Received April 19, 2021, accepted April 28, 2021, date of publication May 3, 2021, date of current version May 10, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3077028

Predictive Analytics for Octane Number: A Novel Hybrid Approach of KPCA and GS-PSO-SVR Model

BAOSHENG LI[1] AND CHUANDONG QIN[1,2]

[1]School of Mathematics and Information Science, North Minzu University, Yinchuan 750021, China
[2]Ningxia Key Laboratory of Intelligent Information and Big Data Processing, Yinchuan 750021, China

Corresponding author: Chuandong Qin (qinchuandong123@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 62066001, and in part by the National Natural Science Youth Science Foundation of China under Grant 61907012.

ABSTRACT Octane number is the most important indicator of reflecting the combustion performance, and a great deal of research has been devoted to improving it. In this paper, a new analytical framework is proposed to predict octane number, kernel principal component analysis (KPCA) is used to reduce the dimension of the variables in the process of Fluid Catalytic Cracking (FCC), support vector regression (SVR) is used to construct the gasoline octane number prediction model and the particle swarm optimization algorithm (PSO) is used to select the optimal combination of parameters for the model. The experiments show that the octane number can be improved under a given production environment with a guaranteed desulfurization effect of gasoline products. Furthermore, several key attributes that have a significantly positive or negative correlation with the improvement of gasoline product quality are identified through computing the feature score. The findings can help engineers adjust operational variables to obtain a series of high-quality products.

INDEX TERMS Gasoline octane number, kernel principal component analysis, support vector regression, particle swarm optimization.

I. INTRODUCTION

With the increasing awareness of environmental protection, gasoline cleaning technologies have received close attention from people all over the world. From the perspective of the gasoline pool, Fluid Catalytic Cracking (FCC) has been the main harmonic component of gasoline. To effectively utilize the heavy oil resource, China has vigorously developed the light-weight technology with FCC as the core, converting the heavy oil into gasoline, diesel oil, and low carbon olefins. More than 70% of the gasoline is produced by FCC, which brings approximately 95% of the sulfur and olefins in the refined gasoline. Therefore, the main tendency is to reduce the content of olefins and aromatics while keeping sulfur content less than or equal to 10mg/kg [1]. Especially, the major research institutions of the petroleum and petrochemical companies in the world have invested heavily in the development and improvement of FCC technologies. Likewise, building on the full implementation of CHN IV in 2014, China has brought forward the implementation of CHN V from 1 January 2017 to 1 January 2018 and will implement CHN VI-A and VI-B on 1 January 2019 and 1 January 2023, respectively, which are presented in Table 1.

TABLE 1. The gasoline standard, which is organized from [1], [2].

Standards	Octane	Sulfur (μg/g) $\not>$	Benzene (%) $\not>$	Aromatic (%) $\not>$	Olefin (%) $\not>$
CHN IV	90-97	50	1	40	28
CHN V	89-95	10	1	40	24
CHN VI-A	89-95	10	0.8	35	18
CHN VI-B	89-95	10	0.8	35	15
EUR V	95	10	1	35	18
EUR VI	95	10	1	35	18
JPN	89-96	10	1	–	–
WWFC V	95	10	1	35	10

Octane number is the most influential indicator of the combustion performance of gasoline, which is reduced in the process of desulfurization and olefin reduction. Every unit reduction in octane number is equivalent to the loss of

The associate editor coordinating the review of this manuscript and approving it for publication was Juan A. Lara.

about $22.38 per ton. Due to the constraints of measurement technology and complex operating conditions, it is difficult to obtain the octane number in real-time. Given this situation, it is valuable to develop an accurate prediction model for octane number in the gasoline refining process. There are two main areas of research on measuring the octane number, one is the quantitative structure-property correlation studies on octane number of gasoline components, including group contribution method, topological indices [2]. The other is to carry out instrumental analysis of each gasoline component, component absorbance, and content of each component, and use the collected parameters to construct a predictive model for gasoline octane number. Based on the different relationships between the attributes of the dataset, the methods can be divided into linear and non-linear regression models. For linear regression prediction methods, there are several analytical techniques, including the combination of principal component regression and partial least square regression [3], [4], infrared spectroscopy, near-infrared (NIR) spectroscopy analysis, Raman spectroscopy analysis, and nuclear magnetic resonance spectroscopy analysis [5]–[8]. If the operating variables have a high degree of nonlinearity and coupling, non-linear functions can be introduced for non-linear regression prediction, such as artificial neural network (ANN) [9]–[13], support vector regression (SVR) [14], deep learning (DL) [15] etc. They have strong capabilities of processing non-linear systems and are widely implemented in chemical engineering, biological sciences, image processing, and other fields.

Support vector machine (SVM) is based on the classic theories of Vapnik-Chervonenkis (VC) dimension and structural risk minimization (SRM) principle. SVM works well in tackling small sample problems, non-linear problems, and high-dimensional pattern recognition problems [16]. The support vector regression is popularized by SVM and used for regression prediction problems which have the advantages of strong generalization ability, high prediction accuracy, and prevention of overfitting [17], [18]. Moreover, it has been applied to a certain extent in chemometrics and chemical analysis. By comparing SVR to partial least square (PLS) for spectral regression applications, U. Thissen *et al.* [19] concluded that SVR not only outperforms PLS, but also is more robust to non-linear effects induced by the external environment. Roman Balabin *et al.* [13] raised and answered the question of how far SVM-based technologies can replace ANN-based approaches in real-world (industrial/scientific) applications. The results demonstrated that SVR and LS-SVM are comparable to neural networks in terms of prediction accuracy, but the former has higher robustness. After processing the gasoline samples with principal component analysis (PCA) and fuzzy C-means (FCM) algorithms, K. Brudzewski *et al.* [14] introduced hybrid neural networks and support vector machines to obtain the predicted octane numbers. With the continuous exploration of SVR in chemometrics, many scholars have focused on combining various feature processing techniques and optimization algorithms with SVR [21]–[23], [25], [26] to improve prediction accuracy and convergence speed. For example, F. Melgani *et al.* [21] proposed a novel classification approach based on particle swarm optimization (PSO) to improve the accuracy and the generalization performance of SVM. J.C.L. Alves *et al.* [22] applied SVM to NIR spectroscopy data and introduced the new ideas for optimizing SVR models based on genetic algorithm (GA). Moreover, the cross-validation (CV) and grid search procedure are used to tune the parameters of the SVR in [23]. Qiao *et al.* [24] proposed five models, namely MLP-based hybrid models of IWO-MLP, DA-MLP, ES-MLP, GA-MLP, and ICA-MLP, which can well recognize the local minima help improving better prediction of monthly natural gas consumption. Also, M.S. AL-Musaylh *et al.* [25] presented the support vector regression based on the two-phase particle swarm and obtained high predictive utility. In addition, a novel combination of the improved genetic algorithm (IGA) and PSO was employed to select the optimal parameters of the SVM discriminators for specifying the quality of Dian Hong black tea in [26].

However, there is scarce research on the modeling and analysis of operational variables and material properties in FCC. Furthermore, it is incapable to obtain satisfactory results from the traditional models because of its slow response to process optimization. Therefore, KPCA-SVR combining grid search and PSO algorithm to optimize parameters is proposed for the octane number prediction in this paper. This study aims to predict the octane number in real-time and identify the key attributes that can contribute to the octane number, thereby helping the engineers adjust the operating variables to improve the octane number while maintaining the desulfurization effect.

The paper is structured as follows. Part 2 briefly introduces the background of KPCA and summarizes the basic theory of SVR. The third part presents the grid search strategy and the PSO optimization algorithm, which are applied to acquire the optimal parameters of SVR. Part 4 reports the experimental design and process in detail and gives the main experimental results and analysis. Part 5 shows the main conclusions and makes recommendations for future research methods.

II. MATERIALS AND METHOD

A. KERNEL PRINCIPAL COMPONENT ANALYSIS

Kernel principal component analysis (KPCA) is a powerful tool for extracting information from high-dimensional non-linear datasets, which are readily performed for the pattern recognition in [27]. By the implementation of ''Kernel Trick'', the non-linear variables are mapped into a high-dimensional feature space, and PCA is performed by using iterative algorithms that estimate principal components in this high-dimensional space.

Given a set of centered data,

$$X = (x_1, \ldots, x_m)^T, x_i \in R^n, \quad i = 1, \ldots, m. \tag{1}$$

where $\sum_{i=1}^{m} x_i = 0$, $\frac{1}{m}\sum_{i}^{m}(x_i - \bar{x})^2 = 1$. Let us consider a feature space H, related to the input domain by a map,

$$\Phi : R^n \to H, \quad x \mapsto X \tag{2}$$

which is probably non-linear. Again, $\sum_{i}^{m} \Phi(x_i) = 0$, the covariance matrix takes the form in H,

$$C = \frac{1}{m}\sum_{i=1}^{m} \Phi(x_i)\Phi(x_i)^T = \frac{1}{m}\Phi(X)\Phi(X)^T \tag{3}$$

Let H be an infinite-dimensional feature space, and we think of $\Phi(x_i)\Phi(x_i)^T$ as the linear operator in H which maps $X \in H$ to $\Phi(x_i)(\Phi(x_i) \cdot X)$. Then, we have to find eigenvalues $\lambda \geq 0$, and eigenvectors $V \in H$ satisfying,

$$\lambda V = CV \tag{4}$$

where all solutions V with $\lambda \neq 0$ lie in the span of $\Phi(x_1), \ldots,$
$\Phi(x_m)$. We can get the set of equations,

$$\lambda \Phi(x_i)^T V = \Phi(x_i)^T CV, \quad i = 1, \ldots, m \tag{5}$$

and there exist coefficients α_i,

$$V = \sum_{i}^{m} \alpha_i \Phi(x_i) = \Phi(x)\alpha, \quad i = 1, \ldots, m \tag{6}$$

Let $\lambda_i, \alpha_i, i = 1, \ldots, m$ denote the eigenvalues of K and the corresponding complete set of eigenvectors, respectively, with λ_j being the first nonzero eigenvalue. We normalize $\alpha_j, \ldots, \alpha_m$ by requiring that the corresponding vectors in H be normalized.

$$V_p^T V_p = 1, \quad p = j, \ldots, m. \tag{7}$$

$$1 = \alpha^T \Phi(x)^T \Phi(x)\alpha = \alpha^T K \alpha = \lambda_p \alpha^T \alpha \tag{8}$$

For the aims of principal component extraction, we need to compute the projections onto the eigenvectors V_p in H $(p = j, \ldots, m)$. Let x be a test point, with an image $\Phi(x)$ in H, then,

$$\Phi(x_i)^T V_p = \sum_{i,j=1}^{m} \alpha_{pi} \Phi(x_i)^T \Phi(x_j) = \sum_{i=1}^{m} \alpha_{pi} K \tag{9}$$

can be called its non-linear principal components. In particular, the first principal component is obtained by solving,

$$\underset{\alpha}{\operatorname{argmin}} \sum_{i=1}^{m} \|\alpha_i K\|^2, \|\lambda_1 \alpha^T \alpha\| = 1 \tag{10}$$

Certainly, the amount of retained principal components is always confirmed by the cumulative percentage variance (CPV), which is defined as,

$$\mathrm{CPV}(p) = (\sum_{i=1}^{p} \lambda_i)/(\sum_{i=1}^{m} \lambda_i) \times 100\% \tag{11}$$

and the scores [28] can be calculated as,

$$S_p = \frac{1}{\sqrt{\Lambda_p}} V_p^T K \tag{12}$$

where S_p can be called the principal component scores.

B. SUPPORT VECTOR REGRESSION

Support vector regression is an extension of the support vector machine, which applies the support vectors to regression functions for solving non-linear regression problems [29]. The target of SVR is to obtain a hyperplane for fitting multidimensional input vectors to output values. And the outcome is used to predict the unknown output values. Given a set of data (x_i, y_i), $x_i \in R^n$, $y_i \in R$, $i = 1, \ldots, l$, the decision function f is defined in (13).

$$f(x) = w^T \Phi(x) + b \tag{13}$$

where Φ represents non-linear mapping, and x is mapped into a high-dimensional feature space from low-dimensional space, while w, b are slope and intercept of the hyperplane, respectively. By introducing an ε-insensitive region around the function, called the ε-tube and shown in (14),

$$|y_i - (w^T \Phi(x_i) + b)| \leqslant \varepsilon \tag{14}$$

the sparsity of SVM can be preserved, which means a small number of support vectors can respresent the decision function f. The ε-insensitive loss function is formally defined in (15), where $\varepsilon > 0$. When the difference between the observed value y and the predicted value $f(x)$ at the sample point x does not exceed a given ε, the predicted value is considered to be void of loss. In other words, the loss is 0 when x in the ε-tube, whereas a loss occurs only when x is outside this tube.

$$c(x, y, f(x)) = \max\{0, |y - f(x)| - \varepsilon\} \tag{15}$$

By introducing the penalty parameter C and the slack variables $\xi^{(*)} = (\xi_1, \xi_1^*, \ldots, \xi_l, \xi_l^*)^T$, the following convex quadratic programming problem is obtained.

$$R(w, \xi^{(*)}) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{16}$$

subject to,

$$(w^T \Phi(x_i) + b) - y_i \leq \varepsilon + \xi_i, \quad i = 1, 2, \ldots, l, \tag{17}$$

$$y_i - (w^T \Phi(x_i) + b) \leq \varepsilon + \xi_i^*, \quad i = 1, 2, \ldots, l, \tag{18}$$

$$\xi_i^{(*)} \geq 0, \quad i = 1, 2, \ldots, l \tag{19}$$

To obtain the solution of (17)-(19), the Lagrange function is introduced first, and the problem is transformed into the following dual problem (20), (21) which is convenient to solve.

$$W(\alpha, \alpha^*) = \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i, x_j) + \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{l} y_i(\alpha_i^* - \alpha_i) \tag{20}$$

subject to,

$$\sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0, 0 \leq \alpha_i, \alpha_i^* \leq C \tag{21}$$

If the optimal solution to the problem is $\bar{\alpha} = (\bar{\alpha}_1, \bar{\alpha}_1^*, \ldots, \bar{\alpha}_l, \bar{\alpha}_l^*)^T$, the decision function is constructed,

$$f(x) = \sum_{i=1}^{l} (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x_i, x_j) + \bar{b} \tag{22}$$

where the calculation method of $\bar{b}$ is to select $\bar{\alpha}_j$ or $\bar{\alpha}_k^*$. If $\bar{\alpha}_j$ is selected, the result $\bar{b}$ is,

$$\bar{b} = y_j - \sum_{i=1}^{l} (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x_i, x_j) + \varepsilon \tag{23}$$

or $\bar{\alpha}_k^*$ is selected, the result $\bar{b}$ is,

$$\bar{b} = y_k - \sum_{i=1}^{l} (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x_i, x_k) - \varepsilon \tag{24}$$

Here, three common kernel functions $K(x_i, x_j)$ are given in this paper. Generally speaking, functions satisfying the Mercer theorem can be used as kernel functions. where $\gamma = 1/\sigma^2$, r is a constant, d is a positive integer. In Table 2, γ, r, d are all built-in parameters of these kernels.

TABLE 2. Overview of common kernel functions.

Kernel name	Formula
Linear	$x_i^T x_j$
Polynomial	$(\gamma x_i^T x_j + r)^d$
Radial basis	$e^{-\gamma(\|x_i - x_j\|^2)}$

III. PARAMETER OPTIMIZATION

Since there are hyper-parameters in the SVR model, we need to optimize the parameters.

A. PARTICLE SWARM OPTIMIZATION ALGORITHM

Particle swarm optimization (PSO) is an effective global search algorithm, which is based on the theory of swarm intelligence and was first introduced by [30]. PSO can be described as: let the particle swarm fly in an n-dimensional searching space, a population composed of m particles $Z = \{z_1, \ldots, z_m\}$, the position of each particle $Z_i = \{z_{i1}, z_{i2}, \ldots, z_{in}\}$ represents a potential solution to an optimization question. Every particle can search for a new solution by constantly adjusting its position, and remember its optimal position p_{gd}. Besides, each particle has its velocity $V_i = \{v_{i1}, v_{i2}, \ldots, v_{in}\}$ and updates its velocity by (25), (26).

$$v_{id}^{(t+1)} = w v_{id}^{(t)} + c_1 r_1 (p_{id}^{(t)} - x_{id}^{(t)}) + c_2 r_2 (p_{gd}^{(t)} - x_{gd}^{(t)}) \tag{25}$$

$$z_{id}^{(t+1)} = z_{id}^{(t)} + v_{id}^{(t+1)} \tag{26}$$

where w is the inertia weight, and the ability to search for better solutions will be deprived when $w = 0$. r_1, r_2 are uniform random numbers in the interval (0,1). c_1 is called the "cognitive factor" since it stores in its best solution, and c_2 is called the "social factor" since it takes into account the global behavior [31]. When $c_1 = 0$, the rate of convergence accelerates, cognitive performance declines, and there is a tendency to fall into local extremes. Similarly, when $c_2 = 0$, the particles are in an independent relationship with each other, which makes it challenging to find the optimal solution. Therefore, we set the coefficients to the same value $c_1 = c_2 = 1.5$.

The selection of parameters has an essential influence on the performance and efficiency of the algorithm. In recent years, many scholars have proposed different strategies for the adjustment of inertia weight. For instance: Random Inertia Weight; Sigmoid Increasing Inertia Weight; Sigmoid Increasing Inertia Weight; Linear Decreasing Inertia Weight; Simulated Annealing Inertia Weight [31]–[34]. Since the linearly decreasing inertial weighting algorithm allows PSO to converge quickly and the performance is not sensitive to population size, we will choose this strategy to tune w.

$$w_g = w_{\max} - \frac{w_{\max} - w_{\min}}{iter_{\max}} \times g \tag{27}$$

where g represents the latest generation index of evolutionary generations,$iter_{\max}$ donates a predefined maximum number of allowable iterations, and $w_{\max}$, $w_{\min}$ are the greatest and least values of the inertia weight, which are usually set to 0.9 and 0.4, respectively.

B. KPCA-SVR OPTIMIZED BY GRID SEARCH AND PSO

From the above descriptions, it is abundantly apparent that, although PSO has a great advantage in local search capability, it is prone to local convergence to the extent that it lacks the ability to optimize globally. In contrast, the grid search has a high probability of finding the global optimum with a large search range and a small step size. But its computational complexity is expensive due to the utilization of cross-validation. Therefore, we propose an improved method for combining grid search strategy and PSO, and (25) is restructed as,

$$v_{id}^{(t+1)} = w v_{id}^{(t)} + (c_1 r_1 + c_2 r_2)(p_{id}^{(t)} - x_{id}^{(t)}) \tag{28}$$

where $p_{gd}^{(t)} = p_{id}^{(t)}, x_{gd}^{(t)} = x_{id}^{(t)}$. with this adjustment, the searching speed of particles will be accelerated, the searching range will be increased, and the probability of finding the optimal solution will be higher. Fig. 1 illustrates the detailed process of KPCA-based SVR with grid search and PSO algorithm.

IV. EXPERIMENT DESIGN

In this section, a brief experiment design is presented, including data collection and analysis, feature selection, performance metrics, and the detailed analytical framework.

A. DATA COLLECTION AND ANALYSIS

The experiments are based on historical data accumulated over 4 years (2017.04-2020.05) from the desulfurization unit of an FCC gasoline plant in China. This dataset has the following attributes: 7 feedstock properties; 2 raw adsorbent properties; 2 regenerated adsorbent properties; 2 product properties; and 354 operating variables, which are strongly

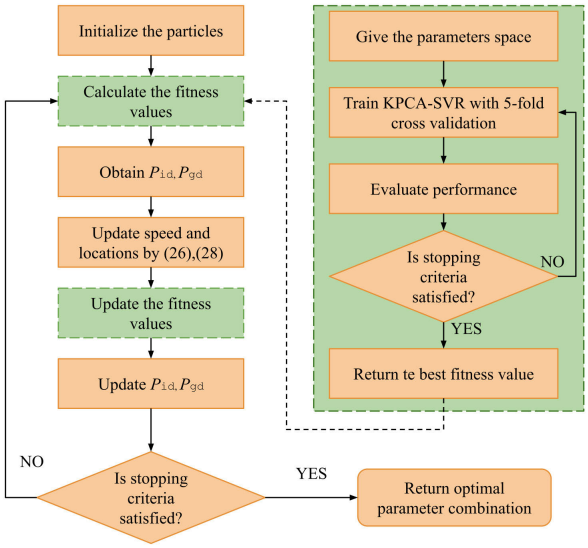


FIGURE 1. The flowchart of KPCA-SVR with grid search and PSO.

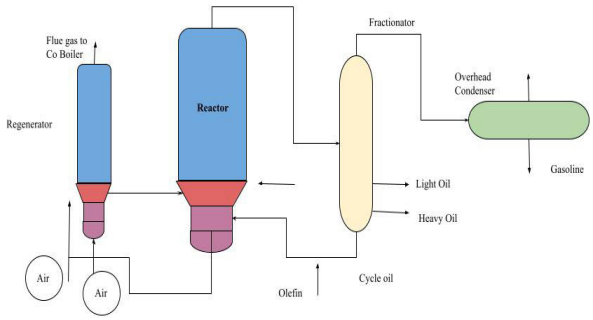


FIGURE 2. FCC unit.

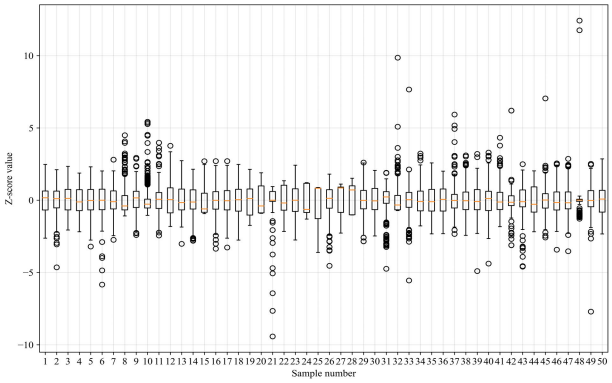


FIGURE 3. Box chart of the first 50 features, where the Y-axis is the Z-score value, the orange lines in the box are the mean value, and the circles are defined as the abnormal points.

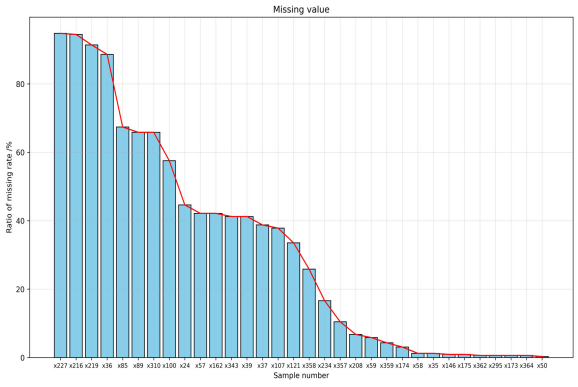


FIGURE 4. The distribution of missing values. By computing the missing rate, we can easily decide between direct deletion and the mean fill method.

non-linear and coupled. As we can see from Fig. 2, the FCC unit is committed to converting the heavy oil into gasoline, diesel oil, and low carbon olefins.

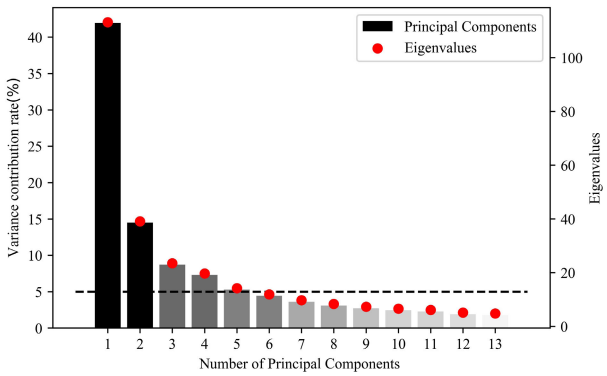
Data cleaning is the premise of data processing, including data formatting, anomaly detection and elimination, etc. We use Z-score ($\frac{x-\mu_x}{\sigma_x}$) method to standardize the original dataset. and then box chart is used for visualization, so that the outliers could be found intuitively.

As we can see from Fig. 3, the outliers are the circles, which will be removed directly.

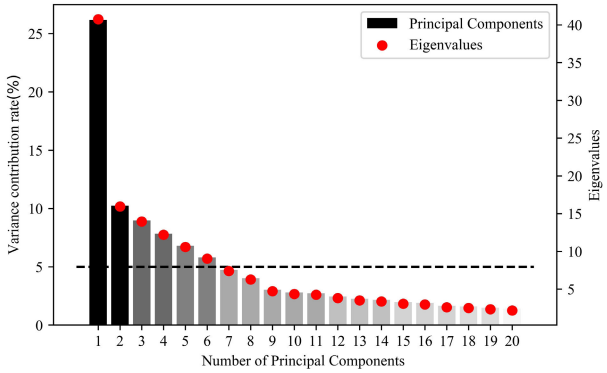
There are many possible reasons for missing values, such as equipment failure, power outage, parts damage, etc. The treatment of missing values is very important for model building, and the general treatment methods are usually, direct deletion, mean filling, regression filling, expert prediction, etc [35]. For the case of serious missing features (such as more than 60% missing rate), we choose to delete directly, and for the case of less missing features, we choose to use the mean fill method. The detailed distribtion of missing values is presented in Fig. 4.

B. FEATURE SELECTION

As a consequence of the complexity of the heavy oil cracking process and the diversity of equipment, the operational variables in the production process are non-linear and strongly coupled. Therefore, kernel principal component analysis is used to select the important variables. Fig. 5 demonstrates the ratio of explained variance for the first principal component of PCA and KPCA are 42.29%, 28.47%, respectively. However, the variance contribution rate of the latter principle components is lower, expressing that the first principal feature occupies key information. Furthermore, the number of exceeding 5% in Fig. 5(b) is more than Fig. 5(a), which indicates that KPCA requires more principal components to acquire key characteristics [28]. The total ratio of the variance is 96.05% in PCA, which explains that 13 principle components can embody the primary information of 367 attributes. Also, 95.51% in KPCA implies that 20 principal components represent the principal information of all the attributes. Especially, it is worth exploring whether the performance of the model can be improved by increasing the number of principal components. Experimentally, we find that the performance metrics of the model are not significantly improved when increasing the number of principal components of PCA, whereas the performance metrics of the model can be improved when the number of principal



(a) PCA



(b) KPCA

FIGURE 5. The ratio of the explained variance and the eigenvalue of the covariance matrix in PCA and KPCA.

components of KPCA is set to 20. Finally, the number of principal components is set to 13 and 20, respectively.

C. PERFORMANCE METRICS

R-squared (R^2) is donated as the coefficient of determination, which is always smaller than 1 and usually larger than 0. (29) describes the correlation between predictive variables and response variables. If R-squared approaches 0, it expresses all the dependent variables can not account for the predictive variables. Root Mean Square Error (RMSE) is calculated in (30), which measures the deviation between the values observed and the values estimated. Mean Absolute Percentage Error (MAPE) [36] is always employed as a loss function for regression models since it can interpret the relative error intuitively, and in the interval $[0, +\infty]$, defined in (31). The objective of GS-PSO is to improve the prediction accuracy and the generalization performance of the regressor. Therefore, RMSE is used as the fitness function.

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)^2} \tag{29}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2} \tag{30}$$

$$\text{MAPE} = \frac{100\%}{n}\sum_i \left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{31}$$

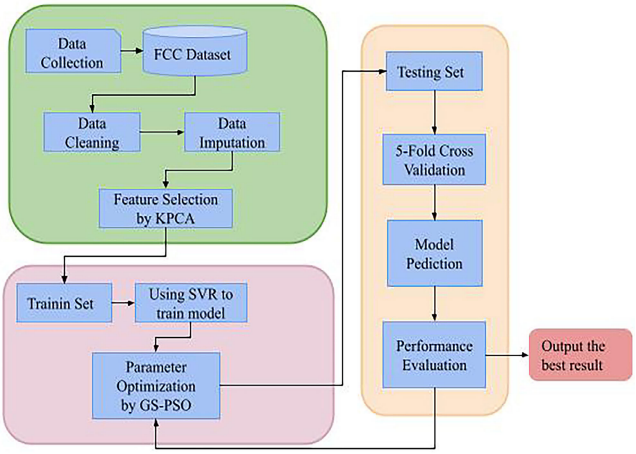


FIGURE 6. Schematic illustration of grid search and particle swarm optimization combined with SVR octane number prediction.

D. ANALYTICAL FRAMEWORK

As shown in Fig. 6, Numerical experiments will be carried out in the following aspects:

i. Training SVR with linear, polynomial, and RBF kernel functions.

ii. Performing PCA and KPCA dimensionality reduction techniques for preprocessing the raw data.

iii. Making a comparison between PCA-based SVR and KPCA-based SVR regarding the forecasting result.

iv. Optimizing the parameter set of KPCA-based SVR with RBF kernel function by the implementation of the hybrid of grid search and PSO.

v. Comparing KPCA-SVR (RBF) optimized by grid search and PSO with other non-linear regression methods, including ridge regression, PLS, Adaboost, BP neural network, and Long Short-Term Memory (LSTM) [15], [37].

All programs are processed in python 3.7 and run on Windows10, 8G memory, Intel® Core i5-8500CPU@ 3.00GHz. Note that the datasets can be found in http://dx.doi.org/10.17632/ y2hp539ww3.1, and the code of these experiments can be found in https://github.com/daishu-li/Predictive-analytics-for-octane-number.

V. RESULTS AND DISCUSSIONS

This numerical experiment starts with randomly splitting the raw data set into an 80% training set and a 20% test set. Subsequently, the default parameters provided by the *scikit-learn 0.23.* ($C = 1$, $d = 3$ and γ is calculated by (32)) are applied to construct the model [38].

$$\gamma = \frac{1}{m\sigma^2} \tag{32}$$

where m is the number of features, and σ^2 is the variance of samples.

A. MODEL TRAINING

The first step of model training is to select appropriate kernel parameters. We can see the esult of SVR with linear, polynomial, and RBF in Fig. 7.

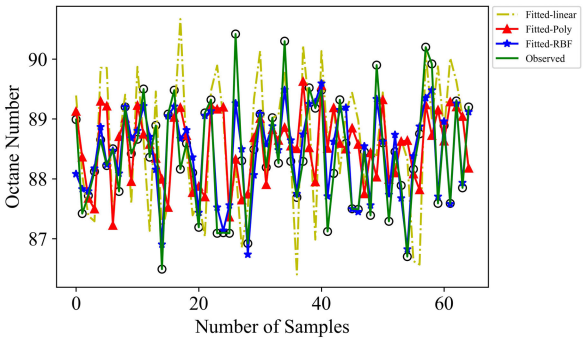


FIGURE 7. Comparison of predicted and observed octane number by introducing linear, polynomial, and RBF kernel functions to SVR, respectively.

We can see the comparison of predicted and observed octane number. In the first subgraph, the green circles are always outside or inside the red triangle, which indicates a large gap between the fitted and observed values, and the same goes for Fig. 7(b). Unlike the first two, Fig. 7(c) presents a smaller gap between the green circles and the red triangles. More specifically, Table 3 demonstrates that SVR with the RBF kernel achieves such best performance as the average deviation between the observed and fitted value is only 0.4212, and the minimal forecasting error compared with the other approaches. The underlying reason is its capability of achieving non-linear mapping.

TABLE 3. The results of each model with default parameters.

Model	Performance Metrics		
	R^2	RMSE	MAPE
Linear	0.7851	0.4392	0.3379
Polynomial	0.7350	0.4877	0.3729
RBF	**0.8024**	**0.4212**	**0.3162**

By comparing the two feature extraction techniques, we can determine which method has a higher feature extraction capability, thus improving the model prediction performance. Concerning the data after dimension reduction with PCA, the performances of SVR with RBF kernel are the most remarkable than linear and polynomial kernel functions, as shown in Table 4 and Fig. 8. It is apparent that this conclusion is consistent with the analysis in Table 3.

Similarly, KPCA-based SVR with RBF kernel has the most accurate forecasting capability compared to other kernel functions. Comparing all models above, KPCA-based SVR with RBF kernel performs best. The reason lies in the fact that KPCA has the characteristics of computing non-linear principal features by mapping the original input data to high-dimensional feature space, while PCA lacks the capability of handling non-linear features. In view of this case, we are dedicated to optimizing the parameters of KPCA-SVR by the implementation of the hybrid grid search and PSO algorithm.

TABLE 4. The performance metrics of PCA-SVR and KPCA-SVR with linear, polynomial, and RBF kernel functions.

Model	Performance Metrics		
	R^2	RMSE	MAPE
PCA-SVR(Linear)	0.7350	0.4878	0.4212
PCA-SVR(Polynomial)	0.6638	0.5494	0.4791
PCA-SVR(RBF)	**0.8616**	**0.3366**	**0.3052**
KPCA-SVR(Linear)	0.8024	0.4212	0.3574
KPCA-SVR(Polynomial)	0.7604	0.4637	0.3898
KPCA-SVR(RBF)	**0.8630**	**0.3349**	**0.2995**

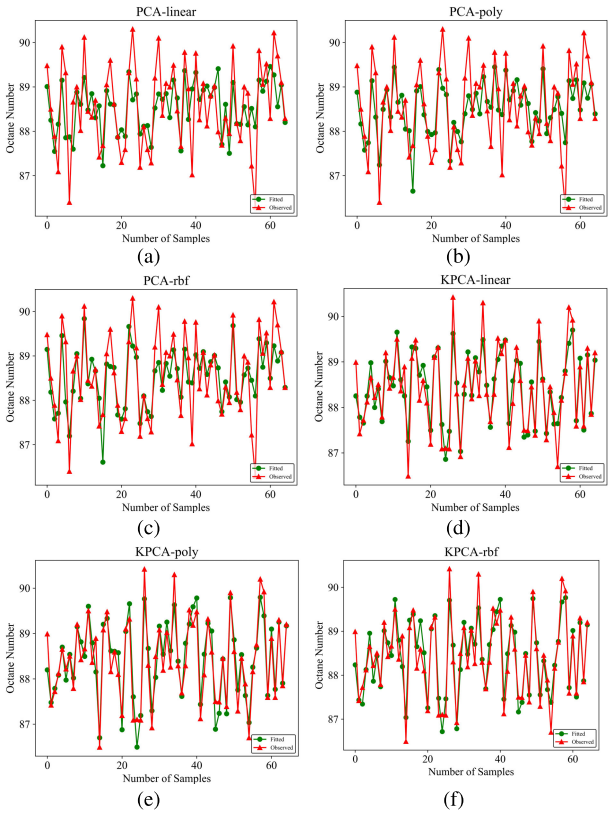


FIGURE 8. The comparison of PCA-SVR and KPCA-SVR with linear, polynomial and RBF kernel functions on the capability of forecasting.

TABLE 5. The parameters setting of grid search and PSO.

Grid search		PSO				
C	γ	$w_{\min}$	$w_{\max}$	n	p	c_1=c_2
[0.1,10000]	[0.01,100]	0.4	0.9	50	200	1.5

B. MODEL OPTIMIZATION

At the beginning of this experiment, it is worth suitably setting the initial parameters of PSO. Table 5 shows the detailed parameters setting of GS and PSO. And the result is presented in Fig. 9.

The first two sub-charts illustrate the forecasting result, where Fig. 9(a) is the predicted outputs of the model in the

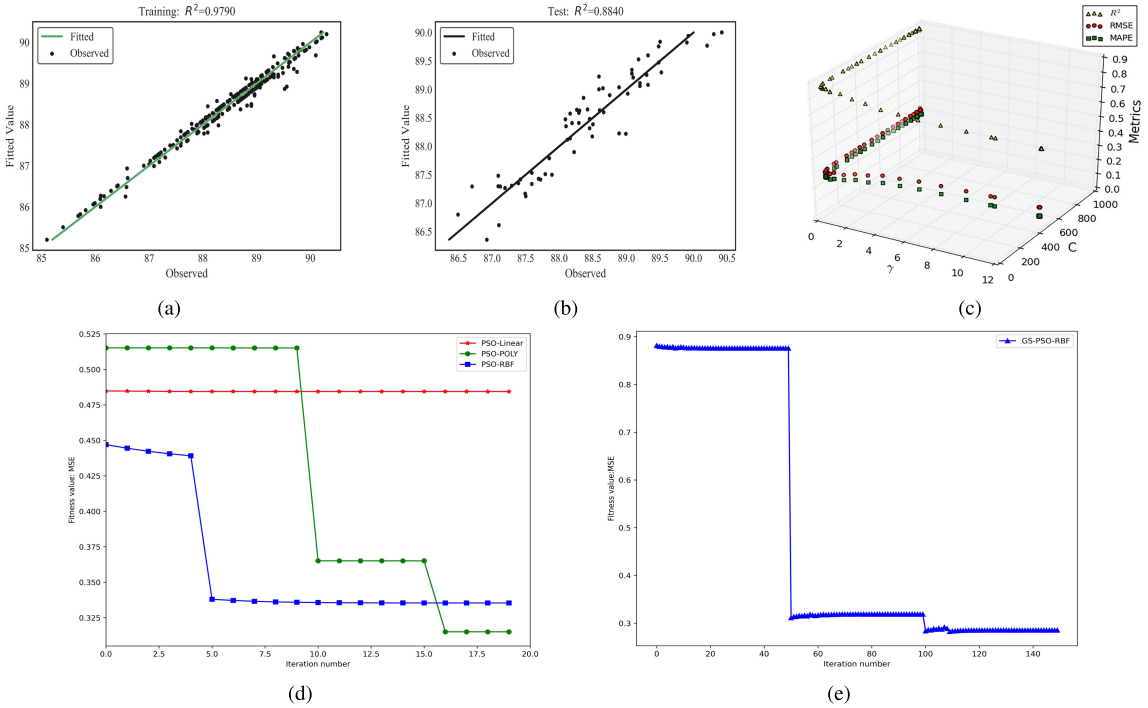


FIGURE 9. R-squared, RMSE, and MAPE of KPCA-SVR (RBF) optimized by the hybrid of the grid search strategy and particle swarm algorithm. And the trajectory of particle swarm in the parameters space composed by C and γ in subgraph (c). Moreover, the fitness value of each iteration number is presented in (d) and (e).

training dataset, and Fig. 9(b) is the prediction result of the test set. We can see that the predicted and true values almost coincide with each other in the during the training. As we can see from Fig. 9(c), the particle swarm (50 particles) moves quickly along the γ-axis with a slight adjustment along the C-axis. Subsequently, they turn to move along the C-axis in a leaping manner. Finally, through mutual information transmission [31], they obtain the best-fit parameter set at a much faster rate, which verifies that the hybrid of the grid search strategy and particle swarm algorithm is highly convergent.

As we can see from Fig. 9(d), PSO-KPCA-SVR with linear kernel function keeps stable during iteration. Note that we use the mean squared error as the fitness value. This demonstrates that linear function fails to handle nonlinear data. On the contrary, polynomial function and RBF function are capable of dealing with the nonlinear and coupled data. After the twentieth iteration, the RBF achieves the best fitness value. As we can see from Fig. 9(e), With each iteration, the particle swarm is searching for the optimal value, and by applying the adaptive weight acceleration factor, the particle swarm is more capable of jumping out of the local optimum. Moreover, the global search ability of the particle swarm is significantly enhanced after grid search.

To investigate the superiority of this novel proposed integrated strategy, the-state-of-the-art regressors are used for comparison, namely ridge regression, PLS (partial least square), Adaboost (adaptive boost), BP neural network, LSTM, and random forest. Also, before building these models, KPCA is performed to extract the main features.

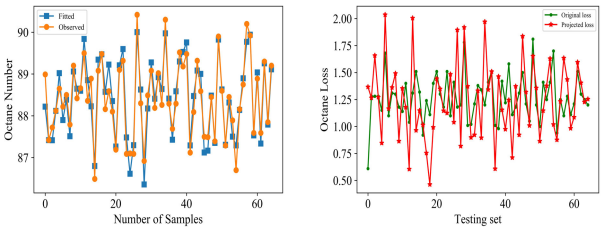
Table 6 presents the performance metrics of the above models, which can be grouped into two categories. One of them is the models with weak generalization ability, including KPCA-Adaboost, KPCA-BP neural network, LSTM, KPCA-LSTM, and KPCA-Random Forest. Although the forecasting accuracy is higher when processing the training set, the capability of handling the unknown data or test sets will decline. Conversely, the remaining models have strong generalization abilities since the prediction accuracy will increase when processing the test set. On a comparative basis, PSO-KPCA-SVR has the best performances ($R^2 = 0.8838$, RMSE $= 0.3084$, MAPE $= 0.2852$) and the strongest generalization capability. Moreover, the performance metrics of PSO-KPCA-SVR with RBF kernel are best among the above models by adopting the optimal parameter combination.

VI. THE PROJECTED OCTANE AND THE CRITICAL ATTRIBUTES

The parameters of the model are always being tuned in order to improve the prediction accuracy. Nonetheless, there is an interesting phenomenon that the mean projected octane number from this model is always greater than the product octane number, which implies that the product octane number in this petrochemical enterprise can be upgraded in a larger room. Here, we give the forecasting octane number

TABLE 6. Comparison results of multiple models. To avoid contingency, we conducted 20 experiments for each algorithm. The left side of ± is the average value and the right side is the variance.

Model	Performance Metrics					
	R^2		RMSE		MAPE	
	Train	Test	Train	Test	Train	Test
KPCA-Ridge	0.6812±0.02	0.8451±0.01	0.5631±0.03	0.3515±0.009	0.4643±0.03	0.3025±0.008
KPCA-PLS	0.7152±0.03	0.8613±0.0075	0.5317±0.0044	0.3369±0.0038	1.1662±0.1	1.1927±0.1
KPCA-Adaboost	0.8475±0.0083	0.8223±0.0012	0.3912±0.0036	0.3813±0.0013	0.3692±0.0014	0.3465±0.0021
KPCA-BP neural network	0.9336±0.02	0.7492±0.02	0.3854±0.0023	0.4908±0.001	0.3093±0.0034	0.4146±0.001
LSTM	0.9204±0.003	0.8326±0.001	0.2461±0.001	0.3627±0.002	0.1823±0.0002	0.3378±0.001
KPCA-LSTM	0.9672±0.003	0.8633±0.0028	0.2201±0.001	0.3436±0.002	0.1795±0.0002	0.3126±0.001
KPCA-Random Forest	0.9484±0.001	0.8012±0.002	0.2276±0.0024	0.4034±0.003	0.1849±0.001	0.3443±0.002
PSO-KPCA-SVR (RBF)	**0.9790±0.002**	**0.8838±0.0021**	0.2194±0.0031	**0.3084±0.0011**	0.1709±0.0008	**0.2852±0.001**



(a) (b)

FIGURE 10. The mean of the fitted result from PSO-KPCA-SVR (RBF) is 88.6988, while the mean of the original octane from the testing data set is 88.4136 in (a). Moreover, the comparison of the original loss of the octane number and projected loss of the octane number in (b).

and the loss of octane number of PSO-KPCA-SVR with RBF kernel.

As we can see from Fig. 10(b), most of the red stars are underneath the green squares, which implies that the loss of the projected octane number is less than the original octane. Under the premise of ensuring the desulfurization effect of gasoline products (The sulfur content should not be greater than 10μg/g in EUR VI and CHN VI standards, but to leave room for the operation of desulphurization unit, the sulfur content of the product is required to be less than 5μg/g), it is necessary to minimize the loss of gasoline octane as much as possible to 30% or more. The loss reduction and the sulfur content of projected octane are presented in Fig. 11(a).

Interestingly, although the locations of the original data sets are scrambled during data processing, their indices are not changed. Hence, it is more likely to confirm the sample numbers which are presented in Table 7. The following assignment is to identify the key attributes by analyzing the principal components of the KPCA again. Fig. 5(b) illustrated that the explained variance and eigenvalue of the first principal component is largest than that of the remaining principal components, which means that the main information of the original dataset can be obtained by only analyzing the first principal component. Since the variables correlated with the principal components can be represented by the score matrix

TABLE 7. These sample numbers meet the requirements for the sulphur content of less than 5μg/g and a loss reduction of more than 30%.

Sample number	Product properties	
	Loss reduction (%)	Sulfur content (μg/g)
16	40.10	3.2
18	36.15	3.2
183	41.84	3.2
197	39.87	3.2
207	35.85	3.2
231	58.33	3.5
261	30.00	4.9
268	33.19	4.5

TABLE 8. 25 key attributes out of 367 raw attributes and their scores.

Attributes	Scores	Attributes	Scores
PC_2105.PV	0.6395	FT_1202.TOTAL	-0.6101
PC_1603.PV	0.6168	TC_2201.PV	-0.6104
FT_9101.TOTAL	0.6304	PC_3501.DACA	-0.6117
FT_3301.TOTAL	0.6439	TE_2001.DACA	-0.6317
FT_9201.TOTAL	0.6322	TE_7508.DACA	-0.6111
FT_9301.TOTAL	0.6372	.PT_7508.DACA	-0.6101
FT_9402.TOTAL	0.6538	TE_7506.DACA	-0.6087
FT_9403.TOTAL	0.6229	TXE_2203A.DACA	-0.6110
FT_3201.DACA	0.6549	TXE_2202A.DACA	-0.6112
PT_2106.DACA	0.6411	AT-0011.DACA.PV	0.6401
AT-0007.DACA.PV	0.6461	PT_2106.DACA.PV	0.6572
AT-0008.DACA.PV	0.6577	FT_5204.PV	0.6465
AT-0009.DACA.PV	0.6476	–	–

calculated in (12), the ultimate goal is to obtain the score vector of the first principal component.

As shown in Fig. 11(b), the scores which are not in the range [-0.6, 0.6] are highlighted in red. Clearly, 25 critical

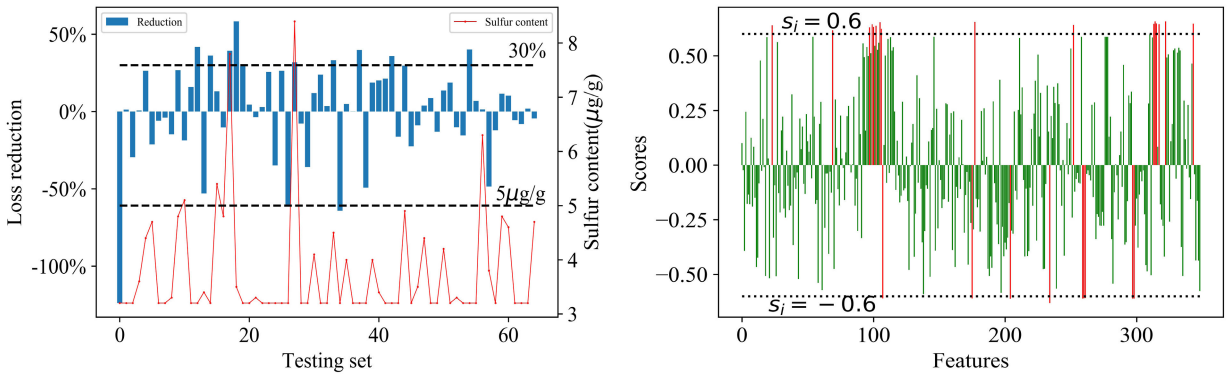


FIGURE 11. The mean of the fitted result from PSO-KPCA-SVR (RBF) is 88.6988, while the mean of the original octane from the testing data set is 88.4136 in (a). Moreover, the comparison of the original loss of the octane number and projected loss of the octane number in (b).

attributes or features can be confirmed from the original 367 attributes of the dataset, and they have a positive or negative effect on the first principal component. The scores corresponding to these 25 key attributes are presented in Table 8.

VII. CONCLUSION & FUTURE WORK

This study was committed to finding the key attributes that intuitively affect the gasoline octane number refined by FCC. we proposed a novel approach of combining grid search and PSO to optimize the KPCA-SVR model for the loss of octane number prediction. In the first phase, KPCA was implemented to extract 20 non-linear principal components of the historical dataset gathered by the FCC gasoline refining process. Next, the strategy of combining grid search and PSO was applied to acquire the optimal parameter set. Meanwhile, five-fold cross-validation was implemented in the grid search to avoid overfitting. The results demonstrated that this model not only improved the prediction accuracy but also displayed the strong capabilities of non-linear mapping and generalization.

Furthermore, the projected results indicated that the mean loss of the predicted octane number was less than that of the original octane number. Therefore, it was sufficient to make the following comparisons, including the fitted and observed octane number, projected and original octane number loss, reduction in predicted loss, and the sulfur content. After these comparisons, we obtained the sample numbers that met the requirements. Additionally, 25 critical attributes of the dataset were detected by analyzing the score vectors of KPCA.

For future research, there are several interesting directions to work. First, how to further improve the performance in the fitting of SVR. Second, how to reduce the computational complexity of KCPA-SVR [25], and process large-scale data set. Finally, with the convince of the applications of SVR is the tip of the iceberg, we will try to combine SVR with different intelligent optimization algorithms (e.g., wavelet transform [15]), and implement it in practical industrial productions.

ACKNOWLEDGMENT

The authors are grateful to all the reviewers and the Editor-in-Chief for their insightful comments on this article.

REFERENCES

[1] T. Wang, W. Li, X. Chang, Y. Xiang, and X. Bao, "Progress in research and development of FCC gasoline cleaning technology," *Chem. Ind. Eng. Prog.*, vol. 38, no. 1, pp. 196–207, 2019, doi: 10.16085/j.issn.1000-6613.2018-1339.

[2] L. Pan, P. Liu, and Z. Li, "A discussion on China's vehicle fuel policy: Based on the development route optimization of refining industry," *Energy Policy*, vol. 114, pp. 403–412, Mar. 2018, doi: 10.1016/j.enpol.2017.12.010.

[3] W. Dabelstein, A. Reglitzky, A. Schütze, K. Reders, and A. Brunner, "Automotive Fuels," in *Ullmann's Encyclopedia of Industrial Chemistry*. Atlanta, GA, USA: ACS, 2016, pp. 1–41.

[4] J. J. Kelly, C. H. Barlow, T. M. Jinguji, and J. B. Callis, "Prediction of gasoline octane numbers from near-infrared spectral features in the range 660-1215 nm," *Anal. Chem.*, vol. 61, no. 4, pp. 313–320, Feb. 1989, doi: 10.1021/ac00179a007.

[5] J. B. Cooper, K. L. Wise, J. Groves, and W. T. Welch, "Determination of octane numbers and reid vapor pressure of commercial petroleum fuels using FT-Raman spectroscopy and partial least-squares regression analysis," *Anal. Chem.*, vol. 67, no. 22, pp. 4096–4100, Nov. 1995, doi: 10.1021/ac00118a011.

[6] S. Lee, H. Choi, K. Cha, and H. Chung, "Random forest as a potential multivariate method for near-infrared (NIR) spectroscopic analysis of complex mixture samples: Gasoline and naphtha," *Microchem. J.*, vol. 110, pp. 739–748, Sep. 2013, doi: 10.1016/j.microc.2013.08.007.

[7] A. Daryasafar, N. Daryasafar, M. Madani, M. K. Meybodi, and M. Joukar, "Connectionist approaches for solubility prediction of n-alkanes in supercritical carbon dioxide," *Neural Comput. Appl.*, vol. 29, no. 1, pp. 295–305, Jan. 2018, doi: 10.1007/s00521-016-2793-7.

[8] A. Janani and M. Sasikala, "Investigation of different approaches for noise reduction in functional near-infrared spectroscopy signals for brain–computer interface applications," *Neural Comput. Appl.*, vol. 28, no. 10, pp. 2889–2903, Oct. 2017, doi: 10.1007/s00521-017-2961-4.

[9] A. G. A. Jameel, V. Van Oudenhoven, A.-H. Emwas, and S. M. Sarathy, "Predicting octane number using nuclear magnetic resonance spectroscopy and artificial neural networks," *Energy Fuels*, vol. 32, no. 5, pp. 6309–6329, May 2018.

[10] N. Nikolaou, C. E. Papadopoulos, I. A. Gaglias, and K. G. Pitarakis, "A new non-linear calculation method of isomerisation gasoline research octane number based on gas chromatographic data," *Fuel*, vol. 83, nos. 4–5, pp. 517–523, Mar. 2004.

[11] T. Dias, R. Oliveira, P. Saraiva, and M. S. Reis, "Predictive analytics in the petrochemical industry: Research octane number (RON) forecasting and analysis in an industrial catalytic reforming unit," *Comput. Chem. Eng.*, vol. 139, Aug. 2020, Art. no. 106912.

[12] B. S. N. Murty and R. N. Rao, "Global optimization for prediction of blend composition of gasolines of desired octane number and properties," *Fuel Process. Technol.*, vol. 85, no. 14, pp. 1595–1602, Sep. 2004, doi: 10.1016/j.fuproc.2003.08.004.
[13] R. M. Balabin and E. I. Lomakina, "Support vector machine regression (SVR/LS-SVM)—An alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data," *Analyst*, vol. 136, no. 8, pp. 1703–1712, 2011, doi: 10.1039/c0an00387e.
[14] G. Najafi, B. Ghobadian, A. Moosavian, T. Yusaf, R. Mamat, M. Kettner, and W. H. Azmi, "SVM and ANFIS for prediction of performance and exhaust emissions of a Si engine with gasoline–ethanol blended fuels," *Appl. Thermal Eng.*, vol. 95, pp. 186–203, Feb. 2016, doi: 10.1016/j.applthermaleng.2015.11.009.
[15] W. Qiao and Z. Yang, "Forecast the electricity price of U.S. Using a wavelet transform-based hybrid model," *Energy*, vol. 193, Feb. 2020, Art. no. 116704, doi: 10.1016/j.energy.2019.116704.
[16] V. N. Vapnik, "Complete statistical theory of learning," *Autom. Remote Control*, vol. 80, no. 11, pp. 1949–1975, Nov. 2019, doi: 10.1134/S000511791911002X.
[17] M. Awad, R. Khanna, M. Awad, and R. Khanna, "Support vector regression," in *Efficient Learning Machines*, Berkeley, CA, USA: Apress, 2015, pp. 67–80.
[18] S. Ding, J. Yu, B. Qi, and H. Huang, "An overview on twin support vector machines," *Artif. Intell. Rev.*, vol. 42, no. 2, pp. 245–252, Aug. 2014, doi: 10.1007/s10462-012-9336-0.
[19] U. Thissen, M. Pepers, B. Üstün, W. J. Melssen, and L. M. C. Buydens, "Comparing support vector machines to PLS for spectral regression applications," *Chemometric Intell. Lab. Syst.*, vol. 73, no. 2, pp. 169–179, Oct. 2004, doi: 10.1016/j.chemolab.2004.01.002.
[20] K. Brudzewski, A. Kesik, K. Kołodziejczyk, U. Zborowska, and J. Ulaczyk, "Gasoline quality prediction using gas chromatography and FTIR spectroscopy: An artificial intelligence approach," *Fuel*, vol. 85, no. 4, pp. 553–558, Mar. 2006, doi: 10.1016/j.fuel.2005.07.019.
[21] F. Melgani and Y. Bazi, "Classification of electrocardiogram signals with support vector machines and particle swarm optimization," *IEEE Trans. Inf. Technol. Biomed.*, vol. 12, no. 5, pp. 667–677, Sep. 2008, doi: 10.1109/TITB.2008.923147.
[22] J. C. L. Alves and R. J. Poppi, "Biodiesel content determination in diesel fuel blends using near infrared (NIR) spectroscopy and support vector machines (SVM)," *Talanta*, vol. 104, pp. 155–161, Jan. 2013, doi: 10.1016/j.talanta.2012.11.033.
[23] M. Wauters and M. Vanhoucke, "Support vector machine regression for project control forecasting," *Autom. Construct.*, vol. 47, pp. 92–106, Nov. 2014, doi: 10.1016/j.autcon.2014.07.014.
[24] W. Qiao, H. Moayedi, and L. K. Foong, "Nature-inspired hybrid techniques of IWO, DA, ES, GA, and ICA, validated through a k-fold validation process predicting monthly natural gas consumption," *Energy Buildings*, vol. 217, Jun. 2020, Art. no. 110023, doi: 10.1016/j.enbuild.2020.110023.
[25] M. S. AL-Musaylh, R. C. Deo, Y. Li, and J. F. Adamowski, "Two-phase particle swarm optimized-support vector regression hybrid model integrated with improved empirical mode decomposition with adaptive noise for multiple-horizon electricity demand forecasting," *Appl. Energy*, vol. 217, pp. 422–439, May 2018, doi: 10.1016/j.apenergy.2018.02.140.
[26] G. Ren, J. Ning, and Z. Zhang, "Multi-variable selection strategy based on near-infrared spectra for the rapid description of dianhong black tea quality," *Spectrochimica Acta A, Mol. Biomol. Spectrosc.*, vol. 245, Jan. 2021, Art. no. 118918, doi: 10.1016/j.saa.2020.118918.
[27] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998, doi: 10.1162/089976698300017467.
[28] A. Rehman, A. Khan, M. A. Ali, M. U. Khan, S. U. Khan, and L. Ali, "Performance analysis of PCA, sparse PCA, kernel PCA and incremental PCA algorithms for heart failure prediction," in *Proc. Int. Conf. Electr., Commun., Comput. Eng. (ICECCE)*, Jun. 2020, pp. 1–5, doi: 10.1109/ICECCE49384.2020.9179199.
[29] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, Aug. 2004, doi: 10.1023/B:STCO.0000035301.49549.88.
[30] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. Int. Conf. Neural Netw.*, vol. 4, Nov. 1995, pp. 1942–1948, doi: 10.4018/ijmfmp.2015010104.
[31] B. Chopard and M. Tomassini, "Particle swarm optimization," in *Natural Computing Series*. Cham, Switzerland: Springer, 2018, pp. 97–102.
[32] R. Poli, J. Kennedy, and T. Blackwell, "Particle swarm optimization: An overview," *Swarm Intell.*, vol. 1, no. 1, pp. 33–57, Oct. 2007, doi: 10.1007/s11721-007-0002-0.
[33] M. A. Ahmadi and S. R. Shadizadeh, "RETRACTED ARTICLE: Prediction of asphaltene precipitation by using hybrid genetic algorithm and particle swarm optimization and neural network," *Neural Comput. Appl.*, vol. 23, no. 2, p. 561, Aug. 2013, doi: 10.1007/s00521-012-0920-7.
[34] J. Xin, G. Chen, and Y. Hai, "A particle swarm optimizer with multi-stage linearly-decreasing inertia weight," in *Proc. Int. Joint Conf. Comput. Sci. Optim.*, Apr. 2009, pp. 505–508, doi: 10.1109/CSO.2009.420.
[35] F. Arteaga and A. Ferrer, "Dealing with missing data in MSPC: Several methods, different interpretations, some examples," *J. Chemometrics*, vol. 16, nos. 8–10, pp. 408–418, Aug. 2002, doi: 10.1002/cem.750.
[36] S. Kim and H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts," *Int. J. Forecasting*, vol. 32, no. 3, pp. 669–679, Jul. 2016, doi: 10.1016/j.ijforecast.2015.12.003.
[37] Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen, and J. Liu, "LSTM network: A deep learning approach for short-term traffic forecast," *IET Intell. Transp. Syst.*, vol. 11, no. 2, pp. 68–75, Mar. 2017, doi: 10.1049/iet-its.2016.0208.
[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

BAOSHENG LI received the B.S. degree from the Henan University of Economics and Law, in 2019. He is currently pursuing the master's degree with North Minzu University. His main research interests include big data analysis and machine learning.

CHUANDONG QIN received the M.S. and Ph.D. degrees from Xidian University, in 2008. He is currently an Associate Professor with North Minzu University. His main research interests include data mining, machine learning, and intelligent information processing.