IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 30, 2019, accepted February 18, 2019, date of publication February 25, 2019, date of current version March 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2901257

Predicting the Energy Consumption of Residential Buildings for Regional Electricity Supply-Side and Demand-Side Management

HUILING CAI[1], SHOUPENG SHEN[1,2], QINGCHENG LIN[1], XUEFENG LI[1], (Member, IEEE), AND HUI XIAO[1]

[1] Department of Control Science and Engineering, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China
[2] Hucheng Information Technology Ltd., Company, Shanghai 200050, China

Corresponding authors: Xuefeng Li (lixuefeng@tongji.edu.cn) and Hui Xiao (xiaohui@tongji.edu.cn)

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 22120180189 and Grant 22120170260.

ABSTRACT Energy consumption predictions for residential buildings play an important role in the energy management and control system, as the supply and demand of energy experience dynamic and seasonal changes. In this paper, monthly electricity consumption ratings are precisely classified based on open data in an entire region, which includes over 16 000 residential buildings. First, data mining techniques are used to discover and summarize the electricity usage patterns hidden in the data. Second, the particle swarm optimization-K-means algorithm is applied to the clustering analysis, and the level of electricity usage is divided by the cluster centers. Finally, an efficient classification model using a support vector machine as the basic optimization framework is proposed, and its feasibility is verified. The results illustrate that the accuracy and F-measure of the new model reach 96.8% and 97.4%, respectively, which vastly exceed those of conventional methods. To the best of our knowledge, the research on predicting the electricity consumption ratings of residential buildings in an entire region has not been publicly released. The method proposed in this paper would assist the power sector in grasping the dynamic behavior of residential electricity for supply and demand management strategies and provide a decision-making reference for the rational allocation of the power supply, which will be valuable in improving the overall power grid quality.

INDEX TERMS Residential buildings, energy consumption prediction, clustering analysis, support vector machine.

I. INTRODUCTION

As a result of the desire to improve living standards and residential comfort, the energy consumption of residential buildings accounts for the second largest proportion of the entire increase in energy consumption [1]–[4]. Energy management and control on the supply-side conduct comprehensive analyses based on electricity usage, weather forecasts and the characteristics of the heating and cooling systems used in the buildings to determine the optimal operation and control scheme. In addition, the demand-side management aims to guide the users' electricity usage in a scientific and reasonable way by adjusting the user loads or users' behavior of electricity consumption through economic subsidies, compulsory legal means or publicity means [5]–[7]. Due to the impact of dynamic real-time changes on both supply and demand sides, it is particularly important to classify and predict the energy consumption of residential buildings from historical data to provide a sufficient decision-making basis for planning power transmission configuration patterns that meet regional characteristics. The energy consumption prediction is of decisive importance for the improvement of the power grid quality and the rational allocation of the power supply, which contributes to the enhancement of life quality and the optimization of energy usage. Thus, there are efforts in related works by researchers around the world that are geared towards improving energy consumption predictions.

Research on data mining in intelligent data analysis technology has been a popular area of interest in recent years [8]–[11]. The analysis of building energy consumption based on data mining has been widely regarded by experts and scholars [12], [13]. At present, the prediction

The associate editor coordinating the review of this manuscript and approving it for publication was Kashif Munir.

method of building energy consumption mainly focuses on the use of linear regression algorithms [14], decision tree algorithms [15], neural network (NN) algorithms [16], [17] and support vector machines (SVMs) [18], [19] to generalize the mapping relationships between the input features and output predictions. In the residential sector, the research reported by Biswas *et al.* [20] showed that the daily energy consumption prediction of the designed house fit better with NN models than with linear regression analyses due to the ability to perform nonlinear analyses. Farzana *et al.* [21] conducted a study to predict the future yearly residential energy demand in urban areas of Chongqing, a city in China and found that an ANN model forecast more acceptably than grey models and regression models, with 97.14% R^2 value according to regression statistics. The test in [22] indicated that the SVM modeling method could predict the annual energy consumption of 59 residential buildings in China with higher accuracy than that produced by back propagation (BP) neural networks, as the SVM method overcame the local best and curse of dimensionality that exists in neural networks and traditional machine learning algorithms. Recently, Son and Kim [23] provided a precise model based on support vector regression (SVR) and fuzzy-rough feature selection with particle swarm optimization (PSO) algorithms for monthly forecast of residential electricity demand using historical data in South Korea, and the mean absolute percent-age error (MAPE) of this model is 2.13 which is smaller than other models.

Genetic algorithms (GA) [24] and particle swarm optimization (PSO) algorithms [25] are powerful approaches for predicting building energy consumption as well. In practice, hybrid models combining GA and PSO with machine learning methods are widely utilized in electricity prediction application. Li and Su [26] used the GA-hierarchical adaptive network to predict the daily air-conditioning consumption with 0.0893 coefficient of variation (CV), which outperformed the BPNNs in terms of accuracy. Afterwards, Li *et al.*'s [27] team developed an improved PSO-ANN model for building electricity consumption predictions, and the CV of this model was 0.0791. Selakov *et al.* [28] coupled the SVM with PSO for predicting the short-term electricity load, and the MAPE(%) of the hybrid method for case 1 was 6.15. The results all showed that these hybrid models improved the performance. In addition, massive researchers have made use of clustering methods (i.e., K-means clustering) to analyze daily and seasonal electricity behaviors for load classification. Three cluster methods were investigated by McLoughlin *et al.* [29] in order to segment the households into clusters according to the electricity usage pattern across the day.

Until now, research on the energy consumption of residential buildings has not been elaborated to the same extent as that of commercial buildings due to the insufficiency of the residential energy-use databases and the greater freedom of user behaviors. Moreover, most studies on residential buildings involve short-term energy consumption predictions (i.e., sub-hourly, hourly) of a certain building to help users decrease electricity usage during the peak time of the day to prevent blackouts or involve long-term predictions (i.e., yearly) to identify requirements for national planning and investment [30]. However, there are relatively fewer studies on monthly prediction for electricity energy consumption of a large number of residential buildings for regional electricity supply-side and demand-side management. To our knowledge, the research on quarterly dividing the electricity consumption of residential buildings in the entire region into different levels on the basis of the electricity consumption per unit area has not been publicly released. Moreover, predicting the monthly electricity consumption ratings according to mass data of architectural characteristics apart from weather information and historical energy consumption has not received attention. The lack of a uniform electricity rating prediction for residential buildings within a region can be an obstacle for the promotion of electricity demand-side management and the rational allocation of the power supply, which improves the power grid quality and encourages general users to manage their energy usage scientifically and reasonably. It is difficult for the power sector to use energy simulations for predicting the energy consumption of an entire region due to the unavailable of some detailed data which requires large quantities of various sensors. As such, the use of open data for modeling research and predictions is a good choice. Therefore, the ability to use historical data to predict electricity consumption ratings for regional residential buildings will be valuable.

In this paper, an optimized SVM model based on a combined feature engineering algorithm and a sampling algorithm is introduced for the classification and prediction of electricity usage for residential buildings. A comparative analysis is performed to compare the classification performance of the back propagation (BP) neural network and gradient boosting decision tree (GBDT) with the proposed new method. The precise classification results help control the main supply to guarantee the stable provision of electricity for comfortable living in the entire region, especially during the peak power seasons: summer and winter. The main contributions of this paper are listed as follows:

1) We quarterly divide the electricity consumption of residential buildings in an entire region into different levels on the basis of the electricity consumption per unit area of each residential building. We extract the cluster centers using the improved PSO-K-means method to determine the classification criterion of electricity energy consumption ratings by quarterly division—not annually—according to the different seasons.
2) We design a framework converting the problem of predicting electricity energy consumption into that of predicting electricity energy consumption levels, for classifying and predicting electricity consumption ratings for the next month in an entire region. In feature engineering, apart from weather information data (14 dimensions), electricity consumption data (1 dimension) and natural gas consumption data

(1 dimension), mass data of architectural characteristics (14 dimensions) are used. The one-month-ahead prediction for electricity demand-side situation and distribution of the region are obtained most accurately than other existing studies through the method we proposed.

The organization of this paper is as follows: The next section (II) discusses the methodology. In section (III), the experiment is presented with data, evaluation indices and results analysis. Finally, we conclude the research and discuss future work in section (IV).

II. METHODOLOGY

A. RESEARCH OUTLINE

This paper provides solutions for the energy consumption prediction problem of residential buildings. The research technical route is shown in Figure 1. First, it is important to preprocess the data acquired from the databases, which includes the characteristics of the residential buildings, the weather information and the energy consumption, to remove the noise, outliers and missing values from the data. The sensors used for building energy consumption measurement systems are multi-sourced and asynchronous, and the measurement and control system may encounter network fluctuations or network interruption in the long-time operation process, which results in some abnormal and missing values. After the characterization, the features of all the data types are divided into the same standard. It is necessary to extract the effective information from the data through feature engineering because of the high-dimensional data characteristics of the residential buildings and weather information; this process is followed by the feature dimensionality reduction. Before predicting the electricity energy consumption ratings of residential buildings, this paper conducts a clustering analysis by using the PSO-K-means algorithm on the electricity consumption data of each quarter and then divides the electricity consumption values of each quarter into corresponding levels according to the clustering center points, which is a feature of the classification model. After sampling, the classifier adopts the SVM model, as the inputs are the characteristics of the residential buildings, weather information, and historical energy consumption (i.e., the natural gas and electricity consumption in the previous month). The predicted outputs are the electricity consumption ratings for the next month.

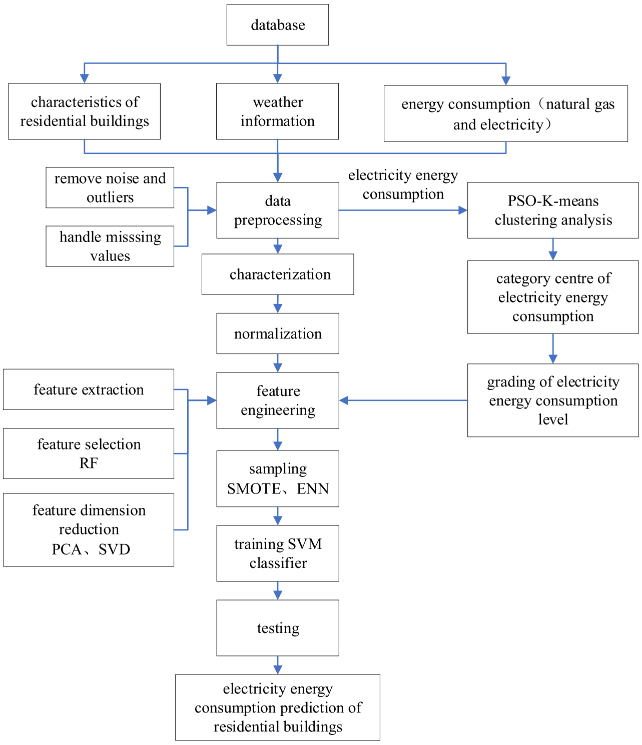

database
characteristics of residential buildings
weather information
energy consumption（natural gas and electricity）
remove noise and outliers
handle misssing values
data preprocessing
electricity energy consumption
PSO-K-means clustering analysis
characterization
category centre of electricity energy consumption
normalization
feature extraction
feature engineering
grading of electricity energy consumption level
feature selection RF
sampling SMOTE、ENN
feature dimension reduction PCA、SVD
training SVM classifier
testing
electricity energy consumption prediction of residential buildings

FIGURE 1. Technical route of the research.

B. COMBINED FEATURE ENGINEERING ALGORITHM

Since the feature dimension is usually high after constructing the basic feature vector, the training process for the conventional classifier is time-consuming, and the irrelevant features that have little or no influence will affect the experimental results. The data we used in this paper are characterized by high-dimensional feature attributes and ample feature information. Therefore, a single feature engineering algorithm is not sufficiently accurate and a feature engineering algorithm combining principal component analysis (PCA) and singular value decomposition (SVD) with random forests (RFs) named RF-PCA-SVD is introduced in this research.

Feature selection is important in feature engineering, which can directly determine the results of model training. Random forests (RFs) are ensemble learning algorithm frameworks that were proposed by Brieman [31] in 2001. In RFs, the decision tree is treated as the minimum unit, and the nodes are randomly selected from the feature space as the split nodes. A bagging algorithm is used to construct a decision tree for multiple training sets. RFs are only applicable to the classification problem as a global feature selection method. RFs adopt the method of quantifying the importance of features to select the features with the largest amount of information.

Assuming that the proportion of k-class samples in the current sample dataset D is p_k $(k = 1, 2, \ldots, m)$,the Gini impurity is expressed by the following formula:

$$Gini(\mathrm{D}) = \sum_{k=1}^{m} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{m} p_k^2 \tag{1}$$

For each feature, the sum of the Gini impurity in the branch nodes formed by the feature in each tree of the RF is counted to evaluate the feature importance. Then, the features that are larger than the threshold are selected.

If the data after feature selection are directly used for model training, problems occur due to the large dimension of the data feature matrix, such as the increase in computation time and training time. Principal component analysis (PCA) is an unsupervised linear transformation technique that is prominent for dimensionality reduction. Highly correlated features are converted to a set of linearly independent characteristic vectors through orthogonal transformation and the transformed variables are called the main components of the original vectors [32]. In data mining, excessive characteristic vectors can bring the curse of dimensionality. The purpose of PCA is to find the dimension with the greatest data difference to help reduce the cost, extract effective information and remove noise.

The approach of PCA is summarized in a few steps:

1. Standardize the features in the dataset D $= \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$ prior to the PCA because it is highly sensitive to data scaling.
2. Construct the covariance matrix $\boldsymbol{XX}^{\mathrm{T}}$.
3. Decompose $\boldsymbol{XX}^{\mathrm{T}}$ into the eigenvalues and eigenvectors.
4. Select d eigenvectors that correspond to the d largest eigenvalues and sort them in a descending order, where d is a lower dimension.
5. Construct the matrix $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_d)$.
6. Obtain the matrix $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}$ after the dimensionality reduction for further use.

Singular value decomposition (SVD) is used to transform a high-dimensional matrix into a low-dimensional matrix, thus achieving feature reduction [33]. SVD removes unwanted singular values and remodels the matrix using valid singular values. SVD is the common solution used to resolve the problem of extracting effective information that describes a matrix that is not square.

Assume that the matrix $\boldsymbol{M}$ can be decomposed as follows:

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{v}^{\mathrm{T}} \tag{2}$$

where $\boldsymbol{U}$ and $\boldsymbol{v}^{\mathrm{T}}$ can be viewed as the rotation operation and $\boldsymbol{\Sigma}$ can be regarded as the zoom operation. The dimensionality reduction can be attained through adjustment to the matrix $\boldsymbol{\Sigma}$. Although the effect of SVD can be gained by PCA dimensionality reduction, SVD has better stability and a wider range of application.

C. PSO-K-MEANS CLUSTERING ANALYSIS

Prior to predicting the electricity energy consumption ratings of the residential buildings, the distribution rule of the data is analyzed with the K-means clustering algorithm. This research divides the electricity energy consumption data into quarters and then extracts the values of three months in each quarter as the eigenvector. For example, the component vector of the first quarter is {electricity energy consumption in January, electricity energy consumption in February, electricity energy consumption in March}.

The K-means clustering algorithm has many characteristics, such as fast convergence and good stability [34]. However, the clustering process is unable to determine the number of clustering centers. This paper introduces the S_DBW fitness value as the evaluation index to determine clustering results. S_DBW not only considers the compactness within the category but also refers to the problem of density between two categories. The smaller the fitness value is, the better the clustering effect will be, which means the intra-cluster connection is closer and the inter-cluster separation is greater.

Let D $= \{V_i | i = 1, 2 \ldots, c\}$ be a partitioning of a dataset S into c convex clusters, where v_i is the center of each cluster. The calculation formula of the average distance from the center to a point of the cluster *Stdev* is shown hereafter [35]:

$$Stdev = \frac{1}{c}\sqrt{\sum_{i=1}^{c} \|\delta_i\|} \tag{3}$$

The validity index $\mathrm{S_{DBW}}$ is defined in the following equation [35]:

$$\begin{aligned} &\mathrm{S_{DBW}} \\ &= \frac{1}{c}\sum_{i=1}^{c} \|\delta(v_i)\| \,/\, \|\delta\,(\mathrm{S})\| \\ &\quad + \frac{1}{c(c-1)} \sum_{i=1}^{c} \sum_{\substack{j=1 \\ i \neq j}}^{c} \frac{\sum_{x \in V_i \cup V_j} f(x, u_{ij})}{\max(\sum_{x \in V_i} f(x, v_i), \sum_{x \in V_j} f(x, v_j))} \end{aligned} \tag{4}$$

where $\delta(v_i)$ is the variance of cluster V_i, $\delta(\mathrm{S})$ is the variance of a dataset, v_i and v_j are the centers of clusters V_i and V_j, respectively, and u_{ij} is the middle point of the line segment defined by the centers of V_i and V_j. If the distance $\mathrm{d}(x, u)$ between center u and point x is larger than the average standard deviation of the clusters *Stdev*, $f(x, u) = 0$; otherwise, $f(x, u) = 1$.

The clustering effect depends on the choice of initial clustering centers. To solve this problem in the paper, an improved particle swarm optimization (PSO) algorithm is used as the previous step of the K-means clustering algorithm, wherein the smallest global fitness points are selected as the initial clustering centers instead of being randomly generated. The improved PSO algorithm continuously searches for the global optimal cluster centers and dynamically adjusts the weighting factor according to the number of iterations to enhance the global search performance [36].

Assume that there are a swarm of m particles moving in the n-dimensional space of the problem solutions. At this point, each particle's own best position $pbest_k$ and global best particle position $gbest$ among all the particles have been found. For each particle k, a position X_k and a flight velocity V_k are adjusted according to the following equations [36]:

$$\begin{aligned} V_k(t+1) = w(t)V_k(t) + c_1 r_1(pbest - X_k) \\ + c_2 r_2(gbest - X_k) \end{aligned} \tag{5}$$

$$X_k(t+1) = X_k(t) + H_0(1 - t/t_{\max})V_k(t+1) \tag{6}$$

The linear adjustment strategy is adopted to dynamically adjust the weight in the formula [36]:

$$w(t) = w_{\max} - (w_{\max} - w_{\min})t/t_{\max} \tag{7}$$

where $w(t)$ is the inertia weight function, c_1 is the cognition weight factor, c_2 is the social weight factor, r_1 and r_2 are two random numbers that are uniformly distributed in the range of [0, 1], $w_{\max}$ and $w_{\min}$ are the initial and final inertia weight factors, respectively, $t_{\max}$ is the maximum iteration number and t is the current iteration number.

In the later iteration stages, the search speed of the optimization algorithm tends to be slow, and the fitness value tends to be stable. In other words, it is described as a precocious particle swarm problem. Accordingly, the research introduces the threshold of fitness variance for the aim of finishing the iteration; this formula is shown as follows [36]:

$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}\left[f(x_i) - f_{avg}\right]^2 \tag{8}$$

where m is the number of particles in the swarm, $f(x_{\mathrm{i}})$ is the fitness value of a single particle, and f_{avg} is the average fitness of the particle swarm.

The improved PSO-K-means algorithm is adopted to conduct cluster analyses on the energy consumption data of residential buildings; the PSO-K-means algorithm makes full use of the global search ability of the improved PSO algorithm and finds the smallest fitness values as clustering centers from the beginning. Thus, the dependence on the initial clustering centers can be avoided. The PSO-K-means algorithm is described in **Algorithm 1**.

D. SAMPLING ALGORITHM

Apart from the SVM proposed as the classification strategy, an imbalanced data classification issue is taken into consideration, which has not received adequate attention in previous works. In this paper, undersampling and oversampling have been carried out for the majority classes and minority classes, respectively, such as the edit nearest neighbor (ENN) algorithm [37] and synthetic minority technique (SMOTE) algorithm [38]. In the actual prediction, the accuracy of minority classes that contain important information determines the generalization ability of the model. It will cost a lot to divide an instance incorrectly.

ENN undersampling looks for adjacent samples around a specific dataset of majority classes and deletes the samples from the original dataset to reduce the imbalance ratio if most of them are different from their own categories, which may have an impact on later experiments. The ENN algorithm is described in **Algorithm 2**.

SMOTE oversampling is used to balance samples by randomly inserting new samples between data points of minority classes. To some extent, SMOTE can avoid overfitting of the classifier and improve the classification ability and prediction accuracy.

In this research, the sampling algorithm integrates ENN with SMOTE to balance the datasets, which are divided into different electricity energy consumption levels. The data in the majority classes are undersampled (a few points are deleted), and the data in the minority classes are oversampled (new data points are added) to improve the prediction results of the subsequent SVM classifier. The SMOTE algorithm is described in **Algorithm 3**.

Algorithm 1 PSO-K-Means Algorithm

```
Input: clustering dataset (residential electricity consumption) $S = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{\mathrm{w}}\}$, number of clustering centers $c$, size of particle swarm $m$, maximum iteration number $t_{\max}$
Output: cluster partition $D = \{V_1, V_2, \ldots, V_{\mathrm{c}}\}$
Process:
1. Iterate over S, find the maximum and minimum of each dimension as the position range $[\boldsymbol{x}_{\min}, \boldsymbol{x}_{\max}]$, wherein the velocity range is $[-\boldsymbol{x}_{\max}, \boldsymbol{x}_{\max}]$.
   Randomly select $c$ initialization centers from S, and then repeat and generate $m$ particle swarms.
   Calculate the fitness $S_{\mathrm{DBW}}$ of each of the particles using (4).
   Initialize $pbest_k$ and $gbest$.
2. for (number of iterations $< t_{\max}$) do
   for ($k = 1, 2, \ldots, m$) do
      Update the velocity and position of the particles according to (5) and (6) and control the velocity and position in the range.
      Dynamically adjust the weight according to (7).
   end for
   for (data point = 1, 2, . . . , w) do
      Divide each data point into the nearest cluster using the Euclidean distance.
      Compute the fitness value $S_{\mathrm{DBW}}$ using (4).
      If $S_{\mathrm{DBW}}$ < the individual extreme value, then update $pbest_k$ and $gbest$.
   end for
   Compute the group fitness variance using (8)
   If $\sigma^2$ > threshold, return 3.
end for
3. Get the best number of clustering centers $c$ and $gbest$.
Execute the K-means algorithm.
4. Select $p_c$ particles as the initial cluster centers with the minimum $S_{\mathrm{DBW}}$ from the PSO algorithm results.
5. for (data point $= 1, 2, \ldots, w$) do
      Divide each data point into the nearest cluster using the Euclidean distance.
   end for
   for (clusters $= 1, 2, \ldots, c$) do
      Update the average value of each cluster and mark the center points.
   end for
   If the center points are not changed, then return the cluster partition $D = \{V_1, V_2, \ldots, V_{\mathrm{c}}\}$.
```

E. SVM METHOD

SVM is a theoretical machine learning classification technique that is widely applied in binary classification and multiclassification [39]–[41]. Owing to the advantages of SVMs in solving non-linear problems, they can be used to predict

Algorithm 2 The ENN Algorithm

```
Input: original dataset T, number of adjacent samples $K$
Output: dataset $T_{ENN}$ after oversampling
Process:
1. $i = 0$, $T_{ENN} = T$.
2. while ($i <$ number of samples) do
      Compare the categories of $K$ adjacent samples around $x_i$ in dataset T and the category of $x_i$, and, if they are different, then delete $x_i$.
   end while
```

Algorithm 3 The SMOTE Algorithm

```
Input: dataset $T_{ENN}$, number of adjacent samples $K$,
       oversampling ratio $n$
Output: dataset $T_{SMOTE}$ after undersampling
Process:
1. $i = 0$
2. while ($i <$ number of samples) do
      Find $K$ adjacent samples near $x_i$, choose $n$ samples $x_{ij}$ for $j = 1, 2, \ldots, n$.
      Compound the new minority class samples $y_j$ according to the equation: $y_j = x_i + (x_i - x_{ij})\mathrm{rand}(0, 1)$.
   end while
3. Add $y_j$ into the dataset $T_{SMOTE}$.
```

residential energy consumption and high accuracy can be achieved in the medium-term and short-term prediction [42]. The SVM method has a simple training process that requires few inputs; however, the low calculation efficiency of SVMs obstructs their use in largescale building energy analyses [42]. Recently, immense efforts have been paid to shorten the calculation time of SVMs by optimizing their structure and by developing hybrid models that are combined with clustering algorithms [43], [44].

When the data are linearly separable, the SVM solves the optimization problem as follows:

$$\begin{cases} \min \ \frac{1}{2}\|\boldsymbol{w}\|^2 \\ s.t. \ y_i(\boldsymbol{w}^{\mathrm{T}}x + \boldsymbol{b}) \geq 1, \quad i = 1, 2, \ldots, n \end{cases} \tag{9}$$

In SVM theory, the Lagrangian multiplier is usually introduced to the objective function, and the latter is easily solved in its dual formulation. The Lagrangian function is formed as follows:

$$L(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i \left(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i + \boldsymbol{b} \right) - 1 \right) \tag{10}$$

where $\boldsymbol{w}$ and $\boldsymbol{b}$ are acquired by the calculation of α, $||\boldsymbol{w}||$ is the Euclidean norm, and α_i $(i = 1, 2, \ldots, n)$ is the Lagrange multiplier. The dual problem is defined as follows:

$$\begin{cases} \max \ L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j) \\ s.t. \ \sum_{i=1}^{n} \alpha_i y_i = 0, \quad \alpha_i \geq 0 \end{cases} \tag{11}$$

Then the dual problem is transformed into the minimax problem of the objective function:

$$\max_{\boldsymbol{\alpha}} \min_{\boldsymbol{w}, \boldsymbol{b}} L(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\alpha}) \tag{12}$$

The solution of the classification hyperplane is shown in the following formula:

$$\begin{cases} \boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i^* \boldsymbol{x}_i y_i \\ \boldsymbol{b}^* = -\frac{1}{2} \boldsymbol{w}^* (\boldsymbol{x}_r + \boldsymbol{x}_s) \end{cases} \tag{13}$$

where α_i^* is the optimal Lagrange multiplier and $\boldsymbol{x}_{\mathrm{r}}$ and $\boldsymbol{x}_{\mathrm{s}}$ are any pair of support vectors in two categories.

To separate the nonlinear data on the plane that is not separable, the SVM completes the calculation in the low-dimensional space, maps the input space to the high-dimensional feature space through a nonlinear change $\varphi(\boldsymbol{x})$, and finally constructs the optimal separation hyperplane. The dot product in (10) is represented by the kernel function $\varphi(\boldsymbol{x}_i)^{\mathrm{T}}\varphi(\boldsymbol{x}_j)$ that is defined as $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. In this paper, the Gaussian kernel shown in the following equation is used:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\frac{\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2}{2\sigma^2}) \tag{14}$$

In this research, the prediction of the residential energy consumption ratings is a multi-classification task, and a multi-classification task can be split into several binary classification tasks. The split strategy takes the form of OvR (One-vs.-Rest). We can train one classifier per class using OvR, where the particular class is regarded as the positive class and all the other samples are considered as the negative class. In the test n classifiers are used and the class label with the highest confidence is assigned to the particular sample.

III. EXPERIMENT

A. DATA AND EVALUATION INDICES

In this paper, the proposed method is tested using data information that consists of the characteristics of residential buildings and the monthly energy consumption of electricity and natural gas; these data come from the website of Open Energy Information (Open EI) for cities in the USA [45]. The experiment is implemented by Python on a Windows 10 operating system with an Intel Core i7 2.7 GHZ processor and 16GB of RAM; the software platform is PyCharm and Anaconda. The original data package contains six Microsoft Excel files, wherein two of the files are the characteristics of

TABLE 1. The dataset for the characteristics of the residential buildings and weather information.

Order	Headers	Order	Headers
1	Year completed	1	Average daily maximum temperature
2	Num floors	2	Average daily minimum temperature
3	Facility type id	3	Average daily temperature
4	Gross floor area	4	Total rainfall
5	Zip code	5	Total snowfall
6	Climate zone id	6	Number of rainy days
7	Heating-sys	7	Number of snowy days
8	Asset-heating-fuel-id	8	Number of sunny days
9	Heating-sys-eff	9	Number of foggy days
10	Heating-sys-eff-flag	10	Number of cloudy days
11	Cooling-sys	11	Number of thunderstorms
12	Asset-cooling-fuel-id	12	Number of snowstorms
13	Cooling-sys-eff	13	Average daily humidity
14	Cooling-sys-eff-flag	14	Average daily air pressure

these residential buildings, which are named 1_ResCharacteristics_DataJam and 2_Headers_DataJam; the headers are listed in Table 1. In addition, the monthly and annual electricity and gas usage of these residential buildings are included in the data package. Monthly weather information is collected according to the zip code of these residential buildings at the National Weather Service website [46], and the headers are listed in Table 1.

The prediction performance is measured by considering three frequently used evaluation indices: accuracy, precision and recall. In addition, the F-measure is used to evaluate the prediction performance of the classification algorithm model. A larger F-measure indicates a better prediction performance, and vice versa [47].

Accuracy measures the proportion of the samples that are predicted correctly. The accuracy is defined as follows [47]:

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{15}$$

where TP is true positive, FN is false negative, FP is false positive and TN is true negative.

Precision represents the proportion of the positive samples that are predicted correctly. The mathematical formula for precision is shown as follows [47]:

$$precision = \frac{TP}{TP + FP} \tag{16}$$

Recall is the proportion of the positive samples that are predicted as positive. The recall is defined as follows [47]:

$$recall = \frac{TP}{TP + FN} \tag{17}$$

The F-measure determines the harmonic average of precision and recall, which makes the evaluation standard more robust by combining two indices. The F-measure is defined as follows [47]:

$$F - measure = \frac{2 \times precision \times recall}{precision + recall} \tag{18}$$

B. CATEGORIZATION OF THE ELECTRICITY ENERGY CONSUMPTION IN RESIDENTIAL BUILDINGS

When the fitness value is at the minimum value, the clustering results with different colors of each quarter are shown in Figure 2. The clustering centers are calculated according to the best fitness value, as shown in Table 2, and the category of the monthly electricity energy consumption of each residential building is divided. For example, the residential buildings which the electricity energy consumption per unit area in the first quarter below 0.287 kWh/m^2 are divided into level 1.

TABLE 2. The cluster centers of electricity energy consumption per unit area in each quarter.

Cluster centres	Centre 1	Centre 2	Centre 3	Centre 4
First quarter	0.287	0.631	1.118	2.034
Second quarter	0.327	0.666	1.185	/
Third quarter	0.447	0.892	1.474	/
Fourth quarter	0.330	0.719	1.404	/

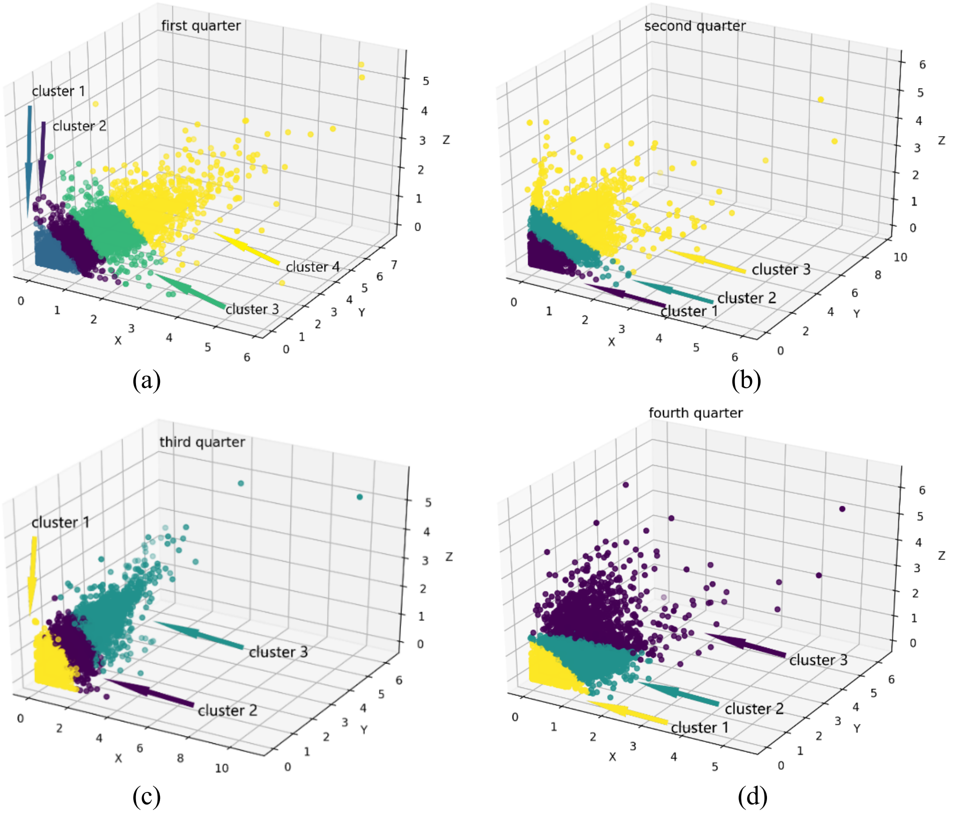
TABLE 3. The statistical number of clusters in each quarter.

Quarter	Level 1	Level 2	Level 3	Level 4	Level 5
First quarter	3901	5582	1256	248	16
Second quarter	4999	7259	3583	719	/
Third quarter	3104	7738	4795	923	/
Fourth quarter	6015	8101	2271	173	/

Afterwards, the number of residential buildings in different levels of each quarter is counted, as shown in Table 3. The first quarter can be divided into five categories from high to low electricity consumption, whereas the other three quarters can be divided into four categories. The statistical results indicate that a class imbalance problem exists and that an imbalance in the distribution of training data will affect the accuracy of the results.

C. UNDERSAMPLING AND OVERSAMPLING

The categorization of the electricity energy consumption of residential buildings is solved by using the clustering algorithm, but, according to the above content, a class imbalance problem emerged. In this paper, the undersampling and oversampling method are combined to address the imbalance problem of the electricity energy consumption data, and the SMOTE-ENN algorithm is used to sample the imbalanced class. The comparison diagrams between the original one and sampling one of the four quarters are shown in Figure 3. To facilitate the display of the post-sampling points, both the original data and the post-sampling data which consist of 14 architectural variables, 14 weather variables and 2 historical energy variables (natural gas consumption values and electricity consumption ratings in the previous month) are reduced to three dimensions by PCA (the coordinate axis has no physical meaning).



(a) (b)

(c) (d)

FIGURE 2. Clustering results of each quarter. Different colors represent different clusters. The x, y and z coordinates represent electricity energy consumption per unit area (the units are given in kWh/m^2) of three months in each quarter. (a) Clustergram of the first quarter; (b) clustergram of the second quarter; (c) clustergram of the third quarter; and (d) clustergram of the fourth quarter.

D. RESULTS AND COMPARISON

The data of Gainesville in Alachua County, Florida in the USA are selected according to the zip codes in the original data set from the website of Open Energy Information (Open EI). Within this region, there are 16560 residential buildings used for prediction in the second, third and fourth quarter; however, there are 11003 residential buildings used for prediction in the first quarter due to the missing values. For predicting the monthly electricity energy consumption ratings, 80% of the processed data set is used for training and 20% is used for test. GBDT, BP and SVM are trained and tested individually with the data to compare their classification performances. The experiment for each model is repeated 20 times, and the average performances and the least performances of accuracy, precision, recall and F-measure are used for comparison. The results in Table 4 illustrate which model performs most efficiently when predicting the electricity energy consumption for residential buildings. As shown in Table 4, the SVM exhibits slightly better performance in almost all evaluation indices compared with GBDT and BP. Moreover, the mean standard deviation of four evaluation indices is around 0.012 using GBDT and BP model, 0.0066 using SVM model and 0.0047 using SMOTE-ENN + SVM model. It is indicated that SVM is better than the other models to predict the electricity consumption ratings of regional residential buildings. As shown in Table 5, the SVM method consumes the longer CPU runtime than GBDP and BP method, that is, the algorithm complexity is high, but it performs the best indicated in Table 4. The reason for the high complexity is to map the low-dimensional data to higher-dimensional data through the function of the kernel function which is invisible in the feature space. Experiences have shown that the higher dimension and the greater algorithm complexity mean that it is more difficult for the training, and it takes more time to consume CPU resources. In the prediction for monthly electricity consumption ratings, we want the prediction accuracy to be as high as possible, so the CPU runtime is sacrificed.

As there are an unequal number of each level in the classification of electricity energy consumption for residential buildings, the problem of imbalance classification arises. The SMOTE-ENN sampling algorithm greatly improves the accuracy of the classification, and the evaluation indices of the SVM method with SMOTE-ENN are much higher than those without the sampling algorithm, which indicates that the imbalanced data have an impact on the experimental results. For instance, in the condition of the least accurate predictions among 20 times, the mean percentages of the

TABLE 4. Classification performance comparison of the GBDT, BP, SVM and SMOTE-ENN + SVM models. The evaluation indices are: accuracy, precision, recall and F-measure.

Classification performance	Evaluation index 1: Accuracy								Evaluation index 2: Precision							
	First quarter		Second quarter		Third quarter		Fourth quarter		First quarter		Second quarter		Third quarter		Fourth quarter	
	Mean	Min	Mean	Min	Mean	Min	Mean	Min	Mean	Min	Mean	Min	Mean	Min	Mean	Min
GBDT	0.799	0.785	0.707	0.693	0.723	0.702	0.757	0.746	0.797	0.782	0.716	0.700	0.724	0.703	0.756	0.747
BP	0.799	0.785	0.707	0.693	0.723	0.703	0.765	0.760	0.81	0.788	0.752	0.740	0.771	0.748	0.776	0.762
SVM	0.822	0.822	0.747	0.735	0.774	0.758	0.777	0.770	0.818	0.818	0.749	0.737	0.775	0.758	0.774	0.765
SMOTE-ENN+SVM	0.934	0.930	0.983	0.979	0.98	0.980	0.975	0.970	0.959	0.951	0.990	0.986	0.989	0.987	0.989	0.982

Classification performance	Evaluation index 3: Recall								Evaluation index 4: F-measure							
	First quarter		Second quarter		Third quarter		Fourth quarter		First quarter		Second quarter		Third quarter		Fourth quarter	
	Mean	Min	Mean	Min	Mean	Min	Mean	Min	Mean	Min	Mean	Min	Mean	Min	Mean	Min
GBDT	0.815	0.792	0.747	0.735	0.774	0.758	0.777	0.770	0.797	0.781	0.703	0.689	0.719	0.698	0.752	0.741
BP	0.816	0.794	0.747	0.735	0.774	0.758	0.77	0.760	0.812	0.791	0.747	0.737	0.768	0.746	0.774	0.761
SVM	0.822	0.822	0.747	0.735	0.774	0.758	0.777	0.770	0.818	0.818	0.746	0.734	0.772	0.756	0.773	0.766
SMOTE-ENN+SVM	0.934	0.930	0.983	0.979	0.978	0.976	0.974	0.968	0.945	0.939	0.987	0.983	0.983	0.981	0.982	0.976

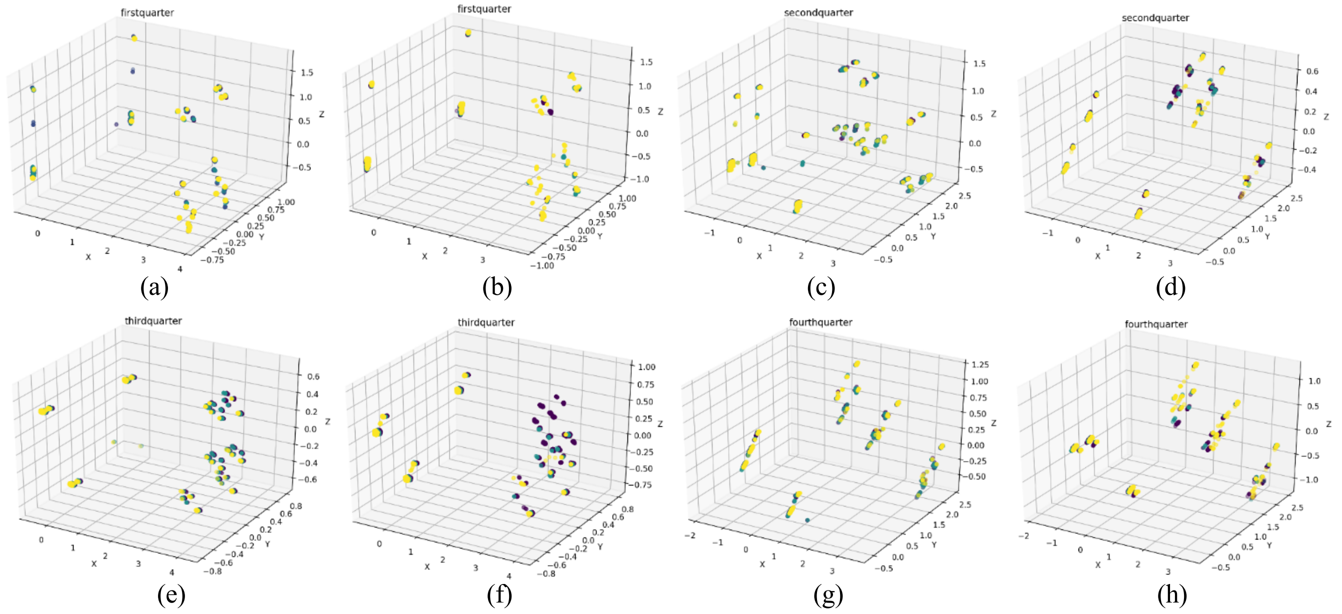


FIGURE 3. Comparison diagrams of the sampling algorithm. The left diagram in the same row is the original one, and the right diagram is the sampling one after processing by the algorithm. To facilitate the display of the post-sampling points, both the original data and the post-sampling data are reduced to three dimensions by PCA (the coordinate axis has no physical meaning). The colors of points in Figure 3 represent three kinds of electricity usage level for demonstration purposes: green color represents level 1, yellow color represents level 2 and purple represents level 3.
(a) and (b) Sampling comparison in the first quarter; (c) and (d) sampling comparison in the second quarter; (e) and (f) sampling comparison in the third quarter; and (g) and (h) sampling comparison in the fourth quarter.

wrong predicted cases of three months in the same quarter are shown in Figure 4. The horizontal axis is the difference between the predicted and actual levels for electricity energy consumption forecast using SMOTE-ENN + SVM method, that is, 1,2, 3, 4 in the first quarter and 1, 2, 3 in the other three quarters. From the histograms, we can see the vast majority of the wrong predictions are one level different from the actual levels. In summary, the average accuracy, precision, recall and F-measure of the new model are up to 96.8%, 98.2%, 96.7% and 97.4%, respectively.

TABLE 5. The comparison of CPU runtime for GBDT, BP, SVM and SMOTE-ENN + SVM models. The unit is given in second (s).

CPU runtime	First quarter	Second quarter	Third quarter	Fourth quarter
GBDT	8.5	10.5	15.8	14.4
BP	2.6	4.1	4.6	2.7
SVM	201.3	798.1	657.2	640.7
SMOTE-ENN+SVM	211.6	813.0	673.1	654.8

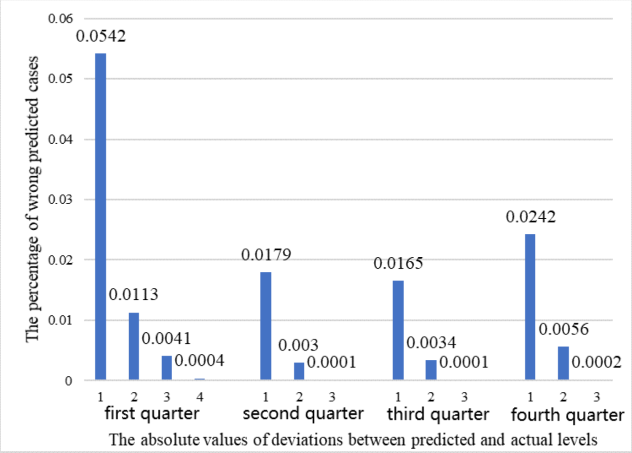


FIGURE 4. The percentages of the wrong predicted cases of four quarters in the condition of the least accurate predictions among 20 times. The horizontal axis is the difference between the predicted and actual levels for electricity energy consumption forecast using SMOTE-ENN + SVM method.

The comparison results demonstrate that the optimized SVM model based on the SMOTE-ENN improves the classification performance by an average of 24.1% in terms of accuracy, 26.03% in terms of precision, 24.01% in terms of recall and 25.35% in terms of F-measure.

For the one-month-ahead forecast of residential electricity demand, one relevant research was done by Son and Kim [23]. They constructed 5 social variables, 14 weather variables and monthly electricity consumption for the proposed model based on SVR and fuzzy-rough feature selection with PSO algorithms. They predicted the monthly electricity energy consumption for the overall South Korea; however, in our study, we predict the monthly electricity energy consumption ratings for each residential building in a city of Florida, USA utilizing the architectural features instead of social variables. As such, we can predict the regional electricity usage distribution of residential buildings monthly, then assist the power sector in providing a decision-making reference for the rational allocation of the power supply. Since the performance in [23] was mainly evaluated by mean absolute percent-age error (MAPE) and root mean squared error (RMSE) which compared the deviation between the predicted value and actual value, the model in [23] couldn't be compared directly with our work.

IV. CONCLUSION AND FURURE WORK

This research provides an optimized SVM model with an accuracy that was increased by using the SMOTE-ENN sampling algorithm to solve the problem of imbalance classification, affording a better understanding of the quarterly electricity energy consumption for residential buildings. First, a combined RF-PCA-SVD feature engineering algorithm is used due to the sparse high-dimensional characteristics of the data. Second, the electricity energy consumption data are analyzed quarterly by the improved PSO-K-means clustering algorithm. Finally, by adopting an SVM and comparing its classification performance with that of the conventional approaches, i.e., GBDT and BP, this paper demonstrates the superiority of SVMs with the sampling algorithm in the monthly prediction of electricity consumption ratings for residential buildings. These findings supply reference opinions for the monthly decision to rationally allocate the power supply in an entire region at macro level. Besides, it can help improve power grid quality to guarantee the stable provision of electricity for comfortable living, especially during the peak power seasons: summer and winter.

The fields of intelligent buildings and smart cities are currently working on promoting the applications of sensor networks to optimize energy utilization. However, medium-term (i.e., monthly) or even long-term (i.e., yearly) residential energy consumption prediction require a relatively larger amount of data that are stably and sufficiently metered with sensors. Besides, uncertainties in the medium-term and long-term prediction is more remarkable than in short-term prediction since many changes may occur in the supply and demand sides over a long period of time. Despite the above challenges, medium-term and long-term energy consumption prediction models are essential. Applying the new model in intelligent control systems can make for accurate regional power configuration construction, and guide changes in residential user behaviors for demand-side management, which can ultimately improve the quality of life.

For future work, we are planning to improve the sampling algorithm for imbalance classification and introduce deep learning methods to the research on the classification and prediction of electricity usage in residential buildings. This will help increase the accuracy and efficiency. Moreover, we will study the electricity energy scheduling strategy according to the prediction for electricity energy consumption ratings of residential buildings based on an entire region.

REFERENCES

[1] T. A. Nguyen and M. Aiello, "Energy intelligent buildings based on user activity: A survey," *Energy Buildings*, vol. 56, pp. 244–257, Jan. 2013. doi: 10.1016/j.enbuild.2012.09.005.

[2] A.-H. Mohsenian-Rad, V. W. S. Wong, J. Jatskevich, R. Schober, and A. Leon-Garcia, "Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid," *IEEE Trans. Smart Grid*, vol. 1, no. 3, pp. 320–331, Dec. 2010. doi: 10.1109/TSG.2010.2089069.

[3] L. G. Swan and V. I. Ugursal, "Modeling of end-use energy consumption in the residential sector: A review of modeling techniques," *Renew. Sustain. Energy Rev.*, vol. 13, no. 8, pp. 1819–1835, Oct. 2009. doi: 10.1016/j.rser.2008.09.033.

[4] M. Zhang and C. Y. Bai, "Exploring the influencing factors and decoupling state of residential energy consumption in Shandong," *J. Clean. Prod.*, vol. 194, pp. 253–262, Sep. 2018. doi: 10.1016/j.jclepro.2018.05.122.

[5] N. Javaid *et al.*, "An intelligent load management system with renewable energy integration for smart homes," *IEEE Access*, vol. 5, pp. 13587–13600, Jun. 2017. doi: 10.1109/ACCESS.2017.2715225.

[6] W.-T. Li *et al.*, "Demand response management for residential smart grid: From theory to practice," *IEEE Access*, vol. 3, pp. 2431–2440, 2015. doi: 10.1109/ACCESS.2015.2503379.

[7] M. H. Albadi and E. F. El-Saadany, "A summary of demand response in electricity markets," *Electr. Power Syst. Res.*, vol. 78, no. 11, pp. 1989–1996, 2008. doi: 10.1016/j.epsr.2008.04.002.

[8] L. Wang and T. YuanFeng, "Application research of data mining technology in power dispatching management system," in *Proc. Int. Conf. Smart Grid Elect. Automat. (ICSGEA)*, Zhangjiajie, China, Aug. 2016, pp. 1–4.

[9] E. Corchado, M. Woźniak, A. Abraham, A. C. P. L. F. de Carvalho, and V. Snášel, "Recent trends in intelligent data analysis," *Neurocomputing*, vol. 126, pp. 1–2, Feb. 2014. doi: 10.1016/j.neucom.2013.07.001.

[10] Y. Zhang and Z. Chuansheng, "A new clustering algorithm based on probability," in *Proc. 1st Euro-China Conf. Intell. Data Anal. Appl.*, Shenzhen, China, Jun. 2014, pp. 119–126.

[11] A. Yassine, S. Singh, and A. Alamri, "Mining human activity patterns from smart home big data for health care applications," *IEEE Access*, vol. 5, pp. 13131–13141, 2017. doi: 10.1109/ACCESS.2017.2719921.

[12] D. Zhao, M. Zhong, X. Zhang, and X. Su, "Energy consumption predicting model of VRV (Variable refrigerant volume) system in office buildings based on data mining," *Energy*, vol. 102, pp. 660–668, May 2016. doi: 10.1016/j.energy.2016.02.134.

[13] C.-F. J. Kuo, C.-H. Lin, and M.-H. Lee, "Analyze the energy consumption characteristics and affecting factors of Taiwan's convenience stores-using the big data mining approach," *Energy Buildings*, vol. 168, pp. 120–136, Jun. 2018. doi: 10.1016/j.enbuild.2018.03.021.

[14] F. S. Westphal and R. Lamberts, "Regression analysis of electric energy consumption of commercial buildings in Brazil," in *Proc. 10th Conf. Int. Buildings Perform. Simulation Assoc.* Beijing, China: Tsinghua Univ, Sep. 2007, pp. 1543–1550.

[15] Z. Yu, F. Haghighat, B. C. M. Fung, and H. Yoshino, "A decision tree method for building energy demand modeling," *Energy Buildings*, vol. 42, no. 10, pp. 1637–1646, Oct. 2010. doi: 10.1016/j.enbuild.2010.04.006.

[16] G. K. F. Tso and K. K. W. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," *Energy*, vol. 32, no. 9, pp. 1761–1768, Sep. 2007. doi: 10.1016/j.energy.2006.11.010.

[17] S. Karatasou, M. Santamouris, and V. Geros, "Modeling and predicting building's energy use with artificial neural networks: Methods and results," *Energy Buildings*, vol. 38, no. 8, pp. 949–958, Aug. 2006. doi: 10.1016/j.enbuild.2005.11.005.

[18] R. K. Jain, K. M. Smith, P. J. Culligan, and J. E. Taylor, "Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy," *Appl. Energy*, vol. 123, pp. 168–178, Jun. 2014. doi: 10.1016/j.apenergy.2014.02.057.

[19] S. Paudel, P. H. Nguyen, W. L. Kling, M. Elmitri, B. Lacarrière, and O. Le Corre, "Support vector machine in prediction of building energy demand using pseudo dynamic approach," in *Proc. 28th Int. Conf. Efficiency, Cost, Optim., Simulation Environ. Impact Energy Syst.*, Pau, France, Jun. 2015.

[20] M. A. R. Biswas, M. D. Robinson, and N. Fumo, "Prediction of residential building energy consumption: A neural network approach," *Energy*, vol. 117, pp. 84–92, Dec. 2016. doi: 10.1016/j.energy.2016.10.066.

[21] S. Farzana, M. Liu, A. Baldwin, and M. U. Hossain, "Multi-model prediction and simulation of residential building energy in urban areas of Chongqing, South West China," *Energy Buildings*, vol. 81, pp. 161–169, Oct. 2014. doi: 10.1016/j.enbuild.2014.06.007.

[22] Q. Li, P. Ren, and Q. Meng, "Prediction model of annual energy consumption of residential buildings," in *Proc. Int. Conf. Adv. Energy Eng. (ICAEE)*, Beijing, China, Jun. 2010, pp. 223–226.

[23] H. Son and C. Kim, "Short-term forecasting of electricity demand for the residential sector using weather and social variables," *Resour., Conservation Recycling*, vol. 123, pp. 200–207, Aug. 2017. doi: 10.1016/j.resconrec.2016.01.016.

[24] H. Sadeghi, M. Zolfaghari, and M. Heydarizade, "Estimation of electricity demand in residential sector using genetic algorithm approach," *Int. J. Ind. Eng. Prod. Res.*, vol. 22, pp. 43–50, Mar. 2011.

[25] H. Nazari, A. Kazemi, M.-H. Hashemi, M. M. Sadat, and M. Nazari, "Evaluating the performance of genetic and particle swarm optimization algorithms to select an appropriate scenario for forecasting energy demand using economic indicators: Residential and commercial sectors of Iran," *Int. J. Energy Environ. Eng.*, vol. 26, pp. 345–355, Dec. 2015.

[26] K. Li and H. Su, "Forecasting building energy consumption with hybrid genetic algorithm-hierarchical adaptive network-based fuzzy inference system," *Energy Buildings*, vol. 42, pp. 2070–2076, Nov. 2010. doi: 10.1016/j.enbuild.2010.06.016.

[27] K. Li, C. Hu, G. Liu, and W. Xue, "Building's electricity consumption prediction using optimized artificial neural networks and principal component analysis," *Energy Buildings*, vol. 108, pp. 106–113, Dec. 2015. doi: 10.1016/j.enbuild.2015.09.002.

[28] A. Selakov, D. Cvijetinović, L. Milović, S. Mellon, and D. Bekut, "Hybrid PSO–SVM method for short-term load forecasting during periods with significant temperature variations in city of Burbank," *Appl. Soft. Comput.*, vol. 16, pp. 80–88, Mar. 2014. doi: 10.1016/j.asoc.2013.12.001.

[29] F. McLoughlin, A. Duffy, and M. Conlon, "A clustering approach to domestic electricity load profile characterisation using smart metering data," *Appl. Energy*, vol. 141, pp. 190–199, Mar. 2015. doi: 10.1016/j.apenergy.2014.12.039.

[30] K. Amasyali and N. M. El-Gohary, "A review of data-driven building energy consumption prediction studies," *Renew. Sustain. Energy Rev.*, vol. 81, pp. 1192–1205, Jan. 2018. doi: 10.1016/j.rser.2017.04.095.

[31] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324.

[32] D. X. Tien, K.-W. Lim, and L. Jun, "Comparative study of PCA approaches in process monitoring and fault detection," in *Proc. 30th Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, Busan, South Korea, Nov. 2004, pp. 2594–2599.

[33] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *IEEE Comput.*, vol. 42, no. 8, pp. 30–37, Aug. 2009. doi: 10.1109/MC.2009.263.

[34] J. Xu and H. Liu, "Web user clustering analysis based on KMeans algorithm," in *Proc. 2010 Int. Conf. Inf., Netw. Automat. (ICINA)*, Kunming, China, Oct. 2010, pp. V2-6–V2-9.

[35] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *Proc. IEEE Int. Conf. Data Mining*, Sydney, NSW, Australia, Dec. 2010, pp. 911–916.

[36] X. Yuan, A. Su, Y. Yuan, H. Nie, and L. Wang, "An improved PSO for dynamic load dispatch of generators with valve-point effects," *Energy*, vol. 34, no. 1, pp. 67–74, Jan. 2009. doi: 10.1016/j.energy.2008.09.010.

[37] M. Lindenbaum, S. Markovitch, and D. Rusakov, "Selective sampling for nearest neighbor classifiers," *Mach. Learn.*, vol. 54, no. 2, pp. 125–152, 2004. doi: 10.1023/B:MACH.0000011805.60520.fe.

[38] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002. doi: 10.1613/jair.953.

[39] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995. doi: 10.1007/BF00994018.

[40] A. S. Ahmad *et al.*, "A review on applications of ANN and SVM for building electrical energy consumption forecasting," *Renew. Sustain. Energy Rev.*, vol. 33, pp. 102–109, May 2014. doi: 10.1016/j.rser.2014.01.069.

[41] S. U. Jan, Y. D. Lee, J. Shin, and I. Koo, "Sensor fault classification based on support vector machine and statistical time-domain features," *IEEE Access*, vol. 5, pp. 8682–8690, May 2017. doi: 10.1109/ACCESS.2017.2705644.

[42] Y. Wei *et al.*, "A review of data-driven approaches for prediction and classification of building energy consumption," *Renew. Sustain. Energy Rev.*, vol. 82, pp. 1027–1047, Feb. 2018. doi: 10.1016/j.rser.2017.09.108.

[43] H. X. Zhao and F. Magoulès, "Parallel support vector machines applied to the prediction of multiple buildings energy consumption," *J. Algorithms Comput. Technol.*, vol. 4, no. 2, pp. 231–249, Jun. 2010. doi: 10.1260/1748-3018.4.2.231.

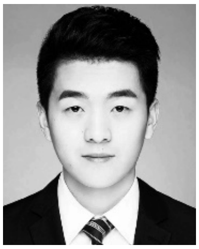
[44] L. Xuemei, D. Yuyan, D. Lixing, J. Liangzhong, "Building cooling load forecasting using fuzzy support vector machine and fuzzy c-mean clustering," in *Proc. Int. Conf. Comput. Commun. Technol. Agricult. Eng. (CCTAE)*, Chengdu, China, Jun. 2010, pp. 438–441.

[45] Office of Energy Efficiency & Renewable Energy. *DOE Buildings Performance Database, Sample Residential Data*. Accessed: Sep. 2018. [Online]. Available: https://openei.org/datasets/dataset/doe-buildings-performance-database-sample-residential-data

[46] *National Weather Service*. Accessed: Sep. 2018. [Online]. Available: https://www.weather.gov/

[47] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation," *J. Mach. Learn. Technol.*, vol. 2, pp. 37–63, Dec. 2011.

HUILING CAI received the B.S. degree from Jiangnan University, Wuxi, China, in 2017. She is currently pursuing the Ph.D. degree in control theory and control engineering with the Department of Control Science and Engineering, Tongji University, Shanghai, China. Her research interests include machine learning and intelligent control.



SHOUPENG SHEN received the B.S. degree from the Changshu Institute of Technology, Changshu, China, in 2015, and the M.S. degree from Tongji University, Shanghai, China, in 2018. Since 2018, he has been an Engineer with Hucheng Information Technology Ltd., Company, Shanghai. His research interests include machine learning and data mining.

QINGCHENG LIN received the B.S. degree from Northeastern University, Shenyang, China, in 2018. She is currently pursuing the Ph.D. degree in control theory and control engineering with the Department of Control Science and Engineering, Tongji University, Shanghai, China. Her research interests include machine learning and intelligent control.

XUEFENG LI (M'17) received the B.S. degree from the Shenyang Institute of Engineering, China, in 1999, and the M.S. and Ph.D. degrees from the Fukuoka Institute of Technology, Fukuoka, Japan, in 2004 and 2007, respectively.

From 2007 to 2013, he was a Postdoctorate and an Assistant Researcher with Waseda University, Kitakyushu, Japan. Since 2010, he has been an Associate Professor with Tongji University. His research interests include sensors and intelligent control.

HUI XIAO received the B.S., M.S., and Ph.D. degrees from Tongji University, Shanghai, China, in 1992, 1998, and 2007, respectively.

From 1992 to 1997, she was a Lecturer with Tongji University. From 1997 to 2011, she was an Associate Professor at Tongji University. Since 2011, she has been a Professor with Tongji University. Her research interests include intelligent building, intelligent control, life optics, and application novel techniques.