International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 7949-7961

https,//doi.org/10.70517/ijhsa463680

Research on abnormal power consumption detection method for grid users based on support vector machine

Wei Zhang[1,*], Qiong Cao[1], Shuai Yang[1] and Yinlong Zhu[1]
[1] State Grid Shanxi Electric Power Company Marketing Service Center, Taiyuan, Shanxi, 030000, China
Corresponding authors: (e-mail: 18733974318@163.com).

Abstract Traditional methods for abnormal power usage detection of power users cannot effectively deal with complex power usage patterns and sudden abnormalities, resulting in low detection efficiency and poor accuracy. In this paper, a support vector machine (SVM)-based abnormal power usage detection method for grid users is proposed. First, a series of indicators for characterizing abnormal power consumption are constructed by extracting features such as user power consumption changes, power consumption differences, and line losses. Then, the improved K-medoids clustering algorithm is used to preprocess and cluster analyze the power consumption data to filter out the abnormal power consumption data. Finally, SVM was utilized for the classification and detection of abnormal electricity consumption. The experimental results show that after processing the data of 3305 electricity users, the proposed method achieves 99.6% in detection accuracy, which is significantly better than the traditional DT-SVM method and PSO-SVM method (87.6% and 92.5%, respectively). In addition, the proposed method also shows a large advantage in training time, which is only 18.21 seconds, compared with 53.62 seconds for DT-SVM and 45.26 seconds for PSO-SVM, which is a significant efficiency improvement. The experiment verifies the effectiveness and superiority of the method in abnormal power usage detection of grid users.

Index Terms support vector machine, abnormal electricity usage, K-medoids clustering, detection accuracy, grid users, classification

I. Introduction

Various types of losses occur during the transmission and distribution of electricity, i.e., technical and non-technical losses. Technical losses are due to energy dissipation in transmission lines, transformers and electrical equipment [1]. Whereas, non-technical losses are due to malicious theft of electricity by customers, meter malfunction, communication system failure, illegal connections, billing violations, unpaid bills and malfeasance of employees within the power system, i.e., abnormal electricity usage behavior of customers [2], [3]. It has been estimated that the annual economic losses of power utilities worldwide due to illegal customer abnormal power usage have exceeded $25 billion [4]. It has been reported that the annual cost of electricity lost due to abnormal user power usage can fully supply 77,000 households for one year [5]. In the whole power grid, abnormal power usage behavior may lead to overloading of generating units and transformers, and the power grid companies do not really keep statistics on the loads of normal and abnormal power users, so it can affect the power system's power supply quality problems, bring three-phase imbalance problems, etc [6], [7]. For power users, circuit overloading may lead to overvoltage, which affects the performance of user equipment and reduces the service life of the equipment, and there are safety hazards such as electrocution or even death of the operator due to user power theft [8], [9].

Therefore, there is an urgent need for a method that can effectively detect abnormal power usage detection of power users. At present, the grid inspection staff mainly check whether the user's meter is altered and whether the wires are connected privately one by one in the high-loss stations, which is not only time-consuming, but also results in a waste of manpower and material resources. Nowadays, artificial intelligence technology is booming, and combining user data collected by smart meters with abnormal power user detection is the latest research direction. Starting from a large amount of measured user big data, using intelligent technology, we can dig out the different characteristics of normal users and abnormal power users from the power user data, and then establish an abnormal power user detection model, and make targeted arrangements for grid power use. The inspectors will check the meters and lines of the users through special instruments, and then finally determine the abnormal power users [10]-[14].

With the popularization of smart grid and the growth of power demand, abnormal power usage in the power grid has become a serious problem, especially power theft. Power theft not only affects the economic benefits of power companies, but also may cause uneven loads in the power grid and even jeopardize the security and stability of the power system. Traditional methods for detecting abnormal power usage, such as threshold methods and rule-based methods, often cannot cope with the complex and changing power usage behaviors, so there is an urgent need for a more accurate and efficient detection method.

Support vector machines (SVMs), as a common classification tool, have been widely used in many fields, especially in pattern recognition and anomaly detection.SVMs can effectively distinguish different categories of data in high-dimensional space by constructing an optimal classification hyperplane. However, the traditional SVM has certain limitations when dealing with large-scale data, such as long training time and weak data processing capability. In order to solve these problems, this paper combines the K-medoids clustering algorithm with the SVM method, and proposes an improved method for abnormal electricity consumption detection.The K-medoids algorithm has strong data processing capability, and is able to efficiently preprocess and cluster analyze the electricity consumption data, so as to improve the classification efficiency and accuracy of the SVM. The first step is to carry out feature extraction and preprocessing of the user's electricity consumption data, extracting a number of features, including changes in electricity consumption, differences in electricity consumption, line losses, etc., to form a complete data set of electricity consumption features; then, the data are initially screened and clustered by the improved K-medoids clustering algorithm, which provides high-quality training data for the subsequent SVM classification. Through this process, it can effectively distinguish normal electricity consumption from abnormal electricity consumption (e.g., electricity theft), and finally realize accurate abnormal electricity consumption detection. In terms of experimental validation, the electricity consumption data of 3305 power users are selected in this paper, and the advantages of the proposed method in terms of detection accuracy and training efficiency are verified by comparing it with the traditional DT-SVM and PSO-SVM methods. The experimental results show that the proposed method can not only effectively improve the detection accuracy, but also significantly reduce the training time, which is an efficient scheme for abnormal power consumption detection in the power grid.

II. Method

II. A.Extraction of Characteristic Parameters of Abnormal Electricity Consumption by Users

II. A. 1) Indicators of changes in electricity consumption by consumers

Based on the characteristics of electricity consumption change, the trend of decline in electricity consumption of abnormal users over time and the indicator of difference in electricity consumption of users with similar electricity consumption pattern are used as the indicator of change in electricity consumption of users.

(1) Electricity consumption decline rate indicator

When abnormal electricity consumption, the trend of the slope of electricity consumption over time is used to define the decline rate of electricity consumption. Let the period within x days before and after day i be the period of abnormal occurrence of user electricity consumption, and assuming that abnormal electricity consumption is found on day i, the trend rate of change of electricity consumption on day i can be characterized as:

$$k_i = \frac{\sum_{d=i-x}^{i+x}\left(Q_d - \frac{1}{2x+1}\sum_{d=i-x}^{i+x} Q_d\right)\left(d - \frac{1}{2x+1}\sum_{d=i-x}^{i+x} d\right)}{\sum_{d=i-x}^{i+x}\left(d - \frac{1}{2x+1}\sum_{d=i-x}^{i+x} d\right)^2} \tag{1}$$

where, Q_d is the user's electricity consumption on day d . x is the number of days before and after the electricity consumption statistics. k_i is the rate of decline of electricity consumption of users.

The change in the slope of electricity consumption per day is counted and analyzed, and the indicator of the decline rate of electricity consumption in $2x+1$ days is defined as:

$$M = \sum_{d=i-x+1}^{i+x} D(d) \tag{2}$$

where, M is the indicator of electricity consumption decline rate. $D(d)$ is the judgment result of whether there is a decrease in electricity consumption on the d day, and the value of 1 is taken when the rate of decrease in electricity consumption on the d day is less than that on the $i-1$ day, and vice versa is taken as 0 [15].

(2) Electricity consumption difference indicator

In the abnormal power consumption situation power users of similar power consumption law will change, the current user and the user average value of power consumption difference can be expressed as:

$$\alpha(j,d)=\frac{\frac{1}{2x+1}\sum_{d=i-x}^{i+x}\sum_{h=1}^{24}m_j^{d,h}}{\frac{1}{r}\left(\frac{1}{2x+1}\sum_{d=i-x}^{i+x}\sum_{h=1}^{24}m_r^{d,h},\forall r\in O\right)} \tag{3}$$

where, α is the electricity consumption similarity rate. m_j^d is the hourly electricity consumption of the j th user on the d th day. h is the time of electricity consumption. O is the set of similar users. r is the number of similar users with electricity usage pattern.

II. A. 2) Line loss indicators

Consumer electricity consumption is closely related to the line loss rate, and the loss on the line will increase when the consumer experiences abnormal electricity consumption. Assuming that there is N user on a line, the line loss rate of the line on a particular day is:

$$L_L=\frac{Q_L-\sum_{r=1}^{N}Q_r}{Q_L}\times 100\% \tag{4}$$

where, L_L is the line line loss rate. Q_L and Q_r are the line transmission power and user electricity consumption, respectively.

Since there are certain fluctuations in the user's electricity consumption in different time periods, the line loss rate changes accordingly, so to accurately characterize the abnormal electricity consumption behavior by the line loss rate, it is necessary to take into account the rate of change of the line loss rate in the whole time period to characterize the line loss index, which is as follows:

$$(L_{L1}-L_{L2})/L_{L2}\times 100\%>\eta \tag{5}$$

where, L_{L1} is the average of the line loss rate between the day of counting and the previous day. L_{L2} is the average value of line loss rate between the latter $2x-1$ days. η is the actual line loss rate change rate.

II. A. 3) Voltage and current unbalance indicators

The imbalance I_{ub} and U_{ub} of the user's three-phase voltage and three-phase current can be expressed as:

$$I_{ub}=\frac{\max\left(\left|I_a-I_{avg}\right|,\left|I_b-I_{avg}\right|,\left|I_c-I_{avg}\right|\right)}{I_{avg}} \tag{6}$$

$$U_{ub}=\frac{\max\left(\left|U_a-U_{avg}\right|,\left|U_b-U_{avg}\right|,\left|U_c-U_{avg}\right|\right)}{U_{avg}} \tag{7}$$

where, I_a, I_b, I_c are the three-phase line currents. U_a, U_b, U_c are the three-phase line voltages. U_{avg}, I_{avg} are the average values of voltage and current, respectively.

II. A. 4) Indicators of the number of alarms

When the user abnormal power consumption, not only through the above parameter value changes can be judged, but also through the user terminal installed detection equipment generated by the alarm information to judge. Therefore, the number of alarms generated by user terminals in a certain period of time can be counted as an indicator of abnormal power consumption.

II. B.Electricity usage classification based on improved K-medoids clustering

Extracting raw data such as user power and lines from the power grid, due to the relatively small proportion of abnormal data, it is necessary to clean a part of the user data that does not have the risk of power theft to reduce the complexity of the data. At the same time, it is necessary to make up for the missing data generated by meter failure, measurement communication instability and other factors to improve the authenticity of the data. Therefore, this paper utilizes the Lagrangian interpolation method to make up the data, specifically:

$$L_M(a)=\sum_{x=0}^{M}\left(\prod_{y=0,x\neq y}^{M}\frac{a-a_y}{a_x-a_y}\right)b_x \tag{8}$$

where, i and j are the subscript numbers of non-missing and missing values a, b respectively. M is the number of values before and after the missing data. $L_{M(a)}$ is the result after insertion of missing values.

After cleaning the data to construct the abnormal power consumption characterization index, it is necessary to classify the grid data based on the above index, label the unlabeled data, screen out the data with abnormal features, and then send the feature data with labels to the SVM classifier model for power theft detection and confirmation [16]. The abnormal electricity data detection process is shown in Figure 1.

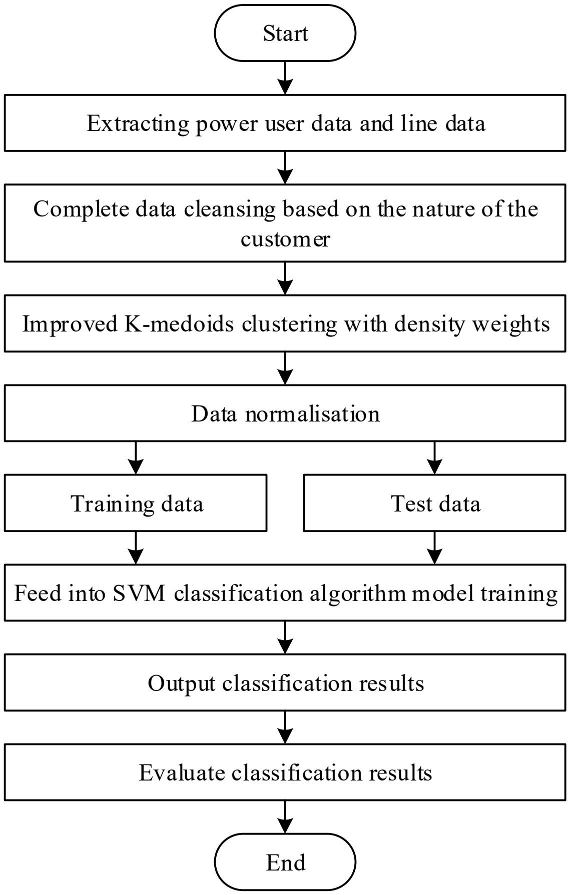

Start
Extracting power user data and line data
Complete data cleansing based on the nature of the customer
Improved K-medoids clustering with density weights
Data normalisation
Training data
Test data
Feed into SVM classification algorithm model training
Output classification results
Evaluate classification results
End

Figure 1: Abnormal data detection process

In order to effectively identify and cluster the anomaly data, this paper adopts the improved K-medoids algorithm with density weight Canopy to screen the anomaly data characterization indexes [17]. The clustering method can be based on the differences in electricity consumption data for preclustering, to get the initial clustering center and the number of clusters, as the initial value of the K-medoids clustering, through this method to effectively improve the accuracy of the clustering of different anomalous electricity consumption data indicators, the specific steps are as follows.

(1) According to the user power consumption change indicator, power consumption difference indicator, three-phase voltage, current imbalance indicator and the number of alarms indicator, assuming that the number of abnormal data clusters that need to be divided into clusters is X, and the number of indicators to be considered for each abnormal data is Y, and the corresponding fuzzy matrix is established as:

$$\begin{cases} X=[a_1,a_2,\cdots,a_X] \\ a_x=[w_{x1},w_{x2},\cdots,w_{xY}] \end{cases} \tag{9}$$

where, $a_1, a_2, \cdots, a_X$ is the number of clusters, a_x is the number of indicators in the i th cluster, and $w_{x1}, w_{x2}, \cdots, w_{xY}$ is the parameter involved in the x th anomalous electricity consumption data.

(2) Calculate the density $\rho_X(w_{xy}, D_{Xw})$ of all data in the electricity consumption data set, ρ_X is defined as the sample region with w_{xy} as the center and D_{Xw} as the radius, and the largest density value of the electricity consumption data $\rho_{X\max}$ is taken as the 1st clustering center, and the expression of the corresponding radius is:

$$D_{Xw} = \frac{2}{X(X-1)} \sum_{x=1}^{X} \sum_{z=x+1}^{X} \sqrt{\sum_{y=1}^{Y} (w_{xy} - w_{zy})^2} \tag{10}$$

where, $x, z \in X$.

(3) Calculate the average distance density value d_X of the power usage data as:

$$d_X = \frac{2}{\rho_x(\rho_x - 1)} \sum_{x=1}^{X} \sum_{z=x+1}^{X} \sqrt{\sum_{y=1}^{Y} (w_{xy} - w_{zy})^2} \tag{11}$$

(4) Classify the different class cluster distances C_X , if the electricity consumption data density $\rho_{X\max}$ corresponding to w_{xy} is the maximum value, defined as $\max d_{xy}$. If the electricity consumption data density $\rho_{X\max}$ corresponding to w_{xy} is not the maximum value, defined as $\min d_{xy}$, $C_X \in [\min d_{xy}, \max d_{xy}]$.

(5) Based on the dataset given in step (1), cluster the dataset by sets, perform steps (2)~(3), and when the clustered dataset is $\varnothing$, end the clustering.

(6) Update the data center with $\sum \min d'_{xy}$ minimum as the new clustering center.

(7)Classify the anomaly data into 3 classes based on the new clustering center using K-medoids clustering.

(8)According to the set minimum threshold ζ_{xy} , when the clustering center is within the running range of the threshold σ_x and the change is small, this center is considered as the clustering center.

(9) Abnormal electricity data determination mechanism, the clustering is completed by forming an array $[a_1, a_2, a_3, \cdots, a_n]$ of electricity data, and the maximum likelihood estimation is used to find the mean value of the array μ and standard deviation of the array σ and according to the 3σ criterion, i.e., $|w_{xi} - \mu| > 3\sigma$ as the threshold to determine whether the outliers become anomalous data.

II. C.Abnormal power usage detection technique based on IALO-DT-SVM

II. C. 1) Support vector machines based on decision tree improvement

Support Vector Machine (SVM) is a more basic and widely used classification algorithm. The data of electricity theft samples are cluttered and belong to the problem of nonlinear irreducibility.SVM for the case of nonlinear irreducibility will generally first transform the data set to the high-dimensional space by the selected polynomial function $\varphi(x)$, and then categorize the data one by one by the constructed decision boundary.

By replacing $\langle x_i, x_j \rangle$ with the kernel function $K(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$, the problem is optimized in the transformed space in the form:

$$\max M(\alpha) = -\frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) + \sum_{i=1}^{l} \alpha_i \tag{12}$$

where $\alpha_i, \alpha_j, y_i, y_j$ are the introduced Lagrange multipliers, and it is also necessary to ensure that $s.t. \sum_{i=1}^{l} \alpha_i y_i = 0$ and $0 \le \alpha_i \le C$ with $i = 1, 2, \cdots, l$. By solving the above problem, we can obtain the optimal classification hyperplane with bounds:

$$f(x) = \sum_{sv} \alpha_i y_i K(x, x_i) + b = 0 \tag{13}$$

where b denotes the displacement term, and in SVM classification, choosing a suitable kernel is a very important. Because the data is essentially nonlinear and the number of features is small, the kernel function needs to be selected to map the data to a high dimensional space. Gaussian Radial Basis Function (RBF) is a commonly used nonlinear kernel, so in this paper RBF kernel function is used and the kernel function can be written as:

$$K(x_i,x_j)=e^{-\frac{\left\|x_i-x_j\right\|^2}{2\sigma^2}} \tag{14}$$

where, x_i is the support vector. x_j is the current data value.

Replacing $1/2\sigma^2$ with γ in the represented radial basis kernel function, the resulting equation is:

$$K(x_i,x_j)=e^{-\gamma\left\|x_i-x_j\right\|^2},\gamma>0 \tag{15}$$

The main goal of this kernel is to construct a decision boundary that divides the training set into two separate parts, the mathematical representation of the decision boundary function of the RBF kernel is:

$$f(x)=w\cdot K(x_i,x_j)+b,w\in x_i,b\in R^m \tag{16}$$

where w is the decision boundary normal vector. b is the regularization parameter in m-dimensional space.

In this section, a binary tree based SVM multilevel classifier is built for power theft detection, and the specific structure of the power theft detection model is shown in Fig. 2.

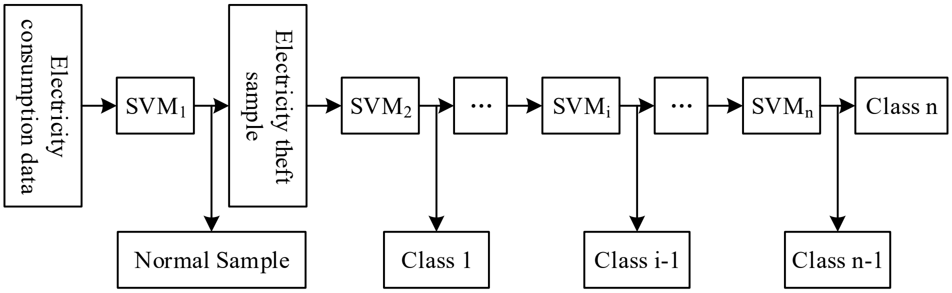


Figure 2: The specific structure of the breakdown model

II. C. 2) Improved Ant-Lion Algorithm

Ant-Lion Optimization Algorithm (ALO) algorithm has stronger global optimization seeking ability than general optimization algorithms, but it also suffers from the problems of premature convergence and local optimality [18]. In this paper, we introduce position update dynamic weight coefficients to try to prevent the algorithm from falling into local optimization, and the main steps to improve the algorithm are as follows:

Step1: Initialize the initial position of ant lions and ant populations, and the number of both populations is set as N, and find the fitness of each individual ant lion by virtue of the fitness function, and the part of ant lions with relatively larger values is defined as the elite ant lions.

Step2: Since the ants are constantly walking, their positions are constantly changing, define new positions:

$$X(t)=\begin{Bmatrix}0,cumsum[2r(t_1)-1],\cdots,cumsum[2r(t_2)-1],\\ \cdots,cumsum[2r(t_{\max})-1]\end{Bmatrix} \tag{17}$$

where $cumsum$ represents the cumulative value, t is the step length of the ant's walk, $t_{\max}$ is the maximum number of iterations, and the random function $r(t)$ is defined as:

$$r(t)=\begin{cases}1,rand>0.5\\ 0,rand\leq 0.5\end{cases} \tag{18}$$

Step3: The lion captures the ants by constructing traps. The smart lion can continuously shorten the area where the ants are walking so that the ants can only be forced to approach the traps in the direction of the traps, and the main relationship is as follows:

$$d_i^t = Antlion_j^t + d^t$$
$$c_i^t = Antlion_j^t + c^t \tag{19}$$

where $Antlion_j^t$ represents the orientation of the j th antlion after t cycles, and d^t and c^t represent the upper and lower bounds of the change of the ant's orientation after t cycles. After continuous loop iterations, the range of the area in which the ants can walk is decreasing, which is defined as follows:

$$c^t = \frac{c^t}{I}, d^t = \frac{d^t}{I}\begin{cases} n = 2, t > 0.1t_{\max} \\ n = 3, t > 0.5t_{\max} \\ n = 4, t > 0.75t_{\max} \\ n = 5, t > 0.9t_{\max} \\ n = 6, t > 0.95t_{\max} \end{cases} \tag{20}$$

where $I = 10^n (t / t_{\max})$, t represents the number of real-time cycles, and $t_{\max}$ represents the number of most cycles.

From $I = 10^n (t / t_{\max})$ and Eq. It can be seen that the range of the area where the ant can walk is decreasing after continuous loop iterations, but the range of the area where the ant walks is fixed during a certain loop iteration, which weakens the algorithm's optimization ability to a certain extent. In order to solve this problem, this paper improves the formula $I = 10^n (t / t_{\max})$ as follows:

$$I = 10^n \frac{t}{t_{\max}}\left(0.5 + \cos\left(\frac{2t\pi}{t_{\max}} \cdot rand\right)\right) \tag{21}$$

where, $rand$ represents the randomly generated value with the value range $[0,1]$, $0.5 + \cos\left((2t\pi / t_{\max}) \cdot rand\right)$ has the value range of $(0.5,1.5)$, and its value change is nonlinear, so the value of I also shows a nonlinear and gradually increasing trend. From Eq. the numerical size of I is inversely proportional to the size of the walking area range of the ants, so that the walking area range of the ants can be continuously reduced, and at the same time improve the algorithm's ability to find the best.

Step4: Re-select new elite ant lions according to the fitness function, and then reconstruct the traps according to the new ant lions that have been selected, defining Eq:

$$Antlion_j^t = Ant_i^t \ if \ f(Ant_i^t) > f(Antlion_j^t) \tag{22}$$

where $Antlion_j^t$ has the same meaning as Eq. while Ant_i^t represents the changing orientation of the i th ant after t cycles, and f represents the fitness function. Through a large number of experiments, it is found that ants usually prefer to choose the walking area around the elite ant lions, which will make the role of the ordinary ant lions more diluted, and also make the algorithm easy to fall into the local optimum, as shown in the following equation:

$$Ant_j^t = \frac{(R_A^t = R_E^t) + R_E^t}{2} = \frac{2R_E^t}{2} = R_E^t \tag{23}$$

where R_A^t represents the area where ants walk near ordinary ant lions, and R_E^t represents the area where ants walk near elite ant lions, in order to solve the above problem, this paper introduces the method of weight proportion coefficient allocation. The improved formula is as follows:

$$\begin{cases} Ant_j^i = \dfrac{\omega_1 R_A^t + \omega_2 R_E^t}{2} \\ \omega_1 = 2 - \omega_2 \\ \omega_2 = 1.9\dfrac{t^3}{t_{\max}^3} \cdot rand \end{cases} \tag{24}$$

where ω_1 and ω_2 are the weighting coefficients, the value range of ω_1 is $[0.1,2)$. At the beginning of the iteration, the value of ω_1 is large, the ants choose to walk around the area near the common ant lion, at this time the algorithm carries out a global search, after many iterations, the value of ω_2 is gradually becoming larger, at this time the ants are more inclined to walk around the elite ant lion, so as to give play to the algorithm's ability to search for the local optimal.

Step5: Determine whether to meet the end of the loop requirements, yes, end, otherwise return to Step2.

II. C. 3) DT-SVM based on IALO optimization

SVM is usually used to divide the data with small sample size, while the sample data size of electricity theft detection is too large, so the simple use of DT-SVM algorithm for electricity theft detection may have the problem of long time-consuming. The optimization of DT-SVM using IALO is based on the idea of optimizing the values of the penalty factor C and the Gaussian radial basis kernel function parameter g through IALO, and optimizing the parameters mainly to improve the accuracy of the model, and at the same time, reduce the training time and improve the efficiency of the implementation of the steps are as follows:

Step1: Initialize IALO, set the number of both ants and ant lions as N, and the maximum number of cycles $t_{\max}$. Also set the SVM parameters.

Step2: Utilize IALO algorithm to perform parameter optimization of SVM.

Step3: Find the fitness of ants and ant lions respectively, and keep the ant lions with the top ranked fitness values as elite ant lions.

Step4: Obtain the optimal ant-lion and calculate its adaptation degree and its orientation.

Step5: Determine whether the termination requirements are met: if yes, terminate the loop, otherwise jump to step 3.

Step6: Optimization ends, obtain the optimal combination (C,g), use the optimal parameters to construct the model.

Step7: Input the test set to detect the classification results of the model.

The flow of IALO optimization DT-SVM algorithm is shown in Figure 3.

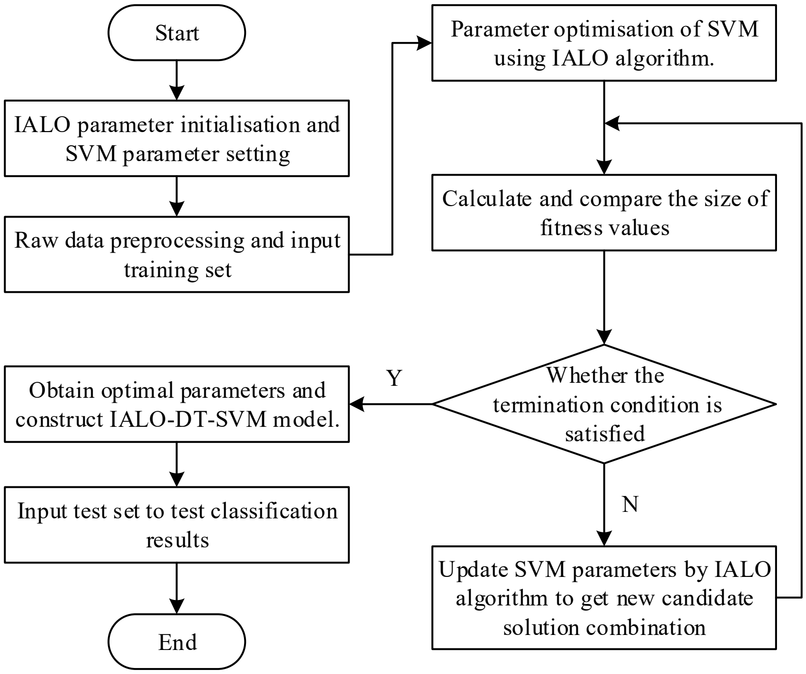

Start
IALO parameter initialisation and SVM parameter setting
Raw data preprocessing and input training set
Parameter optimisation of SVM using IALO algorithm.
Calculate and compare the size of fitness values
Whether the termination condition is satisfied
Y
N
Obtain optimal parameters and construct IALO-DT-SVM model.
Input test set to test classification results
End
Update SVM parameters by IALO algorithm to get new candidate solution combination

Figure 3: IALO optimized the DT-SVM algorithm process

III. Results and discussion

III. A. Classification of electricity consumption by grid users

The experiments in this subsection are conducted in an environment with a processor ADM A8-4555M APU with Radeon(tm) HDGraphics 1.60GHz,4GB of RAM, operating system Windows 10 (64-bit), and Matlab 2015. After data cleaning, missing value interpolation, and normalization of the collected electricity consumption data of 3305 power users, the electricity consumption data of 3082 power users is obtained. Before clustering the electricity consumption data, the validity function $P'(U;c)$ is utilized to determine the optimal number of clusters for the clustering algorithm, and the values of $P'(U;c)$ corresponding to different numbers of clusters are shown in Table 1. From the table, it can be seen that when the number of clusters is 4, the value of $P'(U;c)$ is the largest, so the optimal number of clusters for electricity data in this paper is 4.

Table 1: The different clustering Numbers correspond to $P'(U;c)$ value

Clustering number	2	3	4	5	6	7	8	9
$P'(U;c)$	0.379	0.361	0.533	0.455	0.421	0.409	0.399	0.382

The clustering center curves for the different categories are shown in Figure 4. In this case, the first category contains 1152 electricity users, the second category contains 716 electricity users, the third category contains 531 electricity users and the fourth category contains 683 electricity users.

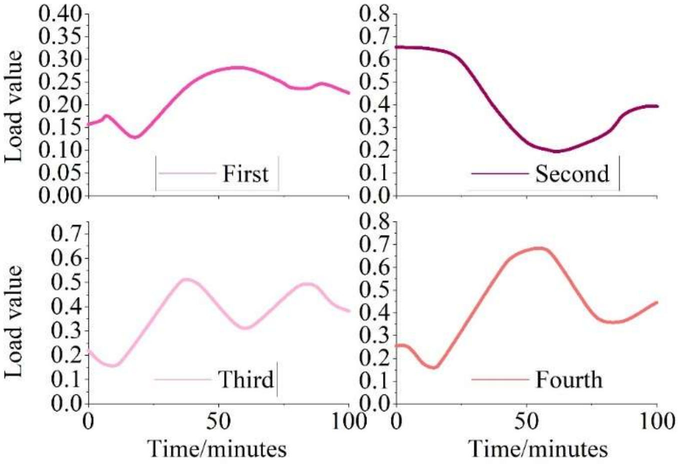


Figure 4: Cluster curve of different categories

As can be seen from the resulting four types of users of electricity consumption characteristics curve, the first type of users throughout the day are relatively low power consumption, night power consumption is slightly more compared to daytime, but the peak load is relatively small, so there should be only small-power appliances at home, this type of user should be the early morning out of the evening return to the office workers. The second type of user power consumption peak in the evening, should be the night-time power users, may be information transmission, computer IT industry workers, the home of high-power appliances are relatively more. The third category for the typical double-peak type of users, and in the morning at eight, nine o'clock and seven, eight o'clock in the evening between the peak load, such users should be living a more stable and regular life from nine to five, and the load value is higher, the home of high-power appliances should be more. The fourth category from nine or ten o'clock in the morning until three or four o'clock in the afternoon are in a higher state of electricity consumption, the evening electricity consumption is lower, so such users should be self-employed, freelancers or the elderly.

III. B. Characteristic dimensionality reduction of electricity consumption data

In this paper, 24 feature indicators are constructed based on the type of feature indicators of electricity consumption data, which reflect the electricity consumption pattern of users in different aspects. These features are as follows: v1-v15 denote the mean, extreme deviation, standard deviation, skewness, and kurtosis of electricity consumption computed over days, weeks, and months, v16 denotes the trend of increasing electricity consumption, v17 denotes the trend of decreasing electricity consumption, v18 denotes the difference between the average of electricity consumption of the user's first r days and the average of the user's second r days, v19 denotes the difference between the user's first r days and the second r days of the fast Fourier transformed coefficient series difference mode, and v20-v24 denote the ratio of daily load rate, daily peak-to-valley difference rate, peak load rate, flat load rate, and valley load rate calculated for the typical load curve of the user and the load characteristic curve. Due to more features of the extracted load data, there will be information overlap between the feature variables and other problems, this paper uses principal component analysis to carry out feature dimensionality reduction on the 24 feature indicator variables extracted.The first 8 principal components of the 4 categories are shown in Table 2.The fragmentation diagrams of the 4 categories are shown in Fig. 5. As can be seen from the table and figure, there are 6 eigenvalues greater than 1 in the first category, with a cumulative contribution rate of 98.179, so the first 6 principal components of the extracted features in the first category are used as the feature set data for anomaly detection. There are 7 eigenvalues greater than 1 in the second, third, and fourth categories, and the cumulative contribution rate is 94.162, 91.105, and 94.13, respectively, so the second, third, and fourth categories all extract the first 7 principal components of the feature as the feature set data for anomaly detection.

Table 2: The top 8 main components of 4 categories

Categories	Eigenvalue	Contribution rate	Cumulative contribution
Categories 1	11.412	47.55	47.55
	3.867	16.113	63.663
	2.574	10.725	74.388
	2.114	8.808	83.196
	1.273	5.304	88.5
	1.152	4.8	93.3
	0.774	3.225	96.525
	0.397	1.654	98.179
Categories 2	9.091	37.879	37.879
	3.332	13.883	51.762
	2.814	11.725	63.487
	2.45	10.208	73.695
	1.722	7.175	80.87
	1.379	5.746	86.616
	1.051	4.379	90.995
	0.76	3.167	94.162
Categories 3	8.19	34.125	34.125
	3.11	12.958	47.083
	2.715	11.313	58.396
	2.443	10.179	68.575
	2.001	8.338	76.913
	1.493	6.221	83.134
	1.059	4.413	87.547
	0.854	3.558	91.105
Categories 4	7.995	33.313	33.313
	3.495	14.563	47.876
	2.989	12.454	60.33
	2.516	10.483	70.813
	1.799	7.496	78.309
	1.54	6.417	84.726
	1.272	5.3	90.026
	0.985	4.104	94.13

III. C. User abnormal power usage detection testing and analysis

III. C. 1) Test parameters

In order to test the performance of the user anomaly detection method, the simulation device is a Huawei PC with an 11th generation core i7 CPU with a main frequency of 2.5 CH and 16 GB of memory, and the simulation platform is the simulation software used for data analysis. A total of 2000 users' electricity consumption data released by a power company for 450 consecutive days (sampling frequency 1h) from 2023 to 2024 are utilized. From the 2000 users, 1200 users are selected as power theft users and their data are replaced. That is, the ordinary user data is converted into eavesdropping users to simulate four actual methods of power stealing (undercurrent method 1, undercurrent method 2, phase-shifting method 3, and spreading difference method 4), 300 users in each case. The parameters of the algorithm are set as follows: ant and ant-lion populations of 50 each, a maximum number of iterations of 110, a shrinkage adjustment factor of 400, and a scaling factor of 20.

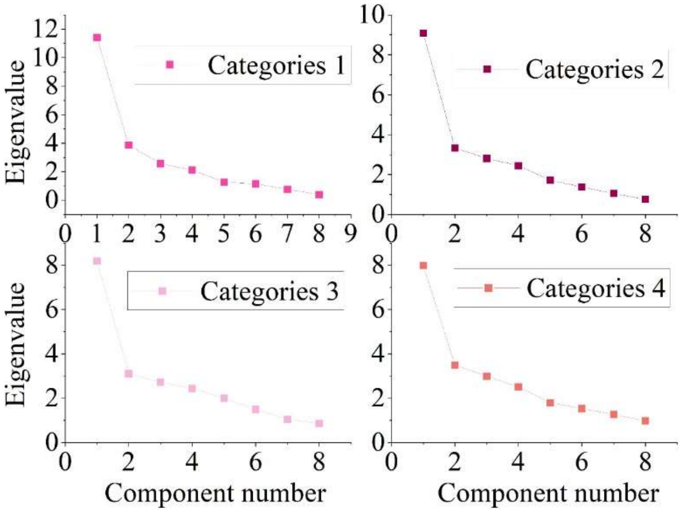
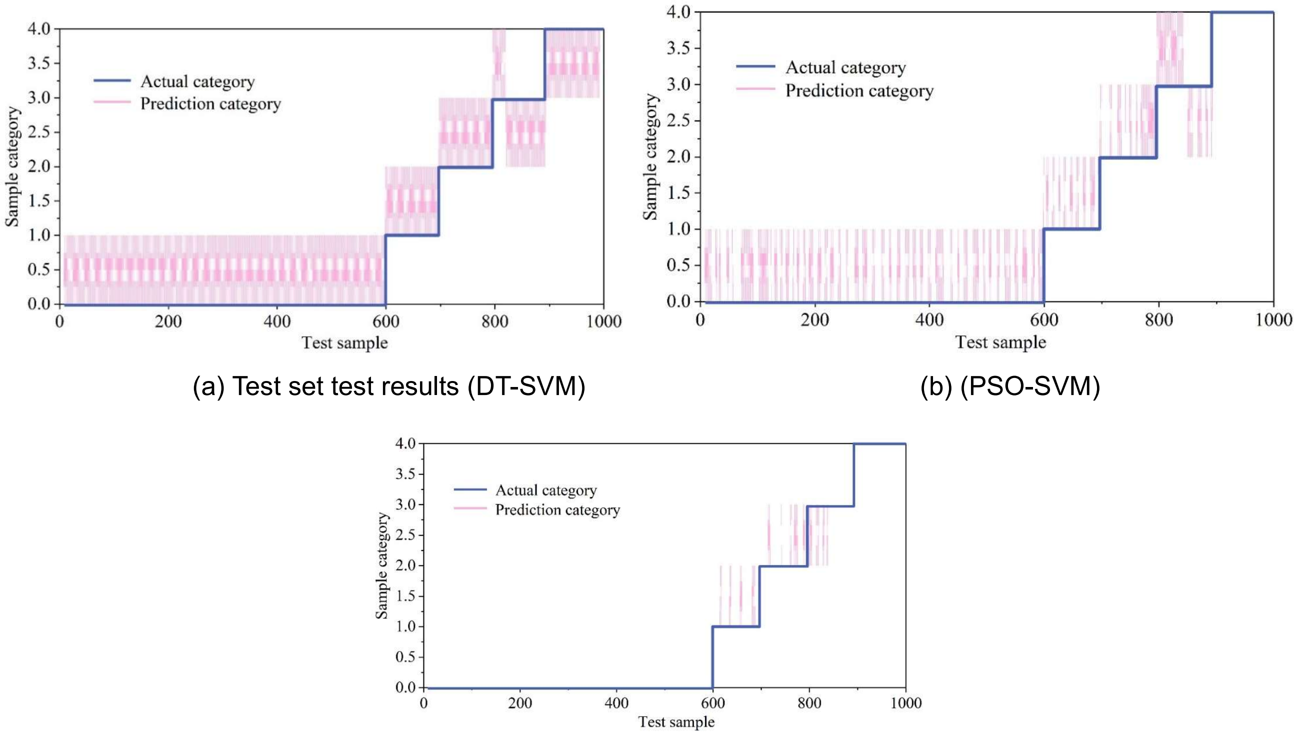


Figure 5: Four categories of rubble

III. C. 2) Test analysis

In order to verify the superiority of the method proposed in the paper, the method is compared and analyzed with the traditional DT-SVM and PSO-SVM detection methods. The DT-SVM and PSO-SVM and the method in the paper are trained by the training set, and the classification effect of different methods is tested in the test set. The comparison of the detection results of different detection methods is shown in Fig. 6 (Figs. a~c show the test set detection results of DT-SVM, PSO-SVM and the method in this paper, respectively). In the figure, users are categorized into normal users, electricity theft users 1, electricity theft users 2, electricity theft users 3 and electricity theft users 4, which are represented by 0, 1, 2, 3 and 4, respectively. From the figure, it can be seen that the detection results of the proposed method in the paper are better than the DT-SVM method and PSO-SVM method.



(a) Test set test results (DT-SVM)

(b) (PSO-SVM)

(c) Test set test results(ours)

Figure 6: Comparison of test results of different detection methods

To test the accuracy of the method, 20 tests were averaged for different methods. The detection results of different methods are shown in Table 3. As can be seen from the table, the proposed method in the paper has the highest detection accuracy of 99.6% in detecting anomalous user behavior, which is 12% better than the DT-SVM detection method and 7.1% better than the PSO-SVM detection method. Also, the proposed method has the shortest training time, which improves the training efficiency. The training time is 18.21 s, which is better than the DT-SVM detection method 35.41 s and better than the PSO-SVM detection method 27.05 s. The superiority of the proposed method in the paper is verified.

Table 3: Test results of different methods

Detection method	Detection accuracy/%	Training time/s
DT-SVM	87.6	53.62
PSO-SVM	92.5	45.26
Ours	99.6	18.21

IV. Conclusion

The support vector machine-based abnormal power usage detection method for grid users proposed in this paper has achieved remarkable results in power user data analysis. By cleaning, interpolating and normalizing the electricity consumption data of 3305 power users, 3082 valid electricity consumption data samples were finally obtained. The experimental results show that using the improved K-medoids clustering algorithm and SVM classifier, the proposed method achieves an accuracy of 99.6% in detecting abnormal electricity consumption behaviors, which is better than the traditional DT-SVM (87.6%) and PSO-SVM (92.5%) methods, and the detection accuracy is improved by 12% and 7.1%. In addition, the training time of the proposed method is only 18.21 seconds, which is much lower than 53.62 seconds for DT-SVM and 45.26 seconds for PSO-SVM, and the training efficiency is greatly improved.

The results show that the improved method combining K-medoids clustering and SVM is able to efficiently perform the detection of abnormal power usage by grid users, and has obvious advantages in accuracy and efficiency. By introducing more power usage features and optimization algorithms, the accuracy and real-time performance of detection can be further improved, providing effective support for the intelligent management of power grids and the prevention and control of power theft.

References

[1] Hussain, Z., Salam, M. H., Shah, S. M., Shah, R. H., Memon, N. A., Memon, M., & Khan, R. A. (2018). Technical Losses Ratio: Analysis of Electric Power Transmission and Distribution Network. International Journal of Computer Science and Network Security, 18(9), 131-136.

[2] Carr, D., & Thomson, M. (2022). Non-technical electricity losses. Energies, 15(6), 2218.

[3] Ahmad, T. (2017). Non-technical loss analysis and prevention using smart meters. Renewable and Sustainable Energy Reviews, 72, 573-589.

[4] Barros, R. M. R., da Costa, E. G., & Araujo, J. F. (2021). Maximizing the financial return of non-technical loss management in power distribution systems. IEEE Transactions on Power Systems, 37(2), 1634-1641.

[5] Qian, Y., Wang, Y., & Shao, J. (2024). Enhancing power utilization analysis: detecting aberrant patterns of electricity consumption. Electrical Engineering, 106(5), 5639-5654.

[6] Kumar, J., Rani, G., & Rani, V. (2025). Energy’s dark side: tracing the impact of electricity theft on power quality deterioration in emerging markets. International Journal of Energy Sector Management.

[7] Cui, J., Fu, T., Yang, J., Wang, S., Li, C., Han, N., & Zhang, X. (2025). An active early warning method for abnormal electricity load consumption based on data multi-dimensional feature. Energy, 314, 134207.

[8] David, J., Ciufo, P., Elphick, S., & Robinson, D. (2022). Preliminary evaluation of the impact of sustained overvoltage on low voltage electronics-based equipment. Energies, 15(4), 1536.

[9] Janthong, S., Duangsoithong, R., & Chalermyanont, K. (2024). Feature Extraction of Risk Group and Electricity Theft by using Electrical Profiles and Physical Data for Classification in the Power Utilities. ECTI Transactions on Computer and Information Technology (ECTI-CIT), 18(1), 51-63.

[10] Nvgui, L. I. N., Lanxiu, H. O. N. G., Daoshan, H. U. A. N. G., Yang, Y. I., Zhixuan, L. I. U., & Qifeng, X. U. (2020). Abnormal electricity consumption behaviors detection based on improved deep auto-encoder. Electric Power, 53(6), 18.

[11] Nabil, M., Ismail, M., Mahmoud, M. M., Alasmary, W., & Serpedin, E. (2019). PPETD: Privacy-preserving electricity theft detection scheme with load monitoring and billing for AMI networks. Ieee Access, 7, 96334-96348.

[12] Bian, J., Wang, L., Scherer, R., Woźniak, M., Zhang, P., & Wei, W. (2021). Abnormal detection of electricity consumption of user based on particle swarm optimization and long short term memory with the attention mechanism. IEEE Access, 9, 47252-47265.

[13] Tang, C., Qin, Y., Liu, Y., Pi, H., & Tang, Z. (2024). An Efficient Method for Detecting Abnormal Electricity Behavior. Energies, 17(11), 2502.

[14] Gou, J., Niu, X., Chen, X., Dong, S., & Xin, J. (2025). Identification of Abnormal Electricity Consumption Behavior of Low-Voltage Users in New Power Systems Based on a Combined Method. Energies, 18(10), 2528.

[15] Zhu Lingkai,Wang Weishuai,Zhang Haijing & Zheng Wei. (2023). Power consumption characteristics of cement industry and parameter analysis of self provided power plant. E3S Web of Conferences,375.

[16] Yimeng Shen,Yiwei Ma,Hao Zhong,Miao Huang & Fuchun Deng. (2025). DTW-based Adaptive K-means Algorithm for Electricity Consumption Pattern Recognition. Engineering Letters,33(1).

[17] Ningtao Liu,Jie Du,Shiliang Chang,Ke Zheng,Ji Xiao,Jiaming Zhang & Feng Zhou. (2024). An automatic diagnosis method of power consumption anomaly of station users based on the k-medoids clustering algorithm. Journal of Physics: Conference Series,2781(1).

[18] Anish Jindal,Amit Dua,Kuljeet Kaur,Mukesh Singh,Neeraj Kumar 0001 & S. Mishra. (2016). Decision Tree and SVM-Based Data Analytics for Theft Detection in Smart Grid.. IEEE Trans. Industrial Informatics,12(3),1005-1016.