Research on Gas Identification of Sensor Arrays Based on Sparrow-Optimized for Support Vector Machines

Wanting Wang[1], Yukang Tang[1], Liyang Xi[1], Yihang Wang[1], Tingting Shao[1,2,*]
[1]*School of Physics and Electronic Information, Yan'an University, Yan'an, China*
[2]*Shaanxi Key Laboratory of Intelligent Processing for Big Energy Data, Yan'an, China*
**Corresponding Author*

Abstract: It is difficult to identify multiple volatile organic compounds (VOC) using a single gas sensor, while integrating sensor arrays with machine learning algorithms can classified. The performance of support vector machines (SVM) depends on judicious parameter selection, but it is hard to achieve global optimal solutions using conventional empirical tuning struggles. The Sparrow Search Algorithm was used for global parameter optimization of SVM (SSA-SVM) to enhance gas classification accuracy in sensor arrays. Training and validation were conducted using publicly available datasets from the University of California, Irvine (UCI) Machine Learning Repository. Results demonstrate that compared to the SVM model, SSA-SVM improves gas recognition accuracy from 97.01% to 99.25%. Integrating sensor arrays with the SSA-SVM model provides valuable reference for gas classification recognition systems and holds practical significance for monitoring mixed-gas pollution in industrial or atmospheric environments.

Keywords: Sensor Array; Gas Identification; Support Vector Machine; Sparrow Search Algorithm

1. Introduction

The variety and concentration of volatile organic compounds (VOC) emitted from industrial production and daily life continue to rise, posing significant threats to human health and atmospheric environment. Extensive research indicates that prolonged exposure to VOC may induce respiratory diseases, neurological damage, and even carry carcinogenic risks[1]. Consequently, achieving efficient monitoring and precise identification of VOC in the environment holds critical importance for safeguarding public health, enhancing industrial safety standards, and advancing ecological conservation. In practical applications, gases often exist as multi-component mixtures. The human olfactory system exhibits low sensitivity to various gases, so it is incapable to detect gas generation or leakage accurately. Traditional single-gas sensors can demonstrate high sensitivity to specific gases but suffer from poor selectivity and severe cross-interference when confronted with complex mixed gas environments, thereby it is fail to meet practical application requirements[2]. Sensor arrays composed of multiple gas-sensitive elements can generate different responses to gases, producing signals with distinguishing features that effectively identify different gas types. However, the output data of sensor array typically exhibits high dimension, strong correlation, and non-linear characteristics, making it difficult to fully extract distinguishing information using traditional signal processing methods alone.

With the rapid advancement of machine learning technologies, integrating sensor arrays with intelligent algorithms has become a research focus for enhancing gas classification and identification performance. By constructing data-driven classification models, machine learning algorithms can automatically extract features from complex sensor responses and establish mappings with gas categories. Artificial neural networks (ANNs) demonstrate superior performance in gas classification recognition due to the robust nonlinear fitting capabilities. However, the practical application of these models is frequently hindered by their complicated architectures, long training times, high parameter complexity, and sensitivity to initial settings, which collectively result in substantial computational burdens and a reliance on large-scale datasets[3-4]. Accordingly, grid search method was used to tune the parameters of the Support Vector Machine (SVM) algorithm, then in order to balance overall recognition

accuracy and code execution time, the Sparrow Search Algorithm is selected to further optimize the Support Vector Machine (SSA-SVM), thereby enhancing the gas classification accuracy of the sensor array.

2. Construction and Processing of Datasets

2.1 Dataset Construction

The data employed herein originates from a publicly available data set of the University of California, Irvine (UCI) Machine Learning Repository[5], requiring no additional fees. The UCI Machine Learning Repository serves as a data generator utilized by databases, theoretical domains, and the machine learning community for empirical analyses in machine learning. This data set pertains to gas sensor arrays under dynamic gas mixtures, comprising records from sixteen chemical sensors, each exposed to two distinct dynamic gas mixtures. For each mixture, signals were continuously collected over a 12-hour period. The dataset was acquired using the Gas Delivery Platform located in the Chemical Signal Laboratory at the Institute for Biological Circuits, University of California, San Diego. This measurement system platform is versatile, enabling the acquisition of target chemical substances at desired concentrations with high precision and repeatability.

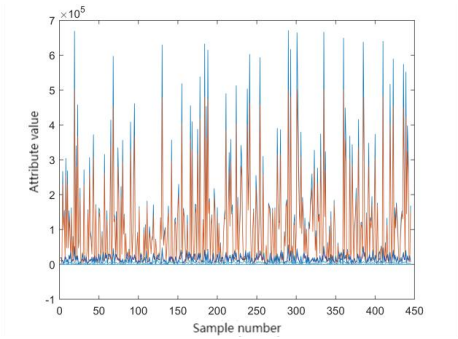
The sensor array comprised 16 chemical sensors (Figaro Inc., USA) of four distinct types: TGS-2600, TGS-2602, TGS-2610, and TGS-2620 (four units per type). The sensors incorporated custom signal conditioning and control equipment. Throughout the experiment, the sensor operating voltage was maintained at 5 V to regulate sensor operating temperature. The sensors' conductivity was continuously acquired at a sampling frequency of 100 Hz. The sensor array was placed within a 60 ml measurement chamber, into which the gas sample was introduced at a constant flow rate of 300 ml/min[6]. Each measurement consisted of uninterrupted signal acquisition from all 16 sensors in the array. For every gas mixture, data were continuously recorded for approximately 12 hours.

2.2 Data Preprocessing

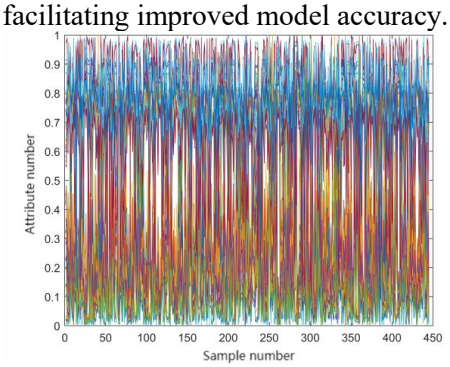
Owing to substantial differences in the response parameters of target gases within the sensor array under dynamic gas mixtures, data normalization is necessary to avoid adverse effects on subsequent algorithm development and prediction accuracy. The preprocessing of the data set and the coding for constructing the recognition model in this paper were completed within the MATLAB environment. The data preprocessing was used the Min-Max Scaling method, employing the MATLAB mapminmax function to normalize the input data according to Equation (1).

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Here, x denotes the raw data, x* represents the normalized data, while x_{min} and x_{max} denote the minimum and maximum values within the raw data respectively. The feature maps of the data raw and after normalization are shown in Figure 1(a) and (b). It can be observed that the raw data exhibits substantial variation, ranging from decimals to hundreds of thousands, which is detrimental to subsequent modelling. The normalized data is confined within the interval [0,1], thereby mitigating the impact of data sets and facilitating improved model accuracy.



(a) Raw data



After data normalization (b)

Figure 1. Data Distribution Chart

2.3 Evaluation Indicators

This paper selects programme execution time and classification accuracy as evaluation indicators of the algorithm, assessing both its predictive process and outcomes. Programme

execution time indicates the code runtime, while accuracy measures the proportion of correctly identified samples among all samples.

3. Gas Identification Algorithm

3.1 Support Vector Machine Algorithm

Support Vector Machines (SVM) constitute a generalized linear classifier employing supervised learning for binary data classification. They were introduced in 1995 by Soviet scholars Vapnik et al[7]. The fundamental model defines the maximum margin linear classifier within the feature space, with the learning strategy centred on maximizing this margin. This ultimately translates into solving a convex quadratic programming problem. SVM applies nonlinear kernel functions to project input data into a high-dimensional feature space, where a linear function with minimal complexity is subsequently learned.

Given the need to classify multiple gases, which is a multi-class classification problem, conventional SVM designed for binary classification, proves inadequate. Hence, the LIBSVM toolbox for algorithm training to address multi-class classification challenges was used in this paper. LIBSVM[8], developed by Chang, Lin et al. from National Taiwan University, is an SVM toolbox with broad applicability across numerous domains, capable of handling classification, regression, and other tasks. Supported in Java, MATLAB, and Python environments, it offers powerful functionality with straightforward operation. During data training, LIBSVM permits customization of numerous parameters.

In constructing the support vector machine, data normalization was performed using MATLAB's mapminmax function. The data set was partitioned into a 7:3 training-to-test split. Fundamental SVM parameters were configured, including the penalty function C and kernel function g. Model training and simulation were conducted using the svmtrain() function. Following training completion, predictions were generated for the test set using svmpredict() as the evaluation function.

SVM demonstrate strong generalization capabilities and adapt well to small sample sizes. However, their classification performance heavily relies on the appropriate selection of parameters. Traditional parameter tuning methods suffer from low computational efficiency and a tendency to converge to local optima. Consequently, the introduction of intelligent optimization algorithms for adaptive parameter tuning in SVM models has increasingly become a research trend.

3.2 Sparrow Search Algorithm

The Sparrow Search Algorithm is an intelligent optimization algorithm that simulates the cooperative search behavior and predator avoidance mechanisms exhibited by sparrows during foraging[9]. This algorithm divides the population into three types of individuals: discoverers, participants, and warning unit. The discoverer leads the population in the search direction, locates high-quality food sources (the range where the optimal solution is located), provides search targets, and assesses dangers at the same time. If predators appear and exceed the threshold, they immediately lead the entire group to evacuate; Followers conduct refined searches, either directly occupying prime positions or foraging in adjacent areas, forming a dual-layer search pattern of “deep central exploration + peripheral hopping” to improve the accuracy and convergence speed of the optimal solution, and participants with high current fitness may also be promoted to discoverers. Warning personnel randomly reset positions to break the population's dependence on local sub-optimal solutions, providing algorithms with an opportunity to escape traps. Through real-time role switching and risk-energy feedback among these three roles, the algorithm cycles through “large-step global exploration, small-step local refinement, and randomly falling into pits and escaping” until converging on the global optimum[10]. The discoverer's position is as follows:

$$x_{i,j}^{t+1} = \begin{cases} x_{i,j}^{t} \cdot \exp\left(-\frac{i}{\alpha \cdot iter}\right), & R_2 < ST \\ x_{i,j}^{t} + Q \cdot L, & R_2 \geqslant ST \end{cases} \quad (2)$$

Here, i represents the index of a sparrow individual, x denotes the new position of the i-th sparrow individual in the j-th dimensional search space, where i is the sparrow's serial number, i=1, 2, ..., n; j represents the dimension , j=1, 2, ..., D; α is a random control parameter, a random decimal generated within the interval (0, 1], used to adjust the randomness of the position update step size. *iter* denotes the maximum iteration count of the algorithm, controlling the decay rate of the exponential term. R_2 is the alert threshold, simulating the probability of sparrows

perceiving "predators/local optimum traps". When $R_2<ST$, the environment is deemed hazardous, prompting the explorer to adjust its search strategy. When $R_2 \geq ST$, the environment is considered safe, and the explorer updates its position conventionally. Q is a random number following a normal distribution, assisting discoverers in escaping local optima to enhance the algorithm's global search capability. L is a D-dimensional all-ones vector (where D is the dimension of the search space, the variables dimension in the optimization problem).

Followers track the discoverer's movements and vie for position. When a follower fails to secure the position, it will fly to another location. Its update formula is as follows:

$$x_{i,j}^{t+1} = \begin{cases} Q \cdot exp\left(\frac{x_{worst}-x_{i,j}^{t}}{i^2}\right), & i > \frac{n}{2} \\ x_p^{t+1} + |x_{i,j}^{t} - x_p^{t+1}| \cdot A^{+} \cdot L, & other \end{cases} \quad (3)$$

x_{worst}^{t} represents the worst position found during iteration t. x_p, at iteration $t+1$, is the optimal position within the discoverer group, guiding joiners to conduct refined searches within the high-quality solution region. A is a random direction vector providing a random direction for position updates of joiners, preventing all searching joiners in the same direction. A^{+} is the pseudo-inverse matrix of the random vector A. n represents the total number of sparrow individuals. When $i>n/2$, it denotes a "losing competitor" whose position is updated using the upper half of the formula. Otherwise, it denotes a "winning competitor" whose position is updated using the lower half of the formula.

The warning unit, comprising 10% or 20% of the sparrow population, conduct investigations and advance notification of hazards. Its position update formula is as follows:

$$x_{i,j}^{t+1} = \begin{cases} x_{best}^{t} + \beta \cdot |x_{i,j}^{t} - x_{best}^{t}|, & f_i > f_g \\ x_{i,j}^{t} + k \cdot \left(\frac{|x_{i,j}^{t} - x_{worst}^{t}|}{f_i - f_w + \varepsilon}\right), & f_i = f_g \end{cases} \quad (4)$$

x_{best} denotes the current global optimum position. β represents a standard normal distribution random number. k is a random number within the range [-1,1]. f_i indicates the fitness value of the current sparrow individual. f_g is the fitness value at the global optimum position. f_w is the fitness value at the global worst position. The minimum parameter ε prevents the denominator from becoming 0. When $f_i>f_g$, the sparrow is vulnerable to predator attacks. When $f_i=f_g$, the sparrow perceives danger and moves closer to other sparrows.

Therefore, the steps of the Sparrow Search Algorithm can be summarized as follows: First, set the initial parameters, including the population size n, the maximum iteration count *itermax*, the proportion of discoverers, the proportion of early warning units, etc. Subsequently, the population is initialized, the fitness values are computed and ranked, and the current global best fitness and position, along with the worst fitness and corresponding position, are determined. According (2), (3), and (4) to update the positions of discoverers, joiners, and warning units, judging whether the maximum iteration count *itermax* has been reached, and outputing the global optimal position and the best fitness value. The flowchart is shown in Figure 2.

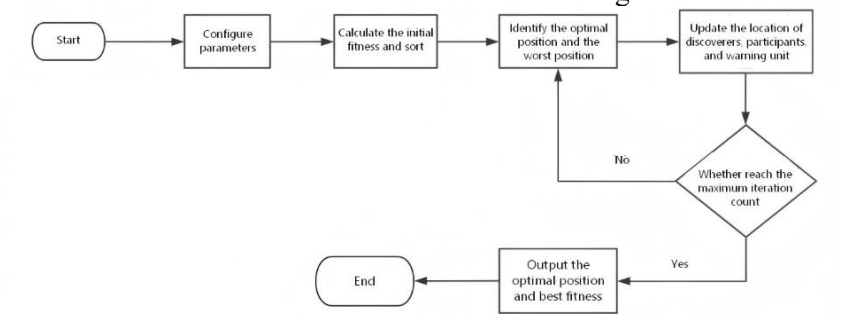


Figure 2. Sparrow Search Algorithm Flowchart

3.3 Optimizing SVM with Sparrow Search Algorithm

In this paper, the sensor array data was normalized before it was be trained. The sensor data and labels are taken as output variables, and the penalty factor C and kernel bandwidth g of SVM are optimized by SSA. The SSA-SVM design flow is shown as Figure 3. The initial parameters of SSA were defined as population size n=30, maximum iteration was 50, discoverer ratio PD is 0.7, warning unit ratio SD is 0.2, and safety value ST is 0.6. In the iterative optimization of SVM parameters C and g using SSA algorithm, the data converges quickly in the 10th iteration, and the fitness value tends to

converge in the 11th iteration. The optimal parameters C is 220.37 and g is 1.29 are obtained.

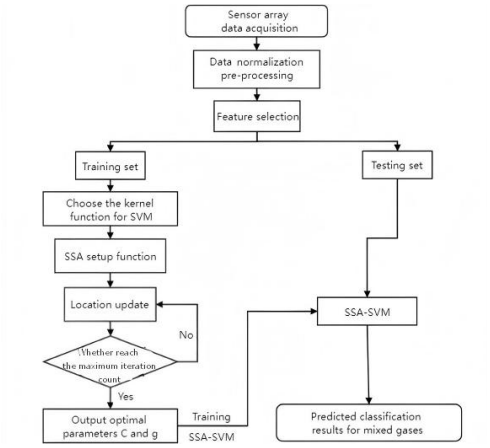
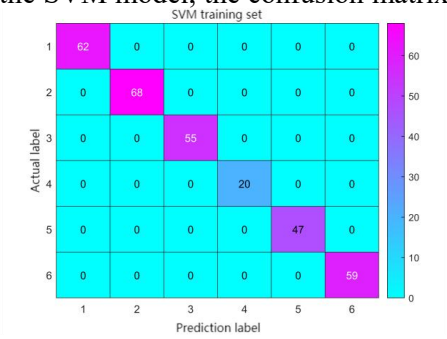


Figure 3. SSA-SVM Design Flow Chart

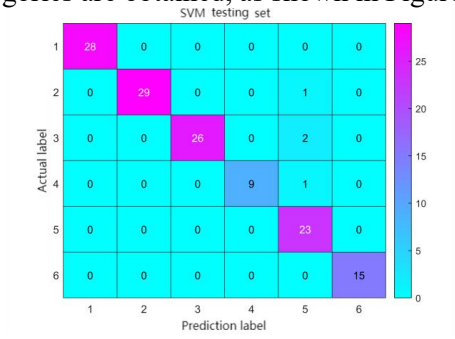
4. Results and Discussion

Using the SVM model, the confusion matrix and classification accuracy of the training set and the test set are shown in Figure 4. It can be found that the prediction results of SVM during the training process are all correct, the prediction category in the classification accuracy line chart completely coincides with the actual category, and the classification accuracy rate is 100%. However, there are four prediction errors in the test set, which classify 1 Class 2, 2 Class 3, and 1 Class 4 gas as Class 5 gas. There is a partial conflict between the predicted category and the actual category in the classification accuracy line chart, and the classification accuracy rate is 97.01%.

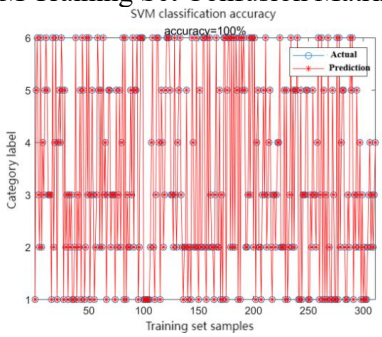
The optimal parameter values C=220.37, g=1.29, which obtained by SSA algorithm, were substituted into the SVM model for regression simulation training. The confusion matrix of the test set and the comparison chart of test categories are obtained, as shown in Figure 5.



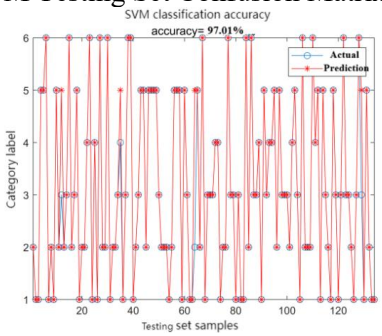
(a) SVM Training Set Confusion Matrix Graph



(b) SVM Testing Set Confusion Matrix Graph

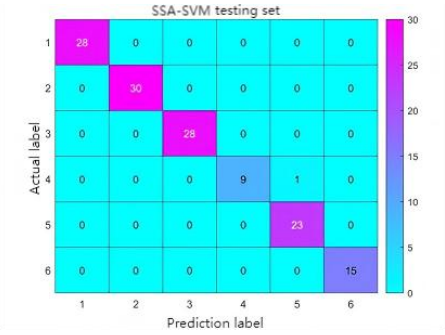


(c) Classification accuracy of SVM training set

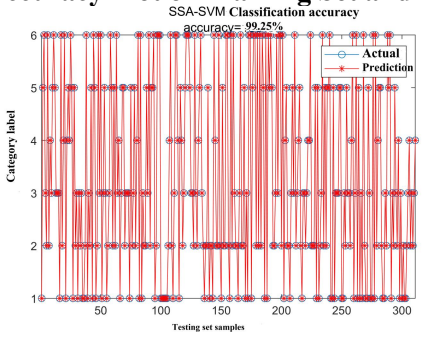


(d) Classification accuracy of SVM testing set

Figure 4. Confusion Matrix Plot and Classification Accuracy Plot of Training Set and Test Set



(a) Testing Set Confusion Matrix Graph



(b) Comparison of test categories

Figure 5. Testing Set Confusion Matrix Graph and Comparison of Test Categories

It can be seen that there are 6 gas categories in the test set samples, and the total number of samples is 134, of which 133 are successful in prediction and 1 is wrong in prediction. The classification accuracy of the model reaches 99.25%, which has a high prediction accuracy.

5. Conclusions

In this study, a gas classification model based on SSA-SVM is developed by employing the sparrow search algorithm to adaptively optimize the parameters of the support vector machine. The experimental results show that the sparrow search algorithm can effectively realize the global optimization of the key parameters of SVM. Compared with the SVM model, the SSA-SVM has significantly improved the accuracy of gas classification and recognition, which verifies the feasibility and effectiveness of combination of the sparrow search algorithm and support vector machine with sensor array. The work in this paper provides a useful reference for building a high-precision and intelligent gas recognition model.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62564014); College Training Program of Innovation and Entrepreneurship of Shaanxi Province (S202410719103).

References

[1] Tang F, Fu J, Xing X, Yan G, Zhang G. Materials for chemical resistance type gas sensors: VOCs identification. Materials Today, 2025, 89: 621-648.

[2] Zhang Y. Design and Optimization of Gas Sensor Based on Machine Learning. Nanjing University of Posts and Telecommunications, 2023.

[3] Liu C. Research on Detection Algorithm of Exhaled Biomarkers in Diabetic Patients Based on Machine Learning. Changchun University Of Technology, 2023.

[4] Song T. Research on the Detection Method of Mixed Gas Based on MOS Gas Sensor Array. Harbin University of Science and Technology, 2022.

[5] Fonollosa, Jordi. Gas Sensor Array under Dynamic Gas Mixtures. UCI Machine Learning Repository. 2015.

[6] Xia Y W. Research on Detection Method of Mixed Gas Based on Sample Unbalance Condition. Harbin University of Science and Technology, 2023.

[7] Vladimir N. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, 1995.

[8] Chang C, Lin C J. Libsvm. ACM Transactions on Intelligent Systemsand Technology, 2011. 2(3): 1-27.

[9] Xue J, Shen B. A Novel Swarm Intelligence Optimization Approach: Sparrow Search Algorithm. Systems Science & Control Engineering, 2020, 8 (1): 22-34.

[10] Li B, Chen Y. Uncertainty-driven portfolio selection via a multi-strategy modified sparrow search algorithm approach. Chaos, Solitons & Fractals, 2025, 201(3): 117349.

