	**0.912**
	AUC	0.677	0.526	0.54	0.599	**0.681**
5#Cooking-type	ACC	0.743	0.713	0.763	0.753	**0.764**
	F1-Score	0.824	0.821	0.825	0.826	**0.834**
	AUC	0.744	0.621	0.621	0.671	**0.746**
6#Children	ACC	0.817	0.803	0.813	0.814	**0.819**
	F1-Score	0.809	0.802	0.818	0.809	**0.821**
	AUC	0.870	0.796	0.814	0.812	**0.886**

[a] The best indexes values are displayed in bold

The results in Table 6 presents that except for the profile of 2#(Residents), the TSVM shows better identification accuracy than other supervised learning methods for other profiles. For 2#(Residents), the ACC and AUC values of TSVM with 5% labeled samples are slightly lower than the supervised learning model with 50% labeled samples, but the difference is not significant. All the classifiers show good performance for 2#(Residents). In terms of F1 score values, for each household profile, the semi-supervised learning method outperforms the other four supervised learning methods. To sum up, only 10 times fewer labeled samples are needed for the semi-supervised learning method to achieve similar or even better identification accuracy than

supervised learning methods, which means the proposed method can maintain the identification accuracy while significantly reduce the cost of sample labeling.

3.4 Discussion

The impact of two factors on the identification performance of the proposed approach is explored in this section. These two factors are: 1) feature selection methods; 2) the resolution of smart meter data.

3.4.1 The impact of the feature selection method

To investigate the impact of feature selection methods on the performance of the proposed TSVM approach, three feature selection methods (filter, wrapper and embedding) are adopted to select the 20 most significant features respectively from the original feature set. The ACC values of identification results are calculated for each feature selection method. The proportion of labeled samples is set to 5%. The comparison results are shown in Fig. 4.

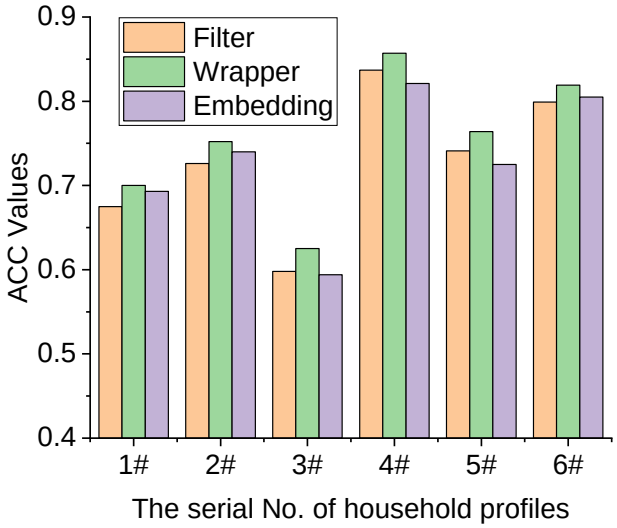


Fig. 4. The comparison results of ACC values using different feature selection methods.

It can be seen from Fig.4 that the wrapper method presents better performance than other two feature selection methods for all household profiles, while filter and embedding methods show inferior performance. The reasons can be explained as follows. Filter methods select features via univariate statistics, the feature selection process is separately performed with the learning model construction and the combination effect of features are not considered. Differently, wrapper methods evaluate all possible combinations of the features and select the combination that produces the best result for a specific machine learning algorithm. Therefore, wrapper methods could typically achieve better performance than filter methods, but at the same time it would lead to more heavy computational burden. Similar to wrapper methods, embedding methods complete the feature selection process within the construction of machine learning algorithm itself. In this case, we find that wrapper methods are more suitable for household profile identification.

3.4.2 The impact of the smart meter data resolution

In order to explore the impact of smart meter data resolution on the performance of the proposed approach, the data resolution is set to change from 0.5h to 2h with the interval of 0.5h. The rate of labeled samples is set to 10% and the wrapper method is chosen for feature selection. The ACC values for different smart meter data resolutions are shown in Fig. 5.

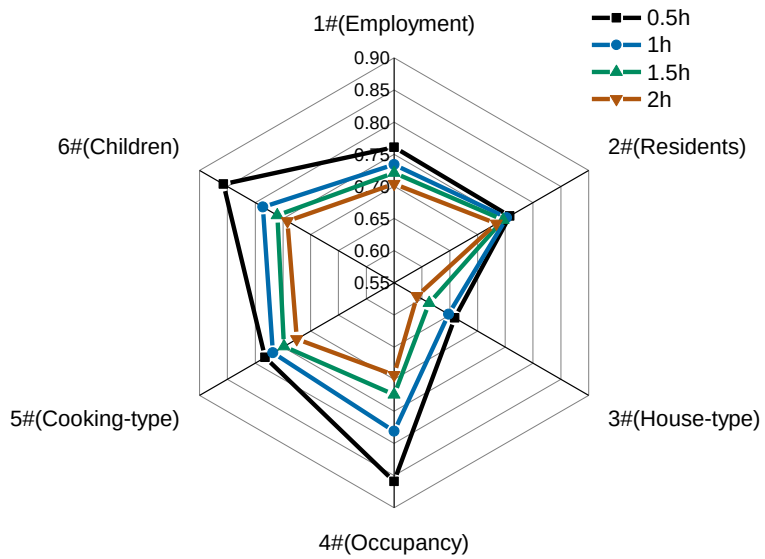


Fig. 5. The comparison of ACC values under different data resolutions

It can be observed that for all household profiles to be identified, the ACC achieves the highest value when the data resolution is 0.5h. The ACC values decrease gradually with the increase of the data acquisition interval (from 1h, 1.5h to 2h). Namely, enhancing data resolution can improve the identification accuracy. This is because that more fine-grained information can be collected with the increase of data resolution, which makes the information used to identify the household profiles richer, thus improving the identification accuracy.

At the same time, it is also observed that different household profiles show different sensitivities to the smart meter data resolution. For the 1#(Employment), 2#(Residents), 3#(House-type) and 5#(Cooking-type), the proposed approach exhibits similar ACC values under different data resolutions, which indicates these profiles are less sensitive to the data resolution. While the data resolution has a greater impact on 4#(Occupancy) and 6#(Children). With the increase of data resolution, the identification accuracy increases faster. This phenomenon indicates that these two household profiles are more sensitive to the smart meter data resolution.

4. Potential applications

In addition to the application in BDR, household profile identification results can also benefit multi-stakeholders including utilities, DR aggregators, households, and policymakers. Several potential applications are discussed as follows.

For utilities, knowing the household profiles of customers is beneficial for them to optimize energy efficiency programs. For example, it is reported from previous studies that some household profiles (e.g., dwelling types, number of occupants, energy consumption attitudes) show strong correlations with the peak load reduction in TOU program [34]. Therefore, with the knowledge of household profiles, suitable customers who have the potential for peak load shaving in TOU program can be found [35]. The targeted services such as new tailor-made tariff schemes can be provided to match customers' specific lifestyles [36]. Additionally, incorporating the household profiles into load forecasting/estimation model could help to improve the accuracy of load forecasting since it is quite important for load dispatching, unit commitment, maintenance planning and energy exchange decisions.

For DR aggregators, household profiles information enables aggregators to know their customers better [37], which can serve as additional informative features to improve clustering, forecasting and optimization tasks in aggregators' daily business, such

as load pattern clustering [38], baseline load estimation [39, 40], DR capacity forecasting [41], optimal bidding and scheduling strategies [42-45], thus enhancing the market competitiveness.

For households, they could receive increasing benefits from the tailored services provides by different companies, who could grasp deep insight into the households' characteristic through the profile identification method [46]. For example, the social interaction-based electricity reduction program can be provided to specific customer groups showing similar household profiles, which can make efficiency-related topics more interesting and therefore increase customer engagement. These services will not only effectively improve residential customers' self-consciousness of energy-saving, get rid of bad electricity consumption habits and reduce customers' electricity bills [47] but also can reduce the carbon emission [48].

For policymakers, knowing the household profiles helps them better understand the energy consumption habits [49], so as to better recognize where do the impacts of electricity consumption come from and then figure out how to formulate new policies to influence and guide (through legislative bans or financial incentives or disincentives) people into desired paths of using electricity more efficiently and friendly [50].

5. Conclusions

In this paper, a semi-supervised learning-based approach is proposed to identify the household profiles from smart meter data. 78 features are extracted from both time and frequency domain. Three feature selection methods are used to select the most relevant features. Case studies using the real-world data from Ireland are performed to verify the effectiveness of the proposed approach. Finally, the impacts of two factors on the performance of the proposed approach are also explored. The main findings are listed as follows:

(1) The proposed semi-supervised learning approach significantly outperforms the well-known supervised learning methods in the case of limited labeled samples. What's more, when the proportion of labeled samples trained by supervised learning method is set to 10 times of semi-supervised learning method, the results show strong advantages over other supervised learning methods.

(2) In terms of feature selection methods, the wrapper method has more advantages than the filter and embedding methods for all household profiles.

(3) The higher the collection resolution (from 2h, 1.5h, 1h to 0.5h) of smart meter data is, the better the identification performance of the household profiles becomes.

The future works are listed as follows:

(1) Testing the semi-supervised approach on more datasets and exploring the applicability of semi-supervised learning method to identify household profiles on resident consumers in different countries, regions or even cities.

(2) More advanced semi-supervised learning methods such as Generative Adversarial Networks (GAN) will be investigated to further improve the identification performance.

(3) The household profiles are identified separately in this paper. A joint household profile identification model will be investigated in our future work.

Acknowledgment

This work was supported by the National Key R&D Program of China (2018YFE0122200), in part by the China Postdoctoral Science Foundation (2020M680552), in part by the Science & Technology Project of State Grid Hebei Electric Power Co., Ltd (SGHEYX00SCJS2000037).

Appendix A

Algorithm: The pseudo-code of TSVM

Input: Labeled sample set $D_l = \{(x_1, y_1), (x_2, y_2), ..., (x_l, y_l)\}$;

Unlabeled sample set $D_u = \{(x_{l+1}), (x_{l+2}), ..., (x_{l+u})\}$;

The weight of labeled sample set C_l, the weight of unlabeled sample set C_u.

1: A initial SVM model is trained by the $D_l = \{(x_1, y_1), (x_2, y_2), ..., (x_l, y_l)\}$;

2: Classify the unlabeled samples in the set $D_u = \{(x_{l+1}), (x_{l+2}), ..., (x_{l+u})\}$, get pseudo-labels $\hat{y}=(\hat{y}_{l+1}, \hat{y}_{l+2}, ..., \hat{y}_{l+u})$;

3: Initialize $C_u \ll C_l$;

4: **while** $C_u < C_l$ **do**

5: Calculate (w,b), ξ with the known D_l, D_u, $\hat{y}$, C_l, C_u according to the formulas (1), (2), (3), (4);

6: **while** $\exists\{i, j \mid (\hat{y}_i \hat{y}_j < 0) \wedge (\xi_i > 0) \wedge (\xi_i + \xi_j > 2)\}$ **do** // Adjust the labels assignment

7: $\hat{y}_i = -\hat{y}_i$;

8: $\hat{y}_j = -\hat{y}_j$;

9: Calculate (w,b), ξ again;

10: **end while**

11: $C_u = \min\{2C_u, C_l\}$ // Gradually increase the value of C_u;

12: **end while**

Output: The final classification results of unlabeled samples: $\hat{y}=(\hat{y}_{l+1}, \hat{y}_{l+2}, ..., \hat{y}_{l+u})$.

References

[1] Mehrjerdi H, Hemmati, R. Energy and uncertainty management through domestic demand response in the residential building. Energy 2019; 192: 116647. https://doi.org/10.1016/j.energy.2019.116647.

[2] Huang S, Abedinia O. Investigation in economic analysis of microgrids based on renewable energy uncertainty and demand response in the electricity market. Energy 2021;225:120247. https://doi.org/10.1016/j.energy.2021.120247.

[3] Hlalele T, Zhang J, Naidoo R, Bansal R. Multi-objective economic dispatch with residential demand response programme under renewable obligation. Energy 2021; 218: 119473. https://doi.org/10.1016/j.energy.2020.119473.

[4] Wang F, Xu H, Xu T, Li K, Shafie-khah M, Catalão JPS. The values of market-based demand response on improving power system reliability under extreme circumstances. Appl Energy 2017; 193: 220-31. https://doi.org/10.1016/j.apenergy.2017.01.103.

[5] Abedinia O, Bagheri M. Power Distribution Optimization Based on Demand Respond with Improved Multi-Objective Algorithm in Power. Energies 2021;14:2961.

[6] Wang F, Li K, Zhou L, Ren H, Shafie-khah M, Catalão JPS. Daily pattern prediction based classification modeling approach for day-ahead electricity price forecasting. Int J Elect Power Energy Syst 2019; 105: 529–40. https://doi.org/10.1016/j.ijepes.2018.08.039.

[7] Khalili T, Jafari A, Abapour M, Mohammadi-Ivatloo B. Optimal battery technology selection and incentive-based demand response program utilization for reliability improvement of an insular microgrid. Energy 2019; 169: 92–104. https://doi.org/10.1016/j.energy.2018.12.024.

[8] Lin J, Marshall KR, Kabaca S, Frades M, Ware D. Energy affordability in practice: Oracle Utilities Opower's business Intelligence to meet low and moderate income need at Eversource. Electr J 2020; 33(2): 106687. https://doi.org/10.1016/j.tej.2019.106687.

[9] Opower "moments that matter". https://www.oracle.com/a/ocom/docs/industries/utilities/utilities-opower-home-energy-report.pdf. [Accessed Nov. 21, 2020]

[10] Opower Reimagines the Home Energy Report. https://www.oracle.com/corporate/pressrelease/oracle-opower-home-energy-report-062220.html. [Accessed Nov. 21, 2020]

[11] Feedough. https://www.feedough.com/what-is-customer-profiling-meaning-elements-examples/. [Accessed Nov. 21, 2020]

[12] Satre-Meloy A. Investigating structural and occupant drivers of annual residential electricity consumption using regularization in regression models. Energy 2019; 174: 148-68. https://doi.org/10.1016/j.energy.2019.01.157.

[13] Melillo A, Durrer R, Worlitschek J, Schütz P. First results of remote building characterisation based on smart meter measurement data. Energy 2020; 200: 117525. https://doi.org/10.1016/j.energy.2020.117525.

[14] U.S. Energy Information Administration. [Online]. Available: https://www.eia.gov/tools/faqs/faq.php?id=108&t=3. [Accessed Nov. 12, 2020].

[15] Smart Energy GB. [Online]. Available: https://www.smartenergygb.org/en/resources/press-centre/press-releases-folder/beis-installation-figures-march-2021?tab=1&docspage=1. [Accessed Mar. 9, 2021].

[16] Kiguchi Y, HeoY, Weeks M, Choudhary R. Predicting intra-day load profiles under time-of-use tariffs using smart meter data. Energy 2019; 173: 959-70. https://doi.org/10.1016/j.energy.2019.01.037.

[17] Kavousian A, Rajagopal R, Fischer M. Determinants of residential electricity consumption: using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behaviour. Energy 2013; 55: 184–94. https://doi.org/10.1016/j.energy.2013.03.086.

[18] Wang Z, Crawley J, Li F, Lowe R. Sizing of district heating systems based on smart meter data: Quantifying the aggregated domestic energy demand and demand diversity in the UK. Energy 2020; 193: 116780. https://doi.org/10.1016/j.energy.2019.116780.

[19] Muhammad F, Alberto S. Analyzing load profiles of energy consumption to infer household characteristics using smart meters. Energies 2019; 12(5): 773. https://doi.org/10.3390/en12050773.

[20] Viegas JL, Vieira SM, Melício R, Mendes VMF, Sousa JMC. Classification of new electricity customers based on surveys and smart metering data. Energy 2016; 107: 804-17. https://doi.org/10.1016/j.energy.2016.04.065.

[21] Zhong S, Tam KS. Hierarchical classification of load profiles based on their characteristic attributes in frequency domain. IEEE Trans Power Syst 2015; 30(5): 2434–41. https://doi.org/10.1109/TPWRS.2014.2362492.

[22] 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. https://doi.org/10.3390/app8091654.

[23] Yan S, Li K, Wang F, Ge X, Lu X, Chen H, Chang S. Time-frequency features combination-based household characteristics identification approach using smart meter data. IEEE Trans Ind Appl 2020; 56(3): 2251-62. https://doi.org/10.1109/TIA.2020.2981916.

[24] Hopf K, Sodenkamp M, Kozlovkiy I, Staake T. Feature extraction and filtering for household classification based on smart electricity meter data. Comput Sci Res Dev 2016; 31: 141–48. https://doi.org/10.1007/s00450-014-0294-4.

[25] Sun G, Cong Y, Hou D, Fan H, Xu X, Yu H. Joint household characteristic prediction via smart meter data. IEEE Trans Smart Grid 2017; 10(2): 1834-44. https://doi.org/10.1109/TSG.2017.2778428.

[26] Wang Y, Chen Q, Gan D, Yang J, Kirschen DS, Kang C. Deep learning-based socio-demographic information identification from smart meter data. IEEE Trans Smart Grid 2019; 10(3): 2593-602. https://doi.org/10.1109/TSG.2018.2805723.

[27] Wang Y, Bennani I, Liu X, Sun M, Zhou Y. Electricity Consumer Characteristics Identification: A Federated Learning Approach. IEEE Trans Smart Grid 2021; Early Access. https://doi.org/10.1109/TSG.2021.3066577.

[28] Albert A, Rajagopal R. Smart meter driven segmentation: what your consumption says about you. IEEE Trans Power Syst 2013; 28(4): 4019-30. https://doi.org/10.1109/TPWRS.2013.2266122.

[29] Beckel C, Sadamori L, Staake T, Santini S. Revealing household characteristics from smart meter data. Energy 2014; 78: 397–410. https://doi.org/10.1016/j.energy.2014.10.025.

[30] Liu T, Yang Y, Huang GB, Yeo YK, Lin Z. Driver distraction detection using semi-supervised machine learning. IEEE Trans Intell Transp 2016; 17(4): 1108-20. 10.1109/TITS.2015.2496157.

[31] Wang Y, Huang ST. Training TSVM with the proper number of positive samples. Pattern Recogn Lett 2005; 26(14): 2187-94. https://doi.org/10.1016/j.patrec.2005.03.034.

[32] Irish Social Science Data Archive. Data from the Commission for Energy Regulation (CER)-smart metering project. [Online]. Available: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/. [Accessed Dec.10, 2017].

[33] Wang F, Li K, Duić N, Mi Z, Hodge BM, Shafie-khah M, Catalão JPS. Association rule mining based quantitative analysis approach of household characteristics impacts on residential electricity consumption patterns. Energy Convers Manage 2018; 171: 839–54. https://doi.org/10.1016/j.enconman.2018.06.017.

[34] Li K, Liu L, Wang F, Wang T, Duić N, Shafie-khah M, Catalão JPS. Impact factors analysis on the probability characterized effects of time of use demand response tariffs using association rule mining method. Energy Convers Manage 2019; 197: 111891. https://doi.org/10.1016/j.enconman.2019.111891.

[35] Andruszkiewicz J, Lorenc J, Weychan A. Seasonal variability of price elasticity of demand of households using zonal tariffs and its impact on hourly load of the power system. Energy 2020; 196: 117175. https://doi.org/10.1016/j.energy.2020.117175.

[36] Li K, Mu Q, Wang F, Gao Y, Li G, Shafie-khah M, Catalão JPS, Yang Y, Ren J. A business model incorporating harmonic control as a value-added service for utility-owned electricity retailers. IEEE Trans Ind Appl 2019; 55(5): 4441-50. https://doi.org/10.1109/TIA.2019.2922927.

[37] Lu X, Li K, Xu H, Wang F. Fundamentals and business model for resource aggregator of demand response in electricity markets. Energy 2020; 204: 117885. https://doi.org/10.1016/j.energy.2020.117885.

[38] Li K, Cao X, Ge X, Wang F, Lu X, Shi M, Yin R, Mi Z, Chang S. Meta-Heuristic Optimization Based Two-stage Residential Load Pattern Clustering Approach Considering Intra-cluster Compactness and Inter-cluster Separation[J]. IEEE Trans Ind Appl 2020; 56(4): 3375-84. https://doi.org/10.1109/TIA.2020.2984410.

[39] Wang F, Li K, Liu C, Mi Z, Shafie-khah M, Catalão JPS. Synchronous Pattern Matching Principle-Based Residential Demand Response Baseline Estimation: Mechanism Analysis and Approach Description[J]. IEEE Trans Smart Grid 2018; 9(6):6972-85. https://doi.org/10.1109/TSG.2018.2824842.

[40] Li K, Wang F, Mi Z, Fotuhi-Firuzabad M, Duić N, Wang T. Capacity and output power estimation approach of individual behind-the-meter distributed photovoltaic system for demand response baseline estimation. Appl Energy 2019; 253: 113595. https://doi.org/10.1016/j.apenergy.2019.113595.

[41] Wang F, Xiang B, Li K, Ge X, Lu H, Lai J, Dehghanian P. Smart households' aggregated capacity forecasting for load aggregators under incentive-based demand response programs. IEEE Trans Ind Appl 2020; 56(2): 1086-97. https://doi.org/10.1109/TIA.2020.2966426.

[42] Wang F, Ge X, Yang P, Li K, Mi Z, Siano P, Duić N. Day-ahead optimal bidding and scheduling strategies for DER aggregator considering responsive uncertainty under real-time pricing. Energy 2020; 213: 118765. https://doi.org/10.1016/j.energy.2020.118765.

[43] Lu X, Li K, Wang F, Mi Z, Sun R, Wang X, Lai J. Optimal Bidding Strategy of DER Aggregator Considering Dual Uncertainty via Information Gap Decision Theory [J]. IEEE Trans Ind Appl 2021; 57(1): 158-169. https://doi.org/ 10.1109/TIA.2020.3035553.

[44] Lu X, Ge X, Li K, Wang F, Shen H, Tao P, Hu J, Lai J, Zhen Z, Shafie-khah M, Catalão JPS. Optimal Bidding Strategy of Demand Response Aggregator Based on Customers Responsiveness Behaviors Modeling under Different Incentives [J]. IEEE Trans Ind Appl 2021; Early Access. https://doi.org/ 10.1109/TIA.2021.3076139.

[45] Waseem M, Lin Z, Liu S, Zhang Z, Aziz T, Khan D. Fuzzy compromised solution-based novel home appliances scheduling and demand response with optimal dispatch of distributed energy resources. Appl Energy 2021; 290:116761. https://doi.org/10.1016/j.apenergy.2021.116761.

[46] Lu Q, Lü S, Leng Y, Zhang Z. Optimal household energy management based on smart residential energy hub considering uncertain behaviors. Energy 2020; 195: 117052. https://doi.org/10.1016/j.energy.2020.117052.

[47] Finck C, Li R, Zeiler W. Economic model predictive control for demand flexibility of a residential building. Energy 2019; 176: 365-79. https://doi.org/10.1016/j.energy.2019.03.171.

[48] Kumar P, Banerjee R, Mishra T. A framework for analyzing trade-offs in cost and emissions in power sector. Energy 2020; 195: 116949. https://doi.org/10.1016/j.energy.2020.116949.

[49] Newsham GR, Bowker BG. The effect of utility time-varying pricing and load control strategies on residential summer peak electricity use: a review. Energy Policy 2010; 38(7): 3289–96. https://doi.org/10.1016/j.enpol.2010.01.027.

[50] Yu B, Tian Y, Zhang J. A dynamic active energy demand management system for evaluating the effect of policy scheme on household energy consumption behavior. Energy 2015; 91: 491-506. https://doi.org/10.1016/j.energy.2015.07.131.