Forecast of Short-Term Electricity Price Based on Data Analysis

Shuang Wu, Li He, Zhaolong Zhang, and Yu Du

1School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan, China
2College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen, China
3College of Mechatronics and Automation, Wuchang Shouyi University, Wuhan, China

Correspondence should be addressed to Li He; heli@szu.edu.cn

1.Introduction

In the power market, the electricity price is an essential element because it influences the behaviour of power generation enterprises, power supply enterprises, and other buyers [1]. In recent years, with the rapid development of smart grids, the increase in the number of power users, and the large-scale development of renewable energy, more and more factors have begun affecting the electricity price, which is making price forecasting more difficult. Therefore, how to extract and mine useful information from the electricity price and its influencing factors are extremely important for accurate and timely forecasting.

Currently, the methods of forecasting include traditional forecasting methods [2–6] and intelligence methods [7–12]. Among these methods, support-vector machine (SVM) has higher generalization ability and good robustness, so it is widely applied to forecast electricity price. A framework composed of time series segmentation, recursive feature elimination, and minimum redundancy maximum relevance was developed based on SVM [13]. Chen et al. [14] proposed a combination of nonlinear regression and SVM to forecast the price in peak and off-peak periods. In [15], a least-squares SVM, combining radial basis function (RBF) and universal SVM kernel function, was applied to forecast electricity price.

Data processing has a great impact on forecasting accuracy. An efficient sparse autoencoder was proposed to extract features and a nonlinear autoregressive network was applied to forecast electricity price [16]. In [17], the improved wavelet transform was applied to process the input data, and the new feature selection was applied to filter the input data. XG-boost, decision tree, recursive feature elimination, and random forest were applied to feature
selection and feature extraction [18]. Jahangir et al.[19] used grey correlation analysis (GCA) to select features, and the data were denoised by using a deep neural network with stacked denoising autoencoders. To the best of the authors’ knowledge, some developments are missing from research on electricity price forecasting based on data analysis, categorized as follows:

(1) Electricity price data are huge and contains a great deal of useless information, which is prone to low accuracy. Moreover, SVM has a high time-cost when using big data. In the past, researchers have carried out data processing on the features, ignoring that too many useless samples will also affect the forecasting accuracy and the time-cost.

(2) Existing forecasting models are generally based on a single sample set composed of features and combined with the models of data processing, which is prone to extracting excessive data, resulting in poor forecasting accuracy.

The main contributions of this paper are as follows:

(1) Sample selection, feature selection, and feature extraction are applied to clean the massive data to decrease the amount of data and make full use of the useful information so as to improve the accuracy of the SVM model and reduce the time-cost.

(2) Considering the periodicity and the temporal correlation of the electricity price, feature classification is proposed to avoid excessive extraction of data to reduce the probability of the forecasting model falling into local optima.

(3) A framework of electricity price forecasting based on big data is proposed to realize 24-hour forecasting, rather than single-point forecasting by 24 times.

The main structure of this paper is as follows: Section 2 introduces the model of electricity price forecasting in this paper; Section 3 introduces the method of data processing; Section 4 shows the numerical results and details of a case study; and Section 5 discusses the results of this study.

2. Price Forecasting Model Based on DE-SVM

2.1. SVM Model. SVM is based on the Vapnik–Chervonenkis (VC) dimension theory [20] and the principle of structural risk minimization [21]. It seeks the best compromise between model complexity and learning ability. Compared with ANN and random forest (RF), SVM has a better generalization ability and higher forecasting accuracy in solving nonlinear problems. Therefore, SVM is selected as a predictor, as shown in [22].

The SVM problem can be transformed into the following expression:

$$\min_{\omega, b} \frac{1}{2} \omega^2 + C \sum_{i=1}^{m} \ell_{i}(f(x_i) - y_i),$$

where $\omega$ is a weight vector; $b$ is a bias; $C$ is the regularization constant; $\ell_{i}$ is the insensitive loss function; $f(x_i)$ is the actual output of sample $i$; $x_i$ is the feature related to the electricity price; and $y_i$ is the actual price.

The insensitive loss function is defined as follows:

$$\ell_{i}z = \begin{cases} 0, & \text{if } |z| \leq \epsilon, \\ |z| - \epsilon, & \text{otherwise.} \end{cases}$$

SVM needs to introduce a kernel function. Because the kernel function based on RBF has strong stability in dealing with nonlinear problems, the RBF kernel function is adopted:

$$K(x_i, x_j) = \exp \left(\frac{x_i - x_j^2}{2\sigma^2}\right),$$

where $\sigma$ is the width coefficient.

The final model is as follows:

$$f(x) = \sum_{i=1}^{m} (\tilde{a}_i - a_i)K(x, x_i) + y_i + \epsilon - \sum_{i=1}^{m} (\tilde{a}_i - a_i)K(x, x_i).$$

2.2. Parameter Optimization Based on DE. DE is an adaptive global optimization algorithm based on population with a simple structure and strong robustness. There are two parameters ($\sigma$ and penalty factor $c$) to be optimized in SVM, which affect the forecasting accuracy of SVM. However, the sample size for training SVM is huge, and it takes a long time for DE to find the optimal parameters. Wang et al.[23] proposed an improved DE for this problem and introduced a scale factor of dynamic adjustment in mutation, which can speed up the optimization process. Therefore, the improved DE is used to optimize the parameters $\sigma$ and $c$.

DE includes four steps: initialization, mutation, crossover, and selection, in which the scale factor of dynamic adjustment is introduced into the mutation operation. The formulas for this step are

$$V_i^{t+1} = X_i^t + F_i (X_i^t - X_{r_1}^t - X_{r_2}^t),$$

$$F_i = F_{\max} - \frac{g(F_{\max} - F_{\min})}{g_{\max}},$$

where $i \neq r_1 \neq r_2 \neq r_3$, $X_i^t$ is the most suitable individual in the $g$th generation; $F_i$ controls the mutation scale of the $i$th iteration; $F_{\max}$ and $F_{\min}$ are the maximum and minimum $F_i$; and $g_{\max}$ is the total number of iterations.

3. Forecasting Model Based on Data Analysis

3.1. Framework of Forecasting. The time-cost and forecasting accuracy of DE-SVM will be affected when the irrelevant data in the samples and features of the electricity price are applied to the forecasting process. Therefore, this paper proposes feature classification and establishes three models of data processing to select and extract the valid samples and
features, which ultimately reduces the time-cost and improves the accuracy. Figure 1 shows the framework of the price forecasting, including four models: sample selection based on GCA, feature selection based on GCA, feature extraction based on principal component analysis (PCA), and price forecasting based on DE-SVM. Each model plays an important role in the framework. First, sample selection based on GCA is applied to select the valid samples, and feature selection based on GCA is applied to select the important features of the samples. Then, feature extraction is used to extract features of the samples. Finally, DE-SVM is used to forecast the electricity price.

3.2. Sample Selection Based on GCA

3.2.1. GCA Method. GCA determines correlation by quantifying the “closeness” between two different data sequences. The more similar the two data sequences, the greater the correlation between them. The result of GCA is stable and fast, and GCA has good performance in removing irrelevant data. Many scholars have researched GCA [24–26]. Therefore, this paper uses GCA to select important samples and their important features.

(1) The comparison sequence matrix \( D \) is defined as

\[
D = \begin{bmatrix}
\lambda_1(1) & \cdots & \lambda_2(1) & \cdots & \lambda_n(1) \\
\vdots & & \vdots & & \vdots \\
\lambda_1(k) & \lambda_2(k) & \lambda_n(k) \\
\vdots & & \vdots & & \vdots \\
\lambda_1(m) & \cdots & \lambda_2(m) & \cdots & \lambda_n(m)
\end{bmatrix},
\]

where \( \lambda_n(k) \) is the \( n \)th feature in the \( k \)th sample.

(2) The reference sequence is defined as

\[
\lambda_0 = [\lambda_01, \lambda_02, \ldots, \lambda_0(m)]^T,
\]

where \( \lambda_0 \) is the target sequence.

(3) Then, the grey coefficient [27] is determined as

\[
\gamma \lambda_0^*(k), \lambda_1^*(k) = \frac{\Delta_{\min} + \xi \Delta_{\max}}{\Delta_0^*(k) + \xi \Delta_{\max}}, \quad \xi \in (0, 1),
\]

where \( \lambda_1^*(k) \) and \( \lambda_0^*(k) \) are the normalized components of the data sequence; \( \Delta_{\min} \) and \( \Delta_{\max} \) respectively, represent the minimum and maximum of the absolute value of the difference between the reference sequence and the comparison sequences; \( \Delta_0^*(k) \) represents the absolute value of the difference between the reference sequence and the \( i \)th comparison sequence; and \( \xi \) is the distinguishing factor, usually set to 0.5 [28].

(4) The final grey correlation grade is expressed as follows:

\[
\Gamma_i = \sum_{i=1}^{m} \gamma \lambda_0^*(k), \lambda_i^*(k)
\]

where \( \Gamma_i \) is the grey correlation grade of the sample set.

3.2.2. Sample Input considering Periodicity. Each sample is composed of multiple features. Sample selection based on GCA should not require too many features. Otherwise, the final sample set will be extremely similar, resulting in poor fault tolerance and ultimately affecting the forecasting accuracy. In this paper, a sample corresponds to a moment, and the electricity price changes periodically with time [29]. Therefore, a small number of features (\( n_1 \) periodic features) that change periodically with time are selected here to form sample set 1 and serve as the input of the model of sample selection based on GCA.

3.2.3. Sample Selection considering Forecasting Period. There are many useless samples in the input samples, which can affect the forecasting accuracy and time-cost of SVM. GCA can quickly determine the importance of different sequences with a stable result, so GCA has good performance for removing useless samples. Therefore, GCA is
applied to eliminate useless samples. An example of GCA sample selection is shown in Figure 2. GCA removes useless samples and finally obtains \( m_1 \) important samples.

For the sample selection based on GCA, it can be applied to single-point forecasting. To forecast the electricity price in a time period, it needs to be forecasted many times. For this reason, the model of electricity price forecasting for a given time period is proposed, and the following contents should be introduced into GCA.

1. The forecasting period \( s \) is introduced.
2. The reference sequence of multiple targets is defined as

\[
\lambda_0 = \begin{bmatrix}
\lambda_{01}(1) & \lambda_{02}(1) & \cdots & \lambda_{0n}(1) \\
\lambda_{01}(2) & \lambda_{02}(2) & \cdots & \lambda_{0n}(2) \\
\vdots & \vdots & \ddots & \vdots \\
\lambda_{01}(s) & \lambda_{02}(s) & \cdots & \lambda_{0n}(s)
\end{bmatrix}^T. \tag{10}
\]

The process is shown in Figure 3. In sample set 1, the important samples at all times in the forecasting period are found and then superimposed and combined to form an important sample set for the forecasting period.

3.3. Feature Selection Based on GCA

3.3.1. Sample Input considering Temporal Correlation.

The number of features required for feature selection based on GCA is different from that for sample selection. The more features are selected, the more comprehensive the factors affecting electricity price. If a single sample set composed of features is successively subjected to sample selection and feature extraction, then the sample set lacks pertinence. At the same time, as the model of the data processing increases, it becomes more likely that excessive data will be extracted, which ultimately affects the forecasting accuracy. Therefore, feature classification is proposed. This paper classifies the features and introduces them into two models, as shown in Figure 4. First, \( n_s \) periodic features, comprising sample set 1, are applied to the sample selection. Next, temporal correlation features, combined with the important sample set to form sample set 2, are applied to the feature selection. The temporal correlation features, including historical electricity price and temperature of the previous day, are related to the electricity price in the adjacent period.

3.3.2. Feature Selection. There are many features of the electricity price, some of which are useless. The data obtained by sample selection are calculated by the GCA to determine the importance between the target price and each feature. Compared with RF, relief, and other methods of feature selection, the calculation speed of GCA is faster and the results are more stable. Therefore, GCA is applied to analyse the price features and remove the useless features. An example is shown in Figure 5. Finally, \( n_s \) important features are retained and used as the input of feature extraction based on PCA.

3.4. Feature Extraction Based on PCA. Two GCAs can remove useless samples and features, but they cannot remove the redundant information between features, and this redundant information will lead to poor forecasting accuracy. PCA can quickly calculate the results and effectively remove the redundancy. Therefore, PCA is applied to reduce the dimensions of the original features into several comprehensive indexes that contain most of the information.

The \( i \)th principal component contribution [30] can be expressed as follows:

\[
\eta_i = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \times 100\%. \tag{11}
\]

where \( \lambda_i \) represents the \( i \)th eigenvalue.

The cumulative contribution is as follows:

\[
\eta_L = \sum_{i=1}^{p} \left( \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \times 100\% \right). \tag{12}
\]

The principal component can be expressed as

\[
F_i = u_{i1}x_1 + u_{i2}x_2 + \cdots + u_{ip}x_p. \tag{13}
\]

3.5. Process of Electricity Price Forecasting. The electricity price forecasting method proposed in this paper classifies features, selects the important samples and features using GCA, extracts the features using PCA, and finally forecasts the price using DE-SVM. The specific steps are as follows:

1. Set \( t = 1 \), \( a = 0 \), and \( i = 1 \).
2. Select periodic features to form sample set 1 similar to equation (6), and the sample sequence \( D = [\lambda(1), \lambda(2), \ldots, \lambda(m)]^T \). The reference sequence of multiple targets is defined by equation (10).
3. Calculate equation (9) using \( D^T \) to obtain \( \Gamma \cdot \lambda_0^*, \lambda^* \).

Then, important samples at time \( t \) are obtained, which are larger than the control threshold \( \mu_1 \) of the first GCA. Set \( t = t + 1 \).
4. If \( t \leq s \), return to Step (3); otherwise, go to Step (5).

Finally, the important sample set of the forecast period \( s \) is obtained.
(5) Considering the temporal correlation features, the important sample set of the forecast period $s$ is expanded to sample set 2 similar to equation (6).

(6) Calculate equation (9) using sample set 2. Then, $n_2$ important features are obtained, which are larger than the control threshold $\mu_2$ of the second GCA.
(7) Calculate equation (12) to obtain $\eta_k$. Set $i = i + 1$.

(8) If $\eta_k < \theta$, calculate the first $i$ principal components $F_i$ using equation (13) and return to Step (7); otherwise, go to Step (9).

(9) Solve equation (4) to obtain $f(x)$ and optimize the parameters $\sigma$ and $c$ using DE [24].

4. Results

4.1. Implementation Platform and Data Selection. This paper uses MATLAB R2016a simulation. During the simulation, MATLAB is running on a platform with an Intel Core i5, 4 GB RAM, and 500 GB hard disk. Location marginal price (LMP) and its features in the New England electricity market from 2017 to 2018 are used to verify the proposed model. Periodic features are shown in Table 1, and temporal correlation features are shown in Table 2.

4.2. Model Evaluation Index. To verify the effectiveness of the method in this paper, mean absolute percent error (MAPE) and root-mean-square error (RMSE) are selected as the evaluation indexes of the forecast model:

$$E_{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \frac{|Y_i - \tilde{Y}_i|}{|Y_i|},$$

$$E_{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} |Y_i - \tilde{Y}_i|^2},$$

(14)

where $Y_i$ is the real value and $\tilde{Y}$ is the predicted value.

4.3. Parameter Settings. The parameters in the framework are set as follows: (i) sample selection based on GCA: the distinguishing factors $\xi_1 = 0.5$, the control threshold $\mu_1 = 0.983$; (ii) feature selection based on GCA: the distinguishing factors $\xi_2 = 0.5$, the control threshold $\mu_2 = 0.64$; (iii) DE-SVM: population size $N_p = 50$, the maximum number of iterations is 50, the scaling factor range is $[0.2, 0.8]$, the crossover probability is 0.5, the insensitivity coefficient of SVM is set to 0.01, and the change range of $c$ and $\sigma$ is set to $[0.01, 10]$.

4.4. Result Analysis

4.4.1. Analysis of Complexity. Both the data processing and the electricity price forecasting have a time-cost. The complexity of each step is analysed as follows.

Regarding the data processing, the complexity of sample selection is $O(m n_1)$, the complexity of feature selection is $O(m n)$, and the complexity of feature extraction is $O(n^2)$, where $n_1 \leq n < n \ll n_1 < m$. When $n_1$, $n$, and $n_2$ change, the complexity is shown in Figure 6. As can be seen from Figure 6, the time-cost of data processing is very small.

As for the electricity price forecasting based on DE-SVM, the complexity of SVM is $O(SV^3)$ and DE-SVM is $O(N_p SV^3)$, where $SV$ is the number of support vectors. However, after data processing, the number of training samples and features decreases greatly, thereby reducing SV and time-cost.

4.4.2. Daily Forecasting Results. (1) Sample Selection Based on GCA. Taking November 29, 2018, as an example, the important samples in sample set 1 are selected, and the grey correlation grades of 24:00 are shown in Figure 7. In the forecasting period, each sample at each time-point can obtain a grey correlation grade, similar to the results presented in Figure 7. The samples with a correlation degree greater than 0.983 are selected, and finally 3,466 important samples are obtained, which avoids the problems of forecasting accuracy and time-cost caused by the samples with large differences.

(2) Feature Selection Based on GCA. This paper proposes feature classification according to periodicity and temporal correlation and inputs the framework twice. The comparison results are shown in Table 3.

In Table 3, we find that the forecasting results of this paper are better. When 3,466 samples are selected, the MAPE is the smallest at 10.05%.
The importance of each feature is shown in Figure 8, and the forecasting results of DE-SVM under different thresholds are shown in Table 4.

In Table 4, when the threshold is 0.64, the forecasting results are the best, with MAPE = 8.83% and RMSE = 5.4862. Some features (1, 2, 3, 5, 6, 20, 21, 22, 23, 24, 25, 26, 34, and 35) are finally removed in Figure 8. Among these features, there are the features of historical price with long time distance, and the larger the time distance is, the weaker the correlation is. Therefore, it is reasonable to remove such features. The removed features also include hour, week, temperature, and season, which are not directly related to the electricity price, so the correlation is not large and can be removed.

(3) Feature Extraction Based on PCA. PCA transforms the 21 original features into 21 principal components. The corresponding contribution and cumulative contribution are shown in Figure 9.

The first nine principal components, containing 95% of the original features, are shown in Figure 9, which can replace the original features as the input of the DE-SVM forecast model.

### Table 2: Temporal correlation features of the electricity price.

| Number | Feature                                       |
|--------|-----------------------------------------------|
| 7      | Day-ahead LMP of the first 12 hours           |
| 8      | Day-ahead LMP of the first 11 hours           |
| 9      | Day-ahead LMP of the first 10 hours           |
| 10     | Day-ahead LMP of the first 9 hours            |
| 11     | Day-ahead LMP of the first 8 hours            |
| 12     | Day-ahead LMP of the first 7 hours            |
| 13     | Day-ahead LMP of the first 6 hours            |
| 14     | Day-ahead LMP of the first 5 hours            |
| 15     | Day-ahead LMP of the first 4 hours            |
| 16     | Day-ahead LMP of the first 3 hours            |
| 17     | Day-ahead LMP of the first 2 hours            |
| 18     | Day-ahead LMP of the first 1 hours            |
| 19     | Day-ahead LMP of the first 1 hours            |
| 20     | LMP of the first 12 hours of one day ahead     |
| 21     | LMP of the first 11 hours of one day ahead     |
| 22     | LMP of the first 10 hours of one day ahead     |
| 23     | LMP of the first 9 hours of one day ahead      |
| 24     | LMP of the first 8 hours of one day ahead      |
| 25     | LMP of the first 7 hours of one day ahead      |
| 26     | LMP of the first 6 hours of one day ahead      |
| 27     | LMP of the first 5 hours of one day ahead      |
| 28     | LMP of the first 4 hours of one day ahead      |
| 29     | LMP of the first 3 hours of one day ahead      |
| 30     | LMP of the first 2 hours of one day ahead      |
| 31     | LMP of the first 1 hours of one day ahead      |
| 32     | LMP one day ahead                             |
| 33     | Forecast load                                 |
| 34     | Dry bulb temperature one day ahead            |
| 35     | Wet bulb temperature one day ahead            |

![Figure 6](image-url)

**Figure 6:** The complexity of data processing. (a) Sample selection. (b) Feature selection. (c) Feature extraction.
Table 3: Comparison of results under two schemes.

| Number of important samples | MAPE (%) | Proposed method |
|-----------------------------|----------|-----------------|
| Use a set of features directly | | |
| 1,686 | 14.14 | 10.58 |
| 2,510 | 12.40 | 11.01 |
| 3,466 | 11.83 | 10.05 |
| 4,359 | 12.60 | 10.19 |

Table 4: Forecasting results based on DE-SVM under different thresholds.

| Threshold | MAPE (%) | RMSE |
|-----------|----------|------|
| 0.61      | 10.79    | 6.226|
| 0.62      | 9.50     | 5.7375|
| 0.63      | 9.09     | 5.5289|
| 0.64      | 8.83     | 5.4862|
| 0.65      | 9.13     | 5.6539|
4. Analysis of Daily Forecasting Results. Figure 10 shows the final forecasting results and the absolute value of the relative error.

In Figure 10, the error of 62.5% is within 10%, and that of 92% is within 20%. However, for most of the data, the forecasting results are within the acceptable range.

In Table 5, four models are applied to forecast the electricity price, and the results are shown in Figure 11.

From Figure 11, the forecasting results of the proposed GGDS and GGPDS are the best, and the forecasting results of DS without data processing are the worst. Compared with DS without data processing, the forecasting accuracy is improved from 81.68% to 91.44%. More detailed results can be seen in Table 6.

In Table 6, the accuracy of DE-SVM will be improved with each addition of processing data: GCA finds important samples, GCA selects features, and PCA extracts features. This is because after each data processing step, the useless information will be further reduced, and the interference will be lessened, thereby improving the forecasting accuracy.

Regarding time, the time-cost of only using DS is 35,074 s, but after the proposed data processing, the time is greatly reduced to 1,809 s.

Finally, regression algorithm (REG), RF, BP neural network (BP), DE-SVM, and the proposed model are each used for forecasting, and the results are shown in Table 7.

Compared with the four benchmark models in Table 7, the results of the proposed method are better, which further reflects the advantages of the method.

4.4.3. Seasonal Forecasting Results. This paper takes any week in each season and forecasts the electricity price of that week. The results are shown in Figures 12–15.

From Figures 12–15, we find that the electricity price of representative weeks in each season fluctuates greatly and their change law is poor. Moreover, the electricity price in Figure 12 is several times greater than that in Figures 13–15.

To better reflect the advantages of this model, we compare it with other models, and the results are shown in Table 8.

From Table 8, it can be seen that the MAPE of the proposed method is smaller than that of other methods. MAPE considers not only the error between predictive value and real value but also the ratio between the error and true value. MAPE is an index of forecasting accuracy in the field of statistics. The proposed method is the most accurate, but in winter and spring, the RMSE of RF and REG is better. This is because RMSE uses average error, which is sensitive to outliers, and the electricity price fluctuation in spring and winter is unstable, and there are a large number of peak load prices, which has a great impact on RMSE.

5. Discussion

In this paper, a period of the electricity price can be forecasted, rather than a single-point price. This provides a valuable tool for period forecasting.

Considering the periodicity and temporal correlation of electricity price, feature classification is proposed, and two kinds of features are input into the framework respectively. Compared with the features as the input directly, this method has higher forecasting accuracy. A key observation point of this paper is sample selection. It can be found that the number of samples has an impact on forecasting results. Therefore, how to automatically select the optimal sample set will be worthy of attention.

The samples and features proposed in this paper are combined to reduce the dimension of data, and DE-SVM is applied to forecast the electricity price of one day (24 time point), namely, the GGPDS model. It was compared with other models (DS, GDS, and GGDS). The time-cost of the DS model is 35,074 s, and the time-cost of the GGDS model is 1,803 s. The time-cost of the GGPDS model, 1,809 s, is slightly higher than that of the GGDS model. However, the total time is greatly reduced, which reflects an advantage of this framework. Overall, this method has high forecasting
Figure 10: The results of four forecasting models. (a) Forecasting results. (b) Absolute value of relative error.

Table 5: Forecast method.

| Abbreviation | Method                                      |
|--------------|---------------------------------------------|
| DS           | DE-SVM                                      |
| GDS          | 1 GCA finds samples + DE-SVM                |
| GGDS         | 1 GCA finds samples + 2 GCA selects features + DE-SVM |
| GGPDS        | 1 GCA finds samples + 2 GCA selects features + PCA + DE-SVM |
Figure 11: The results of four forecasting models. (a) Forecasting results. (b) Absolute value of relative error. (c) Accuracy.

Table 6: Error and training model of each scheme.

| Model   | MAPE (%) | RMSE     | Time (s) |
|---------|----------|----------|----------|
| DS      | 18.38    | 11.0137  | 35,074   |
| GDS     | 10.05    | 5.9629   | 3,918    |
| GGDS    | 8.83     | 5.4862   | 1,803    |
| GGPDS   | 8.50     | 5.4842   | 1,809    |

Table 7: Forecasting results of different models.

| Model   | MAPE (%) |
|---------|----------|
| REG     | 12.22    |
| RF      | 30.41    |
| BP      | 12.17    |
| DE-SVM  | 18.38    |
| Proposed| 8.50     |

Figure 12: The forecasting results of the electricity price of Jan 1–Jan 7.
**Figure 13:** The forecasting results of the electricity price of May 27–June 3.

**Figure 14:** The forecasting results of the electricity price of July 9–July 15.

**Figure 15:** The forecasting results of the electricity price of Oct 8–Oct 14.
accuracy and low time-cost, which provides a new idea for data processing and can be applied to other fields. Moreover, this method will be significant for dealing with the “big data” problem.

To further verify the feasibility and correctness, the results of the proposed models are compared with those of other benchmark methods (REG, RF, BP, and DS). The comparison indicates that the proposed method is superior. Finally, the model is applied to forecast the electricity price of different seasons to verify its robustness. It is found that the results of May 27–June 3 are poor, while the price fluctuation during this part of the year is large and unstable, and it is difficult to find the internal law. However, the prices of July 9–July 15 are relatively stable, and their results are good. Therefore, the proposed method may be more suitable for forecasting electricity prices during periods of the year which have historically shown stable electricity prices.

6. Conclusion

Aiming at the time-cost and accuracy of electricity price forecasting caused by useless samples and features in big data, a forecasting framework based on big data analysis is proposed. The framework includes feature classification considering the periodicity and the temporal correlation, sample selection based on GCA, feature selection based GCA, feature extraction based PCA, and electricity price forecasting based on DE-SVM.

We apply the framework to forecast price in the New England electricity market. The results show that the probability of over extracting data can be reduced. Besides, compared with DS without data processing, the forecasting accuracy is higher and the time-cost is lower, where the forecasting accuracy is improved from 81.68% to 91.44%, and the time-cost decreases from 35,074 s to 1,809 s. The framework provides a method of electricity price forecasting with strong applicability. However, the model can also be applied to other fields of forecasting.

Data Availability

The data used in this paper are all from ISO new England energy: http://www.iso-ne.com.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors’ Contributions

Shuang Wu and Li He contributed to the conception and method. Shuang Wu and Zhaolong Zhang contributed to the computation. All of the authors contributed to the validation.

Acknowledgments

This research was funded by the National Key Research and Development Program of China (No. 2019YFB2102703-004), the National Natural Science Foundation of China (No. 51309094), and the Science, Technology and Innovation Commission of Shenzhen Municipality (No. 20200125).

References

[1] D. Liu and H. H. Qiao, “A review of electricity price forecasting in deregulated electricity markets,” Shaanxi Electric Power, vol. 39, no. 8, pp. 27–32+41, 2011.
[2] D. Bissing, M. T. Klein, R. A. Chinnathambi, D. F. Selvaraj, and P. Ranganathan, “A hybrid regression model for day-ahead energy price forecasting,” IEEE Access, vol. 7, pp. 36833–36842, 2019.
[3] Y. Quan, M. Jiang, H. Sun et al., “Coordination scheduling of wind-hydropower generation and profit allocation based on shapley value method,” Mathematical Problems in Engineering, vol. 2020, Article ID 4785183, 20 pages, 2020.
[4] J.-L. Zhang, Y.-J. Zhang, D.-Z. Li, Z.-F. Tan, and J.-F. Ji, “Forecasting day-ahead electricity prices using a new integrated model,” International Journal of Electrical Power & Energy Systems, vol. 105, pp. 541–548, 2019.
[5] G. P. Girish, “Spot electricity price forecasting in Indian electricity market using autoregressive-GARCH models,” Energy Strategy Reviews, vol. 11-12, pp. 52–57, 2016.
[6] J. P. Gonzalez, A. M. S. Roque, and E. A. Perez, “Forecasting functional time series with a new hilbertian ARMAX model: application to electricity price forecasting,” IEEE Transactions on Power Systems, vol. 33, no. 1, pp. 545–556, 2018.
[7] F. H. Jufri, S. Oh, and J. Jung, “Day-ahead system marginal price forecasting using artificial neural network and similar-days information,” Journal of Electrical Engineering & Technology, vol. 14, no. 2, pp. 561–568, 2019.
[8] J. Portoles, C. Gonzalez, J. M. Moguerza et al., “Electricity price forecasting with dynamic trees: a benchmark against the random forest approach,” Energies, vol. 6, Article ID 1588, 2018.
[9] F. Feijoo, W. Silva, and T. K. Das, “A computationally efficient electricity price forecasting model for real time energy

Table 8: Weekly results of each season in the different models.

| Method | Jan 1–Jan 7 | May 27–June 3 | July 9–July 15 | Oct 8–Oct 14 |
|--------|-------------|--------------|--------------|-------------|
|        | MAPE (%)    | RMSE         | MAPE (%)     | RMSE        |
|        |             |              |              |             |
| Proposed | 18.26      | 66.10       | 28.67        | 28.57       |
| GDS     | 25.28       | 88.39       | 30.20        | 29.37       |
| DS      | 25.96       | 88.22       | 30.70        | 28.62       |
| REG     | 22.39       | 60.82       | 29.00        | 27.06       |
| RF      | 24.29       | 54.16       | 30.45        | 28.06       |
| BP      | 31.70       | 94.88       | 37.96        | 28.75       |

| Method | MAPE (%) | RMSE | MAPE (%) | RMSE | MAPE (%) | RMSE | MAPE (%) | RMSE |
|--------|----------|------|----------|------|----------|------|----------|------|
| REG    | 13.44    | 6.62 | 14.21    | 8.09 | 15.68    | 8.24 | 16.83    | 8.44 |
| DS     | 14.52    | 8.81 | 15.68    | 8.24 | 17.02    | 8.81 | 18.14    | 9.53 |
| RF     | 17.62    | 10.48| 18.54    | 10.90| 19.68    | 11.50| 21.86    | 11.03|
| BP     | 20.07    | 8.81 | 20.07    | 8.81 | 21.86    | 11.03| 23.67    | 11.37|

Mathematical Problems in Engineering 13
markets,” *Energy Conversion and Management*, vol. 113, pp. 27–35, 2016.

[10] W. Ahmad, N. Ayub, T. Ali et al., “Towards short term electricity load forecasting using improved support vector machine and extreme learning machine,” *Energies*, vol. 13, no. 11, Article ID 2907, 2020.

[11] W. B. Qiao and Z. Yang, “Forecast the electricity price of US using a wavelet transform-based hybrid model,” *Energy*, vol. 193, Article ID 116704, 2020.

[12] J. Li, S. Zhu, Q. Wu et al., “A hybrid forecasting model based on EMD-GASVM-RBFNN for power grid investment demand,” *Mathematical Problems in Engineering*, vol. 2018, Article ID 7416037, 17 pages, 2018.

[13] Z. Shao, S. Yang, F. Gao, K. Zhou, and P. Lin, “A new electricity price prediction strategy using mutual information-based SVM-RFE classification,” *Renewable and Sustainable Energy Reviews*, vol. 70, pp. 330–341, 2017.

[14] Z. Ma, H. Zhong, L. Xie, Q. Xia, and C. Kang, “Month ahead average daily electricity price profile forecasting based on a hybrid nonlinear regression and SVM model: an ERCOT case study,” *Journal of Modern Power Systems and Clean Energy*, vol. 6, no. 2, pp. 281–291, 2018.

[15] Y. H. Chen, M. Li, C. H. Li et al., “A hybrid model for electricity price forecasting based on least square support vector machines with combined kernel,” *Journal of Renewable and Sustainable Energy*, vol. 10, no. 5, Article ID 055502, 2018.

[16] S. Mujeeb and N. Javaid, “ESAENARX and DE-RELM: Novel schemes for big data predictive analytics of electricity load and price,” *Sustainable Cities and Society*, vol. 51, Article ID 101642, 2019.

[17] N. Ghadimi, A. Akharimad, H. Shayeghi, and O. Abedinia, “A new prediction model based on multi-block forecast engine in smart grid,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, no. 6, pp. 1873–1888, 2018.

[18] M. Zabih, F. Ahmed, N. Javaid et al., “Electricity price and load forecasting using enhanced convolutional neural network and enhanced support vector regression in smart grids,” *Electronics*, vol. 8, Article ID 122, 2019.

[19] H. Jahangir, H. Tayarani, S. Baghali et al., “A novel electricity price forecasting approach based on dimension reduction strategy and rough artificial neural networks,” *IEEE Transactions on Industrial Informatics*, vol. 16, no. 4, pp. 2369–2381, 2020.

[20] X. Liu, C. Gao, and P. Li, “A comparative analysis of support vector machines and extreme learning machines,” *Neural Networks*, vol. 33, pp. 58–66, 2012.

[21] Q. B. Zhu, J. Liu, Y. G. Li et al., “Study on noise reduction in singular value decomposition based on structural risk minimization,” *Journal of Vibration Engineering*, vol. 18, pp. 204–207, 2005.

[22] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond*, MIT Press, Cambridge, MA, USA, 1986.

[23] K. Wang, C. Xu, Y. Zhang, S. Guo, and A. Y. Zomaya, “Robust big data analytics for electricity price forecasting in the smart grid,” *IEEE Transactions on Big Data*, vol. 5, no. 1, pp. 34–45, 2019.

[24] Z. X. Chen and D. Wang, “A prediction model of forest preliminary precision fertilization based on improved GRA-PSO-BP neural network,” *Mathematical Problems in Engineering*, vol. 2020, Article ID 1356096, 17 pages, 2020.

[25] H. Jahangir, M. A. Golkar, and F. Alhameli, “Short-term wind speed forecasting framework based on stacked denoising auto-encoders with rough ANN,” *Sustainable Energy Technologies and Assessments*, vol. 38, Article ID 100601, 2020.

[26] H. Liu, M. Zeng, T. Pan et al., “The green photovoltaic industry installed capacity forecast in China: based on grey relation analysis, improved signal decomposition method, and artificial bee colony algorithm,” *Mathematical Problems in Engineering*, vol. 2020, Article ID 9892480, 15 pages, 2020.

[27] C.-P. Fung, “Manufacturing process optimization for wear property of fiber-reinforced polybutylene terephthalate composites with grey relational analysis,” *Wear*, vol. 254, no. 3–4, pp. 298–306, 2003.

[28] J. L. Deng, “Introduction to the grey system theory,” *Grey System*, vol. 1, no. 1, pp. 1–24, 1989.

[29] X. Zhang, X. F. Wang, and F. H. Chen, “Short-term electricity price forecasting based on period-decoupled price sequence,” *Proceedings of the CSEE*, vol. 25, no. 15, pp. 4–9, 2005.

[30] Y. Zhang, B. Chen, Y. Zhao, and G. Pan, “Wind speed prediction of IPSO-BP neural network based on lorenz disturbance,” *Ieee Access*, vol. 6, pp. 53168–53179, 2018.