A New Mixed-Gas-Detection Method Based on a Support Vector Machine Optimized by a Sparrow Search Algorithm

Haitao Zhang and Yaozhen Han *

School of Information Science and Electrical Engineering, Shandong Jiaotong University, Jinan 250357, China
* Correspondence: hanyz@sdjtu.edu.cn

Abstract: To solve the problem of the low recognition rate of mixed gases and consider the phenomenon of low prediction accuracy when traditional gas-concentration-prediction methods deal with nonlinear data, this paper proposes a mixed-gas identification and gas-concentration-prediction method based on a support vector machine (SVM) optimized by a sparrow search algorithm (SSA). Principal component analysis (PCA) is applied to perform data dimensionality reduction on the input data, and SSA is adopted to optimize the SVM hyperparameters to improve the recognition rate and gas-concentration-prediction accuracy of mixed gases. For the mixed-gas identification, the classification accuracy is significantly improved under the proposed SSA optimization SVM method (SSA-SVM), compared with random forest (RF), extreme-learning machine (ELM), and BP neural network methods. With respect to gas-concentration prediction, the maximum fitting degrees reached 99.34% for single gas-concentration prediction and 97.55% for mixed-gas-concentration prediction. The experimental results show that the SSA-SVM method had a high recognition rate and high concentration-prediction accuracy in gas-mixture detection.

Keywords: mixed-gas detection; support vector machine; sparrow search algorithm; prediction; classification

1. Introduction

Mixed-gas detection is of great significance in the fields of food safety testing [1], agricultural production management [2], urban environmental quality testing [3], and life science research [4], particularly in industrial production. This is one of the key means to prevent fire, explosion, poisoning, and other safety accidents and ensure the safe production of enterprises. For example, once flammable and explosive gases, such as methane and ethylene, leak, it is likely to cause fire, explosion, and other safety accidents.

Gas sensors are an important component of gas detection. Some scholars focus on improving the performance of gas sensors to enhance the detection level of mixed gases. For instance, metals were doped into nanomaterials to raise the sensitivity of sensors in [5,6]. Polyaniline/single-walled carbon nanotube composites were utilized to increase the selectivity and stability of the sensor at room temperature in [7]. The soft-film-plate method and nano-casting strategy were adopted to enhance the sensitivity, response, and recovery time of porous metal oxide sensors in chemical synthesis in [8].

The performance of gas sensors has been promoted through composites, preparation processes, and doping. However, it is far from enough to consider only the hardware. As a multi-disciplinary interdisciplinary subject, machine learning is a research hotspot in the field of artificial intelligence and pattern recognition. Its application covers all fields of artificial intelligence [9–11]. Machine learning is also applied in the field of mixed-gas detection. Peng et al. [12] proposed a mixed-gas-identification method based on a deep convolutional neural network (DCNN), as DCNNs have high accuracy and robustness.

Due to the complex network structure of DCNNs and the possibility of falling into local optima, Zhang et al. [13] adopted a BP neural network with a simple network structure.
to realize the anti-interference detection of carbon monoxide and methane. However, the optimization objective function of the network was complex, which led to a slow convergence speed, and it was difficult to achieve timeliness in practical applications.

For sake of further improving the time efficiency of gas detection, random forest (RF), and k-nearest neighbor (KNN) algorithms were adopted [14,15]. Researchers experimentally demonstrated that the response time of gas detection is reduced to an extent under RF and KNN methods. Yet, with the increase in training samples, the amount of calculation for the KNN becomes larger, and its fault tolerance to training data is poor. The RF is overly sensitive to small changes in the training dataset and tends to overfit during classification [16].

Support vector machines (SVMs), as a type of machine-learning algorithm, has simple calculations, strong versatility, and robustness. SVMs have been widely applied in pattern classification and nonlinear regression, such as image recognition [17,18], text classification [19], fault detection [20,21], etc. In the field of mixed-gas detection, Rachid et al. [22] compared the effectiveness of partial least squares with the SVM method in monitoring gas concentrations in a confined environment. The results showed that gas concentrations were more accurately estimated under the SVM method. Zhao et al. [23] adopted a BP neural network, SVM, and extreme-learning machine (ELM) for identifying ethanol, acetone, and formaldehyde.

The results showed that SVM had the highest recognition accuracy. Zhang et al. [24] achieved the semi-quantitative detection of toxic and harmful gas mixtures in the kitchen through electronic nose detection. Whether through pattern classification or nonlinear regression, the relevant parameters of the SVM (mainly penalty parameter $c$ and kernel function parameter $g$) must be adjusted to obtain a relatively good prediction effect. The grid search method is commonly used to optimize the selection of SVM parameters [25].

Although the global optimal solution can be found by grid search, the hyperparameter space is not limited. If the scope is expanded to find the global optimal solution, it will be time-consuming. Scholars have proposed different optimization algorithms to solve the problem of SVM hyperparameters, including genetic algorithms (GA) [26,27], particle swarm optimization (PSO) algorithms [28,29], ant colony algorithms [30], grey wolf optimization (GWO) algorithms [31], etc.

In the field of mixed-gas detection, the hyperparameters of support vector regression (SVR) are optimized by PSO to improve the prediction accuracy of mixed gas [32]. Deng et al. [33] proposed a GWO optimization SVM method to effectively suppress the effects of temperature and pressure on carbon monoxide. Li et al. [34] presented the SVR method optimized by GA to increase the prediction accuracy of the gas content in the drilling site. This method achieved precise control of the differences outside the prediction point.

The above optimization algorithms have enhanced the performance of SVM/SVR to an extent. However, PSO and GWO tend to fall into local optima when dealing with complex problems, and the convergence speed is slow in late iterations. The efficiency of GA is usually lower when compared with other traditional optimization algorithms.

Sparrow search algorithm (SSA), a new swarm intelligence optimization algorithm, was proposed by Xue and Shen in 2020 [35]. The algorithm mainly simulates the foraging and anti-predation behavior of sparrows. SSA has attracted wide attention because of its fast convergence speed and strong optimization ability. Wang et al. [36] adopted SSA to solve the optimal configuration model of distributed generations, and the effectiveness and superiority of SSA were verified by experimental simulation.

Song et al. [37] described the SSA optimization ELM method for the evaluation of water quality, which successfully overcomes the instability and nonlinearity of water quality parameter data. Liu et al. [38] optimized the SSA method to solve the problem of unmanned aerial vehicle (UVA) path planning. The modified SSA provides the best route for UAVs in complex 3D flight environments.
In the field of image processing, Wu et al. [39] proposed a modified SSA algorithm to deal with the matter of poor performance of threshold image segmentation in peak-to-noise ratio and feature similarity. Zhang et al. [40] introduced SSA into an adaptive enhancement classifier to improve the potential of lung CT image classification performance and the probability of early-stage cancer detection. Aiming at the problem of a low recognition rate of mixed gas and low prediction accuracy of gas concentration, this paper proposes a mixed-gas-detection method based on SSA-SVM. Its contribution is summarized as follows.

1. An optimization algorithm based on the sparrow search algorithm to optimize the hyperparameters of support vector machines is designed for the identification of mixed gases and the prediction of gas concentration.

2. The parameter selection of SVM/SVR is solved by the SSA algorithm. In mixed-gas identification, the identification accuracy of the mixed gas is improved by combining PCA with SSA-SVM. In gas-concentration prediction, the fitting degree of gas-concentration prediction under single gas and mixed gas is improved by the SSA-SVR. Compared with GWO and GA optimization algorithms, the SSA optimization algorithm improves the convergence speed and prediction fitting degree.

The structure of this paper is organized as follows. Section 2 introduces the mixed-gas-detection method. In Section 3, the mixed-gas identification and gas-concentration prediction are experimentally verified, and the results are analyzed and compared. Finally, our summary is given in Section 4.

2. Mixed-Gas-Detection Methods

In order to further improve the accuracy of mixed-gas identification and concentration prediction, a new mixed-gas-detection method based on SSA-SVM is proposed in this paper. The flow chart of mixed-gas identification based on SSA-SVM is shown in Figure 1, and the flow chart of mixed-gas-concentration prediction based on SSA-SVR is shown in Figure 2. The dataset provided by the University of California Irvine (UCI) machine-learning repository is adopted for simulation tests in MATLAB 2021 b. This section presents the proposed method in detail.

![Figure 1. The SSA-SVM mixed-gas-identification method.](image-url)
2.1. Mixed-Gas Classification and Gas-Concentration Prediction Based on SVM

2.1.1. Mixed-Gas Classification

SVM, as a machine-learning algorithm, is commonly used to realize the pattern recognition of gas sensor arrays. The main idea is to construct a classification hyperplane as a decision surface to distinguish mixed gases. The relationship between a gas sensor and gas mixture is usually nonlinear, such that a correctly divided hyperplane is often not found in the original gas sample space. Thus, the original low-dimensional space is mapped into a high-dimensional space to search for a suitable partition hyperplane.

In the experiment, the mixed-gas classification dataset has 41,785 samples. Then, the training set in the given feature space is represented as

\[ D = \{ (x_1, y_1), \ldots, (x_{41785}, y_{41785}) \} \in (X \times Y)^{41785} \]

where \( x_i \in X, y_i \in Y = \{1, 2, 3, 4\} (i = 1, 2, \ldots, 41785) \); \( x_i \) is the main feature vector extracted by PCA; \( y_i \) is the gas classification label.

The optimal hyperplane of high-dimensional space can be defined as

\[ f(x) = w^T \varphi(x) + b \]

where \( w \) is the normal vector of the classification hyperplane and \( b \) is the intercept.

After the input data is mapped to a high-dimensional space, the SVM optimization problem becomes

\[ \min_{w,b} \frac{1}{2} \| w \|^2 + c \sum_{i=1}^{41785} \xi_i \]

where \( c \) is the regularization coefficient and \( \xi_i \) is the relaxation variable.

SVM maps linear and indivisible low-dimensional space to high-dimensional space by nonlinear transformation, and it has “dimension disaster” when operating in high-dimensional space. The introduction of the kernel function can not only transform the...
feature from low-dimensional space to high-dimensional space but also effectively avoid “dimension disaster” and reduce the amount of computation. For optimization problems, it is necessary to select an appropriate kernel function $K(x, x_i)$ and penalty parameter $c$ and then to construct and solve the optimization problem.

$$\max Q(\lambda) = \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j K(x_i, x_j)$$

s.t. \[
\sum_{i=1}^{m} \lambda_i y_i = 0 , \\
0 \leq \lambda_i \leq c , \\
i = 1, 2, \ldots, m
\]  

Then, the optimal solution is obtained as $\lambda^* = (\lambda_1^*, \lambda_2^*, \ldots, \lambda_m^*)^T$.
According to the optimal solution, $\lambda^*$, $w^*$ is calculated as

$$w^* = \sum_{i=1}^{m} \lambda_i^* y_i K(x, x_i)$$

To choose a positive component $\lambda_j^* (0 < \lambda_j^* < C)$ of $\lambda^*$, the threshold is calculated as

$$b^* = y_i - \sum_{i=1}^{m} y_i \lambda_i^* K(x_i - x_j)$$

Finally, the decision function is constructed as

$$f(x) = \sgn(\sum_{i=1}^{m} \lambda_i^* y_i K(x, x_i) + b^*)$$

2.1.2. Mixed-Gas-Concentration Prediction

SVR, as an important application branch of SVM, is superior to other machine-learning algorithms in gas-concentration prediction. The basic idea of SVR is no longer to find an optimal classification surface to separate the samples but to find an optimal classification surface to minimize the error of all training samples from the optimal classification surface.

For the regression problem, the training data $T = \{(x_1, y_1), \ldots, (x_n, y_n)\}, y_i \in \mathbb{R}$ is given to make $f(x) = w^T x + b$ and $y_i$ as close as possible. Then, the loss is calculated from the difference between the predicted output gas concentration $f(x)$ and the real gas concentration $y_i$. The corresponding loss function is

$$L(z) = \max(0, |z| - \epsilon) = \begin{cases} 0, & \text{if } |z| < \epsilon \\ |z| - \epsilon, & \text{otherwise} \end{cases}$$

Thus, the optimization problem of SVR can be expressed as

$$\min_{w, b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} L(z_i)$$

s.t. \[
|y_i - f(x_i)| \leq \epsilon
\]  

Slack variables $\xi_i$ and $\xi_i^*$ are then introduced to replace the loss function, and Equation (9) is rewritten as

$$\min_{w, b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \left( \xi_i + \xi_i^* \right)$$

s.t. \[
\xi_i \geq 0 , \\
y_i - f(x_i) \leq \epsilon + \xi_i , \\
f(x_i) - y_i \leq \epsilon + \xi_i^* , \\
\xi_i, \xi_i^* \geq 0
\]
where C is the regularization coefficient, which is mainly used to prevent SVR overfitting. If the value of is C too large or too small, the generalization ability of SVR will be deteriorated.

Then, the Lagrange multiplier is introduced to obtain the Lagrange function and the dual problem. The solution of SVR can be obtained by solving the dual problem.

\[ f(x) = \sum_{i=1}^{n} (\lambda_i^* - \lambda_i)x_i^T x_j + b \]  

Finally, the kernel function is added to obtain the decision function of the gas-concentration prediction.

\[ f(x) = \sum_{i=1}^{n} (\lambda_i^* - \lambda_i)K(x,x_i) + b \]  

2.2. Sparrow Search Algorithm

Whether SVM or SVR, the performance depended on the selection of the two parameters. In this paper, an SSA optimization algorithm is introduced to solve the hyperparameter problem and to improve the identification accuracy of mixed gas and the accuracy of gas-concentration prediction. SSA mainly simulates the foraging and anti-predatory behavior of sparrow groups. In the sparrow foraging process, sparrows are divided into finders and joiners according to their locations and energy reserves. The finders are responsible for finding food and for providing the joiners with foraging areas and directions. Joiners find food based on the information provided by the finders.

Assuming that there are sparrows, the sparrow population can be expressed as

\[
A = \begin{bmatrix}
a_{1,1} & a_{1,2} & \ldots & a_{1,d} \\
a_{2,1} & a_{2,2} & \ldots & a_{2,d} \\
\vdots & \vdots & \ddots & \vdots \\
a_{p,1} & a_{p,2} & \ldots & a_{p,d}
\end{bmatrix}
\]  

where \( p \) is the number of sparrows, and \( d \) is the number of optimization parameters.

In the iterative process, the algorithm will continue to search for food within the search range if the finder does not find predators, and it will sound an alarm to alert other sparrows when the finder finds a predator. At this point, the whole flock of sparrows will quickly fly to other safe areas to feed. The location update of the discoverer is described as

\[
a_{t+1}^{i,j} = \begin{cases} 
a_{t}^{i,j} \times \exp \frac{-i}{\eta \times MaxT} & \text{if (} R_2 < ST \text{)} \\
\frac{a_{t}^{i,j} + Q \times L}{|| a_{t}^{i,j} - a_{t+1} \|_p } & \text{otherwise} 
\end{cases}
\]  

where \( t \) is the current number of iterations; \( MaxT \) is the maximum number of iterations; and \( R^2 \) and \( ST \) are the warning value and the safety value, respectively. When \( t < MaxT \), SSA will sort the sparrows according to their fitness and find the current best and worst sparrows.

During foraging, the foraging strategy is determined by their energy level (fitness) of the sparrows. However, it will immediately leave the current location to obtain food if the joiner finds that the finder has found better food. At this time, the location update of the joiners is described as

\[
a_{t+1}^{i,j} = \begin{cases} 
Q \times \exp \frac{a_{t}^{Worst} - a_{t}^{i,j}}{p} & \text{if (} i < \frac{n}{2} \text{)} \\
a_{t+1}^{i,j} + \frac{a_{t}^{i,j} - a_{t}^{Worst}}{p} \times (X^T XX^T)^{-1} \times L & \text{otherwise} 
\end{cases}
\]
danger. However, once the sparrow in the middle of the population is aware of the danger, it will immediately move closer to other sparrows to reduce its own danger. For these sparrows, the location update is described as

$$a_{i,j}^{t+1} = \begin{cases} a_{best}^t + \lambda |a_{i,j}^t - a_{best}^t| & f_i > f_g \\ a_{i,j}^t + K \ast \left( \frac{|a_{i,j}^t - a_{best}^t|}{(f_i - f_w)^+} \right) & f_i = f_g \end{cases}$$

(16)

where $a_{best}$ is the current global optimal position; $\lambda$ is the step size; $f_i$ is the fitness of the current sparrow individual; and $f_g$ and $f_w$ are the current global best and worst fitness, respectively.

During the iteration process, if the sparrow’s new position is better than the previous position, the current position will be updated until the global optimal position and the global optimal fitness are found. During this period, the identity of the sparrow is also constantly updated and alternated. Every sparrow can be a finder if it is well adapted; however, the proportions of finders and joiners in the population are constant.

3. Experimental Results and Analysis

3.1. Data Preparation

This paper uses the data set “Gas sensor arrays in dynamic gas mixtures” from the UCI. The dataset contains 16 chemical sensors of four different types: TGS-2602, TGS-2602, TGS-2600, TGS-2600, TGS-2610, TGS-2620, TGS-2620, TGS-2602, TGS-2602, TGS-2600, TGS-2600, TGS-2610, TGS-2610, TGS-2620, TGS-2620, and TGS-2620. In the experiment, the sensor array composed of 16 gas sensors is placed in a 60 mL sealed box.

Gas samples are injected at a constant flow rate of 300 mL/min. The conductivity of the sensor (S/m) is continuously collected at a sampling frequency of 100 Hz. Each measurement is constructed by continuously acquiring signals from 16 sensor arrays. The concentration of the gas sample varies randomly. The gas samples are mainly composed of methane and ethylene in the air. The concentration range of ethylene is 0–20 ppm, and the concentration range of methane is 0–300 ppm. More details on the dataset can be found in [41]. After sorting and screening this dataset, 41,785 gas samples are finally selected for the experiment.

3.2. Mixed-Gas Identification

3.2.1. Feature Extraction

The input data of this system is the output of 16 sensors in the sensor array. The output data is the gas-classification label, and the desired output is

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

(17)

which represents methane, ethylene, air, and mixed gases.

To improve the recognition accuracy of gas mixtures, the input data needs to be subjected to feature extraction. This eliminates or reduces the effect of the gas concentrations on the sensor array. As a common data analysis method, PCA can be used to extract the main feature vectors of data, achieve dimensionality reduction of high-dimensional data, and maximally preserve the feature information in high-dimensional data. Since the dimensions of the input data are inconsistent, which affects the calculation results, this study needs to be standardized to the original data first.

$$a_i = \frac{x_i - \bar{x}}{\sqrt{\frac{1}{m} \sum_{i=1}^{m} (x_i - \bar{x})^2}}$$

(18)
Table 1. Eigenvalues and the contribution rates of 16 eigenvectors.

<table>
<thead>
<tr>
<th>Principal Components</th>
<th>Eigenvalue</th>
<th>Contribution Rate %</th>
<th>Cumulative Contribution Rate %</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC1</td>
<td>10.650</td>
<td>66.562</td>
<td>66.562</td>
</tr>
<tr>
<td>PC2</td>
<td>3.779</td>
<td>23.620</td>
<td>90.182</td>
</tr>
<tr>
<td>PC3</td>
<td>1.067</td>
<td>6.669</td>
<td>96.851</td>
</tr>
<tr>
<td>PC4</td>
<td>0.413</td>
<td>0.246</td>
<td>99.436</td>
</tr>
<tr>
<td>PC5</td>
<td>0.036</td>
<td>0.224</td>
<td>99.906</td>
</tr>
<tr>
<td>PC6</td>
<td>0.007</td>
<td>0.049</td>
<td>99.955</td>
</tr>
<tr>
<td>PC7</td>
<td>0.005</td>
<td>0.017</td>
<td>99.972</td>
</tr>
<tr>
<td>PC8</td>
<td>0.002</td>
<td>0.012</td>
<td>99.984</td>
</tr>
<tr>
<td>PC9</td>
<td>0.001</td>
<td>0.006</td>
<td>99.990</td>
</tr>
<tr>
<td>PC10</td>
<td>0</td>
<td>0.006</td>
<td>99.996</td>
</tr>
<tr>
<td>PC11</td>
<td>0</td>
<td>0.002</td>
<td>99.999</td>
</tr>
<tr>
<td>PC12</td>
<td>0</td>
<td>0.001</td>
<td>99.999</td>
</tr>
<tr>
<td>PC13</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PC14</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PC15</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PC16</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Standardization transforms the input data into data with a mean of 0 and a standard deviation of 1. This reduces the effect of outliers on the gas-classification results. Then, the input data is

\[
X = \begin{bmatrix}
    a_{1,1} & a_{2,1} & \cdots & a_{16,1} \\
    a_{1,2} & a_{2,2} & \cdots & a_{16,2} \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{1,41785} & a_{2,41785} & \cdots & a_{16,41785}
\end{bmatrix} = \begin{bmatrix} b_1 & b_2 & \cdots & b_{16} \end{bmatrix}
\] (19)

where \( b_1 = [a_{1,1}, a_{1,2}, \ldots, a_{1,41785}]^T \).

After standardized processing, the variance matrix is represented as

\[
C = \frac{1}{41784} X^T X
\] (20)

The eigenvalue \( k_1, k_2, \ldots, k_{16} \) and \( \lambda_1, \lambda_2, \ldots, \lambda_{16} \) eigenvector of the covariance matrix are then obtained. The eigenvalues are sorted from large to small, and the contribution rate of each eigenvalue is

\[
p_i = \frac{k_i}{\sum_{i=1}^{16} k_i}
\] (21)

The cumulative contribution rate is obtained by solving the contribution rate of each eigenvalue. The cumulative contribution rate is the sum of the contribution rates of the first \( n \) eigenvalues. This is reflected the ability of the first \( n \) principal components to explain the original variable. Finally, the required principal components are obtained from the cumulative contribution rate: \( X = (x_1, x_2, \ldots, x_n)^T \). The sixteen eigenvalues and contribution rates obtained from this data set after PCA processing are shown in Table 1.

As shown in Table 1, with the increase in the principal components, the increase in the contribution rate gradually decreases. When there are four principal components, the cumulative contribution rate reaches more than 99%. The four principal components explain more than 99% of the total variance. Thus, the original data can be reduced from 16 to 4 dimensions to represent the original features in this paper. After the data is reduced by PCA, the sample order of the classification dataset is randomly shuffled. According to the ratio of 8:2, this is divided into training sets and test sets.
3.2.2. Classification Results for Mixed Gases

In the experiment, the value range of hyperparameter $c$ and $g$ of SVM are set as $[2^{-2}, 2^{8}]$ and $[2^{-5}, 2^{5}]$, respectively. The kernel function of SVM is the radial basis function. The related parameters of SSA are set as shown in Table 2.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$d$</th>
<th>$lb$</th>
<th>$ub$</th>
<th>$p$</th>
<th>$MaxT$</th>
<th>$ST$</th>
<th>$FD$</th>
<th>$JD$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>2</td>
<td>$[2^{-2}, 2^{-5}]$</td>
<td>$[2^{8}, 2^{5}]$</td>
<td>50</td>
<td>100</td>
<td>0.6</td>
<td>90%</td>
<td>10%</td>
</tr>
</tbody>
</table>

In Table 2, $d$ is the number of optimization parameters; $lb$ is the lower limit of the optimization parameters in SVM; $ub$ is the upper limit of the optimization parameters in SVM; $p$ is the sparrow population size; $MaxT$ is the maximum number of iterations; $ST$ is the safety threshold; $FD$ is the finder; $JD$ is the joiner.

After setting the relevant parameters of SSA, the optimal parameters $c = 53.248$ and $g = 5.979$ are obtained by solving Equations (14)–(16). Then, the best parameter combination is used to train on the training set to obtain the SSA-SVM model, and finally the classification test is performed on the test set. The classification results are shown in Figure 3.

![Confusion Matrix](image)

**Figure 3.** Mixed-gas-identification confusion matrix. ‘1’ is methane, ‘2’ is ethylene, ‘3’ is air, and ‘4’ is mixed gas.

It can be seen from Figure 3 that the accuracy rates of methane, ethylene, air, and mixed gas in the mixed-gas identification are 95.5%, 96.1%, 95.9%, and 96.7%, respectively. The accuracy rate is 96% among 2006 methane test samples, and 96.3% of samples were correctly identified as mixed gas among 1695 mixed-gas test samples. Overall, SSA-SVM correctly classified 8025 out of 8357 test samples. The accuracy rate reached 96%.

According to the model processing tasks, the evaluation criteria were also different. Accuracy is the most simple and intuitive evaluation index in classification problems. The accuracy is susceptible to the influence of the larger category if the sample is unevenly proportioned. In multi-classification, the arithmetic mean of the sample accuracy under each category is generally used as the evaluation index of the model.

$$ACC = \frac{1}{4} \sum_{i=1}^{4} \frac{TP_i}{TP_i + FP_i}$$

(22)
where TP is the number of correct predictions and FP is the number of incorrect predictions.

To further demonstrate the performance of SSA-SVM, we compared it with RF, ELM, and BP neural networks. The relevant parameters of the RF, ELM, and BP neural networks are shown in Table 3. The classification results are shown in Figure 4.

Table 3. Related parameters of the comparative test.

<table>
<thead>
<tr>
<th>Method</th>
<th>Related Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>trees = 500, mtry = 4</td>
</tr>
<tr>
<td>ELM</td>
<td>the number of hidden layer neurons is 100</td>
</tr>
<tr>
<td>BP-NN</td>
<td>the number of neurons in the hidden layer is 10</td>
</tr>
<tr>
<td></td>
<td>the activation function is Sigmoid, three layers</td>
</tr>
</tbody>
</table>

Figure 4. Mixed gas results for four mixed-gas-identification methods. (a) The SSA-SVM gas-identification results; (b) the RF gas-identification results; (c) the ELM gas-identification results; and (d) the BP neural network gas-identification results.

From the combination of Figure 4 and Equation (22), the accuracies of the four methods can be calculated as 96%, 92.9%, 86.8%, and 81.3%. SSA-SVM improved the accuracy by 3.1%, 8.2%, and 14.7% compared with the RF, ELM, and BP neural network.

In order to better reflect the mixed gas recognition ability of the SSA-SVM, RF, ELM, and BP neural networks, the accuracy rate of each category under the above methods is calculated. The comparison results are shown in Figure 5. As shown in Figure 5, the SSA-SVM method is better than the other three methods in identifying both single gases and mixed gases.
3.3. Prediction of Mixed-Gas Concentration

3.3.1. Data Processing

In this paper, the characteristic subsets of each gas concentration are screened from the classification dataset. The input data is unchanged, and the output data is the gas concentration value. In the experiment, the input data and output data are normalized to be between 0 and 1 to improve the calculation efficiency of the regression model fitting process.

\[
y_i = \frac{x_i - \min_{1 \leq j \leq n} \{x_j\}}{\max_{1 \leq j \leq n} \{x_j\} - \min_{1 \leq j \leq n} \{x_j\}}
\]

(23)

Once the data process is completed, the sample order of the feature subset is randomly shuffled. It is divided into a training set and test set with a ratio of 9:1, and five-fold cross-validation will be performed during the training process. The sample sizes for each subset of gas concentration characteristics are shown in Table 4.

<table>
<thead>
<tr>
<th>Data</th>
<th>Single Gas</th>
<th>Mixed Gas</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CH4</td>
<td>C2H4</td>
</tr>
<tr>
<td>Train</td>
<td>9010</td>
<td>8589</td>
</tr>
<tr>
<td>Test</td>
<td>1002</td>
<td>955</td>
</tr>
</tbody>
</table>

3.3.2. Mixed-Gas-Concentration Prediction Results

The value range of the hyperparameter \( q \) is \([2^{-5} - 2^{10}]\), and the value range of \( t \) is \([2^{-5} - 2^{10}]\). The loss function value is set as 0.01. The relevant parameters of SSA are set as Table 5.
Table 5. SSA-related parameters in SSA-SVR.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
<td>2</td>
</tr>
<tr>
<td>lb</td>
<td>$[2^{-5}, 2^{-5}]$</td>
</tr>
<tr>
<td>ub</td>
<td>$[2^{10}, 2^{10}]$</td>
</tr>
<tr>
<td>p</td>
<td>50</td>
</tr>
<tr>
<td>MaxT</td>
<td>100</td>
</tr>
<tr>
<td>ST</td>
<td>0.8</td>
</tr>
<tr>
<td>FD</td>
<td>40%</td>
</tr>
<tr>
<td>JD</td>
<td>60%</td>
</tr>
</tbody>
</table>

In Table 5, d is the number of optimization parameters; lb is the lower limit of optimization parameters in SVR; ub is the upper limit of optimization parameters in SVR; p is the sparrow population size; MaxT is the maximum number of iterations; ST is the safety threshold; FD is the finder; JD is the joiner.

According to Equations (14)–(16), the two hyperparameters of SVR are optimized, and the optimal parameter combination is finally determined. The optimal parameter combination of each feature subset is shown in Table 6.

Table 6. The best parameter combination.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Single Gas</th>
<th>Mixed Gas</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CH4</td>
<td>C2H4</td>
</tr>
<tr>
<td>q</td>
<td>66.6915</td>
<td>5.9072</td>
</tr>
<tr>
<td>t</td>
<td>7.8229</td>
<td>66.3515</td>
</tr>
</tbody>
</table>

Then, the gas-concentration-prediction test is conducted using the test set. The prediction results are normalized to obtain the final concentration prediction results. The SSA-SVR gas-concentration-prediction results are shown in Figure 6.

![Figure 6](image-url)

**Figure 6.** The SSA-SVR gas-concentration prediction results. (a) The predicted concentration of methane under a single gas; (b) the predicted concentration of ethylene under a single gas; (c) the predicted concentration of methane under the mixed gas; and (d) the predicted concentration of ethylene in the mixed gas.

The coefficient of determination $R^2$ is one of the performance evaluation indicators of the regression model, which is applied to reflect the accuracy of the regression model to fit the data. The value range of $R^2$ is 0–1. The closer the value is to 1, the better the model fits the data.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - f(i))^2}{\sum_{i=1}^{n} (y_i - g)^2} \times 100\%$$ (24)
To further validate the regression effect of the SSA algorithm for optimizing SVR, we compared it with the GWO and GA optimization algorithms. The minimum mean square error result is shown in Figure 7. The relevant parameter settings of the GWO and GA optimization algorithms are shown in Tables 7 and 8.

Table 7. The parameters related to the GWO optimization algorithm.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Maxt</th>
<th>dim</th>
<th>lb</th>
<th>ub</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>50</td>
<td>100</td>
<td>2</td>
<td>$[2^{-5}, 2^{-5}]$</td>
<td>$[2^{10}, 2^{10}]$</td>
</tr>
</tbody>
</table>

In Table 7, SA is the number of wolves, Maxt is the maximum number of iterations, dim is the number of optimization parameters, lb is a lower bound on the optimization parameters, and ub is an upper bound on the optimization parameters.

Table 8. The parameters related to the GA optimization algorithm.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Maxgen</th>
<th>ggap</th>
<th>px</th>
<th>pm</th>
<th>select</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nind</td>
<td>50</td>
<td>100</td>
<td>0.9</td>
<td>0.7</td>
<td>0.1</td>
<td>rws</td>
</tr>
</tbody>
</table>

In Table 8, Nind is the population size; Maxgen is the maximum number of genetic generations; ggap is the generation gap; px and pm are the crossover probability and variation probability, respectively; and select is the selection function.

The mean square error of the SSA optimization algorithm reaches the minimum after six iterations under the same population size and iteration numbers. The GWO optimization algorithm achieves the minimum mean square error in about 12 iterations, and the GA optimization algorithm takes about 15 iterations. In comparison, the convergence speed of the SSA optimization algorithm is better than that of the GWO and GA algorithms.
In order to further intuitively reflect the fitting degree of the prediction data of SSA-SVR, GWO-SVR, and GA-SVR, the determination coefficient of each model is calculated. The fitting results of the gas-concentration prediction are shown in Figure 8. Whether for a single gas or a mixture of gases, the concentration prediction fit of SSA-SVR is better than that of GWO-SVR and GA-SVR. The fitting degree of SSA-SVR for single gas-concentration prediction is over 97%. The fitting degrees of SSA-SVR for the prediction of methane and ethylene concentrations under mixed gas are 92.36% and 97.55%.

4. Conclusions

Mixed-gas detection is of great importance in industrial production processes. This type of detection is important to ensure the green development and safe production of enterprises. This paper presents a mixed-gas-identification method based on SSA-SVM. Our conclusions are summarized as follows.

(1) In the mixed-gas identification experiment, SSA-SVM was combined with PCA for mixed-gas identification. PCA was adopted to reduce the original data from 16 dimensions to 4 dimensions to achieve dimensionality reduction of the high-dimensional data. The influence of redundant data on the SSA-SVM model was reduced. The classification accuracy of SSA-SVM on the test set was over 96%. Compared with the RF, ELM, and BP neural network models, the classification accuracy of SSA-SVM was improved by 3.1%, 8.2%, and 14.7%, respectively.

(2) In the gas-concentration-prediction experiment, the prediction fit and convergence iteration speed of SSA-SVR were better than those of GWO-SVR and GA-SVR. For the concentration prediction of a single gas, the fitting degree of SSA-SVR was more than 97%. The fitting degree of the predicted concentration of methane was as high as 99.34%. The fitting degrees of the predicted concentrations of methane and ethylene in mixed gas reached 92.36% and 97.55%.

(3) With the long-term use of the sensor, there will inevitably be sensor drift. Data distortion is prone to affect data analysis. Next, we will focus on this issue and reduce the effect of drift on the system.
Author Contributions: Conceptualization, H.Z. and Y.H.; methodology, H.Z. and Y.H.; software, H.Z.; validation, H.Z. and Y.H.; formal analysis, H.Z.; investigation, H.Z. and Y.H.; resources, H.Z.; data curation, H.Z.; writing—original draft preparation, H.Z.; writing—review and editing, Y.H. and H.Z.; visualization, H.Z. and Y.H.; supervision, Y.H. and H.Z.; project administration, Y.H. and H.Z.; funding acquisition, Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (61803230); Projects of Shandong Province College Youth Innovation Technology Support Program (2019KJN023); the Graduate’s Scientific Research Foundation of Shandong Jiaotong University (2022YK060).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available in a publicly accessible repository that does not issue DOIs. Publicly available datasets were analyzed in this study. These data can be found here: https://archive-beta.ics.uci.edu, (accessed on 30 September 2022).

Acknowledgments: The authors would like to thank the anonymous reviewers and the editors for their helpful comments. Thanks to UCI for the dataset.

Conflicts of Interest: The authors declare no conflict of interest.

References


34. Liu, G.; Shu, C.; Liang, Z.; Peng, B.; Cheng, L. A modified sparrow search algorithm with application in 3d route planning for UAV. *Sensors* 2021, 21, 1224.

