Computers and Electrical Engineering 101 (2022) 108125

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Quantitative combination load forecasting model based on forecasting error optimization ☆

Song Deng [a], Fulin Chen [b], Di Wu [c,*], Yi He [d], Hui Ge [e], Yuan Ge [f]

[a] Institute of Advanced Technology, Nanjing University Post & Telecommunication, Nanjing 210003, China
[b] School of Cyber Science and Engineering, SouthEast University, Nanjing 211189, China
[c] Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, China
[d] Old Dominion University, Norfolk, VA 23462, USA
[e] College of Automation and Artificial Intelligence, Nanjing University Post & Telecommunication, Nanjing 210003, China
[f] Key Laboratory of Advanced Perception and Intelligent Control of High-end Equipment, Ministry of Education, Anhui Polytechnic University, Wuhu 241000, China

ARTICLE INFO

Keywords:
Gene expression programming
Data noise reduction
Load forecasting error
Combination load forecasting

ABSTRACT

Accurate load forecasting is indispensable in various applications of the electric power industry. Although existing load forecasting methods perform well, they cannot handle complicated scenarios where load-related data are highly random and uncertain. To deal with this issue, A Quantitative Combination Load Forecasting model(QCLF) is proposed. Its main idea is to incorporate the load forecasting errors into the forecasting process as an optimization problem, which can significantly reduce the adverse impacts of random and uncertain load-related data. First, we propose an improved K-Means and Least Square-based Load Forecasting Error Model(LFEM-KLS) to improve the availability and effectiveness of load-related data. Second, we employ gene expression programming (GEP) to optimize the proposed LFEM-KLS to achieve highly accurate load forecasting. Experimental results on three load datasets demonstrate that a QCLF model significantly outperforms other related load forecasting models.

1. Introduction

We know that accurate load forecasting is indispensable in various applications of the electric power industry, such as power dispatching, control of generators, infrastructure development, and battery management of electric vehicles [1]. Up to date, existing load forecasting methods consist of classical methods, traditional methods, and combined methods.

Although the classic methods are easy to implement and have high forecasting efficiency, they rely on prior knowledge and have relatively poor forecasting accuracy [2]. The traditional artificial neural network-based methods can significantly improve load forecasting accuracy. However, they cannot obtain the quantitative functional relationships between load and load features, which makes them have poor interpretability [3]. The combined methods make full use of various load forecasting features and combine different load forecasting approaches with different weight coefficients, aiming at possessing muti-advantages in load forecasting [4].

Notably, due to the widespread access of flexible loads in some complicated scenarios, the collected load-related data are highly random and uncertain. Undoubtedly, such random and uncertain data will inevitably cause some forecasting errors. While in existing

☆ Reviews processed and recommended for publication to the co-Editor-in-Chief by Guest Editor Dr. Weihua Ou.

* Corresponding author.
E-mail addresses: dengsong@njupt.edu.cn (S. Deng), 230219312@seu.edu.cn (F. Chen), wudi@cigit.ac.cn (D. Wu), yihe@cs.odu.edu (Y. He), gehuissl@njupt.edu.cn (H. Ge), geyuan@ahpu.edu.cn (Y. Ge).

https://doi.org/10.1016/j.compeleceng.2022.108125
Received 16 March 2022; Received in revised form 14 May 2022; Accepted 25 May 2022
Available online 7 June 2022

combined methods, there is no one to constrain the adverse impacts of such random and uncertain data. As a result, they cannot handle complicated scenarios. Existing combinatorial load forecasting algorithms also have two challenges: (1) the randomness and uncertainty of load data make the error of load forecasting model directly affect the accuracy of the overall load forecasting model; (2) each algorithm in the existing combined load forecasting algorithms exists independently and does not take into account the coupling between algorithms.

To better improve accuracy of load forecasting, we propose a quantitative combination load forecasting (QCLF) model based on gene expression programming (GEP) [5], improved K-means algorithm, and least squares method. Its main idea is to incorporate the load forecasting errors into the forecasting process as an optimization problem, significantly reducing the adverse impacts of random and uncertain load-related data. By conducting comparison experiments on three load datasets, we demonstrate that QCLF model has a greater performance advantage over state-of-the-art models. Besides, we also verify that QCLF model can discover the quantitative functional relationships between load and load features, making it possess excellent reusability.

We try to make the following contributions.

- An improved K-Means and least square-based load forecasting error model (LFEM-KLS) are given to improve the availability and effectiveness of load-related data.
- It proposes a QCLF model based on GEP and LFEM-KLS. It has highly accurate load forecasting because it can significantly reduce the adverse impacts of random and uncertain load-related data.

The remainder of this study is structured as follows. Related work is stated in Section 2. Section 3 presents LFEM-KLS. QCLF model is given in Section 4. Section 5 conducts and analyzes the experiments. Finally, Section 6 concludes this paper.

2. Related work

In this section, we focus on the prior art in the load forecasting models and discuss the pros and cons of these models. In brief, the current load forecasting models can be categorized into the single load forecasting models and combined load forecasting models, with details presented in sequence as follows.

2.1. Single load forecasting models

Many techniques are adopted to model a single load forecasting method, including statistic analysis, machine learning, and intelligent optimization algorithm. First is statistic analysis-based methods. Buzna et al. [6] gave a probabilistic load forecasting model for electric vehicles considering the influence of different geographic regions, which used a standard probabilistic model to forecast the load of electric vehicles in lower-level regions. However, this method increases the time overhead of load forecasting, making it difficult to support the dynamic real-time allocation of electric vehicle loads in the distribution network. In [7], conditional residual model was applied to probabilistic load forecasting. However, load forecasting algorithms based on statistical analysis are not suitable for load forecasting scenarios that require high real-time performance due to the large computational load.

Second is machine learning-based methods. Dabbaghjamanesh et al. [8] proposed a machine learning-based load forecasting of electric vehicle charging. However, this method cannot be applied to more complex types of PHEV load forecasting and lacks certain universality. Bandara et al. [9] used a novel model with sets of related time series based on multiple seasonal patterns, and the simulation results proved that the prediction model has high precision. But, we know that machine learning algorithms are prone to local convergence may lead to poor accuracy of the above mentioned load prediction models based on machine learning. To better optimize the load prediction models, many researchers have used algorithms to optimize the load prediction models to improve the forecasting accuracy. In terms of intelligent optimization algorithms, Xie et al. [10] optimized load forecasting model by using particle swarm algorithm. Although the machine learning-based load forecasting algorithms can improve the efficiency of load forecasting, the machine learning algorithms are prone to fall into local optimum which will eventually affect the accuracy of load forecasting models.

Notably, although these single load forecasting models have high forecasting accuracy, they still fail to discover quantitative functional relationships between load and load characteristics, resulting in their problem of poor reusability.

2.2. Combined load forecasting model

To address the shortcomings of single load forecasting model, the combined methods are proposed to make full use of various load forecasting characteristics and approaches. Hence, they possess muti-advantages in load forecasting in general. Zhang et al. [11] gave an improved electric vehicle charging load simulation method. But the model does not consider factors such as the user base's preference for vehicle type and tolerance for low battery state of charge. Li et al. [12] used a hybrid load forecasting method, which combined with MLR and LSTM neural network. The results demonstrated that the model had the better prediction performance compared with single MLR and single LSTM model. Si et al. [13] used a combined solar forecasting model based on remote sensing images and improved CNN.

Also, there are some other combined load forecasting models. Gilanifar et al. [14] proposed a Bayesian spatiotemporal Gaussian process-based load forecasting model. Based on load data preprocessing, Nie et al. [15] presented a new combination forecasting model by combining individual forecasting algorithm and weight determination theory. Zhang et al. [16] proposed a hybrid load

Algorithm 1: IK-means

Input: $T = \{x_1, x_2, ...x_n\}$, K, C;
Output: $C' = \{C'_1, C'_2, ..., C'_K\}$;
1. $\lambda_1 \leftarrow RanCluCenter(T)$;
2. $C \leftarrow C.add(\lambda_1)$;
3. $Cen \leftarrow \frac{1}{C} \sum_{x \in C} x$;
4. $MaxD(m) \leftarrow MaxCalDis(T, Cen)$;
5. $C \leftarrow C.add(\lambda_m)$;
6. Repeat steps 3 to 5 until all cluster centers are calculated.
7. **for** all i in T **do**
8. $j \leftarrow MinCalDisL(x_i, C)$;
9. $C' \leftarrow C'_i.add(x_j)$;
10. $Cen \leftarrow \frac{1}{C'} \sum_{x \in C'} x$;
11. **end for**
12. return C';

forecasting model by combining SSA, neural networks, and LSSVM. Panapakidis et al. [17] presented a hybrid load forecasting model based on clustering algorithm and neural networks.

Despite the success of these combined load forecasting models, they still have two limits. First, similar to the single load forecasting method, they cannot discover quantitative functional relationships between load and load characteristics. Second, they never constrain the adverse impacts of random and uncertain data during modeling load forecasting. While in some complicated scenarios (e.g., battery management of electric vehicles), random and uncertain data are ubiquitous. Meanwhile, these combinatorial load forecasting models do not take into account the coupling effects between different models.

3. LFEM-KLS

3.1. Improved K_means based on maximum distance

In the traditional K_means algorithm [18], the initial clustering centers are generated randomly, which in turn leads to uncertain clustering effects and the algorithm tends to fall into local convergence. To improve the effect and convergent speed of traditional K_means, we propose an improved K_means based on maximum distance (IK-means). The description of IK-means is shown in Algorithm 1.

Let $T = \{x_1, x_2, \dots x_n\}$ be the clustered dataset. We suppose that the first cluster center λ_1 is randomly generated. And we calculate the Euclidean distance from the remaining data points in T to λ_1. The point with the largest Euclidean distance to λ_1 is regarded as the second cluster center λ_2. And then, the centroid of λ_1 and λ_2 can be calculated. Next, we calculate the distance from the remaining data points in T to the centroid, and the point with the largest distance is the next cluster center λ_3. And so on, until k cluster centers are calculated. The IK-means algorithm yields k clusters such that the clusters have the maximum intra-class similarity and the minimum inter-class similarity.

3.2. LFEM-KLS

Existing load forecasting models will produce certain errors. Due to the randomness and irregularity of the errors, if we directly model and analyze the error data, it will lead to poor accuracy of the model. To address this problem, a load forecasting error model based on an IK-means and the least square method (LFEM-KLS) is proposed in this paper, which aims to remove isolated data in the error dataset. The description of LFEM-KLS is shown as follows.

The core of LFEM-KLS algorithm consists of two aspects. One is to reduce the effect of noise in the error dataset. The other is to model the error dataset using least square after noise reduction.

(1) Noise processing in the error data

The noise processing of the error data is to use the improved K-means to eliminate the data that is off the center point in the error data, and to improve the usability of the error model.

Suppose that the load dataset $Y = \{Y_i\}, i \in [1, n], Y_i = (y_{ij})_{n \times m}$, where y_{ij} represents the load value at a given moment, n represents the size of the load dataset and m represents the number of the features affecting the load. And we suppose that $y_{im} = M y_{i1}, \dots y_{i,m-1}$ is the load forecasting model M obtained by calling Algorithm 3, where $y_{i1}, \dots y_{im-1}$ is the historical load value of the previous $m-1$ moments, and y_m is the forecasting value of the model at the mth time.

For the dataset $X = \{X_i\}, i \in [1, n], X_i = (x_{ij})_{n \times m}$, where $(x_{ij})_{n \times (m-1)}$ means the load value at the previous $m-1$ moments, and $x_{im}, i \in [1, n]$ indicates the difference between the forecasting value of the model M at the mth time and the true load value.

Then we cluster the dataset X into K classes by IK-means algorithm, and calculate the number of data under each class and the proportion $P_i, i \in [1, K]$ in the sample data X. When $P_i > 65\%$, the current dataset is regarded as the new dataset $\widetilde{X} = \{\widetilde{X}_1, \widetilde{X}_2, \dots, \widetilde{X}_h\}, h \le n$.

(2) Error data modeling based on least square

Algorithm 2: LFEM-KLS

Input: $X = \{x_1, x_2, ...x_n\}$, K, C;
Output: $\phi(x_{i1}, ..., x_{i,m-1})$;
1. $C \leftarrow IK - means(X, K)$;
2. $Sum \leftarrow 0$;
3. **for** $i \leftarrow 1$ to K **do**
4. $\quad P_i \leftarrow CalPro(C, i)$;
5. $\quad Sum \leftarrow P_i + Sum$;
6. $\quad$ **if** $Sum > 65\%$ **then**
7. $\quad\quad$ return $\widetilde{X} = \{\widetilde{X}_1, \widetilde{X}_2, ..., \widetilde{X}_h\}$;
8. $\quad\quad$ goto Step 11;
9. $\quad$ **end if**
10. **end for**
11. $x_{im} \leftarrow a_1\widetilde{x}_{i1} + ... + a_{m-1}\widetilde{x}_{i,m-1} + C$;
12. $(a_1, ...a_m, C) \leftarrow Cholesky(x_{im})$;
13. return $\phi(x_{i1}, ..., x_{i,m-1})$;

For error dataset $\widetilde{X}$, we construct the following least squares regression equation. For $m-1$ independent variables and a dependent variable, given n groups of observations, the linear expression is shown as follows

$$x_{im} = \phi(\widetilde{x}_{i1}, \dots, \widetilde{x}_{i,m-1}) = a_1\widetilde{x}_{i1} + \cdots + a_{m-1}\widetilde{x}_{i,m-1} + C. \tag{1}$$

where $a_1, \dots a_m, C$ is the regression coefficient and satisfies the following formula

$$(DD^T)\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_{m-1} \\ C \end{bmatrix} = D\begin{bmatrix} \widetilde{x}_{1m} \\ \widetilde{x}_{2m} \\ \widetilde{x}_{3m} \\ \vdots \\ \widetilde{x}_{n-1,m} \\ x_{nm} \end{bmatrix}. \tag{2}$$

where the matrix D is expressed as follows

$$D = \begin{bmatrix} \widetilde{x}_{11} & \widetilde{x}_{12} & \widetilde{x}_{13} & \dots & \widetilde{x}_{1n} \\ \widetilde{x}_{21} & \widetilde{x}_{22} & \widetilde{x}_{23} & \dots & \widetilde{x}_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \widetilde{x}_{(m-1)1} & \widetilde{x}_{(m-1)2} & \widetilde{x}_{(m-1)3} & \dots & \widetilde{x}_{(m-1)n} \\ 1 & 1 & 1 & \dots & 1 \end{bmatrix}. \tag{3}$$

Finally, we solve $a_1, \dots a_m, C$ according to the Cholesky decomposition method, and obtain the error data model based on the least squares as $\phi(\widetilde{x}_{i1}, \dots, \widetilde{x}_{i,m-1})$.

4. Quantitative combination load forecasting model

4.1. LF-GEP

Load forecasting is essentially to mine the non-linear functional relationship between the factors that affect load forecasting [19]. The traditional function discovery algorithm based on nonlinear regression completely relies on prior knowledge. However, the exact function model cannot be known in advance for load forecasting. GEP is a new artificial intelligent algorithm, which is first proposed by Candida to be used in the function finding. The advantage of GEP is that it can fully automatically mine the corresponding nonlinear load forecasting model without prior knowledge. As a result, a load forecasting based on gene expression programming (LF-GEP) is proposed. The description of LF-GEP is shown in Algorithm 3.

To use GEP to build load forecasting model, for the load dataset $L = \{L_i\}, i \in [1, n], L_i = (l_{ij})_{n\times m}$, we set the function set as $\{+, -, *, /\}$, the terminal set as $\{x_0, x_1, \dots, x_n\}$, and the constant set as $C = Random(0, 1)$. Firstly, we initialize the GEP population, calculate the fitness values of all individuals in the population, and then perform all genetic operations which include selection, mutation, recombination, and transposition. Lastly, the GEP individual with the largest fitness value can be obtained. Based on the regular expression, the GEP individual is converted into a mathematical function relationship, which is a mathematical model of load forecasting.

4.2. QCLF

The essence of quantitative combination load forecasting model (QCLF) is the linear or nonlinear combination of two or more load forecasting methods. In this paper, QCLF is proposed to enhance the load forecasting performance by combining LF-GEP and LFEM-KLS. The framework of QCLF is shown in Fig. 1.

Algorithm 3: LF-GEP

Input: $L = \{L_1, L_2, ..., L_n\}$, P_{size}, M_{gen}, P_s, P_m, P_r, P_t;
Output: $f(x_1, x_2, ..., x_n)$;
1. $P_{op} \leftarrow InitPop(L, P_{size})$;
2. $F_v \leftarrow CalFitVal(P_{op})$;
3. **for** $g \leftarrow 0$ to M_{gen} **do**
4. $P_{op} \leftarrow Select(P_s, P_{op})$;
5. $P_{op} \leftarrow Mutate(P_m, P_{op})$;
6. $P_{op} \leftarrow Recom(P_r, P_{op})$;
7. $P_{op} \leftarrow Trans(P_t, P_{op})$;
8. $F_v \leftarrow CalFitVal(P_{op})$;
9. $BestInd \leftarrow Max(F_v)$;
10. **end for**
11. $f(x_1, x_2, ..., x_n) \leftarrow RegExp(BestInd)$;
12. return $f(x_1, x_2, ..., x_n)$;

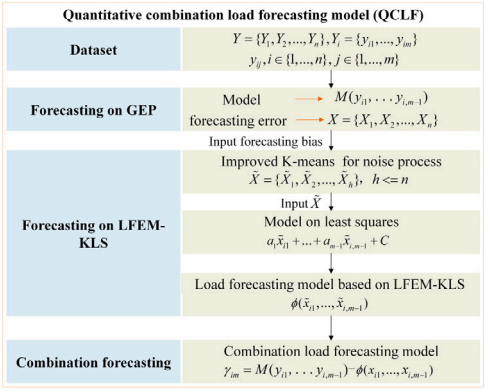

Fig. 1. The framework of QCLF.

Algorithm 4: QCLF

Input: $Y = \{Y_1, Y_2, ..., Y_n\}$, P_{size}, M_{gen}, P_s, P_m, P_r, P_t, K, C;
Output: Combination load forecasting model γ_{im};
1. $M(y_{i1}, ...y_{i,m-1}) \leftarrow LF - GEP(Y, P_{size}, M_{gen}, P_s, P_m, P_r, P_t)$;
2. $R = \{X_1, X_2, ..., X_n\} \leftarrow CalDif(M, Y))$;
3. $\tilde{R} = \{\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_n\} \leftarrow IK - means(X, K, C)$;
4. $\phi(\tilde{r}_{i1}, ..., \tilde{r}_{i,m-1}) \leftarrow \text{LFEM} - KLS(Y)$;
5. $\gamma_{im} \leftarrow M(y_{i1}, ...y_{i,m-1}) - \phi(\tilde{x}_{i1}, ..., \tilde{x}_{i,m-1})$;
6. return γ_{im};

This paper combines the model $My_{i1}, \ldots y_{i,m-1}$ based on LF-GEP with the model $\phi(\tilde{x}_{i1}, \ldots, \tilde{x}_{i,m-1})$ based on LFEM-KLS, and uses the model ϕ to weaken the error generated by the model M. The combination model is regarded as

$$\gamma_{im} = My_{i1}, \ldots y_{i,m-1} - \phi(\tilde{x}_{i1}, \ldots, \tilde{x}_{i,m-1}). \tag{4}$$

The forecasting value obtained by model $My_{i1}, \ldots y_{i,m-1}$ may have errors in each time series. To improve the accuracy of load forecasting, we use an error model $\phi(\tilde{x}_{i1}, \ldots, \tilde{x}_{i,m-1})$ to reduce the impact of the errors generated by the load forecasting model $My_{i1}, \ldots y_{i,m-1}$ on the accuracy of load forecasting. The description of QCLF is shown in Algorithm 4.

5. Experimental analysis

5.1. Dataset

Two real datasets (Hexing 2# and Chengnan 63#) and one public dataset (EUNITE) are selected in experiments. Hexing 2# and Chengnan 63# datasets are supported by State Grid Nantong Power Supply Company, respectively. EUNITE comes from European Network of Intelligent Technology competition load forecast dataset in 1997–1999. All experimental datasets can be downloaded at https://github.com/chenxiaoxin-102496/DataSet.git. Experimental datasets are divided into training and testing dataset, respectively. The detail of experimental datasets is shown in Table 1.

Table 1
Experimental dataset.

Dataset		#Instance	#feature	Date range
Training dataset	Hexing 2#	236	7	08/01/2019–03/23/2020
	Chengnan 63#	232	7	08/05/2019–03/23/2020
	EUNITE	730	7	01/01/1997–12/31/1998
Testing dataset	Hexing 2#	15	7	03/24/2020–04/07/2020
	Chengnan 63#	15	7	03/24/2020–04/07/2020
	EUNITE	31	7	01/01/1999–01/31/1999

Table 2
The comparison of the average running time.

Algorithm	Dataset	#Instance	# Cluster	The averagerunning time (ms)
K-means	Hexing 2#	236	3	78
	Chengnan 63#	232	3	119
	EUNITE	730	3	217
IK-means	Hexing 2#	236	3	**34**
	Chengnan 63#	232	3	**43**
	EUNITE	730	3	**108**

5.2. Metrics

To demonstrate the experimental results, we first use Silhouette-Coefficient and Davies–Bouldin Index to evaluate the performance of the IK-means.

Definition 1 ([20]). Let $p(k)$ be the average distance from the load data L_k to other load data in the same class, denoted as $p(k) = \frac{1}{n}\sum_{i=1,k\neq i}^{n} d(L_k - L_i)$, and $q(k)$ be the minimum average distance from sample L_k to each sample M_{Ci} in other class C, denoted as $q(k) = \min_{C}\{\frac{1}{N_C}\sum_{i=1}^{N_C} d(L_k - M_{Ci})\}$. Then, $S(k) = \frac{q(k)-p(k)}{max\{p(k),q(k)\}}$ is called Silhouette-Coefficient (SC) of sample L_k.

Based on Definition 1, we can calculate the SC of the entire dataset as $S = \frac{1}{n}\sum_{k=1}^{n} S(k)$, where n represents the size of dataset. It can be seen from Definition 1 that the larger the S, the better the clustering effect.

Definition 2 ([21]). Suppose that n is the number of clusters, C_i is the center of the ith class, and E_i is the average Euclidean distance from all load data in class i to the cluster center C_i. Then, $DBI = \frac{1}{n}\sum_{i=1}^{n} \max_{j\neq i}(\frac{E_i+E_j}{dist(C_i,C_j)})$ is called Davies–Bouldin Index (DBI).

It can be seen from Definition 2 that the smaller the DBI, the better the clustering effect. Meanwhile, $RMSE$, MAE, $MAPE$, and R^2 are used to evaluate the performance of load forecasting model.

5.3. Experimental analysis of IK-means

For all experimental datasets, from Table 2, we can see that when the number of cluster is 3, the average running time based on IK-means is better than the traditional K-means. Especially for Chengnan 63# dataset, compared with the traditional K-means, the average running time of IK-means is reduced by about 63.87%.

Table 3 shows that the average Euclidean distance between the all cluster centers based on the IK-means is significantly greater than that of K-means when the number of cluster is 3. From Table 3, compared with K-means, it can be seen that the average Euclidean distance between the cluster centers based on the IK-means is improved by 40.3%, 41.58% and 0.35%, respectively. Experimental results show that the cluster centers obtained based on the IK-means are scattered, which makes the difference between each class larger.

To further illustrate the advantage of IK-means, we also compared DBI and SC values between IK-means, K-means, K-means++ [22], Agglomerative Clustering(AC) [23] and Gaussian Mixture(GM) [24]. Figs. 2 and 3 compare the SC and DBI values of the five clustering algorithms, respectively.

From Fig. 2, we can see that for training datasets in Table 1, under different number of clusters, the SC of the IK-means is obviously superior to other clustering algorithms. When number of clusters are 3, 4, 5, 6, respectively, compared with K-means++, K-means, AC and GM, the SC of IK-means is increased by 12.56%, 14.65% and 7.99%, respectively.

Fig. 3 indicates that when the number of clusters is 3 and 5, for all training datasets, the DBI value of IK-means is the smaller than other algorithms. And from Fig. 3, we can also see that when the number of clusters is 4, for the EUNITE training dataset, the DBI of IK-means is smaller than other clustering algorithms; and when the number of clusters is 6, in all clustering algorithms, the DBI of IK-means is the smallest for EUNITE and Chengnan 63# training dataset. The experimental results demonstrate that IK-means has higher clustering efficiency and better clustering effect.

Table 3
The comparison of the average Euclidean distance.

Algorithm	Dataset	#Instance	#Cluster	The average Euclidean distance
IK-means **(Ours)**	Hexing 2#	50 114 72	3	**258.82**
	Chengnan 63#	38 115 75	3	**124.82**
	EUNITE	223 329 178	3	**359.61**
K-means	Hexing 2#	115 44 77	3	154.52
	Chengnan 63#	139 24 69	3	72.92
	EUNITE	231 324 175	3	358.35

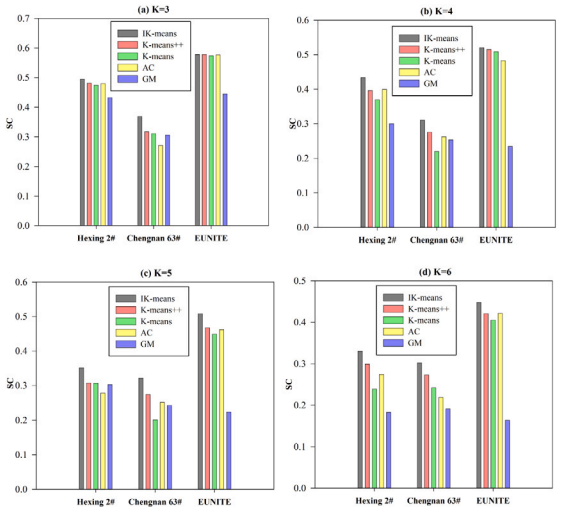

Fig. 2. Comparison of SC with different number of clusters.

5.4. Experimental analysis of LFEM-KLS

To verify the performance of LFEM-KLS, we first compare the load forecasting models based on GEP, GEP-LS and GEP-LFEM-KLS. Tables 4 and 5 compare statistic metrics of load forecasting models based on GEP, GEP-LS and GEP-LFEM-KLS, respectively, for all experimental datasets.

From Table 4, we can find that MAPE, RMSE, MAE of load forecasting model based on GEP-LFEM-KLS is obviously better than that of other algorithms. Compared with GEP and GEP-LS, R^2 of load forecasting model based on GEP-LFEM-KLS is increased by 7.33%, 6.22%; 12.78%, 9.65%; 1.23%, 1.21, respectively. The experimental results indicate that the IK-means algorithm can eliminate the noise in the load forecasting error data such that load forecasting accuracy can be greatly improved. Also, we find that even if the noise in the load forecasting error data is not eliminated, accuracy of combination load forecasting model based on GEP-LS is better than that of single load forecasting model based on GEP.

From Table 5, we can find that for all testing datasets, compared with GEP and GEP-LS, accuracy of load forecasting model based on GEP-LFEM-KLS is the best. Compared with GEP and GEP-LS, R^2 of load forecasting model based on GEP-LFEM-KLS is increased by 7.33%, 6.22%; 12.78%, 9.65%; 1.23%, 1.21%, respectively. Moreover, for all the testing datasets, compared with the GEP and GEP-LS, the MAPE, RMSE and MAE of the load forecasting model based on the GEP-LFEM-KLS are the smallest. This is because the noise in the load forecast error data has a large impact on the forecast accuracy.

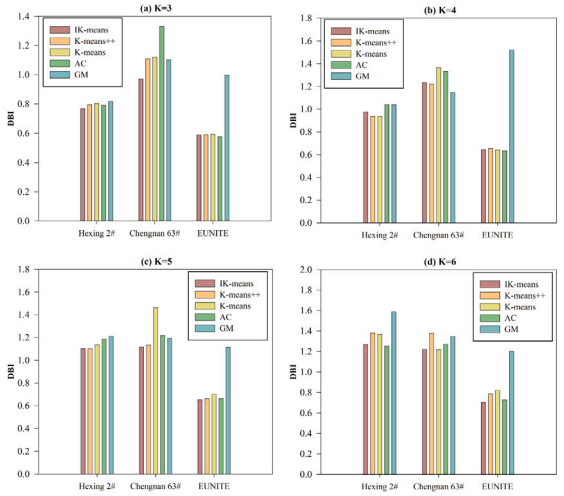

Fig. 3. Comparison of DBI with different number of clusters.

Table 4
Comparison of load forecasting model for training datasets.

Training dataset	Metric	GEP	GEP-LS	GEP-LFEM-KLS (Ours)
Hexing 2#	MAPE	0.1027	0.1017	**0.0231**
	RMSE	9.518	9.423	**2.22**
	MAE	7.319	7.246	**1.742**
	R^2	0.923	0.934	**0.996**
Chengnan 63#	MAPE	0.1042	0.1031	**0.006**
	RMSE	12.2623	12.1397	**0.6726**
	MAE	9.5098	9.4147	**0.5287**
	R^2	0.8719	0.9031	**0.9996**
EUNITE	MAPE	0.0223	0.0221	**0.0143**
	RMSE	14.4401	14.2957	**9.3028**
	MAE	11.2953	11.1824	**7.4371**
	R^2	0.9788	0.979	**0.991**

Table 5
Comparison of load forecasting model for testing datasets.

Testing dataset	Metric	GEP	GEP-LS	GEP-LFEM-KLS (Ours)
Hexing 2#	MAPE	0.139	0.2037	**0.1399**
	RMSE	12.9326	21.51	**12.7206**
	MAE	10.1615	18.0437	**10.203**
	R^2	0.3132	0.2087	**0.3355**
Chengnan 63#	MAPE	0.1355	0.2355	**0.1123**
	RMSE	12.5342	22.5342	**11.5153**
	MAE	11.2582	18.2582	**9.5776**
	R^2	0.2074	0.192	**0.331**
EUNITE	MAPE	0.0183	0.0183	**0.0133**
	RMSE	13.9355	14.4026	**10.9587**
	MAE	11.3855	11.8331	**8.3444**
	R^2	0.7964	0.6988	**0.8741**

5.5. Experimental analysis of QCLF

To manifest the performance of QCLF, in this paper, QCLF is compared with the other load forecasting models based on LSTM [12], LS [25], BP [26], respectively. The performance comparison between QCLF and other three load forecasting models is shown in Tables 6 and 7, respectively. Figs. 4 and 5 show model value and real value based on the above four models for training and testing datasets, respectively.

From Table 6, we can see that MAPE, RMSE, MAE and R^2 of QCLF is superior to that of load forecasting model based on LS, BP and LSTM for training datasets. Specifically, compared with load forecasting model based on LS, BP and LSTM, MAPE, RMSE and MAE of QCLF are reduced by 79.77%, 79.93%, 79.75%; 95.43%, 95.49%, 95.45% and 53.87%, 52.96%, 53.58%, respectively. This

Table 6
Comparison of performance between different load forecasting models for training datasets.

Training dataset	Index	QCLF (Ours)	LS	BP	LSTM
Hexing 2#	MAPE	**0.0231**	0.1142	0.1003	0.1075
	RMSE	**2.22**	11.0621	10.3589	9.9195
	MAE	**1.742**	8.6026	7.2039	7.8212
	R^2	**0.996**	0.896	0.913	0.9202
Chengnan 63#	MAPE	**0.006**	0.1172	0.1065	0.1313
	RMSE	**0.6726**	12.6046	12.589	14.909
	MAE	**0.5287**	10.1508	9.5962	11.625
	R^2	**0.9996**	0.8646	0.8649	0.8106
EUNITE	MAPE	**0.0143**	0.0294	0.031	0.0201
	RMSE	**9.3028**	17.8322	19.7761	13.3766
	MAE	**7.4371**	15.2403	16.0199	10.3445
	R^2	**0.991**	0.9676	0.9602	0.9818

Table 7
Comparison of performance between different load forecasting models for testing datasets.

Testing dataset	Index	QCLF (Ours)	LS	BP	LSTM
Hexing 2#	MAPE	**0.139**	0.1508	0.1293	0.1838
	RMSE	**12.9326**	15.7051	12.5609	18.7941
	MAE	**10.1615**	11.9408	11.9438	14.8192
	R^2	**0.3132**	0.0129	0.3021	0.0012
Chengnan 63#	MAPE	0.1355	0.1429	0.1417	0.1255
	RMSE	**12.5342**	12.9215	13.6523	14.7829
	MAE	11.2582	11.7054	11.9739	10.8943
	R^2	**0.2074**	0.1576	0.0597	0.0153
EUNITE	MAPE	0.0183	0.0221	0.0172	0.0156
	RMSE	13.9355	15.8813	13.7958	11.9168
	MAE	11.3855	13.9794	10.7576	9.7427
	R^2	0.7964	0.7355	0.8004	0.8511

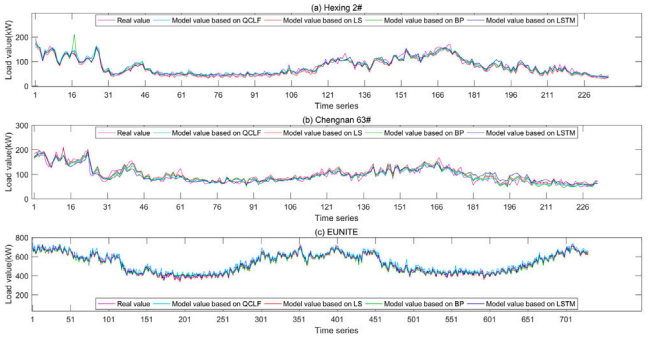

Fig. 4. Comparison between model values and real values of four algorithms for all training datasets.

means that error of QCLF is obvious less than that of load forecasting model based on LS, BP and LSTM. And, in contrast to load forecasting model based on LS, BP and LSTM, R^2 of QCLF is increased by 10.02%, 8.31%, 7.59%; 13.51%, 13.48%, 18.91% and 2.36%, 3.11%, 0.93%, respectively. It demonstrates that QCLF can better fit the load data.

From Table 7, it is found that for Hexing 2# and Chengnan 63# testing datasets, QCLF is superior to load forecasting model based on LS, BP and LSTM. However, for EUNITE testing dataset, the various indexes of QCLF are not as good as load forecasting model based on BP and LSTM. The errors between MAPE, RMSE, MAE and R^2 of QCLF and load forecasting model based on BP, LSTM are 0.001, 0.14, 0.63, 0.004; 0.003, 2.02, 1.64, 0.055, respectively.

And, it can be analyzed from Fig. 4 that compared with load forecasting model based on LS, BP and LSTM, the maximum and minimum errors between the model value of QCLF and the real value are the smallest, which are 24.2359, 0.0001; 36.5175, 0.046; 58.7922, 0.0497; 28.1526, 0.0256, respectively.

Fig. 5 illustrates that in contrast to load forecasting model based on LS, BP and LSTM, the maximum and minimum errors between the model value of QCLF and the real value are also the smallest, which are 18.2157, 0.0065; 19.7071, 1.0688; 29.5954, 0.0113; 19.2157, 0.0156, respectively, for all testing datasets.

We know that LS-based load forecasting requires the construction of a mathematical model in advance, which relies entirely on a priori knowledge. And the BP and LSTM-based load prediction model is a black box. We do not know the specific load forecasting

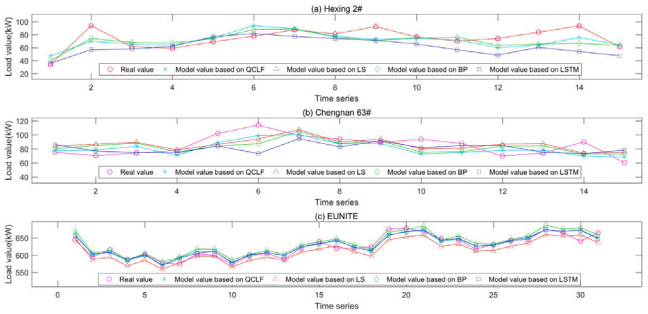

Fig. 5. Comparison between model values and real values of four algorithms for all testing datasets.

Table 8
Mathematical model of load forecasting based on QCLF.

Dataset	Mathematical model of load forecasting
Hexing 2#	$-5.082 + 0.098x_0 + 0.771x_1 - 0.012x_2 + 0.122x_3 + 0.218x_4 + 0.339x_5$
Chengnan 63#	$-0.118 + 0.185x_0 + 0.011x_1 + \frac{14.591x_1}{-0.557+x_0} + 0.203x_2 - \frac{0.213x_1}{1.459x_1 - x_3}$ $- \frac{6.18+x_3}{-0.902+0.132x_3} + 0.0295x_3 + 0.195x_4 + 0.984x_5 - \frac{21.619x_5}{x_2}$
EUNITE	$-7.362 + 0.043x_0 + 0.018x_1 + 0.092x_2$ $+0.57x_3 - 0.357x_4 + 0.613x_5$

model at all. However, QCLF model does not rely on any prior knowledge, and the load forecasting mathematical model is completely discovered from the load dataset. For experimental datasets, mathematical formulas of load forecasting based on QCLF are shown in Table 8. From Table 8, we can also find that in addition to discovering load forecasting mathematical models from load datasets without any prior knowledge, QCLF also has a certain feature reduction ability.

6. Conclusions

To address the problem that the traditional single load forecasting model has low accuracy and is sensitive to load forecasting errors, a quantitative combination load forecasting optimization algorithm based on the load forecasting error model is proposed. Specifically, main work of this paper mainly include: (1) to improve the influence of the error data in the existing single load forecasting model on the later combination load forecasting, a load forecasting error model based on an improved K-means and the least square method (LFEM-KLS) is proposed; (2) based on LFEM-KLS, this paper integrates the load forecasting function model mining algorithm based on gene expression programming to construct a quantitative combination load forecasting model (QCLF) that considers the impact of load forecasting error; (3) experiments are conducted on three load datasets, and results demonstrate that compared with the existing load forecasting algorithms, the QCLF model proposed has greater advantages in MAPE, RMSE, MAE and R^2, thus showing better prediction accuracy.

CRediT authorship contribution statement

Song Deng: Conceptualization, Methodology, Software, Writing – original draft, Formal analysis, Funding acquisition. **Fulin Chen:** Conceptualization, Methodology, Software, Writing – original draft, Formal analysis, Funding acquisition, Resources, Investigation. **Di Wu:** Conceptualization, Methodology, Software, Writing – original draft, Formal analysis, Funding acquisition, Data curation, Writing – review & editing. **Yi He:** Data curation, Writing – review & editing, Investigation, Project administration, Supervision. **Hui Ge:** Conceptualization, Methodology, Software, Writing – original draft, Formal analysis, Funding acquisition. **Yuan Ge:** Data curation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The subject is supported by the National Natural Science Foundation of P. R. China (No. 51977113, 62176070, 52077106), Key Laboratory of Advanced Perception and Intelligent Control of High-end Equipment, Ministry of Education (No. GDSC202003/TK220008).

References

[1] Niri MF, Dinh TQ, Yu TF, Marco J, Bui TMN. State of power prediction for lithium-ion batteries in electric vehicles via wavelet-Markov load analysis. IEEE Trans Intell Transp Syst 2020;22(9):5833–48.
[2] Li R, Chen X, Balezentis T, Streimikiene D, Niu Z. Multi-step least squares support vector machine modeling approach for forecasting short-term electricity demand with application. Neural Comput Appl 2021;33:301–20.
[3] Heydari A, Nezhad MM, Pirshayan E, Garcia DA, Keynia F, De Santoli L. Short-term electricity price and load forecasting in isolated power grids based on composite neural network and gravitational search optimization algorithm. Appl Energy 2020;277:115503.
[4] Wu EQ, Hu D, Deng P-Y, Tang Z, Cao Y, Zhang W-M, Zhu L-M, Ren H. Nonparametric bayesian prior inducing deep network for automatic detection of cognitive status. IEEE Trans Cybern 2020;51(11):5483–96.
[5] Ferreira C. Algorithm for solving gene expression programming: a new adaptive problems. Complex Syst 2001;13(2):87–129.
[6] Buzna L, De Falco P, Ferruzzi G, Khormali S, Proto D, Refa N, Straka M, van der Poel G. An ensemble methodology for hierarchical probabilistic electric vehicle load forecasting at regular charging stations. Appl Energy 2021;283:116337.
[7] Wang Y, Chen Q, Sun M, Kang C, Xia Q. An ensemble forecasting method for the aggregated load with subprofiles. IEEE Trans Smart Grid 2018;9(4):3906–8.
[8] Dabbaghjamanesh M, Moeini A, Kavousi-Fard A. Reinforcement learning-based load forecasting of electric vehicle charging station using Q-learning technique. IEEE Trans Ind Inf 2020;17(6):4229–37.
[9] Bandara K, Bergmeir C, Hewamalage H. LSTM-MSNet: Leveraging forecasts on sets of related time series with multiple seasonal patterns. IEEE Trans Neural Netw Learn Syst 2020;32(4):1586–99.
[10] Xie K, Yi H, Hu G, Li L, Fan Z. Short-term power load forecasting based on Elman neural network with particle swarm optimization. Neurocomputing 2020;416:136–42.
[11] Zhang J, Yan J, Liu Y, Zhang H, Lv G. Daily electric vehicle charging load profiles considering demographics of vehicle users. Appl Energy 2020;274:115063.
[12] Li J, Deng D, Zhao J, Cai D, Hu W, Zhang M, Huang Q. A novel hybrid short-term load forecasting method of smart grid using mlr and lstm neural network. IEEE Trans Ind Inf 2020;17(4):2443–52.
[13] Si Z, Yu Y, Yang M, Li P. Hybrid solar forecasting method using satellite visible images and modified convolutional neural networks. IEEE Trans Ind Appl 2020;57(1):5–16.
[14] Gilanifar M, Wang H, Ozguven EE, Zhou Y, Arghandeh R. Bayesian spatiotemporal gaussian process for short-term load forecasting using combined transportation and electricity data. ACM Trans Cyber-Phys Syst 2019;4(1):1–25.
[15] Nie Y, Jiang P, Zhang H. A novel hybrid model based on combined preprocessing method and advanced optimization algorithm for power load forecasting. Appl Soft Comput 2020;97:106809.
[16] Zhang H, Yang Y, Zhang Y, He Z, Yuan W, Yang Y, Qiu W, Li L. A combined model based on SSA, neural networks, and LSSVM for short-term electric load and price forecasting. Neural Comput Appl 2021;33(2):773–88.
[17] Panapakidis IP, Skiadopoulos N, Christoforidis GC. Combined forecasting system for short-term bus load forecasting based on clustering and neural networks. IET Gener Transm Distrib 2020;14(18):3652–64.
[18] Hartigan JA, Wong MA. Algorithm AS 136: A k-means clustering algorithm. J R Stat Soc C 1979;28(1):100–8.
[19] Deng S, Chen F, Dong X, Gao G, Wu X. Short-term load forecasting by using improved GEP and abnormal load recognition. ACM Trans Internet Technol (TOIT) 2021;21(4):1–28.
[20] Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987;20:53–65.
[21] Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1979;(2):224–7.
[22] Yoder J, Priebe CE. Semi-supervised k-means++. J Stat Comput Simul 2017;87(13):2597–608.
[23] Murtagh F, Legendre P. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? J Classification 2014;31(3):274–95.
[24] Vo B-N, Ma W-K. The Gaussian mixture probability hypothesis density filter. IEEE Trans Signal Process 2006;54(11):4091–104.
[25] Dudek G. Pattern-based local linear regression models for short-term load forecasting. Electr Power Syst Res 2016;130:139–47.
[26] Xiao Z, Ye S-J, Zhong B, Sun C-X. BP neural network with rough set for short term load forecasting. Expert Syst Appl 2009;36(1):273–9.

Song Deng received the Ph.D.degree in information network from Nanjing University of Posts and Telecommunications in 2009. He is the Associate Professor of Nanjing University of Posts and Telecommunications. His research interests include data security, data mining and knowledge engineering.

Fulin Chen received the M.S. degree in Automation control from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2021. He is currently working toward the Ph.D.degree in computer science from the SouthEast University, Nanjing, China. His research interests include data mining, multi-agent and load forecasting.

Di Wu received his Ph.D. degree from the Chongqing Institute of Green and Intelligent Technology (CIGIT), Chinese Academy of Sciences (CAS), China in 2019. He is now an Associate Professor of CIGIT. His research interests include machine learning and data mining.

Xi He received his Ph.D. from the University of Louisiana at Lafayette, Louisiana, USA, in 2020. He is an Assistant Professor of Computer Science at Old Dominion University, Virginia, USA. His research interests lie broadly in data mining, artificial intelligence, and optimization theory and specifically in online learning, data stream analytics, and graph learning.

Hui Ge received the Ph.D. degree from Nanjing University of Posts and Telecommunications in 2018, Nanjing, China. He is a lecture of Nanjing University of Posts and Telecommunications. His research interests include networked control systems and CPS security control.

Yuan Ge received Ph.D. degrees from University of Science and Technology of China in 2011. He is a professor in the School of Electrical Engineering at Anhui Polytechnic University, China. His research interests include smart grid, energy internet, and networked control systems.